Resource Article

Siberian larch (*Larix sibirica* Ledeb.) chloroplast genome and development of polymorphic chloroplast markers

Eugeniya I. Bondar[1], Yuliya A. Putintseva[1], Nataliya V. Oreshkova[1,2], Konstantin V. Krutovsky [1,3,4,5]

[1]Laboratory of Forest Genomics, Genome Research and Education Center, Siberian Federal University, Krasnoyarsk, Russian Federation

[2]Laboratory of Forest Genetics and Selection, V.N. Sukachev Institute of Forest, Siberian Branch of Russian Academy of Sciences, Krasnoyarsk, Russian Federation

[3]Department of Forest Genetics and Forest Tree Breeding, Georg-August University of Göttingen, Göttingen, Germany

[4]Laboratory of Population Genetics, Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia

[5]Department of Ecosystem Science and Management, Texas A&M University, College Station, Texas, USA

Abstract

Background: The main objectives of this study were sequencing, assembling and annotation of chloroplast genome of one of the main Siberian boreal forest tree conifer species Siberian larch (*Larix sibirica* Ledeb.) and detection of polymorphic genetic markers – microsatellite loci or simple sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs).

Results: We used data of the whole genome sequencing of three Siberian larch trees from different regions - Urals, Krasnoyarsk, and Khakassia, respectively. Sequence reads were obtained using the Illumina HiSeq2000 in the Laboratory of Forest Genomics at the Genome Research and Education Center in the Siberian Federal University. The assembling was done using the Bowtie2 mapping program and the SPAdes genomic assembler. The genome annotation was performed using the RAST service. We used the SciRoKo program for the SSRs search, and the Bowtie2 and UGENE programs for the SNPs detection. Length of the assembled chloroplast genome was 122,561 bp, which is similar to 122,474 bp in the closely related European larch (*Larix decidua* Mill.). As a result of annotation and comparison of the data with

existing data available only for three larch species - *L. decidua*, *L. potaninii var. chinensis* (complete genome 122,492 bp) and *L. occidentalis* (partial genome of 119,680 bp), we identified 110 genes, 34 of which represented tRNA, 4 rRNA and 72 protein-coding genes. In total, 13 SNPs were detected; two of them were in the *tRNA-Arg* and *Cell division protein FtsH* genes, respectively.

Conclusions: The complete chloroplast genome sequence was obtained for Siberian larch for the first time. The reference complete chloroplast genomes, such as one described here would greatly help in the chloroplast resequencing and search for additional genetic markers using population samples. The results of this research will be useful for further phylogenetic and gene flow studies in conifers.I

Keywords: Chloroplast genome, *Larix sibirica*, Sequencing, Siberian larch, SNPs, SSRs

Correspondence: Konstantin V. Krutovsky, Department of Forest Genetics and Forest Tree Breeding, Georg-August University of Göttingen, Büsgenweg 2, D-37077 Göttingen, Germany; Fax:+49 (551)398367; E-mail: konstantin.krutovsky@forst.uni-goettingen.de; ORCID**:** http://orcid.org/0000-0002-8819-7084

Background

Chloroplast genome in conifers, including larch species [1] has a unique, strictly paternal inheritance via pollen, unlike angiosperms, where it has a maternal inheritance via seeds [2]. It allows tracing paternal gene flow and lineages separately from maternal (mitochondrial genes) and bi-parental (nuclear genes) ones. Therefore, chloroplast DNA sequences are the most important source of genetic markers to study distribution of paternal genes and paternally based molecular phylogenetic relationships in conifers.

Larch species, as well as many other conifer species are the main boreal forest tree species, which comprise ~30% of the world's forested lands [3]. Boreal forests play very important ecological role, but are also affected by the global climate change. On one hand, they suffer

now from more frequent and drastic droughts, but on the other hand their area is expanding in the northern regions, and their tree line is moving towards north creating an ecotone, a highly dynamic transition area [4]. It is important to know how much of paternal associated gene flow by pollen contributes into establishing this zone compared to the maternal and bi-parental contributions by seeds. Such studies require chloroplast markers. Next generation sequencing (NGS) technique allows whole chloroplast genome sequencing in multiple individuals and makes a search for the molecular genetic markers more efficient. For instance, Parks *et al.* [5] nearly completely sequenced chloroplast genomes in 37 pine species using NGS. They found significant amount of variation (especially in two loci *ycf1* and *ycf2*) that provided them with additional data for inferring intrageneric phylogeny of genus *Pinus*.

Whole chloroplast genome comparison across different species and genera allows also studying organelle evolution and how it is associated with speciation and dispersal. Complete chloroplast genome sequences are available in NCBI Genbank for multiple plant species, including conifers. However, most of them represent the *Pinus* genus, and only three chloroplast genomes are available for the *Larix* genus: complete for European (*Larix decidua* Mill.; AB501189.1) and Chinese (*L. potaninii var. chinensis* Beissn.; KX808508) larch and partial for Western larch (*L. occidentalis* Nutt.; FJ899578.1).

Variation in the chloroplast genome is effectively used in phylogenetics at different levels. It allowed discriminating different subgenera and genera. For instance, Cronn *et al.* [6] compared chloroplast genome sequences of seven pine and one spruce species and found three regions that have deletions corresponded to the subgenera specific deletions in three genes: *ycf12* (78 bp at the nucleotide starting position 51051), *psaM* (93 bp at position 51442) and *ndh*I (371 bp at position 101988), respectively. These are common deletions in the chloroplast genome in the pine species of the subgenus *Strobus* (i.e., *P. gerardiana, P. krempfii, P.*

lambertiana, P. longaeva, P. monophylla, P. nelsonii, P. koraiensis); the corresponding genes were present in the subgenus *Pinus* (*P. contorta, P. ponderosa, P. thunbergii*) and in spruce *Picea sitchensis* [6].

Variation in the chloroplast genome can be also effectively used in discriminating different populations of the same species. For instance, Whittall *et al.* [7] demonstrated a strong differentiation between mainland and island populations of Torrey pine (*Pinus torreyana*) based on 5 SNPs found in the entire chloroplast genome of 120 Kbp.

Methods

We used data of the whole genome sequencing of three Siberian larch trees generated by Illumina HiSeq2000 [8]. DNA samples were isolated from needles and haploid callus of three Siberian larch trees, representing different regions in Russia – Ural Mountains, Krasnoyarsk Region and Khakassia Republic, respectively. *Larix decidua* Mill. [9] and *L. occidentalis* Nutt. [5] chloroplast genomes were used as reference (NCBI Genbank accession numbers AB501189.1 and FJ899578.1, respectively). We did not use the chloroplast genome of *L. potaninii* [10] as a reference, because it was assembled by using the chloroplast genome of *L. decidua* (NC_016058; [9]) as a reference, but we used it in the comparative analysis. The paired-end (PE) and mate-pair (MP) libraries with fragment sizes of 400-500 bp (Ural and Krasnoyarsk trees) and 300-400 bp (Khakassia tree), respectively, were used for sequencing via 2 × 100 cycles by Illumina HiSeq2000.

The sequence reads were mapped to the reference chloroplast genomes using the Bowtie2 software [11], which is good for mapping short sequence reads to medium-sized and large genomes. This software implements an algorithm to derive FM-index based on Burrows-Wheeler Transform. The SPAdes genome assembler has been used to assemble the larch

genome, which implements the De Bruijn graph approach [12]. The Rapid Annotation service with Subsystem Technology (RAST) has been used for annotation [13].

The first step in our assembly procedure consisted of mapping short reads to the available chloroplast genome references of *L. decidua* and *L. occidentalis* using the Bowtie2 software. Then, the aligned reads were assembled by SPAdes. Obtained contigs were aligned again on the reference of *L. decidua* using BLAST. At the third step, the selected contigs were verified to get the "trusted" status. Then, the assembly was carried out using SPAdes. The final step of the assembly was the scaffolding, which was done using the generated contigs and MP reads using the SSPACE program [14].

Considering a well-known fact that chloroplast organelle originated from cyanobacteria, and that, therefore, chloroplast genes are still very similar to the bacterial ones, the RAST service, which was designed for annotation of bacterial and archaeal genomes, was used for the larch genome annotation. The annotation obtained by the RAST contained both confirmed known genes and predicted genes, potentially coding hypothetical proteins. In order to clarify the roles of these hypothetical coding regions our annotation was compared with annotations of two closely related species *L. decidua* and *L. occidentalis*, respectively. In addition, some fragments of the genome have been also selectively aligned with BLAST. Sites of hypothetical proteins confirmed by BLAST were identified and recorded.

SNPs were search using the Bowtie2 and UGENE [15] software (option *Call Variants with SAMtools*). The search was done across the three above mentioned trees. First, reads of Urals and Khakassian trees were mapped to the finally assembled genome of the Krasnoyarsk tree. The resulting *sam*-file together with the assembled genome was used by the UGENE program to search for SNPs.

Results

The total length of the final Siberian larch chloroplast genome assembly was 122,560 bp, which is very close to 122,474 bp in closely related European larch (*Larix decidua*). The annotation through the comparison with available data for *L. decidua* and *L. occidentalis* identified 110 genes, from which 34 represented tRNA genes, 4 rRNA and 72 protein-coding genes. In three trees 13 SNPs were detected. Two of them were found in the coding regions of the *tRNA-Arg* and *Cell division protein ycf2* genes.

We used available software, such as Bowtie2, BLAST and SPAdes to assemble chloroplast genome using reads generated in the whole genome sequencing of Siberian larch project. We used SSPACE for scaffolding and the RAST service for annotation of obtained chloroplast genome. We developed a procedure that allowed us to successfully extract chloroplast genome specific reads and then assemble and annotate the resulting sequences. We identified and verified 110 coding regions representing 38 RNA and 72 protein genes, which is equal to the number of genes in chloroplast sequences of *L. decidua and L. potaninii* and close to 105 genes in a partial chloroplast genome sequence of *L. occidentalis*. A gene map of the genome was generated using OGDRAW [16] and presented in Fig. 1. Search for SNPs using UGENE revealed a relatively small number of SNPs (Fig. 2; Additional file 1), but it is only preliminary data based on a limited sample size.

Discussion

The chloroplast genome variation in most plants is often limited due to a relatively low frequency of mutations in this organelle. For example, the mutation rate of the chloroplast genome in pines is approximately $0.2\text{-}0.4\times10^{-9}$ synonymous substitutions per nucleotide per year [17, 18]. However, with an average length of 120-160 Kbp and 130 genes chloroplast

genomes are sufficiently large and complex and include structural and point mutations that reflect population differentiation and evolutionary divergence [6].

Unlike angiosperms, conifer chloroplast DNA (cpDNA) lacks large inverted repeats (IR), but contains dispersed repetitive DNA that is associated with structural rearrangements. In addition to large dispersed repeated sequences, conifer cpDNA also possess a number of small repeats. It contains variable numbers of tandem repeats of 124 to 150 bp in size, which are associated with polymorphic rearranged region near *trnK-psbA*, where the *psbA* gene has been duplicated [19].

Most variation in the chloroplast genome is associated with microsatellite loci [20, 21]. However, these markers have a too high mutation rate that can lead to the incorrect phylogenetic inferences [22-24]. SNPs could be better markers for phylogenetic inferences, and comparative complete chloroplast genome studies are needed to discover these markers. The reference complete chloroplast genomes, such as one described here would greatly help in the chloroplast resequencing and search for SNPs using population samples.

Conclusions

The complete chloroplast genome sequence was obtained for Siberian larch for the first time. Annotation and comparison of the obtained data with data available only for two other larch species helped us identify and verify 110 coding regions representing 38 RNA and 72 protein genes. Total 13 SNPs were detected; two of them were in the coding regions of the genome. The results of this research will be useful for further phylogenetic and gene flow studies in conifers.

Additional files

Additional file 1: excel file representing the Siberian larch chloroplast genetic variant data.

Acknowledgements

We thank Dr. M. Sadovsky for discussing results and his valuable comments, Dr. I. N. Tretyakova and M. E. Pak for providing with callus culture, Dr. V. L. Semerikov and K. Deich for help with sampling, Dr. A. Ibe – with laboratory analysis, and Dr. D. Kuzmin and S. Makolov - with computing. This study was supported by a research grant No. 14.Y26.31.0004 from the Government of the Russian Federation.

Funding

The presented study was a part of the project "Genomic studies of major boreal coniferous forest tree species and their most dangerous pathogens in the Russian Federation" funded by the Government of the Russian Federation (grant № 14.Y26.31.0004).

Availability of data and materials

The annotated chloroplast genome of *L. sibirica* has been deposited in the NCBI GenBank with the accession number MF795085.

Authors' contributions

EIB and YAP assembled and annotated the chloroplast genome, analysed the data and wrote the draft paper, NVO prepared and sequenced DNA libraries, KVK designed and coordinated research and wrote the paper.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Szmidt AE, Aldén T, Hällgren J-E. Paternal inheritance of chloroplast DNA in *Larix*. Plant Mol Biol. 1987;9:59-64. https://doi.org/10.1007/BF00017987
2. Hipkins VD, Krutovskii KV, Strauss SH. Organelle genomes in conifers: structure, evolution, and diversity. Forest Genet. 1995;1:179-89.
3. Brandt JP, Flannigan MD, Maynard DG, Thompson ID, Volney WJA. An introduction to Canada's boreal zone: ecosystem processes, health, sustainability, and environmental issues. Environm Rev. 2013;21:207–26. https://doi.org/10.1139/er-2013-0040
4. Johnson JS, Gaddis KD, Cairns DM, Konganti K, Krutovsky KV. Landscape genomic insights into the historic migration of mountain hemlock in response to Holocene climate change. Am J Bot. 2017;104(3):439-50. https://doi.org/10.3732/ajb.1600262
5. Parks M, Cronn R, Liston A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 2009;7:84. https://doi.org/10.1186/1741-7007-7-84
6. Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 2008;36(19):e122. https://doi.org/10.1093/nar/gkn502
7. Whittall JB, Syring J, Parks M, Buenrostro J, Dick C, Liston A. Cronn R. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. Mol Ecol. 2010;19:100–14. https://doi.org/10.1111/j.1365-294X.2009.04474.x
8. Krutovsky KV, Oreshkova NV, Putintseva YuA, Ibe AA, Deich KO, Shilkina EA. Preliminary results of *de novo* whole genome sequencing of Siberian larch (*Larix sibirica* Ledeb.) and Siberian stone pine (*Pinus sibirica* Du Tour.). Siberian J Forest Sci. 2014;1:79-83.
9. Wu CS, Lin CP, Hsu CY, Wang RJ, Chaw SM. Comparative chloroplast genomes of Pinaceae: insights into the mechanism of diversified genomic organizations. Genome Biol Evol. 2011; 3:309-19. https://doi.org/10.1093/gbe/evr026
10. Han K, Li J, Zeng S, Liu Z. Complete chloroplast genome sequence of Chinese larch (*Larix potaninii* var. *chinensis*), an endangered conifer endemic to China. Conservation Genet Res. 2017;9:111–3. https://doi.org/10.1007/s12686-016-0633-9
11. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9:357-9. https://doi.org/10.1038/nmeth.1923

12. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455-77. https://doi.org/10.1089/cmb.2012.0021
13. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014;42(Database issue):D206-D214. https://doi.org/10.1093/nar/gkt1226
14. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27:578-9. https://doi.org/10.1093/bioinformatics/btq683
15. Okonechnikov K, Golosova O, Fursov M, the UGENE team. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28:1166-7. https://doi.org/10.1093/bioinformatics/bts091.
16. Lohse M, Drechsel O, Kahlau S, Bock R. OrganellarGenomeDRAW – a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 2013;41:575–81. https://doi.org/10.1093/nar/gkt289
17. Willyard A, Syring J, Gernandt DS, Liston A, Cronn R. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. Mol Biol Evol. 2007;24:90-101. https://doi.org/10.1093/molbev/msl131
18. Gernandt DS, Magallon S, Lopez GG, Flores OZ, Willyard A, Liston A. Use of simultaneous analyses to guide fossil-based calibrations of Pinaceae phylogeny. Int J Plant Sci. 2008;169:1086–99. https://doi.org/10.1086/590472
19. Lidholm J, Gustafsson P. A three-step model for the rearrangement of the chloroplast *trnK-psbA* region of the gymnosperm *Pinus contorta*. Nucleic Acids Res. 1991. 19:2881–87.
20. Provan J, Soranzo N, Wilson NJ, Goldstein D B, Powell W A. A low mutation rate for chloroplast microsatellites. Genetics. 1999;153:943–47.
21. Ebert D, Peakall R. Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. Mol Ecol Res. 2009;9:673–90. https://doi.org/10.1111/j.1755-0998.2008.02319.x

22. Afzal-Rafii Z, Dodd A. Chloroplast DNA supports a hypothesis of glacial refugia over postglacial recolonization in disjunct populations of black pine (*Pinus nigra*) in Western Europe. Mol Ecol. 2007;16:723–36. https://doi.org/10.1111/j.1365-294X.2006.03183.x

23. Höhn M, Gugerli F, Abran P, Bisztray G, Buonamici A, Cseke K., et al. Variation in the chloroplast DNA of Swiss stone pine (*Pinus cembra* L.) reflects contrasting post-glacial history of populations from the Carpathians and the Alps. J Biogeography. 2009;36:1798–1806. https://doi.org/10.1111/j.1365-2699.2009.02122.x

24. Moreno-Letelier A, Piñero D. Phylogeographic structure of *Pinus strobiformis* Engelm. across the Chihuahuan Desert filter-barrier. J Biogeography. 2009;36:121–31. https://doi.org/10.1111/j.1365-2699.2008.02001.x

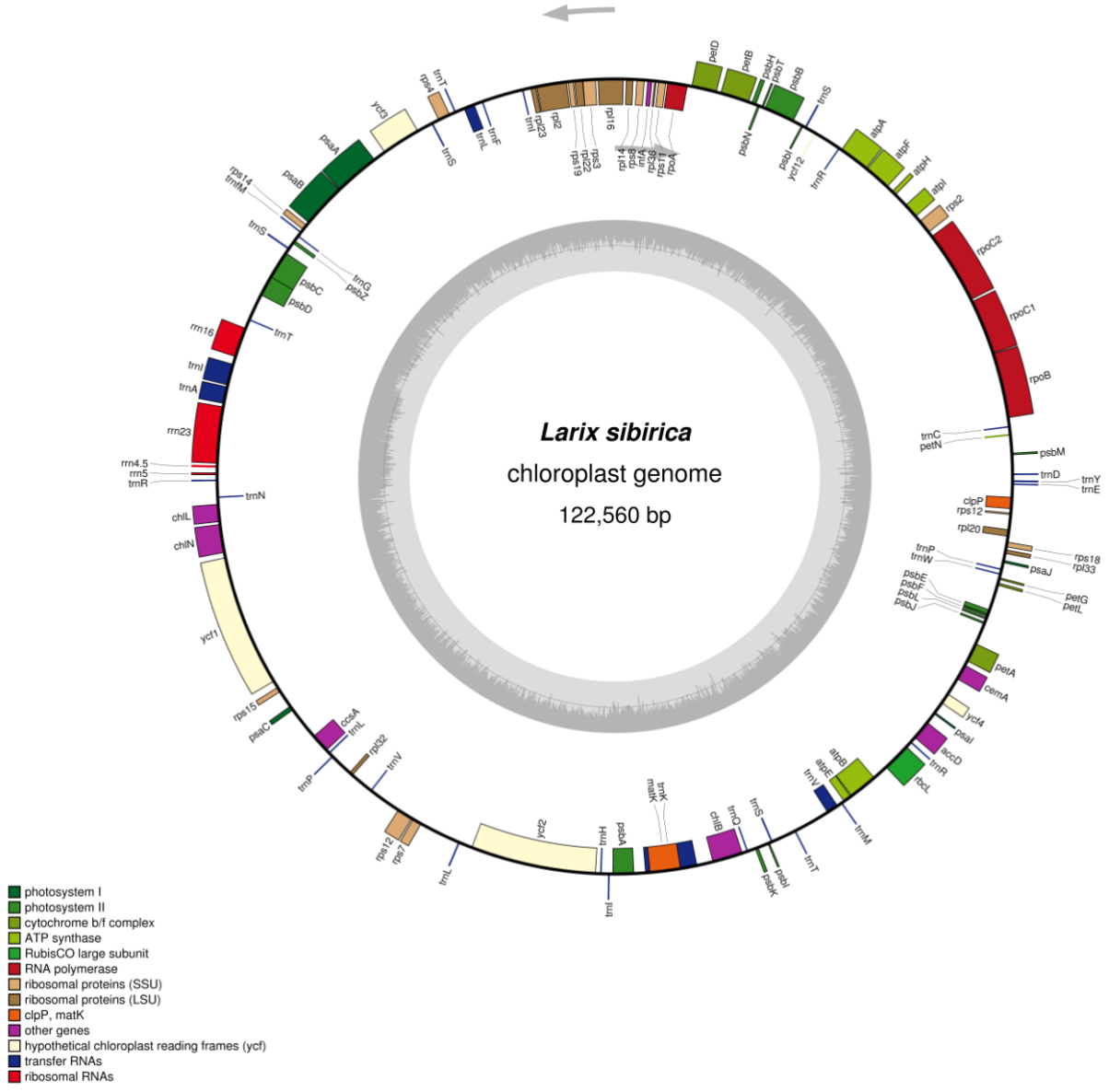

Fig. 1 Gene map of the *Larix sibirica* chloroplast genome. Genes belonging to different functional groups are color-coded. The dark and light grey in the inner circle represents the GC and AT content, respectively

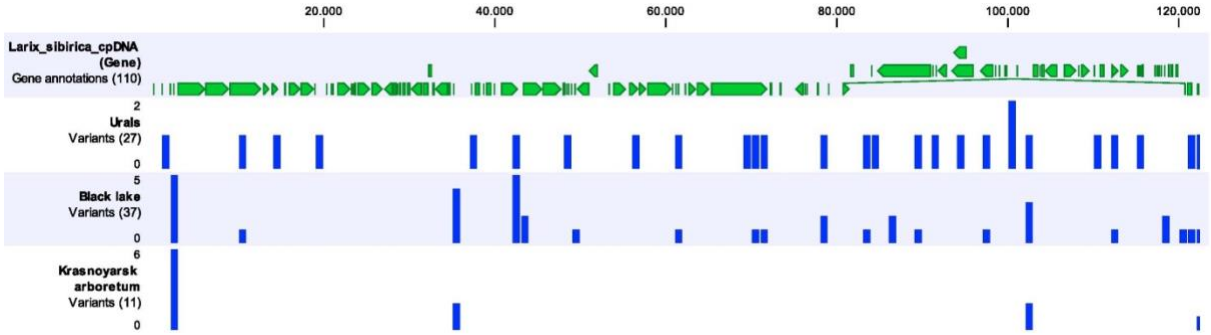

Fig. 2 Variation detected in the *Larix sibirica* chloroplast genome (see also Additional file 1)