Data in Brief 40 (2022) 107721

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Data on draft genome assembly and annotation of Haloxylon salicornicum Moq.

Fadila Al Salameen, Nazima Habibi*, Sami Al Amad, Bashayer Al Doaij

Environment and Life Sciences Research Centre, Kuwait Institute for Scientific Research, Kuwait

ARTICLE INFO

Article history:
Received 2 October 2021
Revised 9 December 2021
Accepted 13 December 2021
Available online 16 December 2021

Keywords:
Whole-genome sequencing
Desert
Native plants
Biodiversity

ABSTRACT

Haloxylon salicornicum Moq. Bunge ex Boiss (Rimth) is one of the main structural elements in Eastern Arabian vegetation associations. The plant is utilized as a food source for domestic stock, stabilizes the soil surface besides providing suitable microclimates for exotic species. It is considered one of the most promising species for re-vegetation. *H. salicornicum* community is under threat from overgrazing leading to a reduction in the percentage of distribution from 22.7% to 2.2% in Kuwait. Therefore, genome characterization of this important Kuwaiti plant is required to formulate strategies for its conservation. Here we report the draft of the *H. salicornicum* genome, which was sequenced on an Illumina HiSeq 2500 platform. BUSCO assessment revealed 69% of the genome was to be complete. Overall, 12960 gene structures, 11280 protein-coding genes, 11309 mRNAs (protein-coding), 51265 exons and 48100 CDSs were predicted. Functional annotation was carried out by interproscan-5.29-68.0. A total of 7222 protein-coding sequences were, annotated out of 11309 by at least one ontology term. All these genes were associated with 11 major biological processes branched into 60 child processes.

* Corresponding author.
E-mail address: nhabibi@kisr.edu.kw (N. Habibi).

https://doi.org/10.1016/j.dib.2021.107721

Specifications Table

Subject	Plant Sciences
Specific subject area	Genomics
Type of data	Tables, Figures
How the data were acquired	Paired-end Illumina Sequencing
Data format	Raw, filtered, analysed
Parameters for data collection	A single specimen growing in its natural habitat (Al Kabd, Kuwait) was used for this study. Genomic DNA for sequencing was extracted from young leaves.
Description of data collection	Genomic DNA was digested using the restriction enzymes PstI+Btgl (New England BioLabs, Inc., Ipswich, MA, United States), and barcoded adapters were ligated to the DNA sample using T4 ligase (New England BioLabs, Inc.). Dual indexed libraries for *H. salicornicum* were pooled and loaded across 4 lanes of a 150 bp paired read sequencing run on an Illumina HiSeq 2500 (Illumina, San Diego, CA).
Data source location	Kuwait Institute for Scientific Research, Kuwait (N-DM-29.64798; E-DM-47.99595)
Data accessibility	Repository name: National Centre for Biotechnology Information Data identification number: PRJNA766761(SRA: SRR16094057) Direct URL to data: https://www.ncbi.nlm.nih.gov/sra/SRX12380181[accn] Supplementary data available at: https://figshare.com/s/a3c215093885a9707613

Value of the Data

- The data provides valuable information on the genome sequences of *Haloxylon salicornicum* and fills in the gap of genomic studies in this genus.
- The genome assembly will be useful for geneticists interested in comparative genomics, conservation, breeding and phylogeny of *Haloxylon*.
- The genome analysis formulates a basis for further high depth sequencing of the species.
- The data can be used to develop molecular markers.

1. Data Description

The loss of biodiversity in arid lands due to harsh climatic conditions is an issue of global concern [1]. Human interventions and encroachment have further added to the effect. The native vegetation of Kuwait is unique with diverse species of desert plants adaptable to the harsh climate, however is degrading at an alarming rate. Native vegetation is crucial for the health of the environment, supporting agricultural productivity as well as the biodiversity that is central to a country's cultural identity. To formulate effective conservation and restoration strategies, advanced molecular research is highly desirable [2–4]. Genome sequencing studies are thus helpful in providing first-hand information on the genome size, repeat content, microsatellite regions and genes involved in local adaptation. Synergistically the knowledge gained can be applied to biodiversity management [5,6].

In the present study, we conducted the whole genome sequencing of the desert shrub *Haloxylon salicornicum* Moq. The perennial herb has a tropical distribution, however, faces the threats of extinction in the Middle-eastern region. A total data of 180 million raw sequences were generated by Illumina Hiseq 2000 sequencing that included 5,323,041,232 paired-end reads of 126 bp each with a GC content of 36.77%. The average Phred score per base was $Q \geq 40$. Raw reads

Table 1
Basic statistics and N50 and GC content of raw and assembled sequences of Haloxylon salicornicum.

Platform	*Illumina Hi Seq 2500*
Total raw reads	*180 million*
Average read length (bp)	*125*
Total no. of Contigs	*533,304*
Max. Contig length	*50,005,871*
Mean Contig length	*50,000,765.65*
N50 value max	*50,000,194*
Sum of bases in contigs	*1,550,023,735*
GC %	*36.77%*
Mean Quality	*Q ≥ 40*

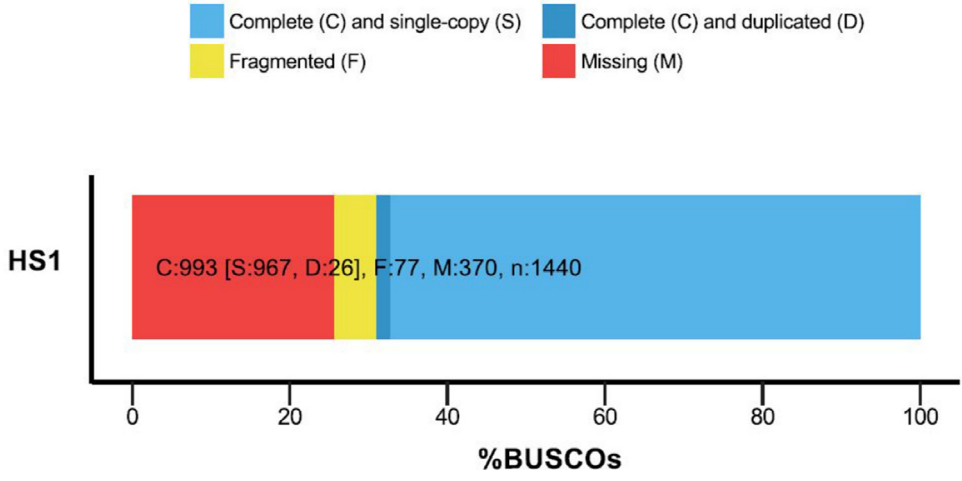

Fig. 1. BUSCO assessment of completeness of genome.

were *de novo* assembled into 533,304 contigs by Abyss yielding a genome of 1.5 Gb. The largest contig size and N50 was 50,005,871 bp 50,000,194, respectively (Table 1). The total number of bases in the contigs amounted to 1,550,023,735.

The BUSCO evaluation of completeness of the *H. salicornicum* genome sequence predicted that it was 67% complete (Fig. 1). A total of 1,440 BUSCO groups were searched in the genome mode. The genome assembly was found to contain 967 complete single-copy BUSCOs, 26 complete duplicated BUSCOs, 77 fragmented BUSCOs, and 370 missing BUSCOs (Fig. 1).

Gene annotation was performed against the *H. ammondendron* transcript assembly. The repeat modeller identified 1,796,653 repeats, 28,963 est2genome, 18,682 protein2genome and 12,690 gene structures. Multiple evidences by MAKER classified the gene structures into 11,280 protein-coding genes, 11,309 mRNAs, 51,263 exons and 48,100 CDs. We compared metrics for the full set of gene models, and the smaller high confidence set for transcript lengths (Fig. 2A), exon lengths (Fig. 2B) and exons numbers per transcripts (Fig. 2C).

The average length of the predicted transcripts (mRNA) was 3,216.49 (%) with a median length of 2,287 bp. The total coding length was 13,021,732. A significant number of non-coding transcripts or introns (36,786) were also predicted with a total length of 20,829,849 as represented in Table 2. The average and median lengths of the introns were 566.24 and 216, respectively.

The 48,090 exons were classified based on their position and length. The exons present at the initial position were 7,582, internal was 30,050, terminal was 2,719, UTR3 were 4,124 and

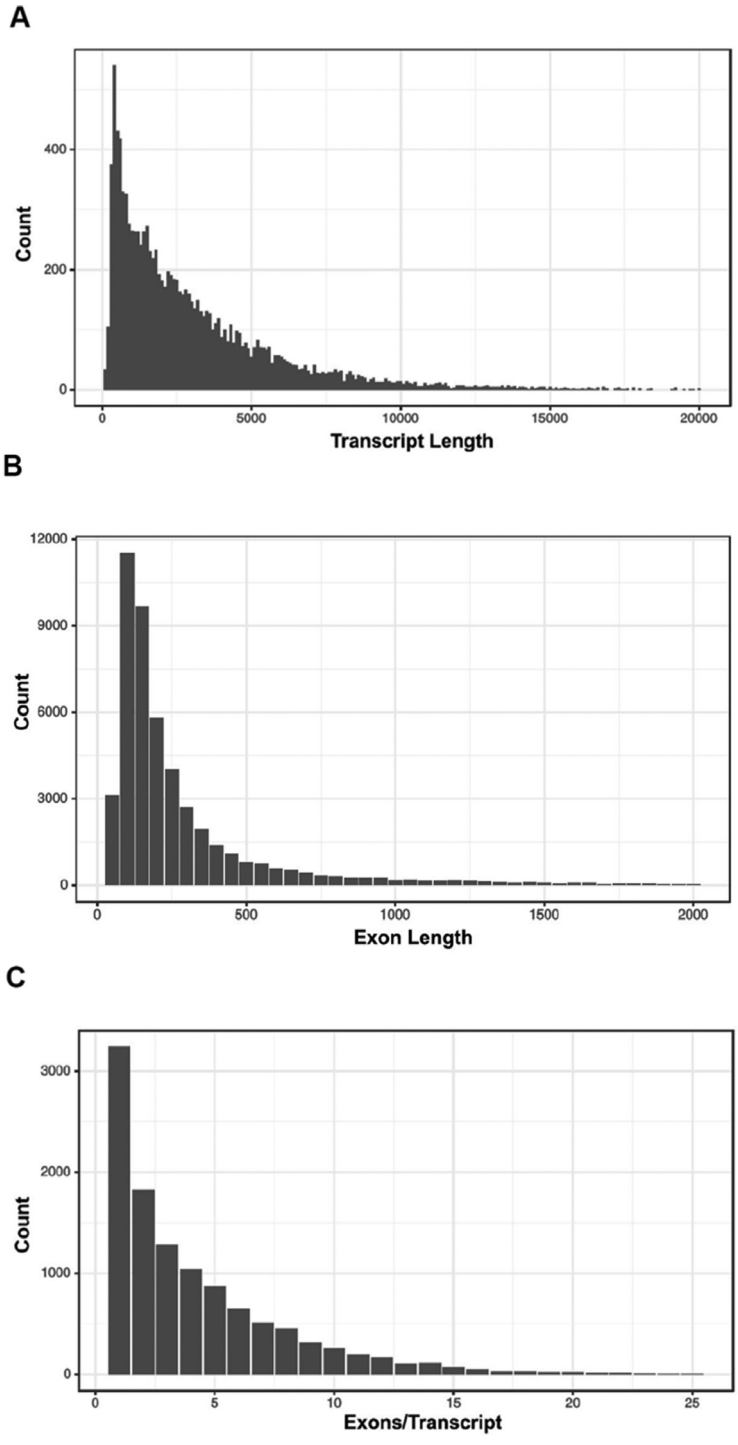

Fig. 2. Bar plots showing (A) Distribution of transcript lengths, (B) exon lengths and (C) exons number per transcript.

Table 2
Feature of coding and non-coding transcripts in Haloxylon salicornicum.

	Coding transcripts	Non-Coding Transcripts
Count	*11309*	*36786*
Average Length	*3216.49*	*566.24*
Median Length	*2287*	*216*
Total length	*36375287*	*20829849*

Table 3
Exon features and their position in the annotated genome of Haloxylon salicornicum.

Exon	All	Initial	Internal	Terminal	Single	UTR3	UTR5
Count	*48090*	*7582*	*30050*	*7739*	*2719*	*4124*	*4105*
Average Length	*270.6*	*326.5*	*181.16*	*344.99*	*891.99*	*264.49*	*147.57*
Median Length	*149*	*198*	*123*	*225*	*660*	*239*	*106*

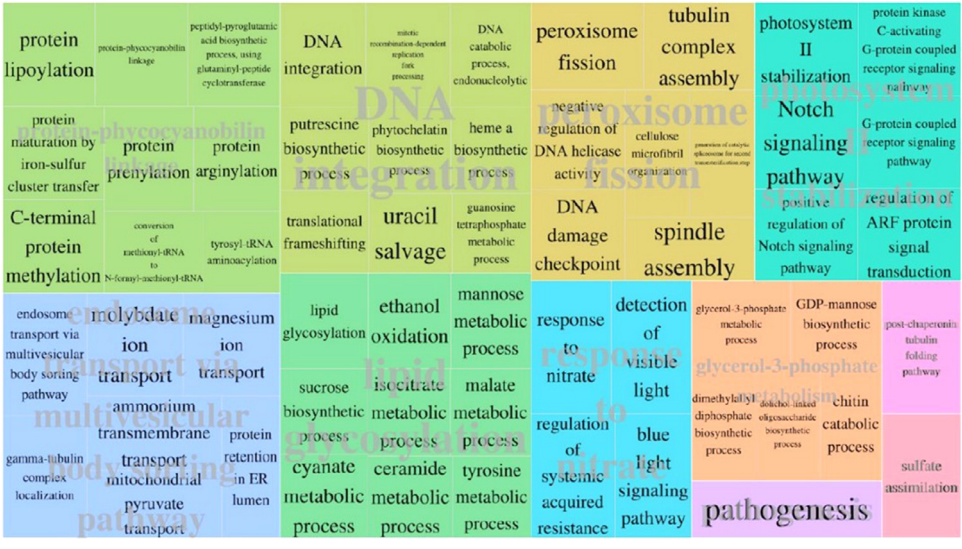

Fig. 3. REVIGO-TreeMAP showing the distribution of gene ontology terms related to each biological process.

UTR5 were 4,105 (Table 3). The largest average size of the exon was ~890 bp represented as single exons in the genome, whereas the minimum was ~150 for UTR5.

A total of 7,222 (64%) protein-coding sequences have been annotated out of 11,309 by at least one ontology term. All these genes were associated with 11 major biological processes branched into 60 child processes (Fig. 3).

2. Experimental Design, Materials and Methods

2.1. Preparation of plant material and DNA extraction

Fresh leaves of *H. salicornicum* were collected from a single specimen growing in the Al Kabd area of Kuwait. GPS coordinate of the collected specimen was recorded. Young leaf samples and shoots were stored in sealed polythene bags and transported on ice to the lab. The sample was appropriately labelled and kept at -80°C until further use. DNA isolation from leaf tissues was

carried out using GenElute™ Plant Genomic DNA Miniprep Kit (Sigma, St. Louis, MO), as described previously [2]. The DNA isolation was done in triplicate. DNA purity (Absorbance ratio A260/A280) and quantity (Absorbance at 260 nm) were measured by the Nanodrop (Thermo Scientific, Carlsbad, CA) and Qubit fluorometer (Thermo Fisher Scientific, Carlsbad, CA). Isolated DNA samples of *H. salicornicum* were run on 0.8% of agarose gel to check the intactness and quality.

2.2. DNA sequencing and assembly

Genomic DNA was digested using the restriction enzymes PstI+BtgI (New England BioLabs, Inc., Ipswich, MA, United States), and barcoded adapters were ligated to each DNA sample using T4 ligase (New England BioLabs, Inc.). Dual indexed libraries were loaded across 1 lane of a 126 bp paired-end read sequencing run on the Illumina HiSeq 2000 at the University of Minnesota Genomics Centre (http://genomics.umn.edu/). The quality of the fastq files was assessed via the FastQC tool and a Q value $\geq$40 was recorded [7] (Fig. S1–S4). Adapters were trimmed using Trimmomatic [8]. The sample was assembled with Abyss 2.0.2 using the "abyss-pe" command setting a kmer size of 64 (k=64). Assembly statistics were generated by QUAST 3.9 [9] (Table S1). Completeness of genome was evaluated using Benchmarking Universal Single-Copy Orthologs Version 2 (BUSCO v3.0.2) [10].

2.3. Gene annotation

Haloxylon ammodendron transcriptome assembly (GSE63970_Trinity.fasta) was used to provide evidence based gene prediction in the MAKER pipeline [11]. A *de novo* gene prediction tool AGUSTUS [12] was trained using a curated dataset to use in the MAKER pipeline. A *de novo* repeat element identification was performed for repeat masking to correctly predict gene structures using Repeat Modeler [13]. Functional annotation was carried out by interproscan-5.29-68.0 [14]. They were classified into Gene ontology categories and visualized using Web Gene Ontology Annotation Plot (WEGO) 2.0 [15].

Ethics Statement

Not applicable.

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.

CRediT Author Statement

Fadila Al Salameen: Conceptualization, Writing – original draft; **Nazima Habibi:** Software, Data curation, Visualization, Writing – review & editing; **Sami Al Amad:** Funding acquisition, Supervision; **Bashayer Al Doaij:** Methodology.

Acknowledgments

We thank the Kuwait Foundation for Advancement of Sciences (KFAS Grant No. P214-42SL-02) and Kuwait Institute for Scientific Research (KISR; Grant No. FB089C) for funding this research.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107721.

References

[1] F. Al-Salameen, N. Habibi, V. Kumar, S. Al-Amad, L. Talebi, B. Al-Doaij, J. Dashti, Genetic characterization of Haloxylon salicornicum and Rhanterium eppaposum native plant species of Kuwait by DNA Markers, Final Report, Kuwait Institute for Scientific Research, Kuwait, 2018 FB089C, KISR # 14599, doi:10.6084/m9.figshare.14851578.v1. In press

[2] F. Al Salameen, N. Habibi, V. Kumar, S. Al Amad, J. Dashti, L. Talebi, B. Al Doaij, Genetic diversity and population structure of Haloxylon salicornicum Moq. in Kuwait by ISSR markers, PloS one 13 (11) (2018) e0207369, doi:10.1371/journal.pone.0207369.

[3] F. Al Salameen, N. Habibi, V. Kumar, S. Al Amad, J. Dashti, L. Talebi, B. Al Doaij, Genetic diversity analysis of Rhanterium eppaposum Oliv. by ISSRs reveals a weak population structure, Curr. Plant Biol. 21 (2020) 100138, doi:10.1016/j.cpb.2020.100138.

[4] N. Habibi, M.H. Rahman, F. Al Salameen, Synoptic overview on application of molecular genetic markers in Acacia, Res. J. Biotechnol 15 (10) (2020) 152–166, doi:10.6084/m9.figshare.14169872.v25.

[5] N. Habibi, DNA marker technology for conservation of plant genetic resources in Kuwait, in: O.P. Yadav, N.R. Panwar (Eds.), Proceedings of the 13th International Conference on Development of Drylands Converting Dryland Areas from Grey into Green, Jodhpur, India, 2019, doi:10.6084/m9.figshare.14673867.v1.

[6] N. Habibi, F. Al Salameen, Role of ISSR markers for conservation of Rhanterium eppaposum Oliv, in: N. Bhat (Ed.), Proceedings of the Kuwait in International Symposium and workshop on Native Seed Restoration of Dryland Ecosystems, Kuwait, Kuwait Institute for Scientific Research, Kuwait, 2017, pp. 69–74, doi:10.6084/m9.figshare.14174231.v1. In press.

[7] S. Andrews, FastQC: a quality control tool for high throughput sequence data. Available online. Retrieved May. 2010;17:(2018). https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[8] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (15) (2014) 2114–2120, doi:10.1093/bioinformatics/btu170.

[9] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics 29 (8) (2013) 1072–1075, doi:10.1093/bioinformatics/btt086.

[10] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (19) (2015) 3210–3212, doi:10.1093/bioinformatics/btv351.

[11] M.S. Campbell, C. Holt, B. Moore, M. Yandell, Genome annotation and curation using MAKER and MAKER-P, Curr. Protoc. Bioinfor. 48 (1) (2014) 4.11.1-4.11.39, doi:10.1002/0471250953.bi0411s48.

[12] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, B. Morgenstern, AUGUSTUS: ab initio prediction of alternative transcripts, Nucleic Acid Res. 34 (suppl_2) (2006) W435–W439, doi:10.1093/nar/gkl200.

[13] A. Smit, R. Hubley, RepeatModeler Open-1.0. Repeat Masker, (2008). Website: http://www.repeatmasker.org.

[14] P. Jones, D. Binns, H. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliams, J. Maslen, A. Mitchel, G. Nuka, S. Pesseat, A.F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S. Yong, R. Lopez, S. Hunter, InterProScan 5: genome-scale protein function classification, Bioinformatics 30 (9) (2014) 1236–1240, doi:10.1093/bioinformatics/btu031.

[15] J. Ye, L. Fang, H. Zhang, Y. Zhang, J. Chen, Z. Zhang, J. Wang, S. Li, R. Li, L. Bolund, J. Wang, WEGO: a web tool for plotting GO annotations, Nucleic Acid Res. 34 (suppl 2) (2006) W293–W297, doi:10.1093/nar/gkl031.