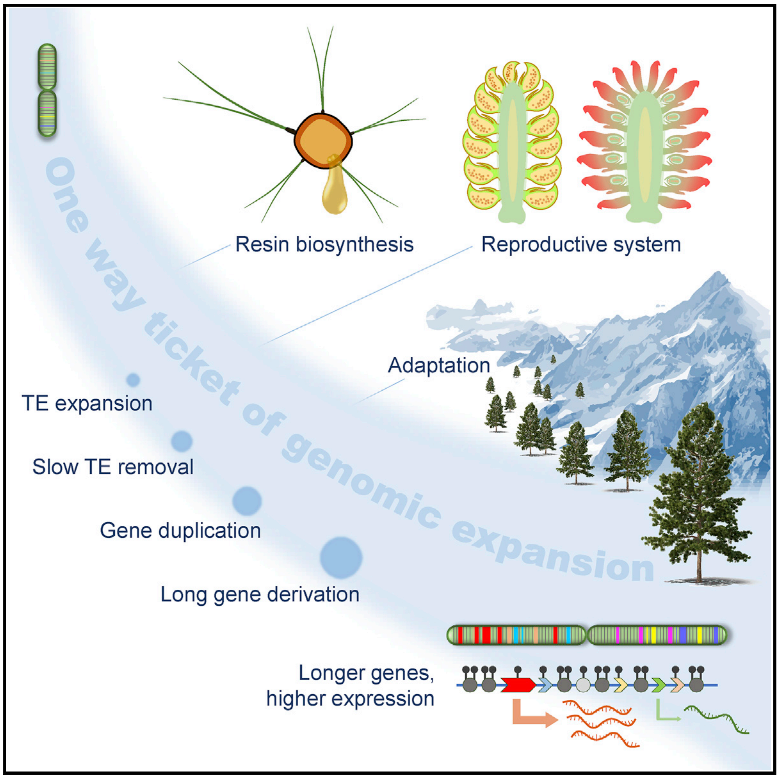
Cell

Resource

The Chinese pine genome and methylome unveil key features of conifer evolution

Graphical abstract



Highlights

- Chromosome-level assembly and methylome of the largest gymnosperm genome so far
- Continuous expansion and slow removal of transposons cause conifer huge genome
- Large genes with ultra-long introns tend to be expressed at higher levels
- Distinctive reproductive evolutionary trajectory compared to angiosperms

Authors

Shihui Niu, Jiang Li, Wenhao Bo, ..., Yue Li, Hairong Wei, Harry X. Wu

Correspondence

arrennew@bjfu.edu.cn (S.N.),
hairong@mtu.edu (H.W.),
harry.wu@slu.se (H.X.W.)

In brief

Assembly of the Chinese Pine giga-genome reveals insights into conifer evolution and provides a resource for studies on conifer adaptation and development.

Niu et al., 2022, Cell *185*, 204–217
January 6, 2022 © 2021 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2021.12.006

CellPress

Resource

The Chinese pine genome and methylome unveil key features of conifer evolution

Shihui Niu,[1,12,*] Jiang Li,[1,12] Wenhao Bo,[1,12] Weifei Yang,[2,12] Andrea Zuccolo,[3,4] Stefania Giacomello,[5] Xi Chen,[1] Fangxu Han,[1] Junhe Yang,[1] Yitong Song,[1] Yumeng Nie,[1] Biao Zhou,[1] Peiyi Wang,[1] Quan Zuo,[1] Hui Zhang,[1] Jingjing Ma,[1] Jun Wang,[1] Lvji Wang,[1] Qianya Zhu,[1] Huanhuan Zhao,[1] Zhanmin Liu,[6] Xuemei Zhang,[2] Tao Liu,[2] Surui Pei,[2] Zhimin Li,[2] Yao Hu,[7] Yehui Yang,[7] Wenzhao Li,[7] Yanjun Zan,[8] Linghua Zhou,[8] Jinxing Lin,[1] Tongqi Yuan,[1,9] Wei Li,[1] Yue Li,[1] Hairong Wei,[10,*] and Harry X. Wu[1,8,11,13,*]

[1]Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, National Engineering Laboratory for Tree Breeding, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Ministry of Education, The Tree and Ornamental Plant Breeding and Biotechnology Laboratory of National Forestry and Grassland Administration, College of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, P.R. China
[2]Annoroad Gene Technology (Beijing) Co., Ltd, Beijing 100180, P.R. China
[3]Center for Desert Agriculture, Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia
[4]Institute of Life Sciences, Scuola Superiore Sant'Anna, 56127 Pisa, Italy
[5]SciLife Lab, KTH Royal Institute of Technology, Tomtebodavägen 23, SE-171 65 Stockholm, Sweden
[6]Qigou State-owned Forest Farm, Pingquan, Hebei Province 067509, P. R. China
[7]Alibaba Group, Hangzhou 311121, P.R. China
[8]Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, Linnaeus väg 6, 901 83 Umeå, Sweden
[9]College of Material Science and Technology, Beijing Forestry University, Beijing 100083, P.R. China
[10]College of Forest Resources and Environmental Science, Michigan Technological University, Houghton, MI 49931, USA
[11]CSIRO National Research Collection Australia, Black Mountain Laboratory, Canberra, ACT 2601, Australia
[12]These authors contributed equally
[13]Lead contact
*Correspondence: arrennew@bjfu.edu.cn (S.N.), hairong@mtu.edu (H.W.), harry.wu@slu.se (H.X.W.)
https://doi.org/10.1016/j.cell.2021.12.006

SUMMARY

Conifers dominate the world's forest ecosystems and are the most widely planted tree species. Their giant and complex genomes present great challenges for assembling a complete reference genome for evolutionary and genomic studies. We present a 25.4-Gb chromosome-level assembly of Chinese pine (*Pinus tabuliformis*) and revealed that its genome size is mostly attributable to huge intergenic regions and long introns with high transposable element (TE) content. Large genes with long introns exhibited higher expressions levels. Despite a lack of recent whole-genome duplication, 91.2% of genes were duplicated through dispersed duplication, and expanded gene families are mainly related to stress responses, which may underpin conifers' adaptation, particularly in cold and/or arid conditions. The reproductive regulation network is distinct compared with angiosperms. Slow removal of TEs with high-level methylation may have contributed to genomic expansion. This study provides insights into conifer evolution and resources for advancing research on conifer adaptation and development.

INTRODUCTION

Forest ecosystems, especially those in boreal and temperate regions, are primarily dominated by gymnosperm trees comprising ~1,000 species (Sederoff, 2013). Phylum Pinophyta, also known as conifers, comprise 615 species contributing 39% of the world's forests (Jin et al., 2021) and serve as backbone components of forest ecosystems.

Conifer forests have adapted to survive in extremely cold and harsh environments (Sander and Meikar, 2009). However, the evolutionary basis of their wide distribution remains elusive. While whole-genome duplication (WGD) played a critical role in adaptive evolution in angiosperms (Wu et al., 2020; Zhang et al., 2020a), few recent WGD events were found in extant gymnosperms (Li et al., 2015) after they diverged from the sister clade of angiosperms about 320 million years ago (MYA) (Smith et al., 2010). The angiosperms now dominate the planet through a species explosion (~300 times more species than gymnosperms) after the divergence. This is partly due to the evolved reproductive advantages, such as flower organ and double fertilization (Sharma et al., 2021). The seed plant reproductive morphology complexity increased in two pulses (400 and 100 MYA) and gymnosperms did not experience the second pulse (Leslie et al., 2021). These indicate that the genome evolution and expansion of gymnosperms may have evolved in a different trajectory from angiosperms with regards to adaptation and reproductive regulatory network.

In the last two decades, more than 1,031 genomes of 788 different plant species have been released and 47 of them have acquired chromosome-scale assemblies as the technologies advance (Sun et al., 2021). However, the genomes of almost all conifers remain poorly assembled, primarily owing to the highly repetitive sequences (70%–80%) and the large genome sizes (17–35 Gb) (Murray, 1998). In addition to assembly obstacles, conifers' iconic long introns also pose great challenges to gene identification and annotation (Warren et al., 2015). Recent RNA sequencing (RNA-seq) studies indicate that the transcriptomes were often substantially underestimated, even in the extensively studied model organisms such as *Arabidopsis thaliana* (Zhang et al., 2020b) and rat (*Rattus norvegicus*) (Ji et al., 2020). Large-scale RNA-seq data could provide direct evidence and resources for a higher-resolution gene annotation (Ji et al., 2020).

To address these issues, we optimized the strategy for genome assembly and annotation, including using large-scale RNA-seq data from 760 biological samples to aid gene identification. With the significant progress in giga-genome assembly and complex gene space annotation, we also unveiled multiple genomic features and molecular mechanisms relevant to genome expansion, reproductive processes, and the adaptive evolution of conifers.

RESULTS

Chromosome-scale assembly of *Pinus tabuliformis* giga-genome

The diploid somatic cells of *P. tabuliformis* have 12 pairs of giant chromosomes, including 20 long metacentric chromosomes and 4 short submetacentric chromosomes (Figure 1A). Based on *k*-mer analysis with 103X coverage short reads, the size of the *P. tabuliformis* genome was estimated to be 25.6 Gb (Figure S1A). The estimated genome size is in an agreement with the DNA content of 25.7 ± 0.13 pg/C that was previously determined with haploid megagametophyte tissue using flow cytometry (Joyner et al., 2001).

The 172 million PacBio long reads (103X) were corrected and assembled using an optimized version of the widely used assembler Canu (Koren et al., 2017). A 25.4 Gb non-redundant assembly was obtained with a contig N50 of 2.6 Mb (Figure 1B; Table 1), which represents the best contiguity among currently released ultra-large (>15 Gb) gymnosperm genome assemblies (Figure 1C). Using 122X coverage reads from nine Hi-C (high-throughput chromosome conformation capture) libraries, a total of 24.4 Gb (96.1%) of the assembled sequences were scaffolded on 12 chromosomes with a super-scaffold N50 length of 2.1 Gb (Figure 1D; Table 1). We measured the relative physical lengths of 12 sets of chromosomes from six somatic cells and found the assembly lengths of all chromosomes were consistent with the observed physical lengths (Figure 1E). These 12 super-long chromosomes ranged from 1.4 to 2.4 Gb (Figure 1F), and only 828 to 1,638 gaps were left in each gigabase chromosome (Table S1).

Large-scale RNA-seq-based high-quality gene space annotation

Based on large-scale RNA-seq data including short reads of 760 biological samples (~40 M reads per sample) spanning across 11 different organs/tissues subjected to various treatments and conditions (Table S3), the medium reads (0.8 M) and long reads of the isoform sequencing (Iso-seq) (2 M circular consensus sequences), 80,495 genes and 144,584 transcripts were annotated, among which 69,599 (86.5%) genes had detectable expression (TPM >1) in at least one sample, and 58,214 (72.3%) genes had orthologs in another 18 selected plants (Key Resources Table). The 22,281 putative species-specific genes, which do not have assigned homologous genes in any of the other 18 species, are functionally associated with biological regulation, stress response, and cell-component organization (Figure S1B).

Our gene-annotation results are consistent with previous studies, in that more genes usually were identifiied in conifers than diploid angiosperms (Mosca et al., 2019). We found that 73,380 (91.2%) genes of *P. tabuliformis* were duplicated, which resulted in the expansion of many gene families (Figure S1C). Classification of the duplicated genes into five different categories (Qiao et al., 2019) revealed that the paralogs in *P. tabuliformis* were mainly derived from dispersed duplication (DSD, 80.7% of genes), and few from WGD (only 0.6% of genes) (Figure S1D).

Lack of recent WGD

We intended to seek evolutionary relics of WGD in *P. tabuliformis* by detecting paralogous synteny gene blocks. However, we only identified 65 blocks and 857 syntenic gene pairs based on all-to-all blastp alignments. We also performed a synteny analysis between *P. tabuliformis* and two other gymnosperm species, *Ginkgo biloba* (Zhao et al., 2019) and *Sequoiadendron giganteum* (Scott et al., 2020), in which chromosome-level genomes have been recently assembled. The result showed that genome reorganization through chromosomal exchanges had occurred during the evolution of gymnosperms; however, only 4% and 2% of homologous in *G. biloba* and *S. giganteum,* respectively, had one-to-more (≥2) synteny orthologous genes in the *P. tabuliformis* genome (Figure S1E). Thus, none of the genetic relics supported a recent WGD event in *P. tabuliformis*.

However, based on the distributions of the synonymous nucleotide substitutions (*Ks*) of syntenic gene pairs, two *Ks* peaks were found (Figure S1F), and both are considered as the ancient WGDs. One ancient WGD event at median *Ks* = ~1.3, is considered as a polyploidization event in the ancestors of all extant seed plants (Leebens-Mack et al., 2019) prior to the divergence of *P. tabuliformis* and *G. biloba* (~320 MYA) (Schneider et al., 2004). After that, there was another ancient WGD event at median *Ks* = ~0.6 in *P. tabuliformis* after the separation of *Pinaceae* and *Cupressaceae* (~260 MYA) (Schneider et al., 2004). Our observation supports the view that independent paleopolyploidy occurred in *Pinaceae* and *Cupressaceae*, respectively (Li et al., 2015).

Unique gene space morphology with multiple long introns

A large number of long introns were observed in the *P. tabuliformis* genome (Table 1). We compared the intron

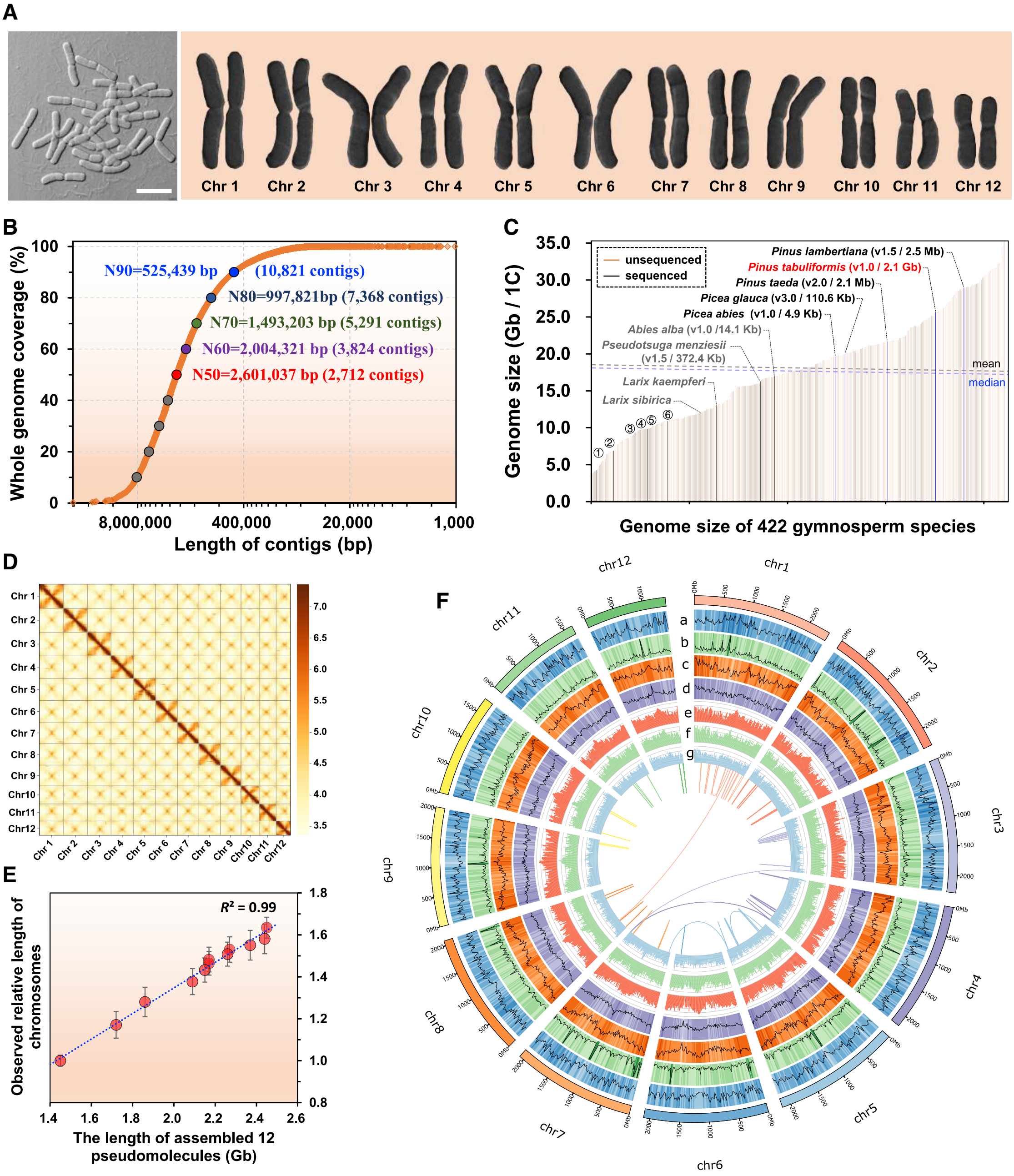


Figure 1. High-quality assembly and genome features of *Pinus tabuliformis*

(A) The karyotype of *P. tabuliformis.* Bar, 1 μm. The separate chromosoms were digitally extracted for comparison.

(B) The statistics for the initial contig assembly.

(C) The genome-size distribution of 422 gymnospers (Table S2); 15 sequenced species are highlighted. The numbers in each pair of parentheses after the species Latin name denote the assembly version/scaffold N50, the mean and median of all known gymnosperms' genome sizes are also indicated. 1–6 refer to *Gnetum montanum*, *Welwitschia mirabilis*, *Sequoiadendron giganteum*, *Ginkgo biloba*, *Taxus wallichiana*, *Taxus chinesis*, respectively.

(D) Twelve pseudomolecules scaffolding with Hi-C data.

(E) The correlation between the assembly lengths and observed physical lengths of all chromosomes. Data are represented as mean ± SD.

(F) Genome features depicted by using 20-Mb-wide bins across the 12 chromosomes. Units on the circumference show megabase values. Track a, gene density (range 5–124 per 20 Mb). Track b, ncRNA density (range 0–200 per 20 Mb). Track c, repeat coverage (48%–88% per 20 Mb). Track d, GC content (35%–42% per 20 Mb). Track e, CG methylation level (84%–96% per 20 Mb). Track f, CHG methylation level (78%–88% per 20 Mb). Track g, CHH methylation (1.7%–2.1% per 20 Mb). Center, curve lines link the syntenic regions that have been retained presumably since the last whole-genome duplication event.

See also Figure S1 and Tables S1, S2, and S4.

Table 1. Assembly and annotation statistics of the Chinese pine genome

Genome assembly	Number of sequences	Total length (bp)	N50 (bp)	N90 (bp)	Longest (bp)
Contigs	22,739	25,421,342,128	2,601,037	525,439	49,421,087
Chromosomes	12	24,405,604,838	2,107,674,557	1,650,012,615	2,364,278,061
Unplaced	7,359	1,017,274,090	153,682	85,662	2,041,681
Final assembly	7,371	25,422,878,928	2,107,674,557	1,650,012,615	2,364,278,061
Number of genes	mean gene length (bp)	mean CDS length (bp)	mean exon length (bp)	mean intron length (bp)	number of introns > 10 kb
80,495	25,170	898	294	10,034	29,883

distribution of *P. tabuliformis* with those of 67 other recently sequenced seed plants (Table S4). We found that all sequenced species share similar average exon length (200–300 bp); however, the length of introns varied greatly. 25,407 (15.4%) introns in *P. tabuliformis* were larger than 20 kb, while few introns of such a length exist in the sequenced angiosperm genomes (Stival Sena et al., 2014). The average of intron length is 10 kb in *P. tabuliformis*, compared with average 0.5 kb in 57 other sequenced angiosperms (Table S4). We found a positive correlation between the ratio of total intron/exon length with the genome's size, especially in the gymnosperm plants (Figure 2A), indicating that the genome expansion not only occurs in the intergenic region but also in the genic region.

We assessed the annotation completeness of *P. tabuliformis* using both BUSCO genome model and protein model. The result showed that BUSCO covered 84% of complete genes in the protein model in contrast to 44.5% in the genome model. We compared the gene sets that could be recognized as complete by both BUSCO protein and genome model (Pc-Gc) with the gene sets that could only be recognized by protein mode (Pc-Gm). We found that most super long genes with multi-introns were not detected under genome model but were recognized in protein model, indicating that multiple long introns are the primary causes of low BUSCO genome completeness (Figure 2B).

We also compared the available proteomes from other sequenced gymnosperm genomes or related pine transcriptomes with angiosperm genomes and observed that the protein annotation completeness of a gymnosperm was always far lower compared with that of angiosperm. Only *P. tabuliformis* and *Gnetum montanum* (4.1 Gb) (Wan et al., 2018) scored more than 80% completeness in this assessment (Figure S2A).

To determine whether these very long introns may be caused by assembly errors, we checked the longest reads for 11,053 ultra-long genes that exceed 20 kb. All of these long genes were supported by at least one ultra-long read (peak at 60 kb). Almost half of them can be assembled by two long reads with 1,914 of which were covered by single reads only (Figure 2C). This suggests that the assembly errors of these ultra-long genes would be very rare. We also manually checked 10 longest full-length genes (>500 kb), which cannot be recognized by the BUSCO genome mode. The solid long-read data support these largest genes in the assembly, and these 10 genes have the similar exon numbers and lengths as their *A. thaliana* homologs, but the intron lengths of these *P. tabuliformis* genes are about 100 times larger than the introns of *A. thaliana* (Figure 2D).

Large genes with long introns are highly expressed

To study whether such extraordinarily long introns would disrupt transcription, we divided genes into two groups by the sizes of the first introns and found that the genes with longer first introns always had relatively higher expression levels in all eleven studied organs/tissues than those with shorter introns (Figure S2B). We further confirmed that the BUSCO complete genes showed a consistent pattern and then analyzed the effects of gene characteristics (intron number and length, exon length, gene with/without introns and TE, total gene length) on gene expression. Many factors that contribute to gene length are positively correlated with higher expression levels (Figure 2E; Figure S2C). Confirmed by regression tree analysis (Loh, 2002), the variance of gene-expression levels can be best and most explained by gene length and intron number (Figure 2F; Figure S2D).

To gain insight into the gene-expression recognition mechanism of small exons from super-long introns in conifers, we manually checked the RNA-junction and DNA methylation patterns of the 10 long genes. Large amounts of RNA-junction data confirm that small exons can be accurately identified and transcribed in a huge DNA that was thousands of times longer than exons (Figure S2E). It is noteworthy that almost all CG and CHG sites in long introns were methylated, whereas exon regions were marked by low methylation levels, especially for the CHG context (Figure S2F), indicating that DNA methylation was probably involved in the accurate exon recognition from super-long introns.

Adaptive evolution of conifers

Based on functional enrichment analysis, we found that the functions of 3,623 significant expanding gene families are mainly associated with biotic and abiotic stress responses (Figure S3A). We manually identified all members of known transcription factor (TF) and transcription regulator (TR) families based on conserved domains and phylogenetic analyses in *P. tabuliformis*. Compared with an average of 1,955 TFs and 390 TRs in the other 197 plant genomes hitherto sequenced, 2,261 TFs and 758 TRs were found in *P. tabuliformis* (Table S5). Based on these resources, we found 188 TFs/TRs that significantly responded to various environmental stresses ($p < 0.01$, fold change >4) including cold (4°C and 10°C), hot (40°C), moderate and progressive drought (Pervaiz et al., 2021), and wound and ultraviolet B (UVB) irradiation (Xu et al., 2021). Notably, among the stress responsive TFs, one-third belong to the AP2/ERF family, which is well known for their functions in abiotic stress responses

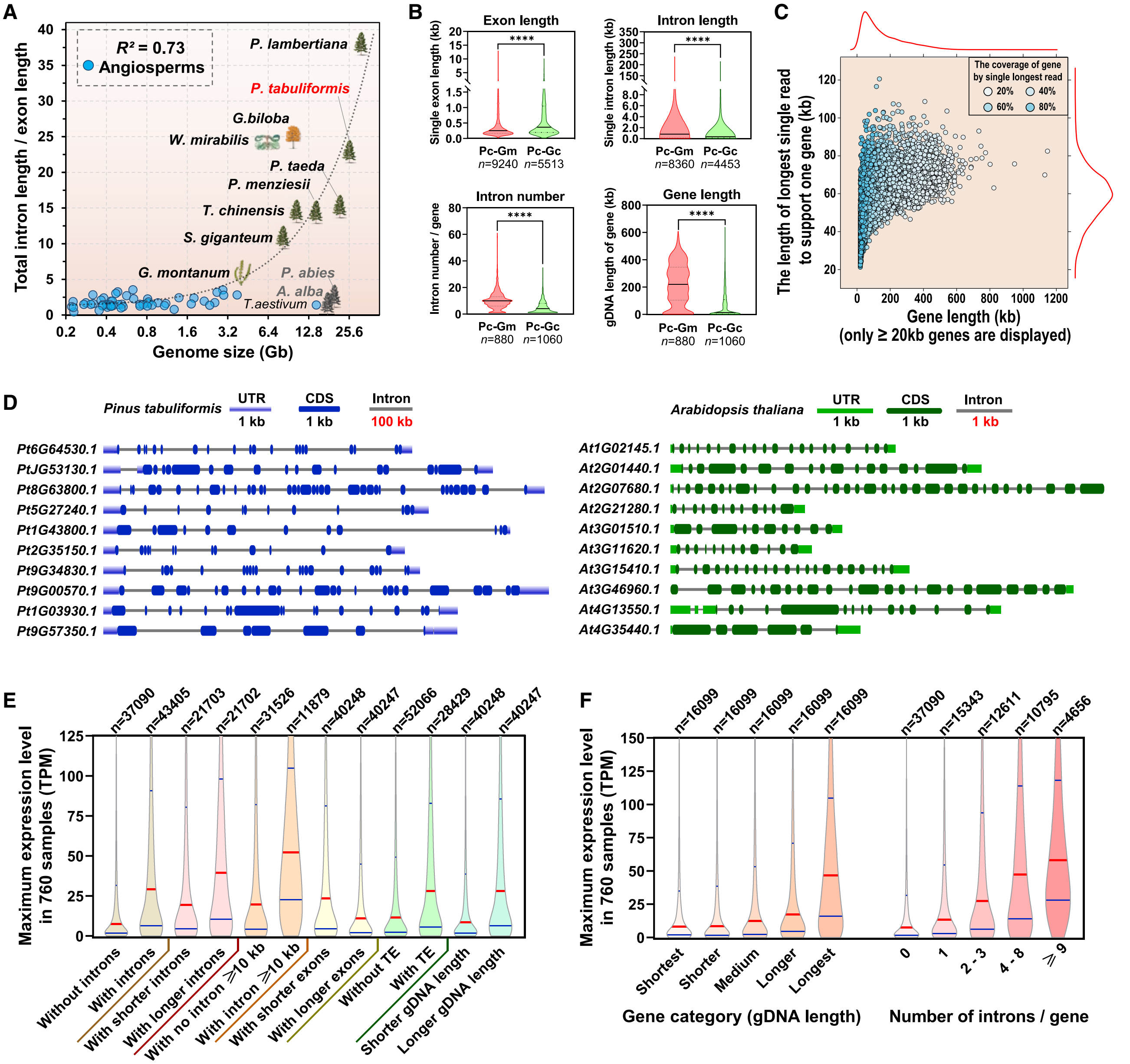


Figure 2. Gene space morphology and complexity in *Pinus tabuliformis* genome

(A) The correlation between the total intron length/total exon length ratio and genome size. 67 randomly selected plant genomes were used for this correlation analysis. The data of *Picea abies* and *Abies alba* were not used to calculate the Pearson correlation coefficient because they were obvious outliers.

(B) The comparison of gene structure between full-length genes that BUSCO recognized and those that were missing in genome mode assessment. Pc-Gm denotes those genes that were recognized in protein mode but were missing in genome mode of BUSCO assessment. Pc-Gc denotes those genes that were recognized in both protein and genome modes of BUSCO assessment. **** refer to $P < 0.0001$.

(C) The longest single reads that supported ultra-long genes (>20 kb). The different colors refer to the percentage of single-read coverage of each gene.

(D) The 10 longest genes that had the consistent exon numbers and lengths as their *A. thaliana* counterparts, but the average intron length of these genes in *P. tabuliformis* is about 100 times that of 10 counterparts in *A. thaliana*.

(E and F) The comparison of expression levels between genes with distinctive structural features.

See also Figure S2 and Table S3.

(Table S5). Nevertheless, there is a lack of C-repeat/DREB binding factors (CBFs) sub-family in *P. tabuliformis* and other conifer genomes (Figure S3B), which are critical for cold acclimation in *A. thaliana* (Zhao et al., 2016) and other angiosperm plants (Shi et al., 2018). However, we found other highly cold-responsive AP2/ERF members such as *PtDREB1* and *PtDREB2* (Figure S3C), which may act as key players in cold acclimation in conifers.

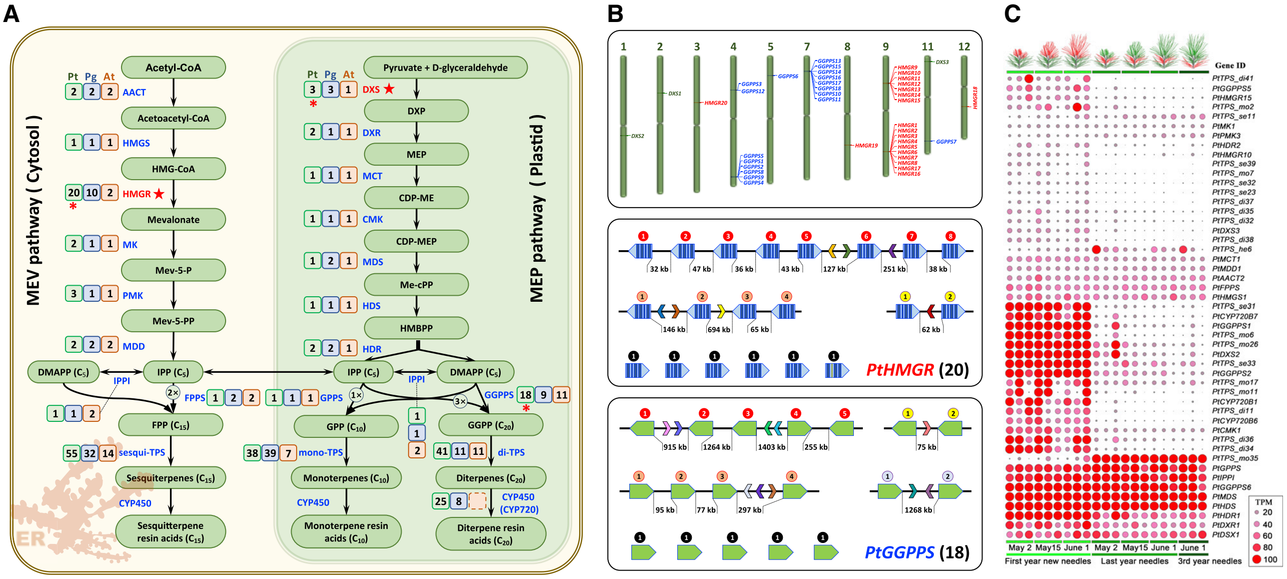


Figure 3. The resin terpene biosynthesis pathways in *Pinus tabuliformis*

(A) The gene numbers of *P. tabuliformis* (green box), *Picea glauca* (blue box), and *Arabidopsis thaliana* (orange box) in the mevalonate (MEV) and methylerythritol phosphate (MEP) pathways. A penta-star represents the rate-limiting steps of isoprenoid biosynthesis. An asterisk denotes the genes that were duplicated in *P. tabuliformis*. AACT, acetyl-CoA acetyltransferase; HMGS, hydroxy methylglutaryl-CoA synthase; HMGR, hydroxy methylglutaryl-CoA reductase; MK, mevalonate kinase; PMK, phosphomevalonate kinase; MDD, diphosphomevalonate decarboxylase; DXS, 1-deoxy-d-xylulose-5-phosphate synthase; DXR, 1-deoxy-d-xylulose-5-phosphate reductoisomerase; MCT, 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase; CMK, 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; MDS, 2-C-methyl-D-erythritol 2, 4-cyclodiphosphate synthase; HDS, 2-C-methyl-D-erythritol 2, 4-cyclodiphosphate synthase; HDR, 2-C-methyl-D-erythritol 2, 4-cyclodiphosphate reductase; IPPI, isopentenyl-diphosphate Delta-isomerase; GPPS, geranyl diphosphate synthase; FPPS, farnesyl pyrophosphate synthase; GGPPS, geranylgeranyl diphosphate synthase; TPS, terpenoids synthase; CYP450, Cytochrome P450.

(B) The distribution of *HMGR*, *DXS*, and *GGPPS* genes in the chromosomes (top graph) and single contigs (middle and bottom graphs). The numbers in middle and bottom graphs refer to the number of genes cluster in a single contig. Genes located on different contig are distinguished by different colors.

(C) The expression profiles of the genes involved in terpene biosynthesis pathways in the needles of *P. tabuliformis*. As evergreen tree needles formed in different years exist on the same branch in *P. tabuliformis*, the red needles represent the part of the branch from which the needle sample was collected and for RNA-seq analysis. Correspondingly, the collection time of each sample is marked at the bottom of this figure. The samples of elongating new needles (no. 1–3 from left), 2-year-old mature needles (no. 4–6 from the left, formed last year) as well as 3-year-old needles (the rightmost one, formed the year before last). The dot sizes and dot colors represent the different absolute expression levels as illustrated by the legend.

See also Figure S3 and Table S5.

The pathway enrichment analysis revealed that the plant-pathogen interaction pathway was the most significant expanded pathway in *P. tabuliformis* (Figure S3D). One example is the terpenoid metabolism related genes in terpenoid biosynthesis pathways. Terpenoid metabolism plays vital roles in defending against pests and pathogens as well as adapting to environmental conditions in conifers (Celedon and Bohlmann, 2019; Liu et al., 2021a). To study the evolution of terpenoid biosynthesis in pines, we identified 221 candidate genes encoding enzymes that catalyze the 22 enzymatic reaction steps of the resin terpene biosynthesis pathway (Figure 3A). All terpenes in conifers are known to be derived from two pathways: the chloroplast methylerythritol phosphate (MEP) and cytosolic mevalonic acid (MEV) pathways (Celedon and Bohlmann, 2019). Most of these steps share similar numbers of catalytic enzyme genes across many different species (Celedon and Bohlmann, 2019; Tholl and Lee, 2011). However, the main rate-limiting enzymes as DXS in the MEP and HMGR in MEV pathway were expanded in pine (Figure 3A). In contrast with only one *DXS* and two *HMGR* genes in *A. thaliana*, *P. tabuliformis* has 3 *DXS* and 20 *HMGR* genes, respectively. Seventeen of the 20 *HMGRs* in *P. tabuliformis* were tandemly arranged in two clusters in chromosome 9 (Figure 3B). Additionally, the 14 of the 18 *GGPPS* genes, which catalyze the last step in the synthesis of the common precursor of all diterpenes, were multiplied in chromosomes 4 and 7 (Figure 3B). Most of these duplicated genes are adjacent on single contigs (Figure 3B). Furthermore, terpene synthase (TPS) family, which catalyzes the universal substrate to various terpenes in plants and fungi (Chen et al., 2016), has expanded considerably to 134 *TPS* genes in several gene clusters of *P. tabuliformis*, much higher than in any of other sequenced plant genomes (Song et al., 2020) (Figure 3A; Figure S3E). Additionally, the cytochrome P450 enzymes (CYP450s), which catalyze the terpenes to resin acids have also expanded considerably (Figure S3F). Most of *TPS* genes are arranged in several respective gene clusters (Figure S3G).

We found that the terpene synthesis related genes had substantially different expression patterns in the needles of different ages. Almost two-thirds of expressed genes were only highly expressed in the first-year new needles, their expression levels then declined sharply in the needles formed 1 year ago (2 year old) and 2 years ago (3 year old) (Figure 3C; Figure S3H), suggesting that new needles are likely the main synthetic sites of terpenes.

The distinct regulatory network for conifer reproductive development

Long-lasting juvenility in conifers is a significant impediment in breeding programs. The termination of juvenility is embodied by flowering. More than 306 flowering-time regulatory genes have been characterized in the model plant *A. thaliana* (Bouché et al., 2016). We identified 77 most conserved orthologs of these flowering time genes in *P. tabuliformis* (Table S5). We then used the RNA-seq data from two groups of 102 samples at reproductive stage to examine whether the roles of these orthologs resemble those in angiosperm species. We did not find significant differentially expressions between different sample groups for these genes (Table S5).

We found that many key regulatory factors, such as *FLOWERING LOCUS T* (*FT*), *FLOWERING LOCUS C* (*FLC*), *FLOWERING LOCUS M* (*FLM*), and *APETALA1* (*AP1*), which represent the most important hub mediators of flowering signals in angiosperms (Blümel et al., 2015), lack orthologs in *P. tabuliformis*. The *FT*/*TERMINAL FLOWER 1* (*TFL1*)-like genes are key flowering integrators in angiosperms (Wickland and Hanzawa, 2015). We identified only two *FT*/*TFL1*-like genes in *P. tabuliformis*, less than the four to six copies reported in other conifers (Liu et al., 2016; Nystedt et al., 2013). We overexpressed both of them in *A. thaliana* and the transgenic plants displayed a substantially late flowering phenotype, suggesting that both of them resemble *TFL1-like* gene in functions (Figure S4A).

The MADS-box TF family has been extensively studied in angiosperms for their crucial functions in reproductive development (Gramzow et al., 2014). We identified a total of 74 *MADS-box* genes with an average length of 70 kb (Figure S4B). All 74 *MADS-box* members had detectable expression (TPM >1), and 71 had highly expressed (TPM >10) in at least one sample. The MADS-box family in *P. tabuliformis* lacks FLC and FLM sub-clusters, which are key negative regulators in vernalization and ambient temperature pathway of angiosperm (Figure S4C). In contrast, we identified 24 SUPPRESSOR OF OVEREXPRESSION OF CO 1 (SOC1)-like proteins in *P. tabuliformis* (Figure S4C). However, overexpression of the *SOC1-like* genes from *P. tabuliformis*, which are close to the *A. thaliana AtSOC1* on a phylogenetic tree failed to complement the late flowering-time phenotype of the *soc1* mutants of *A. thaliana* (Ma et al., 2021), so whether there are conservative function *SOC1-like* genes in conifers is still an outstanding issue.

We conducted yeast two-hybrid assays among 12 highly expressed *MADS-box* genes during the male and female strobili development (Figure 4A) and provided a 12 × 12 pairwise protein-protein interaction (PPI) matrix map (Figure 4B; Figure S4D). Our detailed expression data showed that the two *AGL6-like* genes, *PtDAL1* and *PtDAL14*, had distinct expression patterns in *P. tabuliformis*, and the *PtDAL14* specifically expressed in reproductive organs and widely interacted with other MADS proteins, indicating it may served as bridges between other MADS-box TFs to form a heteromultimer PPI network (Figure 4B). In addition, we found *PtDAL10*, which is highly expressed during reproductive organ development, was also widely interacted with other MADS-box proteins (Figure 4B). The overexpression of *PtDAL10* in *A. thaliana* induced extremely early flowering (Figure S4E), suggesting it might function as a key player involved in the regulatory network for conifer reproductive development. Based on these findings, we have proposed a model to control the development of male and female strobili in *P. tabuliformis* (Figure 4C), which provides a blueprint for future research on conifer reproductive development.

Chromosome-scale methylation landscape largely shaped by high TE content

The methylation level of the genome at chromosome level was significantly correlated with TE coverage in *P. tabuliformis* (Figure S5A). The global average methylation levels of mCG, mCHG, and mCHH in *P. tabuliformis* genome were 88.4%, 81.6%, and 2.0%, respectively (Figure S5B). Both the mCG and mCHG methylation levels are much higher than those in any other previously studied plant species (Ausin et al., 2016; Wang et al., 2019).

The DNA methylation was substantially reduced at the gene transcriptional start sites (TSSs) and the end sites (TESs) (Figure 5A). To investigate the effect of TEs on methylation in genic regions, we separated the genes having TEs inserted in their introns, defined as the TE-bearing genes, from the non TE-bearing genes. We found that the average methylation levels in the genic regions of TEs-bearing genes were much higher than those in the non TE-bearing, but the average methylation levels at TSS and TES regions were always maintained at the equivalent low levels (Figure 5A).

The genic methylation was previously considered to be not involved in the gene-expression regulation in conifers (Ausin et al., 2016). However, by dividing genes into six groups based on their expression levels, we revealed a clear negative correlation between methylation and expression, which is more evident in proximal upstream and downstream regions (Figure 5B). We then divided non-expressed genes identified in the new shoots into two groups, one including genes expressed in at least one of other organs/tissues, the other containing genes with their expression not detectable in any organs/tissues tested so far. We found there was still a moderate methylation reduction at TSS and TES sites of the genes in the first group but such a reduction was substantially lower than the second group (Figure 5B). These results suggest that the methylation level serves as a regulatory constraint on gene transcription in conifers.

Recent burst of LTR-RT and robust methylation maintenance system in *P. tabuliformis*

Based on the *de novo*-constructed species-specific TE library, 69.4% of the *P. tabuliformis* genome was masked. The most prevalent class of TEs is long terminal repeats retroelements (LTR-RTs), which occupied 60.0% of the genome. The *Ty3*/*Gypsy* and *Ty1*/*Copia* elements were the two main classes of LTRs, accounting for 33.6% and 13.5% of the *P. tabuliformis* genome, respectively. Most of TEs were heavily methylated (over 80%) in both mCG and mCHG contexts (Figure 5C), supporting the previous conclusion that DNA methylation plays an influential role in TE-driven genome expansion (Fedoroff, 2012; Zhou et al., 2020). The detailed analysis of LTR-RT insertion

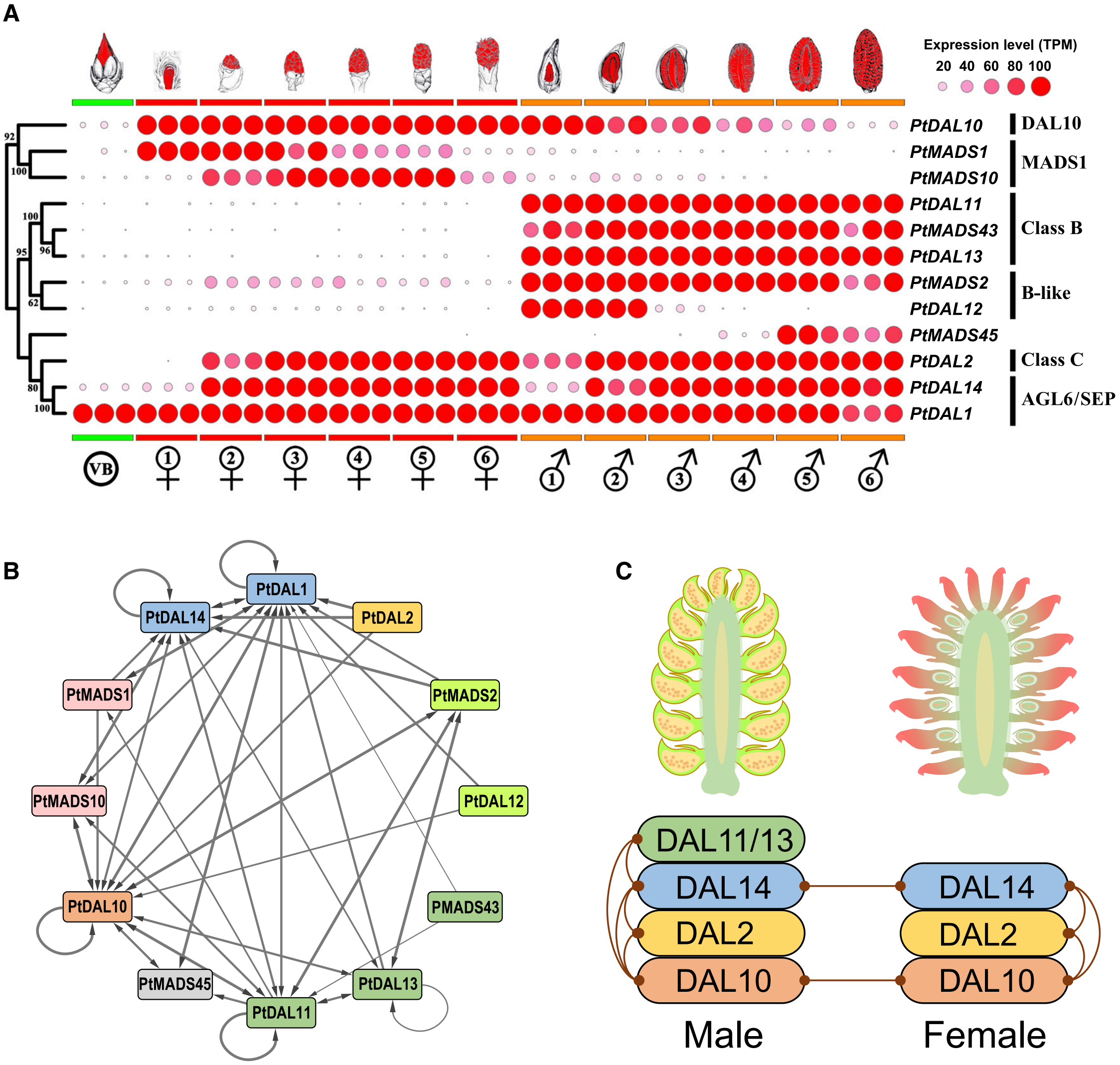


Figure 4. The expression and protein-protein interaction patterns of 12 MADS-box family transcription factors involved in floral development in *Pinus tabuliformis*

(A) The expression profiles of 12 MADS-box TFs involved in reproductive development in developing male and female strobili of *P. tabuliformis*. We divided female and male strobili development into 6 stages to study the ontogeny of conifer reproductive development. VB, vegetative bud; ♀, female cones; ♂, male cones. The numerical order indicates progressing developmental stages.

(B) The protein-protein physical interaction network of 12 MADS-box TFs in *P. tabuliformis* based on yeast two-hybrid assays (Figure S4D).

(C) The proposed model sketching the regulation of the development of male and female strobili in *P. tabuliformis*.

See also Figure S4 and Table S5.

time by Kimura distance method (Kimura, 1980) found that a high proportion of TEs were inserted more recently (within 6 MYA) in *P. tabuliformis* (Figure 5D), after the speciation of almost all pines (Jin et al., 2021). The phylogenetic analyses of both *Ty3/Gypsy* and *Ty1/Copia* superclasses showed many species-specific clades characterized by short branches suggestive of recent TE amplification (Figure 5E).

A negative correlation between the TE insertion time and DNA methylation level was observed in many small angiosperm genomes (Wang et al., 2019; Zhou et al., 2020); however, we did not find evidence that the methylation declined as the insertion age increased in *P. tabuliformis* (Figures 5C and 5F). The LTR-RTs represent the majority of TEs and the unequal recombination (UR) is a major LTR-RT removal mechanism in plants (Cossu et al., 2017). The UR between LTRs leads to removal of intervening sequence and formation of solo-LTRs. We estimated the relative rates of LTR-RT-associated UR within the *P. tabuliformis* genome based on our previously established method (Cossu et al., 2017). We manually examined the three highly abundant classified LTR-RT elements (two *Ty3/Gypsy* and one *Ty1/Copia*) in a 6.1 Mbp contig. As a result, 46 complete, 19 partial complete and 9 solo-LTR elements were identified. These figures point to a ratio of a complete element to solo-LTR that ranges from 5.1 to 7.2, while this ratio in

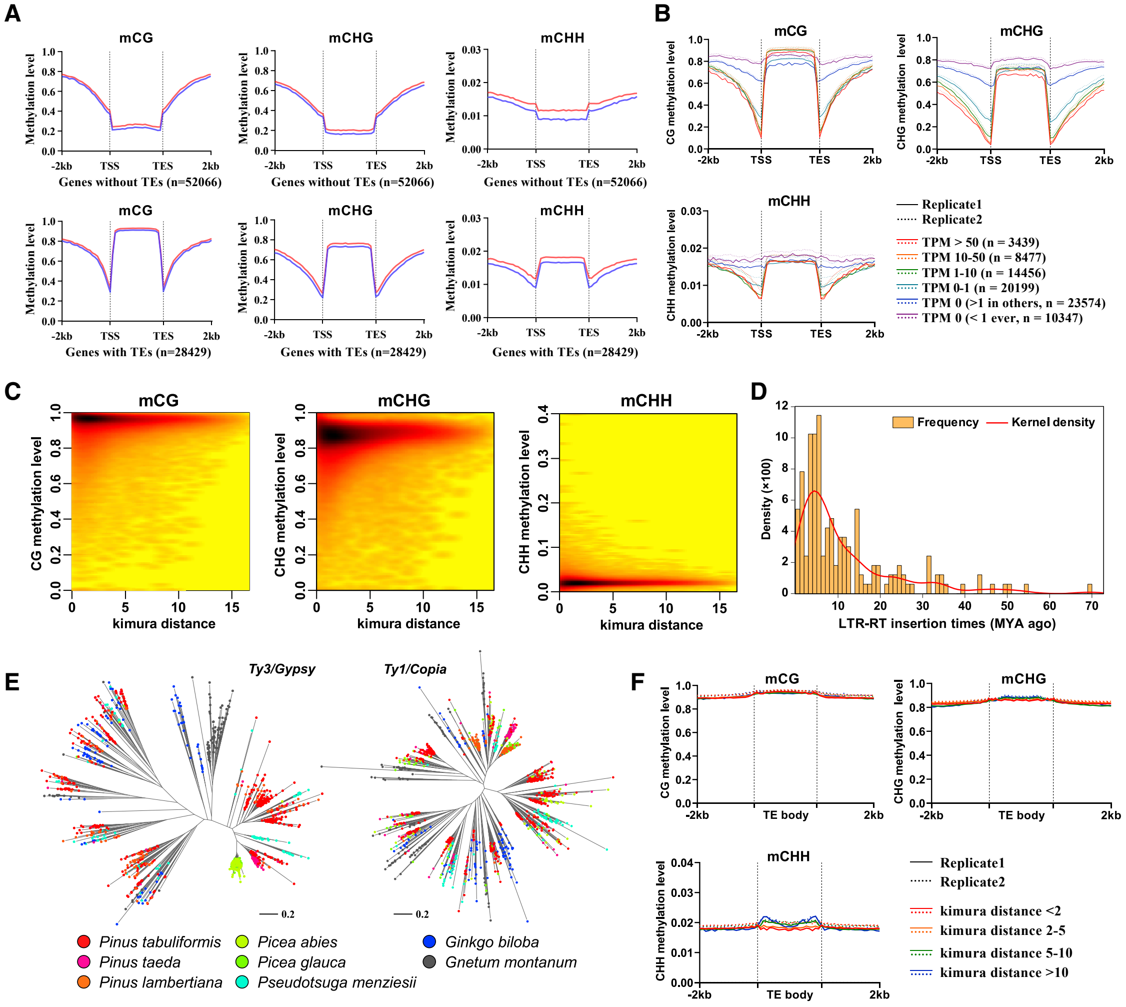


Figure 5. DNA methylome landscapes and transposable element expansion in *Pinus tabuliformis*

(A) DNA methylation patterns of gene bodies and flanking regions in the mCG, mCHG, and mCHH sequence contexts for genes with or without TE-like sequences. The red and blue line represent two replicates.

(B) DNA methylation comparison between gene sets that had different expression levels.

(C) DNA methylation landscape of TEs with different insertion ages. Kimura distance was used as a proxy for TE age.

(D) Distribution of insertion times calculated by manually annotated LTR-RTs in *P. tabuliformis* using mutation rates of 2.2 × 10^{-9} (per base per year).

(E) The phylogenetic trees of LTR-RTs similar to the *Ty3/Gypsy* and the *Ty1/Copia* reverse transcriptase domains from eight gymnosperm species. Heuristic neighbor-joining trees constructed from 500 sequences of *P. tabuliformis* and 100 sequences from each other species.

(F) DNA methylation comparison among TEs of different age groups at the whole-genome level.

See also Figure S5.

angiosperm of small genome is much lower, such as 1.2 in *A. thaliana*, 0.56 in *O. sativa* and 1.1 in *P. trichocarpa* (Cossu et al., 2017). These results indicate the conifer may have a much lower UR rate compared with small genome angiosperms.

For the mechanisms of *de novo* and maintenance of DNA methylation in conifers, the abundant 21- and 22-nt small RNA (sRNA) fragments (Figure S5C) and the expressed *PtSGS3* and *PtRDR6* in most organs/tissues of *P. tabuliformis* indicate that the SGS3-RDR6-RdDM (RNA-directed DNA methylation) pathway (Kim et al., 2021) probably is the primary DNA methylation pathway in conifers (Figure S5D). This contrasts with the well-elucidated 24-nt RdDM pathway that seemingly plays a primary role only in angiosperms (Ma et al., 2015).

DISCUSSION

Giga-genome with TE expansion and methylation

Conifers are renowned for their enormous genomes, especially in pines, which have even larger genome sizes (17–35 Gb) than other gymnosperms (Murray, 1998). Here, we decoded the 25.4 Gb genome of *P. tabuliformis*, which represents the best contiguity among the released ultra-large (>15 Gb) genomes to date (Meyer et al., 2021; Nowoshilow et al., 2018; Wang et al., 2021b). Its giant genome is mostly attributable to huge intergenic regions (93.2%) with high repeat TE content (69.4%). We found a continuously ancient accumulation of TEs with a most recent burst about 4–6 MYA. Although this burst is younger than the estimates of previously released draft gymnosperm genomes (Liu et al., 2021b), *P. tabuliformis* retained more ancient TEs than most angiosperms, as 74 sequenced angiosperm genomes had a median TE insertion time of 2.4 MYA with many less than 1 MYA (Wang et al., 2021a).

The giant genome of conifers raises a striking question: how can conifers endure or bear such large amount of ''parasitic sequences'' intergenic and within genes? We revealed that all TEs in *P. tabuliformis* were consistently targeted by a robust surveillance and methylation maintenance system regardless of insertion age. The epigenetic silencing system was likely to represent a nuclear defense system that had evolved precisely to ''control'' the destructive potential of ''parasitic sequences'' (Fedoroff, 2012; Zhou et al., 2020). Nevertheless, we found that with the genome expansion, not only the intergenic regions, but also the genic regions became less compact. A lot of ultra-long genes were derived by the insertion of TEs in the introns. Noticeably, as in angiopserms, a 5.8 kb T-DNA element inserted in an intron is sufficient to knock out the expression of a target gene (Rosso et al., 2003), but the average intron length of *P. tabuliformis* exceeds 10 kb. We found the boundaries between exons and long introns are marked by elevated cytosine methylation, especially at CHG contexts. Although DNA methylation was originally thought to only affect transcription, emerging evidence shows that it also regulates splicing (Lev Maor et al., 2015). Our observation indicates that DNA methylation probably enabled the accurate exon recognition from super-long introns. Extraordinarily long introns did not show any detrimental effects for the transcription and splicing, on the contrary, the larger genes tend to have a higher expression levels. Interestingly, in the study of African lungfish (*Protopterus annectens*), which also has a giga (40 Gb) genome (Wang et al., 2021b), we noticed that longer genes also have a relatively higher expression level in both brain and lung tissue (Figure S2B in Wang et al., 2021b). These observations indicate there may be a similar strategy adopted to keep transcription efficacy of ultra-long genes in both plants and animals. The precise control of parasitic TEs and efficient transcription systems may result in giga-genomes that are still highly functional.

Compared to conifers, angiosperms had much larger diversity in genome size, but the size distribution is strongly skewed to small genomes with a modal value of just 0.6 Gb (Dodsworth et al., 2015). Why do the conifers collectively have a huge genome? The TEs in angiosperms have a half-life about 3–4 MYA (El Baidouri and Panaud, 2013); however, the half-life of LTR-RTs in conifers is much longer than that (Nystedt et al., 2013). In this study, we observed a considerably lower UR-mediated LTR removal rate in *P. tabuliformis* compared with small-genome angiosperms (Cossu et al., 2017) and gymnosperms (Wan et al., 2021). The DNA methylation levels in *P. tabuliformis* and another conifer genome (Ausin et al., 2016) were very high (76%–88%). The large, repeat-rich conifer genomes may become locked down by epigenetic silencing, reducing the frequency of repeat removal (Fedoroff, 2012; Kelly et al., 2015). Maintenance of high methylation levels in ancient TEs regardless insertion ages in *P. tabuliformis* seems to support this hypothesis.

The adaptative evolution of conifer

Having been noted for their strong adaptability to harsh environments (Sander and Meikar, 2009), conifers generally have large distribution areas covering multiple climate zones and dominating boreal and cold climates in the Northern hemisphere (Sederoff, 2013). The diversification of gymnosperms is mainly associated with increased rates of climatic occupancy evolution, particularly in cooler and/or more arid climatic conditions (Stull et al., 2021). Consistent with most gymnosperm genomes (Li et al., 2015), *P. tabuliformis* lacks recent WGD, which play important roles in microevolution and adaptive evolution in angiosperms (Wu et al., 2020; Zhang et al., 2020a). However, the large-scale gene duplications (i.e., 91.2%) caused gene redundancy that may played the similar roles as WGDs in adaptive evolution. Genes under diversifying selection often show overrepresentation in responses to biotic and abiotic stresses (Van de Peer et al., 2009), and in *P. tabuliformis*, the significantly expanded gene families were also dominated with genes involved in biotic and abiotic stress responses.

TFs/TRs are key regulators that perceive environmental cues and control various developmental processes (Song et al., 2016). *P. tabuliformis* had a larger number of TFs/TRs than the average of 197 plant species, and 188 of these TFs/TRs had significant responses to drought, cold, and other stresses. In particular, 70 TFs/TRs were associated with cold tolerance, which is consistent with the extreme adaptation of conifers to cold climates.

We revealed that the plant-pathogen interaction pathway was the most significant expanded pathway in *P. tabuliformis*. Terpenoid metabolism plays vital roles in defending against pests and pathogens and adapting to environmental conditions in conifers (Celedon and Bohlmann, 2019). For example, *P. contorta* trees with larger resin ducts survived the attack of mountain pine beetle better (Zhao and Erbilgin, 2019), whereas two genes encoding terpene synthases contributted to the defense against the wood nematode attack that causes pine wilt disease in *P. massoniana* (Liu et al., 2021a). The *P. tabuliformis* is named after its rich resin, and its Chinese name ''you song'' just literally means ''resin pine.'' This may be associated with the substantial expansion of key rate-limiting enzyme genes coding terpenoid resin biosynthesis pathway and terpene synthase (TPS). The TPS number in *P. tabuliformis* is the largest in all major tree species sequenced so far (Celedon and Bohlmann, 2019; Song et al., 2020). These highly expanded metabolic genes may underpin the diverse adaptation and high accumulation of resin in *P.tabuliformis*, and its survival after frequent pine caterpillar attacks (Chen, 1990).

Distinctive conifer reproductive system

Over the decades, considerable progress has been made in elucidating the molecular basis of floral transition in model plants (Blümel et al., 2015). Even though the upstream regulatory networks seem to diverge among plant species, the highly conserved core floral integrator genes, such as *FT* and *AP1*, were shared in most of the flowering plants (Wellmer and Riechmann, 2010; Wickland and Hanzawa, 2015). In particular, FT as the long-distance mobile "florigen" (Corbesier et al., 2007), is the most important integrator of floral signals. However, *P. tabuliformis* and other conifers lack the *FT* ortholog (Nystedt et al., 2013), indicating the hierarchy gene regulatory network (hGRN) regulating flowering may be distinct. Our recent studies on *P. tabuliformis* indicated that a gymnosperm-sepecific age pathway (juvenile-mature), which did not involve *FT* genes, exists (Chen et al., 2021; Ma et al., 2021).

The MADS-box TFs play crucial roles in floral induction and reproductive organ identity in both angiosperms and gymnosperms (Gramzow et al., 2014). 74 *MADS-box* genes containing ultral-long introns were identified from the *P. tabuliformis* genome, and 51 of them were newly discovered (Niu et al., 2016). A comprehensive PPI network of 12 MADS-box TFs involved floral development in *P. tabuliformis* was constructed. These data may serve as fundamental resources and facilitate research on conifer reproductive development.

Overall, the nearly complete *P. tabuliformis* genome and gene space annotation provided us with insights into conifer evolution and will facilitate various evolutionary studies on conifer-specific traits of interest, comparative and functional genomics, GWAS, and genomics-assisted breeding.

Limitations of the study

High heterozygosity has great impacts on the quality of haplotype genome assembly. For annual crop/plant species from which inbreeding lines or haploids can be generated, the genome assembly quality is usually much higher than in the outcrossing perennial species with higher heterozygosity. For conifers, the seed endosperms are ideal haploid materials for genome sequencing, but the amount of endosperm material in a single seed of *P. tabuliformis* is insufficient for extracting enough DNA to meet the requirements for long-read sequencing. Due to the lack of suitable genetic transformation tools for conifers at the moment, we do not have sufficient experimental data to explain the reasons why conifers preserve these very long genes. This will be a highly interesting question in future studies of the evolutionary trajectory of conifers. There is a lack of other high-quality conifer genomes for an in-depth comparative genomics analysis, which limits the discovery of more conifer evolutionary trajectories and mechanisms in this study. However, the success in assembly of Chinese pine giga-genome may start a new era for comparative and evolutionary studies in conifers.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Biological materials
- METHOD DETAILS
  - ○ Chromosome (cytogenetic) analysis
  - ○ DNA and RNA extractions
  - ○ Illumina short-read sequencing
  - ○ Genome survey
  - ○ PacBio library construction and sequencing
  - ○ Hi-C library construction and sequencing
  - ○ De novo genome assembly and polishing
  - ○ Genome evaluation
  - ○ Chromosome-level pseudomolecule scaffolding
  - ○ RNA-seq and Iso-seq based gene model annotation
  - ○ *In silico* gene model annotation
  - ○ Gene functional annotation
  - ○ Non-coding RNA gene annotation
  - ○ Annotation of repeats and transposable elements
  - ○ Phylogenetic analysis of TEs
  - ○ Gene family and phylogenetic analysis
  - ○ Gene family expansion and contraction analysis
  - ○ Synteny analysis between Pinus tabuliformis and other gymnosperms
  - ○ Whole-genome duplication and gene duplication analysis
  - ○ Comparison of expression levels between genes with distinctive structural features
  - ○ Annotation of transcription factor and transcriptional regulator families
  - ○ Terpenoid biosynthesis pathway analysis
  - ○ Identification of orthologs of known flowering-time regulatory genes
  - ○ Yeast two-hybrid (Y2H) assay
  - ○ Whole genome bisulfite sequencing (WGBS)
  - ○ DNA methylome analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2021.12.006.

ACKNOWLEDGMENTS

This work was supported by The National Natural Science Foundation of China (no. 31870651 and 31600535) and The Program of Introducing Talents of Discipline to Universities (111 project, B13007). We would like to thank Dongmei Hu for her help in drawing some diagrams and English editing by Dr. Peter Stevens and Dr. Jennifer Sanders.

AUTHOR CONTRIBUTIONS

S.-H.N., J.L., W.-H.B., and W.-F.Y. are joint first authors, and A.Z. and S.G. are joint second authors who contributed to most parts of the work; W.-F.Y., X.-M.Z., T.L., S.-R.P., and Z.-M.L. contributed to the assembly and sequence analysis; A.Z., S.G., Y.-J.Z., and L.-H.Z. contributed to analysis of the repeat sequence and LTR-RT evolution; J.-J.M., F.-X.H., X.C., J.-H.Y., Y.-T.S., Y.-M.N., B.Z., P.-Y.W., Q.Z., and H.Z. contributed to manually annotating all TFs and TRs families; X.C., F.-X.H., J.-H.Y., Y.-T.S., Y.-M.N., B.Z., P.-Y.W., Q.Z., H.Z., Q.-Y.Z., and H.-H.Z. contributed to the pairwise Y2H assay between 12 TFs; and J.W. and L.-J.W. contributed to analyses of the chromosome morphology and genome size. Z.-M.L. selected the elite tree. Y.H., Y.-H.Y., and W.-Z.L. re-engineered the assembler Canu into a Workflow Description Language Canu (WDL-Canu) and quality controlled the assembly. J.-X.L., T.-Q.Y., W.L., and Y.L. contributed to the design and supervision of various parts of the research; H.X.W. conceived this research. H.X.W. and H.-R.W. headed and managed the project; S.-H.N. coordinated the sequencing, assembly, bioinformatics, and writing activities; S.-H.N., H.X.W., and H.-R.W. wrote and edited most of the manuscript. All authors commented on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 27, 2021
Revised: October 23, 2021
Accepted: December 3, 2021
Published: December 28, 2021

REFERENCES

Amborella Genome Project (2013). The Amborella genome and the evolution of flowering plants. Science *342*, 1241089.

Ausin, I., Feng, S., Yu, C., Liu, W., Kuo, H.Y., Jacobsen, E.L., Zhai, J., Gallego-Bartolome, J., Wang, L., Egertsdotter, U., et al. (2016). DNA methylome of the 20-gigabase Norway spruce genome. Proc. Natl. Acad. Sci. USA *113*, E8106–E8113.

Blümel, M., Dally, N., and Jung, C. (2015). Flowering time regulation in crops—what did we learn from Arabidopsis? Curr. Opin. Biotechnol. *32*, 121–129.

Bouché, F., Lobet, G., Tocquin, P., and Périlleux, C. (2016). FLOR-ID: an interactive database of flowering-time gene networks in Arabidopsis thaliana. Nucleic Acids Res. *44* (D1), D1167–D1171.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol. *31*, 1119–1125.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. *18*, 188–196.

Celedon, J.M., and Bohlmann, J. (2019). Oleoresin defenses in conifers: chemical diversity, terpene synthases and limitations of oleoresin defense under climate change. New Phytol. *224*, 1444–1463.

Chan, P.P., and Lowe, T.M. (2019). tRNAscan-SE: searching for tRNA genes in genomic sequences. Methods Mol. Biol. *1962*, 1–14.

Chen, C. (1990). Integrated Management of Pine Caterpillars in China (China Forestry Publishing House).

Chen, X., Köllner, T.G., Jia, Q., Norris, A., Santhanam, B., Rabe, P., Dickschat, J.S., Shaulsky, G., Gershenzon, J., and Chen, F. (2016). Terpene synthase genes in eukaryotes beyond plants and fungi: Occurrence in social amoebae. Proc. Natl. Acad. Sci. USA *113*, 12132–12137.

Chen, X., Zhu, Q., Nie, Y., Han, F., Li, Y., Wu, H.X., and Niu, S. (2021). Determination of conifer age biomarker DAL1 interactome using Y2H-seq. Forestry Res. *1*, 12.

Cheng, C.Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S., and Town, C.D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. *89*, 789–804.

Corbesier, L., Vincent, C., Jang, S., Fornara, F., Fan, Q., Searle, I., Giakountis, A., Farrona, S., Gissot, L., Turnbull, C., and Coupland, G. (2007). FT protein movement contributes to long-distance signaling in floral induction of Arabidopsis. Science *316*, 1030–1033.

Cossu, R.M., Casola, C., Giacomello, S., Vidalis, A., Scofield, D.G., and Zuccolo, A. (2017). LTR retrotransposons show low levels of unequal recombination and high rates of intraelement gene conversion in large plant genomes. Genome Biol. Evol. *9*, 3449–3462.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics *22*, 1269–1271.

Dodsworth, S., Leitch, A.R., and Leitch, I.J. (2015). Genome size diversity in angiosperms and its influence on gene space. Curr. Opin. Genet. Dev. *35*, 73–78.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Eilbeck, K., Moore, B., Holt, C., and Yandell, M. (2009). Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics *10*, 67.

El Baidouri, M., and Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. Genome Biol. Evol. *5*, 954–965.

Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. *16*, 157.

Fedoroff, N.V. (2012). Presidential address. Transposable elements, epigenetics, and genome evolution. Science *338*, 758–767.

Gramzow, L., Weilandt, L., and Theißen, G. (2014). MADS goes genomic in conifers: towards determining the ancestral set of MADS-box genes in seed plants. Ann. Bot. *114*, 1407–1429.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Ji, X., Li, P., Fuscoe, J.C., Chen, G., Xiao, W., Shi, L., Ning, B., Liu, Z., Hong, H., Wu, J., et al. (2020). A comprehensive rat transcriptome built from large scale RNA-seq-based annotation. Nucleic Acids Res. *48*, 8320–8331.

Jin, W.T., Gernandt, D.S., Wehenkel, C., Xia, X.M., Wei, X.X., and Wang, X.Q. (2021). Phylogenomic and ecological analyses reveal the spatiotemporal evolution of global pines. Proc. Natl. Acad. Sci. USA *118*, e2022302118.

Joyner, K.L., Wang, X., Johnston, J.S., Price, H.J., and Williams, C.G. (2001). DNA content for Asian pines parallels New World relatives. Can. J. Bot. *79*, 192–196.

Kelly, L.J., Renny-Byfield, S., Pellicer, J., Macas, J., Novák, P., Neumann, P., Lysak, M.A., Day, P.D., Berger, M., Fay, M.F., et al. (2015). Analysis of the giant genomes of Fritillaria (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. New Phytol. *208*, 596–607.

Kendig, K.I., Baheti, S., Bockol, M.A., Drucker, T.M., Hart, S.N., Heldenbrand, J.R., Hernaez, M., Hudson, M.E., Kalmbach, M.T., Klee, E.W., et al. (2019). Sentieon DNASeq Variant Calling Workflow Demonstrates Strong Computational Performance and Accuracy. Front. Genet. *10*, 736.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915.

Kim, E.Y., Wang, L., Lei, Z., Li, H., Fan, W., and Cho, J. (2021). Ribosome stalling and SGS3 phase separation prime the epigenetic silencing of transposons. Nat. Plants *7*, 303–309.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. *16*, 111–120.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res. *27*, 722–736.

Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L., and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. *20*, 278.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. *33*, 1870–1874.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lawrence, M., Gentleman, R., and Carey, V. (2009). rtracklayer: an R package for interfacing with genome browsers. Bioinformatics *25*, 1841–1842.

Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., and Li, Z.; One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. Nature *574*, 679–685.

Leslie, A.B., Simpson, C., and Mander, L. (2021). Reproductive innovations and pulsed rise in plant complexity. Science *373*, 1368–1372.

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. *47* (W1), W256–W259.

Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: recent updates, new developments and status in 2020. Nucleic Acids Res. *49* (D1), D458–D460.

Lev Maor, G., Yearim, A., and Ast, G. (2015). The alternative role of DNA methylation in splicing regulation. Trends Genet. *31*, 274–280.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178–2189.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, Z., Baniaga, A.E., Sessa, E.B., Scascitelli, M., Graham, S.W., Rieseberg, L.H., and Barker, M.S. (2015). Early genome duplications in conifers and other seed plants. Sci. Adv. *1*, e1501084.

Li, W., Liu, S.W., Ma, J.J., Liu, H.M., Han, F.X., Li, Y., and Niu, S.H. (2020). Gibberellin signaling is required for far-red light induced shoot elongation in *Pinus tabuliformis* seedlings. Plant Physiol. *182*, 658–668.

Lim, J.Q., Tennakoon, C., Li, G., Wong, E., Ruan, Y., Wei, C.L., and Sung, W.K. (2012). BatMeth: improved mapper for bisulfite sequencing reads on DNA methylation. Genome Biol. *13*, R82.

Liu, B., Liu, Q., Zhou, Z., Yin, H., Xie, Y., and Wei, Y. (2021a). Two terpene synthases in resistant Pinus massoniana contribute to defence against Bursaphelenchus xylophilus. Plant Cell Environ. *44*, 257–274.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., and Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv, 1308.2012.

Liu, H., Wang, X., Wang, G., Cui, P., Wu, S., Ai, C., Hu, N., Li, A., He, B., Shao, X., et al. (2021b). The nearly complete genome of Ginkgo biloba illuminates gymnosperm evolution. Nat. Plants *7*, 748–756.

Liu, Y.Y., Yang, K.Z., Wei, X.X., and Wang, X.Q. (2016). Revisiting the phosphatidylethanolamine-binding protein (PEBP) gene family reveals cryptic FLOWERING LOCUS T gene homologs in gymnosperms and sheds new light on functional evolution. New Phytol. *212*, 730–744.

Loh, W. (2002). Regression trees with unbiased variable selection and interaction detection. Stat. Sin. *12*, 361–386.

Ma, L., Hatlen, A., Kelly, L.J., Becher, H., Wang, W., Kovarik, A., Leitch, I.J., and Leitch, A.R. (2015). Angiosperms are unique among land plant lineages in the occurrence of key genes in the RNA-directed dna methylation (RdDM) pathway. Genome Biol. Evol. *7*, 2648–2662.

Ma, J.J., Chen, X., Song, Y.T., Zhang, G.F., Zhou, X.Q., Que, S.P., Mao, F., Pervaiz, T., Lin, J.X., Li, Y., et al. (2021). MADS-box transcription factors MADS11 and DAL1 interact to mediate the vegetative-to-reproductive transition in pine. Plant Physiol. *187*, 247–262.

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–770.

Meyer, A., Schloissnig, S., Franchini, P., Du, K., Woltering, J.M., Irisarri, I., Wong, W.Y., Nowoshilow, S., Kneitz, S., Kawaguchi, A., et al. (2021). Giant lungfish genome elucidates the conquest of land by vertebrates. Nature *590*, 284–289.

Mizutani, M., and Ohta, D. (2010). Diversification of P450 genes during land plant evolution. Annu. Rev. Plant Biol. *61*, 291–315.

Mosca, E., Cruz, F., Gómez-Garrido, J., Bianco, L., Rellstab, C., Brodbeck, S., Csilléry, K., Fady, B., Fladung, M., Fussi, B., et al. (2019). A Reference Genome Sequence for the European Silver Fir (*Abies alba* Mill.): A Community-Generated Genomic Resource. G3 (Bethesda) *9*, 2039–2049.

Murray, B.G. (1998). Nuclear DNA amounts in gymnosperms. Ann. Bot. (Lond.) *82*, 3–15.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935.

Niu, S.H., Li, Z.X., Yuan, H.W., Chen, X.Y., Li, Y., and Li, W. (2013). Transcriptome characterisation of *Pinus tabuliformis* and evolution of genes in the *Pinus* phylogeny. BMC Genomics *14*, 263.

Niu, S.H., Liu, C., Yuan, H.W., Li, P., Li, Y., and Li, W. (2015). Identification and expression profiles of sRNAs and their biogenesis and action-related genes in male and female cones of *Pinus tabuliformis*. BMC Genomics *16*, 693.

Niu, S., Yuan, H., Sun, X., Porth, I., Li, Y., El-Kassaby, Y.A., and Li, W. (2016). A transcriptomics investigation into pine reproductive organ development. New Phytol. *209*, 1278–1289.

Niu, S.H., Liu, S.W., Ma, J.J., Han, F.X., Li, Y., and Li, W. (2019). The transcriptional activity of a temperature-sensitive transcription factor module is associated with pollen shedding time in pine. Tree Physiol. *39*, 1173–1186.

Nowoshilow, S., Schloissnig, S., Fei, J., Dahl, A., Pang, A.W.C., Pippel, M., Winkler, S., Hastie, A.R., Young, G., Roscito, J.G., et al. (2018). The axolotl genome and the evolution of key tissue formation regulators. Nature *554*, 50–55.

Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. Nature *497*, 579–584.

Ou, S., and Jiang, N. (2019). LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob. DNA *10*, 48.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. *20*, 275.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., et al. (2007). The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res. *35*, D883–D887.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics *20*, 289–290.

Pervaiz, T., Liu, S.W., Uddin, S., Amjid, M.W., Niu, S.H., and Wu, H.X. (2021). The transcriptional landscape and hub genes associated with physiological responses to drought stress in *Pinus tabuliformis*. Int. J. Mol. Sci. *22*, 9604.

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., and Paterson, A.H. (2019). Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. Genome Biol. *20*, 38.

Roach, M.J., Schmidt, S.A., and Borneman, A.R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics *19*, 460.

Rosso, M.G., Li, Y., Strizhov, N., Reiss, B., Dekker, K., and Weisshaar, B. (2003). An Arabidopsis thaliana T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. Plant Mol. Biol. *53*, 247–259.

Sander, H., and Meikar, T. (2009). Exotic Coniferous Trees in Estonian Forestry after 1918. Allg. Forst Jagdztg. *180*, 158–169.

Schneider, H., Schuettpelz, E., Pryer, K.M., Cranfill, R., Magallón, S., and Lupia, R. (2004). Ferns diversified in the shadow of angiosperms. Nature *428*, 553–557.

Scott, A.D., Zimin, A.V., Puiu, D., Workman, R., Britton, M., Zaman, S., Caballero, M., Read, A.C., Bogdanove, A.J., Burns, E., et al. (2020). A reference genome sequence for giant sequoia. G3 (Bethesda) *10*, 3907–3919.

Sederoff, R. (2013). Genomics: A spruce sequence. Nature *497*, 569–570.

Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: assessing genome assembly and annotation completeness. Methods Mol. Biol. *1962*, 227–245.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. *16*, 259.

Shao, M., and Kingsford, C. (2017). Accurate assembly of transcripts through phase-preserving graph decomposition. Nat. Biotechnol. *35*, 1167–1169.

Sharma, V., Clark, A.J., and Kawashima, T. (2021). Insights into the molecular evolution of fertilization mechanism in land plants. Plant Reprod. *34*, 353–364.

Shi, Y., Ding, Y., and Yang, S. (2018). Molecular regulation of CBF signaling in cold acclimation. Trends Plant Sci. *23*, 623–637.

Smith, S.A., Beaulieu, J.M., and Donoghue, M.J. (2010). An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. Proc. Natl. Acad. Sci. USA *107*, 5897–5902.

Song, L., Huang, S.C., Wise, A., Castanon, R., Nery, J.R., Chen, H., Watanabe, M., Thomas, J., Bar-Joseph, Z., and Ecker, J.R. (2016). A transcription factor hierarchy defines an environmental stress response network. Science *354*, g1550.

Song, X., Wang, J., Li, N., Yu, J., Meng, F., Wei, C., Liu, C., Chen, W., Nie, F., Zhang, Z., et al. (2020). Deciphering the high-quality genome sequence of coriander that causes controversial feelings. Plant Biotechnol. J. *18*, 1444–1456.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics *24*, 637–644.

Stival Sena, J., Giguère, I., Boyle, B., Rigault, P., Birol, I., Zuccolo, A., Ritland, K., Ritland, C., Bohlmann, J., Jones, S., et al. (2014). Evolution of gene structure in the conifer Picea glauca: a comparative analysis of the impact of intron size. BMC Plant Biol. *14*, 95.

Stull, G.W., Qu, X.J., Parins-Fukuchi, C., Yang, Y.Y., Yang, J.B., Yang, Z.Y., Hu, Y., Ma, H., Soltis, P.S., Soltis, D.E., et al. (2021). Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. Nat. Plants *7*, 1015–1025.

Sun, Y., Shang, L., Zhu, Q.H., Fan, L., and Guo, L. (2021). Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci. S1360-1385(21)00281-8. https://doi.org/10.1016/j.tplants.2021.10.006.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. Science *320*, 486–488.

Tholl, D., and Lee, S. (2011). Terpene specialized metabolism in Arabidopsis thaliana. Arabidopsis Book *9*, e0143.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science *313*, 1596–1604.

Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y., Coppens, F., and Vandepoele, K. (2018). PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. *46* (D1), D1190–D1196.

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. *10*, 725–732.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE *9*, e112963.

Wan, T., Liu, Z.M., Li, L.F., Leitch, A.R., Leitch, I.J., Lohaus, R., Liu, Z.J., Xin, H.P., Gong, Y.B., Liu, Y., et al. (2018). A genome for gnetophytes and early evolution of seed plants. Nat. Plants *4*, 82–89.

Wan, T., Liu, Z., Leitch, I.J., Xin, H., Maggs-Kölling, G., Gong, Y., Li, Z., Marais, E., Liao, Y., Dai, C., et al. (2021). The Welwitschia genome reveals a unique biology underpinning extreme longevity in deserts. Nat. Commun. *12*, 4247.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. *40*, e49.

Wang, C., Liu, C., Roqueiro, D., Grimm, D., Schwab, R., Becker, C., Lanz, C., and Weigel, D. (2015). Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. Genome Res. *25*, 246–256.

Wang, L., Shi, Y., Chang, X., Jing, S., Zhang, Q., You, C., Yuan, H., and Wang, H. (2019). DNA methylome analysis provides evidence that the expansion of the tea genome is linked to TE bursts. Plant Biotechnol. J. *17*, 826–835.

Wang, D., Zheng, Z., Li, Y., Hu, H., Wang, Z., Du, X., Zhang, S., Zhu, M., Dong, L., Ren, G., and Yang, Y. (2021a). Which factors contribute most to genome size variation within angiosperms? Ecol. Evol. *11*, 2660–2668.

Wang, K., Wang, J., Zhu, C., Yang, L., Ren, Y., Ruan, J., Fan, G., Hu, J., Xu, W., Bi, X., et al. (2021b). African lungfish genome sheds light on the vertebrate water-to-land transition. Cell *184*, 1362–1376.

Warren, R.L., Keeling, C.I., Yuen, M.M.S., Raymond, A., Taylor, G.A., Vandervalk, B.P., Mohamadi, H., Paulino, D., Chiu, R., Jackman, S.D., et al. (2015). Improved white spruce (Picea glauca) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. Plant J. *83*, 189–212.

Wegrzyn, J.L., Lee, J.M., Tearse, B.R., and Neale, D.B. (2008). TreeGenes: A forest tree genome database. Int. J. Plant Genomics *2008*, 412875.

Wellmer, F., and Riechmann, J.L. (2010). Gene networks controlling the initiation of flower development. Trends Genet. *26*, 519–527.

Wickland, D.P., and Hanzawa, Y. (2015). The FLOWERING LOCUS T/TERMINAL FLOWER 1 gene family: functional evolution and molecular mechanisms. Mol. Plant *8*, 983–997.

Wu, S., Han, B., and Jiao, Y. (2020). Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. Mol. Plant *13*, 59–71.

Xu, J., Nie, S., Xu, C.Q., Liu, H., Jia, K.H., Zhou, S.S., Zhao, W., Zhou, X.Q., El-Kassaby, Y.A., Wang, X.R., et al. (2021). UV-B-induced molecular mechanisms of stress physiology responses in the major northern Chinese conifer Pinus tabuliformis Carr. Tree Physiol. *41*, 1247–1263.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Zhang, L., Wu, S., Chang, X., Wang, X., Zhao, Y., Xia, Y., Trigiano, R.N., Jiao, Y., and Chen, F. (2020a). The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. Plant Cell Environ. *43*, 2847–2856.

Zhang, S., Li, R., Zhang, L., Chen, S., Xie, M., Yang, L., Xia, Y., Foyer, C.H., Zhao, Z., and Lam, H.M. (2020b). New insights into Arabidopsis transcriptome complexity revealed by direct sequencing of native RNAs. Nucleic Acids Res. *48*, 7700–7711.

Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X., and Dai, L. (2012). ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. Biochem. Biophys. Res. Commun. *419*, 779–781.

Zhao, S., and Erbilgin, N. (2019). Larger resin ducts are linked to the survival of lodgepole pine trees during mountain pine beetle outbreak. Front. Plant Sci. *10*, 1459.

Zhao, C., Zhang, Z., Xie, S., Si, T., Li, Y., and Zhu, J.K. (2016). Mutational evidence for the critical role of CBF transcription factors in cold acclimation in Arabidopsis. Plant Physiol. *171*, 2744–2759.

Zhao, Y.P., Fan, G., Yin, P.P., Sun, S., Li, N., Hong, X., Hu, G., Zhang, H., Zhang, F.M., Han, J.D., et al. (2019). Resequencing 545 ginkgo genomes across the world reveals the evolutionary history of the living fossil. Nat. Commun. *10*, 4201.

Zheng, Y., Jiao, C., Sun, H., Rosli, H.G., Pombo, M.A., Zhang, P., Banf, M., Dai, X., Martin, G.B., Giovannoni, J.J., et al. (2016). iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. Mol. Plant *9*, 1667–1670.

Zhou, W., Liang, G., Molloy, P.L., and Jones, P.A. (2020). DNA methylation enables transposable element-driven genome expansion. Proc. Natl. Acad. Sci. USA *117*, 19359–19366.

STAR★METHODS

KEY RESOURCES TABLE

Reagent or resource	Source	Identifier
Biological samples		
Pinus tabuliformis (Chinese pine)	This study	N/A
Critical commercial assays		
NovaSeq 6000 S4 Reagent Kit V1.5(300cycles)	Illumina	20028312
NovaSeq XP 4-Lane Kit V1.5	Illumina	20043131
SMRTbell Template Prep Kit 1.0	Pacific Biosciences	Cat#100-259-100
SMRTbell Express Template Prep Kit 2.0	Pacific Biosciences	Cat#100-938-900
Sequel Sequencing Kit 3.0 (8 reaction plate)	Pacific Biosciences	Cat#101-597-800
Sequel Binding and Internal Control Kit 3.0	Pacific Biosciences	Cat#101-626-600
Sequel II Sequencing Kit 1.0 (4 rxn)	Pacific Biosciences	Cat#101-717-200
Sequel II Binding and Internal Control Kit1.0	Pacific Biosciences	Cat#101-731-100
Unmethylated Lambda DNA	Promega	D1521
EZDNAMethylation-Goldkit	ZYMO Research	D5006
KAPA HiFi HotStart ReadyMix	KAPA Biosystems	KK2631
KAPA HiFi HotStart Uracil+ Kit	KAPA Biosystems	KK2802
VAHTS mRNA-seq v2 Library Prep kit for Illumina-BOX2	Vazyme	NR601-02
LibPrep kit V2	Annoroad	N/A
Deposited data		
Genome assembly for *Pinus tabuliformis*	This study	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA784915
Genome annotations for *Pinus tabuliformis*	This study	https://figshare.com/articles/dataset/Pinus_tabuliformis_gene_space_annotation/16847146/1
RNA-seq data of 760 samples for gene space annotation of *Pinus tabuliformis*	This study	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA784915
The DNA methylation data of *Pinus tabuliformis*	This study	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA785099
The sRNA-seq data of *Pinus tabuliformis*	This study	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA785122
Reference transcriptome of *Pinus tabuliformis*	Niu et al., 2013	https://www.ncbi.nlm.nih.gov/sra/SRA056887
RNA-seq data of drought treatment of *Pinus tabuliformis*	Pervaiz et al., 2021	http://db.cngb.org/search/project/CNP0002179/
RNA-seq data of *Pinus tabuliformis* at different ages	Ma et al., 2021	http://db.cngb.org/search/project/CNP0001648/
RNA-seq data of light treatment of *Pinus tabuliformis*	Li et al., 2020	http://db.cngb.org/search/project/CNP0000737/
RNA-seq data of *Pinus tabuliformis* at different reproductive-stage	Niu et al., 2019	https://www.ncbi.nlm.nih.gov/bioproject/173457
RNA-seq data of UVB treatment of *Pinus tabuliformis*	Xu et al., 2021	https://www.ncbi.nlm.nih.gov/bioproject/557580
Genome and annotation of *Abies alba*	Wegrzyn et al., 2008	https://treegenesdb.org/FTP/Genomes/Abal/v1.1
Genome and annotation of *Amborella trichopoda*	Amborella Genome Project, 2013	https://phytozome-next.jgi.doe.gov/info/Atrichopoda_v1_0
Genome and annotation of *Arabidopsis thaliana*	Cheng et al., 2017	https://phytozome-next.jgi.doe.gov/info/Athaliana_Araport11
Transcriptome of *Cycas micholitzii*	Van Bel et al., 2018	ftp://ftp.psb.ugent.be/pub/plaza/plaza_gymno_01/Fasta/proteome.cmi.csv.gz

(*Continued on next page*)

Continued

Reagent or resource	Source	Identifier
Genome and annotation of *Ginkgo biloba*	Zhao et al., 2019	http://gigadb.org/dataset/100613?tdsourcetag=s_pcqq_aiomsg
Genome and annotation of *Gnetum montanum*	Wegrzyn et al., 2008	https://treegenesdb.org/FTP/Genomes/Gnmo/v1.0
Genome and annotation of *Oryza sativa*	Ouyang et al., 2007	https://phytozome-next.jgi.doe.gov/info/Osativa_v7_0
Genome and annotation of *Physcomitrella patens*	Van Bel et al., 2018	ftp://ftp.psb.ugent.be/pub/plaza/plaza_gymno_01/Fasta/proteome.ppa.csv.gz
Genome and annotation of *Picea abies*	Wegrzyn et al., 2008	https://treegenesdb.org/FTP/Genomes/Paab/v1.0b
Genome and annotation of *Picea sitchensis*	Van Bel et al., 2018	ftp://ftp.psb.ugent.be/pub/plaza/plaza_gymno_01/Fasta/proteome.psi.csv.gz
Genome and annotation of *Pinus lambertiana*	Wegrzyn et al., 2008	https://treegenesdb.org/FTP/Genomes/Pila/v1.5
Transcriptome of *Pinus pinaster*	SustainPineDB	https://www.scbi.uma.es/sustainpinedb
Transcriptome of *Pinus sylvestris*	Van Bel et al., 2018	ftp://ftp.psb.ugent.be/pub/plaza/plaza_gymno_01/Fasta/proteome.psy.csv.gz
Genome and annotation of *Pinus taeda*	Wegrzyn et al., 2008	https://treegenesdb.org/FTP/Genomes/Pita/v2.01
Genome and annotation of *Populus trichocarpa*	Tuskan et al., 2006	https://phytozome-next.jgi.doe.gov/info/Ptrichocarpa_v4_1
Genome and annotation of *Pseudotsuga menziesii*	Wegrzyn et al., 2008	https://treegenesdb.org/FTP/Genomes/Psme/v1.0
Genome and annotation of *Sequoiadendron giganteum*	Wegrzyn et al., 2008	https://treegenesdb.org/FTP/Genomes/Segi/v2.0
Transcriptome of *Taxus baccata*	Van Bel et al., 2018	ftp://ftp.psb.ugent.be/pub/plaza/plaza_gymno_01/Fasta/proteome.tba.csv.gz
Protein datasets of *Abies alba*	Wegrzyn et al., 2008	https://treegenesdb.org/org/Abies-alba
Protein datasets of *Pinus taeda*	Wegrzyn et al., 2008	https://treegenesdb.org/org/Pinus-taeda
Protein datasets of *Pinus lambertiana*	Wegrzyn et al., 2008	https://treegenesdb.org/org/Pinus-lambertiana
Protein datasets of *Picea abies*	Wegrzyn et al., 2008	https://treegenesdb.org/org/Picea-abies
Protein datasets of *Pseudotsuga menziesii*	Wegrzyn et al., 2008	https://treegenesdb.org/org/Pseudotsuga-menziesii
Protein datasets of *Sequoiadendron giganteum*	Scott et al., 2020	https://nealelab.ucdavis.edu/redwood-genome-project-rgp
Experimental models: Cell lines		
Y2HGold (Yeast)	Weidi Biotechnology	CAT#: YC1002
Experimental models: Organisms/strains		
Arabidopsis thaliana: 35S::PtTFL1	This study	N/A
Arabidopsis thaliana: 35S::PtTFL2	This study	N/A
Arabidopsis thaliana: 35S::PtDAL10	This study	N/A
Software and algorithms		
ImageJ	NIMH	https://imagej.nih.gov/ij/
Jellyfish v2.2.0	Marçais and Kingsford, 2011	https://github.com/gmarcais/Jellyfish
GCE v1.0.2	Liu et al., 2013	https://github.com/fanagislab/GCE
Canu v1.8	Koren et al., 2017	https://canu.readthedocs.io/en/latest/
Purge haplotigs multiBAM	Roach et al., 2018	https://github.com/skingan/purge_haplotigs_multiBAM
Pilon v1.23	Walker et al., 2014	https://github.com/broadinstitute/pilon/releases/
BWA v0.7.9a	Li and Durbin, 2009	https://github.com/lh3/bwa/releases/
SAMtools v0.1.19	Li et al., 2009	https://github.com/samtools/samtools
Sentieon DNASeq variant calling workflow v201911	Kendig et al., 2019	https://github.com/Sentieon/sentieon-dnaseq
BUSCO v4.1.4	Seppey et al., 2019	https://gitlab.com/ezlab/busco
Bowtie2 v2.2.3	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml
HiC-Pro v2.7.8	Servant et al., 2015	https://github.com/nservant/HiC-Pro

(*Continued on next page*)

Continued

Reagent or resource	Source	Identifier
LACHESIS	Burton et al., 2013	https://github.com/shendurelab/LACHESIS
SMRT Link v8.0	Pacific Biosciences	https://www.pacb.com/support/software-downloads/
Lima v2.2.0	Pacific Biosciences	https://github.com/PacificBiosciences/barcoding
minimap2 v2.17	Li, 2018	https://github.com/lh3/minimap2
HISAT2 v2.2.0	Kim et al., 2019	http://daehwankimlab.github.io/hisat2/
StringTie2 v2.1.2	Kovaka et al., 2019	https://github.com/skovaka/stringtie2
Scallop v0.10.5	Shao and Kingsford, 2017	https://github.com/Kingsford-Group/scallop
TransDecoder v5.2.0	Brian Haas	https://github.com/TransDecoder/TransDecoder
PASApipeline v2.3.3	Brian Haas	https://github.com/PASApipeline/PASApipeline
Augustus v3.3.3	Stanke et al., 2008	https://github.com/Gaius-Augustus/Augustus/releases
MAKER v3.01.03	Cantarel et al., 2008	http://www.yandell-lab.org/software/maker.html
RepeatMasker v4.0.6	Arian Smit & Robert Hubley	http://repeatmasker.org/RepeatMasker/
tRNAscan-SE v1.3.1	Chan and Lowe, 2019	https://github.com/UCSC-LoweLab/tRNAscan-SE
Infernal v1.1.3	Nawrocki and Eddy, 2013	https://github.com/EddyRivasLab/infernal
EDTA	Ou et al., 2019	https://github.com/oushujun/EDTA
LTR_FINDER_parallel	Ou and Jiang, 2019	https://github.com/oushujun/LTR_FINDER_parallel
MUSCLE v3.6	Edgar, 2004	http://www.drive5.com/muscle/
MEGA7	Kumar et al., 2016	https://www.megasoftware.net/
OrthoMCL v1.4	Li et al., 2003	https://github.com/stajichlab/OrthoMCL
PhyML v3.0	Guindon et al., 2010	https://www.softpedia.com/get/Science-CAD/PhyML.shtml
PAML v4.4	Yang, 2007	http://abacus.gene.ucl.ac.uk/software/paml.html
CAFE v2.1	De Bie et al., 2006	https://github.com/hahnlab/CAFE
JCVI v0.84	Tang et al., 2008	https://github.com/tanghaibao/jcvi/wiki/MCscan-Python-version
MCScanX	Wang et al., 2012	https://github.com/wyp1125/MCScanX
ParaAT v2.0	Zhang et al., 2012	https://ngdc.cncb.ac.cn/tools/paraat
DupGen_finder	Qiao et al., 2019	https://github.com/qiao-xin/DupGen_finder
APE v5.5	Paradis et al., 2004	https://cran.r-project.org/web/packages/ape/index.html
rtracklayer	Lawrence et al., 2009	https://bioconductor.org/packages/release/bioc/html/rtracklayer.html
ggplot2	Hadley Wickham	https://github.com/tidyverse/ggplot2
iTAK v1.7a	Zheng et al., 2016	https://github.com/kentnf/iTAK/releases
SMART	Letunic et al., 2021	http://smart.embl-heidelberg.de/
iTOL	Letunic and Bork, 2019	https://itol.embl.de/itol.cgi
OrthoFinder v2.2.6	Emms and Kelly, 2015	https://github.com/davidemms/OrthoFinder
Bismark v0.20.0	Krueger and Andrews, 2011	https://www.bioinformatics.babraham.ac.uk/projects/bismark/
BatMeth2	Lim et al., 2012	https://github.com/GuoliangLi-HZAU/BatMeth2

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Harry X. Wu (harry.wu@slu.se).

Materials availability

This study did not generate any new unique reagents.

Data and code availability

- The genome sequences and raw genome and transcriptome sequencing data for *Pinus tabuliformis* have been deposited at the NCBI under the BioProject: PRJNA784915. Please note that due to the limitation of NCBI on the length of a single chromosome, we must divide the five longest chromosomes into two halves, named Chr1.1/Chr1.2 to Chr5.1/Chr5.2, respectively. In order to facilitate users with different needs, we also upload the undivided version to the CNSA of China National GeneBank Database: CNP0001649. The annotation gff3 file have been deposited at the Figshare (https://figshare.com/articles/dataset/Pinus_tabuliformis_gene_space_annotation/16847146/1). The DNA methylation and sRNA sequencing data are available in NCBI under the BioProject: PRJNA785099 and PRJNA785122, respectively.
- All original codes have been deposited at Github and is publicly available as of the date of publication. Links are listed in the Key Resources Table.
- Any additional information required to reanalyze the data reported in this work paper is available from the Lead Contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Biological materials

We sequenced a 35-year-old root-grafted elite tree clone of *P. tabuliformis*, which was collected in a Chinese pine breeding seed orchard located in Pingquan City, Hebei Province, China (118°44.6758′ E, 40°98.8784′ N, 560−580 m above sea level). For high quality DNA extraction and Hi-C library construction, the flesh new shoots that had just sprouted and were still covered with bud scales, were sampled on May 10th, 2019. After being harvested, the scales were removed and the new shoots were immediately frozen in liquid nitrogen and then stored under the −80°C refrigerator until further use.

760 biological samples from 11 different organs/tissues under normal and various treatment conditions, including multiple light regimes (Li et al., 2020), temperatures (Niu et al., 2015; Niu et al., 2019), drought (Pervaiz et al., 2021), phytohormones treatment, aging and developmental stages (Niu et al., 2016), were collected in our previous studies and this project (2014 to 2019), and used to conduct RNA-seq experiments to generate multi-organ/tissue transcriptomes. In addition, two biological replicates of new shoots were used to conduct bisulfite sequencing for profiling genome-wide DNA methylation.

METHOD DETAILS

Chromosome (cytogenetic) analysis

Root tips of one-centimeter long from the *P. tabuliformis* seedlings were excised and pretreated in 0.05% colchicine solution for 12 h at 25°C in the dark. After rinsing with ddH_2O, the root tips were fixed in fresh Carnoy's solution (3:1 ethanol: glacial acetic acid). The fixed roots were washed 3 times in ddH_2O and 0.1 M citric buffer (pH 4.8) for 10 min, respectively. Subsequently, root tips were cut off and immediately transferred into an enzyme mixture containing 2.0% cellulase and 1.0% pectinase for 1 h at 37°C. After that, the tips were washed using ddH_2O and fixed with fresh Carnoy's solution again. Each of the fixed tip was placed on a cleaned slide and then squashed under a cover glass in 45% acetic acid. The slides were observed and imaged using confocal microscope (Leica-SP8) with differential interference contrast (DIC) optics system. The ImageJ (https://imagej.nih.gov/ij/) was used to quantify size of each chromosome.

DNA and RNA extractions

High-quality genomic DNA was extracted from the new shoots of *P. tabuliformis* following the protocol of DNeasy Plant Mini Kit (QIAGEN). The integrity of the DNA was verified with an Agilent 4200 Bioanalyzer (Agilent Technologies, Palo Alto, California). Total RNA from different tissues of *P. tabuliformis* was isolated by standard TRIzol protocol. The integrity of the RNA was determined with the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, California, USA).

Illumina short-read sequencing

Qualified DNA fragmentation was carried out by Ultrasonic Processor to yield the DNA-fragment lengths of approximately 350 bp, which were verified with an Agilent 2100. Then the sequencing libraries were constructed following the procedures provided by the Nextera XT DNA Library Prep Kit (Illumina, Inc., San Diego, CA, USA), which included, but were not limited to, terminal repair, base A addition, sequence adaptor addition, DNA purification, and PCR amplification. Following that, preliminary quantitative by Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA), and then, each library was diluted to a concentration of 1 ng/ul. The insert size of each library was verified by an Agilent 2100 (Agilent Technologies, Santa Clara, CA, USA), Subsequently, qPCR was implemented to ensure the effective quantitative concentrations of the libraries. The Illumina DNA libraries were then sequenced on an Illumina NovaSeq 6000 (Illumina Inc, CA, USA) platform to generate 150-bp paired-end reads; 2,638 Gb clean sequencing data were obtained, accounting for nearly 103 × genome coverage for subsequent analysis.

Genome survey

To estimate the genome size, heterozygosity and repeat content, Jellyfish v2.2.0 (Marçais and Kingsford, 2011) (https://github.com/gmarcais/Jellyfish) was used to generate a 21 *K*-mer frequency distribution. Depth of *K*-mer = 1 is considered as an error, and this error rate was used to calculate and correct the genome size. The formula for estimating the genome size is: Genome size = (K-mer num/main peak depth) × (1-Error rate). The heterozygous ratio and repeat sequence ratio were estimated by the GCE v1.0.2 (Liu et al., 2013).

PacBio library construction and sequencing

DNA were sheared using g-Tubes (Covaris), and concentrated with AMPure PB magnetic beads. Each SMRTbell library was constructed using the Pacific Biosciences SMRTbell express template prep kit 2.02. The constructed libraries were size-selected on a BluePippin system for molecules 20 Kb, followed by primer annealing and the binding of SMRTbell templates to polymerases with the DNA/Polymerase Binding Kit.

For cDNA library construction, 4 μg RNA was used to synthesize to cDNA using the Clontech SMARTer PCR cDNA Synthesis Kit (Takara Biotechnology, Dalian, China), and subsequently amplified to generate double-stranded cDNA. SMRTbell library was constructed following the protocol from the manufacturer: 1 μg cDNA was used with the Pacific Biosciences SMRTbell template prep kit. The binding of SMRT bell templates to polymerases was conducted using the Sequel II Binding Kit. After primer annealing was performed, the sequencing was carried out on the Pacific Bioscience Sequel II platform by Annoroad Gene Technology Company (Beijing, China).

Hi-C library construction and sequencing

The Hi-C libraries were prepared using freshly sampled new shoots following the published procedure (Wang et al., 2015). Briefly, the nuclear DNA was cross-linked *in situ* in 2% formaldehyde before the nuclei were extracted; the nuclei extracted were then digested by MboI restriction endonuclease. The sticky ends of the digested fragments were biotinylated, diluted and ligated randomly. The biotinylated DNA fragments were enriched to construct nine sequencing libraries, and the sequencing of these libraries was conducted on Illumina NovaSeq 6000 platform with 2 × 150-bp reads. We eventually obtained a total of 3,132 Gb sequencing data.

De novo genome assembly and polishing

We used about 172 Mb PacBio sequel II reads comprising 2.6 Tb subreads with contig N50 = 23kb for the primary assembly. The 172 million PacBio long reads (103 ×) were first corrected with Canu v1.8 (Koren et al., 2017) and nearly 2 Tbp corrected subreads were collected. Initially, Canu was not able to assemble the genome using such large dataset at 2 Tb. We modified the Canu v1.8 using Workflow Description Language (WDL) and Intel Advanced Vector Extensions 512 (AVX-512) to accelerate several computation steps in parallel to save both CPU and wall time. Therefore, the original Canu pipeline was optimized into a WDL-Canu by implementing schedule, algorithm and imput/output optimizations (detailed strategies and procedures see Supplementary method 1). We finally obtained an assembly by WDL-Canu (v1.8, corOutCoverage = 80, maxSortJobs = 50, maxBucketJobs = 50) with contig50 (= 2.6 Mb) using 3.1 million CPU hours. To identify the haplotigs (haplotypes) from a high heterozygous genome, we used purge_haplotigs_multiBAM (Roach et al., 2018) (https://github.com/skingan/purge_haplotigs_multiBAM) with the default parameters. After removing one copy of haplotigs, we performed two rounds of corrections to the assembly using nearly 30 × Illumina reads by Pilon tool v1.23 (Walker et al., 2014) with default parameters. In brief, we mapped the Illumina reads sequenced from the same genomic DNA samples using bwa v0.7.9a (Li and Durbin, 2009) (-M -k 30), and sorted the mapping bam using SAMtools v0.1.19 (Li et al., 2009) (sort -m 1G). Second, we employed Pilon to correct the misassembles of the small proportion of SNPs and indels. We had the Illumina short reads mapping rate of 96.73% to the genome with a final mapping accuracy of 99.973%. After two rounds of corrections, we collected the final *P. tabuliformis* genome assembly.

Genome evaluation

To evaluate the single nucleotide quality of genome, we employed Sentieon DNASeq variant calling workflow v201911 (Kendig et al., 2019) using nearly 30 × Illumina reads to identify the variation of single nucleotide variant (SNV). After filtering the low-quality variants, we collected the high quality SNVs and the small Indels to define the wrong nucleotide positions. Finally, 275 SNVs were identified per Mb that is equal to an accuracy of 99.973%. In addition, the Benchmarking Universal Single-Copy Orthologs (BUSCO, v4.1.4)(Seppey et al., 2019) with embryophyta_odb10 and eukaryota_odb10 database was used to check the assembly quality and the gene annotation with genome and protein modes, respectively. The available proteomes from other sequenced gymnosperm genomes or related pine transcriptomes in the database TreeGenes(Wegrzyn et al., 2008) and PLAZA (Van Bel et al., 2018) were also evaluated by the same pipeline for comparison.

Chromosome-level pseudomolecule scaffolding

The Illumina clean pair-end reads yielded from nine Hi-C libraries were mapped to contigs by Bowtie2 v2.2.3 (Langmead and Salzberg, 2012) (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml). HiC-Pro v2.7.8 (Servant et al., 2015) (https://github.com/nservant/HiC-Pro) was used to process the mapped Hi-C reads to obtain the valid reads pairs, which were used to generate the normalized contact maps. Pseudo-chromosomes were organized using LACHESIS (Burton et al., 2013) (https://github.com/shendurelab/LACHESIS), with key parameters including CLUSTER_MIN_RE_SITES = 985, CLUSTER_MAX_LINK_DENSITY = 9, CLUSTER_NONINFORMATIVE_RATIO = 8, and then followed by manual correction. The final assembly consisted of twelve pseudo-chromosomes and 96% of the contigs (24.4 Gb) were anchored and oriented successfully.

RNA-seq and Iso-seq based gene model annotation

We used SMRT Link v8.0 with the following parameters:–min-passes 3–min-length 50–max-length 15000–min-rq 0.99, to correct the circular consensus sequence (CCS) subreads and then collected the high-quality reads. Lima v2.2.0 was used to classify the full-length reads with the parameters:–isoseq–dump-clips–peak-guess. The isoseq3 was used to collect the final full length Iso-seq transcripts using the refine (parameters:–require-polya–min-polya-length 20) and cluster (parameters:–verbose–use-qvs) models. The full-length transcripts from PacBio Iso-seq reads were first aligned to the *P. tabuliformis* genome by minimap2 v2.17 (Li, 2018) with the following parameters: -t 30 -ax splice:hq -G 2000k -uf–secondary = no genome.fa input.ccs.fasta -o out.aln.sam. All short reads resulting from 760 RNA-seq samples were merged and then aligned to *P. tabuliformis* genome by HISAT2 v2.2.0 (Kim et al., 2019) with the default parameters. The aligned reads were separated or merged for assembling using two state-of-the art assemblers, StringTie2 v2.1.2 (Kovaka et al., 2019) and Scallop v0.10.5 (Shao and Kingsford, 2017), respectively. The different versions of transcriptomes were assessed by the 273 complete coding regions that were cloned by PCR in our previous studies (in house data). The StringTie2 was used to assemble the transcriptomes using the combined long reads and short reads, and the version of transcriptome with the most complete coding regions of 273 reporter genes was finally selected for subsequent analyses. The TransDecoder v5.2.0 was used to identify the coding sequence with default parameters.

In silico gene model annotation

For gene structure annotation, a strategy combining a homology-based method and a transcriptional evidence-based method was adopted. Previously assembled reference transcriptomes based on 454 pyrosequencing libraries (Niu et al., 2013) were used for *ab initio* gene prediction. For *ab initio* prediction, the PASApipeline v2.3.3, which can exploit the spliced alignments of expressed transcript sequences to automatically model gene structures, was used to identify and classify all splicing variations by the alignments of full-length transcripts. All complete gene structures produced with the PASApipeline v2.3.3 were used to train a gene model using Augustus v3.3.3 (Stanke et al., 2008) with default parameters. Final gene models to predict whole genome gene structure were prepared using Augustus v3.3.3 in the MAKER v3.01.03 pipeline (Cantarel et al., 2008). For homology-based mapping, all proteins of ten species from databases were mapped to the *P. tabuliformis* genome using the MAKER v3.01.03 pipeline. For transcriptional evidence, final full-length transcripts were used as cDNA sequence for MAKER v3.01.03 pipelines. A repeat library (repbase20.05) downloaded from the Repbase database (https://www.girinst.org/downloads/) was used to mask the TE repeats in the genome using RepeatMasker v4.0.6 in MAKER pipelines. MAKER v3.01.03 was then used to identify gene structures based on the three resources (trained gene model, homologous proteins, and full-length transcripts) with the parameters: alt_peptide = C, max_dna_len = 1000000, min_contig = 1000, pred_flank = 200, pred_stats = 0, AED_threshold = 1, min_protein = 50, alt_splice = 1, map_forward = 0, keep_preds = 0, split_hit = 10000, single_exon = 1, single_length = 250, correct_est_fusion = 0, always_complete = 0. Only gene models with the highest confidence, suggested by the zero annotation edit distance (AED) metric (Eilbeck et al., 2009), were selected. Finally, we determined the final whole genome gene structural annotation by merging two high quality gene structure annotation GFF3 files from MAKER.

Gene functional annotation

Functional annotation for each gene was carried out by mapping the protein sequence to the following databases using BLAST v2.2.28 with parameter -num_alignments 1: NR and NT (https://www.ncbi.nlm.nih.gov/), KEGG (https://www.genome.jp/kegg), GO (https://www.uniprot.org/), Pfam (http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/) and SwissPro (https://www.uniprot.org/downloads).

Non-coding RNA gene annotation

Non-coding RNA species including miRNA, tRNA, rRNA, and snRNA were annotated using several methods. tRNA species were predicted using tRNAscan-SE v1.3.1 (Chan and Lowe, 2019) with default parameters. rRNA species were identified by mapping *Arabidopsis thaliana* rRNA sequences to the *Pinus tabuliformis* genome using BLASTN-short v2.2.28. miRNA and snRNA were identified using Infernal v1.1.3 (Nawrocki and Eddy, 2013) with default parameters.

Annotation of repeats and transposable elements

To identify transposable elements (TEs) across the whole genome, we constructed a special TE library for *Pinus tabuliformis*. The TE library was obtained by running the pipeline extensive *de-novo* TE annotator (EDTA) (Ou et al., 2019) (https://github.com/oushujun/EDTA) on a random batch of contigs for a total of more than 250 Mbp. Since TEs are highly repetitive and randomly distributed in the conifer genome, this library is sufficient to identify the medium or highly repetitive fraction. The output of EDTA was filtered to retain only LTR-RTs and DNA-TEs, including MITEs. The *P. tabuliformis* TE library was extensively compared to other conifer TE libraries available in house (*Picea abies, Pinus teada, Picea glauca*) and proved both specific and comprehensive. The *P. tabuliformis* TE library was therefore used as a TE library for TE annotation and TE abundance evaluation. The library contained 628 entries and 559 of them are LTR-RTs (300 Ty3-gypsy RTs, 129 Ty1-copia RTs, 130 not assigned to either of the two super families). Full length

LTR-RTs were identified using LTR_FINDER_parallel (Ou and Jiang, 2019) with the following parameters: -w 2 -C -D 15000 -d 1000 -L 7000 -l 100 -p 20 -M 0.85.

Phylogenetic analysis of TEs

Phylogenetic analyses were carried out focusing on 100 amide acid (AA) residue-long tracts of the reverse transcriptase domains of both *Ty1-copia* and T*y3-gypsy* LTR-RTs. The tract was used as a query to carry out tBLASTN v2.2.28 searches of the following datasets: *Ginkgo biloba* (https://treegenesdb.org/FTP/Genomes/Gibi/v1.0/genome/Gibi.1_0.fa.gz), *Picea abies* (ftp://plantgenie.org/Data/ConGenIE/Picea_abies/v1.0/FASTA/GenomeAssemblies/Pabies1.0-genome.fa.gz), *Picea glauca* (https://www.ncbi.nlm.nih.gov/nuccore/ALWZ000000000.4), *Pinus taeda* (https://treegenesdb.org/FTP/Genomes/Pita/v2.01/genome/Pita.2_01.fa.gz), *Pinus lambertiana* (https://treegenesdb.org/FTP/Genomes/Pila/v1.5/annotation/Pila.1_5.cds.fa.gz), *Pseudotsuga menziesii* (https://treegenesdb.org/FTP/Genomes/Psme/v1.0/genome/Psme.1_0.fa.gz), and *Gnetum montanum* (https://www.ncbi.nlm.nih.gov/nuccore/MNCI01026832.1). All significant hits covering at least 80% of the query lengths were retrieved from the tBLASTN v2.2.28 output using an *ad hoc* Perl script (available upon request). For both the *copia* and *gypsy* subfamilies, 500 'rooted tree (RT)' paralogs were randomly extracted. The number of random paralogs per species was set to 100. Multiple sequence alignments were carried out using MUSCLE v3.6 (Edgar, 2004). MEGA7 (Kumar et al., 2016) was then used to create Neighbor-Joining phylogenetic trees with complete deletion; 1,000 replicates were used for bootstrap analysis, and the cutoff value was set to 50%.

Gene family and phylogenetic analysis

To infer the evolutionary history of *P. tabuliformis*, we selected eight species, including four angiosperms (*Oryza sativa* (ftp://ftp.ensemblgenomes.org/pub/plants/release-51/fasta/oryza_sativa/), *Arabidopsis thaliana* (https://www.arabidopsis.org/), *Nymphaea tetragona* (https://data.jgi.doe.gov/refine-download/ phytozome?organism = Ncolorata&expanded = 566), *Amborella trichopoda* (Amborella Genome Project, 2013), two gymnosperms (*Ginkgo biloba* (https://figshare.com/articles/dataset/annotation_of_Ginkgo_biloba/14759223), *Sequoiadendron giganteum* (http://www.ncbi.nlm.nih.gov/assembly/GCA_007115665.2,GCA_007115665.2/?&utm_source=None)), the lycophyte *Selaginella moellendorffii* (ftp://ftp.ensemblgenomes.org/pub/plants/release-51/fasta/selaginella_ moellendorffii/dna/)), and the moss *Physcomitrella patens* (https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Phytozome). The longest proteins for each species were collected and aligned with each other by all-versus-all BLASTP v2.2.28 with an E-value of 1E-5. OrthoMCL v1.4 (Li et al., 2003) was employed to identify 34,635 orthologs and paralogs for all species with the parameter (-I = 1.5). To construct the phylogenetic tree of *P. tabuliformis* and other eight species, we collected 65 single-copy gene families with orthologs and aligned the orthologs of each family using MUSCLE v3.6 (Edgar, 2004). Then, we built a super alignment matrix and used it to construct a maximum likelihood phylogenetic tree using PhyML v3.0 (Guindon et al., 2010). To estimate the divergence time between species or clade, we employed mcmctree, a sub-program of PAML v4.4 (Yang, 2007) with parameters: RootAge = 500, model = 4, alpha = 0, clock = 3, sample frequency = 2, burn-in = 20000, nsample = 100000, finetune = "0.00876 0.03724 0.06828 0.00789 0.44485." The divergence time was also corrected with the known calibration points sourced from Timetree (http://timetree.org/).

Gene family expansion and contraction analysis

We identified the expansion and contraction of orthologous groups using computational analysis of gene family evolution (CAFE v2.1) (De Bie et al., 2006) according to the difference in gene number of each orthologous group of each species with parameters: -p 0.05 -t 1 -r 10000. The probabilistic graphical model (PGM) was used to estimate the size of each orthologous group at each ancestral node in the phylogenetic tree of *P. tabuliformis* and other eight species. To determine significance for expansion and contraction of orthologous groups, we calculated the *P*-value for each orthologous group based on a Monte Carlo resampling procedure; the threshold for significant expansion and contraction was set to *P*-value < 0.05. The *P value* is the probability or chance of observing at least x number of genes out of a short list is annotated to a particular GO term, thus the smaller the *P value*, the less likely the observation of these genes was caused by random sampling. Since the *P value* may be too small for the top list of the enriched GO terms to draw and show in plot, we converted it with -log10(*P value*) in Figures S1B, S3A, and S3D.

Synteny analysis between Pinus tabuliformis and other gymnosperms

Syntenic gene pairs between *P. tabuliformis*, *G. biloba* and *S. giganteum* were identified using JCVI v0.84 (Tang et al., 2008) (a Python version of MCscan, https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)). Using coding sequence (CDS) and annotation gff3 files as input data, the syntenic blocks for each pair species were identified using 'jcvi.compara.catalog ortholog' with a parameter of–cscore = 0.8. The syntenic blocks were filtered using 'jcvi.compara.synteny screen' with parameters:–minspan = 30–simple. Synteny pattern was detected using 'jcvi.compara.synteny depth–histogram'.

Whole-genome duplication and gene duplication analysis

To determine if there was a recent whole-genome duplication (WGD) in *P. tabuliformis*, we analyzed the distribution of synonymous substitutions per site (*Ks*) for each paralog in *P. tabuliformis*. All proteins sequences of *P. tabuliformis* were aligned all-versus-all with BLAST v2.2.28 (-e 1e-10 -num_alignments 5). The syntenic regions with collinearity of paralog pairs were identified by MCScanX (Wang et al., 2012) (https://github.com/wyp1125/MCScanX) with default parameters. ParaAT v2.0 (Zhang et al., 2012) was used to construct multiple protein-coding DNA alignments with the default parameters (-m muscle -f paml), and the result was used as input data for the codeml of PAML v4.9h (Yang, 2007) to calculate the *Ks* value for each paralog pair with default parameters (verbose = 0, icode = 0, weighting = 0, commonf3x4 = 0, ndata = 1). Finally, using the all-versus-all blast result and the gff3 files as input data, the DupGen_finder (Qiao et al., 2019) (https://github.com/qiao-xin/DupGen_finder) was employed with default parameters to identify different modes of duplicated gene pairs.

Comparison of expression levels between genes with distinctive structural features

The genome annotation files (gff3 or gtf format) from different species were used as input for R packages APE(Analyses of Phylogenetics and Evolution, v5.5) (Paradis et al., 2004) and rtracklayer (Lawrence et al., 2009) for gene structure related data extraction. Considering that many genes are alternatively spliced, only the longest transcript for each gene was chosen for analysis.

We sorted the genes according to the lengths of different genic structures (e.g., intron, exon, TE, nonTE), and then equally divided them into two gene structure sets (Figures 2E and S2C) or five gene structure sets (Figures 2F and S2D). For two gene sets, the length of any gene structures in the first set is shorter than that of any gene structures in the second set; we the used the "shorter" to mark the first gene structure sets, and used the "longer" to mark the other gene sets. Similarly, if the genes were equally divided into five gene structure sets, then the five sets are ranked from the shortest to the longest, namely, "shortest," "shorter," "medium," "longer," and "longest," respectively.

The maximum or average expression level of each genes which form different gene sets in 760 samples was used to compare the effects of different gene structures on expression levels.

Annotation of transcription factor and transcriptional regulator families

The program iTAK v1.7a (Zheng et al., 2016) was used to identify transcription factors (TFs) and transcriptional regulators (TRs) from protein sequences, and then the individual TFs and TRs were classified into different gene families. The known plant TFs and TRs from 197 plant species present in the iTAK database 18.12 (http://itak.feilab.net/cgi-bin/itak/online_itak.cgi) were used as a reference. We found that some non-conserved members would be missed by iTAK, therefore, to avoid missing family members with low identity value, the protein sequences of each putative TF or TR from *P. tabuliformis* and *Arabidopsis thaliana* were used as queries in a BLASTP v2.2.28 search of the *P. tabuliformis* protein dataset with an E-value of 1.0 x e^{-30}. The conserved domains in all the putative TF/TR identified were examined with CD-Search (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) and SMART (Letunic et al., 2021) (http://smart.embl-heidelberg.de/). TFs and TRs found to contain corresponding family domains were used for phylogenetic analysis. The same TF and TR families from other species were part of the analysis as well. These species include: *Chlamydomonas reinhardtii*, *Physcomitrella patens, Selaginella tamariscina*, *Marchantia polymorpha*, *Oryza sativa* (Ouyang et al., 2007), *Arabidopsis thaliana* (Cheng et al., 2017), and *Populus trichocarpa* (Tuskan et al., 2006). Multiple alignments of each family were carried out using Muscle v3.6 (Edgar, 2004). MEGA7 (Kumar et al., 2016) was then used to create maximum likelihood phylogenetic trees; bootstrap values were obtained by 200 bootstrap replicates with the cutoff value set to 50%. The tree was visualized using iTOL (Letunic and Bork, 2019) (https://itol.embl.de/itol.cgi). Each member of TFs/TRs was named according to the relative positions of branches on the phylogenetic tree and manually annotated according to the subfamily clusters and potential functions, predicted by the homologs studied in other species.

Terpenoid biosynthesis pathway analysis

The homologs encoding 22 enzymatic steps of oleoresin terpene biosynthesis in *Arabidopsis thaliana* (Tholl and Lee, 2011) and *Picea glauca* (Warren et al., 2015) served as a reference to identify putative functional homologs in *P. tabuliformis* using BLASTP v2.2.28 with an *e*-value of 1.0 x e^{-30}. Conifer genes encoding terpene synthases (TPS) (Celedon and Bohlmann, 2019; Warren et al., 2015) and P450 family proteins (Celedon and Bohlmann, 2019; Mizutani and Ohta, 2010; Warren et al., 2015) were used as references to screen homologs in *P. tabuliformis* using BLASTP v2.2.28 with an *e*-value 1.0 x e^{-30}. The conserved domains of all resulting proteins were manually checked with CD-Search (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) and SMART (Letunic et al., 2021) (http://smart.embl-heidelberg.de/). Multiple alignments of each family were carried out using Muscle v3.6 (Edgar, 2004). MEGA7 (Kumar et al., 2016) was then used to create maximum likelihood phylogenetic trees; bootstrap values were obtained by 200 bootstrap replicates with the cutoff value set to 50%. The tree was visualized with iTOL (Letunic and Bork, 2019) (https://itol.embl.de/itol.cgi).

Identification of orthologs of known flowering-time regulatory genes

The orthogroups between *P. tabuliformis* and *A. thaliana* were identified using OrthoFinder v2.2.6 (Emms and Kelly, 2015). *P. tabuliformis* orthologs of 306 regulatory genes, whose counterparts in *A. thaliana* are presented in the Flowering Interactive Database (FLOR-ID) (Bouché et al., 2016) (http://www.phytosystems.ulg.ac.be/florid/), are known for their involvement in flowering time regulation in *A. thaliana*. Since *P. tabuliformis* and *A. thaliana* diverged more than 300 million years ago, many mutations have accumulated in protein sequences in the two species. To identify conserved orthologous pairs, Reciprocal Best Hit BLASTP was employed to analyze proteins from these two species; 77 conserved orthologs were identified as having the highest sequence similarities in the reciprocal best-hit analysis. The regulatory roles of these genes are most likely consistent with the angiosperm regulatory network. We then used the RNA-seq data from two groups of 102 samples at reproductive-stage to verify whether the expression level of these 77 conserved orthologs is associated with reproductive development in *P. tabuliformis*. One group included buds of 12 early flowering seedlings (3 year old) and 6 normal flowering seedlings, and the other groups including the needles and male cone buds at three time points in the spring which were collected from 12 earlier and late pollen shedding trees (Niu et al., 2019).

Yeast two-hybrid (Y2H) assay

Total RNA was extracted from male and female buds of *P. tabuliformis* and reverse transcribed into cDNA. Then, we cloned CDS of *PtDAL1* (Pt6G35050), *PtDAL2* (Pt3G43800), *PtDAL10* (Pt2G41770), *PtDAL11* (Pt8G42570), *PtDAL12* (Pt8G42560), *PtDAL13* (Pt8G42540), *PtDAL14* (Pt1G04470), *PtMADS1* (PtJG10100), *PtMADS2* (Pt0G35690), *PtMADS10* (PtJG10040), *PtMADS43* (Pt8G42530) and *PtMADS45* (Pt7G35070) from the cDNA of male or female buds, and then the CDS were inserted into pGBKT7-BD (Clontech, USA, Code No.630443) and pGADT7-AD (Clontech, USA, Code No.630442), respectively. We transferred recombinant plasmids into the yeast strain using Y2HGold (Weidi Biotechnology Co. Ltd, Shanghai, China). All transformants were placed on SD-Leu-Trp plates and incubated for 2d at 30°C. To ensure reliable results, we selected at least six single colonies from the same plate to test the interaction. Interactions were tested on SD-Trp-Leu-His-Ade plates and incubated for 4-5d at 30°C.

Whole genome bisulfite sequencing (WGBS)

The total volume of the combined gDNA sample and unmethylated lambda DNA control was adjusted to 80 μL with 1x TE, and then the DNA was fragmented to 300 bp with an ultrasonic disruptor. The blunt ends of fragments were created by filling in single-stranded overhangs using a mixture of T4 DNA polymerase and the Klenow fragment. In addition, 3′ ends were dA-tailed and fragments were phosphorylated to ensure that fragments were suitable for ligation. This single A overhang enables ligation to adaptors with single T overhangs. Methylated adaptors containing sequences for the downstream sequencing workflow were ligated to the dA-tailed fragments. The 300-600bp fragments were selected by gel electrophoresis and recovered using magnetic bead separation. The bisulfite conversion technique involved treating DNA with bisulfite, during which unmethylated cytosines were converted into uracils. Methylated cytosines remained unchanged during the treatment. The uracil-binding pocket of KAPA HiFi DNA Polymerase is inactivate enabling amplification of uracil-containing DNA. A Qubit® 3.0 Fluorometer was used for the library quantitation for whole genome bisulfite sequencing (WGBS). An Agilent 2100 Bioanalyzer was used to confirm the insert size of libraries. A StepOnePlus Real-Time PCR system was then used to check the molality of libraries (> 10mM). WGBS libraries were sequenced on an Illumina NovaSeq 6000 sequencer with a S4 flow cell as paired-end 150-bp reads.

DNA methylome analysis

All WGBS reads and the lambda DNA (as control) were mapped to the *P. tabuliformis* genome using Bismark v0.20.0 (Krueger and Andrews, 2011). The reads from each biological sample (new shoots) were aligned independently with the specified options (-q–score-min L, 0, −0.2 –directional–ignore-quals–no-mixed–no-discordant–dovetail–maxins 500–bowtie2). Methylated cytosines (Cs) were called from the uniquely mapped reads using BatMeth2 (Lim et al., 2012) under default parameters. Methylation ratios of each cytosine covered by at least five reads were calculated as the number of Cs divided by Cs plus Ts. The bisulfite conversion rate was estimated by lambda genome methylation levels. To calculate the correlation between two biological samples for WGBS data, the *P. tabuliformi*s genome was split into 5 kb bins and the methylation level for each bin was calculated. Then, the Pearson correlation coefficient was calculated for the two biological replicates. For gene and TE methylation analyses, the gene body and upstream and downstream 2kb regions were divided into 20 bins, as 100 bp in each bin for upstream or downstream and gene body of each genes was equally divided into 20 bins. The average methylation level was calculated for each bin and plotted.

To test the correlation of DNA methylation levels with gene expression levels, genes were divided based on their expression levels, and the average methylation level of the gene body and upstream and downstream 2kb regions was calculated and plotted. Since we found that the presence or absence of TE in introns has no significant correlation with gene expression, this analysis did not involve the influence of TE in introns.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data are presented as mean ± SD for all data in figures and Results. All statistical analyses including testing the normality of data distribution were performed using Excel and GraphPad Prism software. The Pearson correlation coefficient was used to summarize the strength of the linear relationship between two data groups. The threshold for significant was set to *P*-value < 0.05. n in the Figures 2, 5, S2, and S4 represents number of genes, n in the Figure S5 represents number of biological replicates. Quantification approaches and statistical analyses used in the genome sequencing and assembly, genome quality assessment, evolutionary analysis and comparative transcriptome analysis can be found in the relevant sections of the Method details.

Supplemental figures

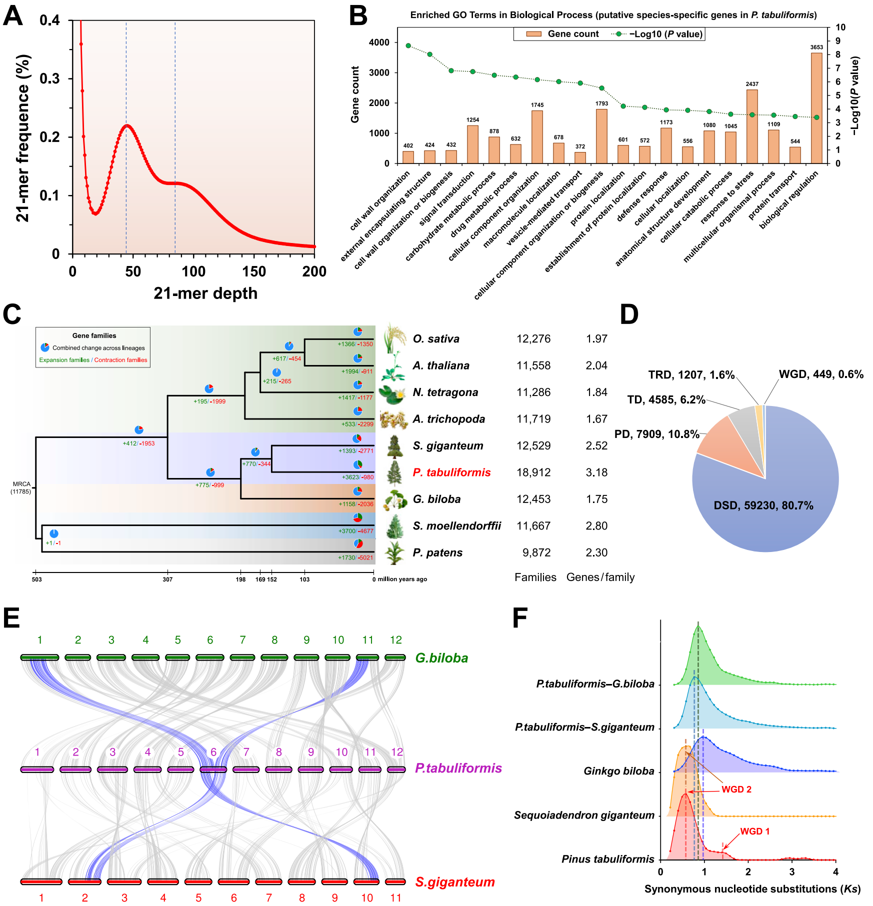


Figure S1. Genome and gene-family evolution *of Pinus tabuliformis*, related to Figure 1

(A) The *k*-mer distribution for genome size estimation and polymorphism of *P. tabuliformis.* (B) Gene ontology (GO) enrichment analysis of 22,281 putative species-specific genes in *P. tabuliformis* which were unassigned to any orthogroups in other 18 selected plant species. (C) The evolution of gene families in *P. tabuliformis*. (D) The number of genes derived from different modes of duplication. WGD whole-genome duplication, TD tandem duplication, PD proximal duplication, TRD transposed duplication, DSD dispersed duplication. (E) Collinearity between the chromosomes of *Ginkgo biloba*, *P. tabuliformis* and *Sequoiadendron giganteum.* Lines depict homologous genome blocks. The blue line indicates an example of chromosome 6 in *P. tabuliformis* which have experienced minor reorganizational exchanges with other chromosomes in *G. biloba* and *S. giganteum.* (F) The distribution of *Ks* values of the syntenic gene pairs within and among species of the *P. tabuliformis*, *G. biloba and S. giganteum.*

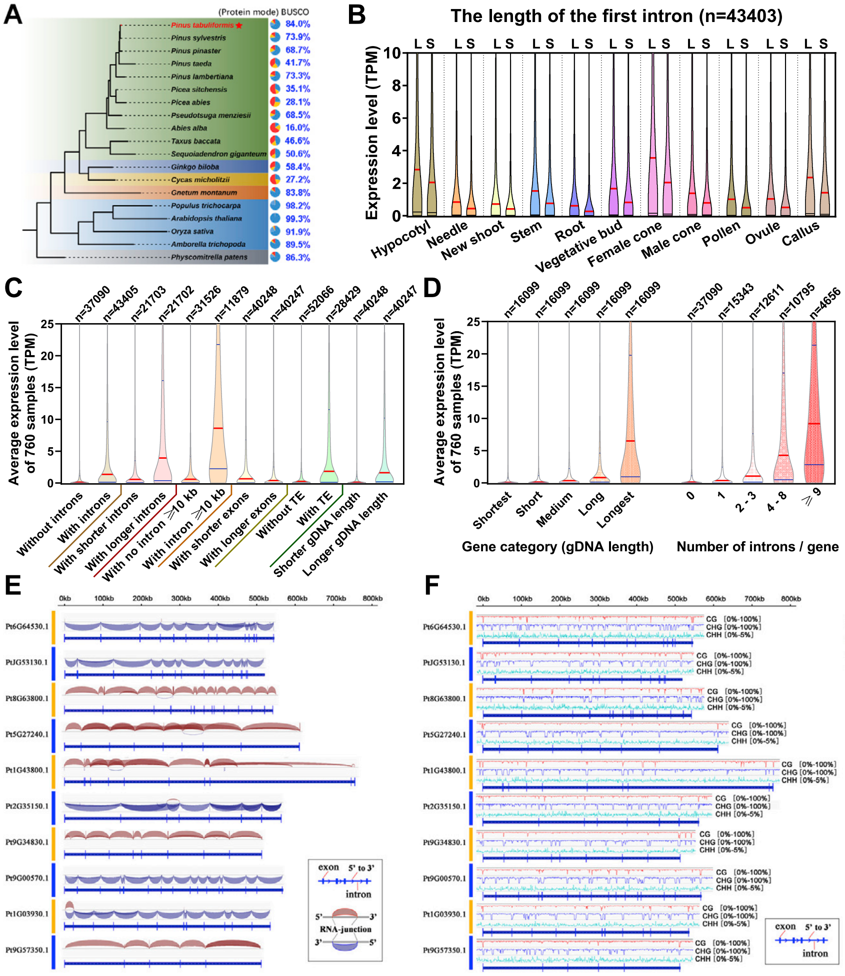


Figure S2. The effect of large genes with ultra-long introns on annotation completeness assessment and transcription, related to Figure 2

(A) The assessment of annotation completeness based on 1614 benchmarking universal single-copy orthologs (BUSCO) using protein mode. The color of pies refer to: complete and single-copy BUSCOs (steel blue), complete and duplicated BUSCOs (light blue), fragmented BUSCOs (yellow), missing BUSCOs (red). (B) Genes with longer first introns tended to have relatively higher expression levels in all 11 tested organs/tissues. All organ/tissue specifically expressed genes were split into two equal groups, based on first intron lengths; the group with longer half of introns is shown at left and the group with smaller half is shown at right. (C, D) The comparison of expression levels between two groups of genes with distinctive structural features. The expression level of each gene was represented by the average expression level calculated based on 760 sample transcriptomes. (E) The RNA-junction data enabled an unbiased structure identification for genes with super long introns. (F) Low CG and CHG methylation acted as recognition markers of exons in super-long genes in *Pinus tabuliformis*.

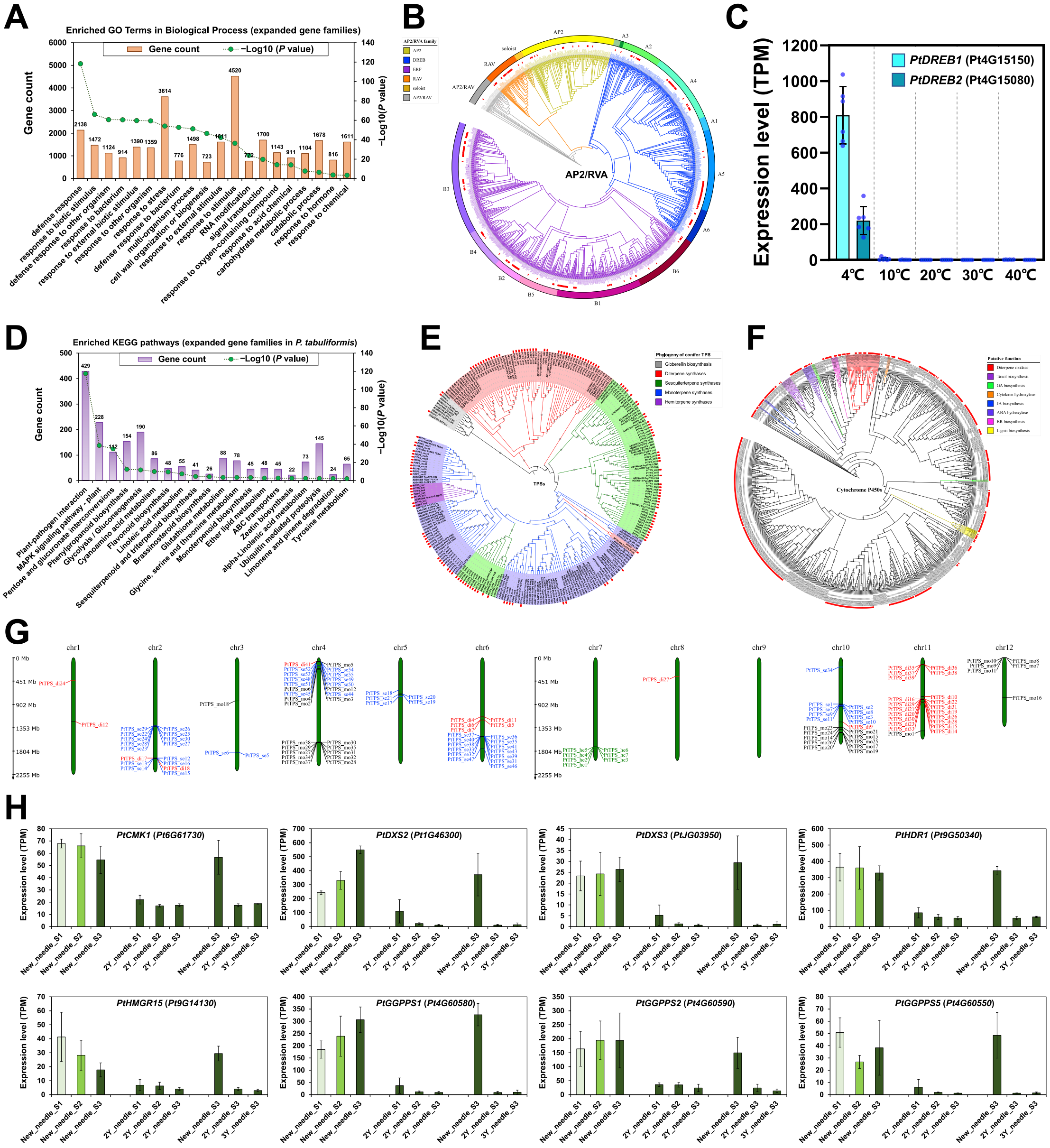


Figure S3. The genomic architecture and expression of terpene biosynthesis pathway genes, related to Figure 3

(A) Gene ontology (GO) enrichment analysis of expanded gene families in *P. tabuliformis*. (B) Maximum likelihood phylogenetic tree of AP2/ERF family proteins in plants. The red bars outside the IDs denote *P. tabuliformis* homologs identified in this project. (C) The cold-specific responses of *PtDREB1*, *2* in *P. tabuliformis*. (D) The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of expanded gene families in *P. tabuliformis*. (E, F) Maximum likelihood phylogenetic tree of terpene synthases (TPS) and cytochrome P450 enzymes (CYP450s) in conifers. (G) Schematic representations delineating the chromosomal distribution of terpene synthases (TPS) genes in *P. tabuliformis*. Black, red, green and blue colored IDs represent monoterpene, diterpene synthases, hemiterpene synthases and sesquiterpene synthases, respectively. (H) The terpene synthases genes were mainly expressed in first year needles in *P. tabuliformis*. New needle represent newly sprouted needles of that year; 2Y_needle and 3Y_needle represent needles sprouted one year earlier and two years earlier, respectively. S1, S2 and S3 indicates time points when the new needles grew to 1/3, 2/3 and the same length of the last year needles, respectively.

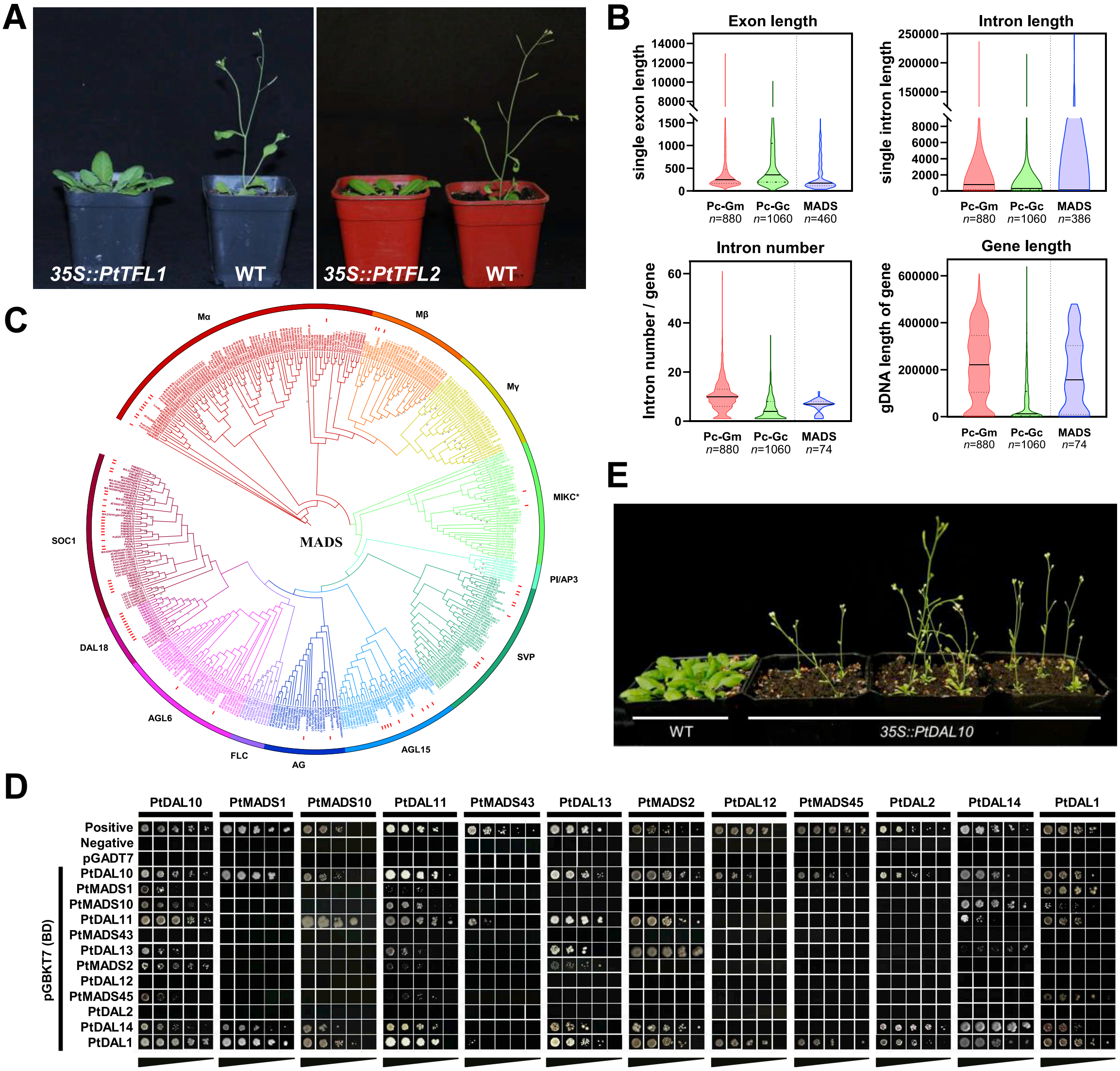


Figure S4. Functional characterization of *PtTFL1-like* and *MADS-box* genes in *Pinus tabuliformis*, related to Figure 4

(A) Effect of ectopic expression of the *PtTFL1*(Pt8G34150) and *PtTFL2* (Pt9G45140) in *Arabidopsis thaliana* (Columbia ecotype). (B) The comparison of gene structure between MADS genes and full-length genes that BUSCO (Benchmarking Universal Single-Copy Orthologs) recognized and failed to recognize in genome mode assessment. Pc-Gm denotes those genes that BUSCO recognized with the protein mode only, but failed to recognize with the genome mode. Pc-Gc denotes those genes that BUSCO recognized with both protein and genome modes. (C) Maximum likelihood phylogenetic tree of MADS-box family proteins in plants. The protein sequences from *Chlamydomonas reinhardtii* (green algae), *Marchantia polymorpha* (liverwort), *Selaginella moellendorffii* (selaginella), *Physcomitrella patens* (moss), *A. thaliana* (herbaceous angiosperm), *Populus trichocarpa* (woody angiosperm) were analyzed. The different colors in the color ring represent different sub-families. The red bars outside the IDs denote *P. tabuliformis* homologs identified in this project. (D) Y2H assays of physical interactions among between 12 reproductive related MADS-box proteins in *P. tabuliformis*. The interactions were tested on SD-Trp-Leu-His-Ade plates and incubated for 4-5 d at 30°C. AD-T and BD-p53 were used as positive control; AD-T and BD-Lam were used as negative control. (E) Effect of ectopic expression of the *PtDAL10* (Pt2G41770) in *A. thaliana*.

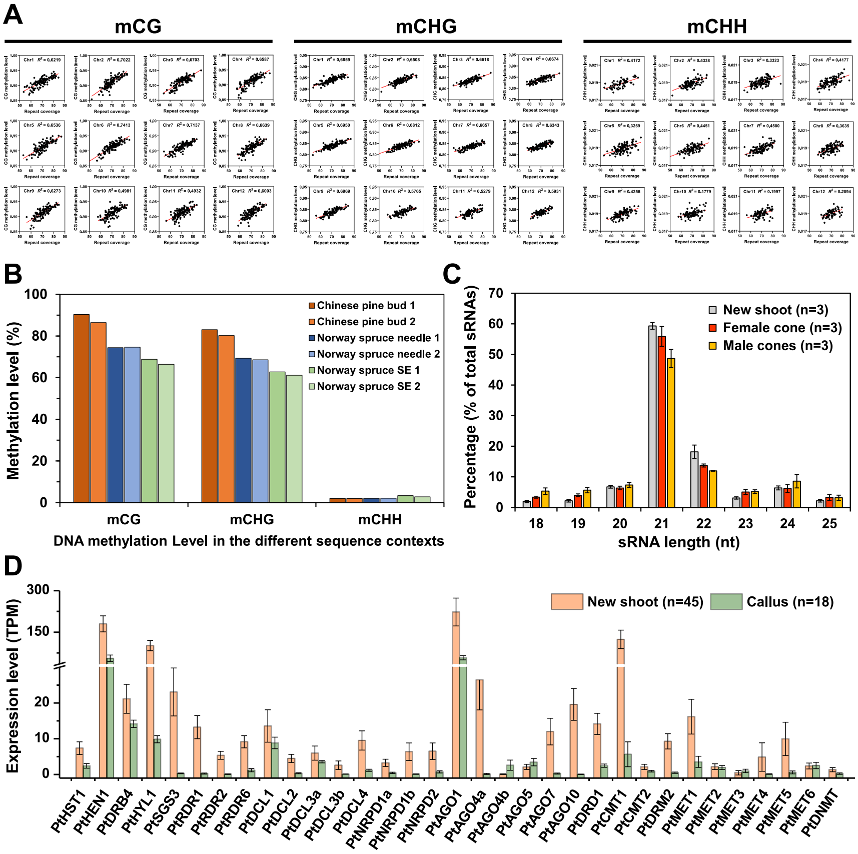


Figure S5. The DNA methylome landscape and pathway genes in *Pinus tabuliformis*, related to Figure 5

(A) Correlation between DNA methylation level and repetitive sequence coverage in 20-Mb width bins across the 12 chromosomes in *P. tabuliformis*. (B) Global average DNA methylation level comparison between *P. tabuliformis* and Norway spruce (data from reference Ausin et al., 2016). SE indicates somatic embryogenesis culture cells. (C) The small RNA (sRNA) length distribution in vegetative and reproductive tissues of *P. tabuliformis*. (D) The expression profiles of all putative DNA methylation pathway genes in *P. tabuliformis*. Noticeably, almost all tested genes had lower expression level in the calli compared with new shoots *in P. tabuliformis*, that may be associated with the lower DNA methylation level in conifer calli than other tissues (Ausin et al., 2016).