Data Article

Dataset of genome identification and characterization of microsatellite markers loci in *Atriplex atacamensis* and *Atriplex deserticola*

Francisco Correa, Jorge Pérez-Díaz, Pamela Rojas, Claudia Torres, Manuel Paneque, Boris Sagredo, Adriana Bastías

**A R T I C L E  I N F O**

Article history:
Received 20 May 2019
Received in revised form 7 July 2019
Accepted 8 July 2019
Available online 9 August 2019

**Keywords:**
Atriplex
SSR
Microsatellite
Molecular markers
Simple sequence repeat

**A B S T R A C T**

In this work, we partially sequenced genomes of two *Atriplex* species (*A. deserticola* Phil. and *A. atacamensis* Phil.), using Illumina technology (Hiseq 2500 paired-end system) and *de novo* assembly strategy. Raw data of *A. deserticola* and *A. atacamensis* are available from NCBI-Bioproject, PRJNA495747 and PRJNA495763 accessions, respectively. A total of 127086 and 134984 microsatellite or simple sequence repeat (SSR) markers were identified within *A. deserticola* and *A. atacamensis* genomic DNA, respectively. In addition, predicted putative genes in *A. deserticola* and *A. atacamensis* sequences are also presented in this article.

© 2019 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Data

Raw partial genome sequencing data for A. deserticola and A. atacamensis was produced by de novo sequencing using a HiSeq 2500 System - Illumina. The data was then quality trimmed, filtered and assembled (assembly statistics are present in Tables 1 and 2).

Dinucleotide to hexanucleotide repeat microsatellite sequences were identified for A. deserticola and A. atacamensis (Table 3). However, only SSRs with a repeat motif size ranging from 2 to 8 bp and a length ≥12 bp were considered. This includes dinucleotide repeats >6, trinucleotide repeats >4, and tetra-, penta-, hexa-, hepta- and octanucleotide repeats >3. We analyzed the distribution of A. deserticola and A. atacamensis SSRs data with regard to motif length, type and number of repeats (Tables 4 and 5, Fig. 1). Primer pairs were designed from flanking sequences of di-to octanucleotide microsatellites of A. deserticola and A. atacamensis (S1 and S2 Tables).

Table 1

| Atriplex sp. | Before cleaning | After cleaning | % total reads |
|-------------|-----------------|----------------|--------------|
|             | Total reads     | GC (%)         | Total reads  | GC (%)         |
| A. deserticola | 876,957,994   | 35             | 803,100,062 | 36             | 91             |
| A. atacamensis | 874,405,882    | 35             | 761,570,465 | 35             | 87             |
The contig and singleton *A. deserticola* and *A. atacamensis* genomic sequences were analyzed by AUGUSTUS software [1,2] using *A. thaliana* as a model organism to predict putative genes (Table 6). For functional annotation, the potential coding regions data were analyzed by WEGO [3], leading to consistent gene annotations, gene names, gene products and Gene Ontology (GO) numbers (Fig. 2 and S3 Table).

Table 2
Data on contig measurements that were assembled by SOAPdenovo2 software after high-quality reads.

|                  | *A. deserticola* | *A. atacamensis* |
|------------------|------------------|------------------|
| Contigs          | 274,412          | 302,895          |
| N50              | 1256             | 1229             |
| Count ≥ N50      | 68816            | 76561            |
| Max contig       | 29233            | 28538            |
| Min contig       | 500              | 500              |
| Total length     | 310,278,579      | 338,284,897      |
| Average contig   | 1113             | 1117             |

Table 3
Microsatellite (SSRs) searches of *A. deserticola* and *A. atacamensis* using MicroSAtellite identification.

| Category                        | *A. deserticola* | *A. atacamensis* |
|---------------------------------|------------------|------------------|
| Total number of sequences examined | 274412           | 302895           |
| Total size of examined sequences (bp) | 310,278,579     | 338,284,897     |
| Total number of identified SSRs | 127086           | 134984           |
| Number of SSR containing sequences | 98939            | 104235           |
| Number of sequences containing more than 1 SSR | 35721            | 38381            |

Table 4
Distribution of microsatellites di-to octonucleotide motifs in the assembled genomic DNA of *A. deserticola* and *A. atacamensis*.

| Motif length | N° loci identified | Frequency (%) | Density (SSR/Mb) |
|--------------|--------------------|---------------|------------------|
|              | *A. deserticola*   | *A. atacamensis* | *A. deserticola*   | *A. atacamensis* |
| Di           | 18605              | 19915         | 14.64            | 14.75            | 59.96          | 58.87          |
| Tri          | 41175              | 44645         | 32.40            | 33.07            | 132.71         | 131.98         |
| Tetra        | 38789              | 41160         | 30.52            | 30.49            | 125.02         | 121.67         |
| Penta        | 16298              | 17645         | 12.82            | 13.07            | 52.53          | 52.16          |
| Hexa         | 6448               | 7313          | 5.07             | 5.42             | 20.78          | 21.62          |
| Hepta        | 5278               | 3806          | 4.15             | 2.82             | 17.01          | 11.25          |
| Octa         | 493                | 500           | 0.39             | 0.37             | 1.59           | 1.48           |
| Total/mean   | 127086             | 134984        | 100              | 100              | 409.60         | 399.03         |

Table 5
Summary of the frequency of SSRs from *A. deserticola* and *A. atacamensis* with different numbers of tandem repeats.

| Motif length | Largest SSRs | Highest % | *A. deserticola* | *A. atacamensis* |
|--------------|--------------|------------|------------------|------------------|
|              | *A. deserticola* | *A. atacamensis* | *A. deserticola* | *A. atacamensis* |
| Di           | (CT)35       | (AG)36     | AT/AT (7.78)     | AT/AT (7)        |
| Tri          | (AAT)29      | (TTA)29    | AAT/ATT (13.31)  | AAT/ATT (12.07)  |
| Tetra        | (TTTA)14     | (TTAA)13   | AAAT/ATT (11.35) | AAAT/ATT (10.14) |
| Penta        | (AAAAAT)8    | (TAAAT)10  | AACTG/AGTTC (2.61)| AAAAT/ATTT (2.36)|
| Hexa         | (TAAAAA)8    | (TTATT)8   | AAAAT/ATTTTTT (0.84) | AAAAT/ATTTTTT (0.78) |
| Hepta        | (AAACCCCT)12 | (CCTAAACC)12 | AAATAA/ATTATTTT (1.82) | AAACCCCT/AGGTTTTT (0.58) |
| Octa         | (GTCAATGT)5  | (GAGAAGA)6 | AAAAAT/ATTTTTTTT (0.1) | AAAAAG/CTTTTTTT (0.11) |
2. Experimental design, materials, and methods

2.1. Plant material and DNA extraction

Samples of *A. deserticola* and *A. atacamensis* were collected at the Las Cardas Agricultural Experimental Field Station, University of Chile, (Coquimbo, Chile). Genomic DNA (approximately 100 mg) was extracted from leaves of *A. deserticola* and *A. atacamensis* plants with the DNeasy Plant Mini Kit (Qiagen).
Inc., Valencia, CA, USA), following the manufacturer’s protocols. DNA quality and quantity were checked by agarose gel electrophoresis and spectrophotometric measurement of UV absorption at wavelengths of 260 and 280 nm and absorbance ratios of 260/280 and 260/230, using an Infinitive M200Pro Nanoquant (Tecan Group US, Inc., Morrisville, NC, USA).

2.2. Next-generation sequencing

The Illumina paired-end library was prepared with the Illumina TruSeq DNA PCR-Free350 bp Library Preparation Kit (Illumina, San Diego, CA, USA). The paired-end library was sequenced using Illumina HiSeq 2500 Sequencer (Macrogen Inc., Seoul, Korea) using the TruSeq rapid SBS kit or Truseq SBS Kit v4 (Illumina, San Diego, CA, USA). The read sequence length was 126 nts from one end of the fragment to the other.

The sequence quality of raw genomic data was assessed using FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). The data was quality trimmed and filtered using PRINSEQ v0.20.4 (http://prinseq.sourceforge.net/) [4]. Reads containing more than 5% of unknown nucleotides, low-quality reads (those with more than 50% bases with Q-value ! 20) and unpaired reads were discarded. Short reads (<35 bp) were removed from the filtered data.

Raw data of A. deserticola and A. atacamensis are available from NCBI-Bioproject, PRJNA495747 and PRJNA495763 accessions, respectively.

2.3. In silico identification of putative SSRs and primer design

We analyzed perfect and imperfect SSRs. The contig sequences obtained in FASTA files were screened with a repeat motif size range of 2–6 bp and a length of >12 bp. This included dinucleotide repeats !6, trinucleotide repeats !4, and tetra-, penta-, hexa-, hepta- and octanucleotide repeats !3, using MiCroSAtellite identification software [7,8]. The program allows for direct primer design using PRIMER 3 [9] by searching for microsatellite repeats and primer annealing sites in the flanking regions (S1 and S2 Tables).

2.4. Putative A. deserticola and A. atacamensis gene prediction

Putative genes were predicted with AUGUSTUS software [1,2], analyzing contig and singleton genomic sequences from A. deserticola and A. atacamensis. The program is based on a hidden Markov model and is used for the ab initio prediction of protein coding genes in eukaryotic genomes. Arabidopsis thaliana (L.), Heynh. was used as the model organism. WEGO (Web Gene Ontology Annotation Plot) software was then used to functionally annotate potential coding sequences or predicted genes [3]. A manual inspection of the predicted genes was performed to maximize the accuracy of gene prediction. The genes encoding predicted proteins were scored using the NCBI non-redundant (NR), Uniprot, and GO database. Matches were selected with the value e ! 1xe-5 and with 40% sequence identity.

Acknowledgments

This work was supported by Fondecyt Project 11150551 and Fondecyt Project 1161377. Plant materials were obtained from Las Cardas Experimental Station (Universidad De Chile, IV Region).

Conflict of 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.
Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2019.104258.

References

[1] M. Stanke, S. Waack, Gene prediction with a hidden Markov model and a new intron submodel, Bioinformatics Suppl. 2 (2003) ii215-225.
[2] M. Stanke, B. Morgenstern, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints, Nucleic Acids Res. 33 (2005) W465–W467, https://doi.org/10.1093/nar/gki458.
[3] J. Ye, Y. Zhang, H. Cui, J. Liu, Y. Wu, Y. Cheng, H. Xu, X. Huang, S. Li, A. Zhou, L. Bolund, Q. Chen, J. Wang, H. Yang, L. Fang, C. Shi, WEGO 2.0: a web tool for analyzing and plotting GO annotations, Nucleic Acids Res. 46 (2018). W1, W71–W75, https://doi.org/10.1093/nar/gky400.
[4] R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets, Bioinformatics 27 (6) (2011) 863–864.
[5] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, et al., SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler, GigaScience 1 (2012) 1–6, https://doi.org/10.1186/2047-217X-1-18.
[6] R. Chikhi, P. Medvedev, Informed and automated K-mer size selection for genome assembly, Bioinformatics 30 (1) (2014) 31–37. https://doi.org/10.1093/bioinformatics/btt310.
[7] T. Thiel, W. Michalek, R.K. Varshney, A. Graner, Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.), Theor. Appl. Genet. 106 (3) (2003) 411–422. https://doi.org/10.1007/s00122-002-1031-0.
[8] S. Beier, T. Thiel, T. Münch, U. Scholz, M. Mascher, MISA-web: a web server for microsatellite prediction, Bioinformatics 33 (16) (2017) 2583–2585. https://doi.org/10.1093/bioinformatics/btx198.
[9] S. Rozen, H. Skaltsky, Primer3 on the WWW for general users and for biologist programmers, Methods Mol. Biol. 132 (2000) 365–386.