nature genetics

Article

https://doi.org/10.1038/s41588-024-01944-y

Haplotype-resolved genome assembly and resequencing analysis provide insights into genome evolution and allelic imbalance in *Pinus densiflora*

Received: 12 May 2023

Accepted: 10 September 2024

Published online: 20 October 2024

Check for updates

Min-Jeong Jang[1,10], Hye Jeong Cho[1,10], Young-Soo Park[1], Hye-Young Lee[2,3], Eun-Kyung Bae[4], Seungmee Jung[5], Hongshi Jin[5], Jongchan Woo[5], Eunsook Park[5], Seo-Jin Kim[1], Jin-Wook Choi[1], Geun Young Chae[1], Ji-Yoon Guk[1], Do Yeon Kim[1], Sun-Hyung Kim[1], Min-Jeong Kang[4], Hyoshin Lee[4], Kyeong-Seong Cheon[4], In Sik Kim[4], Yong-Min Kim[6], Myung-Shin Kim[7], Jae-Heung Ko[8], Kyu-Suk Kang[9], Doil Choi[2], Eung-Jun Park[4] ✉ & Seungill Kim[1] ✉

Haplotype-level allelic characterization facilitates research on the functional, evolutionary and breeding-related features of extremely large and complex plant genomes. We report a 21.7-Gb chromosome-level haplotype-resolved assembly in *Pinus densiflora*. We found genome rearrangements involving translocations and inversions between chromosomes 1 and 3 of *Pinus* species and a proliferation of specific long terminal repeat (LTR) retrotransposons (LTR-RTs) in *P. densiflora*. Evolutionary analyses illustrated that tandem and LTR-RT-mediated duplications led to an increment of transcription factor (TF) genes in *P. densiflora*. The haplotype sequence comparison showed allelic imbalances, including presence–absence variations of genes (PAV genes) and their functional contributions to flowering and abiotic stress-related traits in *P. densiflora*. Allele-aware resequencing analysis revealed PAV gene diversity across *P. densiflora* accessions. Our study provides insights into key mechanisms underlying the evolution of genome structure, LTR-RTs and TFs within the *Pinus* lineage as well as allelic imbalances and diversity across *P. densiflora*.

Recent advances in sequencing technologies with computational methods allow the generation of genome assemblies of individual haplotypes in both animals and plants[1–3]. Haplotype-resolved or allele-aware assembly is an approach to assemble haplotype sequences via accurate separation of allelic variations, which has been omitted in the consensus genome assembly[2,4]. In plants, haplotype-resolved assembly has mainly been implemented for complex genomes such as autopolyploid or highly heterozygous diploid species[5–10]. For example, allele-defined genome sequences revealed accurate chromosomal inversions in autohexaploid sugarcane[5] and allele-specific insertion or deletion of trait-related genes controlling apple color[6] and vanillin compound[7], respectively. Moreover, accurate assessment of allelic variation in the alfalfa genome enabled rapid and precise application of genome-editing methods, suggesting that haplotype-resolved genome data will facilitate functional analysis and breeding-related research in plants[8]. However, haplotype-resolved assembly has thus far only

A full list of affiliations appears at the end of the paper. ✉e-mail: pahkej@korea.kr; ksi2204@uos.ac.kr

been applied to a few angiosperm plants or crop species, even though precise characterization of allelic variation is essential for successful functional and evolutionary studies of the large and complex gymnosperm genomes.

The pine (genus *Pinus*) is a major conifer that contains 113 diverse species predominantly distributed in northern temperate forests[11,12]. *Pinus* species play multiple environmental roles including water conservation, soil stabilization and preservation of natural habitats in the global forest ecosystem. The wood of *Pinus* species has many uses in construction, furniture and manufacturing industries[13,14]. Pine trees would also be one of the resources to mitigate the adverse effects of increased levels of atmospheric carbon dioxide, with long-term impacts on global forest economic and ecological systems[15,16].

Although the genomes of *Pinus* species vary in size from 20 to 40 Gb, they tend to have well-preserved diploid karyotypes (2n = 24)[17,18]. To date, complete draft genomes of *Pinus taeda*[19] and *Pinus lambertiana*[15] and the recently published chromosome-scale genome of *Pinus tabuliformis*[20] have been reported. These genomic data revealed that retrotransposons with large introns played a role in the evolution of the *Pinus* genome and identified candidate genes involved in the pathogen-induced stress response and floral development[15,19,20]. However, many questions remain about the specific components of genome evolution, the comprehensive repertoire and evolutionary process of trait-related genes and haplotype characteristics. To understand these issues, we generated a chromosome-scale and haplotype-resolved genome assembly of *P. densiflora*, the Korean red pine. We demonstrate (1) the genome evolution of *Pinus* species via genome rearrangements and accumulation of specific LTR-RTs, (2) *Pinus*-specific copy number expansion of TF genes and tandem and LTR-RT-mediated duplications of these TFs in *P. densiflora*, (3) haplotype-specific genes and their functional roles in *P. densiflora* and (4) comprehensive allelic diversity across *P. densiflora* accessions.

Table 1 | Genome features and assembly quality of *P. densiflora*

	P. densiflora	
	HA	**HB**
Total length of scaffolds (Mb)	21,738	21,759
Number of scaffolds	2,006	1,849
Scaffold N50 (Mb)	1,792	1,815
Number of gaps per Gb	62.9	64.7
Percentage of Ns in scaffold	0.0006%	0.0006%
Number of contigs	3,370	3,254
Total length of contigs (Mb)	21,738	21,758
Contig N50 (bp)	24,694,369	24,662,732
Contig N80 (bp)	10,773,074	10,665,449
Contig N90 (bp)	5,971,694	6,057,413
Base pair QV	50.3 (HA, 50.2; HB, 50.4)	
k-mer completeness	98.9% (HA, 78.7%; HB, 78.9%)	
Assigned	95.1%	95.5%
Switch errors	2.5%	
LAI	22.0	22.0
Gene number	44,233	44,215
CDS average length (bp)	1,064	1,064
Matched BUSCOs (%)	1,318 (95.9%)	1,311 (95.3%)
Complete BUSCOs (%)	1,252 (91.1%)	1,244 (90.5%)
Missing BUSCOs (%)	57 (4.1%)	64 (4.7%)

CDS, coding sequence; LAI, LTR Assembly Index.

Results

Assembly, annotation and characteristics of *P. densiflora*

We generated a total of 4.35 Tb of raw sequences, including 644 Gb (30×) of PacBio high-fidelity (HiFi) and 1.95 Tb (90×) of Hi-C reads for de novo assembly and 967 Gb (45×) of Illumina paired-end reads with 791 Gb (36×) of 10x Genomics data for quality assessment to construct a chromosome-scale haplotype-resolved assembly representing the *P. densiflora* genome that had approximately an estimated size of 22.7 Gb with a heterozygosity of 1.6% based on 21-mer analyses (Supplementary Table 1 and Extended Data Fig. 1a,b). We performed assembly and phasing using both HiFi and Hi-C reads, generating a total of 21.7 Gb of assemblies with 24.7 Mb of contig N50 in *P. densiflora* haplotypes A and B (HA and HB), respectively (Table 1). After manual curation of order and orientation in chromosomes of each haplotype using a Hi-C map, we verified that 20.7 Gb and 20.8 Gb (95.1% and 95.5%) of HA and HB were anchored to 12 chromosomes ranging in size from 1.2 to 2.3 Gb with 62.9 and 64.7 gaps per 1 Gb, respectively (Fig. 1a, Table 1 and Extended Data Fig. 2a,b). We then evaluated the quality of genome assembly and phasing through multiple approaches. We first estimated base-level accuracy and verified a quality value (QV) of 50.3 and a k-mer completeness of 98.9% (Table 1), indicating high accuracy of the pine genome assembly. To assess phasing accuracy, we verified haplotype-specific *k*-mer frequency data using paired-end and 10x reads and observed evenly bisected peaks and fractions on the haplotype-specific *k*-mer graphs (Extended Data Fig. 1b). This indicates even *k*-mer distribution of paired-end and 10x reads in each haplotype. Moreover, we detected 2.5% of switch errors between haplotypes (Table 1). We predicted a total of 44,233 and 44,215 protein-coding genes in HA and HB, which were found to have benchmarking universal single-copy orthologs (BUSCO) completeness scores of 95.9% and 95.3%, respectively (Table 1). These multiple quality assessments validate the high contiguity and accuracy of the allele-aware chromosome-level assembly and annotation of the *P. densiflora* genome.

To investigate genomic differences between *P. densiflora* and other *Pinus* species[15,19,20], we conducted genome-to-genome alignments between *P. densiflora* HA and three *Pinus* genomes (*P. tabuliformis*, *P. taeda* and *P. lambertiana*) (Fig. 1b and Extended Data Fig. 3a). In total, 23.8 Gb (94%) of the *P. tabuliformis* genome was aligned to 20.7 Gb (95%) of the *P. densiflora* HA genome, suggesting an approximate genome size difference of 3.1 Gb between the genomes of *P. densiflora* and *P. tabuliformis* (Extended Data Fig. 3a). The alignment proportions were gradually decreased within 68% of *P. taeda* and 34% of *P. lambertiana* to 70% and 37% of the *P. densiflora* HA genome, respectively (Fig. 1b and Extended Data Fig. 3a). This indicates increased genomic differences between *P. densiflora* and *P. tabuliformis*, *P. taeda* and *P. lambertiana*, in order. We examined aligned regions of *P. densiflora* HA to the *P. tabuliformis* genome and detected occasional genomic duplication in *P. tabuliformis* (Fig. 1b). For example, the clear genomic duplications on the specific regions in chromosome 9 of *P. tabuliformis* were observed when compared to the corresponding regions in chromosome 9a with the presence of duplicated single-copy orthologous BUSCO genes (Fig. 1b).

When comparing the *P. densiflora* HA and *P. tabuliformis* genomes, we verified genome rearrangements between chromosomes 1 and 3 (Extended Data Fig. 3b). To comprehensively understand the history of genome structure evolution among *Pinus* species, we collected and mapped genetic markers of *P. taeda*[21], *P. densiflora*[22] and *Pinus thunbergii*[23] to *P. densiflora* and *P. tabuliformis* genomes (Fig. 1c and Extended Data Fig. 4a–c). Given the genome structures conserved between *P. taeda* and *P. densiflora*, along with the phylogenetic relationships among *Pinus* species, it appears that the short arms of chromosomes 1 and 3 in *P. taeda* and *P. densiflora* remained conserved but underwent translocation to the short arm of chromosome 3 and translocation with inversion to the long arm of chromosome 1, respectively, in

https://doi.org/10.1038/s41588-024-01944-y

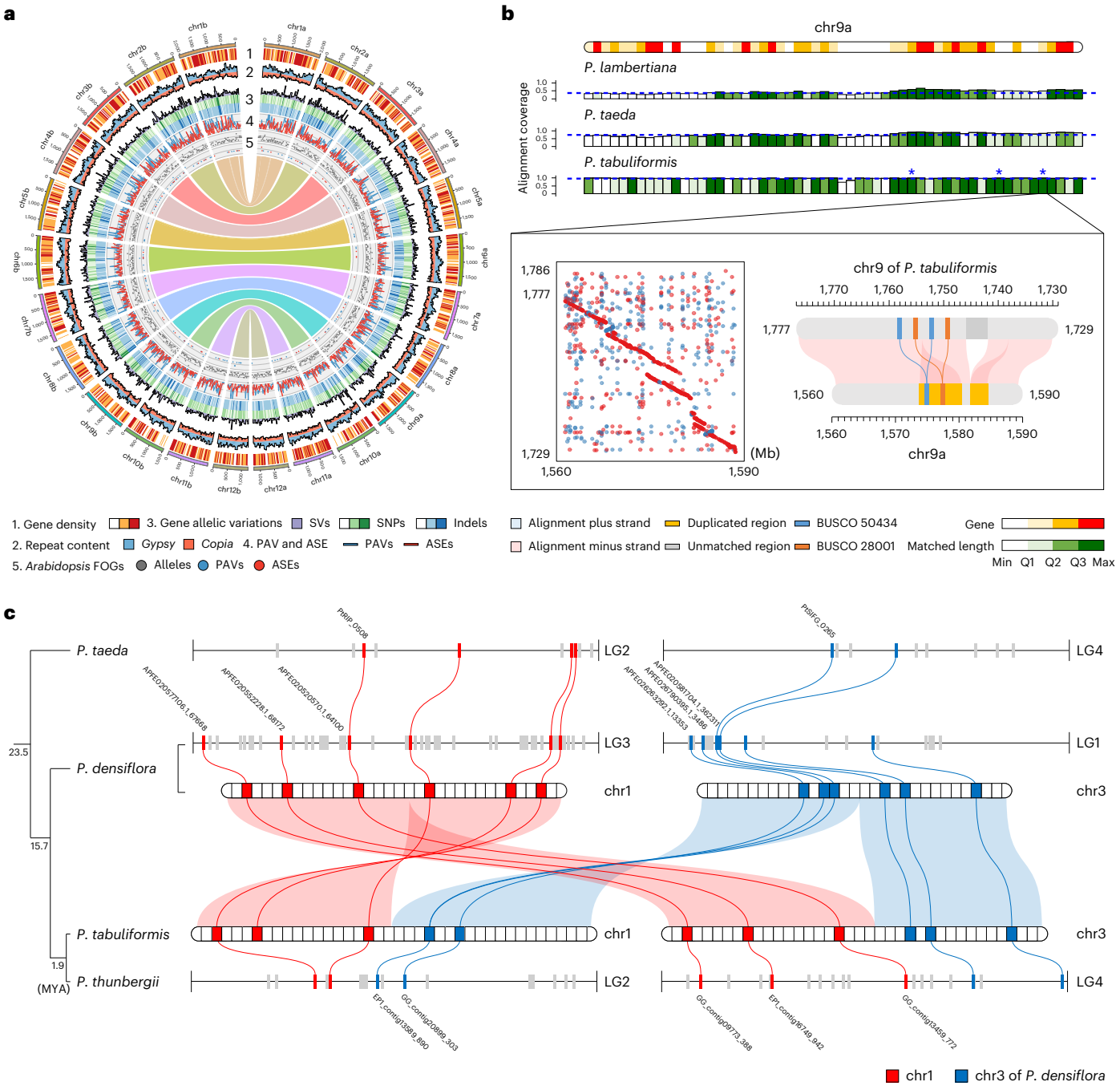


Fig. 1 | Genomic characteristics of *P. densiflora*. **a**, The genomic landscape of HA (right) and HB (left) of *P. densiflora*. The outer track represents the 12 chromosomes (chr). The inner tracks (1–5) represent (1) gene density, (2) repeat content, (3) allelic variations in genic regions, (4) PAV and ASE genes and (5) *Arabidopsis* FOGs using a window size of 30 Mb. **b**, Genome-to-genome alignment of *P. densiflora* HA with other *Pinus* species. The color gradient from white to red represents gene-poor and gene-rich chromosome regions (top). The color gradient from white to green represents the increase in matched alignment length in each *Pinus* genome with the *P. densiflora* genome, while the height of the bar represents the alignment length in the *P. densiflora* genome with other *Pinus* genomes. Q1–Q3, quartiles 1–3; min, minimum; max, maximum. The blue dashed line indicates the total proportion of matched genome sequences in the *P. densiflora* genome from other *Pinus* species. Asterisks (*) denote the top three highly duplicated regions in chromosome 9 of *P. tabuliformis*. Bottom, comparison of specific regions on chromosome 9 between *P. densiflora* HA and *P. tabuliformis* that are duplicated in *P. tabuliformis*. **c**, Genome rearrangements between chromosomes 1 and 3 in *Pinus* species. Physical and genetic maps of *Pinus* species are depicted. The diagram shows representative markers and mapped blocks (with a window size of 60 Mb) in the genomes of *P. densiflora* and *P. tabuliformis*, represented by red or blue colors, respectively, along with marker names. LG, linkage group. In accordance with *P. densiflora*, the red and blue colors for lines, blocks and backgrounds represent chromosomes 1 and 3, respectively.

the shared ancestral lineage of *P. tabuliformis* and *P. thunbergii* after their divergence (Fig. 1c). Our data also suggest an inversion that occurred in the short arm of chromosome 3 in the ancestral lineage of *P. tabuliformis* and *P. thunbergii* while conserving the long arm of chromosome 1 in *P. taeda* and *P. densiflora* (Fig. 1c). These results imply that genome rearrangements in chromosomes 1 and 3 have led to the formation of heterokaryotypes in *P. tabuliformis* and *P. thunbergii*. Taken together, our findings demonstrate evolutionary forces

acting on these *Pinus* species, resulting in variations in genome size and structure.

LTR-RTs driving genome evolution in *Pinus*

LTR-RTs are prominent components of enormous gymnosperm genomes. However, the *Pinus*-specific repertoire of LTR-RTs remains poorly characterized. We identified a total of 13.1 Gb (60% of the total genome) of LTR-RTs in each haplotype of *P. densiflora*, comprising *gypsy* (44%) and *copia* (16%) elements (Supplementary Table 2). Specifically, we found distinct repertories of LTR-RTs in *Pinus* compared to other gymnosperms. The *tat* of *gypsy* and *oryco* of *copia* elements were abundant in *Pinus* (Fig. 2a and Supplementary Table 2). Phylogenetic analysis using LTR-RTs in *P. densiflora* HA and eight other gymnosperm genomes revealed that a large number of specific LTR-RTs clustered as distinct lineages in *Pinus* (Fig. 2b). We verified that *crm*_3, *reina*_3, *athila*_2, *tat*_2 and *tat*_4 of *gypsy* with *tork*_1, *sire*_3, *oryco*_3 and *oryco*_4 of *copia* were particularly abundant in *P. densiflora* and other *Pinus* genomes (Fig. 2b,c). These results suggest that specific LTR-RT subgroups evolved via selective amplification in *P. densiflora* and other *Pinus* species, contributing to the extremely enormous genome size of *Pinus*.

To explore the evolutionary history of *Pinus*-dominant subgroups, we estimated the insertion time of LTR-RTs in *P. densiflora* HA and other *Pinus* genomes. The results show that most of the dominant LTR-RT *gypsy* subgroups expanded in an extremely recent period, whereas the *copia* subgroups have accumulated relatively earlier in *P. densiflora* (Fig. 2d). More specifically, LTR-RTs in *reina*_3, *athila*_2 and *tat*_2 (*gypsy*) and *sire*_3 (*copia*) were massively accumulated 2–6 million years ago (MYA), indicating a recent burst of these subgroups, whereas the major insertions of *crm*_3 and *tat*_4 in *gypsy* and *tork*_1, *oryco*_3 and *oryco*_4 in *copia* took place 16–22 MYA, indicating relatively earlier accumulation of these LTR-RTs (Fig. 2d). Compared to other *Pinus* species, we found lower amounts and earlier insertion periods in the *P. lambertiana* and *P. taeda* genomes than those in the *P. densiflora* and *P. tabuliformis* genomes (Fig. 2c,d). Considering the high contiguity of *P. densiflora* and *P. tabuliformis* genomes containing recent LTR-RTs[24], our data provide insight into the comprehensive evolutionary history of the dominant LTR-RTs, generating extremely large pine genomes via both a recent burst and an earlier insertion.

We explored the genomic distribution of the dominant LTR-RT subgroups on the 12 chromosomes of *P. densiflora* HA to investigate which subgroups contribute to the expansion of gene-rich or gene-poor regions (Fig. 2e and Extended Data Fig. 5a,b). *Tat*_2 and *tat*_4 of *gypsy* mainly accumulated in gene-poor regions, whereas *tork*_1, *oryco*_3 and *oryco*_4 of *copia* were frequently found in gene-rich regions (Fig. 2e and Extended Data Fig. 5a). For instance, *tork*_1, *oryco*_3 and *oryco*_4 on chromosome 10 and *tork*_1 on chromosome 12 were enriched in gene-rich regions, indicating co-localized distribution of genes with these subgroups (Fig. 2e). By contrast, *tat*_2 and *tat*_4 exhibited a negative correlation with the number of genes on the same chromosomes (Fig. 2e). Overall, *tat*_2 tended to be associated with gene-poor regions, while *oryco*_3 and *oryco*_4 were frequently observed in gene-rich regions across many chromosomes (Fig. 2e and Extended Data Fig. 5a). These results suggest that specific *gypsy* and *copia* subgroups contributed to the expansion and diversification of both gene-rich and gene-poor regions in the *P. densiflora* genome. Our findings provide insight into the evolution and diversification of enormous genomes in *P. densiflora* and other *Pinus* species, driven by unbalanced insertions and the evolution of specific LTR-RT lineages.

Evolution of TFs in *Pinus*

To investigate the evolutionary process of gene families in *P. densiflora*, we compared 44,233 genes in *P. densiflora* HA with genes in nine angiosperms and ten gymnosperms including three *Pinus* species (Extended Data Fig. 6a). We found substantially expanded gene families in *P. densiflora* and other *Pinus* genomes. Many genes in these gene families contained conserved domains associated with disease resistance genes (nucleotide-binding and leucine-rich repeats) and TFs, particularly AP2, MYB, NAC, MADS box and LFY TFs (Extended Data Fig. 6b). Interestingly, we observed that genes carrying these conserved domains were abundantly co-localized with transposable elements (TEs), suggesting that their evolution and expansion might have been facilitated by TEs in *Pinus* genomes (Extended Data Fig. 6c). However, the numbers of those TFs were notably higher in the *P. densiflora* genome than the average numbers in other *Pinus* genomes, suggesting that fewer TFs were annotated in other *Pinus* genomes (Extended Data Fig. 6d). Due to the inaccuracy of TF annotations, given the underestimation of conifer TFs in the public database[25] as well as the biased number of TFs in annotations among *Pinus* genomes, we updated annotations of TFs among three other *Pinus* species, seven gymnosperms and nine angiosperms (Supplementary Table 3). We newly identified 21,299 TF genes in 19 species including 10,916 (51%) in three *Pinus* genomes (Supplementary Table 3 and Supplementary Data 1).

On average, we identified a total of 21,376, 15,581 and 21,092 TFs consisting of 48 subfamilies in *Pinus*, seven gymnosperm and nine angiosperm genomes, respectively, indicating a 2.4-fold and 2.2-fold higher number in *Pinus* than in other gymnosperms and angiosperms, respectively (Fig. 3a and Supplementary Table 3). In particular, we found that 11 TF families were significantly over-represented in *P. densiflora* and other *Pinus* species compared to gymnosperms as well as angiosperms, implying a *Pinus*-specific copy number expansion of these TF families (Fig. 3b, Extended Data Fig. 7a and Supplementary Table 3). Subsequently, we conducted phylogenetic analyses of the 11 expanded TFs in 20 plant genomes, revealing that certain TF subgroups, such as AP2_2, AP2_11, MYB_4 and MADS(M)_6 (M type MADS box), formed extremely large *Pinus*-specific lineages (Fig. 3c and Extended Data Fig. 7b). This suggests that the expansion of these TF subgroups has resulted in imbalanced TF repertoires between *Pinus* and other plant lineages.

Our investigation unveiled that the *Pinus*-dominant TF subgroups were mainly generated through tandem duplication (TD) or dispersed duplication (DSD) in the *P. densiflora* genome (Fig. 3d and Supplementary Table 4). For example, the AP2_2, AP2_11, MYB_4, NAC_4 and TCP_2 subgroups contained a higher proportion of tandem duplicated genes than the proportion of TDs in whole genes (Fig. 3d and Supplementary Table 5). The genomic distribution of genes in these TD-dominant subgroups revealed a prevalence of tandem arrays on specific chromosomes (Fig. 3e). By contrast, the MADS(M)_6, NF-YA_2, LFY_2, LFY_3 and LFY_4 subgroups expanded mainly via DSD with scattered gene distribution across the *P. densiflora* genome (Fig. 3d,e). Moreover, we found a remarkable co-localization of genes in the DSD-dominant subgroups with LTR-RTs compared to genes in the TD-dominant subgroups (Fig. 3f). Along with the genomic structures of some genes in the DSD-dominant subgroups, these results suggest that TF genes have undergone massive expansion in *P. densiflora* via LTR-RT-mediated retroduplication (Fig. 3g,h). We detected a clear presence of retrogenes as described in previous studies[26] in the DSD-dominant subgroups with diminished exon numbers compared to parental genes, LTRs and polyA tails (Fig. 3g). Overall, retrogenes were abundant in the DSD-dominant subgroups and had a small number of exons compared to normal genes, suggesting that LTR-RT-driven retroduplication was one of the key evolutionary machineries to create a large number of specific TFs in the *P. densiflora* genome (Fig. 3h). LTR-RT-mediated retroduplication has been exemplified by the tandemly arrayed resistance gene families in pepper[26]. In contrast to the previous study[26], our data illustrate that LTR-RTs have led to the expansion of dispersedly located TFs across chromosomes in *P. densiflora*, suggesting that LTR-RTs facilitate the expansion of diverse gene families in a species-specific manner. Our findings elucidate the expansion of *Pinus* TFs and their tandem and LTR-RT-driven gene duplications in the *P. densiflora* genome, resulting in the diversification of TFs in pine.

https://doi.org/10.1038/s41588-024-01944-y

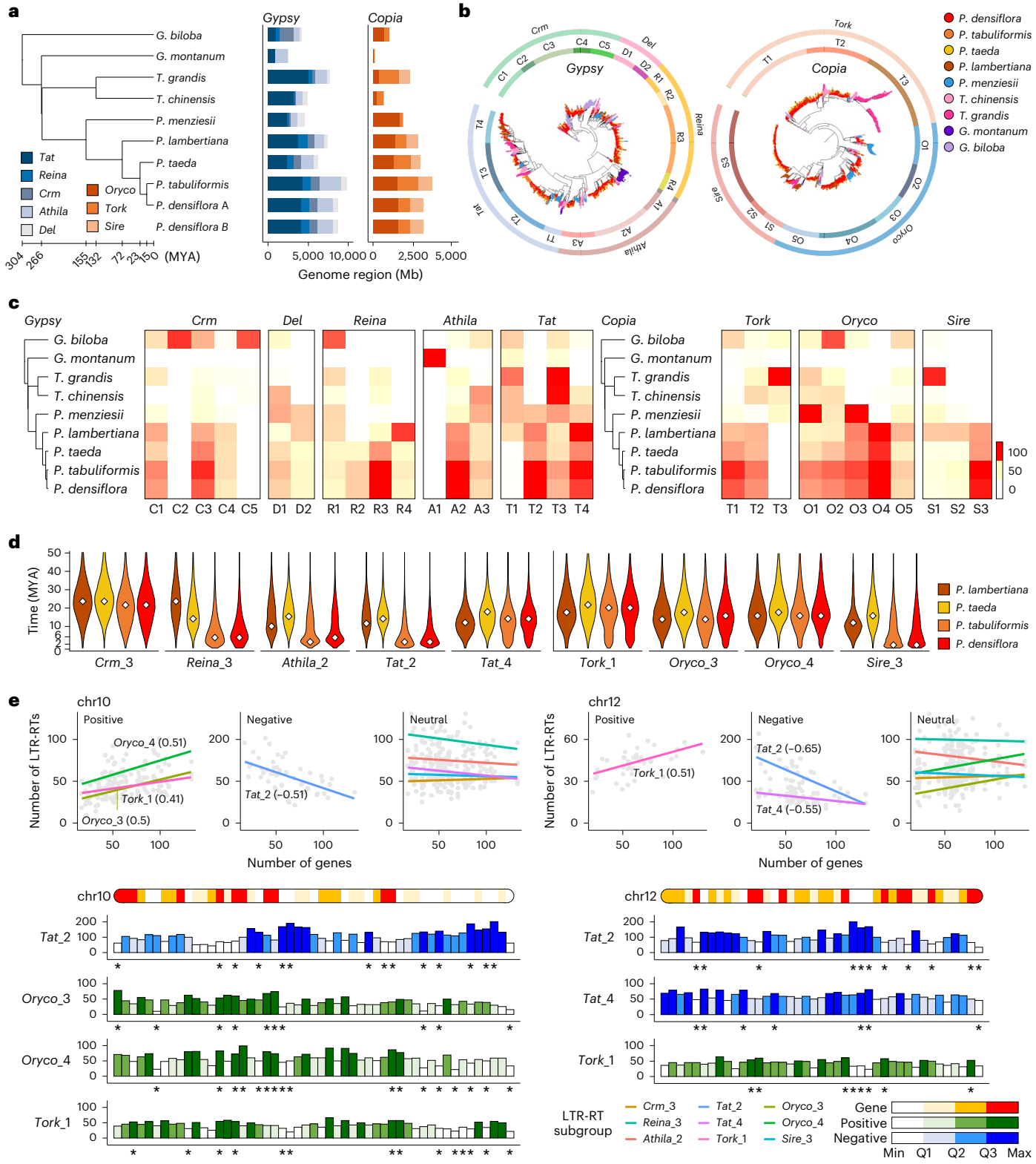


Fig. 2 | Evolutionary dynamics of LTR-RTs in *Pinus*. **a**, Proportions of the *gypsy* and *copia* subfamilies of LTR-RTs in nine gymnosperms. **b**, Phylogenetic relationship of the *gypsy* (left) and *copia* (right) subfamilies in *P. densiflora* HA and other gymnosperms. Colors represent the nine gymnosperms (top right). The outer and inner rings indicate subfamilies and subgroups, respectively. **c**, A heatmap for the number of LTR-RTs in the *gypsy* and *copia* subgroups. **d**, The insertion time (MYA) of dominant subgroups in four *Pinus* species. **e**, The correlation between the number of genes and LTR-RTs in the *gypsy* and *copia* subgroups on chromosomes 10 and 12 of *P. densiflora* HA. For each chromosome, the top correlation plots show LTR-RT subgroups that are positively, negatively or neutrally correlated with gene density. Line colors indicate LTR-RT subgroups. The number of genes (top) and LTR-RTs (bottom) are plotted as density within 30-Mb intervals. The color gradient from white to red, blue and green represents the increased number of genes, negatively correlated LTR-RTs with genes and positively correlated LTR-RTs with genes, respectively. Asterisks (*) denote the major positive and negative regions with genes for each chromosome and subgroup. The scientific names of the species are listed in Methods.

https://doi.org/10.1038/s41588-024-01944-y

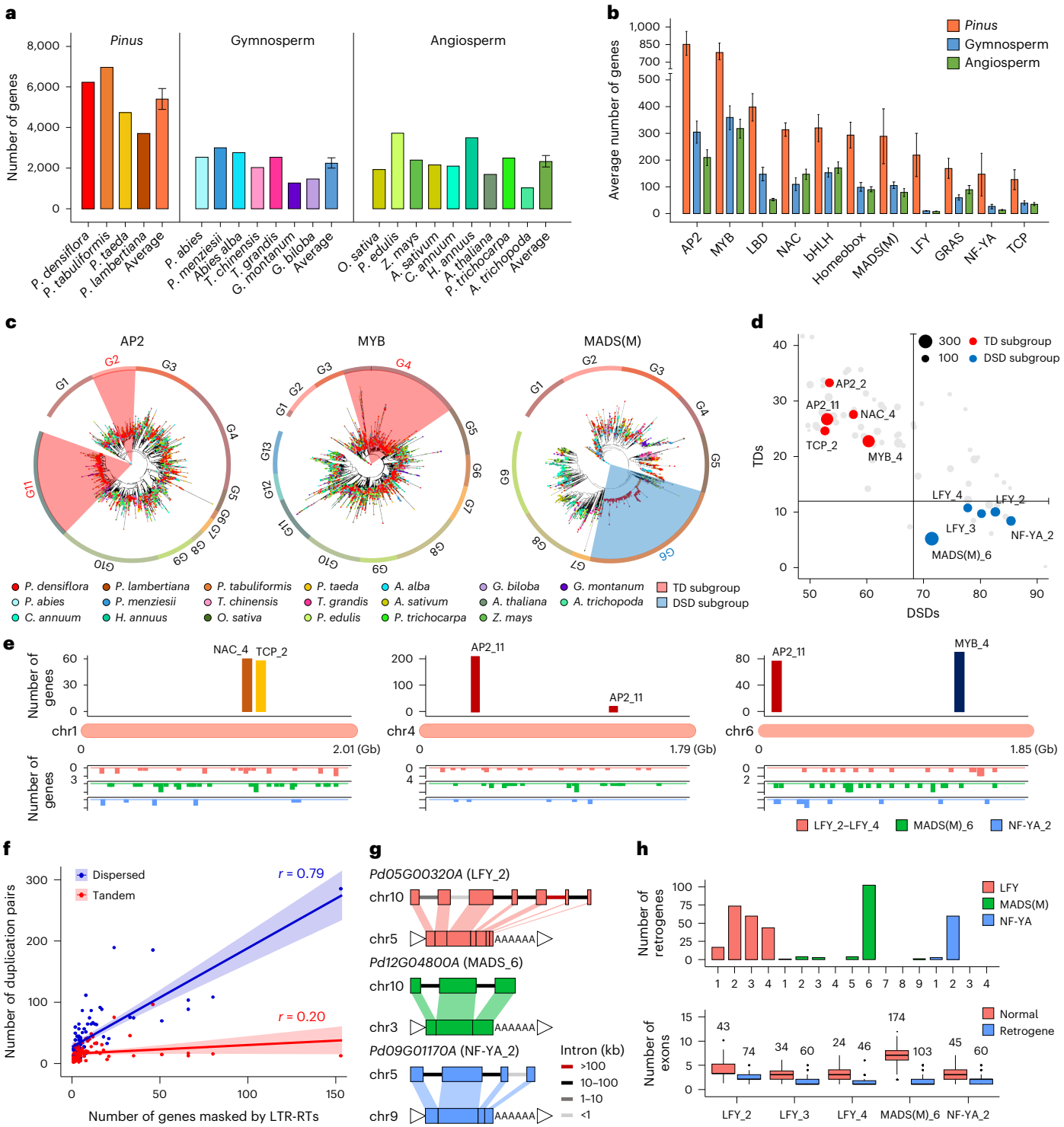


Fig. 3 | Burst of TF gene families by TDs and LTR-mediated DSDs in *P. densiflora*. **a**, The total number of genes in 48 TF families of 20 plant genomes that are categorized into *Pinus*, gymnosperm and angiosperm. **b**, The average number of genes in 11 TF families is significantly higher in *Pinus* based on a one-sided Fisher's test ($P < 0.001$). **a**,**b**, *Pinus*, $n = 4$; gymnosperm, $n = 7$; angiosperm, $n = 9$. Individual data points are listed in Supplementary Table 3. Error bars indicate s.e.m. **c**, The phylogenetic relationships of AP2, MYB and MADS(M) in 20 plant species. Dot colors on the nodes represent species, while the ring indicates subgroups. **d**, The ratio of TDs and DSDs of *Pinus*-enriched subgroups in *P. densiflora*. The ratios of DSDs and TDs in whole genes are depicted on the *x* and *y* axes, respectively. The number of TFs in each subgroup is represented by the relative dot size. **e**, Chromosomal distribution of tandem (top) or dispersed (bottom) duplicated subgroups on chromosomes 1, 4 and 6. **f**, The correlation between the number of genes co-localized with LTR-RTs and duplication pairs in the same TF subgroups. The red and blue colors in dots, lines and backgrounds indicate TD and DSD subgroups, respectively. The correlation coefficients (*r*) between the number of genes co-localized with LTR-RTs and duplication pairs are displayed. **g**, Specific examples of retrogenes in dispersed duplicated TFs. Boxes and lines indicate exons and introns, respectively. **h**, The number of retrogenes in the LFY, MADS(M) and NF-YA subgroups (top). A comparison of the number of exons between retrogenes and normal genes in DSD-dominant subgroups (bottom). Box plots represent the 25th percentile, the median and the 75th percentile, with whiskers from the minimum to the maximum. The number of compartments is shown above the box plots. The scientific names of the species are listed in Methods.

Allelic variations between two haplotypes

We identified a total of 6,915,720 and 6,937,699 variations in 31,277 and 31,243 gene regions of *P. densiflora* HA and HB, respectively (Extended Data Fig. 8a). Similar to heterozygous plant genomes[27], 29,492 (52%) and 29,411 (52%) genes in *P. densiflora* HA and HB had nonsynonymous, indel and/or structural variations (SVs) in the coding sequences (Extended Data Fig. 8a). This suggests that the amino acid alteration resulting from these variations led to the generation of gene imbalances between the two haplotypes. Next, we identified 2,634 and 2,616 PAV genes in HA and HB, respectively, and 1,324 allele-specific expressed (ASE) genes in both haplotypes in addition to 40,275 common allelic genes (Extended Data Fig. 8b and Supplementary Table 6). Gene ontology (GO) descriptions and domain repertoires revealed that the majority of PAV and ASE genes have potential involvement in biological processes relevant to biotic and abiotic stresses, including TFs and disease resistance genes containing MYB, AP2 and leucine-rich repeat domains (Extended Data Fig. 8c,d). We then verified the sequences of several PAV TFs and their corresponding regions in the *P. densiflora* genome (Extended Data Fig. 9a). We observed frameshift mutations leading to the generation of premature stop codons in MADS box, NAC, MYB, NF-YA and AP2 TFs in certain haplotypes, therefore indicating the haplotype-specific presence of these genes in the opposite haplotype (Extended Data Fig. 9a). Validation experiments of representative ASE genes revealed significant haplotype-preferred expression, consistent with the expression pattern observed in the RNA-seq data (Extended Data Fig. 9b).

Utilization of orthologous genes from model plants accelerates the identification of important trait-related genes in relatively less-characterized plant species such as gymnosperms[28]. We detected 2,344 and 2,367 functional orthologous genes (FOGs) of *Arabidopsis* in HA and HB, respectively, as potential candidates for functional genes in *P. densiflora* (Extended Data Fig. 9c and Supplementary Table 7). Among them, FOGs consisted of 173 PAV genes, including 75 in HA and 98 in HB, 69 ASE genes and 2,200 common alleles, respectively (Extended Data Fig. 9c). This indicates that there are 144 distinct FOGs in HA and 167 in HB. GO analysis detected diverse cellular functions and multicopy gene families among these FOGs, including protein kinases, leucine-rich repeats and TFs (Extended Data Fig. 9d). Our findings emphasize that haplotype-resolved genome assembly enables the identification and characterization of allelic variation and haplotype-specific patterns of gene expression in the large diploid genome of *P. densiflora*.

Functional relevance of TFs and *Arabidopsis* FOGs

Given the presence of many TFs in the repertoire of PAV genes, we first conducted functional experiments on two validated PAV TFs, MADS box and NAC, to elucidate their contributions to biological traits in *P. densiflora* (Fig. 4a,b). *Pd03G22920A* (MADS box) was specifically annotated in HA, where a 1-bp insertion (A) in the first exon led to premature translation termination in the corresponding region (Extended Data Fig. 9a). To clarify the function of *Pd03G22920A* in *P. densiflora*, we generated transgenic *Arabidopsis* plants overexpressing *Pd03G22920A* (*Pd03G22920A-ox*). Three independent alleles of *Pd03G22920A-ox* *Arabidopsis* exhibited earlier flowering than the non-transgenic wild type (Fig. 4a). To validate these morphological differences, transgenic T_2 lines were generated, and stably inherited flowering phenotypes were obtained with a significantly reduced number of leaves (Fig. 4a). This indicates that *Pd03G22920A* may play a key role in the flowering characteristics of *P. densiflora*. Next, we performed a comprehensive comparison of protein structure and amino acid sequences between known *Pinus* functional genes and their orthologous genes in *P. densiflora*. Of them, *PdNAC3* (*Pd05G27460A*), the orthologous gene of *PpNAC3* (*Pinus pinaster* NAC3)[29], was specifically identified in *P. densiflora* HA (Fig. 4b). Due to a nonsynonymous mutation (C/T) between HA and HB, premature translation termination in the third exon of *PdNAC3* in *P. densiflora* HB was generated (Extended Data Fig. 9a). Although the C-terminal region of NAC TFs are intrinsically disordered, which can cause a decrease in template modeling (TM) score due to the inability to align predicted structures between protein pairs[30,31], the TM score of the entire proteins encoded by *PpNAC3* and *PdNAC3* was 0.50, indicating a highly conserved protein structure (Fig. 4b). Furthermore, the amino acid similarity between these proteins was 95.3%, and the TM score for the N-terminal region containing the NAM domain region (DR) was 0.99, indicating a high probability of functional equivalence with *PpNAC3* (Fig. 4b). In addition, we observed that *PdNAC3* was abundantly expressed in roots and significantly induced under abiotic stress conditions such as salinity and cold, consistent with the expression pattern of *PpNAC3* (ref. 29) (Fig. 4b). These collective findings suggest a comparable role for PdNAC3 in *P. densiflora*. In addition, we also conducted overexpression analysis of the AP2 allele gene (*Pd03G15540*) in *Arabidopsis* and observed a prominent late flowering phenotype (Fig. 4c). Specifically, *Pd03G15540-ox* plants exhibited a notable delay in the transition from vegetative to reproductive stages compared to both wild types at the same time point (7 weeks old) and at a later time point (4 weeks old). This finding suggests a potential role of *Pd03G15540* in the regulation of flowering in *P. densiflora*.

To speculate about the function of *Arabidopsis* FOGs in *P. densiflora*, we selected orthologous genes of *AtTOM5* (*AT5G08040*), *AtTOM6* (*AT1G49410*), *AtEXPA15* (*AT2G03090*) and *AtABCI12* (*AT3G21580*) to use established function–location data in *Arabidopsis*[32,33] and performed subcellular localization analysis of the corresponding proteins. The results showed that both PdTOM5 (*Pd06G27970*) and PdTOM6 (*Pd09G30900*) co-localized with mitochondrial markers, while PdEXPA15 (*Pd05G28920*) and PdABCI12 (*Pd02G22940*) co-localized with the plasma membrane marker and chloroplast marker in *Nicotiana benthamiana* leaves, respectively (Fig. 4d). This indicates that their association with different cellular compartments is consistent with the subcellular distribution patterns observed for *Arabidopsis* gene products[34–37]. Taken together, our multiple approaches to characterize the gene function of PAV TFs as well as allele TF and *Arabidopsis* FOGs highlight the practical importance of the haplotype-resolved *P. densiflora* genome as an invaluable resource for functional genomics.

PAV gene diversity across wild accessions of *P. densiflora*

Haplotype genome-based resequencing analysis facilitates precise detection of allelic variations, particularly beneficial for species with a large genome[38,39]. To explore allelic imbalance across *P. densiflora* species, we conducted haplotype-informative variation detection by integrating linear and graph-based mapping of 8.8 Tb of resequencing data from 30 wild accessions (13.5× coverage per individual) aligned to the haplotype-resolved *P. densiflora* genome (Fig. 5). When we examined the number of allele, PAV and absent reference genes across accessions by calculating the cumulative count of genes, we observed a gradual decrease in allele genes accompanied by a corresponding increase in PAV genes (Fig. 5a). Specifically, we verified 21,216 allele genes consistently present in all accessions, while 20,374 allele genes exhibited imbalanced patterns and transformed into 20,162 PAV genes across one or more accessions, resulting in the generation of 25,412 pan-PAV genes in 31 accessions, including the reference genome (Fig. 5a and Supplementary Table 8). This implies that many allele genes could exist as PAV genes in different accessions, highlighting the importance of the haplotype genome-based variation analysis to accurately capture the pool of allelic variation. Domain repertoires of PAV genes revealed that changing from allele to PAV genes was predominantly observed in multicopy gene families, playing crucial roles in biological processes, such as disease resistance genes and TFs, including leucine-rich repeat, PPR repeat family, MYB and AP2 genes (Fig. 5b). Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis illustrated that several key genes related to metabolic pathways, such as those for proline and acetate, are present as allele genes in the reference genome; however, they were altered to PAV genes in one or more accessions (Fig. 5c). These data indicate that haplotype-informative variation detection enabled prediction

https://doi.org/10.1038/s41588-024-01944-y

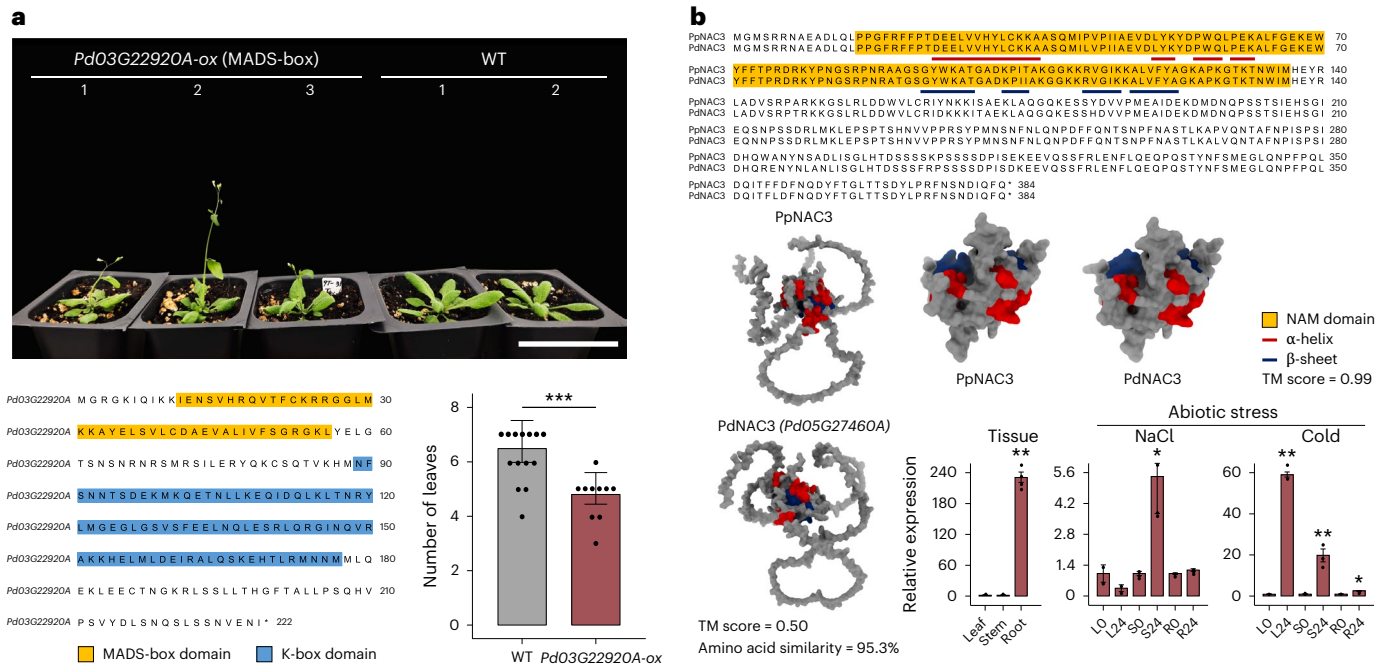


Fig. 4 | Functional characterization of PAV TFs and *Arabidopsis* FOGs in *P. densiflora*. **a**, The effect of ectopic expression of *Pd03G22920A* (MADS box) in *Arabidopsis* (top). The amino acid sequences encoded by *Pd03G22920A* (bottom left) and the number of rosette leaves in the *Pd03G22920A-ox* transgenic T_2 line and the non-transgenic wild type (bottom right) are represented. Asterisks (***) denote a significance level of $P = 0.0006$ based on a two-sided unpaired Student's *t*-test (wild type (WT), $n = 14$; *Pd03G22920A-ox*, $n = 10$). Error bars indicate s.d. Scale bars, 5 cm. **b**, Comparison of amino acid sequence similarity (top) and protein structure prediction for the entire region and DR (middle) of PpNAC3 and PdNAC3 (*Pd05G27460A*). Expression of *PdNAC3* (bottom) in each tissue and under different stress conditions in *P. densiflora*. Expression levels in the leaf tissue and at 0 h were set to 1. Asterisks (*, **) denote significance levels of $P < 0.05$ and $P < 0.01$, respectively, based on a one-sided unpaired Student's *t*-test. At least two biological replicates were used. Error bars indicate s.e.m. L0, leaf 0 h; L24, leaf 24 h; S0, stem 0 h; S24, stem 24 h; R0, root 0 h; R24, root 24 h. **c**, Late flowering and leaf morphology phenotypes of *Pd03G15540* (AP2) in stable transgenic *Arabidopsis* of the gene under the 35S overexpressing promoter. Scale bars, 5 cm. **d**, Subcellular localization of *Arabidopsis* proteins encoded by FOGs fused to fluorescence proteins, including PdTOM5–mRuby3 (*Pd06G27970*), PdTOM6–mRuby3 (*Pd09G30900*), PdEXPA15–moxVenus (*Pd05G28920*) and PdABCI12–moxVenus (*Pd02G22940*) transiently expressed in tobacco leaf epidermal cells (PdTOM5, $n = 16$; PdTOM6, $n = 15$; PdEXPA15, $n = 18$; PdABCI12, $n = 22$). At least two biological replicates were used. Scale bars, 20 μm.

of PAV genes in accessions that can be functionally important. This approach also demonstrates efficiency, particularly for complex giga genomes to avoid omission of haplotype-specific variation.

To explore allele and PAV gene diversity across *P. densiflora* species, we investigated the genome-wide distribution of allele and PAV reference genes in the *P. densiflora* genome and 30 accessions (Fig. 5d and Extended Data Fig. 10). When examining the density level of PAV genes, we observed that many allele genes in the reference genome appear as PAV genes throughout chromosome regions in accessions (Fig. 5d and Extended Data Fig. 10). Observing PAV-rich regions with high PAV gene density, we noticed that these regions were sporadically verified in specific or most accessions (Fig. 5d).

https://doi.org/10.1038/s41588-024-01944-y

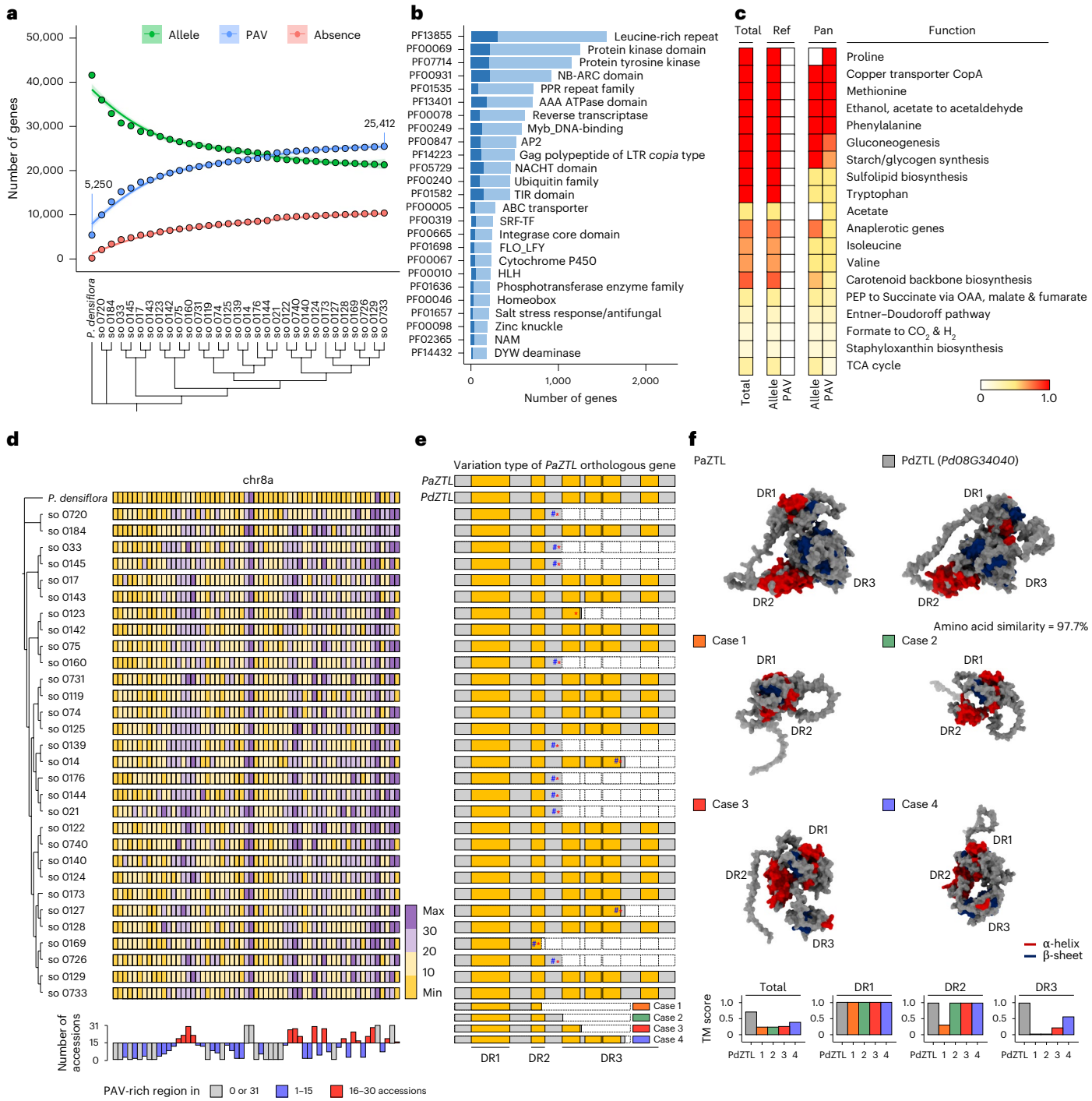


Fig. 5 | Allele and PAV gene diversity across 30 wild accessions of *P. densiflora*. **a**, The cumulative number of allele, PAV and absent genes is plotted in a graph ordered by the phylogenetic relationships among *P. densiflora* and 30 wild accessions. The green, blue and red dots indicate allele, PAV and absent genes, respectively. **b**, The domain repertoire of pan-PAV genes in *P. densiflora* and 30 wild accessions. The dark blue and blue bars represent the number of PAV genes in the reference genome and genes that have transitioned from allele to PAV genes in one or more accessions, respectively. **c**, A heatmap for metabolic pathway completeness of the reference (ref) and pan genes based on KEGG annotation. The color scale from white to red indicates low to high pathway completeness. OAA, oxaloacetate; PEP, phosphoenolpyruvate; TCA, tricarboxylic acid cycle. **d**, Allelic and PAV gene distributions for chromosome 8a of *P. densiflora* HA and 30 wild accessions. The color gradient from yellow to purple represents the ratio of PAV genes within each accession, based on a window size of 30 Mb (top). The bar graph illustrates the number of accessions with a high density of PAV regions (>20% PAV ratio of total genes within the respective window) on chromosome 8a (bottom). **e**, Variation types of *PdZTL* (*Pd08G34040*) in *P. densiflora* and 30 wild accessions. Yellow boxes represent the DRs, while the dashed boxes indicate the truncated regions. Frameshift mutations and stop codons are denoted by hashtags (#) and asterisks (*), respectively. **f**, Predicted protein structures of PaZTL, PdZTL and hypothetical cases distinguished by four variation types (top). The graph shows the TM scores for the total region and each DR of PdZTL and four hypothetical cases. TM scores were calculated by comparing each DR with those of PaZTL.

This suggests that PAV gene diversity across accessions has primarily been generated mainly by conversion of allele to PAV genes in PAV-rich regions. For example, it was reported that overexpressing PaZTL (*Picea abies* ZTL), which contains three functional domains, LOV (DR1), F box (DR2) and Kelch motifs (DR3), inhibits flowering in transgenic *Arabidopsis*[40]. The orthologous gene of *PaZTL*, *PdZTL* (*Pd08G34040*), exists as an allele in the reference genome. However, this gene was predicted to be a PAV gene in many accessions due to a variety of variation types leading to premature termination, thus excluding its crucial domain (Fig. 5e). The variation types were categorized into four cases based on the location of the premature stop codon caused by the frameshift mutation or stop acquisition. Protein structure and sequences of PaZTL and PdZTL were found to be very similar, showing a TM score of 0.71 for the entire region, with 97.8% amino acid similarity and an average TM score of 0.99 for the three DRs (Fig. 5f). This suggests highly conserved protein sequences and structures between the proteins. By examining the protein structural features for each variation type, it was found that the Kelch motif structure in DR3, one of the functional domains, is not formed in cases 1 and 2 with TM scores of 0.23 and 0.24, respectively (Fig. 5f). In cases 3 and 4, the Kelch motif structure in DR3 is partially formed with one and three Kelch repeats, respectively, with TM scores of 0.25 and 0.39. This suggests that the incomplete protein structure, lacking a stable β-propeller in functional DRs, may contribute to abnormal gene formation in one haplotype of some *P. densiflora* accessions. Our findings highlight variable allele and PAV gene diversity among *P. densiflora* accessions and demonstrate the necessity of allele-aware variation studies using accurate haplotype-resolved assembly, especially for complex giga genomes.

Discussion

Allele-defined genome-based research is becoming increasingly feasible and indispensable for comprehensive analysis of extremely large and complex plant genomes. Here, we present a haplotype-resolved assembly of the *P. densiflora* genome with high contiguity, omission-reduced gene annotation and accurate haplotype separation. This represents an important stride toward enabling effective haplotype-resolved functional and breeding studies of gymnosperm genomes. We comprehensively demonstrated that genomic duplications, chromosomal rearrangements and the expansion of specific LTR-RT subgroups have contributed to the diversity in genome size and structure among *Pinus* species. The intensive comparison of TF families in *Pinus*, gymnosperms and angiosperms using updated annotations unveiled remarkably expanded TF repertoires in *Pinus* and their evolutionary processes via a burst of specific TF families by TD and LTR-RT-mediated DSD in the *P. densiflora* genome. Our findings with the updated resources provide insight into the mechanism underlying extreme copy number evolution of important gene families. Characterization of haplotype variation elucidated allelic imbalances between haplotypes with their functional roles in shaping *P. densiflora* traits, such as flowering regulation and abiotic stress resistance. Moreover, haplotype-aware variant detection allowed us to construct an unbiased pool of allelic information across *P. densiflora* accessions, suggesting the importance of a haplotype genome-based approach to accurately understand gene diversity among individuals within species. Finally, our findings with the haplotype-resolved genome assembly provide crucial insights for a comprehensive understanding of *Pinus* genome evolution and haplotype characteristics in *P. densiflora*.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01944-y.

References

1. Garg, S. et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* **39**, 309–312 (2021).
2. Guk, J. Y., Jang, M. J., Choi, J. W., Lee, Y. M. & Kim, S. De novo phasing resolves haplotype sequences in complex plant genomes. *Plant Biotechnol. J.* **20**, 1031–1041 (2022).
3. Garg, S. Computational methods for chromosome-scale haplotype reconstruction. *Genome Biol.* **22**, 101 (2021).
4. Kong, W., Wang, Y., Zhang, S., Yu, J. & Zhang, X. Recent advances in assembly of plant complex genomes. *Genomics Proteomics Bioinformatics* **21**, 427–439 (2023).
5. Zhang, J. et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* **50**, 1565–1573 (2018).
6. Sun, X. P. et al. Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat. Genet.* **52**, 1423–1432 (2020).
7. Hasing, T. et al. A phased *Vanilla planifolia* genome enables genetic improvement of flavour and production. *Nat. Food* **1**, 811–819 (2020).
8. Chen, H. et al. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat. Commun.* **11**, 2494–2504 (2020).
9. Sun, H. et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat. Genet.* **54**, 342–348 (2022).
10. Zhou, Q. et al. Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat. Genet.* **52**, 1018–1023 (2020).
11. Farjon, A. & Filer, D. *An Atlas of the World's Conifers: an Analysis of their Distribution, Biogeography, Diversity and Conservation Status* (Brill, 2013).
12. Jin, W.-T. et al. Phylogenomic and ecological analyses reveal the spatiotemporal evolution of global pines. *Proc. Natl Acad. Sci. USA* **118**, e2022302118 (2021).
13. Moctezuma Lopez, G. & Flores, A. Economic importance of pine (*Pinus* spp.) as a natural resource in Mexico. *Rev. Mex. Cienc. Forestales* **11**, 161–185 (2020).
14. Murthy, R., Dougherty, P. M., Zarnoch, S. J. & Allen, H. L. Effects of carbon dioxide, fertilization, and irrigation on photosynthetic capacity of loblolly pine trees. *Tree Physiol.* **16**, 537–546 (1996).
15. Stevens, K. A. et al. Sequence of the sugar pine megagenome. *Genetics* **204**, 1613–1626 (2016).
16. Pan, Y. et al. A large and persistent carbon sink in the world's forests. *Science* **333**, 988–993 (2011).
17. Kirst, M. et al. Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **100**, 7383–7388 (2003).
18. OBrien, I. E. W., Smith, D. R., Gardner, R. C. & Murray, B. G. Flow cytometric determination of genome size in *Pinus*. *Plant Sci.* **115**, 91–99 (1996).
19. Neale, D. B. et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* **15**, R59–R71 (2014).
20. Niu, S. et al. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* **185**, 204–217 (2022).
21. Echt, C. S. et al. An annotated genetic map of loblolly pine based on microsatellite and cDNA markers. *BMC Genet.* **12**, 17 (2011).
22. National Institute of Forest Science. SNP marker set for individual identification and population genetic analysis of *Pinus densiflora* and their use. KR patent 1020200045790 (2021).
23. Hirao, T. et al. Construction of genetic linkage map and identification of a novel major locus for resistance to pine wood nematode in Japanese black pine (*Pinus thunbergii*). *BMC Plant Biol.* **19**, 424 (2019).

24. Liu, H. L. et al. The nearly complete genome of *Ginkgo biloba* illuminates gymnosperm evolution. *Nat. Plants* **7**, 748–763 (2021).
25. Jin, J. et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* **45**, D1040–D1045 (2017).
26. Kim, S. et al. New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. *Genome Biol.* **18**, 210–220 (2017).
27. Wang, P. et al. Genetic basis of high aroma and stress tolerance in the oolong tea cultivar genome. *Hortic. Res.* **8**, 107 (2021).
28. Berardini, T. Z. et al. The *Arabidopsis* Information Resource: making and mining the 'gold standard' annotated reference plant genome. *Genesis* **53**, 474–485 (2015).
29. Pascual, M. B., Canovas, F. M. & Avila, C. The NAC transcription factor family in maritime pine (*Pinus pinaster*): molecular regulation of two genes involved in stress responses. *BMC Plant Biol.* **15**, 254 (2015).
30. Kang, M. et al. The C-domain of the NAC transcription factor ANAC019 is necessary for pH-tuned DNA binding through a histidine switch in the N-domain. *Cell Rep.* **22**, 1141–1150 (2018).
31. Chakravarty, D. & Porter, L. L. AlphaFold2 fails to predict protein fold switching. *Protein Sci.* **31**, e4353 (2022).
32. Millar, A. H., Carrie, C., Pogson, B. & Whelan, J. Exploring the function–location nexus: using multiple lines of evidence in defining the subcellular location of plant proteins. *Plant Cell* **21**, 1625–1631 (2009).
33. Han, J. et al. All-in-one: a robust fluorescent fusion protein vector toolbox for protein localization and BiFC analyses in plants. *Plant Biotechnol. J.* **20**, 1098–1109 (2022).
34. Murcha, M. W., Kubiszewski-Jakubiak, S., Wang, Y. & Whelan, J. Evidence for interactions between the mitochondrial import apparatus and respiratory chain complexes via Tim21-like proteins in *Arabidopsis*. *Front. Plant Sci.* **5**, 82 (2014).
35. Lister, R. et al. A transcriptomic and proteomic characterization of the *Arabidopsis* mitochondrial protein import apparatus and its response to mitochondrial dysfunction. *Plant Physiol.* **134**, 777–789 (2004).
36. Samalova, M. et al. Hormone-regulated expansins: expression, localization, and cell wall biomechanics in *Arabidopsis* root growth. *Plant Physiol.* **194**, 209–228 (2023).
37. Voith von Voithenberg, L. et al. A novel prokaryote-type ECF/ABC transporter module in chloroplast metal homeostasis. *Front. Plant Sci.* **10**, 1264 (2019).
38. Li, W. et al. Plant pan-genomics: recent advances, new challenges, and roads ahead. *J. Genet. Genomics* **49**, 833–846 (2022).
39. Leonard, A. S. et al. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *Nat. Commun.* **13**, 3012 (2022).
40. Karlgren, A., Gyllenstrand, N., Kallman, T. & Lagercrantz, U. Conserved function of core clock proteins in the gymnosperm Norway spruce (*Picea abies* L. Karst). *PLoS ONE* **8**, e60110 (2013).

[1]Department of Environmental Horticulture, University of Seoul, Seoul, Republic of Korea. [2]Plant Immunity Research Center, Seoul National University, Seoul, Republic of Korea. [3]Department of Horticulture, Gyeongsang National University, Jinju, Republic of Korea. [4]Department of Forest Bioresources, National Institute of Forest Science, Suwon, Republic of Korea. [5]Department of Molecular Biology, College of Agricultural, Life Sciences and Natural Resources, University of Wyoming, Laramie, WY, USA. [6]Plant Systems Engineering Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Republic of Korea. [7]Department of Biosciences and Bioinformatics, Myongji University, Yongin, Republic of Korea. [8]Department of Plant and Environment New Resources, Kyung Hee University, Yongin, Republic of Korea. [9]Department of Agriculture, Forestry and Bioresources, Seoul National University, Seoul, Republic of Korea. [10]These authors contributed equally: Min-Jeong Jang, Hye Jeong Cho. ✉e-mail: pahkej@korea.kr; ksi2204@uos.ac.kr

Methods

Plant material and genome sequencing

A 15-year-old tree, grafted with a scion from *P. densiflora* (Jeongipumsong, Korean Natural Monument No. 103), was maintained at the Korea Forest Research Institute for genome sequencing[41]. High-molecular-weight genomic DNA (gDNA) was extracted from cambium tissue using the modified CTAB method[42]. After assessing the quality and quantity of the gDNA using a Femto Pulse system with the Agilent 2100 Bioanalyzer (Agilent Technologies), it was used for sequencing. For PacBio HiFi long-read sequencing, a total of 15 µg gDNA was prepared for library construction with a preliminary evaluation of gDNA length using the Femto Pulse system (Agilent Technologies). HiFi SMRTbell libraries were constructed using the SMRTbell prep kit 3.0 (Pacific Biosciences) following size selection performed with the BluePippin system (Sage Science). Subsequently, sequencing was conducted using SMRT cells (Pacific Biosciences) on the Revio sequencing platform. After sequencing, consensus HiFi reads were generated using CCS software with default parameters. The same DNA samples were also sequenced on the Illumina NovaSeq 6000 platform with a read length of 150 bp and an insert size of 350–550 bp. The 10x linked read library prepared using the Chromium Genome Reagent Kit and the Hi-C library prepared using the Dovetail Omni-C Kit were also sequenced on the Illumina NovaSeq 6000 platform with 2 × 150-bp paired-end reads following the manufacturer's protocol. Total RNA was isolated from four tissues (current-year shoots, needles, female flowers and immature cones) using TRIzol reagent[43]. After confirming the purity and the integrity of the RNA, the full-length complementary DNA (cDNA) library was generated using the Clontech SMARTer PCR cDNA Synthesis Kit, following the isoform sequencing (Iso-seq) protocol, and was subsequently used for Iso-seq on the PacBio Sequel II platform. Resequencing data for 30 wild accessions of *P. densiflora* were obtained from the Korea Forest Research Institute (Supplementary Table 8). In summary, fresh leaves were collected from 30 randomly distributed wild accessions across Korea for DNA extraction. Qualified gDNA was extracted using the DNeasy Plant Mini Kit (Qiagen) and used to construct paired-end sequencing libraries with the Illumina TruSeq DNA Nano Kit. The libraries were then sequenced on the Illumina NovaSeq 6000 platform with 151-bp paired-end reads, aiming for a target coverage of 14×.

Genome assembly and phasing

The initial contig assembly for the two sets of haplotypes (HA and HB) was performed by Hifiasm version 0.19.5-r587 with default parameters using the PacBio HiFi and Hi-C reads. Redundant haplotigs were eliminated during the implementation of Hifiasm with purge_dups. The contigs within each haplotype were reordered and clustered to generate chromosome-level scaffolding using the ALLHiC pipeline[44] (default parameters were employed for all steps except '-k 12' for the number of chromosomes). *P. tabuliformis*, a closely related species with a well-assembled genome, was used for the 'prune' function. Finally, we manually checked and curated order and orientation errors based on Hi-C contact maps using Juicebox[45] to achieve a chromosome-level haplotype-resolved assembly of *P. densiflora*.

Gene annotation

To annotate protein-coding genes in the two haplotypes, we applied several strategies, including whole-gene annotation using the MAKER version 2.31.10 (ref. 46) pipeline, structure-based orthologous gene annotation with *Arabidopsis* functional genes (FOGs) using GeMoMa version 1.6.1 (ref. 47) and finally in-depth reannotation of TF gene families using TGFam-Finder version 1.01 (ref. 48). First, we primarily used the MAKER pipeline to annotate whole-gene models, which integrates protein homology evidence, transcript mapping and ab initio predictions. The input data for the MAKER pipeline were prepared as follows. The protein-coding sequences of *P. tabuliformis*, *P. taeda*, *P. lambertiana*, *A. trichopoda*, *P. trichocarpa*, *A. thaliana* and *O. sativa* were used for alignment to support protein homology evidence. PacBio Iso-seq was used to assemble high-confidence full-length cDNA transcripts using pbmm2 version 1.10.0 (a Minimap2 wrapper for PacBio data) and Iso-seq collapse version 3.8.2. These transcripts served as transcript evidence. For ab initio prediction, SNAP[49] and AUGUSTUS version 3.2.3 (ref. 50) were used with an in-house training set comprising full-length genes from transcriptome and protein-based annotation data. After gene prediction using MAKER, we further refined the gene models by filtering them based on the annotation edit distance scores defined by MAKER and their deficiencies in transcript or homology evidence. To accurately incorporate *Arabidopsis* FOGs into the annotation of the two haplotypes, we performed homolog-based gene prediction using 10,356 *Arabidopsis* functional genes identified in TAIR (ftp://ftp.arabidopsis.org) following the default parameters of the GeMoMa version 1.6.1 (ref. 47) pipeline. Furthermore, TF genes were annotated using TGFam-Finder version 1.01 (ref. 48). The Pfam IDs of target DNA-binding domains, designated as 'TARGET_DOMAIN_ID', were assigned based on the PlantTFDB[25] classification (https://planttfdb.gao-lab.org). Considering the relatively longer length of introns in gymnosperms, '$EXTENSION_LENGTH' and '$MAX_INTRON_LENGTH' were set to 1 Mb. Finally, the annotation results from multiple approaches were integrated to generate a final set of annotated protein-coding genes for each haplotype.

For functional annotation, we used InterProScan version 5.22-61.0 (ref. 51) (-f tsv -appl Pfam) and hmm-search to identify conserved domains using the Pfam[52] and HMM[53] databases. GO analysis was conducted to analyze the putative functions of protein-coding genes using the OmicsBox platform with Blast2GO[54] annotation and mapping algorithms.

Assessment of genome assembly and annotation quality

We performed multiple assessments to validate the quality of genome assembly, phasing and annotation. Base pair quality was assessed using Merqury version 1.3 (ref. 55) software. Meryl databases were created using Illumina and 10x Genomics reads with a 21-mer size. Subsequently, Merqury was used to estimate QV scores and *k*-mer completeness. We also used the LTR Assembly Index[56] tool, wrapped in the LTR_retriever[57] pipeline with default parameters, to obtain standardized quality index values for each haplotype genome. Using Juicebox[45], we manually curated each chromosome based on the quality of Hi-C contact maps to consolidate contiguous Hi-C hits. For gene annotation quality assessment, we used BUSCO version 5.0 (ref. 58) with the embryophyta_odb10 database to evaluate the quality of gene annotations in *P. densiflora* (Table 1). In addition, we evaluated genome phasing by calculating switch errors between the two haplotype assemblies using calc_switchErr[59] with default parameters. We also assessed haplotype phasing using the copy number spectrum results from Merqury. The 'spectra-asm' plot was employed to track the multiplicity of k-mers found in each haplotype, and we manually confirmed the balance of peaks and fractions depicted in the haplotype-specific *k*-mer graphs.

Annotation of TEs and LTR-RTs

To identify repeat sequences in the two haplotypes of the *P. densiflora* genome, we generated the de novo repeat library using RepeatModeler2 version 2.0.1 (ref. 60) and annotated the library based on model databases of TEs in DeepTE[61] with default parameters. Using the de novo library, we annotated TEs in *P. densiflora* HA and HB using RepeatMasker version 4.1.1 (ref. 62). Moreover, we used the LTR_retriever[57] pipeline for in-depth structural annotation of LTR-RTs. Specifically, LTRharvest[63] (-maxlenltr 7000 -mindistltr 100 -similar 80) was used to identify genomic positions of LTR-RTs, and LTRdigest[64] was used to annotate internal features of LTR-RTs for further evolutionary analysis. For accurate analyses, we annotated TEs and LTR-RTs in three *Pinus* species and five other gymnosperm species using the same pipeline and parameters.

Comparison of *P. densiflora* and other *Pinus* genomes

Homologous synteny blocks between *P. densiflora* HA and *P. tabuliformis* were identified using MCScanX[65] (-s 3 -m 50). Syntenic blocks were visualized using the Circos[66] software package. For whole-genome-versus-genome alignment, the genomes of *P. tabuliformis*, *P. taeda* and *P. lambertiana* were used as query sequences, while the *P. densiflora* HA genome served as the reference sequence in Minimap2 (ref. 67) (-x asm20). Next, the alignment coverage (nonredundant aligned length/window size ratio) was estimated based on the genomic location of the HA genome in *P. densiflora*, with a window size of 30 Mb.

To accurately examine chromosomal rearrangements between chromosomes 1 and 3 of *P. densiflora* and *P. tabuliformis*, we performed comparative mapping of genetic markers from *P. densiflora*, *P. taeda* and *P. thunbergii* to the genome of *P. densiflora* and *P. tabuliformis*. We obtained 1,007 SNP marker sequences for *P. densiflora* from the National Institute of Forest Science[22]. We also collected 341 and 333 genetic marker sequences for *P. taeda*[21] and *P. thunbergii*[23] from previous studies, respectively. The marker sequences were searched against the entire genomes of *P. densiflora* and *P. tabuliformis* using BLASTN with an e-value cutoff of 1×10^{-10}. Only the top hit for each marker was subjected to comparative mapping. We then linked the same marker hits and compared their locations in each genome based on the physical positions of the linkage group on the genetic map. The divergence time between *Pinus* species was estimated using TimeTree5 (https://timetree.org/).

Comparative and evolutionary analyses of LTR-RTs

To explore the phylogenetic relationships among LTR-RTs in gymnosperms, we designed reverse transcriptase (RT) sets for *P. densiflora* HA and eight other gymnosperm genomes to represent the tremendous number of RT domains in these gymnosperm species as an input dataset. We examined motif combinations within individual RT domains using the MEME version 5.1.1 (ref. 68) (-V -time 180000000 -mod zoops -nmotifs 150 -minw 10 -maxw 50 -objfun se -markov_order 0) and MAST[69] algorithms.

Given the abundance of motif combinations, we randomly selected and subsequently reduced the RT domain set. The amino acid sequences of the RT domain in *gypsy* and *copia* were aligned using the fftns module in MAFFT version 7.470 (ref. 70), and poorly aligned regions were removed using TrimAl version 1.4.rev22 (ref. 71) (-gt 0.1). RAxML version 8.2.12 (ref. 72) was used to build a maximum likelihood tree with prior selection of the best suitable model, DUMMY2F and DUMMY2 for *gypsy* and *copia* elements, respectively (-m PROTGAMMAAUTO -p 12345). The maximum likelihood trees for *gypsy* and *copia* elements were supported by 500 bootstraps with random parsimony seeds (-m DUMMY2F/DUMMY2 -p 12345 -x 12345 -# 500). The excluded RT domain sequences in the phylogenetic tree were assigned to each subgroup using BLASTP against selected RT domain sequences in the tree (-evalue 1e-30 -outfmt 7 -num_threads 1 -max_target_seqs 100 -perc_identity 80). The results were visualized using heatmaps by subgroup.

To calculate the insertion time for LTR-RTs, we extracted the 5′ and 3′ LTR sequences for each LTR-RT and aligned them using PRANK version 170427 (ref. 73) (-showtree -f=paml). We then estimated the nucleotide distance (K) between 5′ and 3′ LTRs through maximum likelihood analysis using the baseml module from PAML version 4.9 (ref. 74) with default parameters. The insertion time (T) for each LTR-RT was calculated using the formula $T = K/2r$, where the nucleotide substitution rate (r) was assumed to be 2.2×10^{-9} based on a previous study[24]. The species tree and divergence times among the selected gymnosperms were predicted using TimeTree5 (https://timetree.org/).

Reannotation and phylogenetic and duplication analyses of TFs

The reannotation of TFs in 19 plant genomes (three *Pinus*, seven other gymnosperms and nine angiosperms listed in Supplementary Table 9) was performed using TGFam-Finder version 1.01 (ref. 48) with the same parameters used for the annotation of TFs in *P. densiflora*.

For comparative evolutionary analyses of TFs in *P. densiflora* HA and 19 species (*Pinus tabuliformis*, *Pinus taeda*, *Pinus lambertiana*, *Picea abies*, *Pseudotsuga menziesii*, *Abies alba*, *Ginko biloba*, *Gnetum montanum*, *Taxus chinensis*, *Torreya grandis*, *Amborella trichopoda*, *Zea mays*, *Oryza sativa*, *Allium sativum*, *Capsicum annuum*, *Arabidopsis thaliana*, *Helianthus annuus*, *Phyllostachys edulis* and *Populus trichocarpa*; Supplementary Table 9), we first performed phylogenetic analysis following the methods described in previous studies[75,76]. In detail, TF genes encoding an intact DNA-binding domain were selected from 11 *Pinus*-dominant TF families. The protein sequences of these intact TFs were aligned using the fftns module from MAFFT version 7.470 (ref. 70) and trimmed with TrimAl version 1.4.rev22 (ref. 71) (-gt 0.3). Phylogenetic trees were constructed using IQTREE version 2.0.6 (ref. 77) (-msub nuclear -alrt 1000 -B 1000 -safe), and subgroups were assigned based on the phylogenetic relationships and widely known subfamily classifications defined in previous studies. The remaining partial TFs excluded from the tree were assigned to each subgroup through BLASTP similarity searches (outfmt 7 -seg yes -evalue 1e-10 -max_target_seqs 50).

To identify gene duplication patterns across the 11 TF subgroups, we used DupGen_finder[78] (https://github.com/qiao-xin/DupGen_finder) to predict recently duplicated gene pairs and their duplication types. Initially, we conducted an all-by-all BLAST for whole genes in *P. densiflora* HA using default parameters to prepare the input dataset for DupGen_finder. Subsequently, the duplication types of gene pairs within the 11 TF subgroups were identified using in-house Perl scripts.

Identification of allelic variation and PAV and ASE genes between *P. densiflora* haplotypes

To identify genomic variation between haplotypes, we used the nucmer program in MUMmer version 4.0 (ref. 79) to align the HA and HB genomes using default parameters. SVs were reported using the 'show-diff' function in both HA and HB genomes. SNPs and indels were also identified using the 'show-snps' function with default parameters. Moreover, SNPs and indels within gene regions were annotated using SnpEff version 5.1 (ref. 80), assuming an upstream and downstream length of 2 kb (-ud 2000).

To classify PAV and ASE genes, we first defined allelic genes in *P. densiflora* HA and HB based on reciprocal BLAST best matches and the genomic location of genes. When the gene was annotated in only one haplotype, it was classified as a PAV if counterpart genomic regions did not contain its allelic gene after verification using Exonerate version 2.2.0 (ref. 81) (–model protein2genome–percent 50). For ASE genes, we conducted transcriptome analysis across various tissues of *P. densiflora* and categorized them by organs including (1) stem for immature stem-derived cambium, mature stem-derived cambium, immature stem-derived developing xylem, mature stem-derived developing xylem and immature whole stem, (2) leaf for young needle, (3) root for main root and (4) shoot apical meristem[82]. RNA-seq expression profiles were obtained following the New Tuxedo[83] protocol. Differentially expressed genes in each tissue were identified using DESeq2 (ref. 84) in the R module with $|\log_2(\text{fold change})| > 1$. Finally, ASE genes are defined as genes with alleles differentially expressed in at least one tissue.

Identification of *Arabidopsis* functional orthologous genes in *P. densiflora* haplotypes

To determine the orthologous relationship between *Arabidopsis* functional genes and annotated orthologous genes from the *Arabidopsis* functional genes using GeMoMa version 1.6.1 (ref. 47) in two haplotypes of the *P. densiflora* genome, we performed reciprocal BLAST searches with 10,356 *Arabidopsis* functional genes and putative *Arabidopsis* FOGs in both HA and HB. When a putative FOG, classified as a common allele or a PAV or ASE gene, was reciprocally best matched to a specific

Arabidopsis functional gene, we inferred an orthologous relationship and categorized it accordingly as a common allele, PAV and ASE FOG.

Experimental validation of PAV and ASE genes

Nucleic acids were extracted from *P. densiflora*. gDNA from leaf tissue was isolated using the DNeasy Plant Mini Kit (Qiagen), following the protocol provided by the manufacturer. Total RNA was extracted from leaf, stem and root tissues using the CTAB method. Subsequently, cDNA was synthesized using the PrimeScript II 1st Strand cDNA Synthesis Kit (Takara). For sequence confirmation of PAV genes and validation of ASE gene expression, the concentration (μg ml^{-1}) and quality (A260/A230 and A260/A280 ratios) of the nucleic acids were determined using the NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific) and visually verified on a 1.0% agarose gel. Primers were designed based on haplotype-specific variations in gDNA and cDNA sequences (Supplementary Table 10). PCR for sequence validation of PAV genes was performed with an initial denaturation at 95 °C for 5 min, followed by 40 cycles of denaturation at 95 °C for 30 s, annealing at 58 °C for 30 s and extension at 72 °C for 1 min, with a final elongation step at 72 °C for 5 min. PCR products were used after purification and subsequently subjected to sequencing. To ensure accurate detection of SNPs based on overlapping nucleotide peaks, chromatograms were validated using KB Basecaller version 1.4.1 (ref. 85) with quality values assigned to each base pair in the PCR product. Quantitative real-time PCR was used to identify tissue-abundant and haplotype-unbalanced expression of ASEs. Each reaction was analyzed in triplicate using an Exicycler 96 Real-Time Quantitative Thermal Cycler (Bioneer) and initiated with a pre-denaturation step at 95 °C for 10 min, followed by 40 cycles of denaturation at 95 °C for 5 s, annealing at 60 °C for 25 s and extension at 72 °C for 30 s, with a final extension at 65 °C for 5 min. Transcript levels of the genes were measured, and relative quantification was calculated using the $2^{-\Delta\Delta Ct}$ algorithm.

Functional validation of *P. densiflora* PAV genes and *Arabidopsis* orthologous genes

For ectopic overexpression of *Pd03G22920A* (MADS box) and *Pd03G15540* (AP2) in transgenic *Arabidopsis*, *Agrobacterium* containing the corresponding genes under the 35S promoter in the binary vectors was used for *Arabidopsis* transformation via the floral dip method[86]. *Col-0* was used as wild type. One-week-old T_1 seedlings resistant to 8 mg l^{-1} glufosinate ammonium (MilliporeSigma) were transferred to soil (3 BM6:1 vermiculite:1 perlite) from ½ Murashige and Skoog (MilliporeSigma) supplemented with 1% sucrose (MilliporeSigma), adjusted to pH 5.7 and hardened with 2.4% Phytagel (MilliporeSigma). Plants were grown in a controlled environmental growth room with a 16/8 h light/dark cycle under 130 μmol m^{-2} s^{-1} light at 22–24 °C and 45–55% humidity. In the given growth condition, wild-type plants flowered after 2–3 weeks with 6.2 rosette leaves on average. T_2 seeds were harvested from a *Pd03G22920A* (MADS box)-transgenic T_1 plant and grown to quantify flowering time to confirm the early flowering phenotype of *Pd03G22920A*-overexpressing transgenic *Arabidopsis* plants. An unpaired Student's *t*-test was performed to examine statistical significance with Prism 10 (GraphPad).

For subcellular localization of *Arabidopsis* FOGs in *N. benthamiana* plants, *Arabidopsis* FOGs were cloned in a binary vector to be expressed as fusion proteins with various fluorescent proteins ectopically in *N. benthamiana* leaves under the 35S constitutively overexpressing promoter. Fusion proteins were expressed in plant cells by *Agrobacterium*-mediated transient gene expression[87]. Corresponding subcellular compartment markers were coexpressed as references. MT-moxVenus and PM-mCherry constructs indicate markers for mitochondria and plasma membrane, respectively[88]. A transgenic *N. benthamiana* plant expressing NRIP1–mCerulean was used as a reference for the chloroplastic localization of ABCI12 (ref. 89). Confocal microscope images were acquired with Leica TCS SP8 STED and Zeiss LSM 980 laser scanning confocal microscopes. All images were acquired with a ×40 1.2-NA C-Apochromat water-immersion objective with 405-nm (mCerulean, 1.5%), 514-nm (moxVenus, 15–20%) and 561-nm (mCherry, 15–20% and mRuby3, 50%) laser lines. Images were processed using Fiji ImageJ[90]. mCherry and mRuby3 are pseudocolored in magenta, while moxVenus and cerulean are pseudocolored in green and cyan, respectively.

Comparison of *P. densiflora* PAV genes with known *Pinus* functional genes was conducted by considering similarity in both sequence and protein structure. Protein structure prediction was carried out using AlphaFold2 (ref. 91), and visualization was accomplished using ChimeraX[92]. The similarity between the *Pinus* functional gene and its orthologous PAV genes in *P. densiflora* was quantified based on protein structure using the TM score[93]. We estimated the expression of orthologous PAV genes in *P. densiflora* under abiotic stress conditions. Briefly, the propagated clones were subjected to salinity and cold stress treatments by exposing them to 250 mM NaCl and 4 °C for 0 and 24 h, respectively. The control group comprised clones that were not exposed to any abiotic stress. Each treatment time point consisted of at least two biological replicates, including the control group. Other environmental conditions were maintained at a constant level. To analyze gene expression, fresh leaves, stems and roots were collected from each group. Quantitative real-time PCR was performed using the same protocol as that used for ASE validation, and the primers used for assessing abiotic stress expression are listed in Supplementary Table 10.

Linear and graph-based analyses of resequencing data for *P. densiflora* accessions

Given recent studies that applied disparate methods for variant detection depending on the type of variant to enhance accuracy[94–96], we used a linear reference-based approach to detect small variants such as SNPs or indels and a graph reference-based analysis to explore large SV variants. To analyze SNPs and indels (1–49 bp), we performed variant calling via linear reference genome alignment for HA and HB using resequencing data of 30 accessions. Raw resequencing reads from 30 *P. densiflora* accessions were trimmed using CLC Assembly Cell (CLC bio) (-c 20 -f 33 -m 70). The clean reads were iteratively mapped to each haplotype using BWA-MEM[97] with default parameters. Mapped reads were sorted using SAMtools[98], and duplicated reads were marked using MarkDuplicates from Picard (available at https://broadinstitute.github.io/picard/; Picard Toolkit 2019). Subsequently, variants were called using Genome Analysis Toolkit (GATK) version 4.1.6.0 (ref. 99) with the following functions: call germline with HaplotypeCaller (-ERC GVCF), import and merge GVCFs of 30 accessions with GenomicDBImport, assign genotype with GenotypeGVCF and filter only SNPs and indels with VariantFiltration ('QD < 2.0||MQ < 40.0||FS > 60.0||SOR > 3.0 || MQRankSum<−12.5 || ReadPosRankSum<−8.0||GQ < 20.0' for SNPs; all criteria the same with SNPs but 'FS > 200.0||SOR > 10.0' for indels).

Because it was known that graph reference-based alignment mitigates reference bias, especially in SV detection[100–102], we benchmarked this approach by graphically reflecting our haplotype variation in the reference genome. We constructed a haplotype variation-aware graph reference for HA and HB using PanGenome Graph Builder[103] (default parameters except '-n 2' for number of haplotypes). We converted and indexed the graph reference (GFA format) appropriately to use in vg toolkit[101] as follows: we converted GFA to VG format using 'vg convert' and indexed the graph using 'vg autoindex' to obtain the topological (xg) index. All reads from 30 accessions were then aligned to the haplotype-aware graph reference using 'vg giraffe' with default parameters. The compressed read coverage index was calculated using 'vg pack' with mapping quality filtration (-Q 5), snarls were computed using 'vg snarls', and variants were genotyped using 'vg call' with default parameters. We retained only large SVs (≥50 bp) in VCF. Finally, variant information of SNPs, indels and SVs from each accession was

incorporated into a single VCF format considering nonredundant positions. A SNP-based maximum likelihood phylogenetic tree was constructed using RAxML version 8.2.12 (ref. 72) with the best amino acid model PMB and empirical base frequency.

To examine the diversity of allele and PAV genes in the reference genome across accessions, we predicted the types of variations in the reference genes within each accession. Variants were annotated with SnpEff version 5.1 (ref. 80) using default parameters and categorized based on their size and impact on the gene. In addition, the annotated variants were filtered to retain only those predicted to be high-impact variants, which included protein-disrupting variants such as frameshift variants, stop acquisition and loss of exons including a stop or start codon, etc. We assumed that, if a particular gene contained one or more high-impact variants in its coding region, it could be omitted. Finally, we classified reference genes in each accession as allele, PAV or absent genes. The completeness of various metabolic pathways was determined using KEGG Decoder[104].

Statistics and reproducibility

Statistical details, including the number of replicates, statistical methods and *P* values, are reported in the legends of Figs. 1–5 and Extended Data Figs. 1–10 or are described in the main text. For each experiment, at least two biological replicates were performed. No statistical method was used to predetermine sample size, and no data were excluded from the analyses. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

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

Data availability

The genome assembly and annotation data for the haplotypes of *P. densiflora* and the genotype information generated from resequencing analysis of *P. densiflora* accessions have been deposited in figshare+ (https://doi.org/10.25452/figshare.plus.25546534)[105]. Resequencing data have been deposited at the NCBI SRA under BioProject accession number PRJNA1089250.

Code availability

The related code has been deposited in GitHub (https://github.com/minjeongjj/pinus_densiflora_haplotype_genome) and Zenodo (https://doi.org/10.5281/zenodo.12791823)[106]. All software used in this study is publicly available as described in Methods and Reporting Summary.

References

41. Lee, S., Hong, Y., Kwon, H. & Kim, Z. Population genetic studies on indigenous conifers in Korea. *For. Sci. Technol.* **2**, 137–148 (2006).
42. Inglis, P. W., Pappas, M. C. R., Resende, L. V. & Grattapaglia, D. Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *PLoS ONE* **13**, e0206085 (2018).
43. Meng, L. & Feldman, L. A rapid TRIzol-based two-step method for DNA-free RNA extraction from *Arabidopsis* siliques and dry seeds. *Biotechnol. J.* **5**, 183–186 (2010).
44. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
45. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
46. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
47. Keilwagen, J. et al. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89 (2016).
48. Kim, S. et al. TGFam-Finder: a novel solution for target-gene family annotation in plants. *New Phytol.* **227**, 1568–1581 (2020).
49. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
50. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
51. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
52. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
53. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
54. Gotz, S. et al. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).
55. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245–271 (2020).
56. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
57. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
58. Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
59. Zhang, X. T. et al. Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nat. Genet.* **53**, 1250–1274 (2021).
60. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
61. Yan, H., Bombarely, A. & Li, S. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. *Bioinformatics* **36**, 4269–4275 (2020).
62. Tempel, S. Using and understanding RepeatMasker. *Methods Mol. Biol.* **859**, 29–51 (2012).
63. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18–31 (2008).
64. Steinbiss, S., Willhoeft, U., Gremme, G. & Kurtz, S. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.* **37**, 7002–7013 (2009).
65. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e62 (2012).
66. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
67. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
68. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
69. Bailey, T. L. & Gribskov, M. Combining evidence using *p*-values: application to sequence homology searches. *Bioinformatics* **14**, 48–54 (1998).

70. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
71. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
72. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
73. Loytynoja, A. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* **1079**, 155–170 (2014).
74. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
75. Jang, M. J., Hong, W. J., Park, Y. S., Jung, K. H. & Kim, S. Genomic basis of multiphase evolution driving divergent selection of zinc-finger homeodomain genes. *Nucleic Acids Res.* **51**, 7424–7437 (2023).
76. Chae, G. Y., Hong, W. J., Jang, M. J., Jung, K. H. & Kim, S. Recurrent mutations promote widespread structural and functional divergence of MULE-derived genes in plants. *Nucleic Acids Res.* **49**, 11765–11777 (2021).
77. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
78. Qiao, X. et al. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **20**, 38–60 (2019).
79. Marcais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
80. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain. *Fly* **6**, 80–92 (2012).
81. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31–41 (2005).
82. Kim, M. H. et al. Wood transcriptome analysis of *Pinus densiflora* identifies genes critical for secondary cell wall formation and NAC transcription factors involved in tracheid formation. *Tree Physiol.* **41**, 1289–1305 (2021).
83. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
84. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550–570 (2014).
85. Hyman, R. W., Jiang, H., Fukushima, M. & Davis, R. W. A direct comparison of the KB™ Basecaller and phred for identifying the bases from DNA sequencing using chain termination chemistry. *BMC Res. Notes* **3**, 257 (2010).
86. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998).
87. Norkunas, K., Harding, R., Dale, J. & Dugdale, B. Improving agroinfiltration-based transient gene expression in *Nicotiana benthamiana*. *Plant Methods* **14**, 71 (2018).
88. Park, E., Lee, H. Y., Woo, J., Choi, D. & Dinesh-Kumar, S. P. Spatiotemporal monitoring of effectors via type III secretion using split fluorescent protein fragments. *Plant Cell* **29**, 1571–1584 (2017).
89. Caplan, J. L. et al. Chloroplast stromules function during innate immunity. *Dev. Cell* **34**, 45–57 (2015).
90. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
91. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
92. Meng, E. C. et al. UCSF ChimeraX: tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).
93. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
94. Talenti, A. et al. A cattle graph genome incorporating global breed diversity. *Nat. Commun.* **13**, 910 (2022).
95. Li, N. et al. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat. Genet.* **55**, 852–860 (2023).
96. He, Q. et al. A graph-based genome and pan-genome variation of the model plant *Setaria*. *Nat. Genet.* **55**, 1232–1242 (2023).
97. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
98. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
99. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
100. Gao, Y. et al. A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).
101. Hickey, G. et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
102. Siren, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).
103. Garrison, E. et al. Building pangenome graphs. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.05.535718 (2023).
104. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J.* **12**, 1861–1866 (2018).
105. Kim, S., Jang, M.-J. & Cho, H. J. Chromosome-level haplotype-resolved genome assembly of *Pinus densiflora*. *Figshare+* https://doi.org/10.25452/figshare.plus.25546534 (2024).
106. Jang, M.-J., Cho, H. J. & Kim, S. Code for chromosome-level haplotype-resolved genome assembly of *Pinus densiflora* (v1.0). *Zenodo* https://doi.org/10.5281/zenodo.12791823 (2024).

Acknowledgements

This study was supported by the Basic Science Research Program through a National Research Foundation of Korea grant funded by the Korean government (NRF-2022R1C1C1004918) to S.K., by the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, and Forestry through the Digital Breeding Transformation Technology Development Program funded by the Ministry of Agriculture, Food, and Rural Affairs (322075-3) to S.K., by a grant from the Korea Forest Service of the Korean government through the R&D Program for Forestry Technology (2014071H10-2022-AA04) to S.K. and E.-J.P. and by a National Institute of Forest Science grant (Forest Science Research project number FG0603-2021-01-2022) to E.-J.P. This research was supported in part by the Plant Biotic Interaction program of the National Science Foundation of the USA (NSF-IOS-2126256) to E.P. and J.W. and the intramural research program of the US Department of Agriculture, National Institute of Food and Agriculture Hatch Capacity (7000762) to E.P. We appreciate S. J. Lee of DNA Link, who provided support for genome sequencing analyses during this project.

Author contributions

S.K. designed and organized the study as a lead contact. S.K. and E.-J.P. initiated the project. M.-J.J., H.J.C., Y.-S.P., S.-J.K., J.-W.C., G.Y.C., J.-Y.G. and S.K. performed data generation and/or bioinformatics analysis. M.-J.J., G.Y.C. and Y.-M.K. performed de novo genome assembly and annotation. H.J.C., Y.-S.P. and J.-W.C. performed TF

gene analysis. M.-J.J. and H.J.C. performed haplotype variation and ortholog-based analysis. J.-W.C., J.-Y.G. and J.-H.K. performed the transcriptome analysis. M.-J.J., Y.-S.P., H.J.C. and M.-S.K. performed the resequencing analysis. D.Y.K., H.J.C., M.-J.J., S.-J.K. and S.-H.K. performed PAV and ASE validation. H.J.C. and E.-K.B. performed protein structure and abiotic stress analyses. E.-K.B., M.-J.K., H.L., K.-S.C., I.S.K., K.-S.K. and E.-J.P. prepared plant material, DNA and RNA samples and raw resequencing data. H.-Y.L., S.J., H.J., J.W., E.P. and D.C. performed ectopic expression and subcellular localization analyses. H.-Y.L. and E.P. designed, organized and drafted the initial manuscript of the ectopic expression and subcellular localization analyses. M.-J.J. and H.J.C. wrote the initial draft manuscript. S.K., M.-J.J. and H.J.C. reviewed and edited the manuscript. All authors approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

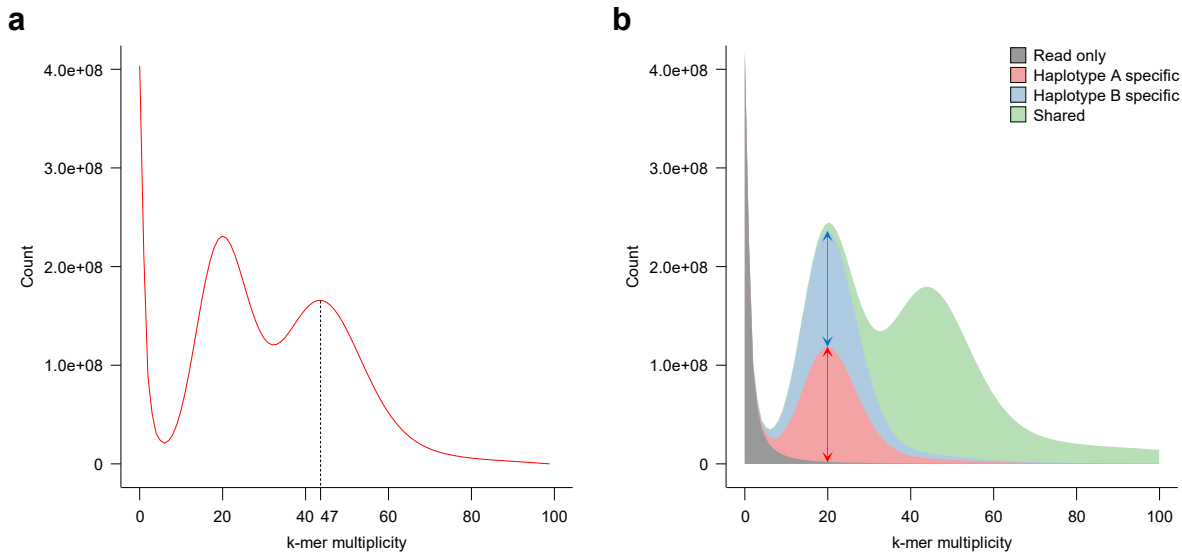
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-01944-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01944-y.

Correspondence and requests for materials should be addressed to Eung-Jun Park or Seungill Kim.

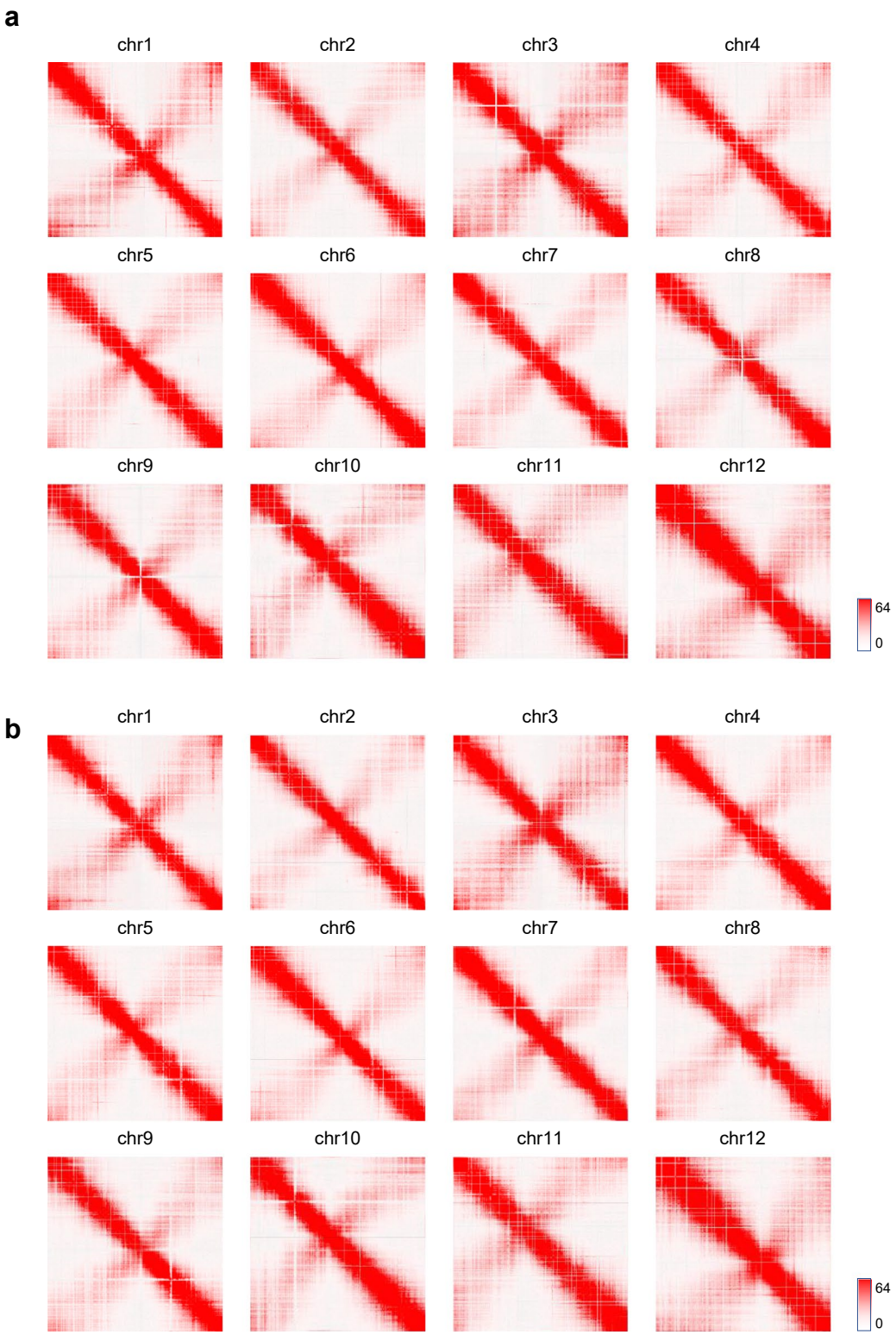
Peer review information *Nature Genetics* thanks De-Zhu Li, Yuanyuan Liu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-024-01944-y



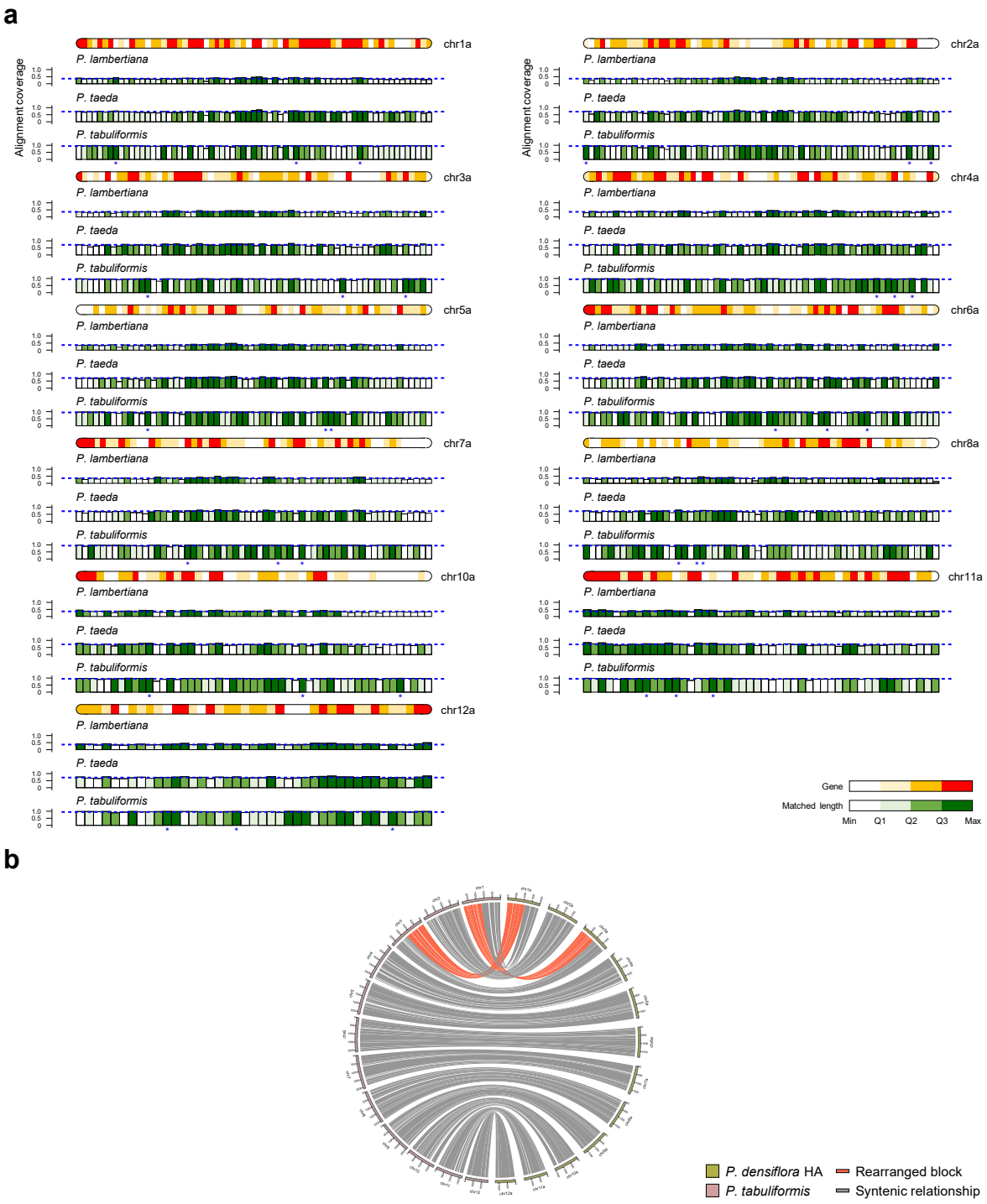
Extended Data Fig. 1 | The k-mer analysis for genome size estimation and polymorphism of *P. densiflora*. **a**, 21 k-mer depth distribution for genome size estimation. The x-axis indicates the k-mer depth and the y-axis indicates the frequency of k-mers. The dotted line represents the peak value. **b**, Haplotype-specific k-mer assembly spectrum (spectra-asm) plot from Merqury results. The graph shows k-mer proportion of haplotype-specific (red and blue), shared (green), and read-only (grey). The red and blue lines represent the evenly bisected the haplotype specific portion of k-mers, respectively.

https://doi.org/10.1038/s41588-024-01944-y



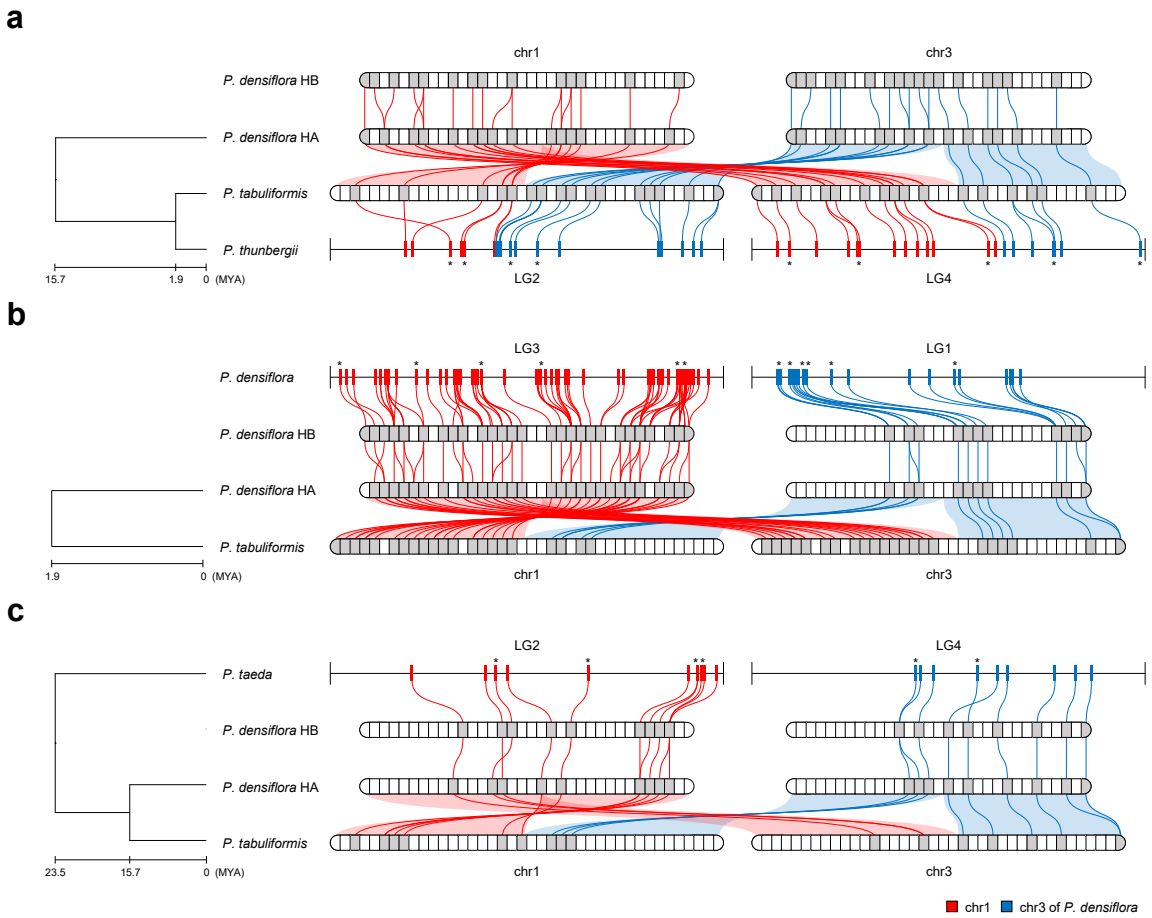
Extended Data Fig. 2 | Hi-C contact map for each of the 12 chromosomes in the *P. densiflora* genome. a, Haplotype A (HA). **b**, Haplotype B (HB). The color scale from white to red indicates low to high contact probability.

https://doi.org/10.1038/s41588-024-01944-y



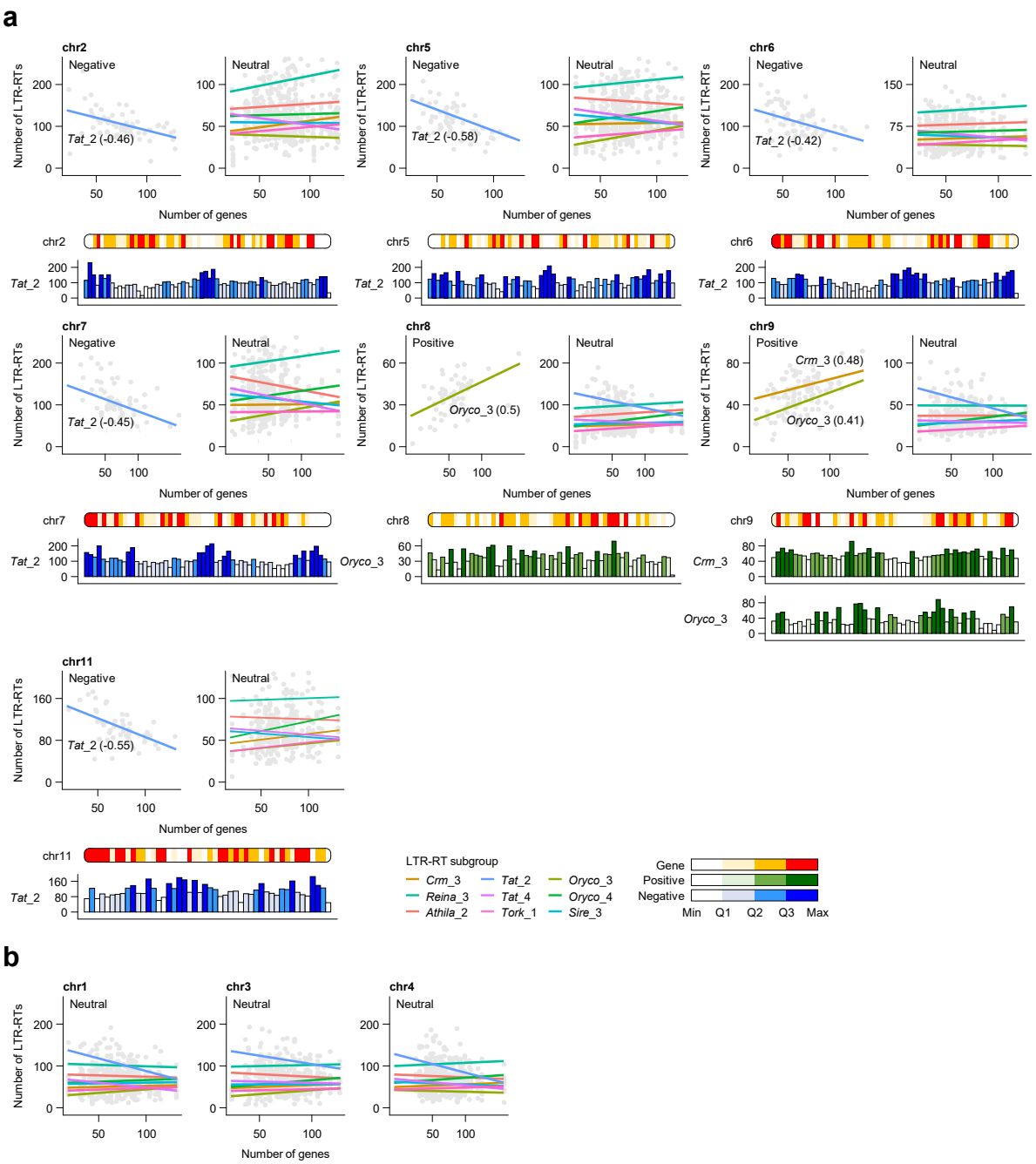
Extended Data Fig. 3 | Comparison of genomic analysis between *P. densiflora* and other *Pinus* species. a, Genome-to-genome alignment of *P. densiflora* HA and other *Pinus* species. The color gradient from white to red represents gene-poor and gene-rich chromosome regions. The color gradient from white to green represents the increase in matched alignment length in each *Pinus* genome by the *P. densiflora* genome, while the height of the bar represents the alignment length in the *P. densiflora* genome by other *Pinus* genomes. The blue dotted line indicates the total proportion of matched genome sequences in the *P. densiflora* genome from other *Pinus* species. Asterisks (*) denote the top three highly duplicated regions in each chromosome of *P. tabuliformis*. **b**, Synteny comparison between *P. densiflora* HA and *P. tabuliformis*.

https://doi.org/10.1038/s41588-024-01944-y



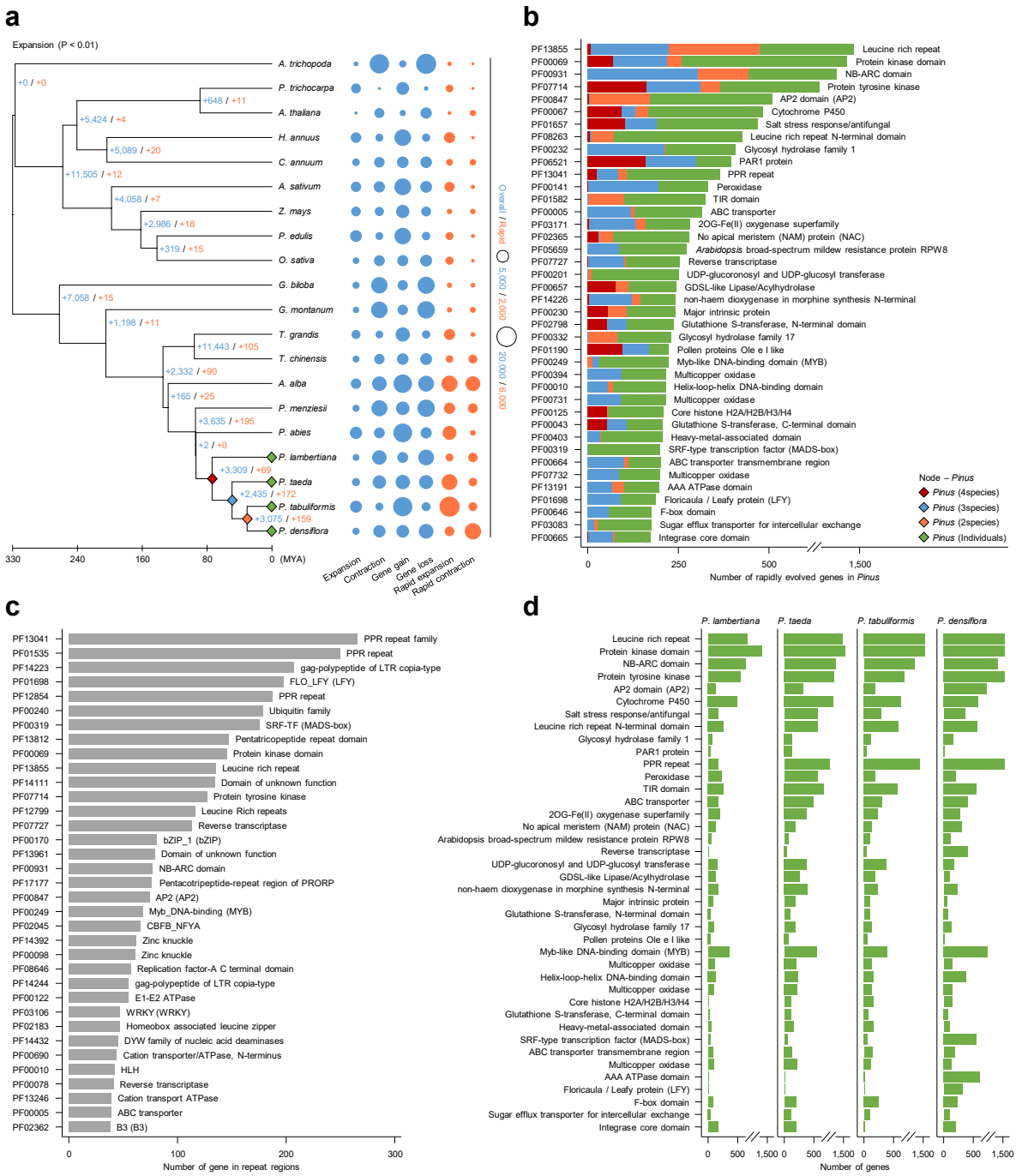
Extended Data Fig. 4 | Genome rearrangements between chromosomes 1 and 3 in *Pinus* species. Physical and genetic maps of *Pinus* species are depicted. The blocks (with a window size of 60 Mb) in the *P. densiflora* and *P. tabuliformis* genomes, mapped by representative markers of **a**, *P. thunbergii*, **b**, *P. densiflora*, and **c**, *P. taeda* are illustrated in grey color. LG, linkage group. In accordance with *P. densiflora*, the red and blue color for line, background, and marker represent chromosomes 1 and 3, respectively. Asterisks (*) denote the markers shown in Fig. 1c.

https://doi.org/10.1038/s41588-024-01944-y



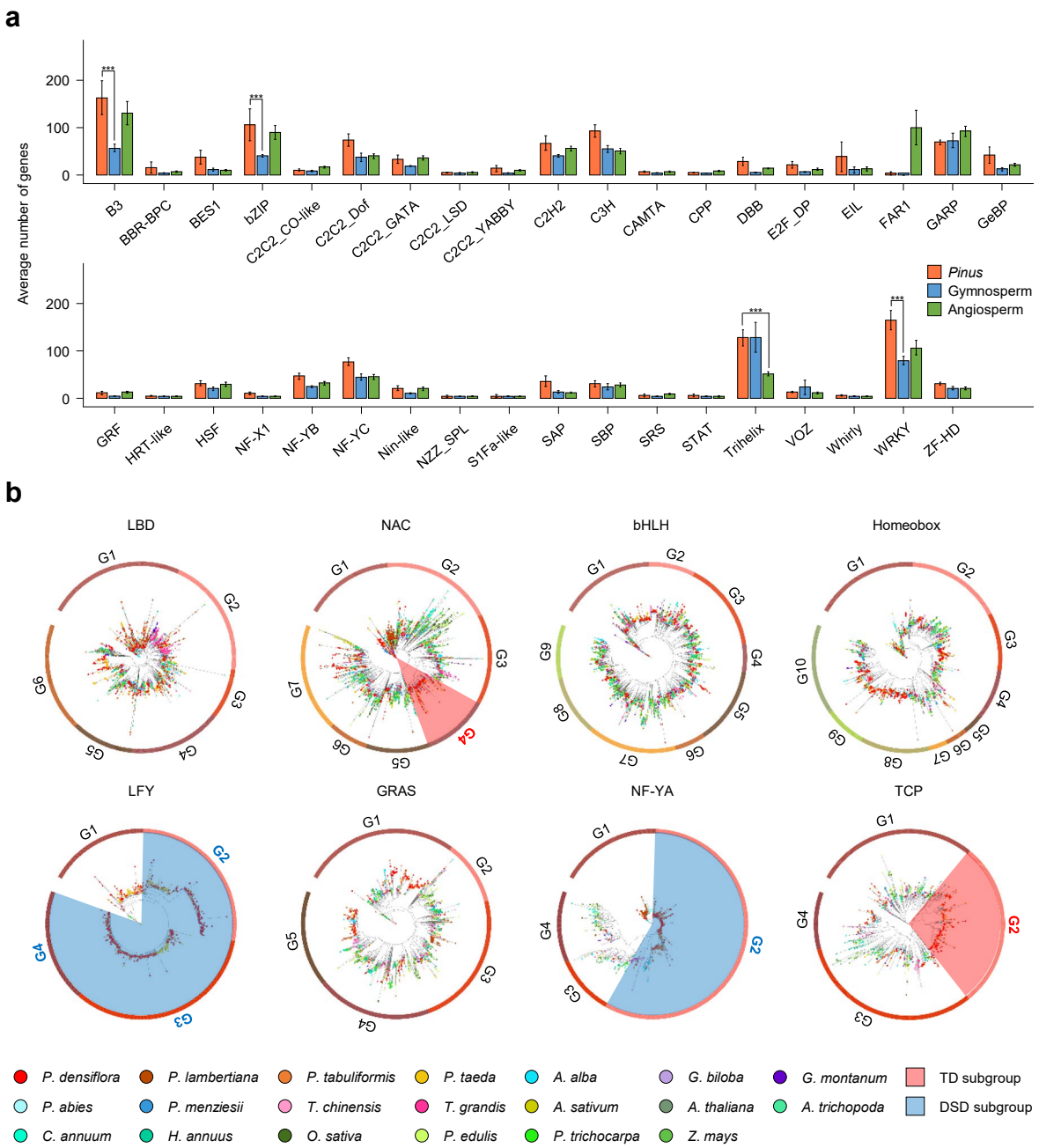
Extended Data Fig. 5 | The correlation between the number of genes and LTR-RTs in the *gypsy* and *copia* subgroups on chromosomes of *P. densiflora* HA. For each chromosome: **a**, upper correlation plots show LTR-RT subgroups that positively, negatively, or neutrally correlated with gene density and **b**, only neutrally correlated with gene density. The line colors indicate each LTR-RT subgroup. The number of genes (top) and LTR-RTs (bottom) were plotted as density within 30 Mb intervals. The color gradient from white to red, blue, and green represents increased number of genes, negatively correlated LTR-RTs with genes, and positively correlated LTR-RTs with genes, respectively.

https://doi.org/10.1038/s41588-024-01944-y



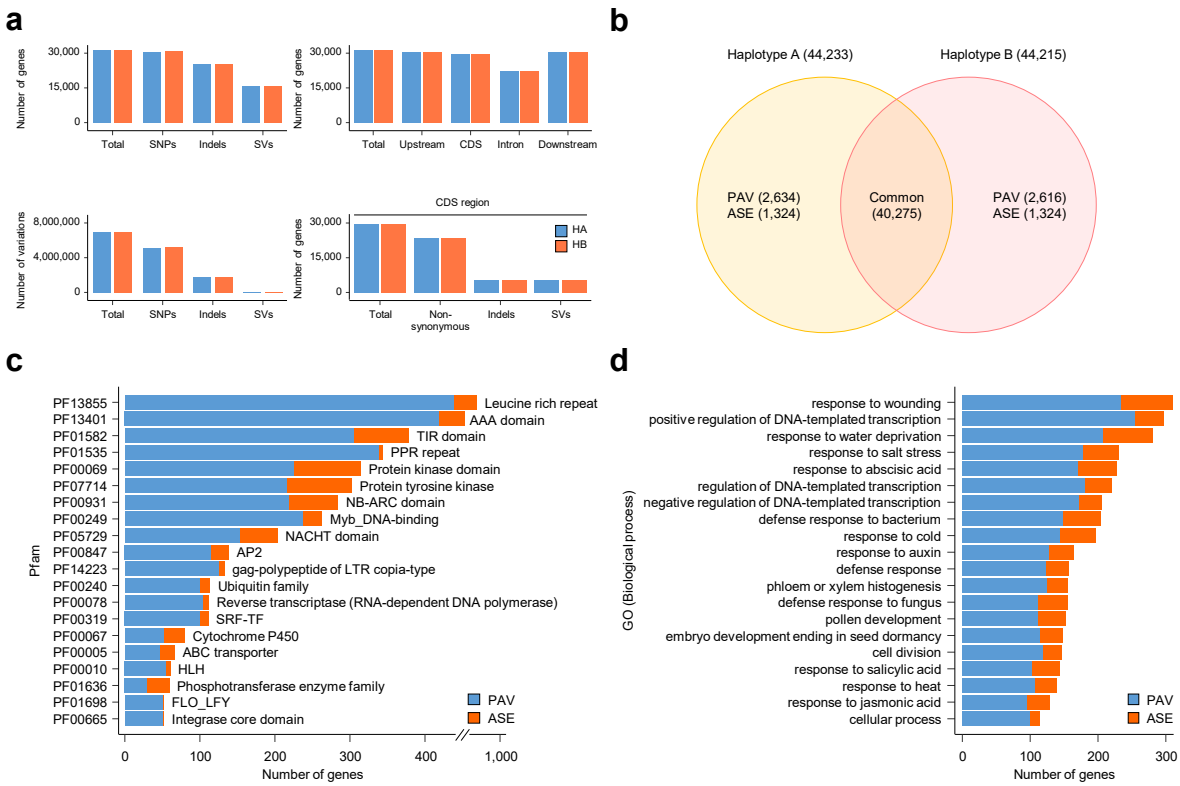
Extended Data Fig. 6 | Evolutionary analysis of *Pinus* and other plant species. **a**, Evolution of gene families in 20 plant species. The numbers in blue and orange, separated by a slash, indicate the expanded and contracted gene families and rapidly evolved gene families, respectively (a two-sided P < 0.01) (left). On the right, the size and color of the circles indicate the number of rapid or not gene family expansion/contraction and gene gain/loss in each species. **b**, Domain repertoire of rapidly evolved genes in 4 *Pinus* species. The red, blue, orange, and green diamonds indicate 4, 3, 2, and individual *Pinus* species, respectively. **c**, Domain repertoire of genes in repeat regions of *P. densiflora*. **d**, Domain repertoire of rapidly evolved genes in individual *Pinus* species.

https://doi.org/10.1038/s41588-024-01944-y



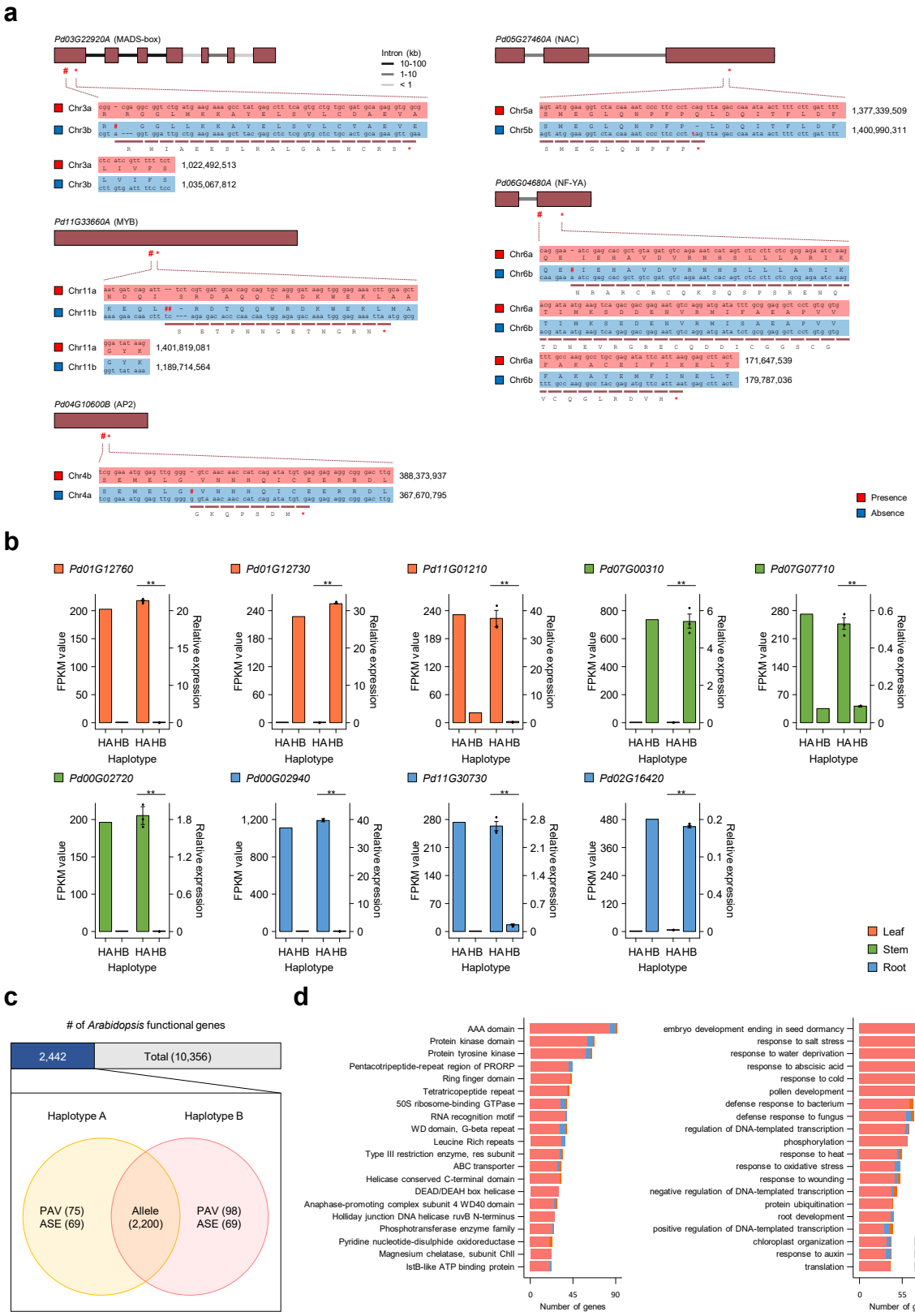
Extended Data Fig. 7 | Enrichment analysis and phylogenetic relationships of TFs in *Pinus*, gymnosperm, and angiosperm. a, The average number of TFs in *Pinus* (n = 4), gymnosperm (n = 7), and angiosperm (n = 9). Asterisks (***) denote a significance level of P < 0.001 based on a one-sided Fisher's test. The individual data points are listed in Supplementary Table 3. Error bars indicate the standard error (SE). **b**, The phylogenetic relationships of TFs in 20 plant species. Dot colors on the nodes represents species, while the ring indicates subgroups.

https://doi.org/10.1038/s41588-024-01944-y



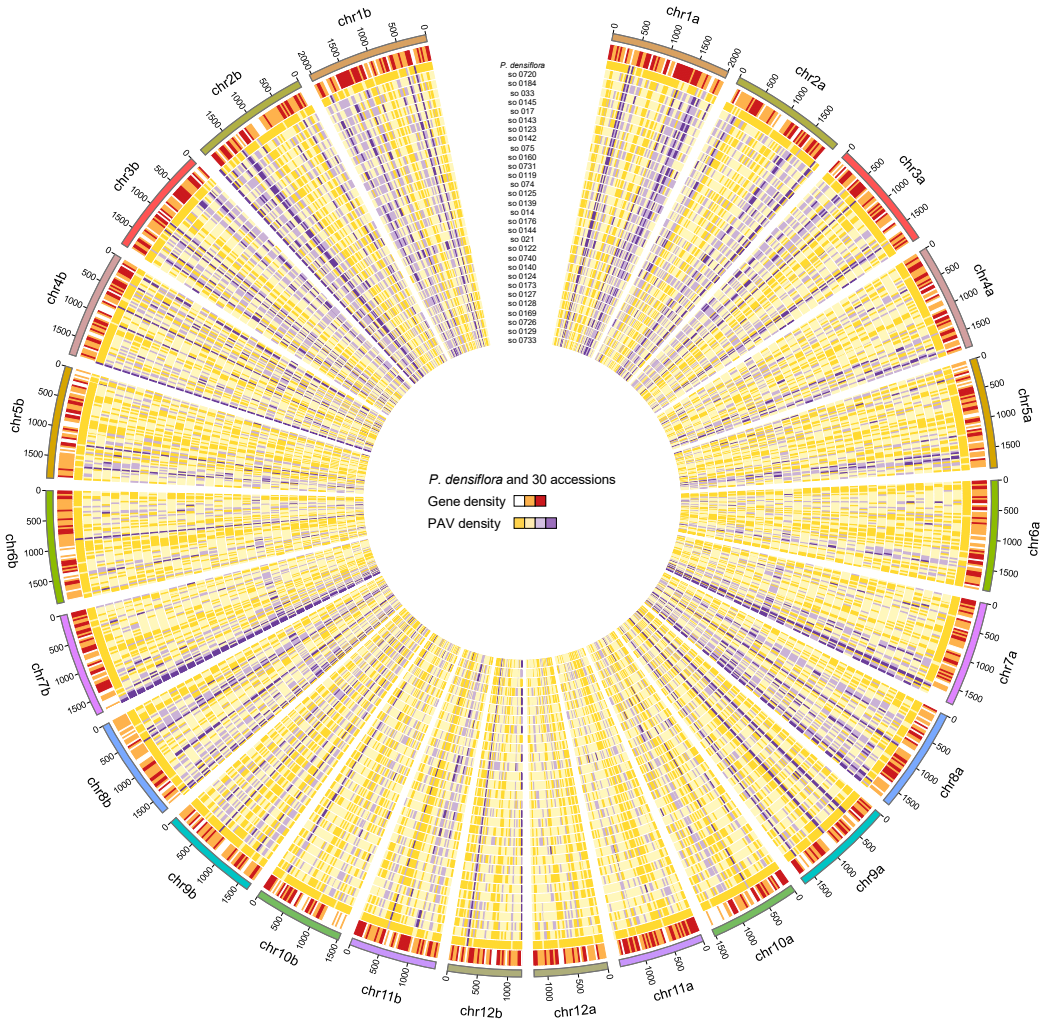
Extended Data Fig. 8 | Allelic imbalance of genic regions. a, Genomic variations in genes including 2 kb upstream and downstream between *P. densiflora* HA and HB. The bar graphs show the total number of variations and overall number of genes containing SNPs, indels, and SVs in each genic region. **b**, Allelic gene categorization of *P. densiflora*. **c**, Domain repertoire of PAV and ASE genes. **d**, GO descriptions in biological process of PAV and ASE genes.

https://doi.org/10.1038/s41588-024-01944-y



Extended Data Fig. 9 | Validation of PAVs and ASEs in *P. densiflora* and characterization of *Arabidopsis* functional orthologous genes (FOGs). **a**, Sequence validation of haplotype-specific presence of PAVs. **b**, Tissue abundant and haplotype unbalanced expression of ASEs in leaf, stem, and root. Asterisks (**) denote a significance level of P < 0.01 based on a one-sided unpaired Student's *t*-test. At least two biological replicates are used. Error bars indicate the SE. The brown boxes and black lines indicate exons and introns, respectively. The red and blue backgrounds of amino acids indicate presence and absence, respectively. The hashtags (#) and asterisks (*) indicate frameshift mutation and stop codon, respectively. **c**, *Arabidopsis* FOGs annotated in *P. densiflora* HA and HB. **d**, Domain repertoire (left) and GO descriptions in biological process (right) of allele, PAV, and ASE genes. The pink, blue, and orange bars indicate allele, PAV, and ASE genes, respectively.

https://doi.org/10.1038/s41588-024-01944-y



Extended Data Fig. 10 | Genome-wide distribution for allele and PAV genes of *P. densiflora* and 30 wild accessions. The outer track represents 12 chromosomes. The inner tracks represent gene density and PAV density for each accession.

nature portfolio

Corresponding author(s): Seungill Kim

Last updated by author(s): Jul 30, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data.
Data analysis	The related code has been deposited in GitHub (https://github.com/minjeongjj/pinus_densiflora_haplotype_genome) and Zenodo (https://doi.org/10.5281/zenodo.12791823). All software used in this study is publicly available as described in the Methods and Reporting Summary. The software used in this manuscript includes: Hifiasm v0.19.5-r587, ALLHiC pipeline, MAKER v2.31.10, GeMoMa v1.6.1, TGFam-Finder v1.01, pbmm2 v1.10.0, Iso-Seq collapse v3.8.2, SNAP, Augustus v3.2.3, InterproScan v5.22–61.0, Merqury v1.3, LTR_retriever pipeline, BUSCO v5.0, calc_switchErr pipeline, RepeatModeler2 v2.0.1, RepeatMasker v4.1.1, LTR_retriever, LTRharvest, MCScanX, Circos, Minimap2, TimeTree5 (https://timetree.org/), MEME v5.1.1, MAST, MAFFT v7.470, TrimAl v1.4.rev22, RAxML v8.2.12, BLASTP, PRANK v.170427, IQTREE v2.0.6, DupGen_finder (https://github.com/qiao-xin/DupGen_finder), Exonerate v2.2.0, CLC Assembly Cell, BWA-MEM, SAMtools, Picard MarkDuplicates (https://broadinstitute.github.io/picard/), Genome Analysis Toolkit (GATK) v4.1.6.0, PanGenome Graph Builder (PGGB), vg-toolkit, RAxML v8.2.12, SnpEff v5.1, KEGG Decoder.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The genome assembly and annotation data for the haplotypes of P. densiflora, and the genotype information generated from resequencing analysis of P. densiflora accessions have been deposited in the Figshare plus (https://doi.org/10.25452/figshare.plus.25546534). The resequencing data have been deposited at NCBI SRA under BioProject accession number PRJNA1089250.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This is not relevant to our study.
Reporting on race, ethnicity, or other socially relevant groupings	This is not relevant to our study.
Population characteristics	This is not relevant to our study.
Recruitment	This is not relevant to our study.
Ethics oversight	This is not relevant to our study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	30 wild accessions of P. densiflora used in the resequencing analysis were collected from the mountains in Korea. The number of wild accessions was based on the collectibility.
Data exclusions	No data was excluded from the analysis.
Replication	At least two biological replicates were used for entire analyses, including ectopic overexpression, PCR, and qRT-PCR. All attempts at replication were successful.
Randomization	Randomization was not applicable, as this study was not a clinical trial.
Blinding	Blinding was not applicable, as this study was not a clinical trial.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☐	☒ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☒	☐ Public health
☒	☐ National security
☒	☐ Crops and/or livestock
☒	☐ Ecosystems
☒	☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
☒	☐ Demonstrate how to render a vaccine ineffective
☒	☐ Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐ Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐ Increase transmissibility of a pathogen
☒	☐ Alter the host range of a pathogen
☒	☐ Enable evasion of diagnostic/detection modalities
☒	☐ Enable the weaponization of a biological agent or toxin
☒	☐ Any other potentially harmful combination of experiments and agents