The complete chloroplast genome sequencing data of *Juniperus sabina* L. (Cupressaceae Bartl.) from Kazakhstan

Shyryn Almerekova\textsuperscript{a}, Moldir Yermagambetova\textsuperscript{a,b}, Saule Abugalieva\textsuperscript{a,b}, Yerlan Turuspekov\textsuperscript{a,b,*}

\textsuperscript{a}Molecular Genetics Laboratory, Institute of Plant Biology and Biotechnology, Almaty, Kazakhstan

\textsuperscript{b}Faculty of Biology and Biotechnology, Al-Farabi Kazakh National University, Almaty, Kazakhstan

**ABSTRACT**

The genus *Juniperus* L. (Cupressaceae Bartl.) is consisting of about 75 species that are divided into sections *Caryocedrus* Endl., *Sabina* (Miller) Spach, and *Juniperus* (syn: sect. *Oxycedrus* Spach). *Juniperus sabina* L. from section *Sabina* is an important shrub for the maintenance of the ecosystem in mountainous regions and a source of medicinal compounds. The species is monoecious, rarely dioecious, and distributed in Europe, Central Asia, China, and Mongolia. The goal of the present study was to sequence and reconstruct the complete chloroplast genome of *J. sabina*. De novo chloroplast (cp) genome assembly for *J. sabina* was conducted using Illumina paired-end reads. The assembled cp genome size is 127,646 bp in length and has a typical circular DNA molecule. The genome encodes 118 genes, including 82 protein-coding genes, 32 tRNA genes, and 4 rRNA genes, the overall GC content is 34.36%. The complete cp genome nucleotide sequence of *J. sabina* was deposited to the NCBI (National Center for Biotechnology Information) under accession number OL467323. The raw data in fastq format was deposited.
Specifications Table

| Subject                      | Plant Science: General |
|------------------------------|------------------------|
| Specific subject area        | Genomics, Forest ecosystem, Environmental science |
| Type of data                 | Figure, table          |
| How the data were acquired   | Chloroplast genome of *J. sabina* was sequenced using Illumina Novaseq 6000 (San Diego, USA) platform. |
| Data format                  | Raw sequencing reads (fastq) and analyzed data (fasta) |
| Description of data collection | The young leaves were collected from Ile Alatau Mountain (North Tian Shan), Kazakhstan. Total genomic DNA was extracted using the CTAB protocol [1]. Genomic DNA was used for sequencing by an Illumina Novaseq 6000 (Illumina, San Diego, CA, USA) platform in Macrogen (Macrogen, Seoul, Korea). Chloroplast genome assembly was performed using SPAdes 3.13.0 [3]. The *J. sabina* complete chloroplast genome was annotated using Prokka [4] and the circular map was illustrated using OGDRAW [5]. |
| Data source location         | Institution: Institute of Plant Biology and Biotechnology |
| City: Almaty                 | Country: Kazakhstan    |
| Data accessibility           | Repository name: National Center for Biotechnology Information |
| Data identification number   | OL467323 (Nucleotide) |
|                             | PRJNA767752 (Bioproject) |
|                             | SRR21515769 (Sequence Read Archive) |
| Direct URL to data           | https://www.ncbi.nlm.nih.gov/nuccore/OL467323 (Nucleotide) |
|                             | https://www.ncbi.nlm.nih.gov/sra/PRJNA767752 (SRA) |

Value of the Data

- The data provides information on the whole chloroplast genome data of *J. sabina* that can be useful in phylogenetic comparisons with the related species and in clarification of the molecular taxonomy of the genus.
- The obtained data will be a benefit for researchers in molecular botany as a valuable resource for evaluating taxonomy, population structure, and evolution of the genus *Juniperus*.
- The data is expected to be useful in detecting the clarification of the chloroplast (cp) genome structure of the genus and in phylogenomic studies of *Juniperus* species.
- The cp data will be important in the development of molecular genetic studies related to the use of next-generation sequencing technologies for the wild flora of Kazakhstan.

1. Data Description

The circular gene map of *J. sabina* chloroplast (cp) genome is presented in Fig. 1. Chloroplast genome sequencing was carried out using the platform Illumina and sequencing by synthesis. In total 4,01 Gb, Illumina Novaseq 6000 reads data were generated. The length of the assembled chloroplast genome of *J. sabina* is 127,646 bp, with GC content of 34.36%. The sequencing quality values Q20 and Q30 were 98.28% and 93.43%, respectively. The whole cp genome of *J. sabina* contained 118 genes, including 82 protein-coding genes, 32 tRNA genes, and 4 rRNA
The assembled chloroplast genome sequence (OL467323) and raw data in fastq formats (SRR21515769) were deposited in the NCBI nucleotide database and SRA database, respectively. To elucidate the phylogenetic position of *J. sabina*, *matK* and *rbcL* sequences from *J. sabina* and 19 samples of *Juniperus* cp genome from NCBI were aligned and the phylogenetic tree was reconstructed. The phylogenetic tree demonstrated that the *matK* and *rbcL* nucleotide sequences grouped sample of *J. sabina* from Kazakhstan into the same clade as *J. sabina* from NCBI (Fig. 2). The tree is clearly separated species in sections Juniperus and Sabina.

The cp genome of the *J. sabina* has 15 intron-containing genes, 14 of them (9 protein-coding and 5 tRNA genes) have a single intron and *ycf3* contain two introns (Table 1).
Fig. 2. Neighbor joining of the concatenated phylogenetic tree based on matK and rbcl. nucleotide sequences. Denotes species analyzed in this study.
Table 1
List of genes in the sequenced *J. sabina* chloroplast genome.

| Category                        | Group of genes                     | Name of genes                                                                 |
|---------------------------------|------------------------------------|-------------------------------------------------------------------------------|
| Self-replication                | Transfer RNA                       | trnA-UGC, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnFM-CAU, trnG-GCC,      |
|                                 |                                    | trnH-GUG, trnI-CAU (x2), trnI-GAU', trnK-UUU', trnL-CAA, trnL-UAA, trnL-UAG,   |
|                                 |                                    | trnM-CAU, trnN-GUU, trnP-UGG, trnQ-UUG (x2), trnP-AGG, trnR-UCU,            |
|                                 |                                    | trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC, trnV-UAC',      |
|                                 |                                    | trnW-CCA, trnY-GUA                                                           |
|                                 | Ribosomal RNA                       | rps12, rps14, rps15, rps19                                                  |
|                                 | Small subunit of ribosome           | rpl2, rps3, rps4, rps7, rps8, rps11, rps12*, rps14, rps15, rps18, rps19   |
|                                 | Large subunit of ribosome           | rpl14, rpl16*, rpl22, rpl23*, rpl32, rpl33, rpl36                           |
|                                 | DNA-dependent RNA polymerase        | rpoA, rpoB, rpoC1*, rpoC2*                                                   |
|                                 | Translational initiation factor     | infA                                                                        |
| Genes for photosynthesis        | Rubisco                            | rbcL                                                                        |
|                                 | Photosystem I                      | psaA, psaB, psaC, psaL, psaJ, psaM                                          |
|                                 | Photosystem II                     | psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbl, psbK, psbL, psbl,     |
|                                 | ATP synthase                       | psbK, psbL, psbM, psbN, psbl, psbZ                                         |
|                                 | Subunits of cytochrome             | petA, petB, petD, petG, petL, petN                                          |
|                                 | Chlorophyll biosynthesis            | chlB, chlL, chlN                                                            |
|                                 | NADH dehydrogenase                 | ndhA*, ndhB*, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhl, ndhj, ndhk         |
| Other genes                     | Maturase                           | matK                                                                        |
|                                 | Protease                           | clpP                                                                         |
|                                 | Envelope membrane protein          | cemA                                                                        |
|                                 | Subunit of acetyl-CoA               | accD                                                                        |
|                                 | C-type cytochrome synthesis gene   | ccsA                                                                        |
| Genes of unknown function       | Conserved open reading frames       | ycf1, ycf2, ycf3**, ycf4                                                     |

* genes with one intron,
** genes with two introns

2. Experimental Design, Materials and Methods

The young leaves were collected from the Kim Asar gorge of Ile Alatau Mountain (Kazakhstan, North Tian Shan) and dried in silica gel. GPS coordinates of collected samples: 43.162500 N, 77.093889 E, altitude 2264 m. Total genomic DNA was extracted from approximately 70 mg of dried leaves using CTAB (cetyl trimethylammonium bromide) protocol [1]. The concentration and quality of the isolated DNA were measured using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA) and 1% agarose gel electrophoresis.

Genomic DNA was sent for sequencing to Macrogen (Macrogen, Seoul, Korea). The TruSeq Nano DNA kit (Illumina, USA) was used to construct an Illumina paired-end library. Paired-end sequencing (2 × 150 bp) was run on an Illumina Novaseq 6000 (Illumina, San Diego, CA, USA). Raw data were filtered by quality for accurate assembly. The reads of which 90% of bases had phred score 20 or higher were used for assembly. After quality filtering, poly-G trimming was conducted using fastp 0.19.4 with qualified quality phred option as 10 and unqualified percent limit as 50. Adapter trimming was conducted using Trimmomatic v0.38 [2]. Trimmed reads were used for chloroplast genome assembly by using SPAdes 3.13.0 [3] assembler approach. The complete genome contigs were combined into one contig by joining overlapping DNA segments of each contig. The *J. sabina* complete chloroplast genome was annotated using Prokka [4] and the circular map was illustrated using OGDRAW [5]. The scientific names for Juniperus sections were given according to Adams [6].
The Neighbor-Joining method [7] with 1000 bootstrap values was used for the reconstruction of the phylogenetic tree of Juniperus species in MEGA 11 [8]. The nucleotide sequences of matK and rbcL genes from cp genome of J. sabina and 19 samples of Juniperus from NCBI were downloaded for the reconstruction of the phylogenetic tree.

Ethics Statement

This study did not require an official ethics review.

CRediT Author Statement

Shyryn Almerekova: Sample collection, Investigation, Data curation, Writing an original draft, Fund acquisition Moldir Yermagambetova: Sample collection, Investigation, Data curation Saule Abugalieva: Conceptualization, Methodology, Writing- Reviewing Yerlan Turuspekov: Sample collection, Conceptualization, Methodology, Writing- Reviewing & Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Juniperus sabina chloroplast, complete genome (Original data) (NCBI).

Acknowledgments

This research was funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP09259027)

References

[1] J.J. Doyle, J.L. Doyle, A rapid DNA isolation procedure for small quantities of fresh leaf tissue, Phytochemistry Bulletin 19 (1987) 11–15.
[2] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120, doi: 10.1093/bioinformatics/btu170.
[3] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A.V. Pyshkin, A.V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, P.A. Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477, doi: 10.1089/cmb.2012.0021.
[4] T. Seemann, Prokka: rapid prokaryotic genome annotation, Bioinformatics 30 (2014) 2068–2069, doi: 10.1093/bioinformatics/btu153.
[5] M. Lohse, O. Drechsel, S. Kahlau, R. Bock, OrganellarGenomeDRAW - a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets, Nucleic Acids Res. 41 (2013) W575–W581, doi: 10.1093/nar/gkt289.
[6] R.P. Adams, The junipers of the world: The genus Juniperus, Trafford Publishing, Bloomington, 2014 fourth ed.
[7] N. Saitou, M. Nei, The neighbor-joining method: A new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425.
[8] K. Tamura, G. Stecher, S. Kumar, MEGA11: Molecular Evolutionary Genetics Analysis version 11, Mol. Biol. Evol. 38 (2021) 3022–3027.