scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A phased, near-telomere-to-telomere chromosome-scale reference genome of the Moroccan argan tree

Hanan El Idrissi[1,2], Anestis Gkanogiannis[3], Driss Iraqi[1], Siham Khoulassa[4], Mohamed Fokar[5], Bouabid Badaoui[2], Rachid Moussadek[6], Rachid Mentag[1]✉ & Slimane Khayi[1]✉

The argan tree (*Argania spinosa*), endemic to Morocco, holds major ecological and economic value. We generated a high-quality, chromosome-scale, phased reference genome using PacBio HiFi long reads and Hi-C scaffolding. The assembly resolves two haplotypes, each organized into 11 pseudochromosomes (2n = 22). Haplotype-1 spans 621 Mb (scaffold N50 = 50 Mb; GC = 33.79%), and Haplotype-2 spans 615 Mb (scaffold N50 = 51 Mb; GC = 33.77%). BUSCO completeness is 97.8% for Hap1 and 98.1% for Hap2, with Merqury QV values of 75 for both, indicating high consensus accuracy and strong phasing. Telomeric repeats (AAACCCT)n appear at both ends of most chromosomes, and only small terminal gaps remain, so we conservatively classify the assemblies as near-T2T. We annotated 35,183 gene loci producing 39,805 mRNA isoforms and 410 tRNA genes, with 76.46% of loci functionally characterized. Repeats represent 61.65% of the genome, dominated by LTR retrotransposons. All raw data, assemblies, and annotations are publicly accessible, providing a robust genomic foundation for conservation genetics, breeding, and evolutionary studies in *A. spinosa*.

Background & Summary

The argan tree (*Sideroxylon spinosum*), synonym (*Argania spinosa*), is a native species of the west-central region of the Kingdom of Morocco. It thrives in arid and semiarid areas and is the unique representative of the Sapotaceae family in North Africa[1]. This forest, designated as a UNESCO Biosphere Reserve[2], plays a vital socioeconomic and ecological role, particularly in arid regions[3]. It provides essential resources, such as food for both humans and livestock through its fruits and forage, as well as the globally valuable argan oil, extracted from its seeds[4]. Additionally, the argan tree contributes to combating desertification and soil erosion, while supplying wood for various uses[3,5]. Despite its importance, natural populations of the argan tree have been in continuous decline since the 19th century due to significant threats, including environmental changes such as climate change, overpopulation, and overexploitation. These factors have severely diminished the tree's ability to regenerate naturally[6–8].

The governmental "Generation Green 2020–2030" initiative aims to enhance development of agricultural sector in Morocco, building on the successes of the previous Green Morocco Plan. One of its goals is to double argan oil production to 10,000 tons by 2030, alongside efforts to promote the planting of 50,000 Ha of modern argan trees fields, and rehabilitation of 400,000 Ha of the forest[9]. Developing a high-quality reference genome of *Argania spinosa* is crucial for understanding its genetic diversity, resilience to harsh environmental conditions, and unique biochemical pathways, particularly those involved in oil production. This genomic resource would facilitate conservation efforts, improve breeding programs for enhanced productivity and stress tolerance, and support sustainable exploitation of this valuable species. Additionally, it would provide insights into the evolutionary adaptations of this endemic Moroccan tree, aiding in its restoration and conservation.

[1]Biotechnology Research Unit, Regional Center of Agricultural Research of Rabat, National Institute of Agronomic Research, Rabat, Morocco. [2]Department of Biology, Faculty of Science, Mohammed V University, Rabat, Morocco. [3]International Center for Biosaline Agriculture, Dubai, UAE. [4]Oasis System ResearchOasis System Research Unit, Regional Center of Agricultural Research of Errachidia, National Institute of Agricultural Research, Rabat, Morocco. [5]Center for Biotechnology and Genomics, Texas Tech University, Lubbock, USA. [6]International Center for Agricultural Research in the Dry Areas, Rabat, Morocco. ✉e-mail: rachid.mentag@inra.ma; slimane.khayi@inra.ma

Fig. 1 Specimen of *Argania spinosa* (TAGUERTE) used in this study: Morphological characteristics and habitat.

	PacBio HiFi	Hi-C	RNA-seq
Read number	4,726,911	676,616,214	409,755,010
Data volume (Gb)	86.22	101.25	41.3
N50(b)	20,247	—	—
Coverage depth (x)	130	152	61

Table 1. Summary of sequencing data used for genome assembly.

In this study, we present the first phased near-T2T reference genome assembly of *Argania spinosa*. The assembly was generated using PacBio HiFi long reads in combination with Hi-C data, resulting in two fully phased haplotypes (Hap1 and Hap2), each comprising 11 chromosomes and totaling 621 Mbp and 615 Mbp, respectively. The assembly achieved a high contiguity, with an N50 of 51 Mb, and a GC content of 33.77%. Assembly completeness was assessed using 2326 eudicot BUSCOs, revealing 97.8% complete genes, with 5.6% duplication and only 0.6% missing for Hap1 and respectively 98.1% complete, 5.4% duplicated and 0.6% missing genes for Hap2. Both haplotypes exhibited high consensus accuracy, with QV scores of 75.7%. A total set of genes models were predicted in *A. spinosa* accounting for 35,183 genes, among which 76.46% were annotated to at least one of the publicly available databases. This near-T2T phased genome assembly establishes a valuable platform for understanding the genetic basis of adaptation and resilience in adverse environments, as well as advancing conservation and genetic improvement efforts. All sequencing data supporting the assembly and annotations are publicly available, ensuring reproducibility and facilitating further research.

Methods

Sample collection, genomic DNA extraction. The sequenced *Argania spinosa* tree, named TAGUERTE (S7P2), was selected for its biological and ecological properties, it was collected from the Souss Plain Valley, Taguerte commune located at (31°01′52.3″N 9°22′43.1″W) (Fig. 1). Fresh young leaves of *A. spinosa* were collected and immediately frozen in liquid nitrogen and stored in a -80°C deep freezer for genome sequencing. High-Molecular-Weight (HMW) DNA (>50 kb fragments) was extracted using the QIAGEN MagAttract HMW DNA kit. DNA quality was assessed using NanoDrop, Qubit fluorometer, and agarose gel electrophoresis[10].

For transcriptomics, *Argania spinosa* seeds were obtained from the University of Arizona Arboretum. To initiate germination, the hard shells were removed, and the kernels were surface sterilized by soaking in 70% ethanol for 5 minutes, followed by 2–3 rinses with sterile distilled water. The kernels were then immersed in a 5% sodium hypochlorite solution with a few drops of Tween-20 for 10–15 minutes and subsequently rinsed thoroughly with sterile distilled water five times to eliminate any bleach residue. Sterilized kernels were placed on moist filter paper in Petri dishes and incubated at 25°C under a 12-hour light/dark cycle. After three weeks, germinated seeds were transferred into pots and maintained in a growth chamber under the same conditions. Roots and true leaves were harvested from three-month-old seedlings.

Library preparation and sequencing. For a hybrid de novo assembly approach combining short and long sequencing reads, both PacBio HiFi and Illumina platforms were used (Table 1). The experiments were carried out following PacBio's standard protocol, which included sample quality assessment, library preparation, library quality assessment, and sequencing. The library preparation process involved several key steps: DNA fragmentation using g-TUBE, followed by nick and end repairing. SMRTbell adapters were then ligated to the prepared DNA fragments, and exonuclease treatment was applied to degrade any unsuccessful ligation products. Finally, size-selected fragments were purified using BluePippin and prepared for sequencing. PacBio HiFi (CCS) sequencing was performed using two Revio SMART cells (PacBio Revio, Biomarker Technologies, Hong Kong).

The Hi-C library was prepared through several key steps: First, DNA cross-linking was performed by fixing tissue samples in methanol to preserve protein-DNA interactions and 3D structure. Next, DNA was digested using HindIII restriction enzyme to generate sticky ends. Then the fragment ends were then repaired to blunt ends and biotinylated for DNA capturing and purification. Following this, blunt-ended DNA molecules were circularized through end ligation, bringing interacting DNA fragments together. Finally, cross-links were reversed, and DNA fragments ranging from 300 to 700 bp were purified using magnetic beads for library preparation. The Concentration and insert size of the library were examined by Qubit2.0 and Agilent 2100 respectively. Q-PCR were further processed for accurate DNA quantification to ensure an adequate amount and libraries passed quality assessment were processed for sequencing on Novaseq 6000 2 × 150bp (Biomarker Technologies, Hong Kong). A total of 101.25 (92.24% with >Q30) Gbp data was yielded from 1 Hi-C library corresponding to more than 150X in coverage. Quality of a Hi-C library generally refers to the percentage of valid data. In projects without reference genomes, data assessment is based on a ratio of truncated reads. Truncated reads are defined as reads with an enzyme cutting site, i.e. reads containing two different fragments. In a good Hi-C library, the ratio of truncated reads in total reads is expected to be above 10%[11].

For RNA-seq, total RNA was extracted from root and leaf tissues using the Direct-zol™ RNA Kits (Zymo Research, #R2050, Irvine, California, USA) according to the manufacturer's instructions. RNA from seeds was isolated using the RNeasy PowerPlant Kit (Qiagen, #13500-50), followed by DNase I treatment (Qiagen, #79254, Valencia, California, USA). mRNA was purified from total RNA using the Qiagen RNeasy Pure mRNA Bead Kit (#180244), and RNA quality was assessed with the Agilent TapeStation 2200 using RNA ScreenTape (Agilent Technologies Inc., Santa Clara, California, USA). RNA fragmentation, double-stranded cDNA synthesis, and adapter ligation were performed with the NEBNext Ultra II Directional RNA Library Prep Kit (NEB, #E7760L) following the manufacturer's protocol. PCR-enriched libraries were quantified using PicoGreen (Thermo Fisher Scientific), and equimolar libraries were pooled. Pooled libraries were assessed on the Agilent TapeStation 2200 and quantified by qPCR. Libraries were diluted to 200 pM and spiked with 1% PhiX control libraries (Illumina). Transcriptome sequencing was carried out using barcoded, stranded RNA-Seq libraries on an Illumina NovaSeq 6000 S1 flow cell with paired-end reads (2 × 100 bp).

Genome survey analysis of Argan Tree. After filtering the raw data and removing adapters, the genome size estimation of the argan tree was performed based on the 27-mer frequency distribution from PacBio data. The analysis was carried out using KMC[12] and GenomeScope2[13] using default parameters. According to Fig. 2, the 27-mer distribution is bimodal, showing two peaks at 647 and 12840, indicating the heterozygous nature of the *A. spinosa* genome. Based on the total number of 27-mers, the genome size is estimated to be 645 Mbp with a high heterozygosity rate of 1.48% which is very close to the total size of the assembled genome sequence (Fig. 2).

Contig-level assembly. To generate high-quality contigs from *Argania spinosa* sequencing data, we used hifiasm[14–16] in phased mode, which supports the use of both PacBio HiFi reads and Hi-C data to produce haplotype-resolved diploid assemblies. To improve assembly quality beyond the default parameters, we employed Mabs v2.28[17], a tool that explores multiple parameters sets and selects the best assembly based on the Accurately assembled Genes (AG) score calculated from BUSCO[18] reconstruction. This allowed us to select the most gene-complete and structurally accurate phased assemblies.

The assembler constructs an assembly graph using the overlap-layout-consensus strategy and resolves the haplotypes using Hi-C contact information. Due to the moderate level of heterozygosity estimated from k-mer analysis, we enabled hifiasm's internal default purging mode, which performs three rounds of duplicate sequence removal and correction during graph resolution.

A total of 4.7 million HiFi reads, and 676 million Hi-C read pairs, with a truncated rate of 24%, were used to assemble the genome. Genome characteristics were estimated using GenomeScope2[13] based on k-mer analysis (k = 27) of the HiFi reads. The estimated genome size was approximately 645 Mb, with a heterozygosity level of 1.48%, indicative of substantial allelic variation. The k-mer spectrum showed distinct peaks for homozygous and heterozygous regions, consistent with a diploid genome (Fig. 2).

The final phased contig assemblies consisted of 1389 contigs for Haplotype 1, with a total size of 740 Mb, N50 of 43 Mb, and GC content of 35%. Haplotype 2 consisted of 360 contigs, with a size of 648 Mb, N50 of 42 Mb, and GC content of 34% (Table 2). The base-level quality values (QV) were estimated to be 56.5 and 61.5 for the two haplotypes, respectively. BUSCO analysis against the Eudicotyledons odb10 lineage dataset showed 97.8% complete BUSCOs for Haplotype 1 and 98.1% for Haplotype 2, with low duplication rates (5.6% and 5.4%, respectively).

Purging of duplicated sequences. To further reduce residual redundancy and eliminate duplicated haplotypic segments not fully resolved during contig assembly, we applied purge_dups v1.2.5[19] independently to each haplotype. This tool identifies and removes contig-level duplications based on coverage patterns and sequence similarity.

The purging step removed redundant or overlapping sequences per haplotype, improving haplotype separation and contributing to the overall assembly quality. Post-purging assemblies retained high completeness and QV scores, with BUSCO duplication slightly reduced compared to the unpurged contigs. Haplotype 1 retained 63 contigs with total size pf 649 Mb, N50 of 45 Mb, GC content of 34%, QV of 70.3, 98.0% complete and 5.5% duplicated BUSCOs. Purged Haplotype 2 has 21 contigs with total size of 617 Mb, N50 of 42 Mb, GC content of 34%, QV of 75.2, 98.1% complete and 5.5% duplicated BUSCOs.

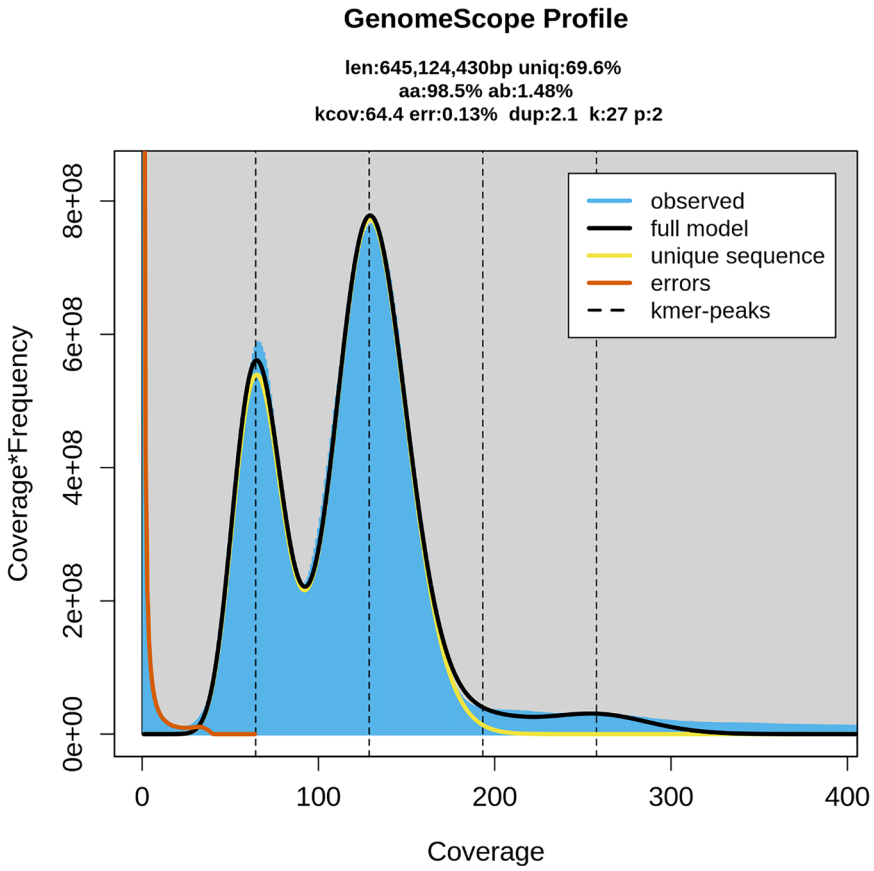

Fig. 2 K-mer profile (k = 27) spectral analysis to estimate genome size of *A. spinosa*. The two peaks that represent two distinct k-mer frequency distributions: The first peak (lower frequency) corresponds to heterozygous k-mers, suggesting variation between homologous chromosomes. The second peak (higher frequency) represents homozygous k-mers, which are present in both copies of the genome.

	hap1_ctg	hap2_ctg	hap1.chr	hap2.chr
Contigs/Scaffolds	1389	360	11	11
Largest Contig	86031520	110790353	115611259	114277299
Total Length	739994486	647600981	621462413	614616934
GC (%)	35.11	34.40	33.77	33.79
N50	43310576	42393772	50247898	50701409
N90	936536	24348970	34570747	33712129
auN	41476630.5	50426414.8	68915520.4	67424947.4
L50	7	6	4	4
L90	23	13	10	10
# N's	0.00	0.00	0.19	0.15

Table 2. Statistics of *A. spinosa* genome assemblies.

Scaffold-level assembly. After generating haplotype-resolved contig assemblies, we scaffolded each haplotype using Hi-C data and the YaHS v1.2.2[20]. Hi-C paired-end reads were first aligned to the purged contig assemblies using BWA v0.7.19[21] to obtain BAM files capturing chromatin proximity. YaHS was then used to construct chromosome-scale scaffolds by leveraging the contact matrix and known Hi-C cut site patterns.

Scaffolded assemblies were indexed and evaluated with gfastats v1.3.10[22] for standard N-statistics and structural summary. Final scaffolded FASTA files were compressed and indexed for downstream applications.

Following scaffolding, we remapped the Hi-C reads to the scaffolded assemblies and generated contact maps using two independent pipelines: PretextMap v0.1.9[23] for visualization in PretextView[24], incorporating telomere annotations and HiFi coverage tracks, and Juicer v1.2.2 format contact maps for manual curation in Juicebox Assembly Tools (JBAT)[25].

To evaluate HiFi read support, reads were aligned back to the scaffold assemblies, and coverage tracks (bedGraph and bigWig) were generated. These were also incorporated into contact map visualizations.

The scaffolded assemblies for both haplotypes of *A. spinosa* were successfully constructed into 11 pseudochromosomes, consistent with the species' karyotype (2n = 22).

Hi-C contact data enabled high-contiguity scaffolding with minimal need for manual correction. Scaffold-level N50 increased to 50 Mb and 51 Mb for haplotypes 1 and 2, respectively, and the assemblies spanned 649 Mb and 617 Mb. L90 values showed that 90% of the genome content was contained in 10 and 10 scaffolds, respectively.

Contact map visualization in both PretextView[24] and Juicebox[25] confirmed strong chromatin contact patterns within individual chromosomes and minimal interchromosomal noise, indicating correct scaffold placement. We detected (AAACCCT)n motifs at both ends of most chromosomes; a few termini lack detectable repeats, likely due to small unresolved gaps; hence we describe the assemblies as near-T2T. Centromeric regions were inferred from characteristic drops in Hi-C signal and HiFi coverage near the center of several scaffolds.

BUSCO completeness remained high in both assemblies post-scaffolding, with 98.1% and 98.0% of genes complete in haplotypes 1 and 2, respectively. Merqury analysis showed high consensus accuracy, with estimated QV values of 70.3 and 75.2. No significant sequence loss was observed during scaffolding.

Manual curation and chromosome-level assembly. To finalize the chromosome-scale assemblies, we manually curated each haplotype after scaffolding. For both haplotypes, preliminary Hi-C contact maps were visualized using Juicebox Assembly Tools and PretextView. We manually inspected scaffold structures, Hi-C link density, and misjoins using *.hic* and *.assembly* files generated by Juicebox, and *.pretext* maps generated with PretextMap[23]. Based on these inspections, we created a curated list of scaffolds to retain in the final assembly.

Selected scaffolds were extracted using seqtk (https://github.com/lh3/seqtk) and compiled into new FASTA files representing the final chromosome-level assemblies. These assemblies were indexed with BWA, evaluated structurally with gfastats, and compressed with bgzip for efficient downstream processing.

Hi-C reads were then remapped to the final assemblies, and contact maps were re-generated with both Juicer and PretextMap pipelines for final visualization. Additionally, HiFi reads were mapped back to assemblies to assess sequence coverage and validate completeness across chromosomes. BedGraph[26] and BigWig[27] tracks were generated to visualize coverage uniformity.

The final chromosome-level assemblies were re-evaluated with BUSCO, Merqury, and gfastats. BUSCO was run using the eudicots_odb10, and Merqury[28] was used to estimate consensus QV and k-mer completeness using the full HiFi read set.

Manual inspection led to the removal of a small number of low-confidence or unanchored scaffolds per haplotype. The final curated assemblies consist of 11 scaffolds per haplotype, corresponding to the expected number of chromosomes (2n = 22). No contamination was detected after inspection, and all retained scaffolds showed consistent Hi-C signal and strong internal contact density.

HiFi coverage was uniformly distributed across each pseudochromosome, except for regions of reduced coverage and contact signal, consistent with centromeric regions. Telomeric repeats were detected at the ends of most chromosomes, supporting the classification of both haplotype assemblies as near near-T2T.

Final chromosome-level assemblies measured approximately 621 Mb and 615 Mb for haplotypes 1 and 2, respectively. BUSCO completeness scores were 97.8% and 98.1%, with low duplication and fragmentation rates. Merqury analysis confirmed high k-mer completeness and base-level accuracy, with consensus QV scores of 75.9 and 75.7, respectively.

Contact maps generated from the final assemblies showed high intra-chromosomal contact density and minimal inter-chromosomal noise, further confirming the structural accuracy of the curated chromosomes. These assemblies are therefore considered high-quality, phased, and near-complete chromosome-level references for Argania spinosa. Final contact maps and JBAT Hi-C visualization plots are shown in Fig. 3c and d.

Repeats annotation. Repetitive elements in the *Argania spinosa* genome assembly were identified and annotated using a combination of de novo and library-based approaches. We first employed RepeatModeler v2.0.6[29] to construct a custom repeat library from the assembly. This tool integrates algorithms such as RECON v1.08[30] and Tandem Repeats Finder v4.09.1[31] to detect and cluster repetitive regions into consensus repeat families without relying on prior knowledge. Identified repeat sequences were classified into known families when possible and labeled as "Unknown" otherwise.

To improve classification of these unknown families, we used the repclassifier script with Repbase and Dfam databases, targeting the Ericales (NCBI taxon ID: 41945) lineage. This step enabled the classification of species-specific unknown repeats based on homology to known elements from phylogenetically related taxa.

Annotation and masking were performed in four iterative rounds using RepeatMasker v4.1.9. The rounds included: Simple repeat masking using lineage-specific libraries; Masking of lineage-specific interspersed repeats from Repbase/Dfam; Masking of classified species-specific repeats from the de novo library; Masking of unclassified species-specific repeats from the de novo library.

The outputs of each round were merged to produce a comprehensive annotation of all repeat types. Final results were summarized using ProcessRepeats and parsed into GFF3 format for visualization and further analysis. We also created soft-masked and hard-masked versions of the genome using BEDTools. Repeat composition summaries were generated using custom scripts to quantify genomic coverage per repeat family.

Repeat annotation of the primary *Argania spinosa* haplotype revealed that 61.65% of the genome assembly consists of repetitive sequences, corresponding to around 379 Mbp. The majority were interspersed repeats, including Retrotransposons (24.24%), primarily composed of LTR elements (18.66%), with Gypsy and Copia being the dominant subfamilies. Non-LTR elements included LINEs (5.57%). DNA transposons and simple repeats constituted 5.86% and 2.79% respectively. A total of 26.07% were designed as Unclassified repeats (Table 3, Fig. 3).

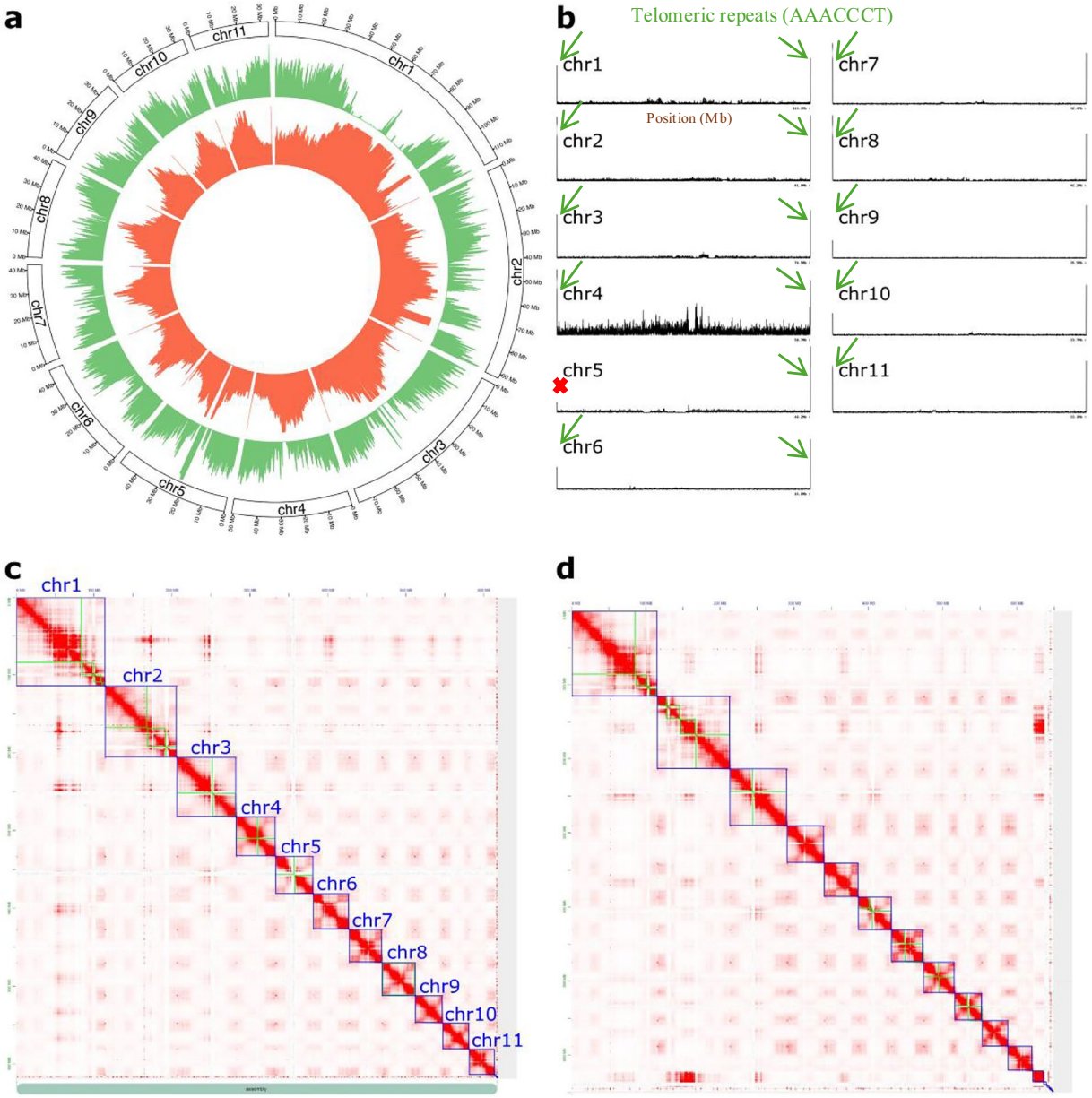

Fig. 3 Circos and Hi-C representations of the *A. spinosa* Chromosomal Assembly. **(a)** Circos plot showing the chromosomal assembly of *A. spinosa* Hap2. The outermost circle represents the 11 chromosomes. The orange track indicates transposable element (TE) density, and the green track shows gene density using a sliding window of 0.5 Mb with a step size of 50 Kb. **(b)** Frequency of the telomeric repeat motif (AAACCCT) across the 11 chromosomes, highlighting telomeric ends in most chromosomes. **(c–d)** Hi-C contact matrices for Hap1 **(c)** and Hap2 **(d)**, each displaying 11 distinct chromosomal interaction blocks.

The repeat-soft-masked assemblies were subsequently used for gene prediction and downstream analyses. BUSCO evaluation on masked versus unmasked assemblies confirmed that repeat masking had minimal effect on gene completeness, with all versions maintaining over 91.7% of conserved single-copy orthologs.

Gene prediction and functional annotation. Structural and functional annotation of the *Argania spinosa* genome was performed using the Funannotate v1.8.17 pipeline[32]. We used a soft-masked version of the chromosome-level haplotype assemblies as input, along with RNA-Seq data from leaf tissue and protein homology evidence from closely related species (*Ericales*, *Vitellaria paradoxa*) and the SwissProt database.

Transcriptome-based training was first carried out using PASA v2.5.3[33] and Trinity v2.15.2, with alignments generated via minimap2 v2.28[34] and BLAT v39 × 1, and a maximum intron length of 50 kb. Gene prediction was then conducted with trained ab initio models (AUGUSTUS v3.5.0, GeneMark-ES v4.71)[35,36], guided by transcript and protein evidence. The predicted models were further refined using RNA-Seq-based strand inference and Funannotate's update module, which adds untranslated regions (UTRs).

Repeat class	Repeat subclass	Number of elements	Total length (bp)	(%)
Retroelements		227433	148952660	24.24
	SINEs:	0	0	0.00
	Penelope:	0	0	0.00
	LINEs:	63240	34254187	5.57
	CRE/SLACS	0	0	0.00
	L2/CR1/Rex	0	0	0.00
	R1/LOA/Jockey	349	248958	0.04
	R2/R4/NeSL	0	0	0.00
	RTE/Bov-B	17420	11254188	1.83
	L1/CIN4	45471	22751041	3.70
	LTR elements:	164193	114698473	18.66
	BEL/Pao	0	0	0.00
	Ty1/Copia	80611	49516075	8.06
	Gypsy/DIRS1	76318	59002054	6.49
	Retroviral	184	134329	0.02
DNA transposons:		136951	36018568	5.86
	hobo-Activator	44954	10374408	1.69
	Tc1-IS630-Pogo	347	158606	0.03
	En-Spm	0	0	0.00
	MULE-MuDR	39364	13971183	2.27
	PiggyBac	0	0	0.00
	Tourist/Harbinger	44504	8730006	1.42
	Other	0	0	0.00
Rolling-circles		36378	7976706	1.30
Unclassified		517559	160206534	26.07
Small RNA		5402	2618163	0.43
Simple repeats		437162	17171639	2.79
Low complexity		113949	5990170	0.97

Table 3. Statistics of TE identified in *A. spinosa* genome sequence.

For functional annotation, motifs and protein domains were identified using InterProScan v5.74-105.0[37] and a locally installed version of EggNOG-mapper v2.1.12[38], which compared sequences against public databases. The annotation results from InterProScan and EggNOG-mapper were then merged using the "funannotate annotate" function with defaults parameters.

We annotated 35,183 gene loci producing 39,805 mRNA transcript isoforms and 410 tRNA genes. The average gene length was approximately 5,275 bp. 75.5% of gene loci had exons supported by RNA-seq alignments. Functional annotation was conducted using InterProScan v5.74-105.0, eggNOG v2.1.12, Pfam[39], and Gene Ontology, with support from DIAMOND v2.1.10, MMseqs2 v13-45111+ds-2, and HMMER v3.4[40,41]. Functional categories were assigned to a large proportion of genes: 24,108 with GO terms, 29,679 with InterPro domains, and 32,706 with eggNOG orthologs.

Data Records

The data that supports the findings (PacBio HiFi, Hi-C and RNA-seq) of this study have been deposited at the Sequence Read Archive (SRA) under accession number SRP565314[42].

The Whole Genome Shotgun project has been deposited at the INSDC under accession JBPKRC000000000[43] and JBPKRD000000000[44] respectively for haplotype 1 and 2. The versions described in this paper are versions JBPKRC000000000.1 and JBPKRD000000000.1

Additionally, the corresponding genome annotation is publicly accessible via Zenodo[45].

Technical Validation

The quality of the *Argania spinosa* genome assembly was assessed using multiple complementary approaches. Hi-C contact maps revealed well-organized interaction patterns along the main diagonal, consistent with a chromosome-scale assembly comprising 11 chromosomes for both haplotypes (Fig. 3). This structural organization aligns with the previously reported karyotype ($2n = 2x = 22$)[46]. In addition, telomeric repeat sequences (AAACCCT) were detected at the ends of nearly all chromosomes, supporting their completeness. Notably, telomeric motifs were missing at one end of chromosome 5 (Fig. 3b).

Assembly completeness was further validated using BUSCO (Benchmarking Universal Single-Copy Orthologs) with the eudicotes_odb10 dataset, showing 97.8% and 98.1% of complete BUSCOs for hap1 and hap2 respectively. These results highlight the high level of gene space completeness. The total assembled genome size (615–621 Mbp) closely matched the genome size estimated by k-mer analysis (645 Mbp), indicating consistency between genome survey and final assembly.

Base-level accuracy was estimated using QV (quality value) scores, with both haplotypes exhibiting high confidence values of 75.5%, reflecting low consensus error rates. Repetitive elements and gene content were annotated using a combination of de novo, homology-based, and transcriptome-supported prediction strategies, resulting in 61.65% of repetitive fraction of *A. spinosa* genome and the identification of 40,630 high-confidence gene models.

Data availability

The data supporting this study, including PacBio HiFi, Hi-C, and RNA-seq datasets, have been deposited in the Sequence Read Archive (SRA) under accession number SRP565314[42]. The Whole Genome Shotgun project has been deposited at the INSDC under the accession JBPKRC000000000[43] and JBPKRD000000000[44] respectively for haplotype 1 and 2. The versions described in this paper are JBPKRC000000000.1 and JBPKRD000000000.1. Additionally, the corresponding genome annotation is publicly accessible in Zenodo[45].

Code availability

The scripts used to generate the figures and run the analysis are available through the following GitHub repository (https://github.com/SolayMane/Argania_sciData/tree/main). All bioinformatics tools used in this study followed their respective protocols and manuals. If specific parameters are not mentioned, the default parameters were used. The versions of the software used are indicated in the Methods section.

Received: 26 February 2025; Accepted: 12 January 2026;
Published online: 11 February 2026

References

1. Khayi, S. *et al*. First draft genome assembly of the argane tree (argania spinosa). *F1000Research* **7**, 1310, https://doi.org/10.12688/f1000research.15719.1 (2018).
2. Lybbert, T. J., Aboudrare, A., Chaloud, D., Magnan, N. & Nash, M. Booming markets for moroccan argan oil appear to benefit some rural households while threatening the endemic argan forest. *Proceedings of the National Academy of Sciences* **108**, 13963–13968, https://doi.org/10.1073/pnas.1106382108 (2011).
3. Charrouf, Z. & Guillaume, D. Sustainable development in northern africa: The argan forest case. *Sustainability* **1**, 1012–1022, https://doi.org/10.3390/su1041012 (2009).
4. Charrouf, Z. & Guillaume, D. Ethnoeconomical, ethnomedical, and phytochemical study of it argania spinosa (l.) skeels. *Journal of Ethnopharmacology* **67**, 7–14, https://doi.org/10.1016/S0378-8741(98)00228-1 (1999).
5. Laaribya, S., Alaoui, A. & Gmira, N. The moroccan forest and sustainable development case of the argan tree it argania spinosa l. skeels in morocco. *Biological Diversity and Conservation* **10**, 1–7, https://dergipark.org.tr/en/pub/biodicon/issue/55731/762173 (2017).
6. Bellefontaine, R., Ferradous, A., Alifriqui, M. & Monteuuis, O. Multiplication vegetative de l'arganier, it argania spinosa, au maroc: le projet john goelet. *Bois et Forets des Tropiques* **(304)**, 47–59, https://doi.org/10.19182/bft2010.304.a20446 (2010).
7. Alaoui, M. M. *et al*. Induction of rhizogenesis of in vitro culture of cuttings of it argania spinosa (l). skeels. In *In Vitro Cellular & Developmental Biology–Plant* volume 54, page S69, New York, Springer (2018).
8. Falasca, G. *et al*. The potential growing areas for it argania spinosa (l.) skeels (sapotaceae) in argentinean drylands. *International Journal of Agronomy*, https://doi.org/10.1155/2018/9262659 (2018).
9. Ministry of Agriculture, Maritime Fisheries, Rural Development, and Water and Forests. Report on oasis and argan zones development (2012–2022). Government report, Exact title and URL to be confirmed (2022).
10. Khayi, S. *et al*. First draft genome assembly of the argane tree (argania spinosa). *F1000Research* **7**, 1310, https://doi.org/10.12688/f1000research.15719.2 (2020).
11. Xie, T. *et al*. De novo plant genome assembly based on chromatin interactions: a case study of *Arabidopsis thaliana*. *Molecular Plant* **8**, 489–492, https://doi.org/10.1016/j.molp.2014.12.015 (2015).
12. Kokot, M., Długosz, M. & Deorowicz, S. Kmc 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761, https://doi.org/10.1093/bioinformatics/btx304 (2017).
13. Vurture, G. W. *et al*. Genomescope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153 (2017).
14. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
15. Cheng, H. *et al*. Haplotype-resolved assembly of diploid genomes without parental data. *Nature Biotechnology* **40**, 1332–1335, https://doi.org/10.1038/s41587-022-01261-x (2022).
16. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nature Methods* **21**, 967–970, https://doi.org/10.1038/s41592-024-02269-8 (2024).
17. Schelkunov, M. I. Mabs, a suite of tools for gene-informed genome assembly. *BMC Bioinformatics* **24**, 377, https://doi.org/10.1186/s12859-023-05499-3 (2023).
18. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. Busco update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).
19. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).
20. Zhou, C., McCarthy, S. A. & Durbin, R. Yahs: yet another hi-c scaffolding tool. *Bioinformatics* 39, https://doi.org/10.1093/bioinformatics/btac808 (2023).
21. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
22. Formenti, G. *et al*. Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics* **38**, 4214–4216, https://doi.org/10.1093/bioinformatics/btac460 (2022).
23. Guan, D., Harry, E. & Sanger Institute Tree of Life. Pretextmap: Paired read texture mapper. GitHub repository, Accessed 2025-10-30, https://github.com/sanger-tol/PretextMap (2025).
24. Guan, D., Harry, E. & Sanger Institute Tree of Life. Pretextview: Opengl powered pretext contact map viewer. GitHub repository, Accessed 2025-10-30, https://github.com/sanger-tol/PretextView (2025).
25. Durand, N. C. *et al*. Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. *Cell Systems* **3**, 99–101, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
26. UCSC Genome Browser. Genome browser bedgraph track format. Online documentation, Accessed 2025-10-30, https://genome.ucsc.edu/goldenpath/help/bedgraph.html (2025).

27. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. Bigwig and bigbed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207, https://doi.org/10.1093/bioinformatics/btq351 (2010).
28. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
29. Flynn, J. M. *et al.* Repeatmodeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
30. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research* **12**, 1269–1276, https://doi.org/10.1101/gr.88502 (2002).
31. Benson, G. Tandem repeats finder: a program to analyze dna sequences. *Nucleic Acids Research* **27**(2), 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
32. Palmer, J. M. & Stajich, J. E. Funannotate v1.8.1: Eukaryotic genome annotation. Zenodo, https://doi.org/10.5281/zenodo.1134477 (2020).
33. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. *Genome Biology* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
34. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
35. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. Augustus: a web server for gene finding in eukaryotes. *Nucleic Acids Research* **32**, W309–W312, https://doi.org/10.1093/nar/gkh379 (2004).
36. Besemer, J. Genemarks: a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research* **29**, 2607–2618, https://doi.org/10.1093/nar/29.12.2607 (2001).
37. Jones, P. *et al.* Interproscan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
38. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggnog-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution* **38**, 5825–5829, https://doi.org/10.1093/molbev/msab293 (2021).
39. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Research* **42**, D222–D230, https://doi.org/10.1093/nar/gkt1223 (2014).
40. Eddy, S. R. Accelerated profile hmm searches. *PLoS Computational Biology* **7**, e1002195, https://doi.org/10.1371/journal.pcbi.1002195 (2011).
41. Sahbou, A.-E., Iraqi, D., Mentag, R. & Khayi, S. Buscophylo: a webserver for busco-based phylogenomic analysis for non-specialists. *Scientific Reports* **12**, 1–5, https://doi.org/10.1038/s41598-022-22461-0 (2022).
42. El Idrissi, H. *et al.* Argania spinosa sequencing reads (hifi/hi-c/rna-seq). *SRA*, Accessed 2025-10-30. http://identifiers.org/insdc.sra:SRP565314 (2025).
43. El Idrissi, H. *et al.* Argania spinosa phased genome assembly, haplotype 1. *INSDC*, http://identifiers.org/insdc:JBPKRC000000000 (2025).
44. El Idrissi, H. *et al.* Argania spinosa phased genome assembly, haplotype 2. *INSDC*, http://identifiers.org/insdc:JBPKRD000000000 (2025).
45. El Idrissi, H. *et al. Zenodo*, https://doi.org/10.5281/zenodo.16017912 (2025).
46. Majourhat, K. *et al.* Karyotype characterization of it argania spinosa (l.) skeel (sapotaceae). *South African Journal of Botany* **73**, 661–663, https://doi.org/10.1016/j.sajb.2007.06.007 (2007).

Acknowledgements

This work benefited from support of the Midterm Research Program (2020 to 2024) of INRA-Morocco and by the MCGP INRA-ICARDA program. We would like to thank Dr. Sadik Muzemil for his help during the bioinformatics analysis.

Author contributions

S.K. and R.M. designed the project and performed the sample collection. S.K., A.G. and H.E. designed the scripts and performed data analysis. S.K., A.G., H.E., R.M., B.B., D.I. and R.Mo. drafted the manuscript. M.F. and S.Kho. performed the sequencing. Authors contributed to data interpretation, edited, and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.M. or S.K.