Building a reference Transcriptome for *Juniperus squamata* (Cupressaceae) based on Single-molecule real-time sequencing

Yufei Wang
Sichuan University School of Life Sciences
Siyu Xie
Sichuan University School of Life Sciences
Jialiang Li
College of Life Sciences, Sichuan University
Jieshi Tang
College of Life Science, Sichuan University
Tsam Ju
College of Life Sciences, Sichuan University
Kangshan Mao (maokangshan@163.com)
Sichuan University https://orcid.org/0000-0002-0071-1844

Data note

Keywords: Juniperus squamata, single-molecule real-time sequencing, simple sequence repeats

Posted Date: May 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-548946/v1

Abstract

Objectives

Cupressaceae is the second largest family of coniferous trees (Coniferopsida) with important economic and ecological values. However, like other conifers, the members of Cupressaceae have extremely large genome (>8 gigabytes), which limited the researches of these taxa. A high-quality transcriptome is an important resource for gene discovery and annotation for non-model organisms.

Data description

Juniperus squamata, a tetraploid species which is widely distributed in Asian mountains, represents the largest genus, *Juniperus*, in Cupressaceae. Single-molecule real-time sequencing was used to obtain full-length transcriptome of *Juniperus squamata*. The full-length transcriptome was corrected with Illumina RNA-seq data from the same individual. A total of 47, 860 non-redundant full-length transcripts, N50 of which was 2, 839, were obtained. Simple sequence repeats for *Juniperus squamata* were also identified. This data presents the first comprehensive transcriptome characterization of Cupressaceae species, and provides an important reference for researches on the genomic evolutionary history of Cupressaceae plants and even conifers in the future.

Objective

Compared with other plant groups, the genome analysis of coniferous species lags behind because of their larger genome [1, 2]. At present, only a few genome-wide datasets are available, such as *Sequoiadendron gigantea*, *Pinus taeda* L. and *Picea abies* [3-5]. Whole genome sequencing of conifers is prohibitively expensive for large genome sizes, and it also produces datasets which are inconvenient to analyze. In contrast, analyses on the dataset produced by transcriptome sequencing is much easier, and it is a convenient and cost-effective method for sequencing coding sequences of complex genomes.

Juniperus squamata is an evergreen shrub of the family Cupressaceae reaching 1-3 meters tall, with brownish-gray bark [6]. It is found in mountains from southwestern China to northeastern Afghanistan, with separate populations east to Fujian and north to western Gansu in China [7]. This tetraploid species is not only of great value to gardening but also of enormous ecological values in subalpine and alpine shrubland ecosystems in Asian mountains. However, very limited genomic information is available for this species. Hence the objective of this work is to generate full-length transcriptome sequences for *Juniperussquamata*. Considering the importance of simple sequence repeats (SSRs) to plant population genetic analysis, we also developed SSRs for this species [8, 9]. The full-length transcriptome data set of *Juniperus squamata* can provide an important reference for its downstream analysis, such as genomic basis of environmental adaptation and genome evolution of Cupressaceae and even conifers.

Data Description

Fresh leaves, stems, and strobiles of one *Juniperus squamata* individual were collected from Kangding, Sichuan Province, China. For each tissue, the short paired reads were sequenced by Illumina platform. We also mixed the samples of each tissue and generated the long reads by the PacBio Sequel platform. Total RNA of the samples was isolated using the Plant RNA kit (Omega bio-Tech., USA) and then treated with RNase-free DNase I (NEB) to remove DNA. RNA degradation and contamination were monitored on 1% agarose gels and RNA purity was checked using the NanoPhotometer® spectrophotometer (IMPLEN, CA, USA). RNA concentration was measured using Qubit® RNA Assay Kit in Qubit® 2.0 Fluorometer (Life Technologies, CA, USA). RNA integrity was assessed using the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). The Single-molecule real-time (SMRT) bell library was constructed with the Pacific Biosciences DNA Template Prep Kit 2.0 and SMRT sequencing was then performed on the Pacific Bioscience Sequel System. The sample used for Illumina sequencing was harvested using the same methods. The library was constructed using Illumina HiSeq X Ten. Adapter clipping and quality filtering of the Illumina raw

reads was done using Trimmomatic version 0.36 [10]. Based on the quality check, the last two base pairs from each read were removed to minimize the overall sequencing error.

The raw full-length transcriptome sequencing data of samples were processed using the SMRT link version 4.0 software (https://www.pacb.com/support/softwaredownloads). Subread BAM files were generate from raw reads, parameters: -minLength 200, -minReadScore 0.75. Circular consensus sequence (CCS) was generated from subread BAM files, parameters: -min_length 50, -max_drop_fraction 0.8, -no_polish TRUE, -min_zscore -9999.0, -min_passes 2, -min_predicted_accuracy 0.8, -max_length 15000. CCS BAM files were output, which were then classified into Full-Length non-chimeric (FLNC) and non-full length (NFL) fasta files by examining the 5' and 3' adapters and the poly(A) tail. Iterative Clustering and Error Correction (ICE) algorithm was utilized to cluster FLNC fasta files to obtain cluster consensus. Quiver from SMRT link (parameters: -hq_uiver_min_accuracy 0.99, -bin_by_primer false, -bin_size_kb 1, -qv_trim_5p 100, -qv_trim_3p 30) were then utilized to polish cluster consensus sequence with NFL fasta files to obtain polished consensus sequence.

To obtain high quality corrected consensus sequence, additional nucleotide errors in polished consensus sequence were corrected using the lllumina RNA-seq data obtained from the same individual with the software LoRDEC version 0.7 [11](parameters: -k 23 -s 3). Any redundancy in corrected consensus sequence was removed by CD-HIT version 4.6.1 [12] (parameters: -c 0.95 -T 6 -G 0 - aL 0.00 -aS 0.99 -AS 30) to obtain final a set of unique transcript isoforms. Benchmarking universal single-copy orthologs (BUSCO) version 3 was used to assess the quality of final transcript isoforms [13]. The summary statistics and length distributions of the PacBio SMART sequencing are shown in Data file 1 (Table S1 and Fig. S1). The results of BUSCO are shown in Data file 1 (Table S2). All three data sets obtained and their NCBI GenBank Accession numbers are listed in Table 1.

MISA version 1.0 was employed to identify SSRs from final unique transcript isoforms of *Juniperus squamata* [14](parameters: definition(unit_size, min_repeats): 1-10 2-6 3-5 4-5 5-5 6-5, interruptions(max_difference_betw-een_2_SSRs): 100). Finally, 57, 393 SSRs were identified which were containing in 42, 273 sequences. The details of SSRs of *Juniperus squamata*, including primer sequences, SSR type, annealing temperature, product size etc., are shown in Data file 2.

Table 1: Overview of data files/sets.

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	*Summary and assessment of the data set*	*MS Word file(.docx)*	*Figshare (10.6084/m9.figshare.14572125)*
Data file 2	*SSRs of Juniperus squamata*	*MS Excel file(.csv)*	*Figshare (10.6084/m9.figshare.14572098)*
Data set 1	*js.fastq.gz*	*fastq (.fastq.gz)*	*NCBI(SRR13966305)*
Data set 2	*juniperus_squamata_final.fastq.gz*	*fastq (.fastq.gz)*	*NCBI(SRR13993906)*
Data set 3	*Juniperus_squamata_final_unique_transcript_isoforms.fastq.gz*	*fastq (.fastq.gz)*	*NCBI(SRR14000623)*

Limitations

There is a shortcoming that we only collected one sample for single-molecule real-time sequencing of transcriptome.

Abbreviations

BUSCO: benchmarking universal single-copy orthologs

CCS: circular consensus sequence

FLNC: full-length non-chimeric

ICE: Iterative Clustering for Error Correction

NFL: non-full length

ROI: reads of insert

SMRT: single-molecule real-time

SSRs: simple sequence repeats

Declarations

Ethics approval and consent to participate

This research received ethics approval and participatory consent.

Consent for publication

The note is approved by all authors for publication.

Availability of data and materials

The data described in this Data note can be freely and openly accessed on ***NCBI*** under ***SRR13966305, SRR13993906*** *and* ***SRR14000623***. Please see table 1 and references ***Data file 1 & 2*** *and* ***Data set 1, 2 & 3*** for details and links to the data.

Competing interests

No conflict of interest exits in the submission of this note.

Funding

National Natural Science Foundation of China (grant numbers U20A2080, 31622015) and Sichuan University (Fundamental Research Funds for the Central Universities, SCU2019D013, SCU2020D003).

Authors' contributions

SX, JL, YJ and KM collected the samples, YW and SX analyzed the data, YW wrote the note. JT, JL, TJ and KM revised the manuscript, KM conceived and designed the program. All authors have read and approved the manuscript.

Acknowledgements

The authors acknowledge financial support by the National Natural Science Foundation of China (grant numbers U20A2080, 31622015) and Sichuan University (Fundamental Research Funds for the Central Universities, SCU2019D013, SCU2020D003).

References

1. De La Torre AR, Birol I, Bousquet J, Ingvarsson PK, Jansson S, Jones SJM, Keeling CI, MacKay J, Nilsson O, Ritland K *et al*: **Insights into Conifer Giga-Genomes**. *Plant Physiology* 2014, **166**(4):1724-1732.
2. Prunier J, Verta JP, MacKay JJ: **Conifer genomics and adaptation: at the crossroads of genetic diversity and genome function**. *New Phytol* 2016, **209**(1):44-62.

3. Lu MM, Krutovsky KV, Loopstra CA: **Predicting Adaptive Genetic Variation of Loblolly Pine (Pinus taeda L.) Populations Under Projected Future Climates Based on Multivariate Models**. *J Hered* 2019, **110**(7):857-865.
4. Scott AD, Zimin AV, Puiu D, Workman R, Britton M, Zaman S, Caballero M, Read AC, Bogdanove AJ, Burns E *et al*: **A Reference Genome Sequence for Giant Sequoia**. *G3 Genes|Genomes|Genetics* 2020, **10**(11):3907-3919.
5. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A *et al*: **The Norway spruce genome sequence and conifer genome evolution**. *Nature* 2013, **497**(7451):579-584.
6. Wu Z, Peter HR, Hong D: **CUPRESSACEAE**. In: *Flora of China*. Edited by Fu L, Yu Y, Aljos F, vol. 4. Saint Louis: Missouri Botanical Garden Press; 1999: 62-77.
7. Adams RP: **Junipers of the World: The Genus Juniperus, 4th Edition**: Trafford Publishing Company; 2014.
8. Vieira MLC, Santini L, Diniz AL, Munhoz CdF: **Microsatellite markers: what they mean and why they are so useful**. *Genetics and Molecular Biology* 2016, **39**:312-328.
9. Zhang Q, Li J, Zhao Y, Korban SS, Han Y: **Evaluation of Genetic Diversity in Chinese Wild Apple Species Along with Apple Cultivars Using SSR Markers**. *Plant Molecular Biology Reporter* 2012, **30**(3):539-546.
10. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
11. Salmela L, Rivals E: **LoRDEC: accurate and efficient long read error correction**. *Bioinformatics* 2014, **30**(24):3506-3514.
12. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**. *Bioinformatics* 2006, **22**(13):1658-1659.
13. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015, **31**(19):3210-3212.
14. Beier S, Thiel T, Münch T, Scholz U, Mascher M: **MISA-web: a web server for microsatellite prediction**. *Bioinformatics* 2017, **33**(16):2583-2585.

Supplementary Files

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

- SSRsofJuniperussquamata.csv
- Summaryandassessmentofthedataset.docx