Annotation of Siberian larch genome draft assembly

Eugeniya I. Bondar*, Nataliya V. Oreshkova, Vadim V. Sharov, Dmitry A. Kuzmin, Sergey I. Feranchuk, Alexander N. Cybin, Tatiana V. Tatarinova, Konstantin V. Krutovsky

*bondar.ev@ksc.krasn.ru

Introduction

Siberian larch is a hardwood cold-resistant tree well-known for its rot-resistant high-quality timber. Growing together with Scots pine, Siberian spruce, and Siberian cedar it forms extensive coniferous forests occupying about 40% of Russia's forested area. The main objective of this work was to provide a verified and reliable annotation for the draft genome assembly of Siberian larch, which will further promote functional and genome-wide association studies.

Table 1. Summary of genome assembly statistics

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

Methods

The MAKER2 annotation pipeline was used for automated gene annotation. RepeatMasker with custom *de novo* repeat library generated by RepeatModeler was used for masking repeated genomic regions. The transcriptome assemblies from five tissue types assembled using the TrinityRnaSeq package were used as species-specific RNA-seq evidence.

Gene prediction was done using AUGUSTUS, which was iteratively trained on the verified set of annotated transcripts, preliminarily assembled with TopHat and Cufflinks. Functional annotation was performed using Blast2GO within the OmixBox Platform (**Fig. 1**).

- Iterative predictor training (Augustus)
- Transcriptome (RNA) data from 4 tissue types
- MAKER annotation was performed on a supercomputer segment with 16GB of RAM per server. It took 22 days, 448 cores at 2.3 GHz/core and 896 GB of RAM with the average processor load of about 61%.

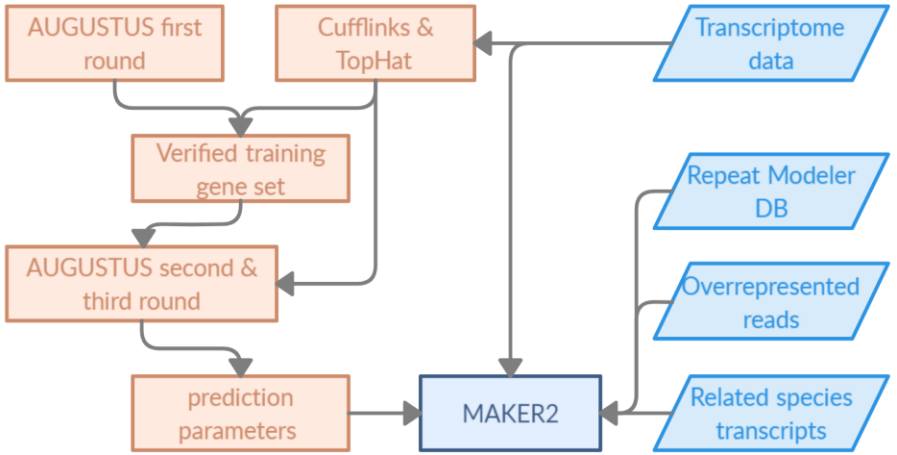

Figure 1. Structural annotation workflow

Laboratory of Forest Genomics, Siberian Federal University, Russia
Laboratory of Genomic Research and Biotechnology, FRC KSC SB RAS, Russia
Department of Forest Genetics and Forest Tree Breeding, Georg-August University of Göttingen, Germany
Department of Biology, University of La Verne, La Verne, CA, USA

Federal Research Center "Krasnoyarsk Science Center of the Siberian Branch of the Russian Academy of Sciences"

Annotation of Siberian larch genome draft assembly

Eugeniya I. Bondar*, Nataliya V. Oreshkova, Vadim V. Sharov, Dmitry A. Kuzmin, Sergey I. Feranchuk, Alexander N. Cybin, Tatiana V. Tatarinova, Konstantin V. Krutovsky

*bondar.ev@ksc.krasn.ru

Results

Repeat analysis and masking

Using a combined repeat library RepeatMasker identified a total of 20.9 million repeating elements with a total size of 4.8 Gbp, which comprises about 39% of the 12.4 Gbp genome assembly. Among classified repeats, Class I retrotransposons of LINE, I, Gypsy, and Copia superfamilies were the most abundant, with LINE elements also having the longest average size and taking the largest fraction of the genome (**Fig. 2**).

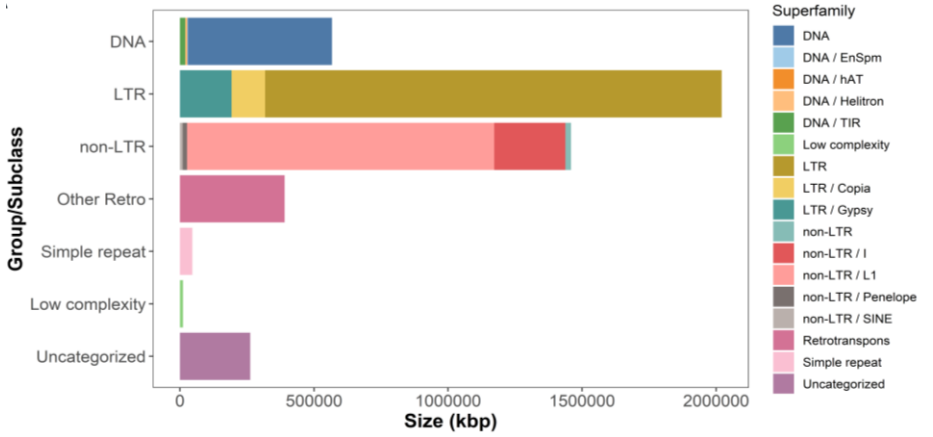

Figure 2. Relative size of the repetitive sequence content of the Siberian larch genome annotated using RepeatMasker and combined library.

- Total number of repeats:

  20.9 million = **4.8** Gbp = **39 %** of the 12.4 Gbp genome
- Class I retrotransposons cover 31.47 % of assembly size
- Class II DNA transposons cover 4.6 % of assembly size

Structural annotation using MAKER2

In total, 50,163 gene models were obtained consisting of 151,838 exons and 101,675 introns. When comparing the top 10% longest intron lengths, larch introns were comparable in length with those of *A. thaliana* and *P. glauca*, although, the longest larch introns were far shorter than those in other conifer species or in the repeat-rich genomes of *Populus thichocarpa, Vitis vinifera* and *Zea mays* (**Fig. 3**). Repeat content in the intron sequences was lower than in the entire genome, 12.9% of intron space are covered by transposable elements. GO category assignment based on InterProScan domains identification and BLAST homology search yielded 39,253 gene models (77%) with at least one assigned GO term.

- 50,163 gene models predicted
- Gene space completeness 45.3 % (BUSCO)
- Introns make up for the 42% (35.7 Mbp) of the gene space and 0.29% of the 12.3 Gbp genome assembly
- 39,253 gene models (77%) with at least one assigned GO term

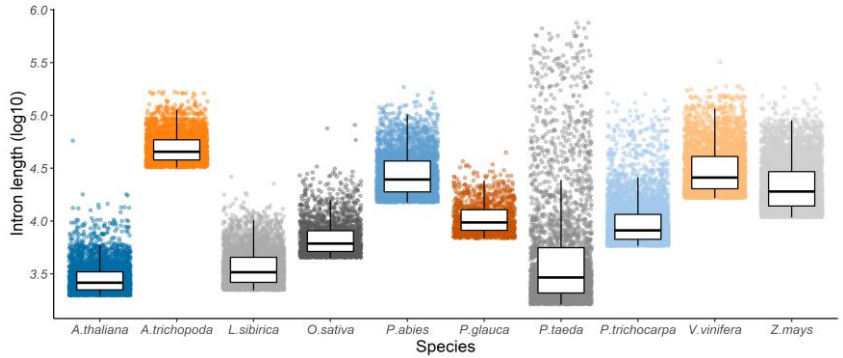

Figure 3. Top 10% of longest introns in plant species

Table 2. Summary of Gene Ontology term assignment

		L. sibirica
Total gene count		50,163
Annotated genes		39,253
GO terms	Biological process	28,944
	Cellular component	18,613
	Molecular function	33,711

Acknowledgments

This study was supported by research grant № 14.Y26.31.0004 from the Russian Federation Government for the "Genomics of the key boreal forest conifer species and their major phytopathogens in the Russian Federation" project.

Laboratory of Forest Genomics, Siberian Federal University, Russia
Laboratory of Genomic Research and Biotechnology, FRC KSC SB RAS, Russia
Department of Forest Genetics and Forest Tree Breeding, Georg-August University of Göttingen, Germany
Department of Biology, University of La Verne, La Verne, CA, USA

СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ | SIBERIAN FEDERAL UNIVERSITY

Federal Research Center "Krasnoyarsk Science Center of the Siberian Branch of the Russian Academy of Sciences"