



Article

Annotation of Siberian Larch (*Larix sibirica* Ledeb.) Nuclear Genome—One of the Most Cold-Resistant Tree Species in the Only Deciduous GENUS in *Pinaceae*

Eugenia I. Bondar ^{1,2} , Sergey I. Feranchuk ¹, Ksenia A. Miroshnikova ^{1,2}, Vadim V. Sharov ^{1,2,3}, Dmitry A. Kuzmin ^{1,3}, Natalya V. Oreshkova ^{1,2,4} and Konstantin V. Krutovsky ^{1,5,6,7,8,9,*} 

- ¹ Laboratory of Forest Genomics, Institute of Fundamental Biology and Biotechnology, Siberian Federal University, 660036 Krasnoyarsk, Russia
- ² Laboratory of Genomic Research and Biotechnology, Federal Research Center “Krasnoyarsk Science Center”, Siberian Branch, Russian Academy of Sciences, 660036 Krasnoyarsk, Russia
- ³ Department of High-Performance Computing, Institute of Space and Information Technologies, Siberian Federal University, 660074 Krasnoyarsk, Russia
- ⁴ Laboratory of Forest Genetics and Selection, V. N. Sukachev Institute of Forest, Siberian Branch, Russian Academy of Sciences, 660036 Krasnoyarsk, Russia
- ⁵ Department of Forest Genetics and Forest Tree Breeding, Georg-August University of Göttingen, 37077 Göttingen, Germany
- ⁶ Center for Integrated Breeding Research, Georg-August University of Göttingen, 37075 Göttingen, Germany
- ⁷ Laboratory of Population Genetics, N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences, 119333 Moscow, Russia
- ⁸ Department of Genomics and Bioinformatics, Institute of Fundamental Biology and Biotechnology, Siberian Federal University, 660074 Krasnoyarsk, Russia
- ⁹ Scientific and Methodological Center, G. F. Morozov Voronezh State University of Forestry and Technologies, 394087 Voronezh, Russia
- * Correspondence: konstantin.krutovsky@forst.uni-goettingen.de; Tel.: +49-551-339-3537



Citation: Bondar, E.I.; Feranchuk, S.I.; Miroshnikova, K.A.; Sharov, V.V.; Kuzmin, D.A.; Oreshkova, N.V.; Krutovsky, K.V. Annotation of Siberian Larch (*Larix sibirica* Ledeb.) Nuclear Genome—One of the Most Cold-Resistant Tree Species in the Only Deciduous GENUS in *Pinaceae*. *Plants* **2022**, *11*, 2062. <https://doi.org/10.3390/plants11152062>

Academic Editors: Noe Fernandez-Pozo and M. Gonzalo Claros

Received: 24 June 2022

Accepted: 26 July 2022

Published: 6 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: The recent release of the nuclear, chloroplast and mitochondrial genome assemblies of Siberian larch (*Larix sibirica* Ledeb.), one of the most cold-resistant tree species in the only deciduous genus of Pinaceae, with seasonal senescence and a rot-resistant valuable timber widely used in construction, greatly contributed to the development of genomic resources for the larch genus. Here, we present an extensive repeatome analysis and the first annotation of the draft nuclear Siberian larch genome assembly. About 66% of the larch genome consists of highly repetitive elements (REs), with the likely wave of retrotransposons insertions into the larch genome estimated to occur 4–5 MYA. In total, 39,370 gene models were predicted, with 87% of them having homology to the *Arabidopsis*-annotated proteins and 78% having at least one GO term assignment. The current state of the genome annotations allows for the exploration of the gymnosperm and angiosperm species for relative gene abundance in different functional categories. Comparative analysis of functional gene categories across different angiosperm and gymnosperm species finds that the Siberian larch genome has an overabundance of genes associated with programmed cell death (PCD), autophagy, stress hormone biosynthesis and regulatory pathways; genes that may play important roles in seasonal senescence and stress response to extreme cold in larch. Despite being incomplete, the draft assemblies and annotations of the conifer genomes are at a point of development where they now represent a valuable source for further genomic, genetic and population studies.

Keywords: angiosperms; annotation; conifer; deciduous; genome; gymnosperms; microsatellites; RNA-seq; repeats; seasonal senescence; Siberian larch; transcriptome; transposons

1. Introduction

Gymnosperms originated approximately 360 million years ago (MYA), when they comprised a prevailing part of the terrestrial vegetation on the earth [1–3]. Today's living

gymnosperms comprise about 1000 species [1], with conifers being the most diverse and abundant group. Being one of the most ancient groups of seed plants, they are considered as a link between angiosperms and pteridosperms. The conifer genomes have a number of features that distinguish them from other plants, the most notable is their enormous genome size, which is not a result of recent polyploidization. It varies among the sequenced species from 4 Gbp in *Gnetum montanum* [4] to 31 Gbp in sugar pine, *Pinus lambertiana* Dougl. [5], much larger than compared to the typical diploid angiosperms, such as 135 Mbp in dicot *Arabidopsis thaliana* [6] or 3.1 Gbp in dicot sunflower *Helianthus annuus* [7], but comparable with some polyploid angiosperms, such as 14.5 Gbp in allohexaploid common wheat *Triticum aestivum* [8] or even smaller, if compared with 150 Gbp in octoploid monocot *Paris japonica* [9]. The gene number also seems to vary in the sequenced conifers, as the number of predicted gene models ranges from 39,370 in Siberian larch, *Larix sibirica* Ledeb. (this study) and 50,172 in loblolly pine, *Pinus taeda* L. [10,11], to 102,915 in white spruce, *Picea glauca* (Moench) Voss [12].

The difference in the conifer genome size has not been shown to be associated with recent polyploidization or a whole genome duplication event [13], however there is a higher gene copy number in gymnosperms than in most angiosperm species, which may be associated with transposed and dispersed duplication events [14]. Another characteristic of conifer genomes is a large proportion of repetitive DNA, estimated as much as 70–82% [10,15–17] of the genome size. It is assumed that insertion and extensive proliferation of transposable elements (TE) were mainly responsible for such enlarged genomes [18]. These factors, the giant genome size and its high complexity due to the prevalence of repeat content, make studies of conifer genomes more challenging than in many other plant species.

Owing to the rapidly developing high-throughput sequencing technologies, eleven conifer species in the Pinaceae family have been sequenced, and their draft genomes have been made available to the community, including those for Norway spruce, *Picea abies* (L.) Karst [19]; white spruce, *P. glauca* [12]; loblolly pine, *Pinus taeda* [10,11]; sugar pine, *P. lambertiana* [5]; Douglas-fir, *Pseudotsuga menziesii* (Mirb.) Franco [16]; European silver fir, *Abies alba* [20]; Siberian larch, *Larix sibirica* Ledeb [21]; Japanese larch, *L. kaempferi* (Lamb.) Carr. [22]; Chinese pine, *P. tabuliformis* Carr. [23]; Engelmann spruce, *P. engelmannii* Parry ex Engelm. (NCBI BioProject PRJNA504036); and Sitka spruce, *P. sitchensis* (Bong.) Carr. (NCBI BioProject PRJNA304257).

The Siberian larch is a cold-resistant deciduous conifer tree native to the east and northeast of European Russia, the Urals, and Western and Eastern Siberia [24]. It forms extensive conifer forests, often growing together with Scots pine, Siberian spruce, and Siberian stone pine, occupying almost 263 million hectares, or about 40%, of Russia's forested areas. The Siberian larch is known for its frost-hardiness, relatively fast growth, and its rot-resistant timber, which makes it especially valuable in construction. Its ecological and economical importance has stimulated exploration of its population structure [25–27] and the development of early genetic markers [28,29]. The whole-genome sequencing made possible the development of additional highly informative species-specific SSR markers in *L. sibirica* [30,31], which can be used in different practical applications, including tracking the timber origin to fight illegal logging [32]. The release of the first nuclear [21], chloroplast [33] and mitochondrial [34] genome assemblies for Siberian larch, and recently for Japanese larch [22], has contributed to the development of the genomic resource for the larch genus. Here, we present an initial annotation for the draft Siberian larch genome assembly.

2. Results

2.1. Transcriptome Assembly

To provide RNA-seq support for gene prediction, five tissues (buds, needles, cambium, seedling, and the first-year shoot) were sampled from a reference Siberian larch tree and used to construct a reference transcriptome. The total RNA was sequenced, using the Illumina MiSeq platform (Illumina, San Diego, CA, USA). All of the clean reads, trimmed

with Trimmomatic (9-bp headcrop, minimum read quality of $Q = 23$, and minimum read length of 35 bp; [35]), were used for the transcriptome assembly. A total of 46,618; 626,542; 59,317; 174,805; and 590,240 transcripts were obtained for the buds, cambium, needles, seedling, and the first-year shoot tissues, respectively, using the TrinityRnaSeq package [36]. The N50 length and average read length of the assembled sequences were 357–790 bp. The reads for all of the tissues are deposited at NCBI SRA under accession numbers SRX9464971, SRX14986114, SRX14997110, SRX14997111 and SRX14997112, respectively. The transcriptome assemblies are available under accession numbers GIXH00000000, GJYD00000000, GJYL00000000, GJYN00000000, and GJYW00000000.

2.2. Repeat Content

The species-specific de novo repeat library, generated using RepeatModeler [37], contained 1721 mobile elements that were found in the current assembly of the Siberian larch. Assembling the consensus sequences of ~21 million clusters (cluster size threshold of 200 reads per cluster) with Inchworm from TrinityRnaSeq package [36] resulted in ~31,000 consensus sequences that likely represent the repeated regions of the Siberian larch genome. To validate these sequences, we compared them to the RepeatModeler-derived library and to a PIER repeat library. Homologs were found for ~12,000 consensus sequences among the RepeatModeler-derived library, and for ~7000 sequences among the PIER database. Reciprocal BLAST showed that 1045 out of the 1721 RepeatModeler-derived sequences had a close homology to the clustering-derived consensus sequences. The separate species-specific RepeatModeler-derived library, as well as the combined custom repeat library used for TE identification, are deposited in figshare with DOI 10.6084/m9.figshare.19785913 or can be also found at <https://hpccloud.sfu-kras.ru/owncloud/index.php/s/GMBabOGEgqOD4JX> (accessed on 12 July 2022).

The proportion of the classified families observed in the Siberian larch genome was similar to those previously described for other conifers. The total number of repetitive elements (REs) in the genome assembly, identified using the RepeatMasker [38] with combined repeat library, was 20.9 million with the total size of 4.8 Gbp, which comprises about 40% of the 12 Gbp genome (Table S1, Supplementary Materials). The fraction covered by repeats in the portion of the Oxford Nanopore long reads was 65.98%, as estimated by RepeatMasker. The rough estimation of the repeat coverage, without considering overlaps, and the nested structure of some repeats was 83.8% (Table S2, Supplementary Materials).

The use of the TEclass allowed for a better reconstruction of the TE groups and families composing a large portion of the genome. Among the classified mobile elements, Class I retrotransposons LINE, I, Gypsy, and Copia superfamilies were the most abundant, with LINE elements also having the longest average size and taking the largest part of the genome (Figure 1A; Table S1, Supplementary Materials).

Class I Long terminal retrotransposons (LTR), presented mostly by the Gypsy and Copia elements, comprised the largest fraction of all of the mobile elements. Substantial portions of the LTRs were homologues to a loblolly pine bacterial artificial chromosome (BAC) library and fosmid sequences [39,40]. PtTalladega (3646 copies in the Siberian larch genome), PtOuachita (1025), IFG (990), PtAppalachian (773), PtConagree (731), and eight more repeat families were identified (Table S3, Supplementary Materials). However, most of the LTR-retrotransposons have not been classified into specific families (“Unclassified LTR” in Table S1, Supplementary Materials). Among the non-LTR retrotransposons LINE/L1, I, Penelope, and SINE together comprise about 97.84% of all of the non-LTRs, which cover 11.98% of the Siberian larch assembly length. The majority of the repeats among the different repeat families was relatively small in length, less than 1 Kbp. A small part of the longest repeats reached almost 15 Kbp; they belonged to the LINE elements and uncharacterized LTR. The most frequent TEs for each family were shorter than 1 Kbp. Some of the repeat groups have a bimodal length distribution (Gypsy, DIR, LINE/I, Helitron, Penelope), but both of the peaks in the distributions were less than 1 Kbp (Figures S1–S5, Supplementary Materials).

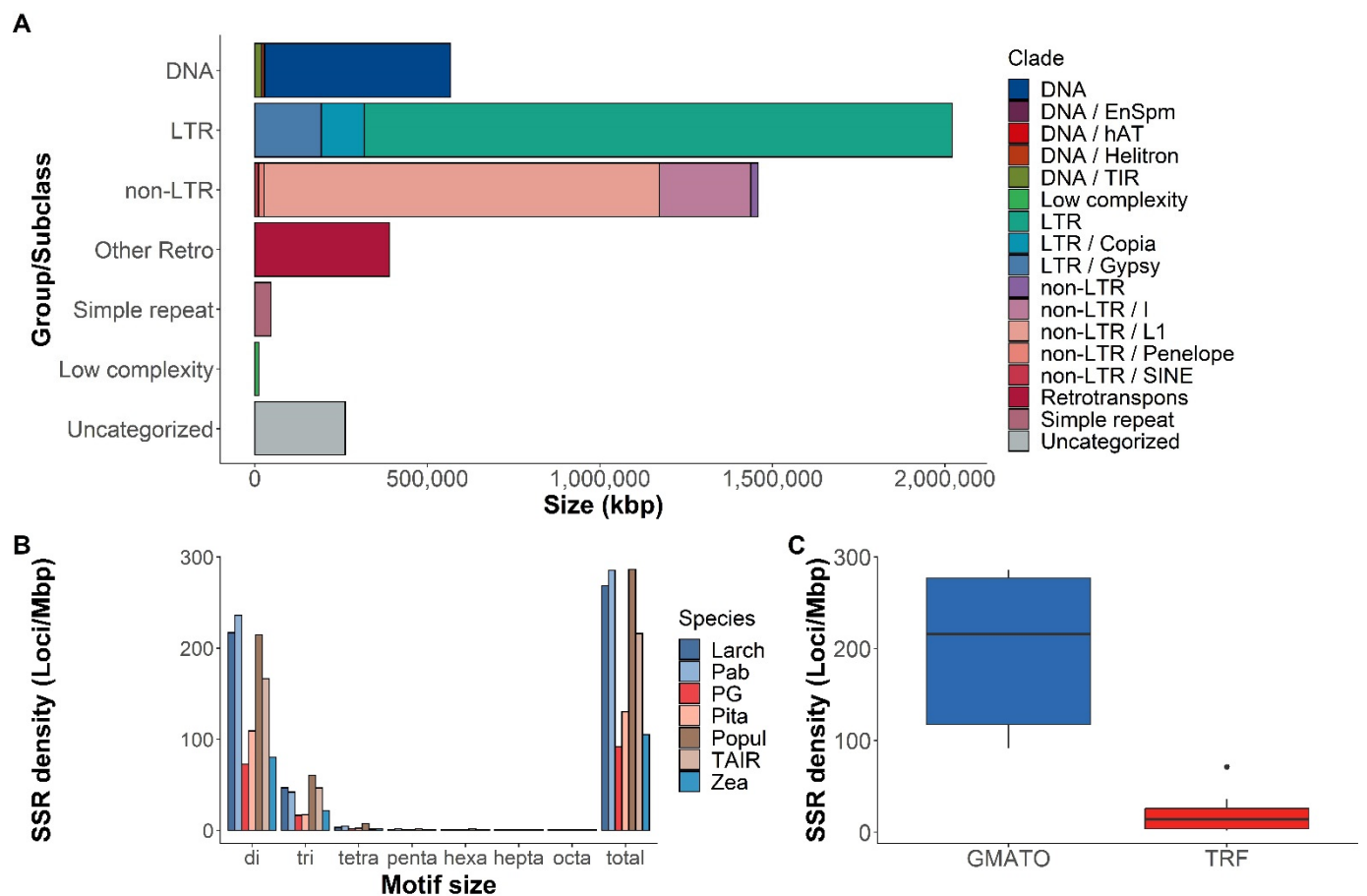


Figure 1. (A)—relative size of the repetitive sequence content of the Siberian larch genome annotated using RepeatMasker and combined library, comprising the RepeatModeler-derived library classified with TEclass, RepBase, MIPS, CPRD and PIER v1.0 libraries; (B)—microsatellite (SSR) density (number of microsatellite loci with di-, tri-, tetra-, penta-, hexa-, hepta- and octanucleotide motifs per 1 Mbp) for several conifer and angiosperm species found using the GMATo program (Larch—*Larix sibirica*; Pab—*Picea abies*; PG—*Picea glauca*; Pita—*Pinus taeda*; Popul—*Populus trichocarpa*; TAIR—*Arabidopsis thaliana*; Zea—*Zea mays*; (C)—box plots for number of all microsatellite loci found in all species listed in B using the GMATo and TRF programs.

Class II DNA transposons cover 4.76% of the assembly size, and 4.49% of them were not classified by TEclass (“Unclassified” in Table S1, Supplementary Materials). Among classified transposons the most numerous were terminal, inverted repeats (TIR, 0.16% of DNA transposons), Helitron (0.06%), EnSpm (0.02%), hAT (<0.01%).

In total, 1,129,244 microsatellite loci with motif size 2–8 bp were detected by the GMATo program [41] in the Siberian larch genome, with an average density of 268.7 loci per megabase. Compared to other species, the larch genome assembly also had a relatively high SSR density, similar to the Norway spruce and black cottonwood genomes (Figure 1B). Wegrzyn et al. (2014) and Neale et al. (2014) reported a SSR density of 10–20 loci/Mbp for *Pinus taeda*, *Picea abies*, and *Picea glauca*, discovered by the TRF program [42]. We also scanned the larch genome with TRF, which yielded 17,145 loci with the same motif size and with overall density of 4.1 loci per megabase. On average, GMATo discovered ninefold more SSR loci than TRF, based on seven plant species (Figure 1C; mean 197 and 21 loci/Mbp for GMATo and TRF, respectively), and has proved to be more efficient for the processing of the large genome sequences.

2.3. LTR-RT Insertion Time Estimate

The LTRharvest [43] with LTR_retriever [44] identified 347 LTR elements and 36 intact LTRs in the Siberian larch draft assembly. These 36 intact LTRs were combined with 367 identified by Zhou et al. (2021). A possible overlap was checked, using blastn against *Larix* LTRs from Zhou et al. (2021). The probable insertion wave of retrotransposons into the larch genome likely occurred 4–5 MYA, as estimated based on 403 LTRs (Figure 2C). Although the Copia (PR-INT-RT) and Gypsy (PR-RT-INT) superfamilies had slightly different profiles, their mean and median values were very close (mean = 3.16 MYA and median = 3.03 MYA for Copia; mean = 3.11 MYA and median = 2.96 MYA for Gypsy) (Figure 2D). The LTRs with different flanking motifs, typical 5'-TG ... CA-3' and other fewer common variants were also compared in terms of their time insertion. Similarly, the LTRs with TG ... CA flanking motifs had a slight peak with a median at 2.56 MYA and LTRs with other flanking motifs had a median at 2.60 MYA (Figure 2E). When compared to the profiles of the other gymnosperms, the larch exhibited the most ancient burst of LTRs insertion, even compared to *Gnetum* and *Ginkgo* (Figure 2B).

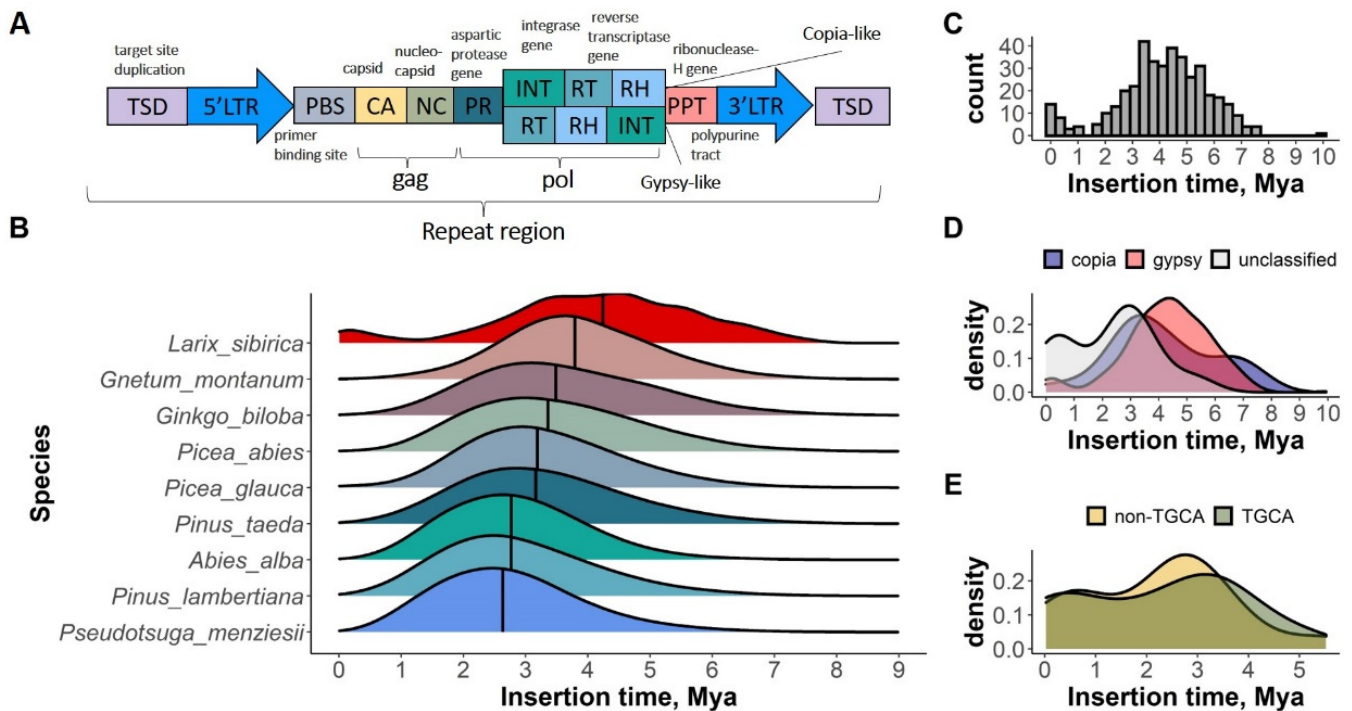


Figure 2. (A)—structure of Copia-like and Gypsy-like LTR retrotransposons; (B)—estimation of the insertion time of the LTR-RT elements in genomes of nine gymnosperm species; estimation of the insertion time of LTR-RT (C); Copia and Gypsy superfamilies (D); and TGCA/non-TGCA LTRs (E) in the genome of Siberian larch. X-axis is in million years (MYA).

2.4. Identification of LRR Genes

In all of the tissues of the Siberian larch, 4482 transcripts containing the LRR domain were detected, using the hidden Markov Model method (HMM) HMMER3 [45] to correctly assign the homologous sequences to one or more Pfam families of LRR. The largest number of LRR domains were contained in the shoot transcriptome (1846), slightly less were in the cambium transcriptome (1599), but their proportions in the total number of transcripts for each tissue (presented in Table S4, Supplementary Materials) were approximately the same for all of the tissues and no more than 2%. The LRR-1, LRR-4, LRR-8, and LRR-6 families encompassed the largest portion of the putative LRR domains identified in the larch transcriptomes (Figure S6A–E, Supplementary Materials). As can be seen in Figure S6A–E (Supplementary Materials), the largest number of transcripts contained the LRR-4 family in all of the tissues of Siberian larch.

The NBS-LRR proteins play an important role in plant defense responses against various classes of pathogens, including bacterial, fungal, viral, nematodes, and insects. Their length usually ranges from ~860 to ~1900 amino acids, but most of the transcripts containing the LRR domain were shorter than 300–400 amino acids (Figure S6F, Supplementary Materials). We filtered out the sequences shorter than 850 and searched for the NBS domain. In total, 56 putative NBS-LRR proteins were found in the transcriptome of the shoot, 18 in the cambium, 5 in the seedling, and 2 in the autumn bud (Table S4, Supplementary Materials). The OmicsBox [46,47] functional annotations confirmed the presence of the domains NB-ARC and LRRs in the identified transcript sequences. The functional annotation by InterProScan [48] did not reveal the presence of other functional domains in these sequences. The NB-ARC- and LRRs-containing sequences in Siberian larch are likely to be resistance genes, because they include P-loop NTPase and LRRs families. The sequences containing identified LRR and NB-ARC domains are deposited in figshare with DOI 10.6084/m9.figshare.19785913 or can be found at repository <https://hpccloud.sfu-kras.ru/owncloud/index.php/s/GMBabOGEgqOD4JX> (accessed on 12 July 2022).

2.5. Structural Annotation Using MAKER2

The benchmarking with the BUSCO package [49] found 317 complete and 307 fragmented genes out of 1614 single-copy orthologs. This makes it possible to estimate the number of complete (not fragmented) genes in this annotation at the level of 32,000 genes, with gene space completeness estimated at 38.6% (Tables S5 and S6, Supplementary Materials). The high fragmentation of the scaffolds could explain the relatively high proportion of the fragmented genes in the Siberian larch genome (19% fragmented vs. 38.6% total) identified in the BUSCO assessment with respect to other conifer genomes (7.5% vs. 80.9% for *Pinus lambertiana*; 11.5% vs. 32.6% for *Picea glauca*).

Using the transcripts from several tissue types, transcriptome shotgun assemblies (TSAs) from other conifer species, and proteins' references from Uniprot as a starting point for the MAKER2 annotation pipeline [50] allowed us to obtain 39,370 gene models in 37,206 scaffolds, composed of 134,271 exons, and 94,901 introns (Table 1; Table S7, Supplementary Materials). Among them, 24,551 gene models were full-length, and 14,819 were partial (6476 truncated from the beginning, 7545 truncated from the end, and 798 truncated from both sides). The mean length of the genes was about 1841 bp containing 3.41 exons on average, with two being the most common number of exons, which is in a good agreement with prediction of ~four exons per gene for *Pinus taeda* [51]. The maximum CDS length was 7216 bp, which is less than the length of the longest intron of 10,153 bp (Table 1).

Table 1. Summary of genome assembly and gene annotation statistics for Siberian larch genome.

Parameter	<i>Larix sibirica</i>
Number of chromosomes	12
Estimated genome size (1C), Gbp	12.03 (12.30 pg) ¹
Assembly length, Gbp	5.59 ² /12.34 ³
Assembly N50, bp	3098 ² /6443 ³
GC content, %	35.41
Repeat content, %	65.98
Number of predicted gene models	39,370
Number of full-length gene models	24,551
Average CDS length, bp	244.29
Average intron length, bp	360.93
Longest intron length, bp	10,153

¹ <https://cvalues.science.kew.org/search/gymnosperm> (accessed on 25 July 2022); ² based on contigs without gaps; ³ based on scaffolds with gaps.

The mapping of the transcriptome reads to the genome resulted in 77.9% to 88.8% overall alignment rate per tissue (Table S5, Supplementary Materials). This suggests that the portion of the complete and partial transcriptome-derived genes identified in the genome can be estimated from 21% in needle tissue to 80.4% in shoot tissue.

MAKER2 uses annotation edit distance (AED), a quality control metric initially introduced in the Sequence Ontology project [52,53], where it was originally used to compare and score different releases of the same annotation. Here, instead of assessing the distance between annotations, it measures the congruency between a gene model and its corresponding evidence [50]. For the Siberian larch annotation, the AED computed by MAKER2 was below 0.5 for 95% of the gene models, which is comparable to the mouse genome release GRCm37 and maize chromosome 4 [50]. However, considering the scarce amount of species-specific supporting data that could be used as evidence in the gene prediction and for quality control, this score could be overestimated to some extent.

For the regions identified by RepeatMasker as repeats, intersections with the CDS from the predicted gene models were also found. In total, 6884 genes had at least 20% overlap with a repeat (Figure S7A, Supplementary Materials). Those gene models were consequently marked as ‘repeat associated’; 2247 (33%) of them were overlapping with the Non-LTR I family, 241 (3%) with LINE, 571 (8%) with LTR Gypsy, 523 (8%) with Copia, and 312 (5%) with Simple repeats. The most frequent functional annotations for the repeat-overlapping genes were receptor-like protein kinases, leucine-rich repeat (LRR) proteins, transcription factors, ATP-binding cassette transporters (ABC transporters), synthase, reductase, esterase and peroxidase enzymes, Cytochrome C and Cytochrome P450 proteins, and others (Figure S7B, Supplementary Materials).

Similar to larger genome sizes, the average intron lengths were also longer in conifers than in angiosperms [54]. In the MAKER2-derived annotation, 94,901 introns were identified in the 36,183 genes in total, with an average length of 361 bp and the longest intron of 10,153 bp, which is less than in other conifer species; 289 introns were longer than five Kbp. When comparing the top 10% of the longest introns, the larch introns were comparable in length with those of *A. thaliana* and *P. taeda*, although, the longest larch introns were far shorter than those in other spruce species, such as *P. abies* and *P. glauca*, or in the repeat-rich genomes of *Populus thichocarpa*, *Vitis vinifera*, and *Zea mays* (Figure 3B). In total, the introns made up to 47% (34,25 Mbp) of the gene space (Figure 3A), and 0.29% of the 12.3 Gbp genome assembly. The repeat content of the introns was lower than in the genome in general; for instance, only 4.59 Mbp (12.9% of the intron sequence space) are covered by TEs compared to 4800 Mbp in the genome (65.98% of genome sequence space excluding introns). In the larch introns, the most abundant were Class I retrotransposons LINE and I (6135 and 2195 elements, respectively), followed by LTR Gypsy and Copia (1879 and 1214 elements, respectively). Among the Class II DNA transposons, the most frequent were the TIRs and EnSpm elements (362 and 245 elements, respectively).

2.6. Functional Annotation

Based on the sequence similarity to the *Arabidopsis thaliana* protein set, 87% of the predicted larch gene models (34,358 out of 39,370) had an alignment with at least 10^{-5} e-value, 20% query coverage, and 20% identity (Figure 4A). The proportion of the mapped proteins in Siberian larch was among the highest compared to other gymnosperms (second only to *P. tabuliformis*), while being lower than in some model angiosperm plants, such as black cottonwood, grape, and common oak (Figure 4B).

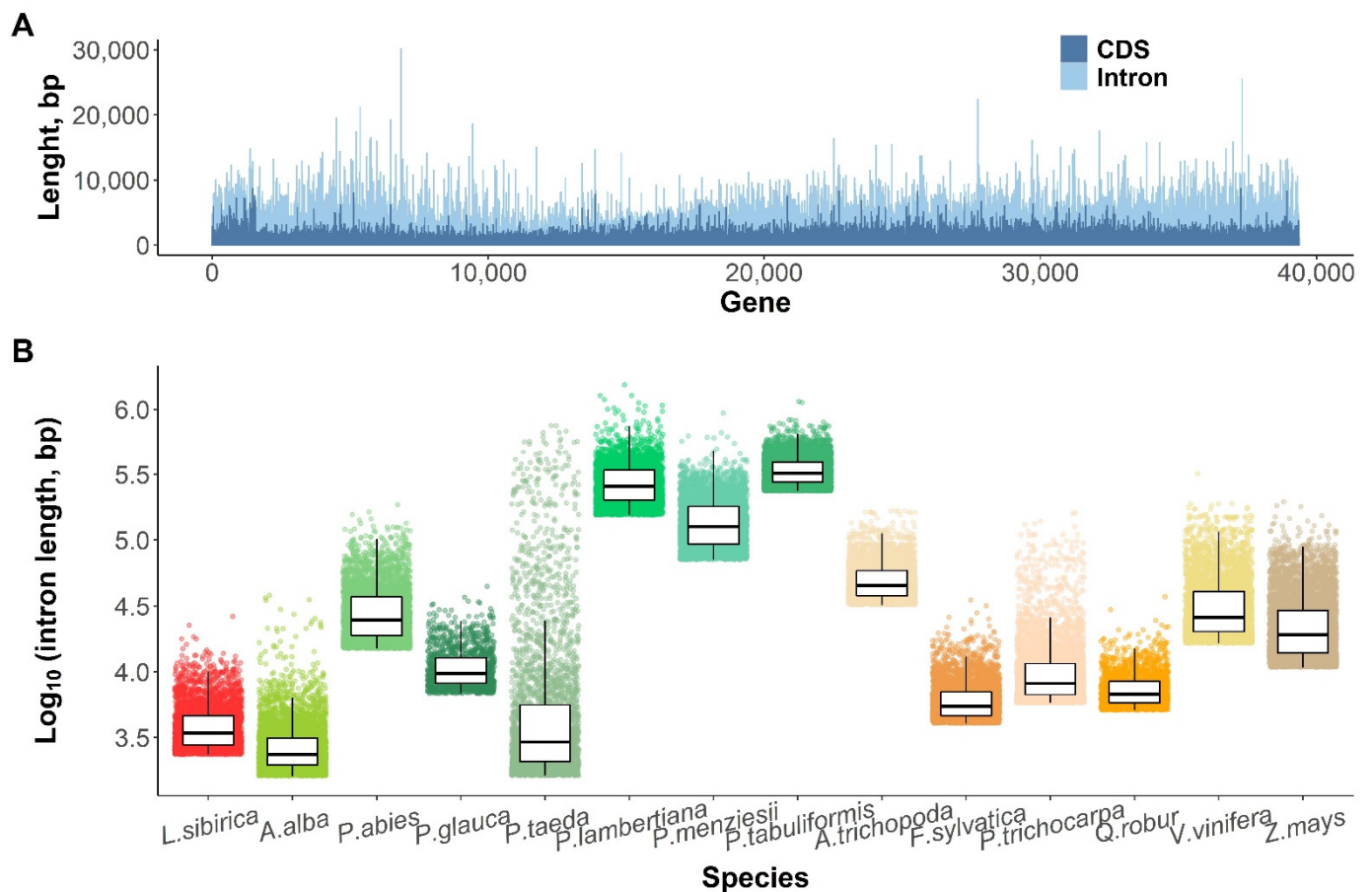


Figure 3. (A)—proportion of coding and intronic parts per every gene model in the Siberian larch genome according to the MAKER2 annotation; (B)—top 10% of the longest introns across 11 plant species.

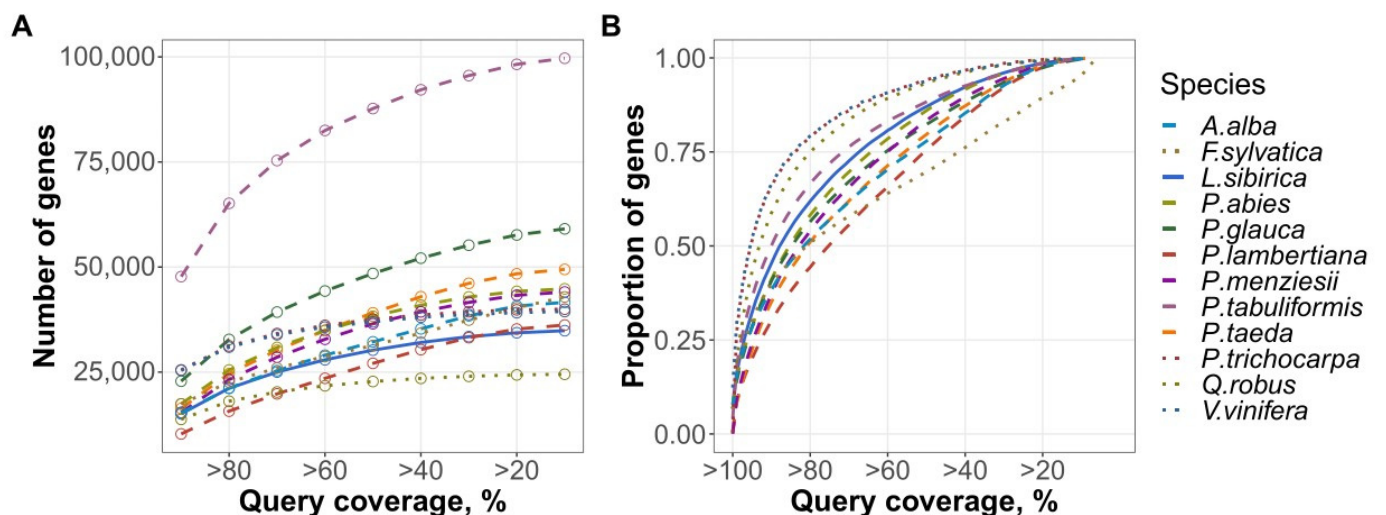


Figure 4. Cumulative number (A) and proportion (B) of genes aligned to the *Arabidopsis* protein set using *qcovhsp* above a given coverage threshold. Gymnosperm species are presented by dashed lines; angiosperms—by dotted lines; Siberian larch—by the solid blue line.

The GO category assignment was based on InterProScan domains identification and BLAST homology search, which yielded 30,512 annotated gene models (78%), with at least one assigned GO term. To analyze the annotated gene pool in more detail, it was

divided into 20 functional categories. The functions were classified according to the most recent GO dictionary: five categories in Biological process, six in Molecular function, five in Cellular component (Figure 5A). All of the proteins from the respective category were mapped to the *Arabidopsis* protein database with BLASTP and $e\text{-value} \leq 10^{-5}$, $\text{pident} > 50$ and $\text{qcovhsp} > 50$. From 50% (in Transcription activity) to 85% (in Transporter activity, Mitochondrion, and Chloroplast) of the Siberian larch annotated proteins were found to be homologs to *Arabidopsis* proteins (Figure 5B).

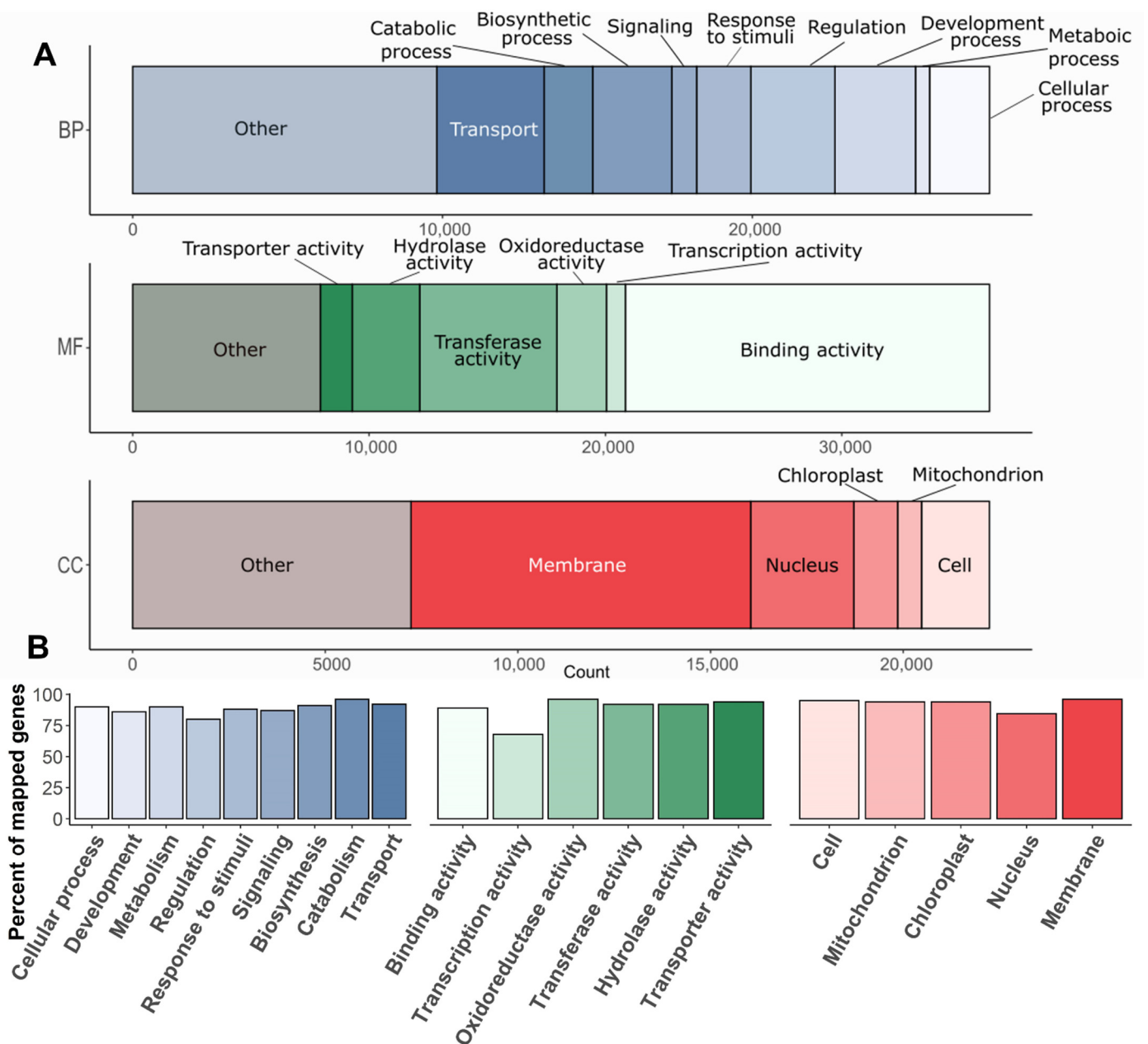


Figure 5. Functional annotation of Siberian larch genes: (A)—proportion of predicted larch genes in three functional categories: BP—biological process; MF—molecular function; and CC—cellular component; (B)—percentage of larch proteins in different functional categories mapped to the *Arabidopsis* non-redundant protein set with a BLASTP match parameters of $e \leq 10^{-5}$, $\text{pident} > 20$ and $\text{qcovhsp} > 20$.

2.7. Comparing GO Annotations between Conifer and Angiosperm Species

Among the 6937 GO terms shared by 11 species compared in this study, 2080 (29.9%) had significant differences in the number of genes annotated with the corresponding GO term (q -value < 0.005 and adj. p -value < 0.01). It was reported recently that the deciduous and evergreen trees differ in the number of genes associated with dormancy and stress response leucine-rich repeats receptor-like kinases (LRR-RLK) [55].

In 12 terms associated with metabolism and the organization of the cell wall and its components, the number of annotated genes was higher in all of the gymnosperms than in the angiosperms (Table S8, Supplementary Materials). For four terms associated with apoptosis and autophagy, the relative gene number in angiosperms and Siberian larch was higher than in the rest of the gymnosperms. Among the 15 terms related to ABA, JA, and ETH metabolism, regulation and response were identified in all 11 species; Siberian larch had the highest number of annotated genes in four of them, compared to both the gymnosperms and angiosperms: jasmonic acid (JA) biosynthetic process (GO:0009695); abscisic acid (ABA)-activated signaling pathway (GO:0009738); ABA binding (GO:0010427); and response to hormone (GO:0009725), respectively. Several of the genes related to the response to JA were generally higher in the gymnosperms than in the angiosperms. In water channel activity (GO:0015250), water transport (GO:0006833), nucleosome (GO:0000786), and nucleosome assembly (GO:0006334), the Siberian larch had a larger number of mapped genes. Among the conifers, the Siberian larch had the highest number of genes in response to light stimulus (GO:0009416) and light harvesting in the photosystem I (GO:0009768) GO categories.

3. Discussion

Work with such large genomes as those found in conifers is often hindered by the limit of computational resources, such as the computation time and memory space needed to process the genomic data. The structural annotation of the draft whole-genome assembly of the Siberian larch with MAKER2 pipeline on a 448 core cluster took 22 days to generate a complete set of predicted gene models, and running RepeatMasker separately on the genome assembly of 40 cores took 20 days. This, and the complex genome structure enriched with numerous repeated regions, still makes genomic studies a challenge for plant species with exceptionally large genomes.

3.1. Repeat Content and LTR Insertion Time Estimate

There are two main factors responsible for a large genome size in higher plants: polyploidy and amplification of TEs. The latter not only contributes to a genome-size expansion, but also presents a source of genetic variation, increasing the mutation rate and affecting the gene expression by altering coding parts and regulatory regions. A characteristic feature of the conifer genomes is a large number of Res, including TEs. The irreversibility of the repeat accumulation process in the genomes of angiosperms and conifers, also called genome obesity, is discussed in the literature [18,19]. The underlying causes for this are under debate. Some researchers consider this to be a result of bursts in the activity of transposable elements [56–60], while others suggest that the large genomes containing many diverse repeats may have acquired them over time by a steady accumulation process, which may also imply that a repeat elimination process could be slower or less efficient, leading to slow genome contraction [18,61,62].

The types of identified repeats and their distribution in the Siberian larch genome are consistent with those found in other conifers. However, the proportion of the genome represented by simple repeats and mobile elements is one of the smallest among all of the gymnosperms; only 40% of the 12 Gbp genome size can be explained by repeat expansion. This estimate is lower than in all of the other gymnosperms. However, the fraction of the genome covered by repeats in the portion of nanopore long reads was 65.98% bp, as estimated by RepeatMasker, which suggests that the part of the Siberian larch repeatome was too fragmented to be included in the final scaffolded assembly.

LTR comprised the largest fraction of all of the mobile elements with prevailing number of the Gypsy TEs, a characteristic also widely observed in the other conifers [15–17,19] and angiosperm species [63]. While the LINEs and SINEs are common for plant genomes, Penelope-like elements (PLEs) were long considered to be a feature of animal and fungi genomes [64,65] until multiple Penelope (EN(+))PLE type or Dryad elements were found in the loblolly pine [66], and AdLINE3 RTEs were found in a number of flowering plant species [67]. A phylogenetic analysis of Dryad and AdLINE3 suggested a horizontal transfer of TEs, possibly between arthropods and a conifers' ancestor approximately 340 Mya [66,67].

Class I retrotransposons proliferate by integrating their RNA intermediate into the host genome via retrotranscription to cDNA, using the host transcription machinery and their own enzymes. The coding part of the repeat, placed between two long terminal repeats, contains *gag* and *pol* genes [68,69]. The latter includes protease, reverse transcriptase, ribonuclease-H, and integrase that are responsible for the cleaving of the Pol protein and RNA (protease and ribonuclease H), copying the retrotransposons RNA into cDNA (reverse transcriptase) and integrating the cDNA into the host genome (integrase), respectively (Figure 2A) [70].

In higher plants, the repeats containing direct LTRs prevail [71,72]. When a retroelement has just been inserted, the two flanking LTRs on its 5' and 3' ends are identical [73], but, with time, they accumulate mutations, and notably their mutation rate is most likely higher since the repeats, unlike the genes, are not under selection. The number of sequence dissimilarities between the two flanking LTRs can be used as a proxy to estimate the time when the element was inserted into the genome. The estimation of the insertion time of the LTR-RT elements can shed light on the evolutionary aspects of the genome organization, and potentially the date expansion events.

The residual DNA of the LTRs was much higher than the intact LTR-RTs, which suggests that, after a massive proliferation of the retrotransposons, a DNA loss might have occurred in the Siberian larch genome. The typical insertion time estimates in the plant genomes range from 1 to 2.5 MYA for the angiosperms [74–79]. In the gymnosperms, 10–15 MYA insertion times have been reported [80]. Based on the LTRs identification by Zhou et al. [81], the approximate time of LTR expansion in the gymnosperms can be estimated to be 2–4 MYA. In the larch, the time estimate can be influenced either by the efficient repeat elimination mechanism combined with the true ancient repeat insertion, or by the fragmented nature of the draft assembly, and consequently, the low number of found LTRs. However, the draft genomes of the Norway spruce and silver fir have comparable assembly contiguity (N50 = 6443, 5206 and 14,051 bp for the Siberian larch, Norway spruce, and silver fir, respectively), and their estimated insertion times are also similar, despite the noticeably different number of identified LTRs (403 in larch, 31,016 in Norway spruce, and 34,952 in silver fir).

Leucine-rich repeats (LRRs) have been found in many functionally diverse proteins. They form horseshoe structures and might be involved in protein–protein interactions. The most representative class of plant-pathogen resistance proteins or R proteins in plants are the NBS-LRR proteins. As their name implies, the NBS-LRR proteins include a nucleotide-binding site (NBS), also called a central nucleotide-binding domain NB-ARC, and a domain containing LRR. The LRR regions demonstrate a high variation in the type and number of the LRR units between and within species, which provides the specificity of pathogen molecules' recognition. The LRR-containing sequences are associated with immune response of plants to biotic stress [82–85]. Their further study will also help us better understand the genetic mechanisms of disease resistance in the larch and other plants.

3.2. Structural Annotation Using AUGUSTUS and MAKER2 Pipeline

The number of predicted gene models in Siberian larch (39,370) is similar to that in the loblolly pine (50,172) and the Douglas fir (54,830), but much lower than in the silver fir (94,205), sugar pine (71,117), white spruce (102,915), Norway spruce (70,968),

and Chinese pine (80,495). Among them, 77% were supported either by RNA-seq data from several tissues, or by homology to the genes from the plant subset of the NCBI *nr* database. The BUSCO analysis suggests that only 39% of the genes had been recovered (Table 1; Table S6, Supplementary Materials). This relatively low percentage can indicate that not all of the genes were fully sequenced and included in the genome assembly. The genome assembly based only on the contigs without gaps covered 5.59 Gbp, about half of the entire genome, while the genome assembly based on scaffolds with gaps covered 12.34 Gbp, which corresponds well to the entire genome. Therefore, we believe that the gaps are not due to the insufficient sequencing coverage, but represent highly repetitive regions that are excluded during assembling procedure, due to their high complexity, while the coding parts of the genome are mostly sequenced and easily assembled, due to their uniqueness. Moreover, comparison to *Arabidopsis* gene set (Figure 5B) shows that a large proportion of the genes have been identified and characterized. Thus, the annotation can still be used as a good resource and a primary reference for further genomic studies.

The average intron length in the larch genome was 1.8 to 3.2 times shorter than in the other conifers. In comparison to the top 10% of longest introns, the larch introns were far shorter than those in other conifer species, such as *P. abies*, *P. glauca*, and *P. taeda* (Figure 3B). The discrepancy could be explained by (1) underestimation of the intron size in the MAKER pipeline [12] that uses the threshold values for resolving the exon–intron structure, or (2) the naturally occurring differences within the conifer clade. The assembly contiguities for *L. sibirica* and *P. abies* are close, as evidenced by their N50 (Table S7, Supplementary Materials; Figure 3B), while their average and maximum intron length differ by 2.8 and 6.7 fold, respectively. Likewise, the average intron size in *L. sibirica* and *P. glauca* were close, differing by 1.8 fold (Table S7, Supplementary Materials), while their N50s differ by 7.2 fold. Nevertheless, it is still possible that the variation in intron sizes in the conifer genome can be explained by a difference in the assembly contiguity.

3.3. Functional Annotation

The conifers differ from flowering woody plants in a number of ways, including the absence of vessels in xylem and sieve tubes and companion cells in phloem, different wood structure, haploid megagametophyte (unlike triploid endosperm in angiosperms), megasporophylls, and reproductive structures presented by cones rather than flowers. The conifers, unlike angiosperms, are mostly evergreen plants with very few of them having a seasonal needle abscission (seasonal senescence), particularly in *Glyptostrobus*, *Metasequoia*, *Taxodium*, *Pseudolarix*, and *Larix* genera. The larches are the dominant species in the boreal forests in Western Siberia and Canada, occupying much colder habitats than the other boreal woody plants.

3.3.1. Cell Wall and Phenylalanine Metabolism

The conifers and flowering woody plants differ in their cell wall structure, which influences their woody properties. In woody plants, the common components of a cell wall are hemicellulose, pectin (in the primary cell wall), and lignin (in the secondary cell wall) [86]. In the woody angiosperms, lignin in the secondary cell walls is made up of guaiacyl (G) and syringyl (S) units, while in the gymnosperms, the lignins are homogeneous, consisting primarily of p-hydroxyphenyl (H) and guaiacyl units [87]. It was proposed that the syringyl units have better ability to strengthen cell walls than the guaiacyl units and are advantageous for anti-fungal defenses [88,89]. The primer building blocks for lignin in conifers are p-coumaryl alcohol and coniferyl alcohol [90].

The differences between the conifers and angiosperm woody species can be clearly seen in the number of genes in GO terms related to cell wall organization and lignin catabolism (Figure 6). The cell-wall enzymes involved in the biosynthesis of lignin from p-coumaryl alcohol and coniferyl alcohol were identified in all six of the analyzed conifer species.

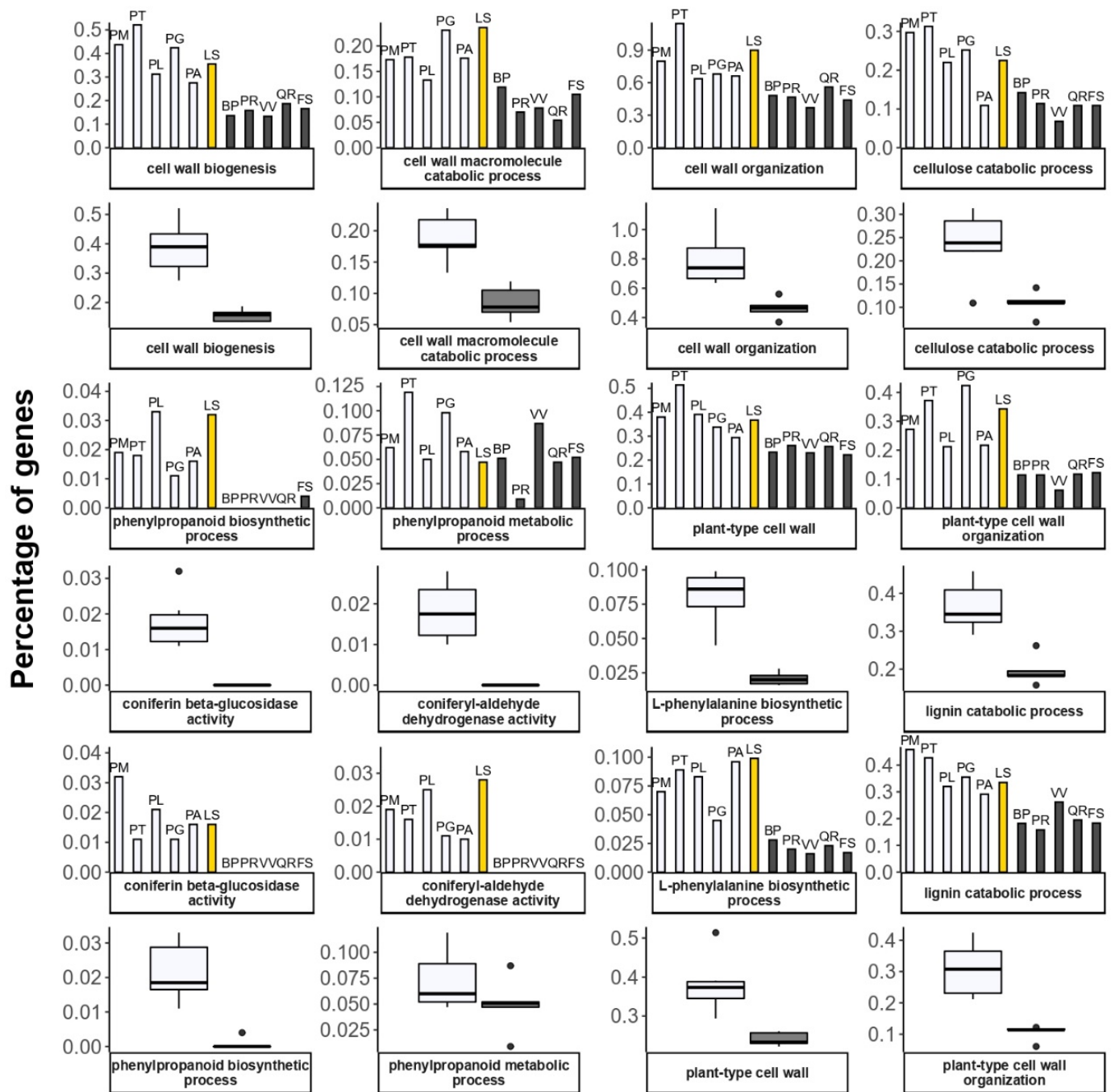


Figure 6. Percentage of genes annotated with GO terms related to cell-wall maintenance. Angiosperm species are represented by solid black columns, gymnosperms by transparent columns, Siberian larch—by yellow. Boxplots demonstrate the difference in gene number between two groups, evergreen (transparent) and deciduous (gray). Angiosperms: BP—*Betula pendula*; FS—*Fagus sylvatica*; PR—*Populus trichocarpa*; QR—*Quercus robur*; VV—*Vitis vinifera*). Gymnosperms: PM—*Pseudotsuga menziesii*; PT—*Pinus taeda*; PL—*Pinus lambertiana*; PG—*Picea glauca*; PA—*Picea abies*; LS—*Larix sibirica*.

One of the major precursors for lignin biosynthesis and production of secondary metabolites, such as phenylpropanoids, flavonoids and anthocyanins, is phenylalanine. Conifers utilize large amounts (30–40%) of carbon fixed during the photosynthesis for lignin biosynthesis and wood formation, and their secondary metabolism is remarkably complex [91]. Apart from their building function, phenylpropanoid, in the form of pheno-

lics, terpenoids, and alkaloids, plays an important role in the defense mechanisms against insect and microbial pathogens, functioning as antifeedants and toxins [91–93].

The biosynthesis of phenylalanine and phenylpropanoids also demonstrates the difference between the conifers and angiosperms (Figure 6). In plants, the main enzymes participating in the biosynthesis of phenylalanine are prephenate-aminotransferase (PAT), that converts prephenate to arogenate, and arogenate dehydratase (ADT) that transforms arogenate to phenylalanine, that is lastly converted by phenylalanine ammonia-lyase (PAL) into cinnamic acid in cytosol, the first component of the phenylpropanoid pathway [91,94]. It was shown previously that the conifers have more diverse families of ADT and PAT genes, compared to angiosperms [95,96].

3.3.2. Programmed Cell Death and Autophagy

Programmed cell death (PCD, or apoptosis in animals) is an organized and genetically regulated process of cellular suicide that can be induced by external environmental stresses or occurs during the organism's development. Unlike animals, plant cells have a rigid cell wall that prevents the formation of apoptotic bodies [97], and lack the classical caspases that act as the main inducers of PCD, and phagocytosis or macrophages that could eliminate the remains of a dead cell. Instead, plant cells use caspase-like proteases (metacaspases) to induce PCD [98], and utilize vacuoles and vacuolar-lytic enzymes to digest their cell contents [99,100]. The correct classification of the plant PCD types has been debated [97,101–103]. In general, it can be divided into two major types: (1) vacuolar/autolytic/apoptosis-like PCD, characterized by the engulfment of the cytoplasm by lytic vacuoles and the later release of vacuolar hydrolases into the cytosol due to the rupture of the tonoplast; (2) a hypersensitive response PCD that often does not involve swelling of the vacuoles and is characterized by cell shrinkage and increased autophagic activity [102]. The former commonly occurs during the differentiation of the xylem elements, leaf senescence, and megasporogenesis, while the latter is activated in response to pathogen invasion to prevent the further spread of the infection [101]. PCD and autophagy have been shown to be closely related to senescence, since it relies on the degradation and recycling of accumulated nutrients for later use in other parts of organs. Important features of PCD, such as DNA fragmentation and protoplast retraction, were observed during senescence in cucumber [104,105].

In the GO:0012501 category associated with PCD, the Siberian larch has gene numbers more similar to those in deciduous angiosperm trees, rather than in evergreen conifers. The number of genes associated with autophagy is also higher in the deciduous angiosperms and larch than in other conifers (Figure 7). However, the GO:0012502 (induction of PCD) genes were annotated only for the conifer species (20 in Douglas-fir, 6 in Loblolly pine, 2 in Sugar pine, 42 in White spruce, 22 in Norway spruce, and 12 in Siberian larch).

PCD features, including the altered nuclear morphology and DNA fragmentation, were also observed in tomato leaves and flowers, where treatment with ethylene (ETH) triggered PCD in abscission zone cells, and the application of ROS scavengers delayed abscission [106]. It was proposed that PCD is regulated through the ETH signaling pathway, as ETH-biosynthesis defective mutants exhibit increased leaf longevity [107]. Activation of autophagy during PCD and senescence were demonstrated on autophagy-deficient *Arabidopsis* mutants that demonstrated accelerated senescence and PCD [108,109].

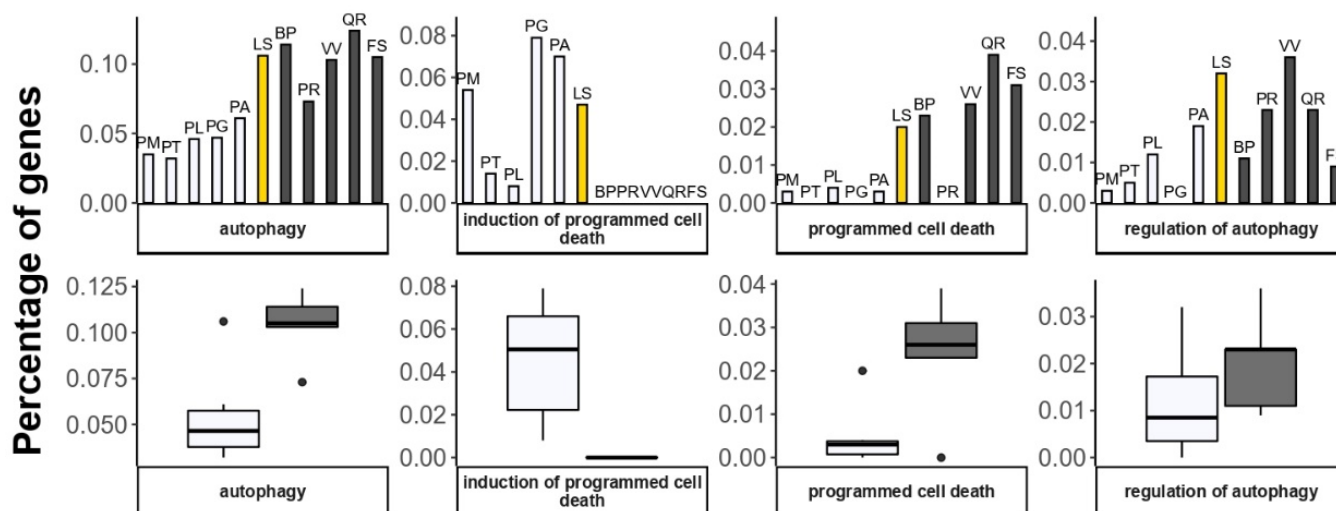


Figure 7. Percentage of genes annotated with GO terms related to programmed cell death (PCD) and autophagy. Deciduous angiosperm species are represented by black solid columns, evergreen gymnosperms—by transparent columns, Siberian larch—by yellow column. Boxplots demonstrate the difference in gene numbers between two groups, evergreen (transparent) and deciduous (gray). Angiosperms: BP—*Betula pendula*; FS—*Fagus sylvatica*; PR—*Populus trichocarpa*; QR—*Quercus robur*; VV—*Vitis vinifera*). Gymnosperms: PM—*Pseudotsuga menziesii*; PT—*Pinus taeda*; PL—*Pinus lambertiana*; PG—*Picea glauca*; PA—*Picea abies*; LS—*Larix sibirica*.

3.3.3. Hormones

ABA, JA, and ETH are phytohormones that play an important role in response to abiotic and biotic stress and leaf senescence. ABA is a phytohormone that participates in a number of processes critical for plant development and growth, such as the control of bud dormancy and seed germination, fruit development, resilience to abiotic stress and pathogens' infection, and leaf senescence. ABA induces stomata closure, thus reducing water loss via transpiration in response to water deficiency or heat stress. ABA-deficient mutants of *Arabidopsis*, tobacco, tomato, and maize suffer even from relatively moderate dehydration or temperature deviations [110]. ABA is also required for cold-stress tolerance [111], as ABA-deficient mutants of *Arabidopsis* were shown to have reduced freezing tolerance in cold-acclimated plants [112–114]. Exposure to low temperatures was shown to increase the levels of endogenous ABA in *Arabidopsis* and wheat. In many plants, ABA is shown to be involved in leaf senescence [115], and together with ETH and reactive oxygen species plays a major role in leaf abscission [116]. In rice and *Arabidopsis* treatment with exogenous ABA was demonstrated to accelerate leaf yellowing and senescence [117,118], and levels of endogenous ABA were reported to be increasing during leaf senescence in maize and *Arabidopsis* [119,120].

JA and its derivatives, referred to as jasmonates (JAs), are fatty acids belonging to the oxylipins family that function as signaling molecules, regulating the expression of genes in response to various abiotic stresses. The biosynthesis of JAs takes place consecutively in plastid, peroxisome, and cytosol, where it is converted from its precursor, α -linolenic acid, through oxo-phytodienoic acid to JA [121]. The process is triggered by abiotic stress that causes an accumulation of JAs in the cytoplasm of stressed leaves [122,123], and activates the JA-signaling pathways. The higher levels of JAs activate the binding of various transcription factors (TFs) to specific jasmonate-responsive genes that otherwise are silenced by the transcriptional repression complex. This complex includes jasmonate ZIM-domain (JAZ) proteins, transcriptional corepressor TOPLESS (TPL), and the novel interactor of JAZ protein (NINJA) [122]. JAs alleviate the effects of water deficiency and

soil salinity [124–126], low temperature [127,128], excessive UV exposure [129–131], and participate in the pathogen-defense mechanisms in gymnosperms [132,133].

The Siberian larch had the highest number of annotated genes related to response to hormones, JA biosynthetic process, ABA-activated signaling pathway, and ABA binding (Figure 8). On the contrary, a number of genes in ETH binding, ETH receptor activity, and response to ETH were relatively lower in all of the conifers, including Siberian larch, than in most of the angiosperm trees. This is probably not surprising, considering that the last components of the canonical ETH signaling pathway emerged after the separation of the angiosperms from the gymnosperms [134,135].

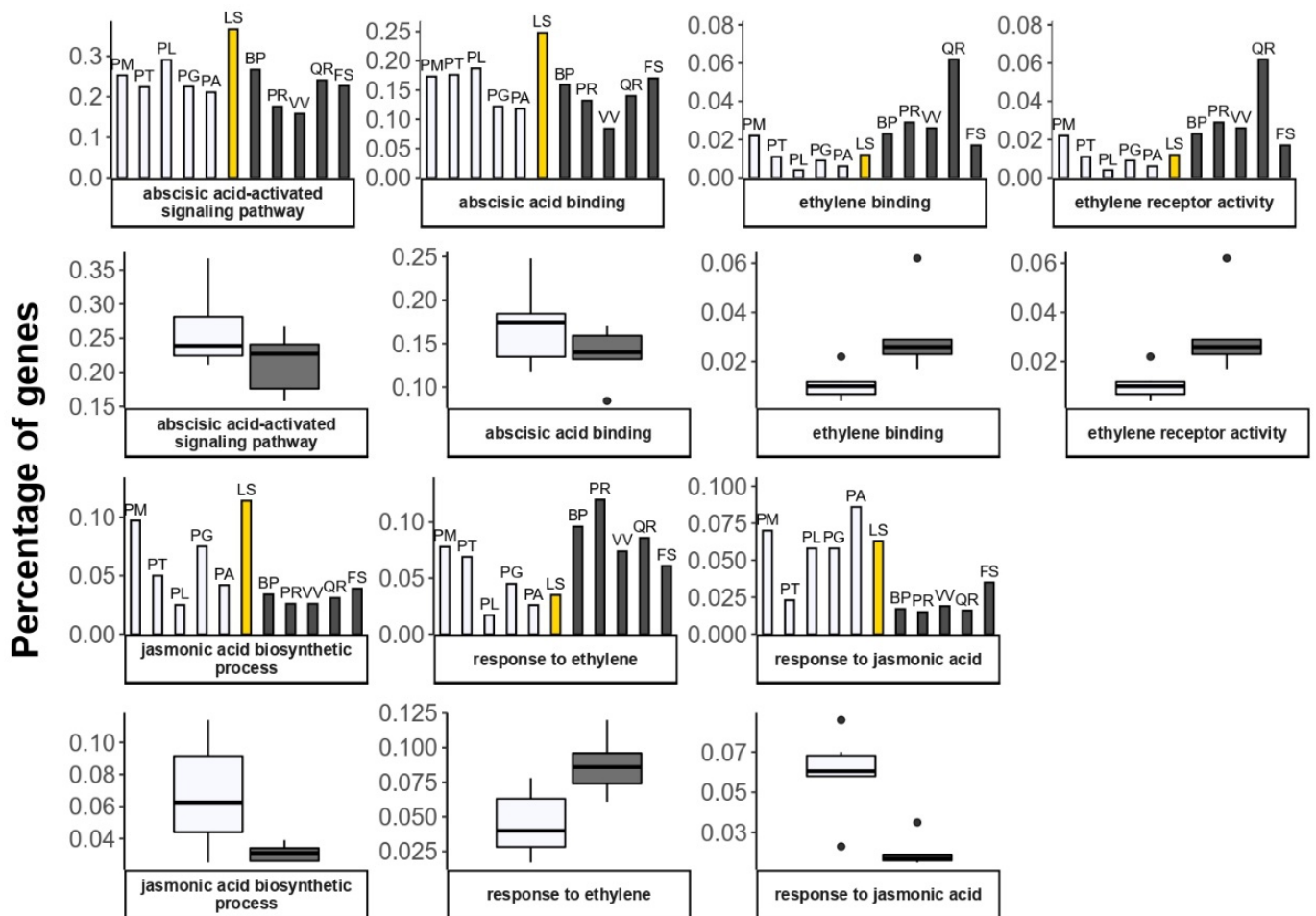


Figure 8. Percentage of genes annotated with GO terms related to hormone signaling and response. Deciduous angiosperm species are represented by black solid columns, evergreen gymnosperms—by transparent columns, Siberian larch—by yellow column. Boxplots demonstrate the difference in gene number between two groups, evergreen (transparent) and deciduous (gray). Angiosperms: BP—*Betula pendula*; FS—*Fagus sylvatica*; PR—*Populus trichocarpa*; QR—*Quercus robur*; VV—*Vitis vinifera*). Gymnosperms: PM—*Pseudotsuga menziesii*; PT—*Pinus taeda*; PL—*Pinus lambertiana*; PG—*Picea glauca*; PA—*Picea abies*; LS—*Larix sibirica*.

4. Materials and Methods

4.1. Genome Data

We used a Siberian larch genome assembly v1.0 (NCBI GenBank accession number GCA_004151065.1) with a total length of 12.34 Gb (Table 2), described in detail in Kuzmin et al. (2019) for annotation.

Table 2. The summary statistics of the Siberian larch assembly with a minimum contig length of 200 bp.

Assembly	Number, mln	N50, bp	Maximum Length, bp	Total Length, Gbp
Contigs	12.40	1074	128,642	7.99
Scaffolds	11.33	6443	354,326	12.34

4.2. Transcriptome Sequencing and Assembly

The RNA from the Siberian larch buds, needles, cambium, seedlings, and first-year shoots was isolated, using the Qiagen RNeasy Plant Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocols. The Illumina paired-end (PE) libraries, consisting of 250–400 bp long DNA fragments, were prepared for the larch buds using the TruSeq RNA Sample Preparation v2, and for the needles, cambium, shoots, and seedlings using the TruSeq Stranded RNA Kit with Ribo-Zero Plant kits (Illumina Inc., San Diego, CA, USA). The sequencing was completed at the Laboratory of Forest Genomics, Institute of Fundamental Biology and Biotechnology, Siberian Federal University, Krasnoyarsk, Russia on an Illumina MiSeq platform with 250 cycles in both directions (250×2) using the MiSeq v2 Reagent Kit 500 Cycles PE (Illumina Inc., San Diego, CA, USA).

The FastQC software v. 0.11.9 was used to evaluate the quality of the sequencing data (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc>, accessed on 16 January 2020). The raw sequencing data were processed using the Trimmomatic program v. 0.39 (9-bp headcrop, minimum read quality of $Q = 23$, and minimum read length of 35 bp; [35]). The SortMeRNA version 4.0.0 was used for the ribosomal RNA removal. In addition, Rcorrector was used for the sequencing error correction [136]. The transcripts were assembled using the TrinityRnaSeq package v2.2.0 [36]. For additional filtering, the transcripts were scanned for the presence of open reading frames (ORFs), using TransDecoder v3.0.1 [137], and conservative protein domains using Pfam [137–139]. The transcripts that did not possess coding regions were excluded from the subsequent analysis. For annotation via MAKER2, the pipeline transcriptomes containing rRNA were used.

4.3. Repetitive Elements (REs) Analysis and Masking

To search for REs, RepeatModeler v.1.0.11 [37], based on de novo RE detection programs RepeatScout and RECON [140,141], was used. Since RepeatScout does not use all of the scaffolds or contigs for the analysis, but only a random sample, it was decided to only analyze scaffolds longer than 100 Kbp (2869 scaffolds with summary length of 360 Mbp). RepeatMasker open-4.0.6 [38] was used to mask the low complexity regions and REs. The RepeatMasker edition of RepBase 2017.01.27 [142] was extended by larch-specific repeat library, constructed using RepeatModeler open-1.0.8 which was run with default settings. This combined database was used in MAKER2 pipeline to mask the repeats.

To assess the relative abundance of previously characterized repeat families, RepeatMasker was used on the whole genome assembly (12.34 Gbp). The RepeatModeler-derived de novo library was augmented by the clustering of frequently occurring reads from whole-genome sequencing data. The clusters of reads were assembled with Inchworm from TrinityRnaSeq v2.2.0, which resulted in consensus sequences that should represent the highly repeated regions of the larch genome. The unrecognized elements from the de novo repeat library generated by RepeatModeler and the clusters of frequently occurring reads were classified by TEclass v2.1.3, that classifies transposons using the Support Vector Machines (SVM) and LVQ neural network [143]. The combined library, comprising the RepeatModeler derived library classified with TEclass, RepBase library (edition 2017.01.27), MIPS Repeat Element Database library [144], CPRD Custom Plant Repeat Database [39], and Pine Interspersed Repeats Resource library PIER v1.0 [10,39] was used for sequence similarity search. The portion of long reads from the Oxford Nanopore sequencing available for Siberian larch was used to estimate the repeat abundance. In total, 3,060,509 reads with total length = 7.24 Gbp, minimum length = 25 bp, maximum = 77,840 bp, mean

length = 2365, min quality = 7, were used to search for known repeat families using RepeatMasker and combined repeat database comprising classified RepeatModeler-derived library, RepBase 2017 edition, MIPS, CPRD, and PIER. The classification of TEs was adopted from the RepBase update [145]. The GMATo [41] and TRF [42] programs were used to search for tandem repeats.

4.4. Identification of LRR Genes

The LRRs were searched in ORFs of the Siberian larch transcripts (NCBI SRA accession numbers SRX9464971, SRX14986114, SRX14997110, SRX14997111, and SRX14997112). The ORFs were identified using Transdecoder v.5.5.0 (<https://github.com/TransDecoder>, accessed on 3 August 2020). The ORFs of the transcript sequences were scanned by HMMER 3.2.1 [45] against the Pfam models LRR-1 (ID PF00560), LRR-2 (ID PF07723), LRR-3 (ID PF07725), LRR-4 (ID PF12799), LRR-5 (ID PF13306), LRR-6 (ID PF13516), LRR-8 (ID PF13855), and LRR-9 (ID PF14580). All of the LRR models were obtained from the Pfam 32.0 database and belong to the LRR clan (ID CL0022). The LRR clan also contains other families, but they were excluded because they represent bacteria, animals, and myxomycetes [139]. A search for NBS R-genes (NB-ARC; obtained from the Pfam 32.0 database: PF00931) was additionally performed, to check if some of the sequences with LRRs belong to R-genes.

The OmicsBox Base Platform [46,47] was used for the BLAST search, GO mapping, annotation, and statistical analysis. Gene ontology (GO) terms associated with the obtained BLAST results were extracted, and evaluated GO annotation was obtained. Enzyme codes were inferred by mapping with equivalent GOs, while the InterPro motifs were directly queried at the InterProScan web service on EMBL-EBI.

4.5. AUGUSTUS Training

We used MAKER2 [50] for an automated gene annotation of the Siberian larch genome assembly. The MAKER2 pipeline supports several ab initio gene predictors, including SNAP, GeneMark, and AUGUSTUS [146]. All of them require prior training to obtain species-specific parameters describing patterns of exon–intron structure. We used AUGUSTUS v3.2.1 to generate the preliminary gene models for the large sized genome of Siberian larch. This gene finder is based on a Generalized Hidden Markov Model (GHMM) and demonstrates good performance in ab initio prediction for non-model organisms [147,148], provided it has been trained appropriately, which can be tricky.

The AUGUSTUS training was carried out iteratively in several steps. First, AUGUSTUS was run with pre-calculated parameters for *Arabidopsis thaliana*, which resulted in 916K preliminary gene models and relatively low prediction accuracy. In order to build an initial training gene set, we used RNA-seq data. Transcriptome reads were mapped to the genome using TopHat [149], and the resulting alignments were used with Cufflinks to annotate coding regions [150]. This resulted in 16K transcriptome-derived gene models. Then, two gene sets from AUGUSTUS and Cufflinks, respectively, were merged together to filter the initial prediction set, and the genes with more than one exon were selected. This process of running AUGUSTUS with the transcriptome-derived gene set, filtering the resulting predictions and obtaining improved training parameters was repeated, till the prediction accuracy had become comparable to the average accuracy for AUGUSTUS, and the final training parameters were used with MAKER2 pipeline (Figure 9).

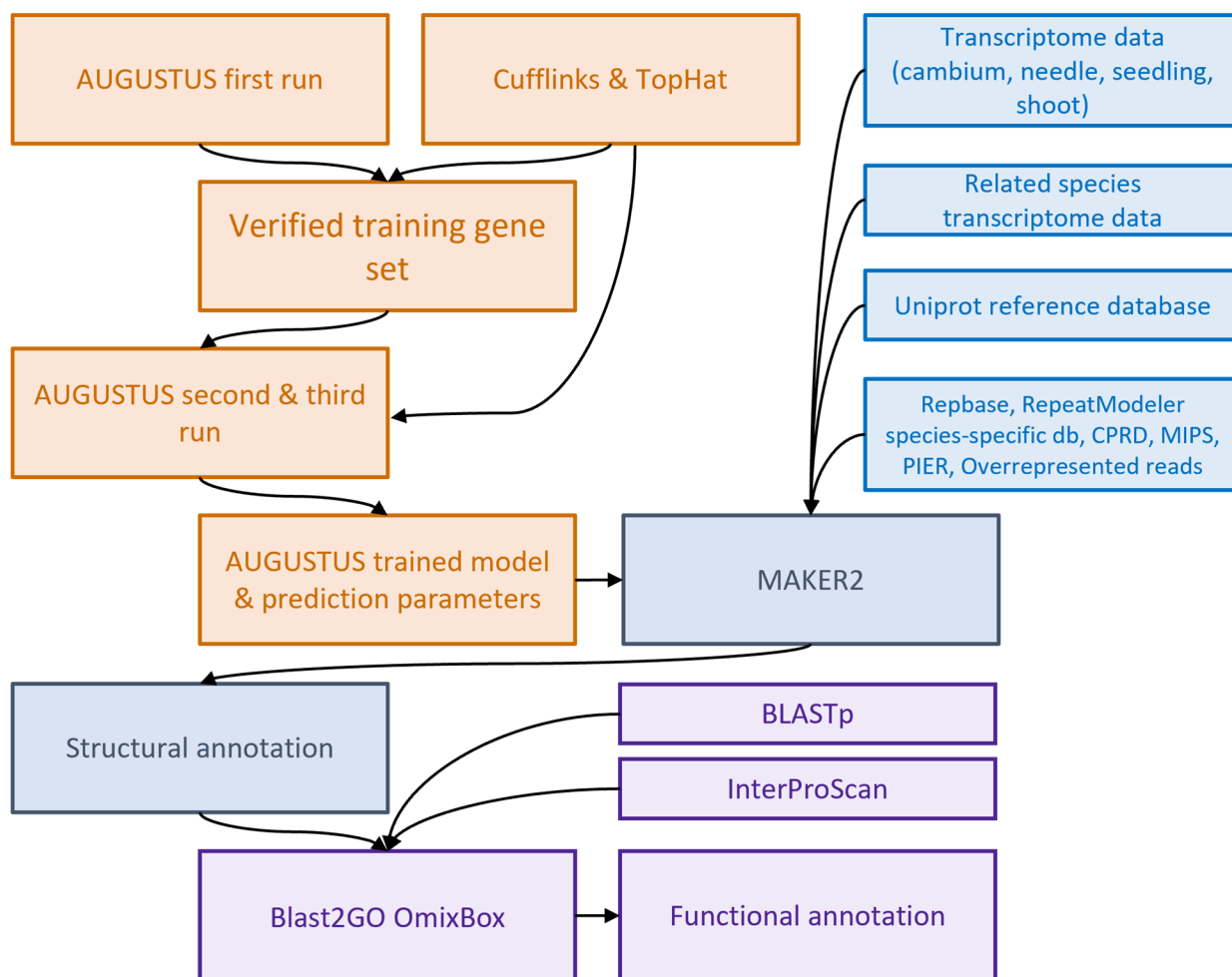


Figure 9. Genome annotation workflow.

4.6. MAKER Annotation

MAKER2 release 2.31.8 was used for producing the final annotation of the larch genome (parameters for MAKER configuration are presented in Text S1, Supplementary Materials). For BLAST, the ncbi-blast-2.2.29+ version was used. The Siberian larch transcriptome data and public transcript assemblies from the related conifer species (*Gnetum gnemon*, *Picea abies*, *Pinus lambertiana*, *Podocarpus macrophyllus*, *Pseudotsuga menziesii*, and *Pinus taeda*), deposited in the PlantGenIE project website (<https://plantgenie.org>, accessed on 1 August 2018), were used as supporting evidence. Uniprot was used as the protein reference database.

The genome annotation, using MAKER2, was performed on a supercomputer segment with 56 IBM BladeCenter HS21 servers (16 GB of RAM per server) at the Department of High-Performance Computing, Institute of Space and Information Technologies, Siberian Federal University, Krasnoyarsk, Russia and took 22 days (excluding the AUGUSTUS setup and the redo database setup). The whole process involved 448 cores at 2.3 GHz/core and 896 GB of RAM with the average processor load of about 61%.

4.7. Assembly Evaluation and Functional Annotation

The assessment of gene completeness was completed using the BUSCO v4.0.5 benchmarking tool [49] with embryophyta reference database and protein sequences derived from the MAKER2 annotation for Siberian larch genome. The protein sets for the other gymnosperm species were taken from the treegenes database (<https://treegenesdb.org>, accessed on 20 June 2021).

It was shown that the distant homologs are usually more likely to be identified when using a smaller database [151]. Thus, the NCBI GenBank *nr* database used to map the gene models and infer a high confidence gene set was filtered by taxonomy ID at embryophyta level. The search for the protein domains was completed using InterProScan on EMBL-EBI webserver. The GO mapping was performed using Blast2GO software integrated within OmixBox Base Platform. All of the predicted genes were mapped to NCBI GenBank *nr* database using blastp, and matches to bacteria, fungi, and archaea were discarded (evalue $< 1 \times 10^{-5}$, pident > 20 , qcovhsp > 20) to eliminate the genes that could potentially represent proteins other than plant genome-derived.

To compare the larch with other gymnosperm and angiosperm species, genome annotations of five other conifers, *Picea glauca*, *Picea abies*, *Pinus lambertiana*, *Pinus taeda*, and *Pseudotsuga menziesii*, and five angiosperms, *Betula pendula*, *Fagus sylvatica*, *Populus trichocarpa*, *Quercus robur*, and *Vitis vinifera*, were used to perform the blast2go GO mapping. To identify the GO terms in which the number of mapped genes differed significantly, a test of proportions was used. Two methods, based on the false discovery rate (FDR) estimation, were used for multiple comparison to correct *p*-values, according to Benjamini and Hochberg [152] and Storey [153], respectively (Figure S8, Supplementary Materials).

4.8. LTR-RT Insertion Time Estimation

There are two common methods for estimating the insertion time of LTR-RTs by (1) measuring the sequence divergence between two flanking LTRs and inferring their divergence time using the species-specific mutation rate, and (2) analyzing the pairwise genetic distances between the RT-encoding sequences that belong to the paralogous elements of the same monophyletic RT group. Whilst they may give different time estimates for some of the LTR lineages, the time distribution profiles produced by both methods are similar [71].

For additional de novo identification of the LTR-RT elements, LTRharvest [43] was used with options “-tis -suf -lcp -des -ssp -sds -dna”. To filter the potential false-positive hits and retrieve a stricter set of LTR-RTs, LTR_retriever [44] was used on the results of LTRharvest with the following settings “-u 1.57e-8 -missrate 0.4 -noanno”.

Zhou et al. [81] have carried out a thorough search for LTR-RTs and generated an LTRs resource for 301 plant genomes, including publicly available draft genomes of 10 gymnosperms. For the *L. sibirica* draft genome, they identified 367 LTR-RT elements. This LTR-RT library was checked against the resulted LTR-RTs found by LTRharvest and LTR_retriever using blastn, and the sequences without matches were added to the final LTR-RT database for *L. sibirica*. The final combined library for *Larix* and LTR-RT libraries for other gymnosperm species from Zhou et al. (2021) were used for the insertion time estimation. The sequence divergence was calculated, using the Jukes–Cantor model [44]:

$$T = \frac{d}{2\mu}, d = -\frac{3}{4} \ln \left(1 - p \frac{4}{3} \right), \quad (1)$$

where *d* is the Jukes–Cantor genetic distance (proxy of divergence rate); μ —the mutation rate; and *p*—the proportion of sequence differences ($p = 1 - \textit{identity}$, where *identity* is approximated using blastn). Insertion times converted into million years assuming a synonymous substitution rate $\mu = 1.57 \times 10^{-8}$ per site per year [154].

5. Conclusions

Despite being fragmented and incomplete, draft assemblies and annotations of conifer genomes still represent a valuable resource for further genomic and genetic studies. The current state of the genome annotations allows the differences between the gymnosperm and angiosperm species to be compared on a genomic level, evidenced by differences in gene abundance in different functional categories, such as cell wall organization and metabolism, PCD, and autophagy, which are related to frost resistance, seasonal senescence, stress hormone biosynthesis, and regulatory pathways.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/plants11152062/s1>, Text S1: Parameters for MAKER configuration; Table S1: Summary of repeat content in Siberian larch genome assembly annotated using RepeatMasker and combined library, comprising the RepeatModeler-derived library classified with TEclass, RepBase, MIPS, CPRD and PIER v1.0 libraries; Table S2: Summary of repeat content in a portion of long nanopore reads annotated using RepeatMasker and combined library, comprising the RepeatModeler-derived library classified with TEclass, RepBase, MIPS, CPRD and PIER v1.0 libraries; Table S3: LTR repeat superfamilies in the loblolly pine BAC library found in the Siberian larch genome; Table S4: The number of transcripts containing the LRR domain among all transcripts, and the number of transcripts containing the ARC domain among transcripts shorter than 850 bp for transcriptomes of different Siberian larch tissues; Table S5: Assessment of gene space completeness using BUSCO; Table S6: Assessment of gene space completeness using BUSCO protein mode and protein sequences for annotated gene models; Table S7: Summary of gene and genome assembly statistics among conifer and angiosperm plant species; Table S8: GO terms associated with cell wall metabolism and organization, programmed cell death (PCD), and hormone response that had a higher number of annotated genes in gymnosperms than in angiosperms; Figure S1: Length distribution of the main LTR-retrotransposons families in the Siberian larch genome; Figure S2: Length distribution of the Gypsy-retrotransposon superfamilies in the Siberian larch genome; Figure S3: Length distribution of the Copia-retrotransposon superfamilies in the Siberian larch genome; Figure S4: Length distribution of the main nonLTR-retrotransposon families in the Siberian larch genome; Figure S5: Length distribution of the main DNA-transposon families in the Siberian larch genome; Figure S6: (A–E)—transcripts with LRRs found using nine LRR families (Lrr1–Lrr9) in the transcriptomes of five tissues (for example, 43 transcripts were found by each of the LRR-1, LRR-4 and LRR-8 families), (F)—the distribution of the lengths of the amino acids sequences of the putative R-genes, LRR—transcripts containing the LRR domain (in blue), LRR and ARC—transcripts containing the LRR and ARC domains (in red); Figure S7: Repeats overlapping with predicted gene models. (A)—number of repeats overlapping with coding parts (CDS) with at least 20% overlap, (B)—the most frequent functional annotations of genes that overlap with repeats; Figure S8: Storey's q -value estimates for FDR in GO comparison. (A)—the estimated proportion of true null hypotheses (π_0) vs. λ ; (B)—the number of significant tests vs. each q -value cutoff; (C)—the q -values vs. the p -values; (D)—the number of expected false positives vs. the number of significant tests.

Author Contributions: Conceptualization, S.I.F. and K.V.K.; methodology, E.I.B., S.I.F., K.A.M. and K.V.K.; software, S.I.F., V.V.S. and D.A.K.; validation, E.I.B. and S.I.F.; formal analysis, E.I.B., K.A.M. and S.I.F.; investigation, E.I.B. and S.I.F.; resources, N.V.O., D.A.K. and K.V.K.; data curation, E.I.B., V.V.S. and S.I.F.; writing—original draft preparation, E.I.B.; writing—review and editing, K.V.K.; visualization, E.I.B. and K.A.M.; supervision, N.V.O. and K.V.K.; project administration, K.V.K.; funding acquisition, K.V.K. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by a research grant No. 14.Y26.31.0004 from the Government of the Russian Federation. No funding agency played any role in the design or conclusion of this study.

Data Availability Statement: Data generated during the study, including gff annotation files, RepeatModeler generated and combined plant repeat libraries, RepeatMasker output, LTR-RT non-redundant library, and LRR-containing sequences, are available in figshare with DOI 10.6084/m9.figshare.19785913 or in SibFU repository at <https://hpccloud.sfu-kras.ru/owncloud/index.php/s/GMBabOGEgqOD4JX> (accessed on 12 July 2022). Transcriptome sequencing data from bud, needle, stem, seedling, and cambium tissues are available at NCBI sequence read archive under accession numbers SRX9464971, SRX14986114, SRX14997110, SRX14997111, and SRX14997112. Transcriptome assemblies are available under accession numbers GIXH00000000, GJYD00000000, GJYL00000000, GJYN00000000, and GJYW00000000.

Acknowledgments: We thank the Department of High-Performance Computing for their help with computing using their HPC cluster at the Siberian Federal University and Thomas Byram for proof-reading the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McLoughlin, S. Gymnosperms. In *Encyclopedia of Geology*, 2nd ed.; Alderton, D., Elias, S.A., Eds.; Elsevier: Amsterdam, The Netherlands, 2021; Volume 3, pp. 476–500. [[CrossRef](#)]
2. Brenner, E.D.; Stevenson, D. Using Genomics to Study Evolutionary Origins of Seeds. In *Landscapes, Genomics and Transgenic Conifers. Managing Forest Ecosystems*; Williams, C.G., Ed.; Springer: Dordrecht, The Netherlands, 2006; Volume 9, pp. 85–106. [[CrossRef](#)]
3. Soltis, P.S.; Soltis, D.E.; Savolainen, V.; Crane, P.R.; Barraclough, T.G. Rate heterogeneity among lineages of tracheophytes: Integration of molecular and fossil data and evidence for molecular living fossils. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 4430–4435. [[CrossRef](#)] [[PubMed](#)]
4. Wan, T.; Liu, Z.-M.; Li, L.-F.; Leitch, A.R.; Leitch, I.J.; Lohaus, R.; Liu, Z.-J.; Xin, H.-P.; Gong, Y.-B.; Liu, Y.; et al. A genome for gnetophytes and early evolution of seed plants. *Nat. Plants* **2018**, *4*, 82–89. [[CrossRef](#)] [[PubMed](#)]
5. Stevens, K.A.; Wegrzyn, J.L.; Zimin, A.; Puiu, D.; Crepeau, M.; Cardeno, C.; Paul, R.; Gonzalez-Ibeas, D.; Koriabine, M.; Holtz-Morris, A.E.; et al. Sequence of the sugar pine megagenome. *Genetics* **2016**, *204*, 1613–1626. [[CrossRef](#)]
6. Berardini, T.Z.; Reiser, L.; Li, D.; Mezheritsky, Y.; Muller, R.; Strait, E.; Huala, E. The Arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. *Genesis* **2015**, *53*, 474–485. [[CrossRef](#)] [[PubMed](#)]
7. Badouin, H.; Gouzy, J.; Grassa, C.J.; Murat, F.; Staton, S.E.; Cottret, L.; Lelandais-Brière, C.; Owens, G.L.; Carrère, S.; Mayjonade, B.; et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **2017**, *546*, 148–152. [[CrossRef](#)] [[PubMed](#)]
8. Zimin, A.V.; Puiu, D.; Hall, R.; Kingan, S.; Clavijo, B.J.; Salzberg, S.L. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *GigaScience* **2017**, *6*, gix097. [[CrossRef](#)]
9. Pellicer, J.; Fay, M.F.; Leitch, I.J. The largest eukaryotic genome of them all? *Bot. J. Linn. Soc.* **2010**, *164*, 10–15. [[CrossRef](#)]
10. 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.; Liechty, J.D.; et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* **2014**, *15*, R59. [[CrossRef](#)]
11. Zimin, A.; Stevens, K.A.; Crepeau, M.W.; Holtz-Morris, A.; Koriabine, M.; Marçais, G.; Puiu, D.; Roberts, M.; Wegrzyn, J.L.; de Jong, P.J.; et al. Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics* **2014**, *196*, 875–890. [[CrossRef](#)]
12. Warren, R.L.; Keeling, C.I.; Yuen, M.M.S.; Raymond, A.; Taylor, G.A.; Vandervalk, B.P.; Mohamadi, H.; Paulino, D.; Chiu, R.; Jackman, S.D.; et al. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *Plant J. Cell Mol. Biol.* **2015**, *83*, 189–212. [[CrossRef](#)]
13. Li, Z.; Baniaga, A.E.; Sessa, E.B.; Scascitelli, M.; Graham, S.W.; Rieseberg, L.H.; Barker, M.S. Early genome duplications in conifers and other seed plants. *Sci. Adv.* **2015**, *1*, e1501084. [[CrossRef](#)] [[PubMed](#)]
14. Qiao, X.; Li, Q.; Yin, H.; Qi, K.; Li, L.; Wang, R.; Zhang, S.; Paterson, A.H. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **2019**, *20*, 38. [[CrossRef](#)]
15. Perera, D.; Magbanua, Z.V.; Thummasuwan, S.; Mukherjee, D.; Arick, M.; Chouvarine, P.; Nairn, C.J.; Schmutz, J.; Grimwood, J.; Dean, J.F.D.; et al. Exploring the loblolly pine (*Pinus taeda* L.) genome by BAC sequencing and Cot analysis. *Gene* **2018**, *663*, 165–177. [[CrossRef](#)] [[PubMed](#)]
16. 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.; Sezen, U.U.; et al. The Douglas-fir genome sequence reveals specialization of the photosynthetic apparatus in Pinaceae. *G3 Genes Genomes Genet.* **2017**, *7*, 3157–3167. [[CrossRef](#)] [[PubMed](#)]
17. Wegrzyn, J.L.; Liechty, J.D.; Stevens, K.A.; Wu, L.-S.; Loopstra, C.A.; Vasquez-Gross, H.A.; Dougherty, W.M.; Lin, B.Y.; Zieve, J.J.; Martínez-García, P.J.; et al. Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics* **2014**, *196*, 891–909. [[CrossRef](#)]
18. Pellicer, J.; Hidalgo, O.; Dodsworth, S.; Leitch, I.J. Genome size diversity and its impact on the evolution of land plants. *Genes* **2018**, *9*, 88. [[CrossRef](#)]
19. Nystedt, B.; Street, N.R.; Wetterbom, A.; Zuccolo, A.; Lin, Y.-C.; Scofield, D.G.; Vezzi, F.; Delhomme, N.; Giacomello, S.; Alexeyenko, A.; et al. The Norway Spruce genome sequence and conifer genome evolution. *Nature* **2013**, *497*, 579–584. [[CrossRef](#)]
20. Mosca, E.; Cruz, F.; Gómez-Garrido, J.; Bianco, L.; Rellstab, C.; Brodbeck, S.; Csilléry, K.; Fady, B.; Fladung, M.; Fussi, B.; et al. A reference genome sequence for the European silver fir (*Abies alba* Mill.): A community-generated genomic resource. *G3 Genes Genomes Genet.* **2019**, *9*, 2039–2049. [[CrossRef](#)]
21. Kuzmin, D.A.; Feranchuk, S.I.; Sharov, V.V.; Cybin, A.N.; Makolov, S.V.; Putintseva, Y.A.; Oreshkova, N.V.; Krutovsky, K.V. Stepwise large genome assembly approach: A case of Siberian larch (*Larix sibirica* Ledeb). *BMC Bioinform.* **2019**, *20*, 37. [[CrossRef](#)]
22. 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.; Zhang, S.-G. The *Larix kaempferi* genome reveals new insights into wood properties. *J. Integr. Plant Biol.* **2022**, *64*, 1364–1373. [[CrossRef](#)]
23. Niu, S.; Li, J.; Bo, W.; Yang, W.; Zuccolo, A.; Giacomello, S.; Chen, X.; Han, F.; Yang, J.; Song, Y.; et al. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* **2022**, *185*, 204–217.e14. [[CrossRef](#)] [[PubMed](#)]
24. Semerikov, V.L.; Lascoux, M. Nuclear and cytoplasmic variation within and between Eurasian *Larix* (Pinaceae) species. *Am. J. Bot.* **2003**, *90*, 1113–1123. [[CrossRef](#)] [[PubMed](#)]

25. Tumenjargal, B.; Ishiguri, F.; Aiso, H.; Takahashi, Y.; Nezu, I.; Takashima, Y.; Baasan, B.; Chultem, G.; Ohshima, J.; Yokota, S. Physical and mechanical properties of wood and their geographic variations in *Larix sibirica* trees naturally grown in Mongolia. *Sci. Rep.* **2020**, *10*, 12936. [CrossRef] [PubMed]
26. Semerikov, V.L.; Semerikova, S.A.; Polezhaeva, M.A.; Kosintsev, P.A.; Lascoux, M. Southern montane populations did not contribute to the recolonization of West Siberian Plain by Siberian larch (*Larix sibirica*): A range-wide analysis of cytoplasmic markers. *Mol. Ecol.* **2013**, *22*, 4958–4971. [CrossRef] [PubMed]
27. Dulamsuren, C.; Hauck, M.; Khishigjargal, M.; Leuschner, H.H.; Leuschner, C. Diverging climate trends in Mongolian taiga forests influence growth and regeneration of *Larix sibirica*. *Oecologia* **2010**, *163*, 1091–1102. [CrossRef]
28. Babushkina, E.A.; Vaganov, E.A.; Grachev, A.M.; Oreshkova, N.V.; Belokopytova, L.V.; Kostyakova, T.V.; Krutovsky, K.V. The effect of individual genetic heterozygosity on general homeostasis, heterosis and resilience in Siberian larch (*Larix sibirica* Ledeb.) using dendrochronology and microsatellite loci genotyping. *Dendrochronologia* **2016**, *38*, 26–37. [CrossRef]
29. Oreshkova, N.V.; Belokon', M.M.; Zham'iansuren, S. Genetic diversity, population structure, and differentiation of Siberian larch, Gmelin larch and Cajander larch on SSR-markers data. *Genetika* **2013**, *49*, 204–213. [CrossRef]
30. Oreshkova, N.V.; Putintseva, Y.A.; Sharov, V.V.; Kuzmin, D.A.; Krutovsky, K.V. Development of microsatellite genetic markers in Siberian larch (*Larix sibirica* Ledeb.) based on the *de novo* whole genome sequencing. *Russ. J. Genet.* **2017**, *53*, 1194–1199. [CrossRef]
31. Oreshkova, N.V.; Bondar, E.I.; Putintseva, Y.A.; Sharov, V.V.; Kuzmin, D.A.; Krutovsky, K.V. Development of nuclear microsatellite markers with long (tri-, tetra-, penta-, and hexanucleotide) motifs for three larch species based on the *de novo* whole genome sequencing of Siberian larch (*Larix sibirica* Ledeb.). *Russ. J. Genet.* **2019**, *55*, 444–450. [CrossRef]
32. Krutovsky, K.V.; Putintseva, Y.A.; Oreshkova, N.V.; Bondar, E.I.; Sharov, V.V.; Kuzmin, D.A. Postgenomic technologies in practical forestry: Development of genome-wide markers for timber origin identification and other applications. *For. Eng. J.* **2019**, *9*, 9–16. [CrossRef]
33. Bondar, E.I.; Putintseva, Y.A.; Oreshkova, N.V.; Krutovsky, K.V. Siberian larch (*Larix sibirica* Ledeb.) chloroplast genome and development of polymorphic chloroplast markers. *BMC Bioinform.* **2019**, *20*, 38. [CrossRef] [PubMed]
34. Putintseva, Y.A.; Bondar, E.I.; Simonov, E.P.; Sharov, V.V.; Oreshkova, N.V.; Kuzmin, D.A.; Konstantinov, Y.M.; Shmakov, V.N.; Belkov, V.I.; Sadovsky, M.G.; et al. Siberian larch (*Larix sibirica* Ledeb.) mitochondrial genome assembled using both short and long nucleotide sequence reads is currently the largest known mitogenome. *BMC Genom.* **2020**, *21*, 654. [CrossRef] [PubMed]
35. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
36. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef] [PubMed]
37. Smit, A.; Hubley, R. *RepeatModeler Open-1.0*. 2008. Available online: <https://www.repeatmasker.org/RepeatModeler> (accessed on 6 June 2018).
38. Smit, A.; Hubley, R.; Green, P. *RepeatMasker Open-4.0*. 2013. Available online: <https://www.repeatmasker.org/RepeatMasker> (accessed on 23 January 2016).
39. Wegrzyn, J.L.; Lin, B.Y.; Zieve, J.J.; Dougherty, W.M.; Martínez-García, P.J.; Koriabine, M.; Holtz-Morris, A.; deJong, P.; Crepeau, M.; Langley, C.H.; et al. Insights into the loblolly pine genome: Characterization of BAC and fosmid sequences. *PLoS ONE* **2013**, *8*, e72439. [CrossRef]
40. Magbanua, Z.V.; Ozkan, S.; Bartlett, B.D.; Chouvarine, P.; Saski, C.A.; Liston, A.; Cronn, R.C.; Nelson, C.D.; Peterson, D.G. Adventures in the enormous: A 1.8 million clone BAC library for the 21.7 Gb genome of loblolly pine. *PLoS ONE* **2011**, *6*, e16214. [CrossRef]
41. Wang, X.; Lu, P.; Luo, Z. GMATo: A novel tool for the identification and analysis of microsatellites in large genomes. *Bioinformatics* **2013**, *9*, 541–544. [CrossRef]
42. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **1999**, *27*, 573–580. [CrossRef]
43. Ellinghaus, D.; Kurtz, S.; Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinform.* **2008**, *9*, 18. [CrossRef]
44. Ou, S.; Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **2018**, *176*, 1410–1422. [CrossRef]
45. 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.* **2013**, *41*, e121. [CrossRef] [PubMed]
46. Conesa, A.; Götz, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genom.* **2008**, *2008*, 619832. [CrossRef] [PubMed]
47. Götz, S.; García-Gómez, J.M.; Terol, J.; Williams, T.D.; Nagaraj, S.H.; Nueda, M.J.; Robles, M.; Talón, M.; Dopazo, J.; Conesa, A. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **2008**, *36*, 3420–3435. [CrossRef] [PubMed]
48. Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **2014**, *30*, 1236–1240. [CrossRef]
49. Seppy, M.; Manni, M.; Zdobnov, E.M. BUSCO: Assessing Genome Assembly and Annotation Completeness. In *Gene Prediction. Methods in Molecular Biology*; Kollmar, M., Ed.; Humana: New York, NY, USA, 2019; Volume 1962, pp. 227–245. [CrossRef]

50. Holt, C.; Yandell, M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* **2011**, *12*, 491. [[CrossRef](#)]
51. Koralewski, T.E.; Krutovsky, K.V. Evolution of exon-intron structure and alternative splicing. *PLoS ONE* **2011**, *6*, e18055. [[CrossRef](#)]
52. Eilbeck, K.; Moore, B.; Holt, C.; Yandell, M. Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinform.* **2009**, *10*, 67. [[CrossRef](#)]
53. Eilbeck, K.; Lewis, S.E.; Mungall, C.J.; Yandell, M.; Stein, L.; Durbin, R.; Ashburner, M. The sequence ontology: A tool for the unification of genome annotations. *Genome Biol.* **2005**, *6*, R44. [[CrossRef](#)]
54. Sena, J.S.; Giguère, I.; Boyle, B.; Rigault, P.; Birol, I.; Zuccolo, A.; Ritland, K.; Ritland, C.; Bohlmann, J.; Jones, S.; et al. Evolution of gene structure in the conifer *Picea glauca*: A comparative analysis of the impact of intron size. *BMC Plant Biol.* **2014**, *14*, 95. [[CrossRef](#)]
55. Batalova, A.Y.; Putintseva, Y.A.; Sadovsky, M.G.; Krutovsky, K.V. Comparative genomics of seasonal senescence in forest trees. *Int. J. Mol. Sci.* **2022**, *23*, 3761. [[CrossRef](#)]
56. Naville, M.; Henriët, S.; Warren, I.; Sumic, S.; Reeve, M.; Volff, J.-N.; Chourrout, D. Massive changes of genome size driven by expansions of non-autonomous transposable elements. *Curr. Biol.* **2019**, *29*, 1161–1168.e6. [[CrossRef](#)] [[PubMed](#)]
57. Belyayev, A. Bursts of transposable elements as an evolutionary driving force. *J. Evol. Biol.* **2014**, *27*, 2573–2584. [[CrossRef](#)] [[PubMed](#)]
58. Zeh, D.W.; Zeh, J.A.; Ishida, Y. Transposable elements and an epigenetic basis for punctuated equilibria. *BioEssays* **2009**, *31*, 715–726. [[CrossRef](#)] [[PubMed](#)]
59. Tsukahara, S.; Kobayashi, A.; Kawabe, A.; Mathieu, O.; Miura, A.; Kakutani, T. Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature* **2009**, *461*, 423–426. [[CrossRef](#)]
60. Piegu, B.; Guyot, R.; Picault, N.; Roulin, A.; Saniyal, A.; Kim, H.; Collura, K.; Brar, D.S.; Jackson, S.; Wing, R.A.; et al. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **2006**, *16*, 1262–1269. [[CrossRef](#)]
61. Wang, W.; Ma, L.; Becher, H.; Garcia, S.; Kovarikova, A.; Leitch, I.J.; Leitch, A.R.; Kovarik, A. Astonishing 35S rDNA diversity in the gymnosperm species *Cycas revoluta* Thunb. *Chromosoma* **2016**, *125*, 683–699. [[CrossRef](#)]
62. Kelly, L.J.; Renny-Byfield, S.; Pellicer, J.; Macas, J.; Novák, P.; Neumann, P.; Lysak, M.A.; Day, P.D.; Berger, M.; Fay, M.F.; et al. Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytol.* **2015**, *208*, 596–607. [[CrossRef](#)]
63. Civián, P.; Švec, M.; Hauptvogel, P. On the coevolution of transposable elements and plant genomes. *J. Bot.* **2011**, *2011*, e893546. [[CrossRef](#)]
64. Arkhipova, I.R. Distribution and phylogeny of Penelope-like elements in eukaryotes. *Syst. Biol.* **2006**, *55*, 875–885. [[CrossRef](#)]
65. Evgen'ev, M.B.; Arkhipova, I.R. Penelope-like elements—A new class of retroelements: Distribution, function and possible evolutionary significance. *Cytogenet. Genome Res.* **2005**, *110*, 510–521. [[CrossRef](#)]
66. Lin, X.; Faridi, N.; Casola, C. An ancient transkingdom horizontal transfer of Penelope-like retroelements from arthropods to conifers. *Genome Biol. Evol.* **2016**, *8*, 1252–1266. [[CrossRef](#)] [[PubMed](#)]
67. Gao, D.; Chu, Y.; Xia, H.; Xu, C.; Heyduk, K.; Abernathy, B.; Ozias-Akins, P.; Leebens-Mack, J.H.; Jackson, S.A. Horizontal transfer of non-LTR retrotransposons from arthropods to flowering plants. *Mol. Biol. Evol.* **2018**, *35*, 354–364. [[CrossRef](#)] [[PubMed](#)]
68. Zhang, L.; Yan, L.; Jiang, J.; Wang, Y.; Jiang, Y.; Yan, T.; Cao, Y. The structure and retrotransposition mechanism of LTR-retrotransposons in the asexual yeast *Candida albicans*. *Virulence* **2014**, *5*, 655–664. [[CrossRef](#)]
69. Wicker, T.; Sabot, F.; Hua-Van, A.; Bennetzen, J.L.; Capy, P.; Chalhoub, B.; Flavell, A.; Leroy, P.; Morgante, M.; Panaud, O.; et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **2007**, *8*, 973–982. [[CrossRef](#)]
70. Aroh, O.; Halanych, K.M. Genome-wide characterization of LTR retrotransposons in the non-model deep-sea annelid *Lamellibrachia luymesii*. *BMC Genom.* **2021**, *22*, 466. [[CrossRef](#)]
71. Mascagni, F.; Usai, G.; Natali, L.; Cavallini, A.; Giordani, T. A Comparison of methods for LTR-retrotransposon insertion time profiling in the *Populus trichocarpa* genome. *Caryologia* **2018**, *71*, 85–92. [[CrossRef](#)]
72. Barghini, E.; Mascagni, F.; Natali, L.; Giordani, T.; Cavallini, A. Identification and characterisation of short interspersed nuclear elements in the olive tree (*Olea europaea* L.) genome. *Mol. Genet. Genom.* **2017**, *292*, 53–61. [[CrossRef](#)]
73. Kumar, A.; Bennetzen, J.L. Plant retrotransposons. *Annu. Rev. Genet.* **1999**, *33*, 479–532. [[CrossRef](#)]
74. Yin, H.; Du, J.; Wu, J.; Wei, S.; Xu, Y.; Tao, S.; Wu, J.; Zhang, S. Genome-wide annotation and comparative analysis of long terminal repeat retrotransposons between pear species of *P. bretschneideri* and *P. communis*. *Sci. Rep.* **2015**, *5*, 17644. [[CrossRef](#)]
75. Yin, H.; Liu, J.; Xu, Y.; Liu, X.; Zhang, S.; Ma, J.; Du, J. TARE1, a mutated Copia-like LTR retrotransposon followed by recent massive amplification in tomato. *PLoS ONE* **2013**, *8*, e68587. [[CrossRef](#)]
76. Zhao, M.; Du, J.; Lin, F.; Tong, C.; Yu, J.; Huang, S.; Wang, X.; Liu, S.; Ma, J. Shifts in the evolutionary rate and intensity of purifying selection between two *Brassica* genomes revealed by analyses of orthologous transposons and relics of a whole genome triplication. *Plant J.* **2013**, *76*, 211–222. [[CrossRef](#)]
77. Buti, M.; Giordani, T.; Cattonaro, F.; Cossu, R.M.; Pistelli, L.; Vukich, M.; Morgante, M.; Cavallini, A.; Natali, L. Temporal dynamics in the evolution of the sunflower genome as revealed by sequencing and annotation of three large genomic regions. *Theor. Appl. Genet.* **2011**, *123*, 779. [[CrossRef](#)] [[PubMed](#)]

78. Paterson, A.H.; Bowers, J.E.; Bruggmann, R.; Dubchak, I.; Grimwood, J.; Gundlach, H.; Haberer, G.; Hellsten, U.; Mitros, T.; Poliakov, A.; et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **2009**, *457*, 551–556. [[CrossRef](#)] [[PubMed](#)]
79. Brunner, S.; Fengler, K.; Morgante, M.; Tingey, S.; Rafalski, A. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* **2005**, *17*, 343–360. [[CrossRef](#)] [[PubMed](#)]
80. Wan, T.; Liu, Z.; Leitch, I.J.; Xin, H.; Maggs-Kölling, G.; Gong, Y.; Li, Z.; Marais, E.; Liao, Y.; Dai, C.; et al. The *Welwitschia* genome reveals a unique biology underpinning extreme longevity in deserts. *Nat. Commun.* **2021**, *12*, 4247. [[CrossRef](#)] [[PubMed](#)]
81. Zhou, S.-S.; Yan, X.-M.; Zhang, K.-F.; Liu, H.; Xu, J.; Nie, S.; Jia, K.-H.; Jiao, S.-Q.; Zhao, W.; Zhao, Y.-J.; et al. A comprehensive annotation dataset of intact LTR retrotransposons of 300 plant genomes. *Sci. Data* **2021**, *8*, 174. [[CrossRef](#)] [[PubMed](#)]
82. Song, H.; Guo, Z.; Hu, X.; Qian, L.; Miao, F.; Zhang, X.; Chen, J. Evolutionary balance between LRR domain loss and young NBS-LRR genes production governs disease resistance in *Arachis hypogaea* cv. Tifrunner. *BMC Genom.* **2019**, *20*, 844. [[CrossRef](#)] [[PubMed](#)]
83. Schaper, E.; Anisimova, M. The evolution and function of protein tandem repeats in plants. *New Phytol.* **2015**, *206*, 397–410. [[CrossRef](#)]
84. Jones, J.D.G.; Dangl, J.L. The plant immune system. *Nature* **2006**, *444*, 323–329. [[CrossRef](#)]
85. Kobe, B.; Kajava, A.V. The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* **2001**, *11*, 725–732. [[CrossRef](#)]
86. Niklas, K.J. The cell walls that bind the tree of life. *BioScience* **2004**, *54*, 831–841. [[CrossRef](#)]
87. Sarkar, P.; Bosneaga, E.; Auer, M. Plant cell walls throughout evolution: Towards a molecular understanding of their design principles. *J. Exp. Bot.* **2009**, *60*, 3615–3635. [[CrossRef](#)] [[PubMed](#)]
88. Li, L.; Cheng, X.F.; Leshkevich, J.; Umezawa, T.; Harding, S.A.; Chiang, V.L. The last step of syringyl monolignol biosynthesis in angiosperms is regulated by a novel gene encoding sinapyl alcohol dehydrogenase. *Plant Cell* **2001**, *13*, 1567–1586. [[CrossRef](#)]
89. Hatfield, R.; Vermerris, W. Lignin formation in plants. The dilemma of linkage specificity. *Plant Physiol.* **2001**, *126*, 1351–1357. [[CrossRef](#)]
90. Wagner, A.; Donaldson, L.; Ralph, J. Chapter 2—Lignification and Lignin Manipulations in Conifers. In *Advances in Botanical Research*; Jouanin, L., Lapierre, C., Eds.; Elsevier: Amsterdam, The Netherlands, 2012; Volume 61, pp. 37–76. [[CrossRef](#)]
91. Pascual, M.B.; El-Azaz, J.; de la Torre, F.N.; Cañas, R.A.; Avila, C.; Cánovas, F.M. Biosynthesis and metabolic fate of phenylalanine in conifers. *Front. Plant Sci.* **2016**, *7*, 1030. [[CrossRef](#)] [[PubMed](#)]
92. Yadav, V.; Wang, Z.; Wei, C.; Amo, A.; Ahmed, B.; Yang, X.; Zhang, X. Phenylpropanoid pathway engineering: An emerging approach towards plant defense. *Pathogens* **2020**, *9*, 312. [[CrossRef](#)]
93. Porth, I.; Hamberger, B.; White, R.; Ritland, K. Defense mechanisms against herbivory in *Picea*: Sequence evolution and expression regulation of gene family members in the phenylpropanoid pathway. *BMC Genom.* **2011**, *12*, 608. [[CrossRef](#)]
94. Vogt, T. Phenylpropanoid biosynthesis. *Mol. Plant* **2010**, *3*, 2–20. [[CrossRef](#)]
95. El-Azaz, J.; de la Torre, F.; Ávila, C.; Cánovas, F.M. Identification of a small protein domain present in all plant lineages that confers high prephenate dehydratase activity. *Plant J. Cell Mol. Biol.* **2016**, *87*, 215–229. [[CrossRef](#)]
96. Bagal, U.R.; Leebens-Mack, J.H.; Lorenz, W.W.; Dean, J.F. The phenylalanine ammonia lyase (PAL) gene family shows a gymnosperm-specific lineage. *BMC Genom.* **2012**, *13*, S1. [[CrossRef](#)]
97. van Doorn, W.G.; Beers, E.P.; Dangl, J.L.; Franklin-Tong, V.E.; Gallois, P.; Hara-Nishimura, I.; Jones, A.M.; Kawai-Yamada, M.; Lam, E.; Mundy, J.; et al. Morphological classification of plant cell deaths. *Cell Death Differ.* **2011**, *18*, 1241–1246. [[CrossRef](#)] [[PubMed](#)]
98. Klim, J.; Gładki, A.; Kucharczyk, R.; Zielenkiewicz, U.; Kaczanowski, S. Ancestral state reconstruction of the apoptosis machinery in the common ancestor of eukaryotes. *G3 Genes Genomes Genet.* **2018**, *8*, 2121–2134. [[CrossRef](#)] [[PubMed](#)]
99. Minina, E.A.; Smertenko, A.P.; Bozhkov, P.V. Vacuolar cell death in plants: Metacaspase releases the brakes on autophagy. *Autophagy* **2014**, *10*, 928–929. [[CrossRef](#)] [[PubMed](#)]
100. Hara-Nishimura, I.; Hatsugai, N. The role of vacuole in plant cell death. *Cell Death Differ.* **2011**, *18*, 1298–1304. [[CrossRef](#)]
101. Kalra, G.; Bhatla, S.C. Senescence and Programmed Cell Death. In *Plant Physiology, Development and Metabolism*; Bhatla, S.C., Lal, M.A., Eds.; Springer: Singapore, 2018; pp. 937–966. [[CrossRef](#)]
102. van Doorn, W.G. Classes of programmed cell death in plants, compared to those in animals. *J. Exp. Bot.* **2011**, *62*, 4749–4761. [[CrossRef](#)]
103. Reape, T.J.; Molony, E.M.; McCabe, P.F. Programmed cell death in plants: Distinguishing between different modes. *J. Exp. Bot.* **2008**, *59*, 435–444. [[CrossRef](#)]
104. Valandro, F.; Menguer, P.K.; Cabreira-Cagliari, C.; Margis-Pinheiro, M.; Cagliari, A. Programmed cell death (PCD) control in plants: New insights from the *Arabidopsis thaliana* deathosome. *Plant Sci.* **2020**, *299*, 110603. [[CrossRef](#)]
105. Delorme, V.G.; McCabe, P.F.; Kim, D.J.; Leaver, C.J. A matrix metalloproteinase gene is expressed at the boundary of senescence and programmed cell death in cucumber. *Plant Physiol.* **2000**, *123*, 917–927. [[CrossRef](#)]
106. Van Hautegeem, T.; Waters, A.J.; Goodrich, J.; Nowack, M.K. Only in dying, life: Programmed cell death during plant development. *Trends Plant Sci.* **2015**, *20*, 102–113. [[CrossRef](#)]
107. Koyama, T. The roles of ethylene and transcription factors in the regulation of onset of leaf senescence. *Front. Plant Sci.* **2014**, *5*, 650. [[CrossRef](#)]
108. Daneva, A.; Gao, Z.; Van Durme, M.; Nowack, M.K. Functions and regulation of programmed cell death in plant development. *Annu. Rev. Cell Dev. Biol.* **2016**, *32*, 441–468. [[CrossRef](#)] [[PubMed](#)]

109. Kim, S.-H.; Kwon, C.; Lee, J.-H.; Chung, T. Genes for plant autophagy: Functions and interactions. *Mol. Cells* **2012**, *34*, 413–423. [[CrossRef](#)] [[PubMed](#)]
110. Zhu, J.-K. Salt and drought stress signal transduction in plants. *Annu. Rev. Plant Biol.* **2002**, *53*, 247–273. [[CrossRef](#)] [[PubMed](#)]
111. Xue-Xuan, X.; Hong-Bo, S.; Yuan-Yuan, M.; Gang, X.; Jun-Na, S.; Dong-Gang, G.; Cheng-Jiang, R. Biotechnological implications from abscisic acid (ABA) roles in cold stress and leaf senescence as an important signal for improving plant sustainable survival under abiotic-stressed conditions. *Crit. Rev. Biotechnol.* **2010**, *30*, 222–230. [[CrossRef](#)]
112. Costa-Broseta, Á.; Perea-Resca, C.; Castillo, M.-C.; Ruíz, M.F.; Salinas, J.; León, J. Nitric oxide controls constitutive freezing tolerance in *Arabidopsis* by attenuating the levels of osmoprotectants, stress-related hormones and anthocyanins. *Sci. Rep.* **2018**, *8*, 9268. [[CrossRef](#)]
113. Wu, J.; Zhang, Y.; Yin, L.; Qu, J.; Lu, J. Linkage of cold acclimation and disease resistance through plant–pathogen interaction pathway in *Vitis amurensis* grapevine. *Funct. Integr. Genom.* **2014**, *14*, 741–755. [[CrossRef](#)]
114. Preston, J.; Sandve, S. Adaptation to seasonality and the winter freeze. *Front. Plant Sci.* **2013**, *4*, 167. [[CrossRef](#)]
115. Song, Y.; Xiang, F.; Zhang, G.; Miao, Y.; Miao, C.; Song, C.-P. Abscisic acid as an internal integrator of multiple physiological processes modulates leaf senescence onset in *Arabidopsis thaliana*. *Front. Plant Sci.* **2016**, *7*, 181. [[CrossRef](#)]
116. Xu, P.; Chen, H.; Cai, W. Transcription factor CDF4 promotes leaf senescence and floral organ abscission by regulating abscisic acid and reactive oxygen species pathways in *Arabidopsis*. *EMBO Rep.* **2020**, *21*, e48967. [[CrossRef](#)]
117. Lee, I.C.; Hong, S.W.; Whang, S.S.; Lim, P.O.; Nam, H.G.; Koo, J.C. Age-Dependent Action of an ABA-Inducible Receptor Kinase, RPK1, as a Positive Regulator of Senescence in *Arabidopsis* leaves. *Plant Cell Physiol.* **2011**, *52*, 651–662. [[CrossRef](#)]
118. Raab, S.; Drechsel, G.; Zarepour, M.; Hartung, W.; Koshiba, T.; Bittner, F.; Hoth, S. Identification of a novel E3 ubiquitin ligase that is required for suppression of premature senescence in *Arabidopsis*. *Plant J. Cell Mol. Biol.* **2009**, *59*, 39–51. [[CrossRef](#)]
119. Yang, J.; Worley, E.; Udvardi, M. A NAP-AAO3 regulatory module promotes chlorophyll degradation via ABA biosynthesis in *Arabidopsis* leaves. *Plant Cell* **2014**, *26*, 4862–4874. [[CrossRef](#)] [[PubMed](#)]
120. Breeze, E.; Harrison, E.; McHattie, S.; Hughes, L.; Hickman, R.; Hill, C.; Kiddle, S.; Kim, Y.-S.; Penfold, C.A.; Jenkins, D.; et al. High-resolution temporal profiling of transcripts during *Arabidopsis* leaf senescence reveals a distinct chronology of processes and regulation. *Plant Cell* **2011**, *23*, 873–894. [[CrossRef](#)] [[PubMed](#)]
121. Han, G.-Z. Evolution of jasmonate biosynthesis and signaling mechanisms. *J. Exp. Bot.* **2017**, *68*, 1323–1331. [[CrossRef](#)]
122. Ali, M.S.; Baek, K.-H. Jasmonic acid signaling pathway in response to abiotic stresses in plants. *Int. J. Mol. Sci.* **2020**, *21*, 621. [[CrossRef](#)] [[PubMed](#)]
123. Truman, W.; Bennett, M.H.; Kubigsteltig, I.; Turnbull, C.; Grant, M. *Arabidopsis* systemic immunity uses conserved defense signaling pathways and is mediated by jasmonates. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 1075–1080. [[CrossRef](#)]
124. Mohamed, H.I.; Latif, H.H. Improvement of drought tolerance of soybean plants by using methyl jasmonate. *Physiol. Mol. Biol. Plants* **2017**, *23*, 545–556. [[CrossRef](#)]
125. Todaka, D.; Shinozaki, K.; Yamaguchi-Shinozaki, K. Recent advances in the dissection of drought-stress regulatory networks and strategies for development of drought-tolerant transgenic rice plants. *Front. Plant Sci.* **2015**, *6*, 84. [[CrossRef](#)]
126. Qiu, Z.; Guo, J.; Zhu, A.; Zhang, L.; Zhang, M. Exogenous jasmonic acid can enhance tolerance of wheat seedlings to salt stress. *Ecotoxicol. Environ. Saf.* **2014**, *104*, 202–208. [[CrossRef](#)]
127. Fan, L.; Wang, Q.; Lv, J.; Gao, L.; Zuo, J.; Shi, J. Amelioration of postharvest chilling injury in cowpea (*Vigna sinensis*) by methyl jasmonate (MeJA) treatments. *Sci. Hortic.* **2016**, *203*, 95–101. [[CrossRef](#)]
128. Zhao, M.-L.; Wang, J.-N.; Shan, W.; Fan, J.-G.; Kuang, J.-F.; Wu, K.-Q.; Li, X.-P.; Chen, W.-X.; He, F.-Y.; Chen, J.-Y.; et al. Induction of jasmonate signalling regulators MaMYC2s and their physical interactions with MalCE1 in methyl jasmonate-induced chilling tolerance in banana fruit. *Plant Cell Environ.* **2013**, *36*, 30–51. [[CrossRef](#)] [[PubMed](#)]
129. Mewis, I.; Schreiner, M.; Nguyen, C.N.; Krumbein, A.; Ulrichs, C.; Lohse, M.; Zrenner, R. UV-B Irradiation changes specifically the secondary metabolite profile in broccoli sprouts: Induced signaling overlaps with defense response to biotic stressors. *Plant Cell Physiol.* **2012**, *53*, 1546–1560. [[CrossRef](#)]
130. Cerrudo, I.; Keller, M.M.; Cargnel, M.D.; Demkura, P.V.; de Wit, M.; Patitucci, M.S.; Pierik, R.; Pieterse, C.M.J.; Ballaré, C.L. Low red/far-red ratios reduce *Arabidopsis* resistance to *Botrytis cinerea* and jasmonate responses via a COI1-JAZ10-dependent, salicylic acid-independent mechanism. *Plant Physiol.* **2012**, *158*, 2042–2052. [[CrossRef](#)] [[PubMed](#)]
131. Svyatyna, K.; Riemann, M. Light-dependent regulation of the jasmonate pathway. *Protoplasma* **2012**, *249* (Suppl. S2), S137–S145. [[CrossRef](#)] [[PubMed](#)]
132. Kozłowski, G.; Buchala, A.; Métraux, J.-P. Methyl jasmonate protects Norway spruce [*Picea abies* (L.) Karst.] seedlings against *Pythium ultimum* Trow. *Physiol. Mol. Plant Pathol.* **1999**, *55*, 53–58. [[CrossRef](#)]
133. Franceschi, V.R.; Krekling, T.; Christiansen, E. Application of methyl jasmonate on *Picea abies* (Pinaceae) stems induces defense-related responses in phloem and xylem. *Am. J. Bot.* **2002**, *89*, 578–586. [[CrossRef](#)]
134. Wang, C.; Liu, Y.; Li, S.-S.; Han, G.-Z. Insights into the origin and evolution of the plant hormone signaling machinery. *Plant Physiol.* **2015**, *167*, 872–886. [[CrossRef](#)]
135. Groen, S.C.; Whiteman, N.K. The evolution of ethylene signaling in plant chemical ecology. *J. Chem. Ecol.* **2014**, *40*, 700–716. [[CrossRef](#)]
136. Song, L.; Florea, L. Rcorrector: Efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* **2015**, *4*, 48. [[CrossRef](#)]

137. Haas, B.J.; Papanicolaou, A.; Yassour, M.; Grabherr, M.; Blood, P.D.; Bowden, J.; Couger, M.B.; Eccles, D.; Li, B.; Lieber, M.; et al. *De novo* transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protoc.* **2013**, *8*, 1494–1512. [[CrossRef](#)]
138. Finn, R.D.; Coghill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; et al. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **2016**, *44*, D279–D285. [[CrossRef](#)] [[PubMed](#)]
139. El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S.R.; Luciani, A.; Potter, S.C.; Qureshi, M.; Richardson, L.J.; Salazar, G.A.; Smart, A.; et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **2019**, *47*, D427–D432. [[CrossRef](#)] [[PubMed](#)]
140. Bao, Z.; Eddy, S.R. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* **2002**, *12*, 1269–1276. [[CrossRef](#)]
141. Price, A.L.; Jones, N.C.; Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **2005**, *21* (Suppl. S1), i351–i358. [[CrossRef](#)] [[PubMed](#)]
142. Bao, W.; Kojima, K.K.; Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **2015**, *6*, 11. [[CrossRef](#)]
143. Abrusán, G.; Grundmann, N.; DeMester, L.; Makalowski, W. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **2009**, *25*, 1329–1330. [[CrossRef](#)]
144. Nussbaumer, T.; Martis, M.M.; Roessner, S.K.; Pfeifer, M.; Bader, K.C.; Sharma, S.; Gundlach, H.; Spannagl, M. MIPS PlantsDB: A database framework for comparative plant genome research. *Nucleic Acids Res.* **2013**, *41*, D1144–D1151. [[CrossRef](#)]
145. Kojima, K.K. Human transposable elements in repbase: Genomic footprints from fish to humans. *Mob. DNA* **2018**, *9*, 2. [[CrossRef](#)]
146. Stanke, M.; Schöffmann, O.; Morgenstern, B.; Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinform.* **2006**, *7*, 62. [[CrossRef](#)]
147. Scalzitti, N.; Jeannin-Girardon, A.; Collet, P.; Poch, O.; Thompson, J.D. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC Genomics* **2020**, *21*, 293. [[CrossRef](#)]
148. Stanke, M.; Keller, O.; Gunduz, I.; Hayes, A.; Waack, S.; Morgenstern, B. AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **2006**, *34*, W435–W439. [[CrossRef](#)]
149. Trapnell, C.; Pachter, L.; Salzberg, S.L. TopHat: Discovering splice junctions with RNA-seq. *Bioinformatics* **2009**, *25*, 1105–1111. [[CrossRef](#)] [[PubMed](#)]
150. Trapnell, C.; Roberts, A.; Goff, L.; Pertea, G.; Kim, D.; Kelley, D.R.; Pimentel, H.; Salzberg, S.L.; Rinn, J.L.; Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **2012**, *7*, 562–578. [[CrossRef](#)] [[PubMed](#)]
151. Pearson, W.R. An introduction to sequence similarity (“homology”) searching. *Curr. Protoc. Bioinform.* **2013**, *42*, 3.1.1–3.1.8. [[CrossRef](#)] [[PubMed](#)]
152. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300. [[CrossRef](#)]
153. Storey, J.D. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2002**, *64*, 479–498. [[CrossRef](#)]
154. De La Torre, A.R.; Li, Z.; Van de Peer, Y.; Ingvarsson, P.K. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Mol. Biol. Evol.* **2017**, *34*, 1363–1377. [[CrossRef](#)]