Repetitive Elements in the Siberian Larch (*Larix Sibirica* Ledeb.) Nuclear Genome

Kseniya A. Miroshnikova
Siberian Federal University
Vasilina S. Akulova
Siberian Federal University
Vladislav V. Biriukov
Siberian Federal University
Eugeniya I. Bondar
Siberian Federal University
Dmitry A. Kuzmin
Siberian Federal University
Natalya V. Oreshkova
Siberian Federal University
Michael G. Sadovsky
Siberian Federal University
Vadim V. Sharov
Siberian Federal University
Konstantin V. Krutovsky (konstantin.krutovsky@forst.uni-goettingen.de)
Georg-August University of Göttingen

Research Article

Keywords: Larix sibirica, leucine-rich repeats, repetitive elements, Siberian larch

Posted Date: August 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-678183/v1

Abstract

Background: Repetitive elements (REs) or repeats are sequences that occur multiple times in the genome. They represent a significant part of the gigantic conifer genomes (70-80%) relative to mammals and other plants and complicate whole genome sequencing and annotation. However, REs play important roles in evolution and adaptation processes in both plants and animals. Moreover, amino acid repeats play an important role in plant immunity being a structural element of the products of some disease resistance genes. Analysis of REs in conifer genomes is an important fundamental task.

Results: REs were identified *de novo* and partly classified in the Siberian larch (*Larix sibirica* Ledeb.) nuclear genome for the first time. In total, 20.9 million REs were detected with the total size of 4.8 Gbp, which comprises about 39% of the 12.3 Gbp larch genome. Resistance genes with leucine-rich repeats (LRRs) were also identified and analyzed in the transcriptome data of autumn buds obtained using RNA-seq.

Conclusions: For the first time, REs were identified and classified in the Siberian larch genome and transcriptome. In addition, LRRs and resistance genes were identified and analyzed in the Siberian larch transcriptomes from autumn buds. The larch genome contains twice as less RE compared to other conifers in the same Pinaceae family (39 vs 70-80%), and it might explain why it also has almost twice as smaller genome size (12 vs 18-31 Gbp).

Background

Conifers are amazingly invaluable representatives of the ancient group of gymnosperms. Their longevity and resistance to harsh conditions contribute to their dominance in different ecosystems of the Northern Hemisphere. Conifer genomes have remarkable gigantic sizes (12–31 Gbp), which is not associated with genome-wide duplication or polyploidization events, at least recent ones [1]. It was suggested that the main cause of the increased sizes is the propagation of repetitive DNA sequences [2]. Thus, conifer genomes are very difficult to study. Nevertheless, a technological breakthrough in DNA sequencing made it possible to assemble and analyse conifer genomes [3]. Repetitive elements (REs) composed 82% of the nuclear genome (22.1 Gbp) in *Pinus taeda* L. [4–7], 79% in the *Pinus lambertiana* Douglas genome (31 Gbp) [8], 70% in the *Picea abies* (L.) Karst genome (19.6 Gbp) [2], 70% in the *Picea glauca* (Moench) Voss genome (22.4 Gbp) [9–12], 78% in the *Abies alba* Mill genome (18.2 Gbp) [3], and 39% in the *Larix sibirica* Ledeb. genome (12.3 Gbp [13]) in this study.

REs in a genome or repeatome consist of diverse highly repetitive sequences including transposons or mobile elements, tandem repeats, various solo elements (such as inactivated transposons [14]), and endogenous viruses. REs can greatly vary in length and number of repeating units. These repeating units can be located sequentially one after another forming tandem repeats. This class of repeats forms simple structures such as microsatellite or simple sequence repeat loci, widely used as genetic markers [15], including in Siberian larch [16, 17]. Another broad class represents interspersed repeats, which are

more complex in structure and practically do not form blocks that follow each other. Interspersed repeats mainly occurred as a result of the mobile element or transposon activity and are scattered throughout the genome with unstable location for active transposons [18, 19].

These repeats are divided into autonomous and nonautonomous transposons. The former ones encode their own enzymes able to excise themselves and paste into another genome location. Nonautonomous elements use enzymes of autonomous transposons for these aims. For example, SINEs are nonautonomous retroelements that use the enzymatic machinery of autonomous LINEs. Many SINE families have been amplified recently leading to insertional polymorphism between closely related individuals, which is a desirable trait for a molecular genetic marker. In addition, this class has been identified in many plant families including *Gramineae*, *Commelinaceae*, *Rosaceae*, *Solanaceae*, *Fabaceae*, and *Brassicaceae* and can be used as a meaningful classification criterion to evaluate phylogenetic relations among species [20].

In addition to noncoding nucleotide repeats, there are also tandem codon repeats in protein-coding genes that encode amino acid repeats in protein sequences. They can represent a structural motif that performs a specific function. 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 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 high variation in type and number of the LRR units between and within species, which provides the specificity of pathogen molecules recognition. Thus, LRRs play an important role in plant immunity [21–24].

In the normal condition, the host defence mechanism supresses transposon activity and does not allow new transposon insertions. The regions of transposon activity are very prone to the mutation accumulation. According to [25], this ultimately leads to the formation of the so-called ‘genomic dark matter’. In such genome regions with high transposon activity, transposable elements (TEs) can overlay each other, which in combination with a higher mutation rate can result in hardly identifiable TEs [19, 25–27].

Nystedt et al. [2] suggested that mechanisms of removing TEs (e.g., via unequal recombination) were less active in conifers than in most other organisms. The chromosomes of the most recent conifer ancestor expanded at a slow but steady pace due to the activity of different mobile elements such as Gypsy and Copia, which are very abundant in conifer genomes. The insertion of TEs into genes can lead to the appearance of large introns and in combination with translocation to an abundant number of pseudogenes.

Siberian larch represents an unusual genus, which is the only genus with deciduous species in the *Pinaceae* family. Like other conifers it has also a relatively large genome size of 12.3 Gbp, but it is much smaller than in pines, spruces, and Douglas-fir [13]. Here we provide pioneering preliminary data on the larch repeatome.

Results

Search and identification of REs

RepeatModeler allowed for the generation of species-specific *de novo* repeat library containing 1721 mobile elements that were found in the current assembly of Siberian larch. Assembling the consensus sequences of 21 million clusters (with cluster size threshold of 200 reads per cluster) with Inchworm from TrinitRnaSeq package resulted in 31 thousand consensus sequences that likely represent repeated regions of Siberian larch genome. To validate these sequences, we compared them to the RepeatModeler-derived library and to a PIER repeat library (Wegrzyn et al. 2013; Neale et al. 2014). Homologs were found for 12000 consensus sequences among the RepeatModeler-derived library, and for 7000 sequences among PIER database. Reciprocal BLAST showed that 1045 out of 1721 RepeatModeler-derived sequences had a close homology to a clustering-derived consensus sequences.

The proportion of classified families observed in the Siberian larch genome was the same as in other conifers. The RepeatMasker and combined repeat library methods allowed to find 20.9 million REs with the total size of 4.8 Gbp, which comprises about 39% of the 12.3 Gbp genome (Table S1). LTRs were the most common REs and comprised 37.5% of all repeats in total. Gypsy superfamily prevailed over the Copia one, which is typical for conifers (Fig. 1). They also abundantly presented in other conifers [2, 7, 28, 29] and angiosperm species [30]. However, most of the LTR-retrotransposons have not been classified into smaller families; they were named "other LTR" in Table S1.

Substantial portion of LTRs was homologues to a loblolly pine bacterial artificial chromosome (BAC) library and fosmid sequences [31, 32], PtTalladega (3646 copies in the Siberian larch genome), PtOuachita (1025), IFG (990), PtAppalachian (773), PtConagree (731) and eight more repeat families were identified (Table 1).

Table 1
LTR repeat superfamilies in the loblolly pine BAC library found in the Siberian larch genome

Superfamily		Number of copies	Length, bp
Gipsy	PtTalladega	3646	1663158
	IFG	990	254098
	PtAppalachian	773	208844
	PtOuachita	1025	191626
	Gymny	476	138565
	PtBastrop	594	94602
	PtOzark	234	24665
	PsAppalachian	108	16203
	PtAngelina	4	497
Copia	PtConagree	731	115681
	Silava_Pta	581	115133
	PtPineywoods	401	64007
	PtCumberland	251	32521

The portion of non-LTR is not much lower than LTR, which distinguishes larch from other conifers. LINE elements take the largest part of the all classified REs in larch and comprised the biggest part of non-LTR retrotransposons and all classified REs in other conifers [2, 4, 7, 8]. As for genome coverage, retrotransposons LINE, Penelope and SINE together comprise the majority in the non-LTR group (Fig. 1).

DNA transposons comprised 11.63% of all REs in total. The largest part of these transposons was not classified to families, and they were marked as "other DNA" (Table S1). The majority of classified DNA repeats consisted of TIR and Helitron repeat families (Fig. 1).

The majority of repeats among different repeat families was also relatively small in length, less than 1 Kbp. A small part of the longest repeats reached almost 15 Kbp, they belong to LINE elements and uncharacterized LTR. The most frequent REs for each family were shorter than 1 Kbp. Some repeat groups have a bimodal length distribution (Gypsy, DIR, LINE/I, Helitron, Penelope), but both of peaks in the distributions were less than 1 Kbp (Figs S1-S5). The most of LTRs homologous to a loblolly pine BAC library did not exceed 500 bp. The relatively large number of the PtTalladega LTRs (relative to other families in the loblolly pine BAC library) were shorter than 1 Kbp and did not reach 4 Kbp (Figs S2 and S3).

LRR domains identification

In total, 195 transcripts containing the LRR domain were detected among 22116 transcripts in the autumn buds of Siberian larch using hidden Markov model (HMM) method HMMER3 to correctly assign homologous sequences to one or more Pfam families of LRR (0.9%). In comparison, in the transcriptome of the Sitka spruce bud, only 51 among 10105 transcripts contained the LRR domain (0.5%), which was significantly less than in the transcriptome of the Siberian larch bud according to the Fisher's exact test (two-tail $p = 0.0002$; Table 2).

Table 2
The number of transcripts containing the LRR domain in the Siberian larch and Sitka spruce autumn bud transcriptomes

Species	Transcripts		
	total	with LRR	%
Siberian larch	22116	195	0.9
Sitka spruce	10105	51	0.5

The LRR-1, LRR-4, LRR-8, and LRR-6 families encompassed the largest portion of the found putative LRR domains identified in both the larch and Sitka spruce transcriptomes (Fig. 2). As it can be seen in Fig. 2, the largest number of transcripts contained LRR that matched several LRR families, and the LRR-4 family was the most common match in both species. There were also a few sets of unique sequences that contained LRR that matched only one of the LRR families (highlighted in green in Fig. 2). The largest set of 24 such sequences contained LRR that was detected only by matching to the LRR-4 family. In Sitka spruce, this family matched LRR in almost all the sequences that were also revealed by other LRR families, except one set of 10 unique sequences found only by using this LRR-4 family in the search (Fig. 2B).

Identification of NB-ARC among all transcripts and transcripts containing LRRs

In total, 38 sequences among all the sequences in the larch transcriptome contained the NB-ARC domain, but only seven of them contained also the LRR domain. Meanwhile in spruce, six sequences with the NB-ARC domain were identified, but none of them contained the LRR domain.

NBS-LRR proteins are among the largest proteins known in plants, usually ranging from ~ 860 to ~ 1,900 amino acids [33], but most of the proteins encoded by transcripts containing the NBS-LRR genes were shorter than 300–400 amino acids in both Siberian larch and Sitka spruce (Fig. 3).

Verification of transcripts with LRRs by OmicsBox

The functional annotation by InterProScan did not reveal the presence of other functional domains in these sequences (Fig. 4). ARC- and LRRs-containing sequences in Siberian larch could be resistance

genes because they include P-loop NTPase and LRRs families. Sequences containing only ARC might be incomplete or truncated in both Sitka spruce and Siberian larch transcriptomes because they included only P-loop NTPase and some parts that were classified as gene families connected with pathogen resistance and playing binding function (Fig. 5). ARC domain-containing sequences mostly consisted of NB-ARC. The LRR-containing sequences in both species mostly were assigned to the LRR superfamily by InterProScan.

Based on the blast data, most of the transcripts with LRR and ARC in both species matched mostly homologous sequences of other conifers (Figs S6 and S7). Among all other best matches were also homologous sequences representing *Amborella trichopoda*, *Nelumbo nucifera*, *Prosopis alba*, *Glycine max*, *Glycine soja*, *Nymphaea colorata*, *Ceratodon purpureus*, and *Selaginella moellendorffii*.

Discussion

The REs classified in the Siberian larch genome were typical for other studied other studied conifers. Among them LTRs, represented by Gypsy and Copia elements, were the most common. A large amount of LINE elements observed in the Siberian larch genome was specific for larch genome and distinguishes larch from other conifers [2]. In other conifers LTRs mainly prevailed over non-LTR and the number of LINE elements was substantially lower (Fig. 2 in [2]).

RE identified in this study cover ~ 39% of the entire genome of larch, which is significantly less compared to other conifers. The reason for this can be due to fragmented assembly and/or smaller genome size. It is possible that we found only fragments of full-size repeats. The bulk of repeats of different families are rather short, less than 1 Kbp (Figs S1-S5).

The largest number of the LRR-containing transcripts were detected using the models LRR-1, LRR-4, LRR-6 and LRR-8. The LRR-4 family was the most frequent with more than 86% of all LRR transcripts, as well as it accounted for 98% of all LRR domains in the Sitka spruce transcriptome based on the autumn bud sample and used merely for comparison.

According to Pfam data, sequences for the LRR-4, LRR-6 and LRR-8 families were distributed almost evenly among the green plants kingdom, *Metozoa* and uncategorised *Eukaryota* (http://pfam.xfam.org/family/PF12799#tabview=tab7, http://pfam.xfam.org/family/PF13516#tabview=tab7, http://pfam.xfam.org/family/PF13855#tabview=tab7). This may indicate non-specificity of larch and spruce LRRs. The kingdom of green plants made the largest contribution to the LRR-1 family (http://pfam.xfam.org/family/PF00560#tabview=tab7).

OmicsBox functional annotations confirmed the presence of domains ARC and LRRs in the identified transcript sequences. Only several sequences contained both ARC and LRR domains. The exact function of the sequences that contained either ARC or LRRs domains is uncertain because most of them are too short to be R-genes, which also must contain both of these domains as well.

Conclusions

For the first time, REs were identified and classified in the Siberian larch genome. Their analysis helps us better understand organization of such large genomes as in conifers. The Siberian larch genome contains twice as less RE compared to other conifers in the same *Pinaceae* family (39 vs 70–80%), and we think that it might explain why it also has almost twice as smaller genome size (12 vs 18–31 Gbp).

In addition, LRRs and resistance genes were identified and analyzed in the Siberian larch transcriptomes from autumn buds. The LRR-containing sequences are associated with immune response of plants to biotic stress. Their further study will also help us better understand genetic mechanisms of disease resistance in larch and other plants.

Methods

Larch RNA-sequencing and transcriptome assembly

The Siberian larch nuclear genome and autumn bud transcriptome were sequenced and assembled in the Laboratory of Forest Genomics of Siberian Federal University (Krasnoyarsk, Russia) [13]. RNA was isolated from autumn buds from a reference Siberian larch tree using the Qiagen RNeasy Mini Kit (Qiagen, Hilden, Germany). The RNA-seq library was prepared using the TruSeq RNA Sample Preparation Kit v2 (Illumina Inc., San Diego, CA, USA). The PE-sequencing of the obtained library was carried out on the Illumina MiSeq platform using the MiSeq Reagent Kit v2 (300-cycles) (Illumina Inc., San Diego, CA). FastQC software v. 0.11.9 was used to evaluate the quality of the sequencing data. The raw sequencing data were processed using Trimmomatic program v. 0.39 [34] (9-bp headcrop, minimum read quality of Q=23, and minimum read length of 35 bp). SortMeRNA version 4.0.0 [35] was used for ribosomal RNA removal. In addition, Rcorrector [36] was used for sequencing error correction. Finally, *de novo* assembler Trinity v. 2.9.1 [37] was used to assemble the *Larix sibirica* transcriptome.

Search and identification of REs

To assess the relative abundance of previously characterized repeat families, RepeatMasker was used on a whole assembly of Siberian larch (12.3 Gbp).

To search for REs, we used RepeatModeler v.1.0.11, which is based on *de novo* RE detection programs RepeatScout and RECON [38, 39]. Since RepeatScout does not use all scaffolds or contigs for the analysis, but only a part of them that is randomly selected, it was decided to analyse 2 869 scaffolds from a larch genome assembly longer than 100 Kbp (Table 3).

Table 3 Siberian larch genome assembly and scaffolds longer than 100 Kbp selected for REs search and identification using RepeatModeler v.1.0.11

Assemblies	Number of scaffolds	Total length, bp	Max length, bp
Genome assembly, bp	11,325,800	12,342,093,815	354,326
Selected scaffolds longer than 100 Kbp	2,869	360,016,106	354,326

This derived RepeatModeler *de novo* library was augmented by clustering of frequently occurring reads from whole-genome sequencing data. Clusters of reads were assembled with Inchworm from TrinityRnaSeq v2.2.0, which resulted in consensus sequences that should represent highly repeated regions of the larch genome. Unrecognized elements from *de novo* repeat library generated by RepeatModeler and frequently occurring reads were classified by TEclass v2.1.3. This program classifies transposons using the Support Vector Machines (SVM) and LVQ neural network [40]. The combined library, comprising the RepeatModeler derived library classified with TEclass, RepBase library (Edition 2017.01.27), MIPS Repeat Element Database library (Nussbaumer et al. 2013), Custom Plant Repeat Database (Wegrzyn et al. 2013) and Pine Interspersed Repeats Resource library PIER v1.0 (Wegrzyn et al. 2013; Neale et al. 2014) was used for sequence similarity search.

Leucine-rich repeats (LRRs)

LRRs were searched in ORFs of the transcripts of Siberian larch and Sitka spruce (NCBI GenBank accession number GACG00000000.1) autumn buds. ORFs were identified using Transdecoder v.5.5.0 (https://github.com/TransDecoder). These ORF of the transcript sequences were scanned by HMMER 3.2.1 [41] 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 LRR models were obtained from the Pfam 32.0 database and belong to the LRR clan (ID CL0022). Also LRR clan contains families LRR-10, LRR-11, LRR-12, FNIP, DUF285, Recep_L_domain, and TTSSLRR, but they were excluded because they represent bacteria, animals, and myxomycetes [42]. Statistics of ORFs in transcriptome assemblies are presented in Table 4.

Table 4 ORFs identified in the Siberian larch and Sitka spruce transcriptomes studied in autumn buds

Species	Number	Total length, bp	Maximum length, bp	N50, bp	N90, bp
Siberian larch	22,116	4,315,585	1,980	207	110
Sitka spruce	10,106	1,827,092	706	192	116

A search for NBS R-genes (NB-ARC; obtained from the Pfam 32.0 database: PF00931) was additionally performed to check if some sequences with LRRs belong to R-genes.

The computing was mostly performed using a 96-core server with symmetric parallel multiprocessing (IBM x3950 X6) and 3 TB of RAM. The computing cluster also included an IBM dx360 M4 hybrid computational server with two NVIDIA Tesla K20 GPUs, as well as an IBM Storwize V3700 48Tb storage

subsystem. The cluster runs on SuSe Linux Enterprise Server 11 with installed parallel file system IBM GPFS, monitoring system Ganglia and Torque batch processing system.

Gene ontology (GO) analysis

The OmicsBox [43] was used for BLASTing, GO mapping, annotation and statistical analysis. Gene ontology (GO) terms associated with the obtained BLAST results were extracted, and evaluated GO annotation was obtained. The annotation step was reduced because graph construction was not possible with sequences extracted from total sequence. Enzyme codes were inferred by mapping with equivalent GOs, while InterPro motifs were directly queried at the InterProScan web service. The GO annotation was visualized by reconstructing the structure of the GO relationships and pathways.

Abbreviations

BLAST: Basic Local Alignment Search Tool; bp: Base pair; CPU: Central processing unit; DNA: Deoxyribonucleic acid; LINE: Long interspersed nuclear element; LTR: Long terminal repeat; Mbp: Million base pair; N50: a weighted median statistic such that 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value; N90: a weighted median statistic such that 90% of the entire assembly is contained in contigs or scaffolds equal in length to or larger than this value; PE-sequencing: Paired-end sequencing; RE: Repetitive element; RNA: Ribonucleic acid; SINE: Short interspersed nuclear element; TE: Transposable element; Gbp: Giga base pair; GO: Gene ontology; LRR: Leucine-rich repeat; NBS-LRR: Nucleotide-binding site–leucine-rich repeat.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work including the study and collection, analysis and interpretation of data, and writing the manuscript was supported by research grant № 14.Y26.31.0004 from the Government of the Russian Federation. The funding agencies played no role in the design of the study and collection material,

analysis and interpretation of data, and in writing the manuscript. Publication cost has been funded by the Open Access Publication Funds of the University of Göttingen.

Authors' contributions

KVK, KAM & MGS designed the study. KVK & NVO administered the project. NVO carried out most of the sequencing. KAM, VSA, VVB, EIB & VVS carried out bioinformatics analysis. VVS & DAK provided computer support. KAM & KVK drafted the manuscript. VSA, VVB, EIB & VVS contributed to analysis and interpretation of data and revised the paper. All authors read and approved the final manuscript.

Acknowledgments

We acknowledge support by the German Research Foundation (DFG) and the Open Access Publication Funds of the University of Göttingen.

Availability of data and materials

The Siberian larch nuclear genome assembly is available in the NCBI GenBank with accession number GCA_004151065.1. The Transcriptome Shotgun Assembly project is available at DDBJ/EMBL/GenBank under the accession number GIXH00000000 (which is currently placed on hold pending manuscript acceptance, but it is available for reviewers on https://hpccloud.sfu-kras.ru/owncloud/index.php/s/YjRgEJf3KCgyhbK).

References

1. Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, et al. Early genome duplications in conifers and other seed plants. Sci Adv. 2015;1:e1501084.
2. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, et al. The Norway spruce genome sequence and conifer genome evolution. Nature. 2013;497:579–84.
3. Mosca E, Cruz F, Gómez-Garrido J, Bianco L, Rellstab C, Brodbeck S, et al. A reference genome sequence for the European Silver fir (*Abies alba* Mill.): A community-generated genomic resource. G3 Genes Genomes Genet. 2019;9:2039–49.
4. Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 2014;15:R59.
5. Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, et al. Sequencing and assembly of the 22-Gb loblolly pine genome. Genetics. 2014;196:875–90.
6. Zimin AV, Stevens KA, Crepeau MW, Puiu D, Wegrzyn JL, Yorke JA, et al. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. GigaScience. 2017;6(1):giw016.

7. Wegrzyn JL, Liechty JD, Stevens KA, Wu L-S, Loopstra CA, Vasquez-Gross HA, et al. Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. Genetics. 2014;196:891–909.
8. Stevens KA, Wegrzyn JL, Zimin A, Puiu D, Crepeau M, Cardeno C, et al. Sequence of the sugar pine megagenome. Genetics. 2016;204:1613–26.
9. Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, et al. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. Bioinformatics. 2013;29:1492–7.
10. Stival Sena J, Giguère I, Boyle B, Rigault P, Birol I, Zuccolo A, 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.
11. Warren RL, Keeling CI, Yuen MMS, Raymond A, Taylor GA, Vandervalk BP, et al. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. Plant J. 2015;83:189–212.
12. Rake AV, Miksche JP, Hall RB, Hansen KM. DNA reassociation kinetics of four conifers. Can J Genet Cytol. 2011;22(1):69–79.
13. Kuzmin DA, Feranchuk SI, Sharov VV, Cybin AN, Makolov SV, Putintseva YA, et al. Stepwise large genome assembly approach: a case of Siberian larch (*Larix sibirica* Ledeb). BMC Bioinformatics. 2019;20:37.
14. Southworth J, Grace CA, Marron AO, Fatima N, Carr M. A genomic survey of transposable elements in the choanoflagellate *Salpingoeca rosetta* reveals selection on codon usage. Mob DNA. 2019;10:44.
15. Schlötterer C. Evolutionary dynamics of microsatellite DNA. Chromosoma. 2001;109:571–2.
16. Oreshkova NV, Putintseva YuA, Sharov VV, Kuzmin DA, Krutovsky KV. Development of microsatellite genetic markers in Siberian larch (*Larix sibirica* Ledeb.) based on the *de novo* whole genome sequencing. Russ J Genet. 2017;53:1194–9.
17. Oreshkova NV, Bondar EI, Putintseva YuA, Sharov VV, Kuzmin DA, Krutovsky KV. 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;55:444–50.
18. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
19. Padeken J, Zeller P, Gasser SM. Repeat DNA in genome organization and stability. Curr Opin Genet Dev. 2015;31:12–9.
20. Deragon J-M, Zhang X. Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. Syst Biol. 2006;55:949–56.
21. Kobe B, Kajava AV. The leucine-rich repeat as a protein recognition motif. Curr Opin Struct Biol. 2001;11:725–32.

22. Jones JDG, Dangl JL. The plant immune system. Nature. 2006;444:323–9.
23. Song H, Guo Z, Hu X, Qian L, Miao F, Zhang X, et al. Evolutionary balance between LRR domain loss and young NBS–LRR genes production governs disease resistance in *Arachis hypogaea* cv. Tifrunner. BMC Genomics. 2019;20:844.
24. Schaper E, Anisimova M. The evolution and function of protein tandem repeats in plants. New Phytol. 2015;206:397–410.
25. Maumus F, Quesneville H. Impact and insights from ancient repetitive elements in plant genomes. Curr Opin Plant Biol. 2016;30:41–6.
26. Bire S, Rouleux-Bonnin F. Transposable Elements as Tools for Reshaping the Genome: It Is a Huge World After All! In: Bigot Y, editor. Mobile Genetic Elements: Protocols and Genomic Applications. Totowa, NJ: Humana Press; 2012. p. 1–28. doi:10.1007/978-1-61779-603-6_1.
27. Piégu B, Bire S, Arensburger P, Bigot Y. A survey of transposable element classification systems – A call for a fundamental update to meet the challenge of their diversity and complexity. Mol Phylogenet Evol. 2015;86:90–109.
28. Perera D, Magbanua ZV, Thummasuwan S, Mukherjee D, Arick M, Chouvarine P, et al. Exploring the loblolly pine (*Pinus taeda* L.) genome by BAC sequencing and Cot analysis. Gene. 2018;663:165–77.
29. Neale DB, McGuire PE, Wheeler NC, Stevens KA, Crepeau MW, Cardeno C, et al. The Douglas-Fir genome sequence reveals specialization of the photosynthetic apparatus in Pinaceae. G3 GenesGenomesGenetics. 2017;7:3157–67.
30. Civáň P, Švec M, Hauptvogel P. On the coevolution of transposable elements and plant genomes. J Bot. 2011:893546. https://doi.org/10.1155/2011/893546.
31. Magbanua ZV, Ozkan S, Bartlett BD, Chouvarine P, Saski CA, Liston A, et al. 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.
32. Wegrzyn JL, Lin BY, Zieve JJ, Dougherty WM, Martínez-García PJ, Koriabine M, et al. Insights into the loblolly pine genome: characterization of bac and fosmid sequences. PLOS ONE. 2013;8:e72439.
33. McHale L, Tan X, Koehl P, Michelmore RW. Plant NBS-LRR proteins: adaptable guards. Genome Biol. 2006;7:212.
34. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
35. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28:3211–7.
36. Song L, Florea L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaScience. 2015;4(1):s13742–015–0089–y.
37. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.

38. Bao Z, Eddy SR. Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12:1269–76.
39. Price AL, Jones NC, Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinformatics. 2005;21(Suppl. 1):i351–8.
40. Abrusán G, Grundmann N, DeMester L, Makalowski W. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics. 2009;25:1329–30.
41. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013;41:e121–e121.
42. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47:D427–32.
43. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A: High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008;36(10):3420-35.

Figures

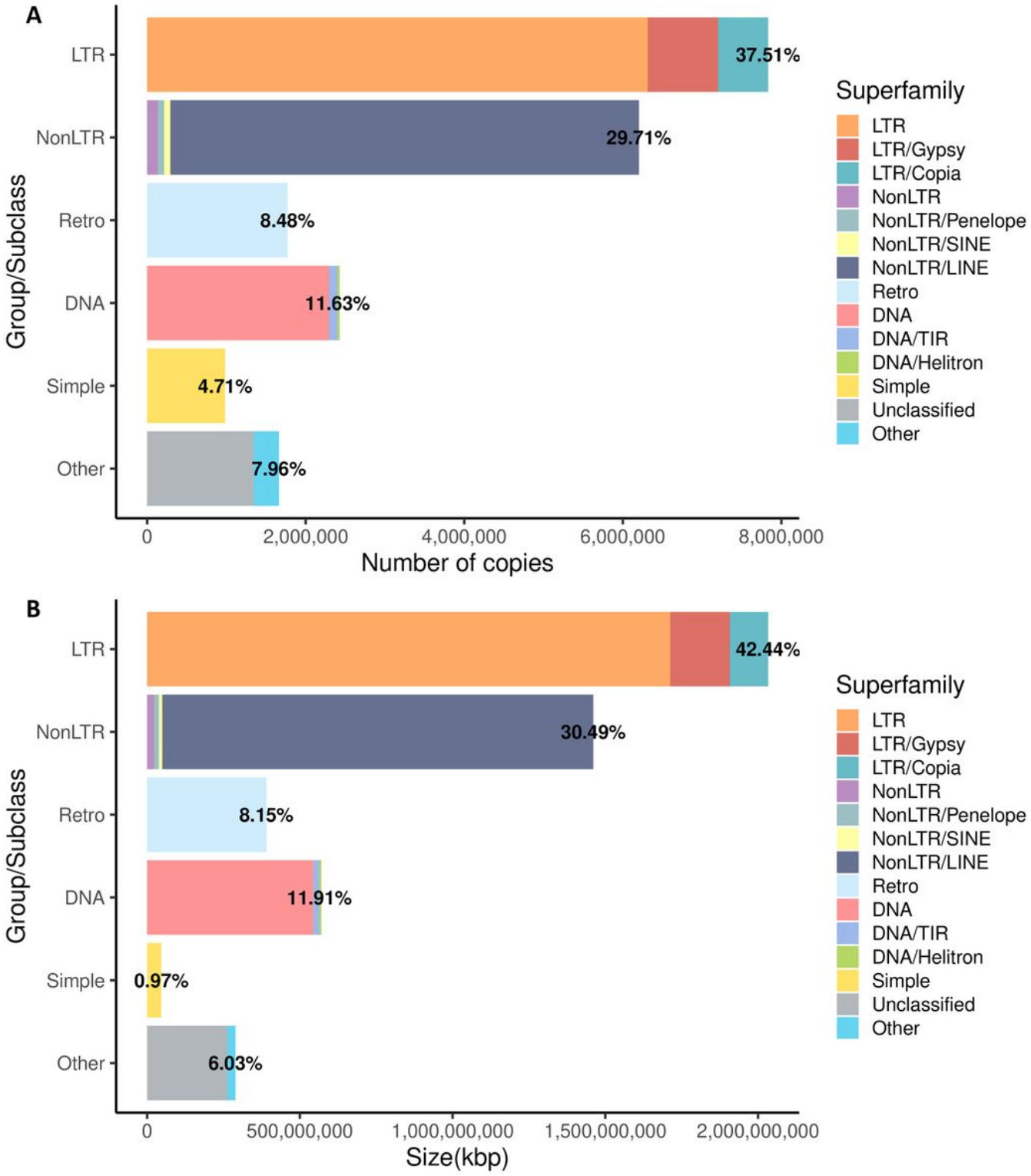

Figure 1

The distribution of the main RE families and groups in the genome of Siberian larch. A - the number of copies of the repeat groups; B - the size of all repeats by groups

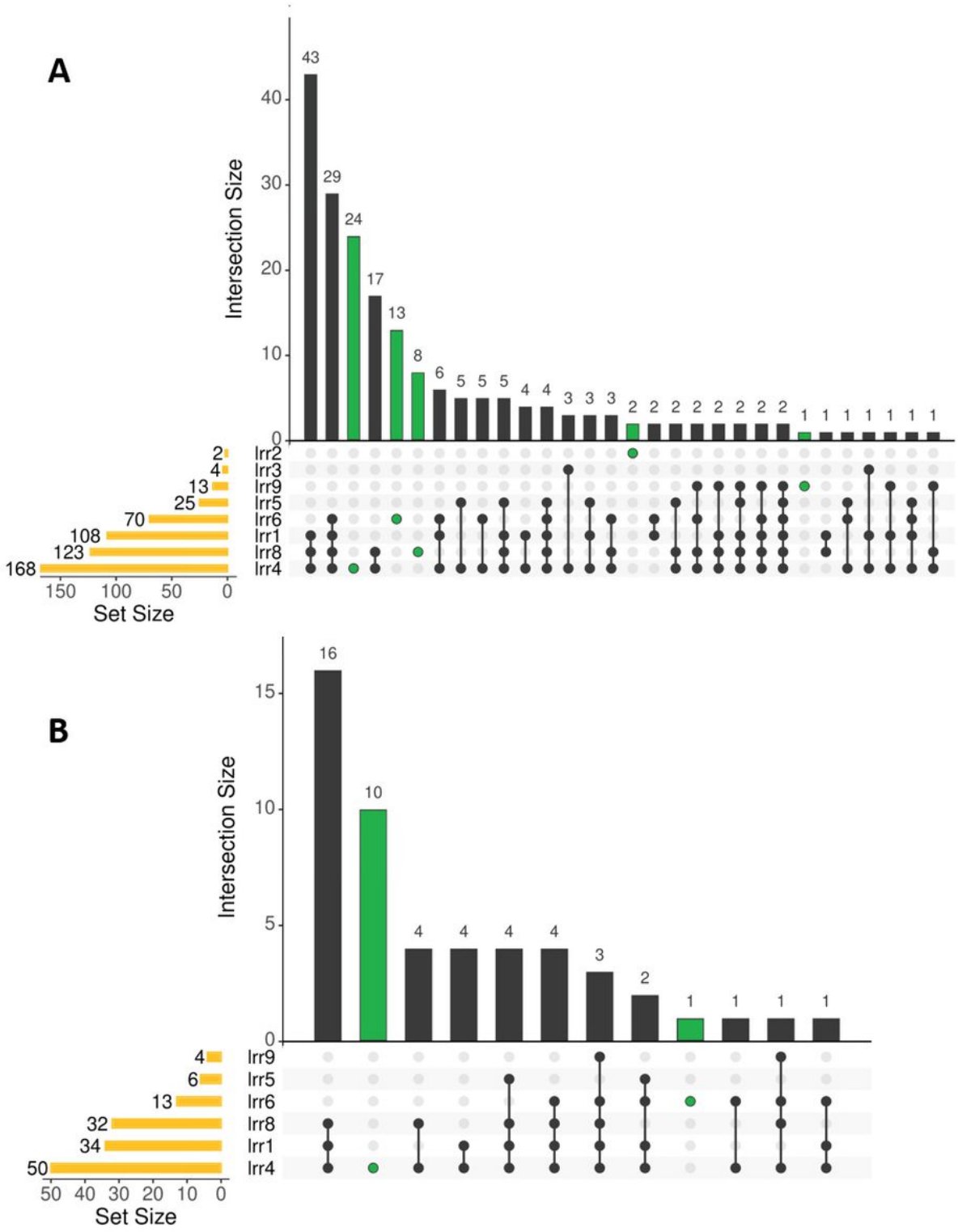

Figure 2

Transcripts with LRRs found using nine LRR families (lrr1-lrr9) in the Siberian larch (A) and Sitka spruce (B) transcriptomes (for example, 43 transcripts were found in the Siberian larch transcriptomes by each of the LRR-1, LRR-4 and LRR-8 families). Transcripts containing LRR that was found by only one of the LRR families are highlighted in green

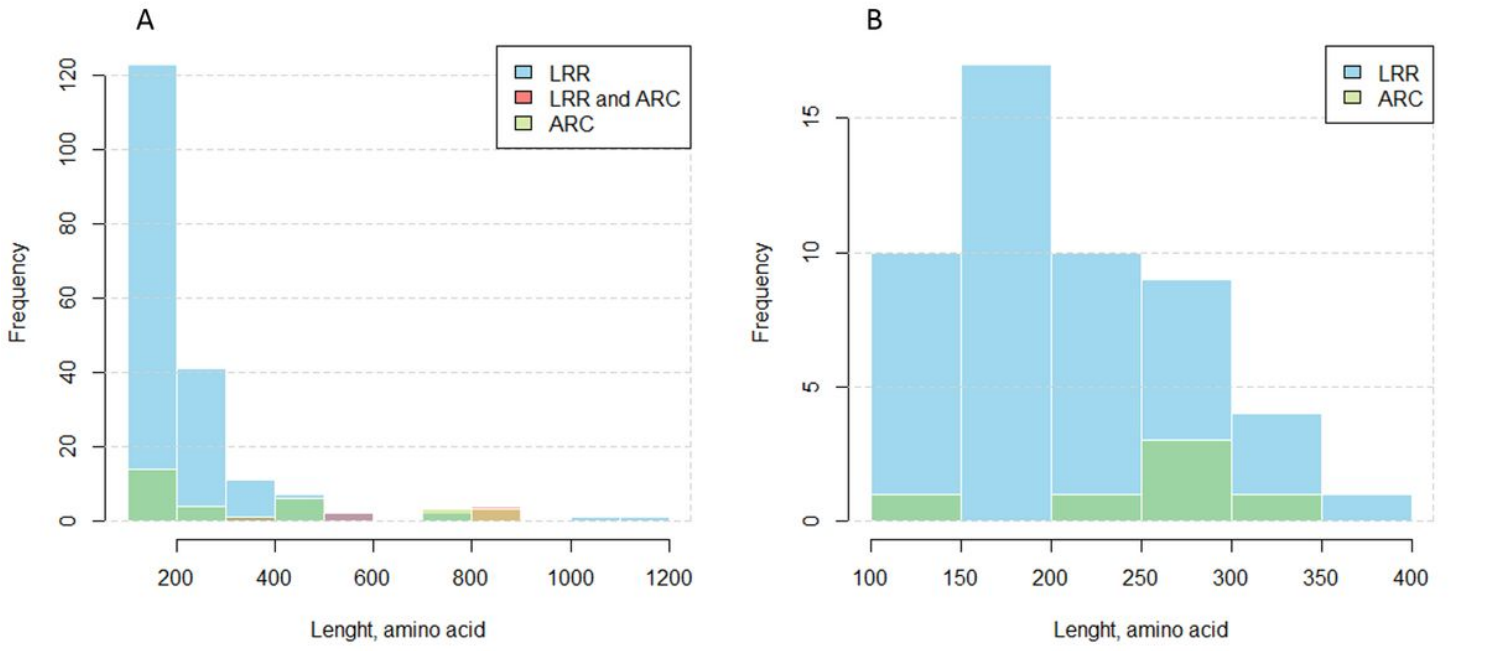

Figure 3

The distribution of the lengths of the amino acids sequences of the putative R-genes in Siberian larch (A) and Sitka spruce (B). LRR — transcripts containing the LRR domain, LRR and ARC — transcripts containing the LRR and ARC domains, ARC — transcripts containing the ARC domain

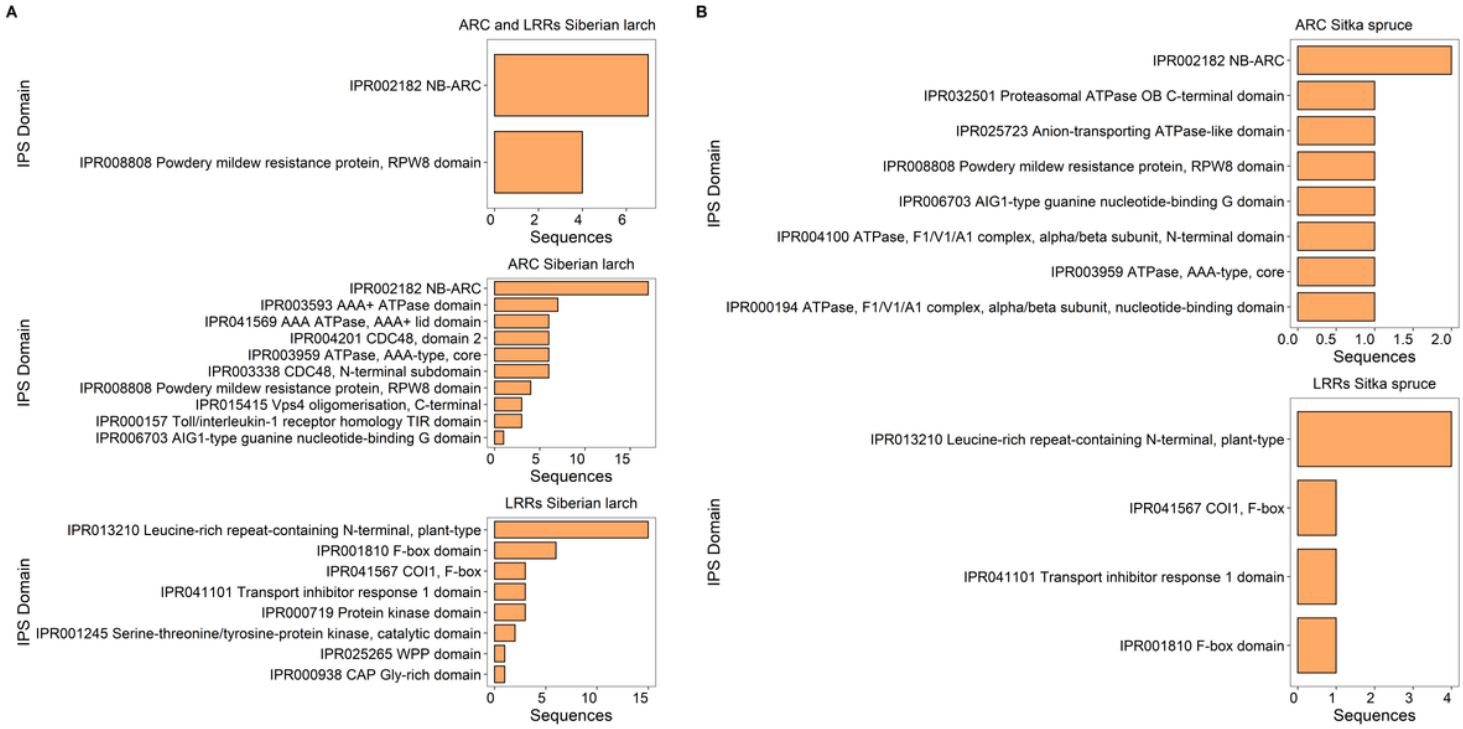

Figure 4

Distribution of the InterProScan domains identified in the Siberian larch (A) and Sitka spruce (B) sequences; ARC and LRR — transcripts containing the ARC and LRR domains; ARC — transcripts containing the ARC domain, LRRs — transcripts containing the LRR domain

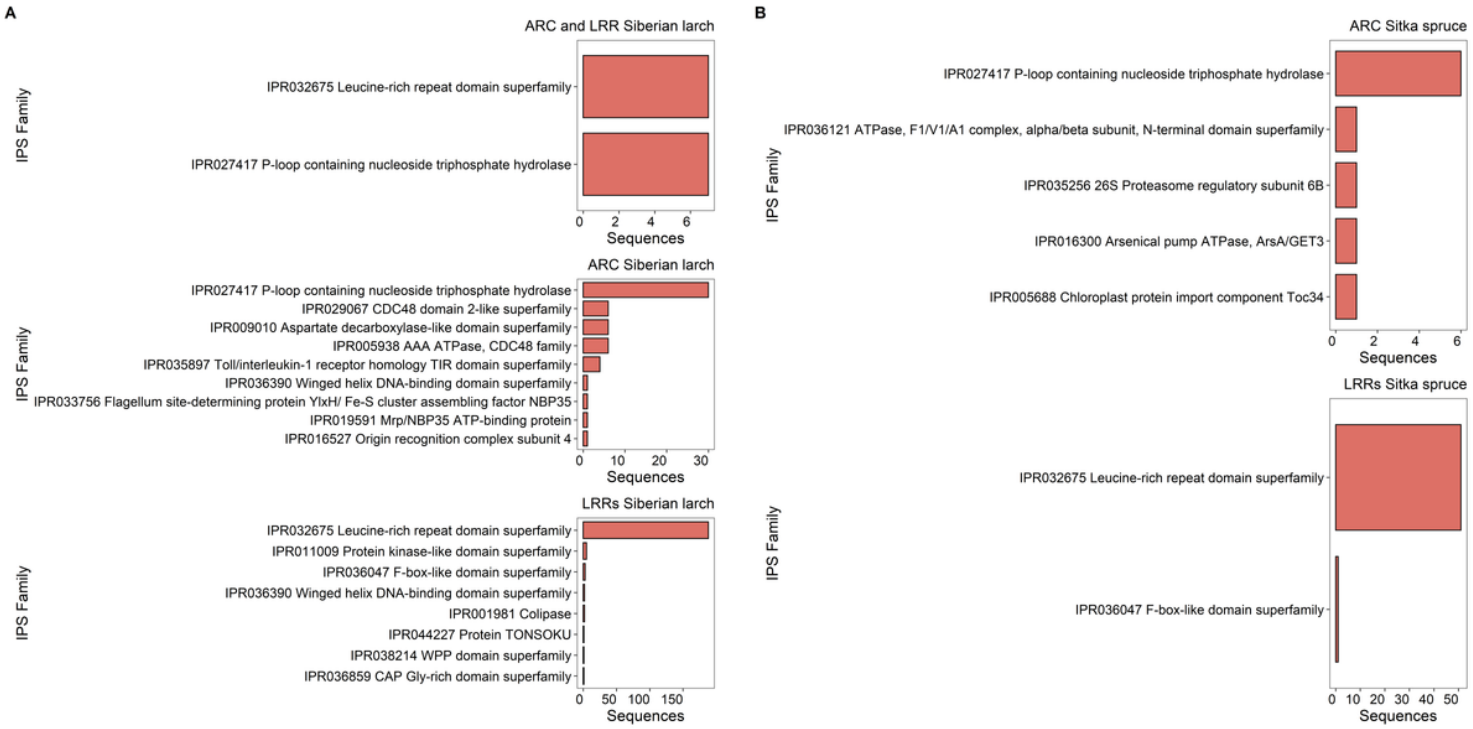

Figure 5

Distribution of the InterProScan families identified in the Siberian larch (A) and Sitka spruce (B) sequences. ARC and LRR — transcripts containing the ARC and LRR domains; ARC — transcripts containing the ARC domain, LRRs — transcripts containing the LRR domain

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MiroshnikovaetalRElarixsibgenomeSupplInf.docx