



OPEN

DATA DESCRIPTOR

Chromosome-level phased genome assembly of the argan tree *Sideroxylon spinosum*

Ivan D. Mateus¹, Abdellatif Essahibi¹, Pamela Nicholson², Mohamed Hijri³, Ahmed Qaddoury⁴, Laurent Falquet^{1,5} & Didier Reinhardt¹

Argan (*Sideroxylon spinosum* L., formerly *Argania spinosa*) is a tree endemic to Morocco, primarily valued for its seed oil. Growing interest in its biology and in genes linked to oil quality and stress resistance highlights the need for high-quality genome and transcriptome models. We integrated PacBio HiFi long-read and Illumina Hi-C sequencing data to generate independently assembled, phased genome models for both parental haplotypes, measuring 636 Mb and 655 Mb, respectively, with BUSCO completeness scores exceeding 97.8%. Each haplotype consists of 11 fully resolved telomere-to-telomere chromosomes, consistent with chromosome numbers in other Sapotaceae species ($n = 10\text{--}13$), and contains approximately 60% repetitive sequences. Annotation predicted ~28,720 protein-coding genes per haplotype. Comparative analyses with other Sapotaceae genomes indicate overall chromosome conservation within the family, alongside repeat expansion and fusion events in the two largest chromosomes (chr1 and chr2). We also independently assembled the complete chloroplast genome. This high-quality assembly provides a valuable resource for future research on argan biology, genetic diversity, and traits relevant to adaptation and oil biosynthesis.

Background & Summary

Argan is a member of the Sapotaceae family (order Ericales), which comprises five tribes and approximately 1,250 species. This family includes economically important species valued for their oil, such as the argan tree (*Sideroxylon spinosum* L., formerly *Argania spinosa* (L.) Skeels), the shea tree (*Vitellaria paradoxa* C.F. Gaertn.), and the miracle fruit tree (*Synsepalum dulcificum* (Schumacher & Thonn.) Daniell). Argan seeds contain an edible oil extensively used for both culinary and cosmetic applications. Argan oil represents a crucial economic resource for many Moroccan families¹, particularly for women, who play a central role along the entire value chain. Argan is a diploid species with a reported haploid chromosome number ranging between 10 and 12².

Argan has not been selectively bred or extensively cultivated, instead, it grows in its natural habitat, where it has been sustainably utilized for centuries. Nevertheless, overgrazing by goats, the intensification of land use in Morocco, and the high demand for argan oil have led to unsustainable exploitation, putting the species at risk¹. In recent years, rising temperatures, prolonged droughts, and an extended dry season have further exacerbated the challenges for the argan-growing region, known as the “Arganeraie”, which has been listed as UNESCO biosphere heritage in 1998.

The economic importance of argan has raised the demand for high-quality genomic resources. A first genome draft containing 75,327 scaffolds³, was followed by a reference annotation with 62,590 genes⁴. The complete mitochondrial and chloroplast genomes have also been released in recent years^{5,6}. Additionally, single-sequence repeat markers⁷ and AFLP markers⁸ have been developed for the use in population genetics studies, revealing high genetic diversity between populations but low diversity within populations^{7–9}. However, a high-quality genome sequence is required for a more comprehensive understanding of argan evolution, physiology, population genomics, as well as for marker-assisted breeding.

¹Department of Biology, University of Fribourg, Chemin du Musée 10, 1700, Fribourg, Switzerland. ²Next Generation Sequencing Platform, University of Bern, Bremgartenstrasse 109a, 3012, Bern, Switzerland. ³African Genome Center, University Mohammed VI Polytechnic (UM6P), Lot 660, Hay Moulay Rachid, 43150, Ben Guerir, Morocco.

⁴Department of Biology, Faculty of Sciences and Techniques, University of Cadi Ayyad, Marrakesh, Morocco. ⁵Swiss Institute of Bioinformatics, Lausanne, Switzerland. ✉e-mail: laurent.falquet@unifr.ch; didier.reinhardt@unifr.ch

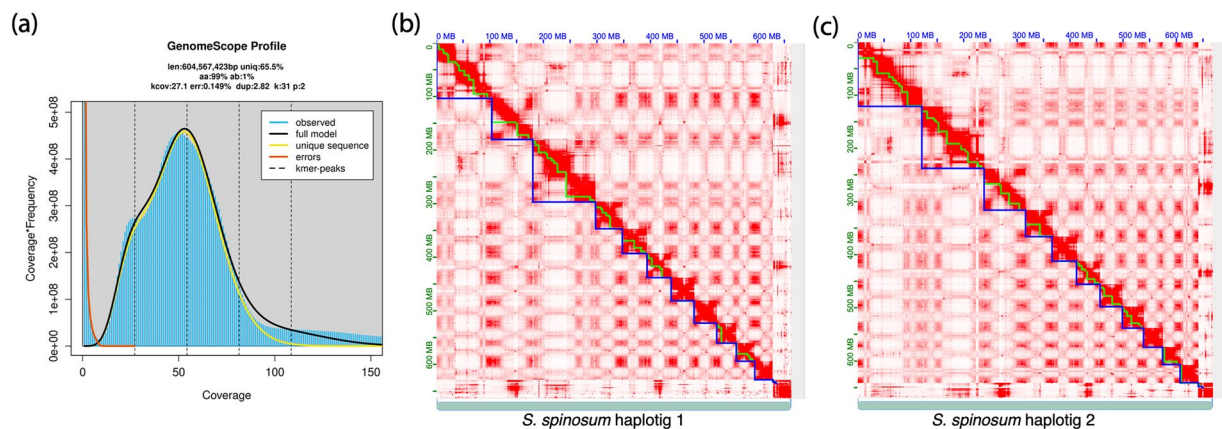


Fig. 1 Global genome characterization for *S. spinosum*. (a) GenomeScope genome size estimates for *S. spinosum*. (b,c) Genome-wide chromosomal Hi-C maps of haplotig 1 (b) and haplotig 2 (c) of *S. spinosum*; chromosomes are delineated in blue, and contigs are highlighted with a green outline.

In this study, we used PacBio HiFi sequencing combined with Illumina Hi-C to generate a complete genome assembly of the argan tree with two independently assembled haplotigs and the chloroplast genome. We compared our phased assembly with genome assemblies of two other Sapotaceae species to investigate chromosome composition and evolution. Our telomere-to-telomere genome assembly provides a high-quality reference for comparative genomics, offering new insights into chromosome evolution in this species.

Methods

Sample collection and sequencing. Leaf samples of *Sideroxylon spinosum* (syn. *Argania spinosa*) were collected from a specimen at the northern border of the UNESCO biosphere reserve “l’Arganaie”, along the National Road 207 east of Essaouira at coordinates 31°32′47.9″ N 9°21′56.6″ W (Fig. S1). We employed a hybrid sequencing strategy that combined PacBio long-read sequencing with Hi-C chromatin conformation data. Genomic DNA was extracted from leaves using a modified CTAB protocol (Supplementary Methods). For Hi-C sequencing, nuclei were isolated using a Sucrose/Percoll gradient centrifugation protocol (Supplementary Methods). To generate highly accurate long reads, we performed Pacific Biosciences (PacBio) Sequel II HiFi sequencing on sheared genomic DNA (gDNA). Hi-C sequencing was carried out using Illumina NovaSeq 6000 (paired-end, 150 bp) on crosslinked chromatin, which was enzymatically digested with DpnII (GATC) and HinfI (G^ANTC) using the Arima Genomics Hi-C kit (Cat. #A510008), followed by proximity ligation to capture three-dimensional chromosomal interactions.

Genome assembly. The genome of *S. spinosum* was assembled by integrating long-read PacBio HiFi sequencing with Hi-C data to establish physical linkage. High sequence coverage allowed for the independent assembly of the two parental genomes, each consisting of 11 chromosomes. PacBio HiFi sequencing reads were quality-checked using FastQC v.0.11.9 and cleaned with fastp v.0.19.5¹⁰. The cleaned HiFi reads were assembled into phased haplotigs using hifiasm v.0.16.1 with `-hom-cov 50` parameters¹¹. K-mer analysis performed with GenomeScope 2.0 (<http://genomescope.org>)¹² revealed a low error rate and a high homozygosity between the two haplotigs (Fig. 1a). The haplotigs were further scaffolded using Illumina Hi-C data. We applied HiCup to map the Hi-C reads to the hifiasm assemblies¹³, and we used Juicer v1.6¹⁴ and YAHS v1.2.2¹⁵ to generate the final telomere-to-telomere (T2T) chromosome assemblies. The YAHS file “telo.c” was customized to include the telomeric sequences of *S. spinosum* TTTAGGG & GGGTGGG and the software was recompiled before generating the final T2T chromosome assemblies. The final assembly was manually curated with JuiceBox (v2.15)¹⁶ (Fig. 1b,c). The telomeric regions were identified using tidk (v0.2.31)¹⁷ with the canonical telomeric sequences TTTAGGG and GGGTGGG (Fig. S2).

We employed Nucmer¹⁸ and dot (<https://github.com/marianatstad/dot>) to compare the two parental argan genomes (haplotig 1 and haplotig 2). This comparison revealed extensive synteny and collinearity, with the exception of an inversion on chromosome 2 (Fig. 2). Taken together, these analyses confirm the genome structure of 11 chromosomes, as independently validated by the two parental genomes. Most chromosomes contained a strong telomere signal at both ends, with few exceptions (Fig. S2).

We used Chromsyn¹⁹ to assess synteny between the two haplotigs and a previously published genome assembly of the species (QLOD00000000.2 from PRJNA294096)³. Overall, we observed strong synteny between haplotig 1 and haplotig 2, with only a few inversions near the centromeric regions of the larger chromosomes (Fig. 3). We also observed that the largest chromosomes (chromosome 1 and chromosome 2) were fragmented in the previously released Argan assembly (QLOD) (Fig. 3).

The result from the Chromsyn tool reflects an absence of synteny in the centromeric regions of chromosomes 1 and 2 between the two haplotigs. While this could be due to differences in transposable element content, a more plausible explanation is the limitation of the Chromsyn tool itself. Chromsyn relies on BUSCO orthologs to identify syntenic blocks²⁰ which may not be effective in regions with extended repetitive sequences, and with very few coding sequences. The lack of annotated genes in these regions likely prevents the detection of synteny

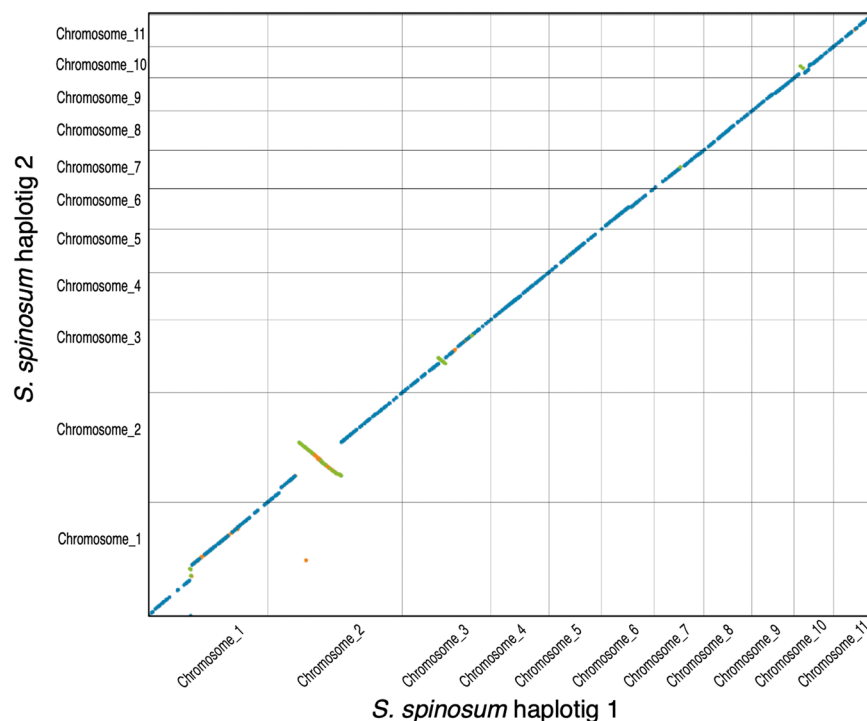


Fig. 2 Collinearity between haplotig 1 and haplotig 2 of *S. spinosum*. Collinearity is shown as a dot plot representation along the eleven chromosomes. All chromosomes were highly collinear, except for chromosome 2 that exhibited an inversion.

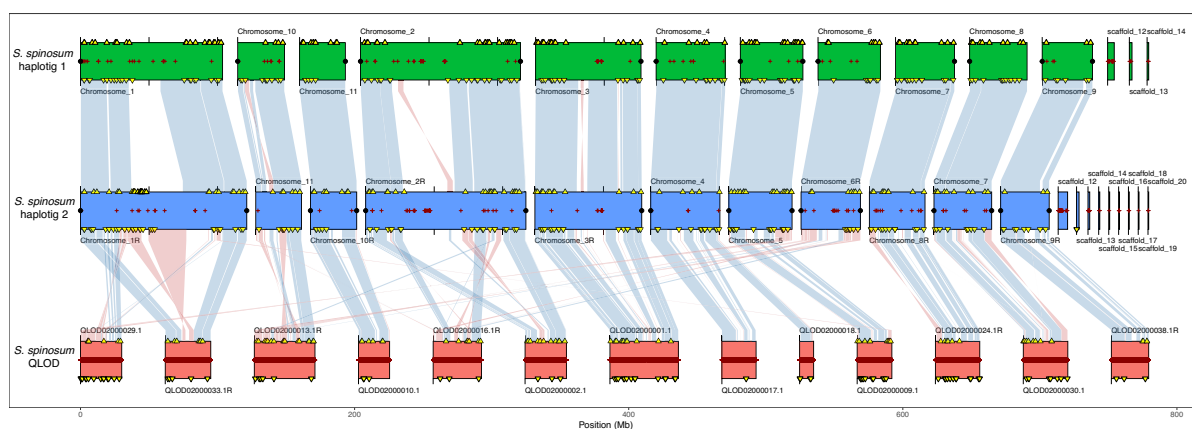


Fig. 3 Synteny between different argan genome assemblies. Haplotig 1 (green) and haplotig 2 (blue) generated in this study were tested for synteny with the previously published genome assembly QLOD2 (orange). Red crosses along chromosomes represent assembly gaps, while black points at the end of chromosomes indicate telomere sequences. Yellow triangles indicate duplicated BUSCO genes in haplotig 1 and haplotig 2. Blue links between assemblies represent collinear regions, while red links between assemblies represent genome inversions.

signals. However, the analysis of collinearity between haplotig 1 and haplotig 2 shows high homology at these regions of chromosomes 1 and 2 (Fig. 2). In general, evaluation and comparison of the genome assemblies with Quast²¹ revealed that all assembly statistics were improved in our two haplotigs relative to the QLOD assembly (Table S1), indicating that our new assembly represents a significant upgrade.

Inversion in chromosome 2. We performed a detailed analysis of the inversion on chromosome 2. We used SAMTools depth²² to identify read coverage in regions flanking the inversion breakpoints. The boundaries of the inversion were identified through pairwise alignment of haplotypes (hap1 and hap2) using D-Genies²³, with coordinates for hap1 defined as follows: start at 27.153 Mb, end at 63.831 Mb, and resumption of collinear sequence at 63.847 Mb (Fig. S3a).

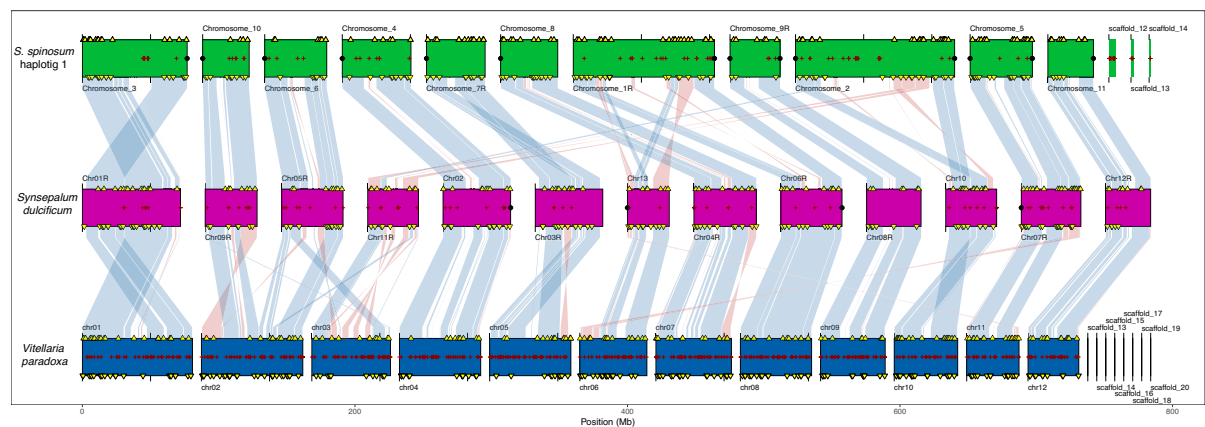


Fig. 4 Synteny between three Sapotaceae species. The chromosomal complement of *S. spinosum* (green), *S. dulcificum* (magenta), and *V. paradoxa* (blue) were aligned. For clarity, only the *V. paradoxa* contigs exceeding 1 Mbp are shown. Red crosses along the chromosomes represent assembly gaps, while black points at the ends of chromosomes represent telomere sequences. Yellow triangles indicate duplicated BUSCO genes. Blue links between assemblies represent collinear regions. Red links between assemblies represent genome inversions.

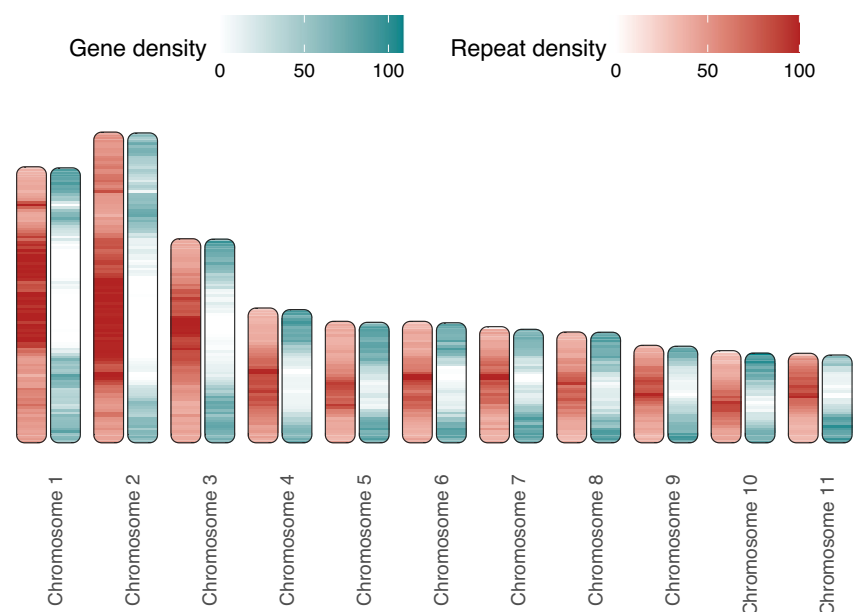


Fig. 5 Gene and repeat density on haplotig 1. Density of predicted protein-coding genes (green) and repetitive sequences (red) are indicated by colour density gradients along the chromosomes of haplotig 1 (compare with Fig. S5).

Coverage analysis revealed a depth of 0 at the start of the inversion in hap1, indicating an absence of mapped reads at this breakpoint. This lack of coverage was visually confirmed with IGV²⁴, hence, precluding to determine the exact location and nature of the beginning of the inversion (Fig. S3b). However, we observed overlapping reads at the end of the inversion and at the beginning of the downstream collinear region (Fig. S3b,c) confirming the presence of the inversion. In conclusion, while no overlapping reads are observed at the beginning of the inversion, the presence of overlapping reads at the end supports the existence of the structural variant, although its exact size remains uncertain.

The comparison with two other published genomes from the Sapotaceae family revealed similar chromosome numbers. The shea tree (*V. paradoxa*) contains 12 chromosomes²⁵ and the miracle fruit tree (*S. dulcificum*) contains 13 chromosomes²⁶. We used OMA 2.6.0²⁷ to build a phylogenetic species tree using an orthologous matrix (OMA) based on 11291 orthologous groups of coding genes. Based on this tree, the shea tree is slightly closer to argan than to the miracle fruit tree (Fig. S4). The three Sapotaceae genomes exhibited extensive synteny and collinearity along their chromosomes, despite multiple chromosomal rearrangements and inversions (Fig. 4). Interestingly, the larger chromosomes in argan (chromosome 1 and chromosome 2) appear to have

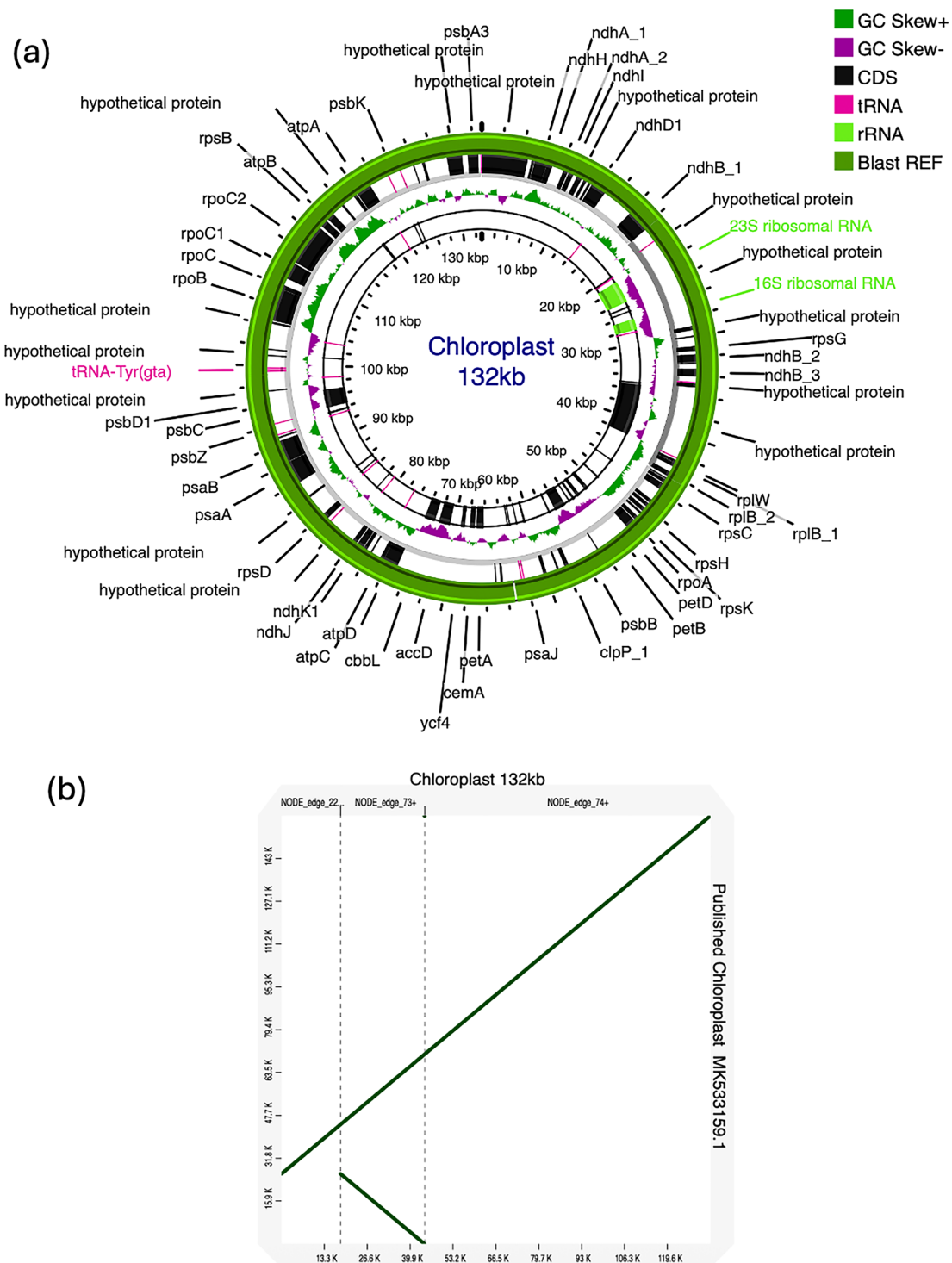


Fig. 6 Chloroplast *de novo* assembly. **(a)** Annotated circular visualization of the assembled chloroplast genome. **(b)** Dot-plot representation of the collinearity between the newly assembled chloroplast genome and a previously published version of the chloroplast genome.

emerged as a result of repeat expansions and chromosome fusions (Figs. 4, 5): Argan chromosome 1 corresponds to shea tree chromosomes 6 and 8, and argan chromosome 2 corresponds to shea tree chromosomes 3 and 10.

Repeat prediction. Repeat elements (including transposable elements) were predicted with RepeatModeler v2.0.1²⁸, followed by soft-masking the genome assemblies with RepeatMasker v4.1.2²⁹. We found that 60.86% of the hap1 and 60.78% of the hap2 genome consisted of repetitive regions (Table S2).

Genome annotation. Genome annotation was performed with BRAKER3³⁰, integrating both proteomic and RNA-seq data. Six different proteomes (*Medicago truncatula*, *V. paradoxa*, *S. dulcificum*, *Diospyros oleifera*, *Camellia sinensis* and *Rhododendron henanense*) were used as proteomic templates for gene prediction. RNA sequencing data were generated from three Argan root samples (PRJEB60382). These three RNA-seq libraries were initially subjected to quality filtering and trimming using fastp. The processed reads were subsequently mapped to the genome assembly with HISAT2³¹. This annotation pipeline resulted in the identification of 28,726 predicted genes for hap1 and 28,712 for hap2. A subsequent functional annotation was carried out using eggNOG³².

A global overview of gene and repeat density along all chromosomes shows a lack of genes and expansion of repetitive elements in the centromeric regions of the chromosomes on haplotig 1 (Fig. 5) and haplotig 2 (Fig. S5). Additionally, the two longest chromosomes in both parental genomes (hap1 and hap2) displayed large expansions of repetitive sequences, each encompassing more than 30 Mb (Fig. 5, S5).

Chloroplast genome assembly. An independent approach was used to assemble the chloroplast genome. First, the HiFi reads were mapped to the published chloroplast reference genome of *S. spinosum* (MK533159.1)⁶ with minimap2³³. The mapped reads were then recovered with SAMTools²² and the chloroplast genome was *de novo* assembled from the recovered reads using flye v.2.9.5³⁴ with parameters—meta—pacbio -g 150k. We then used Bandage v.0.8.1³⁵ to identify sequences that display circularity features. Finally, the assembly was annotated and visualized with the Proksee online tool³⁶, which included Prokka annotation³⁷. Collinearity between our assembly and published assemblies were evaluated with D-genies²³.

The chloroplast sequence consisted of 132,913 bp assembled from 3 contigs with 131 annotated features (Fig. 6a). Our new chloroplast assembly was blasted to a previously published chloroplast reference genome resulting in complete blast homology with excellent collinearity. The only difference with the published chloroplast sequence was a 26 kb duplication on the reference chloroplast assembly (Fig. 6b).

Data Records

The three PacBio HiFi runs, and the Illumina Hi-C data can be accessed at the ENA repository <https://www.ebi.ac.uk/ena/browser/view/PRJEB59883>³⁸ using the identification numbers ERR10968128, ERR10969898, ERR13030747, and ERR13030748, respectively. The genome assembly results are available in the ENA repository <https://www.ebi.ac.uk/ena/browser/view/PRJEB88017> (hap1)³⁹ and <https://www.ebi.ac.uk/ena/browser/view/PRJEB88018>⁴⁰ (hap2). The genome annotations are available on Zenodo (<https://zenodo.org/records/15913223>)⁴¹.

Technical Validation

Genome completeness was assessed using BUSCO v5.4.2⁴² with the eudicots_odb10 plant database, which contains 2,326 BUSCO entries. For hap1 and hap2 we found 97.8% and 98.5%, respectively, of the single-copy BUSCO entries (Tables S3, S4).

Code availability

The genome assembly scripts are available at https://github.com/ivandamg/Argan_Genome. All commands and pipelines used in data processing were executed according to the manual and protocols of the corresponding bioinformatic software.

Received: 9 May 2025; Accepted: 4 August 2025;

Published online: 15 August 2025

References

- 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. *Proc. Natl. Acad. Sci. USA*. **108**, 13963–13968, <https://doi.org/10.1073/pnas.1106382108> (2011).
- El Boukhari, A. *et al.* Flow cytometry and chromosome numbers variation in argan tree *Argania spinosa* (L.) Skeels. *Notulae Scientia Biologicae* **15**, 11451, <https://doi.org/10.55779/nsb15111451> (2023).
- Khay, S. *et al.* First draft genome assembly of the Argane tree (*Argania spinosa*). *F1000Research* **7**, 1310 <https://f1000research.com/articles/7-1310/v2> (2020).
- Rupp, O. *et al.* Genome of *Argania spinosa* L.: insights into oil production and the tocopherol biosynthesis pathway. *Genetic Resources and Crop Evolution* **71**, 4027–4042, <https://doi.org/10.1007/s10722-024-01931-6> (2024).
- Azami, A. I. *et al.* The complete mitochondrial genome data of *Argania spinosa* (L.) Skeels. *Data in Brief* **57**, 110862, <https://doi.org/10.1016/j.dib.2024.110862> (2024).
- Khay, S. *et al.* Complete chloroplast genome of *Argania spinosa*: structural organization and phylogenetic relationships in Sapotaceae. *Plants* **9**, 1354, <https://doi.org/10.3390/plants9101354> (2020).
- Rabeh, K. *et al.* Genome-wide identification of SSR markers from coding regions for endangered *Argania spinosa* L. skeels and construction of SSR database: AsSSRdb. *Database* **2024**, baee118, <https://doi.org/10.1093/database/baee118> (2024).
- Pakhrou, O. *et al.* Study of genetic diversity and differentiation of argan tree population (*Argania spinosa* L.) using AFLP markers. *Australian Journal of Crop Science* **10**, 990–999, <https://doi.org/10.21475/ajcs.2016.10.07.p7680> (2016).
- Chakhchar, A. *et al.* An assessment of genetic diversity and drought tolerance in argan tree (*Argania spinosa*) populations: Potential for the development of improved drought tolerance. *Frontiers in Plant Science* **8**, 1–11, <https://doi.org/10.3389/fpls.2017.00276> (2017).
- Chen, S. F. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**, e107, <https://doi.org/10.1002/imt2.107> (2023).
- Cheng, H. Y., Concepcion, G. T., Feng, X. W., Zhang, H. W. & 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).

12. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**, 1432, <https://doi.org/10.1038/s41467-020-14998-3> (2020).
13. Wingett, S. *et al.* HiCUP: Pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310, <https://doi.org/10.12688/f1000research.7334.1> (2015).
14. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3**, 95–98, <https://doi.org/10.1016/j.cels.2016.07.002> (2016).
15. Zhou, C. X., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808, <https://doi.org/10.1093/bioinformatics/btac808> (2023).
16. 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.2015.07.012> (2016).
17. Brown, M. R., de la Rosa, P. M. G. & Blaxter, M. tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. *Bioinformatics* **41**, btaf049, <https://doi.org/10.1093/bioinformatics/btaf049> (2025).
18. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *Plos Computational Biology* **14**, e1005944, <https://doi.org/10.1371/journal.pcbi.1005944> (2018).
19. Edwards, R. J., Dong, C., Park, R. F. & Tobias, P. A. A phased chromosome-level genome and full mitochondrial sequence for the dikaryotic myrtle rust pathogen, *Austropuccinia psidii*. *BioRxiv*, <https://doi.org/10.1101/2022.04.22.489119> (2022).
20. Simao, 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**, 3210–3212, <https://doi.org/10.1093/bioinformatics/btv351> (2015).
21. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075, <https://doi.org/10.1093/bioinformatics/btt086> (2013).
22. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, <https://doi.org/10.1093/bioinformatics/btp352> (2009).
23. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, 4958, <https://doi.org/10.7717/peerj.4958> (2018).
24. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**, 178–192, <https://doi.org/10.1093/bib/bbs017> (2013).
25. Hale, I. *et al.* Genomic resources to guide improvement of the shea tree. *Frontiers in Plant Science* **12**, 720670, <https://doi.org/10.3389/fpls.2021.720670> (2021).
26. Yang, Z. *et al.* The chromosome-level genome of miracle fruit (*Synsepalum dulcificum*) provides new insights into the evolution and function of miraculin. *Frontiers in Plant Science* **12**, 804662, <https://doi.org/10.3389/fpls.2021.804662> (2022).
27. Altenhoff, A. M. *et al.* OMA standalone: Orthology inference among public and custom genomes and transcriptomes. *Genome Research* **29**, 1152–1163, <https://doi.org/10.1101/gr.243212.118> (2019).
28. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457, <https://doi.org/10.1073/pnas.1921046117> (2020).
29. Smith, A., Hubley, R. & Green, P. <https://www.repeatmasker.org> (2010).
30. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Research* **34**, 769–777, <https://doi.org/10.1101/gr.278090.123> (2024).
31. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, 907–915, <https://doi.org/10.1038/s41587-019-0201-4> (2019).
32. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314, <https://doi.org/10.1093/nar/gky1085> (2019).
33. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, <https://doi.org/10.1093/bioinformatics/bty191> (2018).
34. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**, 540–546, <https://doi.org/10.1038/s41587-019-0072-8> (2019).
35. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352, <https://doi.org/10.1093/bioinformatics/btv383> (2015).
36. Grant, J. R. *et al.* Proksee: in-depth characterization and visualization of bacterial genomes. *Nucleic Acids Res.* **51**, W484–W492, <https://doi.org/10.1093/nar/gkad326> (2023).
37. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069, <https://doi.org/10.1093/bioinformatics/btu153> (2014).
38. ENA European Nucleotide Archive <https://identifiers.org/ena.embl:PRJEB59883> (2025).
39. ENA European Nucleotide Archive <https://identifiers.org/ena.embl:PRJEB88017> (2025).
40. ENA European Nucleotide Archive <https://identifiers.org/ena.embl:PRJEB88018> (2025).
41. Reinhardt, D. *et al.* Genome Annotation of Haplotig 1 and Haplotig 2 of *Sideroxylon spinosum* L. *Zenodo*. <https://doi.org/10.1101/2025.04.06.647421> (2025).
42. 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).

Acknowledgements

We thank the Next Generation Sequencing Platform of the University of Bern for performing the high-throughput sequencing experiments for this challenging project. The work was supported by the Swiss National Science Foundation (Grant No. 310030_200367) and the Pool de Recherche of the University of Fribourg (spring 2020).

Author contributions

L.F., I.D.M. and D.R. contributed to the research design. A.E. and M.H. collected the samples. I.D.M., L.F. and P.N. analyzed the data. L.F., D.R., I.D.M. and A.Q. wrote the draft manuscript and revised the manuscript. All co-authors contributed to this manuscript and approved it.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-025-05768-1>.

Correspondence and requests for materials should be addressed to L.F. or D.R.

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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025