JIPB Journal of Integrative Plant Biology

Research Article
https://doi.org/10.1111/jipb.13265

The *Larix kaempferi* genome reveals new insights into wood properties[∞]

Chao Sun[1,2†], Yun-Hui Xie[1,2†], Zhen Li[3,4†], Yan-Jing Liu[1], Xiao-Mei Sun[1,2], Jing-Jing Li[5], Wei-Peng Quan[5], Qing-Yin Zeng[1*], Yves Van de Peer[3,4,6*] and Shou-Gong Zhang[1,2*]

1. State Key Laboratory of Tree Genetics and Breeding, Chinese Academy of Forestry, Beijing 100091, China
2. Key Laboratory of Tree Breeding and Cultivation of the State Forestry and Grassland Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China
3. Department of Plant Biotechnology and Bioinformatics, Ghent University, Gent B-9052, Belgium
4. VIB Center for Plant Systems Biology, Ghent B-9052, Belgium
5. Nextomics Biosciences Co., Ltd, Wuhan 430073, China
6. Department of Biochemistry, Genetics and Microbiology, Pretoria, South Africa

†These authors contributed equally to this work.
*Correspondences: Qing-Yin Zeng (qingyin.zeng@caf.ac.cn); Yves Van de Peer (yves.vandepeer@psb.vib-ugent.be); Shou-Gong Zhang (sgzhang@caf.ac.cn, Dr. Zhang is fully responsible for the distributions of all materials associated with this article)

Chao Sun

Shou-Gong Zhang

ABSTRACT

Here, through single-molecule real-time sequencing, we present a high-quality genome sequence of the Japanese larch (*Larix kaempferi*), a conifer species with great value for wood production and ecological afforestation. The assembled genome is 10.97 Gb in size, harboring 45,828 protein-coding genes. Of the genome, 66.8% consists of repeat sequences, of which long terminal repeat retrotransposons are dominant and make up 69.86%. We find that tandem duplications have been responsible for the expansion of genes involved in transcriptional regulation and stress responses, unveiling their crucial roles in adaptive evolution. Population transcriptome analysis reveals that lignin content in *L. kaempferi* is mainly determined by the process of monolignol polymerization. The expression values of six genes (*LkCOMT7*, *LkCOMT8*, *LkLAC23*, *LkLAC102*, *LkPRX148*, and *LkPRX166*) have significantly positive correlations with lignin content. These results indicated that the increased expression of these six genes might be responsible for the high lignin content of the larches' wood. Overall, this study provides new genome resources for investigating the evolution and biological function of conifer trees, and also offers new insights into wood properties of larches.

Keywords: evolution, gene expression, genome, *Larix kaempferi*, wood properties

Sun, C., Xie, Y.H., Li, Z., Liu, Y.J., Sun, X.M., Li, J.J., Quan, W. P., Zeng, Q.Y., Van de Peer, Y., and Zhang, S.G. (2022). The *Larix kaempferi* genome reveals new insights into wood properties. J. Integr. Plant Biol. **64:** 1364–1373.

INTRODUCTION

The genus *Larix* Mill., one of the 11 Pinaceae genera, includes 10 species of deciduous trees, and is widely distributed over the temperate and boreal regions of the Northern Hemisphere (Shearer, 2008). This genus is also one of the most widely spread and economically important genera among conifers in the world (Goryachkina et al., 2013). Larch is an important component of forests in Northern Eurasia, and mountain forests in Eastern Europe and Western North America (Schmidt, 1995). In China, the area of larch forest is approximately 10.69 million hectares, of

which 3.14 million hectares are planted forests. China has more larch plantations than any other country in the world.

Japanese larch (*Larix kaempferi*) is naturally distributed on Honshu Island, Japan (Hoshi, 2004). Compared with other larch trees, *L. kaempferi* grows faster at early age, has higher wood density, longer fiber, and can adapt well to different environments (Kurinobu, 2005; Sun et al., 2017). Because of these desirable characteristics, *L. kaempferi* has been introduced into Europe, North China, and Northeast America since 1861 (Park and Fowler, 1983). *L. kaempferi* is now recognized as important for timber production, habitat or food for wildlife, watershed protection, environmental forestry, and also for ornamental purposes (Shearer, 2008). Thus, *L. kaempferi* has become a preferred afforestation trees in the Northern Hemisphere. In addition, *L. kaempferi* is widely used as one of the parental species in tree breeding programs to hybridize with other *Larix* species, such as *L. decidua* from South and East Europe, *L. gmelinii* from East Eurasia, and *L. principis-rupprechtii* from North China (Kurinobu, 2005).

The wood of *L. kaempferi* is excellent material for pulp production and papermaking (Isebrands and Hunt, 2007). These uses mainly depend on the quality of the wood of *L. kaempferi*. The content of lignin could affect the quality of wood (Packman, 1966). As an important structural constituent of the secondary cell wall, lignin also affects the growth and development of trees (Boerjan, Chanoca, and de Vries, 2019). However, in the papermaking process, the high content of lignin in wood will limit the output of pulp, cause discoloration and reduce the whiteness of paper, and increase the cost of papermaking (Baucher et al., 2003). Creating trees with low lignin content is an important goal of *L. kaempferi* breeding. The pathway of lignin biosynthesis had been well studied in angiosperms, such as *Arabidopsis* and *Populus* (Vanholme et al., 2010; Zhao, 2016). However, because the genomes of conifers are very large with sizes ranging between 4 and 35 Gb (Zonneveld, 2012), and are highly repetitive (Neale and Wheeler, 2019), the functional genes, transcription factors, and gene regulatory network responsible for lignin biosynthesis have not been well studied in gymnosperms, such as *L. kaempferi*. Thus, the whole genome sequence will help to reveal the genetic mechanism of wood quality formation, such as the biosynthetic pathway of lignin.

Thirty-one years ago, we developed a control-pollinated population that contain 80 trees yielded from eight full-sib families of *L. kaempferi*. This population is divided into two groups with one group having high lignin content and the other group having low lignin content. Using the single-molecule real-time (SMRT) sequencing technology and a hybrid assembling strategy, this study presents a genome assembly of *L. kaempferi*. Based on the genome and population transcriptome analysis, this study identified the key genes responsible for the high lignin content of *L. kaempferi* wood. In addition, we found that tandem duplication contributed to adaptive evolution of *L. kaempferi*. Overall, this study provides new genome resources for investigating the evolution and biological function of conifer trees, and also offers new insights into larch wood properties.

RESULTS

Genome sequencing, assembly, and validation

A previous flow cytometry analysis showed that the genome size of *L. kaempferi* is 12.9 Gb (Zonneveld, 2012). Using different sequencing technologies, we sequenced genomic DNA of young needles from a 31-year-old branch-grafted individual of the *L. kaempferi* superior clone 'RF27'. In total, we obtained 1.30 Tb (~100×) PacBio CLR data and 0.52 Tb (~40×) Illumina clean data, and scanned 54,913,912 BioNano molecules with total length of 1.64 Tb (~126×). A K-mer analysis based on the Illumina clean data estimated the genome size of *L. kaempferi* at approximately 11.32 Gb, thus slightly smaller than the estimate made by flow cytometry. Heterozygosity was estimated to be 1.8%–2.0% by the K-mer analysis (Figure S1).

Assembling the PacBio and Illumina data (see Materials and Methods), we acquired a 12.95 Gb *L. kaempferi* genome with 65,219 contigs. The contig N50 size is 447.85 Kb, significantly larger than any of the other sequenced gymnosperm genomes so far (Figure S2; Table 1). Next, we used BioNano data to perform error corrections and scaffolding on the contigs, and further eliminated redundant sequences by self-self-BLAST of assembled scaffolds. The clean data of 54,913,912 BioNano molecules were utilized to assemble genome maps by the BioNano Solve v3.0 (https://bionanogenomics.com/). We applied the merged consensus maps to scaffolding and orienting the assembled contigs. After scaffolding, possible redundant/heterozygous scaffolds were removed to further improve the genome assembly (for detailed methods see Methods S2). The final size of the genome assembly was 10.97 Gb, including 507.40 Mb of Ns in order to fill the gaps during scaffolding. The final genome assembly contains 24,640 scaffolds, and the N50 size and the longest scaffold are 1.09 Mb and 11.17 Mb, respectively (Table 1). The GC content of *L. kaempferi* genome is 38.03%, which is consistent with previous observations (35.10%–38.92%) for other gymnosperm genomes (Nystedt et al., 2013; Neale et al., 2014; Zimin et al., 2014; Guan et al., 2016; Stevens et al., 2016; Neale et al., 2017; Wan et al., 2018; Kuzmin et al., 2019; Mosca et al., 2019).

Table 1. Statistics for *Larix kaempferi* genome assembly v1.0

Characteristics	Values
Total length of scaffolds (bp)	10,969,669,938
Number of scaffolds	24,640
N50 size of scaffolds (bp)	1,086,670
N90 size of scaffolds (bp)	176,175
Longest scaffold length (bp)	11,173,318
Total length of contigs (bp)	12,952,415,590
Number of contigs	65,219
N50 size of contigs (bp)	447,849
N90 size of contigs (bp)	88,630
Longest contig length (bp)	6,096,712
GC content (%)	38.03

To assess genome accuracy and integrity, we performed bacterial artificial chromosome (BAC) sequencing, unigene mapping, and Benchmarking Universal Single-Copy Orthologs (BUSCO) evaluation. For the randomly sequenced 37 BAC clones, 25 (67.6%) of them were successfully validated, including 12 perfectly mapped BAC clones, and 13 highly mapped (mapping percentage > 97%) BAC clones (Figure S3). For the 164,300 unigenes assembled from expressed sequence tags (ESTs) (Chen et al., 2015), 156,836 (95.5%) could be appropriately mapped to the *L. kaempferi* genome. The percentage of complete BUSCOs in the *L. kaempferi* genome was 84.63%, much higher than that of other sequenced conifers ranging from 28.07% to 74.78% (Table S1).

Genes, repeats, and noncoding RNAs annotation

We annotated protein-coding genes, repeats, and noncoding RNAs in the *L. kaempferi* genome. Through comprehensive gene-finding by ab initio and homology-based methods (see Materials and Methods), we predicted 45,828 high-confidence protein-coding genes in the *L. kaempferi* genome. These genes are distributed over 15,670 (63.6%) scaffolds and their combined lengths cover 9.64% (1.06 Gb) of the genome. The average length of *L. kaempferi* genes is 23.1 Kb, similar to that in *G. biloba* (22.8 Kb) (Guan et al., 2016), but longer than those in other sequenced gymnosperms and angiosperms (Table S2). To avoid influence from different qualities in genome assembly and annotation of conifer genomes, we further identified a total of 50,139 orthogroups by OrthoMCL using 14 land plant genomes (see Materials and Methods) and compared the features of genes from 3,823 orthogroups shared by seven gymnosperms (*L. kaempferi*, *Gnetum montanum*, *Ginkgo biloba*, *Picea abies*, *Pinus taeda*, *Pinus lambertiana*, and *Pseudotsuga menziesii*). We found that although the average gene length of each gymnosperm varies tremendously, ranging from 4.98 to 33.6 Kb, the average lengths of coding sequences of these genes fall into a relatively narrow range from 1.3 to 1.8 Kb, suggesting that the various average gene lengths in gymnosperms are mainly affected by the considerable differences in average intron lengths (Table S3). For the 45,828 high-confidence protein-coding genes in the genome of *L. kaempferi*, 39,447 (86.1%) of the genes have homologs in at least one public protein database (Table S4). Also, 38,798 (84.7%) of the genes were supported by *L. kaempferi* transcriptome data sets generated from PacBio full-length transcriptome sequencing and Illumina sequencing (Table S5). These results indicated the high quality of assembly and annotation for the *L. kaempferi* genome.

We also annotated a total of 26,633,716 repetitive elements in the *L. kaempferi* genome, covering 66.8% (7.32 Gb) of the genome sequence (Table S6). This percentage is generally consistent with previous observations in other sequenced gymnosperms, ranging from 69.53% in *P. abies* (Nystedt et al., 2013) to 79% in *P. taeda* and *P. lambertiana* (Neale et al., 2014; Zimin et al., 2014; Stevens et al., 2016). Long terminal repeats (LTRs) dominate the repetitive content in the *L. kaempferi* genome, with 69.86% (5.12 Gb) of the total repeats (Table S6). Additionally, we annotated 16,588 non-coding RNAs in the *L. kaempferi* genome, including 210 miRNA precursors, 15,260 tRNAs, 604 snRNAs, and 514 rRNAs.

Genome evolution and phylogeny

To determine the phylogenetic position of *L. kaempferi*, we used the 50,139 orthogroups identified by OrthoMCL as described above. Among these gene families, 5,537 orthogroups are shared by all five conifers, while 1,615 orthogroups are unique to *L. kaempferi* (Figure 1A). We found that genes among the shared orthogroups of conifers are mainly involved in general processes such as protein phosphorylation, oxidation reduction, metabolism, and transmembrane transport, etc. (Table S7). Genes belonging to the 1,615 orthogroups specific to *L. kaempferi* are mainly involved in the processes and/or pathways of DNA replication initiation, arginine biosynthesis, glycine catabolism, and nucleocytoplasmic transport, etc. (Table S8). Only three single-copy genes were shared among all the 14 land plant species. These three single-copy genes as well as 257 low-copy number genes (i.e., the genes that exist in single-copy status in 12 or 13 species) of the 14 land plants were used to construct a phylogenetic tree by the GTRGAMMA model of RaxML. The phylogenetic tree shows that *L. kaempferi* diverged from *P. menziesii* at approximately 65.9 (11.7–111.5) million years ago (Figure 1B). Since their divergence, in *L. kaempferi*, 1,139 gene families have expanded, and 581 gene families became smaller (Figure 1B).

Here, we investigated tandem duplication of genes for the 14 land plants aforementioned. Compared to the other 13 plant genomes, the *L. kaempferi* genome has the largest number (7,450 out of 45,828) of tandem duplicated genes (Table S9). The proportion (16.26%) of tandem duplicated genes in the *L. kaempferi* genome is significantly ($P < 2.80E\text{-}49$, χ^2-test) higher than that of any other sequenced gymnosperm (Figure 2A). This study identified 2,744 tandem repeat clusters (TRCs) in the *L. kaempferi* genome. More interestingly, there are 451 TRCs with more than three genes (Figure S4). GO enrichment analyses of the tandem duplicated genes in *L. kaempferi* show that they are mainly involved in processes such as oxidation reduction, the hydrogen peroxide catabolism, DNA-dependent transcription regulation, response to oxidative stress, carbohydrate metabolism, and response to stress (Figure 2B; Table S10).

Lignin biosynthetic genes for wood properties in *L. kaempferi*

We developed a control-pollinated population of 80 trees from eight full-sib families of *L. kaempferi* 31 years ago. Phenotypes of wood properties and growth traits were measured for each individual (see Materials and Methods). According to the wood lignin content, this population is divided into two groups, one with high lignin content (HLC) and the other with low lignin content (LLC). Each group contains 40 individual larch trees, where the lignin content of HLC trees is significantly ($P < 0.001$, *t*-test) higher than that of LLC

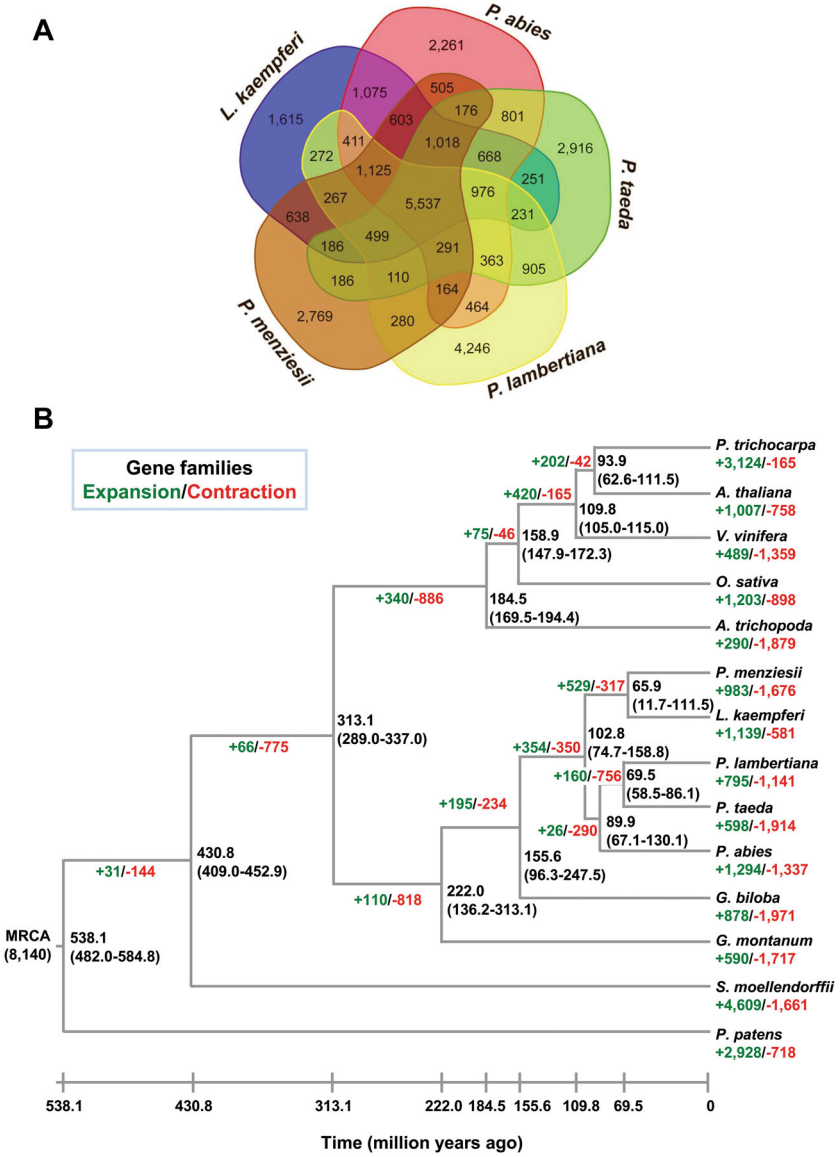

Figure 1. Evolution of *Larix kaempferi* genome

(**A**) Comparison of the numbers of gene families in five conifer trees. (**B**) Phylogenetic tree and gene family expansion/contraction of 14 land plant species. The phylogenetic tree was constructed using three single-copy and 257 low-copy genes of 14 plant species. Expansions and contractions of gene family are indicated in green and red, respectively. MRCA, most recent common ancestor.

trees (Figure 3A). However, no significant differences are found in tree height and breast diameter between the two groups ($P > 0.10$, *t*-test; Figure 3B, C; Table S11).

Lignin biosynthesis includes monolignol biosynthesis and monolignol polymerization, and the genes involved as well as their upstream transcriptional regulators are well deciphered in angiosperms (Zhao, 2016). To study transcriptional regulation of the difference in lignin content in the HLC and LLC groups, we performed population transcriptome analysis on immature xylem of the 80 individual trees. We identified 182 differentially expressed genes (DEGs) between the HLC and LLC groups. Comparing with the LLC group, expression levels of 80 genes are significantly upregulated in the HLC group, while that of 102 genes are significantly downregulated in the HLC group (Figure S6; Tables S12, S13). Functional annotation showed that nine DEGs are involved in the lignin biosynthesis pathway (Figure 4), including four *caffeic acid O-methyltransferases* (*LkCOMT7*, *LkCOMT8*, *LkCOMT15*, and *LkCOMT19*), two *laccases* (*LkLAC23* and *LkLAC102*), and three *peroxidases* (*LkPRX148*, *LkPRX155*, and *LkPRX166*).

Among the nine DEGs involved in the lignin biosynthesis pathway, four genes (*LkCOMT7*, *LkCOMT8*, *LkCOMT15*, and *LkCOMT19*) are involved in monolignol biosynthesis, and five genes (*LkLAC23*, *LkLAC102*, *LkPRX148*, *LkPRX155*, and *LkPRX166*) are involved in the monolignol polymerization

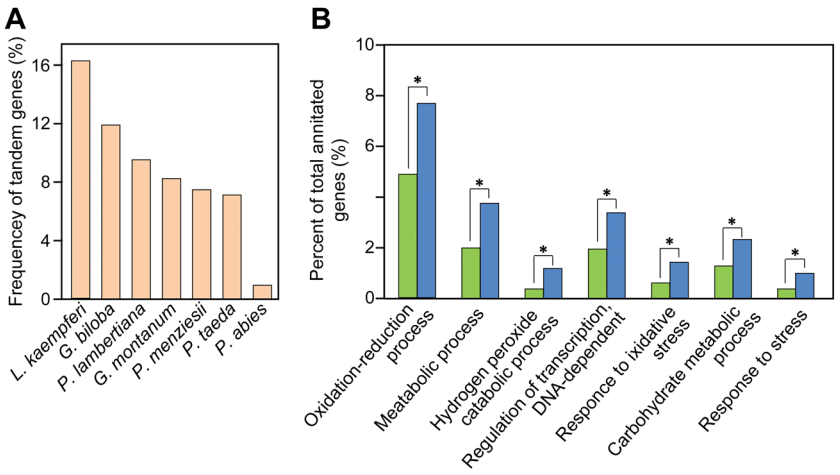

Figure 2. Tandem duplicated genes in *Larix kaempferi*

(**A**) Frequency of tandem genes in seven sequenced gymnosperms. (**B**) Prevalence of tandem genes involved in certain Biological Process. Blue bars indicate the proportion of all tandem genes in the genome belonging to different gene ontology terms. Green bars indicate the proportion of all genes in the genome belonging to different gene ontology terms. The asterisk indicates $P < 1e{-}5$ (Fisher test).

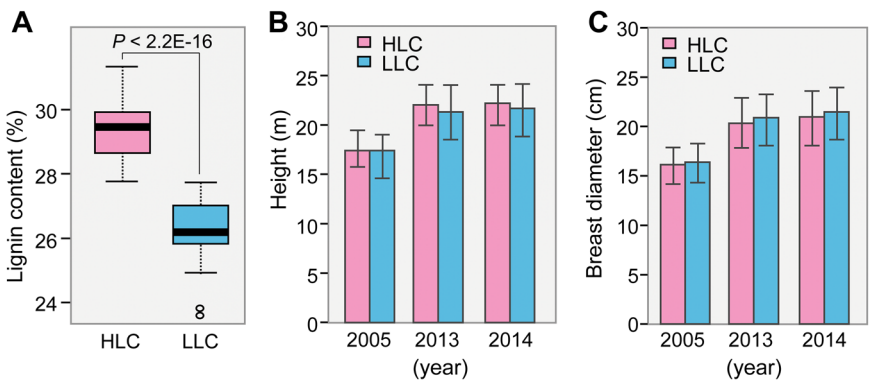

Figure 3. Phenotype analysis of the high lignin content (HLC) and low lignin content (LLC) groups of *Larix kaempferi*

(**A**) Comparison of wood lignin content between HLC and LLC groups. (**B**) Comparison of tree height between HLC and LLC groups. (**C**) Comparison of breast diameter between HLC and LLC groups. The *t*-test was used to analyze the statistical differences in phenotypes between the two groups.

process (Figure 4). Among these nine DEGs, expression levels of six (*LkCOMT7, LkCOMT8*, *LkLAC\23*, *LkLAC102*, *LkPRX148*, and *LkPRX166*) are significantly upregulated in the HLC group compared to the LLC group, while that of three (*LkCOMT15*, *LkCOMT19*, and *LkPRX155*) are significantly downregulated in the HLC group (Figure 5). These results suggest that lignin content differences between HLC and LLC groups is mainly determined by the monolignol polymerization process.

The correlations of lignin content and expression values of the nine DEGs involved in the lignin biosynthesis pathway were estimated (Table 2). We found that the expression values of all the six upregulated DEGs (*LkCOMT7*, *LkCOMT8, LkLAC23*, *LkLAC102*, *LkPRX148*, and *LkPRX166*) had significantly positive correlation ($r > 0.31$, $P < 0.01$) with lignin content. However, three downregulated DEGs (*LkCOMT15*, *LkCOMT19*, and *LkPRX155*) showed significantly negative correlation ($r < -0.44$, $P < 0.001$) with lignin content. These results indicated that the increased expression of these six upregulated DEGs might be responsible for the high lignin content in the HLC group.

DISCUSSION

Compared with most angiosperms, gymnosperms have generally larger genomes (Gaut and Ross-Ibarra, 2008). For example, the genome sizes of conifer trees vary from 4 Gb to 35 Gb, and are highly repetitive (Zonneveld, 2012; Neale and Wheeler, 2019). *L. kaempferi* is a deciduous conifer tree with a genome size of 12.9 Gb (Zonneveld, 2012). In this study, the *L. kaempferi* genome was completely assembled using PacBio sequencing data (~100-fold coverage). The PacBio SMRT sequencing can largely improve the contiguity of genome assembly, because its long sequencing reads can overcome problems caused by repetitive sequences which

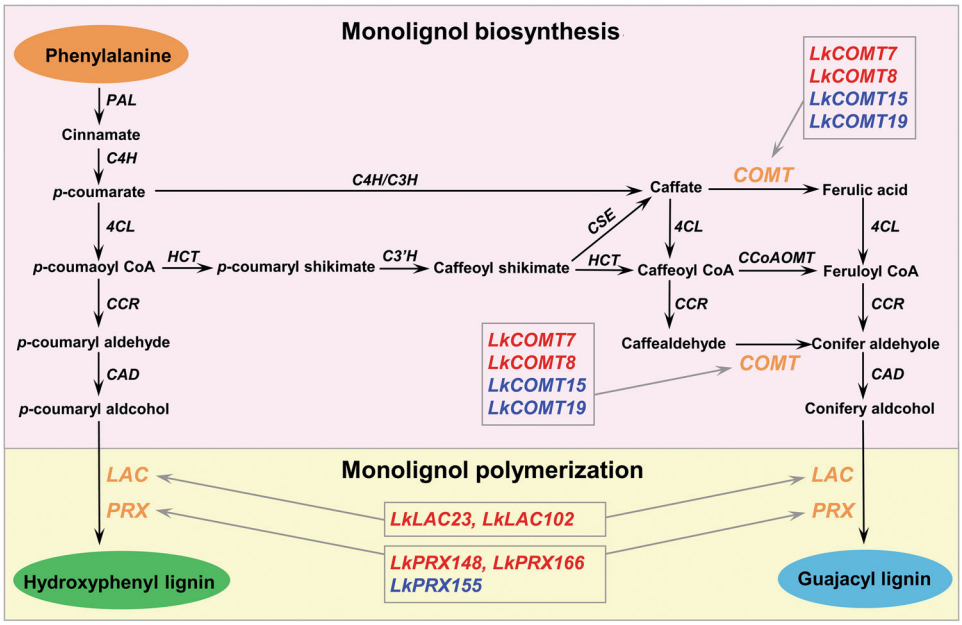

Figure 4. The lignin biosynthesis pathway in *Larix kaempferi*

The orange color letters indicate differentially expressed genes between high lignin content (HLC) and low lignin content (LLC) groups. Compared to the LLC group, the upregulated and downregulated genes in the HLC group were indicated in red and blue color letters, respectively. *4CL*, 4-coumarate CoA ligase; *C4H*, cinnamate 4-hydroxylase; *C3H*, coumarate 3-hydroxylase; *CAD*, cinnamyl alcohol dehydrogenase; *CCR*, cinnamoyl CoA reductase; *HCT*, hydroxycinnamoyl CoA:shikimate hydroxycinnamoyl transferase; *C3′H*, p-coumaroyl shikimate3′-hydroxylase; *CSE*, caffeoyl shikimate esterase; *COMT*, caffeic acid O-methyltransferase; *CCoAOMT*, caffeoyl CoA O-methyltransferase; *LAC*, laccase; *PAL*, phenylalanine ammonia lyase; *PRX*, peroxidase.

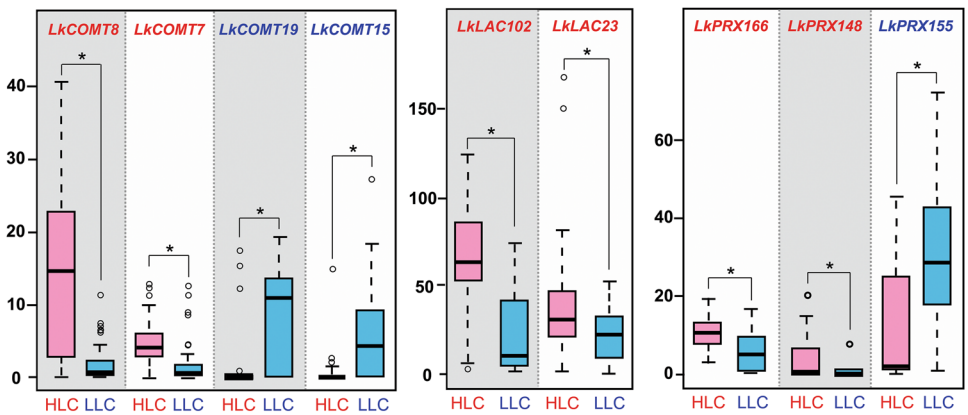

Figure 5. Relative expression levels of DEGs between HLC and LLC groups

The ordinate represents FPKM. The asterisk indicates $P < 0.01$ (*t*-test). DEGs, differentially expressed genes; HLC, high lignin content; LLC, low lignin content.

are mostly longer than the Illumina reads (Zimin et al., 2017a). Indeed, for the *P. taeda* genome, adding approximately 12-fold coverage PacBio data has increased the contig N50 size from 8,206 bp to 25,361 bp (Zimin et al., 2017b). Also here, comparing with other gymnosperm genomes with similar genome sizes as the *L. kaempferi* genome, the application of the PacBio sequencing technology to *L. kaempferi* has resulted in the much continuous genome at the contig level. At the scaffold level, the scaffold N50 size of the *L. kaempferi* genome (1.09 Mb) is much longer than that of other sequenced gymnosperms (Figure S2), except *G. biloba*, which has a slightly longer scaffold N50 size of 1.36 Mb (Guan et al., 2016). This might be caused by the differences in genomic heterozygosity of the sequencing materials of these two species. For the *L. kaempferi*, we sequenced young leaf needles with a diploid genome, of which the heterozygosity is 1.8%–2.0%. For the *G. biloba*, the large endosperm tissue from one ginkgo seed was sequenced. The endosperm tissue was not fertilized and only had a haploid genome (Guan et al., 2016). As a large conifer genome with a size of 12.9 Gb, the assembly of the *L. kaempferi* genome represents one of the best gymnosperm genomes.

Table 2. Pearson correlation coefficients between lignin content and expression values of differentially expressed genes (DEGs) involved in lignin biosynthesis

Gene symbol	*r*	*P*
LkCOMT7	0.382	0.00047
LkCOMT8	0.606	2.64E−9
LkCOMT15	−0.459	1.84E−5
LkCOMT19	−0.475	8.66E−6
LkLAC23	0.315	0.00437
LkLAC102	0.690	1.44E−12
LkPRX148	0.423	9.41E−5
LkPRX155	−0.443	3.84E−5
LkPRX166	0.510	1.33E−6

Gene duplication is a prevailing feature in plant genomes (Ober, 2005). Gene duplications followed by functional divergence are a conspicuous feature of plant genes (Freeling, 2009). Among the sequenced gymnosperms, the *L. kaempferi* genome contained the highest proportion of tandem duplicated genes. The ratio of the number of TRCs with more than three genes to the number of all TRCs in *L. kaempferi* genome is 16.44%. This ratio was much higher than other sequenced gymnosperms (Figure S4). In addition, we found that many of the genes in these TRCs were stress-response genes. One possible explanation was the better quality of the genome assembly and gene annotation in *L. kaempferi*. The contig N50 size of the *L. kaempferi* genome (447.85 Kb) is significantly higher than that of other sequenced gymnosperms (ranging from 0.72 to 48.21 Kb) and its scaffold N50 size (1.09 Mb) is the longest among sequenced conifers. The longer contig length of the *L. kaempferi* genome allowed prediction of more complete gene models, which was critical for tandem duplication detection. The expansion of stress-response genes through tandem duplication in *L. kaempferi* is consistent with previous observation made in eucalypt (Myburg et al., 2014), oak (Plomion et al., 2018), and mangrove (Xu et al., 2017). Thus, tandem duplications might be an important mechanism for adaptive evolution of trees to rapidly changing environments (Hanada et al., 2008).

Lignin biosynthesis is a complex process that involves monolignol biosynthesis and monolignol polymerization (Zhao, 2016). The lignin content might be determined by the interaction of many factors, including phenylalanine content, activities of lignin biosynthetic enzyme, and expression level of lignin biosynthetic genes. Based on population transcriptome analyses, this study showed that six genes (*LkCOMT7, LkCOMT8*, *LkLAC23*, *LkLAC102*, *LkPRX148*, and *LkPRX166*) might be responsible for high lignin content in the HLC group of *L. kaempferi*. Downregulation of *COMT* gene expression in angiosperm trees significantly reduced the amount of lignin S units (Lapierre et al., 1999). In *Medicago sativa*, inhibiting the expression of the *COMT* gene resulted in a decrease in the biosynthesis of G and S units (Guo et al., 2001). The *COMT* gene was involved in methylation of caffealdehyde for G unit biosynthesis (Wang et al., 2015). Conifer trees could not biosynthesize syringyl lignin (Zhao, 2016). In *L. kaempferi*, the upregulation of *LkCOMT7* and *LkCOMT8* gene expression indicated a relatively greater amount of flux into biosynthesis of conifer aldehyole, which resulted in an increase in the accumulation of G unit in the HLC group.

It was well known that monolignol polymerization was mainly catalyzed by laccases and peroxidases (Higuchi, 1985). The significantly positive correlation between *LkLAC23/LkLAC102* and lignin content was consistent with previous observation that laccases were the most abundant proteins in lignin-rich compression wood of *Pinus radiata* (Mast et al., 2010). Previous studies showed that downregulation of *PRX* gene expression could cause a nearly 20% reduction of the lignin content in transgenic poplar (Li et al., 2003). In the catalyzed reaction of PRX protein, coniferyl alcohol was usually the preferred substrate in conifer trees (Marjamaa et al., 2003). Upregulation of the *LkPRX148/LkPRX166* gene expression might enhance the polymerization of G unit, and subsequently increase the content of guajacyl lignin in *L. kaempferi*.

MATERIALS AND METHODS

Plant materials

We sequenced a 31-year-old grafted ramets of the *Larix kaempferi* clone 'RF27', growing in a seed orchard (42°38′ N, 124°86′ E), Dagujia Forest Farm, Liaoning Province, P.R. China, which originated from an elite tree in the plantation of Fushun Region, Liaoning Province, and was collected in 1962. 'RF27' is a superior clone, either as material or paternal parent, with rapid growth, excellent wood properties, and high paternal contributions (Chen et al., 2018). Many copies of the tree are available in seed orchards and breeding gardens, and it has been widely used in Chinese larch breeding programs.

Genome sequencing

Using the modified CTAB method (Allen et al., 2006), genomic DNA was extracted by Qiagen DNeasy Plant Mini Kit (TIANGEN Biotech) from its young needles to construct PacBio, Illumina, BioNano, and BAC sequencing libraries. Moreover, total RNA of 11 tissues (root, phloem, cambium, xylem, branch, spore, sprout, needle, male flower, female flower, cone) in the same tree was extracted by Qiagen RNAprep Pure Kit (TIANGEN Biotech), and then equally mixed into one sample to carry out full-length transcriptome sequencing. In order to quantify gene expression, we performed RNA-Seq on five tissues: root, needle, phloem, immature cone, and immature xylem.

For genomic DNA *de novo* sequencing, we constructed PacBio libraries of young needles following the standard protocol (Pacific Biosciences). A total of 718 SMRT cells with Subreads N50 length of 11.6 Kb were sequenced on the

platforms of RSII (528 cells) and Sequel (190 cells), respectively. This generated 1.30 Tb (~100×) SMRT data. For BioNano optical mapping, high-molecular-weight DNA was extracted, and digested by Nt. BspQI (NEB). The labeled repaired DNA was scanned by Saphyr system (BioNanoGenomics). The BAC DNA libraries were constructed following the method reported previously (Shi et al., 2011). For full-length transcriptome sequencing, we recovered transcripts with different lengths (i.e., < 1, 1–2, 2–3 and >3 Kb) and accordingly built four PacBio sequencing libraries. We sequenced a total of 25 SMRT cells for these four libraries on RSII platform, and then 56 Gb PacBio data generated.

The detailed methods of library construction, genome sequencing, and full-length transcriptome sequencing are described in Method S1.

Genome survey, assembly, and validation

Using the 0.52 Tb Illumina clean data, we first estimated the genome size and heterozygosity of *L. kaempferi* by K-mer analysis (Wuyun et al., 2018). We then used the 1.30 Tb PacBio data to assemble the *L. kaempferi* genome by Falcon v1.8.7 (Chin et al., 2016). To validate the continuity and completeness of the genome assembly, we used three strategies of BAC clone mapping, unigene mapping, and BUSCO (http://busco.ezlab.org/) evaluation (Manni et al., 2021). The detailed methods of genome survey, assembly, and validation are described in Method S2.

Genes, repeats, and noncoding RNA annotation

Two strategies of *de novo* prediction and homology identification were used to annotate repeats. We predicted repeats by RepeatModeler v1.0.8, and the results were combined with the repeat sequences from Repbase (Jurka et al., 2005) and Mips-REdat (Nussbaumer et al., 2012) into one set. Using this set, we annotated repeats in the *L. kaempferi* genome by RepeatMasker v4.0.7. The LTR elements and tandem repeat sequences were identified through LTR_Finder v1.06 and TRF v4.09, respectively.

The integrative gene-finding algorithm (*de novo* prediction, homology identification, and transcriptome analysis) was used to annotated genes (see Figure S5). To identify high-quality isoforms, the full-length transcripts from Iso-seq were mapped on the *L. kaempferi* genome by GMAP v2016-09-23. We then used the collapsed high-quality isoforms to annotate *L. kaempferi* genes by Augustus v3.2.2 and PASA v2.2.0. Using proteins of five gymnosperms, we identified *L. kaempferi* genes by Genewise v2.4.1. The results above were merged through EVM v1.1.1 and manual correction, and the pseudogenes as well as transposons were removed. We used InterProScan v5.25-64.0 and BLASTP v2.2.26 to annotate gene function. Expression levels of the genes were estimated in five sets of *L. kaempferi* transcriptome by genome mapping. In addition, we annotated noncoding RNAs (miRNAs, snRNAs, rRNAs, and tRNAs) by Infernal v1.1.2, RNAmmer Server v1.2 and tRNAscan-SE v1.3.1.

The detailed methods of genes, repeats, and noncoding RNA annotation are described in Method S3.

Genome evolution and phylogeny analysis

This study selected 14 land plants to perform evolution and phylogeny analysis. The gene family clusters were identified through OrthoMCL v2.0.9, and a phylogenetic tree was constructed with three single-copy genes and 257 low-copy number genes by RaxML v8.2.10q through the GTRGAMMA model. We estimated the divergence time of 14 plants by MCMCTREE of PAML v4.9e, and verified it with TimeTree database (http://www.timetree.org/). CAFE v2.2 was used to perform expansion/contraction analysis of the gene families. The tandem genes of 14 land plants were identified through SynOrths v1.0. Using the whole genome genes as background, functional enrichment analysis was performed on tandem genes of *L. kaempferi*. The detailed methods of genome evolution and phylogeny analysis are described in Method S4.

Phenotype measurement and transcriptome sequencing of *L. kaempferi* population

A control-pollinated population that contains 80 trees yielded from eight full-sib families of *L. kaempferi* was developed 31 years ago. Phenotype characteristics (wood lignin content, tree height, and breast diameter) of each individual in this population were determined. Based on lignin content, the population was divided into two groups: HLC and LLC groups. During July 2016, we sampled the immature xylem of the 80 trees. Subsequently, the transcriptomes of these samples were sequenced and analyzed.

The detailed methods of phenotype measurement, sample collection, transcriptome sequencing, and transcriptome analysis are described in Method S5.

Data availability statement

The contig sequence of *L. kaempferi* genome, population transcriptome and transcriptome of five tissues were deposited to the NCBI database with Accession Numbers PRJNA587041 (WOXR00000000.2), PRJNA588099 (SRR11829003–SRR11829082), and PRJNA588100 (SRR11810205–SRR11810219). The *L. kaempferi* genome is available at the National Genomics Data Center (https://ngdc.cncb.ac.cn/) under BioProject no. PRJCA008850 and the LarixGD website (https://www.larixgd.cn/).

ACKNOWLEDGEMENTS

We acknowledge Dr. Shae He, Dr. Leiming Dong, and Dr. Weibo Xiang for their help in preparing sequencing samples. We also thank Ruixue Li for her help in data analysis. This work was funded by the Forestry Industry Research Special Funds for Public Welfare Projects (201504104), National Natural Science Foundation of China (31901335), Fundamental Research Funds for the Central Non-profit Research Institution of CAF (CAFYBB2017QA001, CAFYBB

2018ZY001-4, LYSZX202002), and National Transgenic Major Program (2018ZX08020-003). Z.L. is funded by a postdoctoral fellowship from the Special Research Fund of Ghent University (BOFPDO2018001701).

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

AUTHOR CONTRIBUTIONS

S.G.Z., Q.Y.Z., and X.M.S. designed the project. Y.H.X. collected plant materials and designed the population transcriptome experiment. W.P.Q., J.J.L., and C.S. performed sequencing, genome assembly and annotation. C.S. carried out comparative genomics analysis. Y.X. and C.S. performed transcriptome analysis. Y.J.L., Z.L., and Y.V. d.P. made constructive comments on the work and manuscript. C.S. and Q.Y.Z. wrote the manuscript. All authors read and approved of this manuscript.

Edited by: Xuehui Huang, Shanghai Normal University, China.

Received Mar. 10, 2022; **Accepted** Apr. 18, 2022; **Published** Apr. 20, 2022

OO: OnlineOpen

REFERENCES

Allen, G., Flores-Vergara, M., Krasynanski, S., Kumar, S., and Thompson, W. (2006). A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat. Protoc. **1:** 2320.

Baucher, M., Halpin, C., Petit-Conil, M., and Boerjan, W. (2003). Lignin: Genetic engineering and impact on pulping. Crit. Rev. Biochem. Mol. Biol. **38:** 305–350.

Boerjan, W., Chanoca, A., and de Vries, L. (2019). Lignin engineering in forest trees. Front. Plant Sci. **10:** 912.

Chen, X.B., Xie, Y.H., and Sun, X.M. (2015). Development and characterization of polymorphic genic-SSR markers in *Larix kaempferi*. Molecules **20:** 6060–6067.

Chen, X., Sun, X., Dong, L., and Zhang, S. (2018). Mating patterns and pollen dispersal in a Japanese larch (*Larix kaempferi*) clonal seed orchard: A case study. Sci. China: Life Sci. **61:** 1011–1023.

Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., and Morales-Cruz, A. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods **13:** 1050.

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. Annu. Rev. Plant Biol. **60:** 433–453.

Gaut, B.S., and Ross-Ibarra, J. (2008). Selection on major components of angiosperm genomes. Science **320:** 484–486.

Goryachkina, O.V., Badaeva, E.D., Muratova, E.N., and Zelenin, A.V. (2013). Molecular cytogenetic analysis of siberian *Larix* species by fluorescence in situ hybridization. Plant Syst. Evol. **299:** 471–479.

Guan, R., Zhao, Y., Zhang, H., Fan, G., Liu, X., Zhou, W., Shi, C., Wang, J., Liu, W., and Liang, X. (2016). Draft genome of the living fossil *Ginkgo biloba*. Gigascience **5:** 49.

Guo, D., Chen, F., Inoue, K., Blount, J.W., and Dixon, R.A. (2001). Downregulation of caffeic acid 3-O-methyltransferase and caffeoyl CoA 3-O-methyltransferase in transgenic alfalfa: Impacts on lignin structure and implications for the biosynthesis of G and S lignin. Plant Cell **13:** 73–88.

Hanada, K., Zou, C., Lehti-Shiu, M.D., Shinozaki, K., and Shiu, S.H. (2008). Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. **148:** 993–1003.

Higuchi, T. (1985). Biosynthesis of lignin. In: Higuchi, T., ed. *Biosynthesis and Biodegradation of Wood Components*. Academic Press, Florida, USA. pp. 141.

Hoshi, H. (2004). Forest tree genetic resources conservation stands of Japanese larch (*Larix kaempferi*(Lamb.) Carr.). Forest Tree Genetic Resources Information Special Issue No 1.

Isebrands, J., and Hunt, C. (2007). Growth and wood properties of rapid-grown Japanese larch. Wood Fiber Sci. **7:** 119–128.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. **110:** 462–467.

Kurinobu, S. (2005). Forest tree breeding for Japanese larch. Eurasian J. For. Res. **8:** 127–134.

Kuzmin, D.A., Feranchuk, S.I., Sharov, V.V., Cybin, A.N., Makolov, S.V., Putintseva, Y.A., Oreshkova, N.V., and Krutovsky, K.V. (2019). Stepwise large genome assembly approach: A case of Siberian larch (*Larix sibirica* Ledeb). BMC Bioinfor. **20:** 37.

Lapierre, C., Pollet, B., Petit-Conil, M., Toval, G., Romero, J., Pilate, G., Leplé, J.-C., Boerjan, W., Ferret, V., and De Nadai, V. (1999). Structural alterations of lignins in transgenic poplars with depressed cinnamyl alcohol dehydrogenase or caffeic acidO-methyltransferase activity have an opposite impact on the efficiency of industrial kraft pulping. Plant Physiol. **119:** 153–164.

Li, Y., Kajita, S., Kawai, S., Katayama, Y., and Morohoshi, N. (2003). Down-regulation of an anionic peroxidase in transgenic aspen and its effect on lignin characteristics. J. Plant Res. **116:** 175–182.

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). 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.

Marjamaa, K., Lehtonen, M., Lundell, T., Toikka, M., Saranpää, P., and Fagerstedt, K.V. (2003). Developmental lignification and seasonal variation in β-glucosidase and peroxidase activities in xylem of Scots pine, Norway spruce and silver birch. Tree Physiol. **23:** 977–986.

Mast, S., Peng, L., Jordan, T.W., Flint, H., Phillips, L., Donaldson, L., Strabala, T.J., and Wagner, A. (2010). Proteomic analysis of membrane preparations from developing *Pinus radiata* compression wood. Tree Physiol. **30:** 1456–1468.

Mosca, E., Cruz, F., Gómez-Garrido, J., Bianco, L., Rellstab, C., Brodbeck, S., Csilléry, K., Fady, B., Fladung, M., and Fussi, B. (2019). A reference genome sequence for the European silver fir (*Abies alba* Mill.): A community-generated genomic resource. G3 **9:** 2039–2049.

Myburg, A.A., Grattapaglia, D., Tuskan, G.A., Hellsten, U., Hayes, R.D., Grimwood, J., Jenkins, J., Lindquist, E., Tice, H., and Bauer, D. (2014). The genome of *Eucalyptus grandis*. Nature **510:** 356.

Neale, D.B., McGuire, P.E., Wheeler, N.C., Stevens, K.A., Crepeau, M. W., Cardeno, C., Zimin, A.V., Puiu, D., Pertea, G.M., and Sezen, U.U. (2017). The Douglas-fir genome sequence reveals specialization of the photosynthetic apparatus in Pinaceae. G3 **7:** 3157–3167.

Neale, D.B., Wegrzyn, J.L., Stevens, K.A., Zimin, A.V., Puiu, D., Crepeau, M. W., Cardeno, C., Koriabine, M., Holtz-Morris, A.E., and Liechty, J.D. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. **15:** R59.

Neale, D.B., Wheeler, N.C. (2019). The conifers. In: Neale, D.B. and Wheeler, N.C., eds. *The Conifers: Genomes, Variation and Evolution*. Springer Nature Switzerland AG, Cham, Switzerland. pp. 1–21.

Nussbaumer, T., Martis, M.M., Roessner, S.K., Pfeifer, M., Bader, K.C., Sharma, S., Gundlach, H., and Spannagl, M. (2012). MIPS PlantsDB: A database framework for comparative plant genome research. Nucleic Acids Res. **41:** D1144–D1151.

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

Ober, D. (2005). Seeing double: Gene duplication and diversification in plant secondary metabolism. Trends Plant Sci. **10:** 444–449.

Packman, D. (1966). Pulping of British-grown softwoods. Part II. Preliminary studies on Japanese larch (*Larix leptolepis*). Holzforschung **20:** 110–113.

Park, Y., and Fowler, D. (1983). A provenance test of Japanese larch in eastern Canada, including comparative data on European larch and tamarack. Silvae Genet. **32:** 96–101.

Plomion, C., Aury, J.M., Amselem, J., Leroy, T., Murat, F., Duplessis, S., Faye, S., Francillonne, N., Labadie, K., and Le Provost, G. (2018). Oak genome reveals facets of long lifespan. Nat. Plants **4:** 440–452.

Schmidt, W.C. (1995). Around the world with *Larix*: An introduction. In: Schmidt, W.C. and McDonald, K.J., eds. *Ecology and Management of Larix Forests: A Look Ahead*. USDA Forest Service, Intermountain Research Station, Ogden, USA. pp. 6-18.

Shearer, R.C. (2008). Larix P. Mill.: Larch. In: Bonner, Franklin T., Karrfalt, Robert P., eds. *The Woody Plant Seed Manual*. Agric. Handbook No. 727. Washington, DC. US Department of Agriculture, Forest Service. pp. 637–650.

Shi, X., Zeng, H., Xue, Y., and Luo, M. (2011). A pair of new BAC and BIBAC vectors that facilitate BAC/BIBAC library construction and intact large genomic DNA insert exchange. Plant Methods **7:** 33.

Stevens, K.A., Wegrzyn, J., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., Paul, R., Gonzalez-Ibeaz, D., Koriabine, M., and Holtz-Morris, A. (2016). Sequence of the sugar pine megagenome. Genetics **204:** 1613–1626.

Sun, C., Lai, M., Zhang, S., and Sun, X. (2017). Age-related trends in genetic parameters for wood properties in *Larix kaempferi* clones and implications for early selection. Front. Agr. Sci. Eng. **4:** 482–492.

Vanholme, R., Demedts, B., Morreel, K., Ralph, J., and Boerjan, W. (2010). Lignin biosynthesis and structure. Plant Physiol. **153:** 895–905.

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

Wang, P., Dudareva, N., Morgan, J.A., and Chapple, C. (2015). Genetic manipulation of lignocellulosic biomass for bioenergy. Curr. Opin. Chem. Biol. **29:** 32–39.

Wuyun, T.-n, Wang, L., Liu, H., Wang, X., Zhang, L., Bennetzen, J.L., Li, T., Yang, L., Liu, P., and Du, L. (2018). The hardy rubber tree genome provides insights into the evolution of polyisoprene biosynthesis. Mol. Plant **11:** 429–442.

Xu, S., He, Z., Zhang, Z., Guo, Z., Guo, W., Lyu, H., Li, J., Yang, M., Du, Z., and Huang, Y. (2017). The origin, diversification and adaptation of a major mangrove clade (Rhizophoreae) revealed by whole-genome sequencing. Natl. Sci. Rev. **4:** 721–734.

Zhao, Q. (2016). Lignification: Flexibility, biosynthesis and regulation. Trends Plant Sci. **21:** 713–721.

Zimin, A., Stevens, K.A., Crepeau, M.W., Holtz-Morris, A., Koriabine, M., Marçais, G., Puiu, D., Roberts, M., Wegrzyn, J.L., and de Jong, P.J. (2014). Sequencing and assembly of the 22-Gb loblolly pine genome. Genetics **196:** 875–890.

Zimin, A.V., Puiu, D., Luo, M.C., Zhu, T., Koren, S., Marçais, G., Yorke, J.A., Dvořák, J., and Salzberg, S.L. (2017a). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. **27:** 787–792.

Zimin, A.V., Stevens, K.A., Crepeau, M.W., Puiu, D., Wegrzyn, J.L., Yorke, J.A., Langley, C.H., Neale, D.B., and Salzberg, S.L. (2017b). An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. Gigascience **6:** giw016.

Zonneveld, B. (2012). Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. Nord. J. Bot. **30:** 490–502.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article: http://onlinelibrary.wiley.com/doi/10.1111/jipb.13265/suppinfo

Figure S1. K-mer analysis of genome survey based on the Illumina sequencing data
Figure S2. Comparison of contig N50 and scaffold N50 sizes of *L. kaempferi* and other sequenced gymnosperms
Figure S3. BAC validation of the assembled *L. kaempferi* genome
Figure S4. The proportion of cluster containing different genes in the genome
Figure S5. Gene annotation strategy in this study
Figure S6. Volcano plot of the differential gene expression in immature xylems between HLC and LLC groups
Method S1. Library construction, genome sequencing, and full-length transcriptome sequencing
Method S2. Genome survey, assembly, and validation
Method S3. Genes, repeats, and noncoding RNA annotation
Method S4. Genome evolution and phylogeny analysis
Method S5. Phenotype measurement, sample collection, transcriptome sequencing, and transcriptome analysis
Table S1. BUSCO evaluation for the whole genome sequence among the nine sequenced gymnosperms
Table S2. Comparison of the characteristics of all genes in 14 plant species
Table S3. Comparison of gene characteristics of clusters shared by seven gymnosperms
Table S4. The number of genes assigned to different protein annotation databases
Table S5. Number of *L. kaempferi* genes expressed in five different transcriptome data
Table S6. Composition of repeat sequence in *L. kaempferi* genome
Table S7. The functional enrichment analysis of cluster genes shared by five conifer trees
Table S8. Functional enrichment analysis of *L. kaempferi*-specific genes compared to another four conifer trees: *P. abies*, *P. taeda*, *P. lambertiana*, and *P. menziesii*
Table S9. Number of tandem genes in the 14 sequenced plants
Table S10. Functional enrichment analysis of *L. kaempferi* tandem genes
Table S11. Sample information for population transcriptome analysis
Table S12. Upregulated genes in HLC group compared to LLC group
Table S13. Downregulated genes in HLC group compared to LLC group
Table S14. Genome resource used in this study