scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Long-read *de novo* assembly of the spekboom (*Portulacaria afra*) genome

Wilku B. Meyer[1], Helmien Barkhuizen-Roode[1], Larysha Rothmann[1], Shannon S. Derman[1], Sahr Mian[2], Alvera A. Vorster[3], Carel J. Van Heerden[3], Ilia J. Leitch[2], Ernst Van Jaarsveld[4,5] & Renée Prins[1]✉

Portulacaria afra **(*P. afra*), or Spekboom as it is locally known, is a versatile evergreen shrubby succulent plant in the family Didiereaceae and endemic to the south-eastern parts of southern Africa. *P. afra* is not only known for its nutritional composition and its medicinal properties but has been shown to be pivotal in preventing desertification through limiting soil degradation. In addition, there is currently growing global attention on the use of *P. afra* as a garden ornamental. To address the lack of genetic resources for this increasingly important plant, we report on the *de novo* genome assembly of a Spekboom specimen of the Albany Thicket Biome using Oxford Nanopore Technology long-read sequencing, coupled with Illumina short-read polishing. Flow cytometry and *k*-mer analysis revealed the genome to be tetraploid and supports a haploid assembly size of ~ 0.68 Gb. BUSCO analysis revealed 99.3% complete BUSCO genes, indicative of the high gene completeness of the assembled genome.**

Background & Summary

Spekboom (*Portulacaria afra* L. Jacq.), also known as Pork Bush or Elephant's bush[1], is an indigenous southern African succulent plant belonging to the subfamily Portulacarioideae of the family Didiereaceae (Caryophyllales), with mostly unexplored important properties[2,3]. *Portulacaria afra* (*P. afra*) is one of seven species in the genus *Portulacaria*[4] and its distribution in southern Africa is confined to the eastern parts (Western Cape to the Limpopo Province). Compared to the six other species of *Portulacaria* (from the dry western parts of southern Africa), which are slow to grow outside their natural habitats, *P. afra* has a wide-spread distribution in South Africa (SA) and is an important species for thicket and veld restoration in its optimal growing environments. It is edible by humans[2,3,5] and in its natural habitat it is also an important food source for wild animals such as elephant, buffalo, black rhinoceros and greater kudu[6]. In addition, it is pivotal to the mohair industry in SA, and therefore is an important pillar for the wildlife and ecotourism economies of SA. It is widely used in rural/ethnic communities in combination with other plants for its medicinal properties[7], which include treatment of skin infections[8] and stomach ailments such as diarrhoea[9]. Spekboom is globally popular in ornamental horticulture largely because it is fast-growing and easily propagated and the diversity of growth forms, leaf texture, leaf colour and flower colour are highly sought after[10–12].

Spekboom uses a Crassulacean acid metabolism (CAM) pathway for photosynthesis[13]. CAM is a metabolic and anatomical adaptation that differs from other plant photosynthetic pathways by the net uptake of carbon dioxide at night through open stomata. Therefore, Spekboom is known as a good carbon sequester[14,15]. During the day the stomata remain closed; this significantly reduces water loss and makes Spekboom particularly water-use efficient compared to many other plants[16]. However, it has been shown that Spekboom's CAM response is seasonal, with more CAM activity during the summer months when temperatures are higher[17]. It can also utilise the C3 photosynthetic pathway, photosynthesizing during the day in the cooler months of the year when the climate is more suitable[18]. While radioisotope studies failed to find a direct link between CAM activity, rainfall, and temperature, they did show that CAM operates continuously at low levels even when the plant is photosynthesizing using the C3 pathway (i.e. Spekboom benefits from being able to access two photosynthetic

[1]CenGen (Pty) Ltd, 78 Fairbairn Street, Worcester, South Africa. [2]Royal Botanic Gardens, Kew, Richmond, Surrey, UK. [3]Central Analytical Facility, DNA Sequencing Unit, Stellenbosch University, Stellenbosch, South Africa. [4]Department of Biodiversity and Conservation, University of the Western Cape. Private Bag X17, Bellville, 7535, South Africa. [5]Babylonstoren Farm, P.O Box 167, Simondium, 7670, South Africa. ✉e-mail: cengen@cengen.co.za

pathways depending on the climate). Although these studies have suggested that genotypic differences may be responsible for such phenotypic plasticity, no further genotyping evidence has been reported to date[18].

An extensive review[2] of the claims and 'hype' surrounding Spekboom as a "climate-change-combating plant" concluded that its carbon sequestration ability was comparable to that of other CAM plants. Thus, Spekboom has a significant and expanding value from an evolutionary, agricultural, medicinal, and ornamental perspective. It is therefore crucial that a genomic benchmark for understanding its biodiversity, evolution, medicinal properties, and resilience to climate change is established.

Evolutionary relationships within the subfamily Portulacarioideae has been analysed by several genomic approaches, including the use of six plastid sequences[4] and RNAseq transcriptome data sets, although the source locations were not specified[19,20]. Phylogenetic analysis using the transcriptomic data within the Portulacineae (a higher order clade within the Caryophyllales which includes Didiereaceae), found significant evidence for gene family expansion in genes associated with stress adaptation and within clades found in extreme environments[19]. However, little is known about the genetics of Spekboom, and so far, no genome sequence assembly has been produced. The genetic diversity of Spekboom is thus poorly understood since no genetic data to assess genotype variability for *P. afra* are available. The reported chromosome count for a *P. afra* accession of the Cape province region (*Rowly* s.n.) of 2n = 44, which, based on x = 11 suggests it is a tetraploid[21].

Currently, Spekboom accessions are separated and defined using phenotypic characteristics, chief among them being leaf size. Van Jaarsveld and colleagues have conducted varied assessments and observations of Spekboom across South Africa, revealing a great diversity in leaf sizes and the overall size and shape of the plant, which appears to be genetically controlled as it is retained in common garden cultivation[3]. Nevertheless, the extent to which genetic diversity and/or environmental factors underpin the different phenotypes remains unknown and the development of genomic data to contribute towards dissecting these unknowns will be invaluable.

Here we present the first *de novo* genome assembly of *P. afra* using Oxford Nanopore Technology (ONT) long-read sequencing combined with short-read Illumina data for polishing the genome. We also estimated its genome size (1C-value) to be 0.78 billion base pairs (i.e. 0.78 gigabases (Gb)) by flow cytometry from an accession growing at the Royal Botanic Gardens, Kew, UK. This result was supported by the *k*-mer analysis of the Illumina and ONT reads from a South African accession, from which tetraploidy could be inferred.

Methods

Plant material. For estimating the genome size by flow cytometry, leaf tissue was collected from an individual plant of *Portulacaria afra* (accession number 1989–1484) growing in the Living Collections at the Royal Botanic Gardens, Kew, UK.

For the reference genome construction, a single whole-plant *Portulacaria afra* (28718 EvJ; European Nucleotide Archive (ENA) accession ERS21412505) specimen was collected on 28th February 2020 in the Klein Karoo, South Africa, 21 km south of the highest point of the Robinson Pass (33°41′08.16″ South; 22°08′56.56″ East at 298 meter above sea level). This plant was collected as part of a broader survey conducted by Van Jaarsveld and colleagues spanning forty-five years, representing the full distribution region of *P. afra* across South Africa, under a permit allowing the collection of both protected and unprotected flora for research purposes[2], *P. afra* is unprotected due to its abundance. The plant selected for sequencing was identified taxonomically based on its morphology and distribution and represents a specimen typical of the Gamka thicket biome (AT2) which is nested within the Albany thicket[22] (Fig. 1). This individual now forms part of the Living Collection at Babylonstoren farm (= Accession number 28718 EvJ).

Genome size estimate with flow cytometry. Nuclear genome size was estimated using propidium iodide flow cytometry. Approximately 1 cm^2 of fresh leaf material was co-chopped in a petri dish with the internal standard (*Petroselinum crispum* (Mill) Nyman ex A.E.Hill 'Champion Moss Curled', 1 C = 2 200 Mb)[23] using a new razor blade in 500 µl of OxProtect buffer (Sysmex Europe GmbH, Norderstedt, Germany). A further 1 ml of OxProtect buffer was added and the contents passed through a 30 µm nylon filter. An additional 500 µl of OxProtect was filtered into the sample before staining with 100 µl propidium iodide (1 mg/ml) and incubating on ice for 15 minutes.

Three DNA sample replicates were run, recording up to 1 000 nuclei per fluorescence peak using a Sysmex CyFlow Space (Sysmex Europe GmbH, Norderstedt, Germany) flow cytometer, fitted with a 100-mW green solid-state laser (532 nm). The resulting histograms were analysed with the WindowsTM-based FlowMax software (v. 2.9 2014, Sysmex GmbH) and the average of each sample was used to estimate the genome size[24].

The flow cytometric analysis of *P. afra* resulted in high-quality histograms with the coefficient of variance (CV) for the sample peak being below 4% and the CV of the calibration standard peak below 3% in all samples analysed. Based on the mean G_1 peak positions of the sample and calibration standard, *P. afra* has an estimated genome size of 1 C = 0.79 picograms (pg) or 0.78 Gb pairs.

High Molecular Weight (HMW) Genomic DNA extraction. Spekboom is a highly mucilaginous succulent, which presents a challenge to obtaining pure HMW DNA suitable for whole-genome sequencing (WGS). Mucilage is composed of polar glycoproteins that form a thick, gluey substance that thickens membranes to strengthen water and food storage in the leaves[25,26]. To successfully extract DNA from Spekboom leaf tissue, mucilage and other contaminating compounds, such as phytochemicals[27] need to be removed from the plant material via various clean-up steps. Approximately 45 small young leaves were cut into a mortar filled with liquid nitrogen and pulverised for at least 25 minutes until the leaf material was turned into a smooth, homogenous fine powder. One gram of tissue was then transferred into a 30 ml Nalgene extraction tube. Subsequently, a sorbitol wash method[28] was used to remove the highly abundant polysaccharides before lysing the cells[29,30]. After the

Fig. 1 Whole plant specimen of *Portulacaria afra* (28718 EvJ) used for genome sequencing and assembly.

sorbitol wash, a Nucleon PhytoPure Genomic DNA Extraction Kit (Cytiva RPN8511) was used following manufacturer's instructions (https://doi.org/10.5281/zenodo.10986148).

The DNA was quantified using 1 µl of the extracted DNA sample with both a NanoDrop Eight Microvolume UV-Vis Spectrophotometer and a Qubit 4 Fluorometer (DNA broad range) (ThermoFisher Scientific). The DNA quality was assessed with 260/280 and 260/230 nm absorbance ratios provided by the NanoDrop. Ideally, the absorbance ratio of 260/280 should be between 1.8 and 2 to indicate suitable DNA quality without protein or RNA contamination. The absorbance ratio of 260/230 should be between 2 and 2.2 to indicate there is limited contamination by organic compounds or chaotropic salts. While these ideal absorbance ratios were not obtained, we were confident to proceed with long-read sequencing since the Nanodrop/Qubit DNA concentration (ng/µl) ratio was $\leq$~1.5, indicative of low contamination[31]. Genomic DNA fragment size was checked via a Fragment Analyzer (TapeStation D5000, © 2023 Agilent Technologies, Inc.) and DNA integrity (DIN) scores of >9.5 and fragment lengths of >60 000 bp were obtained, indicating large, intact DNA fragments.

Sequencing. Long-read sequencing runs were performed for the 28718 EvJ Spekboom accession using an ONT PromethION P24 instrument with two PromethION R10.4.1 flow cells (Oxford Nanopore Technologies Ltd, Oxford, UK). HMW genomic DNA libraries were prepared using the ONT Ligation Sequencing Kit (SQK-LSK-114) according to the manufacturer's instructions. Genomic DNA (~3 µg) fragments were repaired, 3′-adenylated and adapters ligated. In the final purification step with AMPure XP beads (Beckmann Coulter, Brea, CA, USA), the library for PromethION flow cell 1 was prepared with LFB wash, while the library for flow cell 2 was prepared with the short fragment buffer wash. The Qubit 4 Fluorometer (DNA broad range) was used to quantify 1 µl of each purified library prior to loading. Two PromethION flow cells (R10.4.1) were prepared by checking for active pores and primed according to the manufacturer's protocol (SQK-LSK114). Purified libraries were mixed with ONT sequencing buffer and loading beads prior to loading. Approximately 20 fmol of HMW DNA libraries were loaded onto each PromethION R10.4.1 flow cell. Flow cell 1 ran for ~72 hours and Flow cell 2 for ~19 hours, after which it was paused, washed according to the manufacturer's instructions using the Flow Cell Wash Kit (EXP-WSH004) and primed before an additional library was loaded and run for ~52 hours.

The PromethION flow cells produced a total of 85.23 Gb of DNA which was then used for base calling with Dorado 0.3.4, SUP (Super Accurate) mode. The majority of the bioinformatic analyses were run on the CHPC (South Africa Centre for High-Performance Computing) through the CSIR (Council for Scientific and Industrial Research) and the CHPC Nvidia GPU allowed base-calling to complete within 48 hours.

PromethION Flow Cell	No. of Reads	Read Max. Length (bp)	N50 (bp)	Yield (Gb)	Mean Read Quality (phred score)
1 (raw)	4,919,667.00	341,637.00	20,724.00	26.3	18.5
2 (raw)	6,781,778.00	327,351.00	23,972.17	41.5	18.4
Combined after filtering	6767,286.00	341,617.00	24,324.00	64.91	19.1

Table 1. Summary of the sequencing results of ONT PromethION flow cells base-called with Dorado SUP model.

Repeat Classes	Number of elements	Length (bp)	Percentage of genome
LINEs:	16,223	11,067,437	1.63
RTE/Bov-B	2,996	3,563,398	0.52
L1/CIN4	13,227	7,504,039	1.1
LTR elements:	72,337	99,950,023	14.68
Ty1/Copia	30,101	51,139,315	7.51
Gypsy/DIRS1	38,535	43,603,883	6.41
Total retroelements	**88,560**	**111,017,460**	**16.31**
DNA transposons	27,234	18,217,458	2.68
hobo-Activator	5,752	3,018,175	0.44
MULE-MuDR	13,716	9,355,951	1.37
Tourist/Harbinger	2,823	1,056,288	0.16
Unclassified	530,275	217,708,849	31.98
Total interspersed repeats		**346,943,767**	**50.97**
Rolling circles	2,242	1,823,777	0.27
Simple repeats:	236,904	12,035,695	1.77
Low complexity:	28,056	1,329,857	0.2

Table 2. Repeat element statistics of the Spekboom draft genome assembly.

Read filtering with Chopper retained reads with a minimum quality threshold of 15 or an estimated error rate of 3.2%[32]. Subsequent quality assessment with NanoPlot (v 1.40.2) indicated an N50 of 20 724 bp and 23 972 bp, a sum of 26.30 and 41.50 Gb and a mean quality score of 18.5 and 18.4 for the PromethION flow cells 1 and 2 respectively giving the total number of reads after filtering of 6.7 million (Table 1). This resulted in a total yield of 64.91 Gb.

Four Illumina libraries of 28718 EvJ Spekboom accession were prepared using the Illumina® DNA PCR-Free Prep Tagmentation kit. The libraries were quantified using the TAKARA- Library Quantification Kit, normalised and sequenced on an SP flow cell of the Illumina Nova Seq. 6000 sequencing platform (Centre for Epidemic Response and Innovation (CERI) at Stellenbosch Tygerberg campus) using the paired-end (150 bp) protocol. A combined estimated coverage of 80X was targeted for the 28718 EvJ Spekboom accession. A total of 302 917 323 reads were generated with a maximum length of 151 bp resulting in a yield of 82.83 Gb with a mean read quality phred score of 34.65.

Draft genome assembly and polishing. The *P. afra* draft genome was assembled with hifiasm v024.0-r703 based on the PromethION R10.4.1 Dorado base-called reads. The assembly was further polished with short-read Illumina data after one long-read polishing step with Medaka[33]. Illumina paired-end reads, with an average Q score of ~35, were used to polish the assembly. The polishing was done with NextPolish (v 1.41) using default parameters[34].

The completeness of the *P. afra* genome sequence assembly was evaluated using the Quality Assessment Tool for Genome Assemblies[35] (Quast 5.0.2), Merqury[36] (Merqury v1.3) using the Illumina data and Benchmarking Universal Single-Copy Orthologs[37] (BUSCO 5.6.1) with Viridiplantae_odb10.2024-01-08, Embryophyta_odb10.2024-01-08 and Eudicots_odb10.2024-01-08 (the closest available lineage dataset to Spekboom)[38].

Identification and removal of contaminant sequences. Possible contaminants and misassembled contigs was identified by self-mapping of ONT reads to the assembly using Minimap2[39] (v2.24) and SAMtools[40] (v1.9). Contigs that appeared more than four times or had coverage less than 25% of the average were identified and removed.

Repeat sequence annotation. Repeat sequences including transposable elements (TEs) were identified using a custom library constructed using RepeatModeler (v 2.05) (https://github.com/Dfam-consortium/RepeatModeler) with the recommended settings and quantified with RepeatMasker (v 4.1.6) (https://github.com/Dfam-consortium/RepeatMasker) in default mode[41]. The genome was found to be moderately repetitive with 362.13 Mb of repetitive elements identified, representing 53.2% of the genome.

The largest percentage of repeat sequences (i.e. 31.98% of the genome) were placed in the 'Unclassified repeat' category, followed by retroelements at 16.31%, consisting of two element types, long terminal repeat (LTR) elements and long interspersed nuclear elements (LINEs). The smallest class of the interspersed repeats was the

Features	All genes	Low similarity	High similarity
Number of coding sequences (CDS)	370,897	269,179	85,726
Mean length of CDS (bp)	212.78	207.53	163.92
Number of exons	370,897	269,179	85,726
Mean length of exons (bp)	212.78	207.53	163.92
Number of protein-coding genes	56,038	35,592	9,202
Mean length of protein-coding genes (bp)	3,399.08	3,805.26	4,413.11
Number of introns	306,021	227,532	75,074
Mean length of introns (bp)	487.07	457.26	415.92
Number of start codons	64,885	41,648	10,654
Number of stop codons	64,879	41,647	10,656
Number of transcripts	64,896	41,653	10,656
Mean length of transcripts (bp)	3,512.89	3,838.98	4,249.01
Total length CDS [Mbp] (% of the genome)	78.92 (11.59%)	55.86 (8.21%)	14.05 (2.06%)
Total length exon [Mbp] (% of the genome)	78.92 (11.59%)	55.86 (8.21%)	14.05 (2.06%)
Total length gene [Mbp] (% of the genome)	190.48 (27.98%)	135.44 (19.90%)	40.61 (5.97%)
Total length intron [Mbp] (% of the genome)	149.05 (21.90%)	104.042 (15.28%)	31.23 (4.59%)
Total length transcript [Mbp] (% of the genome)	227.97 (33.49%)	159.91 (23.49%)	45.28 (6.65%)

Table 3. Summary statistics of the gene prediction and annotation of Spekboom draft genome assembly.

DNA transposons at 2.68%. No satellites nor small RNAs were detected, and low complexity, rolling circles and simple repeats only accounted for 0.2%, 0.27% and 1.77% of the genome respectively (Table 2).

A custom script was used to search for a common plant telomere motif (TTTAGGG) and the five largest contigs (ranging in size from 30 to 26.6 Mb) and three smaller contigs (ranging in size from 24.2 to 22 Mb) were identified to have multiple copies on both ends. As they fall in the same range as the predicted chromosome sizes (estimated to be 35 Mb per chromosome - i.e. 0.78/22) for *P. afra*, it is possible that these contigs correspond to whole chromosomes. An additional 29 contigs were identified having it on a single end, representing 300 Mb of the assembly length, which is only one contig more than the expected 28 [(22 – 8)*2], indicating there may be a duplication which was not filtered correctly during assembly.

Gene annotation. There are currently two publicly available *P. afra* untargeted transcriptomic datasets, generated from leaf tissue, of vouchers DBG 1988-0583-02-1 G20 (https://www.ncbi.nlm.nih.gov/biosample/SAMN10132153)[19] and FUS: Cai F. Zhang & C.-H. Huang ZC7421. (https://www.ncbi.nlm.nih.gov/biosample/SAMN15205528)[20]. BRAKER3[42–52] was used for the *de novo* annotation on the assembly, first only with the dataset from DBG 1988-0583-02-1 G20 as it contained the highest number of paired-end reads with over 50 million providing the greatest sequence depth. The hints generated were then used in a second run using both transcriptomic datasets. A total of 56 038 genes were predicted, corresponding to 64 896 protein-coding transcripts. Among these, 41 653 transcripts showed low similarity to proteins in the reviewed Swiss-Prot database (defined as having $<30\%$ sequence identity, $<50\%$ alignment coverage of the query, and e-values $<$ 1e-10), where 10 656 transcripts showed high similarity (defined as having $\geq 70\%$ sequence identity, $\geq 70\%$ query coverage, and e-values $<$ 1e-10) (Table 3).

Alignment metrics. The quality of the assembly was tested by using the short read Illumina Nova Seq. 6000 data used for polishing the assembly. The reads were mapped to the assembly using the BWA-MEM[53] (v 0.7.17-r1188). 98.82% of the 561 460 725 reads mapped back to the assembly with an average Phred-scaled base quality of 34.5 and coverage of 63.39.

Data Records

The data set for this study have been deposited in the ENA at EMBL-EBI under accession number PRJEB82232. The ENA accession number of the ONT PromethION sequencing fastq.gz data is ERR13936846[54]. The ENA accession number for the Illumina sequencing fastq.gz data is ERR13937131[55]. The polished assembled genome sequence has been deposited at ENA with accession number CAYFBI020000000.2[56], and annotation results of the repeats and genes with accession number ERZ27245339[57]. Clonal cuttings of the Spekboom accession 28718 EvJ and accession number 1989-1484 are available at Babylonstoren Farm, South Africa and the Royal Botanic Gardens, UK respectively.

Technical Validation

Genome size and ploidy estimation via *k*-mer analysis of ONT reads. A histogram of *k*-mer counts was generated with KMC (v3.2.4) using the filtered reads from the ONT PromethION data for accession 28718 EvJ with the following parameters [kmc -cs1000 -k21 -fq; kmc_tools transform prefex histogram][58].

The resulting 21-mer frequency vs. 21-mer coverage graph (Fig. 2) revealed three distinct peaks at coverage values of 41, 85, and 172. The first peak, occurring at 41(1x), represents *k*-mers formed from unique genomic sequences at heterozygous loci and is referred to as the heterozygous peak, which is expected at half of the average coverage of the 2x genome. The second peak, at 85 (2x), typically represents *k*-mers from homozygous loci,

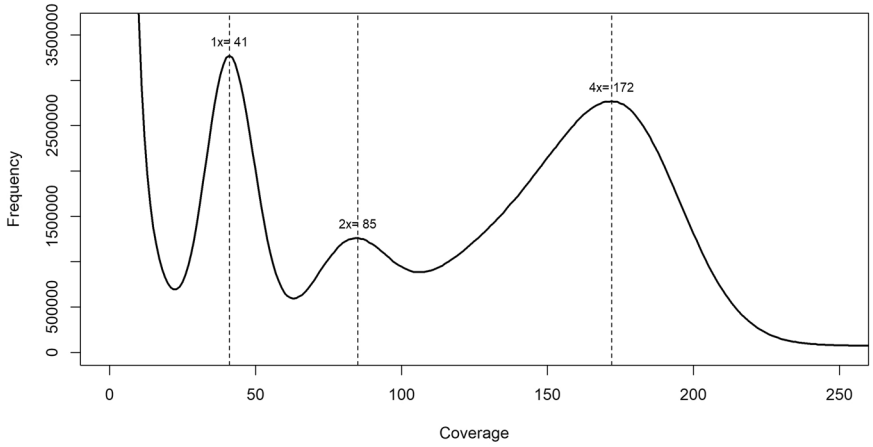

Fig. 2 Frequency of 21-mers vs 21-mer coverage graph based on WGS Oxford Nanopore Technology reads (64 910 614 305 b), with suggested ploidy indicated on the local *k*-mer maximum, estimated un-collapsed genome length (763.65 Mb) for the haploid genome, and expected assembly length for the collapsed genome (377.39 Mb).

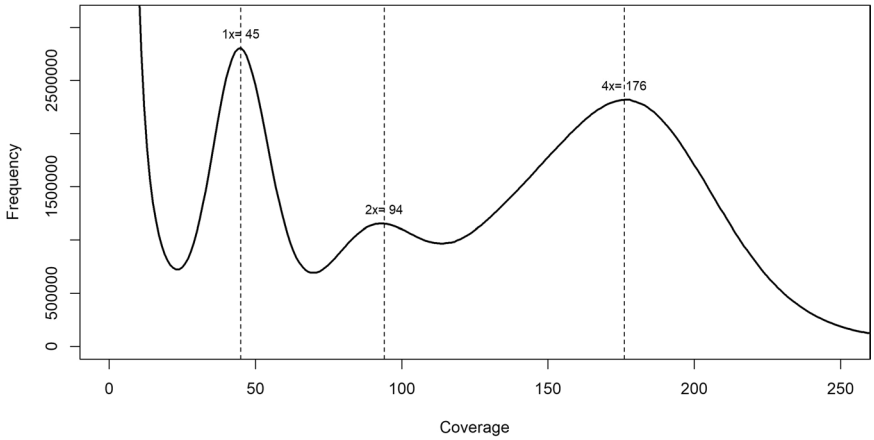

Fig. 3 Frequency of 21-mers vs 21-mer coverage graph based on WGS Illumina short-reads (82 831 991 209 b), with suggested ploidy indicated on the local *k*-mer maximum, estimated un-collapsed genome length (881.19 Mb) and expected collapsed assembly length (470.64 Mb).

however, its frequency is much lower compared to the third peak. The third peak, at 172 (4x), likely represents the true homozygous peak, and the fact that this peak occurs at double the coverage of the second peak suggests that *P. afra* is likely a tetraploid (4x) as the third peak is at approximately double the 2x coverage. An estimate of the genome length can be calculated by dividing the 2x peak at coverage = 85 with our 64.91 Gb of input data, indicating a genome size of about 0.76 Gb (64.91 Gb/85). Based on this, we calculated the size of the monoploid genome to be 0.37 Gb (64.91 Gb /172). The estimated genome size of 0.76 Gb corresponds to the 0.78 Gb estimation for the genome size from flow cytometry and the genome size of 0.68 Gb from the un-collapsed genome assembly generated Hifiasm. At this stage we cannot infer whether it is an auto or allotetraploid.

Genome size and ploidy estimation via *k*-mer analysis of Illumina reads. A histogram of *k*-mer counts was first generated with the short-read WGS data from Illumina for accession 28718 EvJ following the same parameters as above. The resulting frequency of *k*-mers vs *k*-mer coverage graph (Fig. 3) revealed three defined peaks matching the ONT data, with their local maximum at 45, 94 and 176 supporting the ONT analysis that suggests the individual of *P. afra* sequenced is tetraploid. The estimated genome size from the 2x peak coverage of 94 was 0.88 Gb (82.83 Gb/94), slightly larger than the estimate from the ONT data. Similarly, the size of the monoploid genome was also slightly larger at 0.46 Gb (82.83 Gb/176).

Quality assessment of the genome assembly. BUSCO analysis showed that 372 of 425 (87.5%) genes were duplicated *Viridiplantae* only three genes missing with no genes being fragmented from the *P. afra* genome, when considering the *Embryophyta* genes 1383 of the 1614 (85.7%) genes were duplicated with only 11 missing and another 11 being fragmented and when considering the *Eudicots* genes 1969 of the 2326 (84.7%) genes were duplicated and

	Draft genome assembly
Contigs (>= 0 bp)	1,074
Contigs (>= 50000 bp)	243
Largest contig	30,072,207
Total length	680,724,613
GC (%)	37.37
N50 (bp)	22,194,108
N90 (bp)	4,127,385
BUSCO Evaluation (*Viridiplantae*)*	C:99.3%[S:11.8%,D:87.5%],F:0.0%,M:0.7%,n:425
BUSCO Evaluation (*Embryophyta*)*	C:98.6%[S:12.9%,D:85.7%],F:0.7%,M:0.7%,n:1614
BUSCO Evaluation (*Eudicots*)*	C:96.9%[S:12.2%,D:84.7%],F:0.8%,M:2.3%,n:2326
Merqury Quality Value	44.1493
Merqury Error Rate	3.85E-05
Merqury Completeness %	85.4038

Table 4. Assembly statistics of Spekboom draft genome assembly using Quast, BUSCO and Merqury.
* C – Complete; S- Complete and Single Copy; D – Complete and Duplicated; F – Fragmented; M – Missing.

only 56 were missing with 18 being fragmented. This indicates a high level of assembly completeness. Merqury gave a Q score of 44.15 indicating a very low error rate of 3.85^{-5} supporting a highly accurate assembly (Table 4).

Data availability

The data set is accessible without restrictions in the ENA repository. The ENA accession number of the ONT PromethION sequencing fastq.gz data is ERR13936846 (https://www.ebi.ac.uk/ena/browser/view/ERR13936846). The ENA accession number for the Illumina sequencing fastq.gz data is ERR13937131 (https://www.ebi.ac.uk/ena/browser/view/ERR13937131). The polished assembled genome sequence has been deposited at ENA with accession number CAYFBI020000000.2 (https://www.ebi.ac.uk/ena/browser/view/CAYFBI020000000.2), and annotation results of the repeats and genes with accession number ERZ27245339 (https://www.ebi.ac.uk/ena/browser/view/ERZ27245339).

Code availability

All code and data will be made available here: https://github.com/cengenZA.

Received: 27 November 2024; Accepted: 23 September 2025;
Published online: 04 November 2025

References

1. Palmer, E. & Pitman, N. *Trees of South Africa* (A.A. Balkema, 1961).
2. du Toit, A. *et al.* Review of the Underutilized Indigenous *Portulacaria afra* (Spekboom) as a Sustainable Edible Food Source. *Agronomy* **13**(5), 1206, https://doi.org/10.3390/agronomy13051206 (2023).
3. Van Jaarsveld, E. J. & Le Roux, A. *Portulacaria afra* (L.) Jacq.: Variability and distribution. *Bradleya* **2021**(39), 138–152, https://doi.org/10.25223/brad.n39.2021.a13 (2021).
4. Bruyns, P., Oliveira-Neto, M., Melo-de-Pinna, G. F. A. & Klak, C. Phylogenetic relationships in the Didiereaceae with special reference to subfamily Portulacarioideae. *Taxon* **63**, 1053–1064, https://doi.org/10.12705/635.36 (2014).
5. Acocks, J. P. H. *Veld types of South Africa* (Government Printer, 1953).
6. Van der Vyver, M. L., Mills, A. J. & Cowling, R. M. A biome-wide experiment to assess the effects of propagule size and treatment on the survival of *Portulacaria afra* (spekboom) truncheons planted to restore degraded subtropical thicket of South Africa. *PLos ONE* **16**(4), e0250256, https://doi.org/10.1371/journal.pone.0250256 (2021).
7. Teffo, T. K., Ramalepe, P. & Risenga, I. A short communication on the Ethnobotany, Phytochemistry, Pharmacological Evidence and Ecosystem Restoration Potential of South African *Portulacaria afra*. *J. Med. Plants Stud.* **10**(2), 112–115 (2022).
8. De Wet, H., Nciki, S. & van Vuuren, S. F. Medicinal plants used for the treatment of various skin disorders by a rural community in northern Maputaland, South Africa. *J. Ethnobiol. Ethnomed* **9**, 51, https://doi.org/10.1186/1746-4269-9-51 (2013).
9. Mlambo, N. P. The screening of medicinal plants traditionally used to treat diarrhoea, in Ongoye area, KwaZulu Natal. A dissertation submitted to the Faculty of Science at the University of Zululand in fulfilment of the requirement for the degree of Master of Science (2008).
10. Van Jaarsveld, E. J. *Wonderful Waterwise Gardening* (Tafelberg, 2000).
11. Van Jaarsveld, E. J. Shaped by suffering. *Veld and Flora* **87**, 16–19 (2001).
12. Van Jaarsveld, E. J. *Waterwise Gardening in South Africa and Namibia* (Struik Lifestyle, 2010).
13. Guralnick, L. & Jackson, M. The Occurrence and Phylogenetics of Crassulacean Acid Metabolism in the Portulacaceae. *Int. J. Plant Sci.* **162**(2), 257–262, https://doi.org/10.1086/319569 (2001).
14. Osmond, C. B. Crassulacean Acid Metabolism: A Curiosity in Context. *Annu. Rev. Plant Physiol.* **29**, 379–414, https://doi.org/10.1146/annurev.pp.29.060178.002115 (1978).
15. Ting, I. P. & Hanscom, Z. III Induction of Acid Metabolism in *Portulacaria afra*. *Plant Physiol.* **59**(3), 511–514, https://doi.org/10.1104/pp.59.3.511 (1977).
16. Black, C. C. Jr Photosynthetic Carbon Fixation in Relation to Net CO2 Uptake. *Annu. Rev. Plant Physiol.* **24**, 253–286, https://doi.org/10.1146/annurev.pp.24.060173.001345 (1973).
17. Guralnick, L. J., Rorabaugh, P. A. & Hanscom, Z. III. Seasonal Shifts of Photosynthesis in *Portulacaria afra* (L.) Jacq. *Plant Physiol* **76**(3), 643–646, https://doi.org/10.1104/pp.76.3.643 (1984).
18. Guralnick, L. J. & Gladsky, K. Crassulacean acid metabolism as a continuous trait: variability in the contribution of Crassulacean acid metabolism (CAM) in populations of *Portulacaria afra*. *Heliyon* **3**(4), e00293, https://doi.org/10.1016/j.heliyon.2017.e00293 (2017).

19. Wang, N. *et al*. Evolution of Portulacineae marked by gene tree conflict and gene family expansion associated with adaptation to harsh environments. *Mol. Biol. Evol.* **36**(1), 112–126, https://doi.org/10.1093/molbev/msy200 (2018).
20. Zhang, C. *et al*. Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole genome duplication. *Mol. Biol. Evol.* **37**(11), 3188–3210, https://doi.org/10.1093/molbev/msaa160 (2021).
21. Nyananyo, B. L. Chromosome Number Report 96. *Taxon* **36**, 661 (1987).
22. Hoare, D. *et al*. Albany thicket biome. in *The vegetation of South Africa, Lesotho and Swaziland* Strelitzia 19 (eds Mucina, L. & Rutherford, M. C.) Ch. 10 540-567, (SANBI, 2006).
23. Obermayer, R., Leitch, I. J., Hanson, L. & Bennett, M. D. Nuclear DNA C-values in 30 species double the familial representation in pteridophytes. *Ann. Bot.* **90**, 209–217 (2002).
24. Pellicer, J., Powell, R. F. & Leitch, I. J. The application of flow cytometry for estimating genome size, ploidy level endopolyploidy, and reproductive modes in plants. *Methods Mol Biol.* **2222**, 325–361, https://doi.org/10.1007/978-1-0716-0997-2_17 (2021).
25. Chandravanshi, K. *et al*. Isolation of Mucilage from Herbal plants and its Evaluation as a Pharmaceutical Excipients. *Res J. Pharmacognosy and Phytochem* **14**(3), 171–178, https://doi.org/10.52711/0975-4385.2022.00031 (2022).
26. Mondal, S. & Rahaman, S. T. An Overview on Isolation and Characterization of Mucilage from Various Species Related to Various Plant Families. *Saudi J. Med. Pharm. Sci.* **4**(2), 284–288 (2018).
27. Tabassum, S. *et al*. Phytochemical, biological, and in-silico characterization of *Portulacaria afra* Jacq.: A possible source of natural products for functional food and medicine. *S. Afr. J. Bot.* **150**, 139–145, https://doi.org/10.1016/j.sajb.2022.07.009 (2022).
28. Jones, A. & Schwessinger, B. Sorbitol washing complex homogenate for improved DNA extractions. *protocols.io*. https://doi.org/10.17504/protocols.io.beuvjew6 (2020).
29. Souza, H. A. V., Muller, L. A. C., Brandão, R. L. & Lovato, M. B. Isolation of high quality and polysaccharide-free DNA from leaves of *Dimorphandra mollis* (Leguminosae), a tree from the Brazilian Cerrado. *Genet. Mol. Res.* **11**(1), 756–764 (2012).
30. Štorchová, H. *et al*. An improved method of DNA isolation from plants collected in the field and conserved in saturated NaCl/CTAB solution. *Taxon* **4**, 79–84 (2000).
31. Schalamun, M. *et al*. Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from *Eucalyptus pauciflora*. *Mol. Ecol. Resour.* **19**(1), 77–89, https://doi.org/10.1111/1755-0998.12938 (2019).
32. De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**(5), btad311, https://doi.org/10.1093/bioinformatics/btad311 (2023).
33. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.*, gr.214270. 116, https://doi.org/10.1101/gr.214270.116 (2017).
34. Hu, J. *et al*. NextPolish: a fast and efficient genome polishing tool for long read assembly. *Bioinformatics* **36**(7), 2253–2255, https://doi.org/10.1093/bioinformatics/btz891 (2019).
35. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**(13), 142–150, https://doi.org/10.1093/bioinformatics/bty266 (2018).
36. Rhie, A. *et al*. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
37. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**(19), 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).
38. Zdobnov, E. M. *et al*. OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **49**, D389–D393, https://doi.org/10.1093/nar/gkaa1009 (2021).
39. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**(23), 4572–4574, https://doi.org/10.1093/bioinformatics/btab705 (2021).
40. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *GigaScience* **10**(2), giab008, https://doi.org/10.1093/gigascience/giab008 (2021).
41. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA.* **117**(17), 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
42. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**(5), 767769, https://doi.org/10.1093/bioinformatics/btv661 (2016).
43. Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database. *NAR Genomics and Bioinformatics* **3**(1), lqaa108, https://doi.org/10.1093/nargab/lqaa108 (2021).
44. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-genome annotation with BRAKER. In Gene Prediction (pp. 65–95). Humana, New York, NY. https://doi.org/10.1007/978-1-4939-9173-0_5 (2019).
45. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. biotechnol* **37**(8), 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
46. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(16), 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
47. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**(12), 1691–1692, https://doi.org/10.1093/bioinformatics/btr174 (2011).
48. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic acids research* **42**(15), e119, https://doi.org/10.1093/nar/gku557 (2014).
49. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**(1), 59, https://doi.org/10.1038/nmeth.3176 (2015).
50. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**(5), 637–644 (2008).
51. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**(1), 62 (2006).
52. Gabriel, L., Hoff, K. J., Bruna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).
53. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN] (2013).
54. *European Nucleotide Archive* http://identifiers.org/ena.embl:ERR13936846 (2025).
55. *European Nucleotide Archive* http://identifiers.org/ena.embl:ERR13937131 (2025).
56. *European Nucleotide Archive* http://identifiers.org/ena.embl:CAYFBI020000000.2 (2025).
57. *European Nucleotide Archive* http://identifiers.org/ena.embl:ERZ27245339 (2025).
58. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**(17), 2759–2761, https://doi.org/10.1093/bioinformatics/btx304 (2017).

Acknowledgements
The authors acknowledge draft genome sequencing provided by the 1KSA project (www.1KSA.org.za) with support provided by DIPLOMICS through the South African Department of Science and Innovation's South African Research Infrastructure Roadmap (SARIR). The National Integrated Cyber Infrastructure System (NICIS) is acknowledged for 1KSA data management and curation through the Data Intensive Research Initiative of South Africa (DIRISA) and for 1KSA draft genome analysis using the Centre High Performance Computing (CHPC). Dr Anso Le Roux (ABEERA, UNISA; University of Western Cape) is thanked for the text and insight provided during the preparation of the project proposal to DIPLOMICS. Ms Alana Crotz (CenGen) is thanked for technical support.

Author contributions
R.P. devised and led the project. E.v.J. provided the study material. I.J.L. and S.M. did the flow cytometry analysis, H.B.R. and S.S.D. performed DNA extractions, ONT sequencing analysis and data management. C.J.V.H. and A.A.V. performed short-read sequencing. L.R. performed base-calling and read QC. W.B.M. performed all subsequent bioinformatic analyses and final data management duties. W.B.M., H.B.R., L.R., S.S.D., S.M., I.J.L., E.v.J. and R.P. wrote and edited the manuscript. All authors read and approved the final manuscript

Competing interest
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to R.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.