scientific data

Check for updates

OPEN

DATA DESCRIPTOR

# Comprehensive re-assembly and annotation dataset for the argan tree (*Argania spinosa L.*, Sapotaceae) genome

Abdellah Idrissi Azami[1], Stacy Pirro[2], Nihal Habib[1], Turgay Unver[3,4], M. Gonzalo Claros[5,6], Juan de Dios Alché[7,8], Sofia Sehli[1], Zainab EL Ouafi[1], Douae EL Ghoubali[1], Dalila Bousta[9], Najib Al Idrissi[1], Fatima Gaboun[10], Abderrazak Rfaki[11,12], Abdelkhalek Legsyer[13], Chakib Nejjari[14,15], Saaid Amzazi[16,17], Lahcen Belyamani[18], Abdelhamid El Mousadik[19] & Hassan Ghazal[1,2,11,12] ✉

**We present a comprehensive annotation dataset for the scaffold-level nuclear genome assembly of the argan tree (*Argania spinosa* (L.) Skeels, Sapotaceae). Using Illumina whole-genome shotgun reads that we previously generated for the "Argan Amghar" individual and deposited under BioProject PRJNA294096, together with the corresponding GenBank assembly GCA_003260245.2, we re-assembled and curated a 690 Mbp draft with scaffold N50 of 25 Mbp and L50 of 11 large macro-scaffolds. Ab initio gene prediction with AUGUSTUS and GeneMark-ES, integrated by EVidenceModeler, produced 51,078 protein-coding genes and 2,081 non-coding RNA genes, while repeat annotation covers 53.0% of the assembly. Functional annotation combined eggNOG-mapper, InterProScan and BLASTp searches against UniProtKB/Swiss-Prot to assign curated functions, domains and Gene Ontology terms to 32,785 genes and to support 25,484 proteins with UniProt evidence. BUSCO analyses indicate high completeness of the assembly gene space and completeness of the predicted proteome (74.6%). All primary data products, including a unified GFF3 file and the predicted proteome FASTA, are openly available via NCBI and Zenodo (https://doi.org/10.5281/zenodo.17901083).**

[1]Laboratory of Precision Medicine and One Health (MedPreOne), School of Medicine, Mohammed VI University of Sciences and Health (UM6SS), Casablanca, Morocco. [2]Iridian Genome, Silver Spring, MD, USA. [3]Ficus Biotechnology, Ankara, Turkey. [4]Faculty of Engineering, Ostim Technical University, Ankara, Turkey. [5]Institute for Mediterranean and Subtropical Horticulture "La Mayora", Universidad de Málaga-Consejo Superior de Investigaciones Científicas (IHSM-UMA-CSIC), Málaga, Spain. [6]Department of Molecular Biology and Biochemistry, University of Malaga, Malaga, Spain. [7]Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas (CSIC), Granada, Spain. [8]University Institute of Research on Olive Groves and Olive Oils (INUO), Universidad de Jaén, Jaen, Spain. [9]National Agency of Medicinal and Aromatic Plants, 34025, Taounate, Morocco. [10]National Institute for Agricultural Research (INRA), Rabat, Morocco. [11]National Center for Scientific and Technical Research, Rabat, Morocco. [12]Royal Institute of Executive Management (IRFC), Salé, Morocco. [13]Laboratory of Physiology, Genetics, and Ethnopharmacology (LPGE), Faculty of Sciences, University Mohamed Premier, Oujda, Morocco. [14]Euromed Research Center, Euromed University of Fes, Fes, Morocco. [15]Faculty of Medicine, Pharmacy, and Dentistry, Sidi Mohamed Ben Abdellah University, Fes, Morocco. [16]Laboratory of Human Pathologies Biology, Department of Biology, Faculty of Sciences, Mohammed V University, Rabat, Morocco. [17]Genomic Center of Human Pathologies, Faculty of Medicine and Pharmacy, Mohammed V University, Rabat, Morocco. [18]Mohammed VI University of Sciences and Health, Casablanca, Morocco. [19]Laboratory of Biotechnology and Valorization of Natural Resources (LBVRN), Faculty of Science, University Ibn Zohr, Agadir, Morocco. ✉e-mail: hassan.ghazal@fulbrightmail.org

## Background & Summary

*Argania spinosa* (L.) Skeels, commonly known as the argan tree, is a species of significant ecological and economic importance endemic to southwestern Morocco. This drought-tolerant tree occupies approximately 900,000 hectares and has been recognized by UNESCO as a biosphere reserve since 1998[1]. The species is particularly valued for its oil-rich seeds, which produce argan oil used in culinary, cosmetic, and medicinal applications due to its high content of unsaturated fatty acids and antioxidants[1].

Despite the economic importance of *A. spinosa*, comprehensive genomic resources have been limited. Previous genomic studies have focused primarily on organellar genomes, with the chloroplast genome (158,848 bp) and mitochondrial genome (707,441 bp) being sequenced and characterized[2–4]. Initial nuclear genome assemblies have provided basic structural information[5], with a recent annotation published by Rupp *et al.*[6] describing 62,590 predicted genes; however, that work, and earlier draft assemblies lacked comprehensive characterization of non-coding RNAs, detailed repetitive element landscape, and systematic pathway-level functional annotation.

In this Data Descriptor we provide an updated scaffold-level assembly and a comprehensive structural and functional annotation for the "Argan Amghar" nuclear genome. The same individual was originally sequenced by our group using Illumina HiSeq X Ten technology, and the raw reads were deposited in the NCBI Sequence Read Archive under BioProject PRJNA294096[5]. In the present work we re-use these reads to generate a curated scaffold-level reference, and an integrated annotation tailored to downstream comparative and functional genomics.

Our assembly spans 690 Mbp in 186,325 scaffolds with an N50 of 25 Mbp and an L50 of 11 large macro-scaffolds. Genome size estimates from k-mer analysis (698 Mbp), together with previous flow-cytometry estimates, are consistent with this assembly size.

The final gene set contains 51,078 protein-coding genes, representing a balanced gene repertoire that lies between earlier, more expansive annotations (~62,590 genes) and the more conservative predictions for closely related Sapotaceae and Ericales genomes. We additionally annotate 2,081 non-coding RNA genes and extensive repetitive elements covering 53.0% of the assembly. Functional annotation combines orthology-based predictions from eggNOG-mapper v2[7], protein-domain and motif assignments from InterProScan[8,9], and BLASTp[10] searches against UniProtKB/Swiss-Prot[11], together with pathway mapping using the Plant Metabolic Network (PMN)[12]. These yields curated functional descriptions and domain architectures for 32,785 genes and UniProt-validated annotations for 25,484 proteins. Of these, 1,906 are classified as putative transcription factors using the Transcription Factor Prediction pipeline from PlantTFDB v5.0[13], and 7,044 are linked to metabolic pathways, including those involved in lipid and tocopherol metabolism.

Our dataset provides a deeply curated and well-documented annotation of the widely used scaffold-level assembly GCA_003260245.2. Researchers who have already adopted this assembly (Such as Rupp *et al.*[6]) can directly reuse our GFF3, and proteome files without re-mapping to a different reference.

The primary applications for this dataset include Comparative genomics studies within Ericales and Sapotaceae, identification of genes underlying drought tolerance and oil biosynthesis, development of molecular markers for breeding programs, conservation genetics applications, and biotechnological research aimed at understanding and optimizing argan oil production. The dataset is particularly valuable for researchers studying lipid metabolism, tocopherol biosynthesis, stress tolerance mechanisms, and evolutionary genomics of economically important trees.

## Methods

**Plant material and sequencing data source.** The nuclear genomic data analysed in this study originate from the "Argan Amghar" tree (Fig. 1), an adult *Argania spinosa* (L.) Skeels individual sampled in the Souss Valley (southwestern Morocco) and described in our previous work. The voucher specimen (AM-2015) was deposited at the Faculty of Sciences, Agadir, Morocco, under the curation of Professor A. El Mousadik.

Genomic DNA from young leaves of this tree was extracted and sequenced using the Illumina HiSeq X Ten platform (2 × 150 bp paired end). The resulting whole-genome shotgun reads were deposited in the NCBI Sequence Read Archive as part of BioProject PRJNA294096[14] (BioSample SAMN04014715[15]; SRA study SRP077839[16], including runs SRR6062045 and SRR6062046)[5]. In the present Data Descriptor, we re-use and re-analyse these reads, without generating new sequencing data, to produce an updated scaffold-level assembly and a comprehensive annotation dataset.

**Genome size estimation and assembly processing.** Genome size, heterozygosity and repeat content were estimated from the Illumina reads using GenomeScope v2.0[17] with a k-mer size of 27 dataset generated by Meryl[18]. The model indicated a haploid genome size of approximately 698 Mbp, a heterozygosity rate of 1.71%, and 50.1% unique sequence content, in agreement with previous flow-cytometry and k-mer-based estimates for *Argania spinosa*[19].

Raw paired-end reads were quality-filtered using Trimmomatic v0.33[20] with standard parameters for adapter removal and low-quality base trimming. Reads shorter than the default minimum length after trimming were discarded. Assembly was performed using SPAdes v2.5[21] with default parameters, followed by decontamination using NCBI-FCS and NCBI-GX[22] to detect and remove potential contaminant sequences from non-plant taxa. The resulting draft assembly was further processed using Zanfona v1.0[23], which performs scaffolding based on conserved synteny with related plant genomes to improve contiguity while preserving the underlying read-based assembly information.

Assembly quality assessment was conducted using QUAST v5.0[24] to calculate standard metrics, including total assembly size, scaffold and contig N50 values, largest scaffold length, number of scaffolds, GC content and gap statistics. Gene-space completeness of the genome assembly was further evaluated using BUSCO v5 with the

**Fig. 1** Photo of the shrub Argan Amghar (taken by A. El Mousadik).

eudicots_odb10 lineage set in genome mode. These metrics were used to verify that the assembly was consistent with the K-mer based genome size estimation and with previously reported values for *A. spinosa*[5].

**Repeat element annotation.** Repetitive element annotation employed a multi-step approach combining de novo repeat discovery with homology-based classification. A species-specific repeat library was constructed using RepeatModeler[25], which integrates multiple ab initio repeat detection algorithms, including RECON, RepeatScout and LTR_Finder, to identify repetitive sequences specific to *A. spinosa*.

The custom repeat library was used with RepeatMasker v4.1.5[26] on the Galaxy Europe platform to perform soft masking of the genome assembly. RepeatMasker was configured with the Viridiplantae taxonomic clade and DFam as the reference repeat database, so that both lineage-specific and conserved plant repeats were detected. Tandem repeats and simple sequence repeats were identified using Tandem Repeats Finder (TRF)[27] with default parameters, providing additional coverage of satellite DNA and low-complexity regions. The combination of de novo and library-based approaches allowed us to capture both known and novel repeat families before gene prediction.

**Gene prediction and structural annotation.** Structural annotation employed a multi-pronged ab initio gene prediction strategy to maximise accuracy while remaining transparent about the assumptions involved. AUGUSTUS v3.4.0[28] was trained using gene models from three phylogenetically and functionally relevant species:

- *Camellia sinensis* (RefSeq: GCF_004153795.1)[29] – selected for phylogenetic proximity within Ericales, providing evolutionary context for gene structure prediction.
- *Arabidopsis thaliana* (RefSeq: GCF_000001735.4)[30] – included as a model organism with extensively validated gene annotations for core plant gene families.
- *Olea europaea* (RefSeq: GCF_002742605.1)[31] – chosen for functional similarity as an oil-producing species with well-characterised lipid biosynthesis pathways.

For each of these AUGUSTUS species models, gene prediction was carried out on the repeat-masked assembly. In parallel, self-training gene prediction was performed using GeneMark-ES v4[32], which automatically estimates species-specific parameters through unsupervised learning from the genome sequence without requiring pre-existing gene annotations.

Gene sets from individual predictors were integrated using EVidenceModeler (EVM) v2.1.0[33] with differential weighting: the highest weights were assigned to GeneMark-ES and the AUGUSTUS model trained on *C. sinensis*, intermediate weights to the *O. europaea* model, and the lowest weights to the *A. thaliana* model. This weighting scheme reflects the expected phylogenetic and functional relevance of each model to *A. spinosa*. The integrated gene set was processed using AGAT[34] to remove fully overlapping or conflicting predictions, normalise GFF3 formatting and assign consistent gene and transcript identifiers.

Gene set completeness was evaluated using BUSCO v5[35] with the eudicots_odb10 dataset. BUSCO scores were used as an independent measure of how well the predicted gene set recovers universally conserved eudicot orthologues and are reported alongside assembly-level BUSCO scores in the Technical Validation section.

**Non-coding RNA annotation.** Non-coding RNA (ncRNA) elements were identified using a multi-tiered computational approach that combines structural prediction and homology-based methods. Transfer RNA genes were predicted using tRNAscan-SE v2.0[36] with eukaryotic model parameters optimised for plant genomes, yielding a set of high-confidence tRNA loci. Ribosomal RNA genes were identified using Barrnap v0.9[37] with default eukaryotic settings.

Additional ncRNA classes were annotated using Infernal v1.1.4[38] with covariance models from Rfam v15[39]. We applied an E-value cutoff of $\leq 1 \times 10-5$ and a minimum model coverage of $\geq$70% to retain only confident matches. To complement Rfam-based predictions, homology searches were conducted using BLASTn[10] against a curated reference set of *Camellia sinensis* ncRNAs[29], with stringent criteria requiring $\geq$85% identity over $\geq$80% of the sequence length. This cross-species comparison leveraged a phylogenetically related genome to improve detection of conserved ncRNAs in *A. spinosa*.

Final ncRNA annotations were manually curated to reduce redundancy: for overlapping predictions covering more than 50% of the same genomic interval, only the highest-scoring annotation was retained. This procedure yielded a non-redundant set of tRNAs, rRNAs and other ncRNA classes that is consistent with expectations for a woody eudicot genome.

**Functional annotation.** Protein-coding sequences were translated to produce a predicted proteome of 51,078 proteins. Functional annotation followed a three-step strategy designed to combine orthology, domain signatures and sequence similarity.

First, we used eggNOG-mapper v2[7] with the eukaryotic database to assign orthologous groups, preliminary Gene Ontology (GO) terms based on orthology. Second, we ran InterProScan v5.59[9] on the predicted proteome to identify protein domains, signatures and motifs. InterProScan-derived signatures were used to refine GO term assignments, to detect conserved catalytic and binding domains, and to flag proteins with signal peptides or transmembrane segments where relevant. Third, we performed BLASTp[10] searches against UniProtKB/Swiss-Prot[11]. Hits with E-value $\leq 1 \times 10-5$ and alignment coverage $\geq$50% of the query length were considered high-confidence and were used both to refine functional descriptions and to flag UniProt-validated genes.

Using this combined pipeline, 32,785 genes (64.2% of the gene set) were assigned at least one functional description, GO term or domain annotation, and 25,484 proteins (49.9%) have at least one high-confidence UniProtKB/Swiss-Prot match. GO terms derived from both eggNOG-mapper and InterProScan were used to summarise functional categories and pathway mapping.

To explore metabolic pathways, we mapped enzyme-coding genes to the Plant Metabolic Network (PMN)[12] using a BLASTp search against the PMN TeaCyc dataset. This yielded 7,044 genes associated with at least one plant metabolic pathway, including key pathways for triacylglycerol and tocopherol biosynthesis.

Transcription factors (TFs) were identified using the Transcription Factor Predictor from PlantTFDB v5.0 (Plant Transcription Factor Database)[13]. Amino-acid sequences of all 51,078 predicted proteins were submitted to the PlantTFDB v5.0 TF prediction workflow using default parameters. The pipeline classifies proteins into TF families based on curated DNA-binding and regulatory-domain profiles, integrating information from multiple motif and domain databases.

Proteins predicted as TFs by PlantTFDB v5.0 were retained and assigned to their corresponding TF families. This analysis identified 1,906 putative transcription factors spanning major plant TF families.

## Data Records

The complete *Argania spinosa* genome annotation dataset is deposited in multiple complementary repositories to ensure accessibility and long-term preservation[40]:

**Primary genome assembly.** The scaffold-level nuclear genome assembly (690 Mbp; scaffold N50: 25 Mbp; 186,325 scaffolds; largest scaffold ≈ 50 Mbp) used in this study is available in the NCBI Assembly/GenBank database under accession GCA_003260245.2 (assembly ASM326024v2)[41].

A soft-masked version of this assembly, with repeats annotated by RepeatMasker set to lower case, is provided as a FASTA file (Argania_spinosa_assembled_genome.fasta) in the Zenodo repository https://doi.org/10.5281/zenodo.17901083. This file serves as the reference for all structural and functional annotations described in the manuscript.

**Annotation files.** Comprehensive annotation data including protein-coding gene models, repeat annotation in GFF format, and non-coding RNA predictions are deposited in the Zenodo repository (https://doi.org/10.5281/zenodo.15717879). This repository contains:

| Parameter | Contigs | Scaffolds |
|---|---|---|
| Assembled size (Mbp) | 613 | 690 |
| Number of sequences | 259,005 | 186,325 |
| Largest sequence (Mbp) | 0,197 | 50 |
| N50 (Mbp) | 0,005 | 25 |
| L50 | 25,528 | 11 |
| N90 (bp) | 807 | 960 |
| L90 | 161,294 | 83,892 |
| N's per 100 kbp | 462.55 | 11,832 |

**Table 1.** Statistics of the Argania spinosa draft genome assembly, detailing contig and scaffold metrics, as assessed by QUAST v5.0.

- A unified GFF3 file (Argania_spinosa.gff) containing all predicted protein-coding genes, transcripts, exons, CDS features, and non-coding RNAs mapped to the GCA_003260245.2 assembly.
- A tab-delimited/CSV gene annotation table (Argania_spinosa_PCG_annotation.csv) summarising genomic coordinates, domain content and functional attributes for all 51,078 protein-coding genes. Columns include gene identifier, scaffold, start and end positions, strand, product description, and the GO terms.
- The corresponding predicted proteome as an amino-acid FASTA file (Argania_spinosa.proteom.fasta).
- An ncRNA annotation table (Argania_spinosa_RNA_annotation.csv) listing non-coding RNAs with RNA identifier, scaffold, genomic coordinates, strand, RNA type (tRNA, rRNA, miRNA, snoRNA, snRNA, ncRNA) and predicted product. These features are also present in the unified GFF3 file.
- Functional annotation outputs, including UniProt BLASTp results (uniprot_validation.tsv), InterProScan domain/GO assignments, eggNOG-mapper orthology/GO/KEGG annotations and Plant Metabolic Network (PMN) pathway mappings.

**File organization.** Annotation and validation files are organised by data type to facilitate reuse and integration with standard analysis pipelines:

- Genome assembly: repeat-masked genome in FASTA format (GCA_003260245.2, soft-masked version).
- Structural annotation: a unified GFF3 file containing all protein-coding genes, transcripts, exons, CDS features, non-coding RNA, alongside FASTA file of predicted protein sequences.
- Tabular summaries: CSV/TSV tables for protein-coding genes and ncRNAs, each summarising coordinates and key functional attributes.
- Repeat annotation: GFF3 files describing interspersed repeats and tandem repeats, compatible with genome browsers.
- Functional annotation: tabular outputs for UniProt BLASTp hits, InterProScan domain and GO assignments, eggNOG-mapper orthology/GO/KEGG results and Plant Metabolic Network (PMN) pathway mappings.
- Quality assessment: BUSCO result files for the assembly and the predicted proteome.

All files are referenced to the same coordinate system, namely the NCBI assembly accession GCA_003260245.2, and are documented in a README in the Zenodo repository.

**Data formats.** The main data formats provided are:

- FASTA for the repeat-masked genome assembly and the predicted proteome.
- GFF3 for the unified structural annotation element annotations.
- CSV/TSV tables for protein-coding gene attributes, ncRNA features, functional annotations and BUSCO summaries.

These formats are widely supported by current genome browsers and analysis platforms (e.g. JBrowse, Galaxy, Ensembl-compatible tools), ensuring interoperability with standard workflows.

## Technical Validation

**Assembly quality assessment.** Assembly quality was evaluated using multiple complementary approaches. Standard assembly metrics computed with QUAST v5.0 show that the *A. spinosa* nuclear genome assembly (GCA_003260245.2) is a highly contiguous scaffold-level draft spanning approximately 690 Mbp. The scaffold N50 is 25 Mbp, and 11 large macro-scaffolds (L50 = 11) together account for half of the assembled genome length. We emphasise that these macro-scaffolds are treated as long scaffolds rather than fully resolved chromosomes, as the available short-read data do not support chromosome-level reconstruction (Table 1, Fig. 2).

Genome size validation was performed using k-mer based estimation with GenomeScope v2.0, which predicted a haploid genome size of ~698 Mbp and a heterozygosity rate of 1.71%. These values are consistent with independent flow-cytometry estimates (1 C ≈ 724 Mbp) and with previous k-mer based estimates *for A. spinosa*, supporting the conclusion that the assembly captures the bulk of the nuclear genome (Table 2).

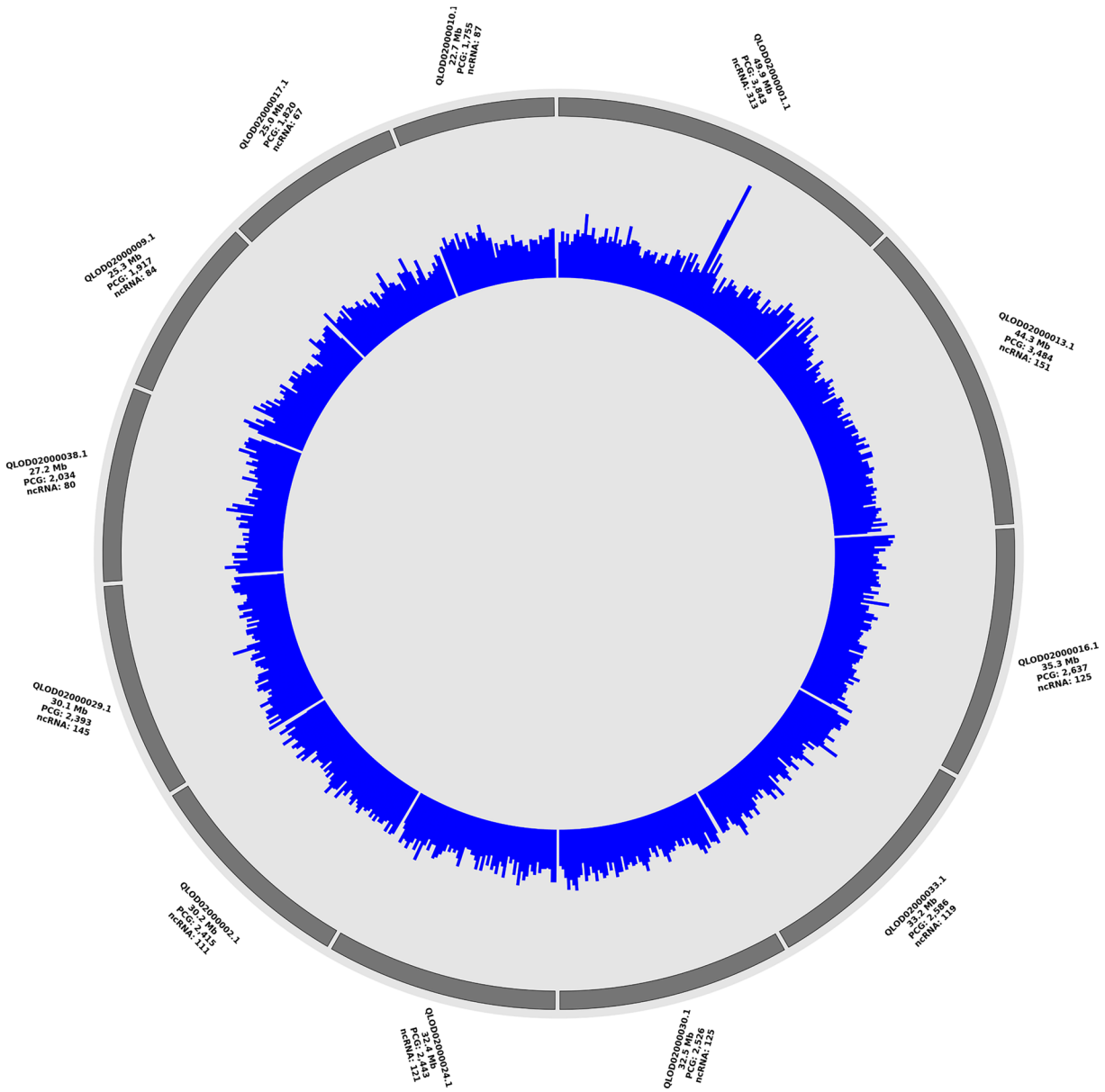


**Fig. 2** Circular feature-density map of the Argania spinosa nuclear genome. The outer grey track represents the 11 largest scaffolds of the GCA_003260245.2 assembly, labelled with scaffold ID, length (Mb) and the number of annotated protein-coding genes (PCGs) and non-coding RNAs (ncRNAs). The inner blue track shows the variation in total feature density (PCGs + ncRNAs) along each scaffold, calculated in non-overlapping genomic windows. Regions with taller blue bars correspond to higher local concentrations of annotated features, whereas lower peaks indicate gene-poor, repeat-rich segments.

| Parameter | Genome size | Proportion of Unique sequences | Percentage of heterozygosity | Duplication rate | Error rate |
|---|---|---|---|---|---|
| GenomeScope estimation | 698 Mbp | 50.1% | 1.71% | 0.48% | 0.2% |

**Table 2.** Summary of the genome characteristics of Argania spinosa derived from 27-mer distribution analysis via GenomeScope v2.0.

Assembly completeness at the gene-space level was further assessed with BUSCO v5 using the eudicots_odb10 dataset. BUSCO analysis of the assembly indicates high completeness (approximately 89.6% complete BUSCOs, with the majority in single copy), comparable to other scaffold-level woody eudicot genomes. These results suggest that most conserved eudicot gene families are represented in the assembly, even though some repeat-rich regions are likely unresolved (Table 3).

| Dataset | BUSCO mode | Complete (C) | Complete single-copy (S) | Complete duplicated (D) | Fragmented (F) | Missing (M) |
|---|---|---|---|---|---|---|
| Genome assembly (GCA_003260245.2) | genome | 2,086 (89.7%) | 2,086 (89.7%) | 2,086 (89.7%) | 2,086 (89.7%) | 2,086 (89.7%) |
| Predicted proteome (51,078 proteins) | protein | 1,736 (74.6%) | 1,736 (74.6%) | 1,736 (74.6%) | 1,736 (74.6%) | 1,736 (74.6%) |

**Table 3.** BUSCO completeness statistics (percentages) for the Argania spinosa genome assembly and predicted proteome (BUSCO v5.7.1; eudicots_odb10 lineage, n = 2,326 BUSCOs).

**Gene prediction validation.** Gene prediction accuracy was evaluated by comparing ab initio predictions generated using different AUGUSTUS training models and GeneMark-ES. Models trained on phylogenetically closer species (*Camellia sinensis*) produced more structurally realistic gene sets than those based on more distant species (*Arabidopsis thaliana*), with higher average exon counts and fewer single-exon artefacts. For example, the *C. sinensis* model yielded gene structures with an average of ~4.8 exons per gene and a reduced proportion of single-exon predictions compared with the *A. thaliana* model (Table S1).

Gene models from AUGUSTUS and GeneMark-ES were integrated using EVidenceModeler (EVM) with a weighting scheme that favoured predictions from the *C. sinensis* model and GeneMark-ES, reflecting their closer phylogenetic and functional relevance to *A. spinosa*. The integrated set was then curated with AGAT to remove overlapping or low-confidence models and to normalise identifiers and GFF3 structure. This process resulted in a final non-redundant set of 51,078 protein-coding genes with realistic exon and intron length distributions.

BUSCO was also run in protein mode on the predicted proteome using the eudicots_odb10 dataset. The proteome achieved 74.6% complete BUSCOs (69.8% single-copy, 4.8% duplicated), 12.0% fragmented and 13.4% missing BUSCOs. The lower completeness of the proteome relative to the assembly-level BUSCO score and to the recently published chromosome-level phased genome reflects the limitations of short-read assemblies and ab initio gene prediction without species-specific RNA-seq support. We highlight this as an inherent limitation of the current dataset and a priority for future refinement (Table 3).

Comparative validation against related Ericales genomes showed that gene density, gene length and exon number distributions for *A. spinosa* fall between those of *Synsepalum dulcificum* and previously reported *A. spinosa* annotations[6]. The final gene count (51,078) lies between earlier estimates (~62,590 genes) and more conservative annotations for other Sapotaceae, suggesting an appropriate balance between sensitivity (capturing genuine genes) and specificity (reducing spurious predictions) (Table S2).

**Repeat annotation validation.** Repetitive element annotation was validated through internal consistency checks and comparison with related plant genomes. The total repeat content of 53.04% falls within the expected range for outcrossing woody perennials and closely matches previous estimates for *A. spinosa*. The relative abundance of major repeat classes, including the ratio of Gypsy to Copia LTR retrotransposons, is consistent with patterns reported in other drought-tolerant Ericales species (Table S3).

A substantial fraction of repeats (37.90%) remains unclassified, reflecting the limited coverage of current repeat databases for non-model species and pointing to the presence of lineage-specific repetitive elements. This pattern is common in under-studied plant taxa and underscores the value of the species-specific repeat library generated here for future comparative repeatomics.

**Non-coding RNA validation.** Non-coding RNA predictions were evaluated by comparing overall counts and composition with those of related plant genomes and by checking for internal consistency. The total number of tRNA genes (706, of which 604 are classified as functional) is similar to that reported for *Camellia sinensis* and Olea europaea, and the distribution of anticodons follows the expected enrichment for amino acids such as leucine, serine and glycine in plant proteomes (Table S4).

The broader ncRNA catalog, which comprises 2,081 loci (including tRNAs, rRNAs, miRNAs, snoRNAs, snRNAs and other ncRNAs), falls within the range observed in other eudicot genomes when similar computational pipelines are applied (Table S5). As with many purely *in silico* ncRNA annotations, we expect that some predicted miRNAs and other regulatory RNAs will require expression and functional data for confirmation. We therefore present this ncRNA catalog as a high confidence but provisional resource rather than a definitive inventory of all regulatory RNAs in *A. spinosa*.

**Functional annotation validation.** Functional annotation quality was assessed through cross-validation between orthology-based, domain-based and similarity-based methods, and by examining coverage of key metabolic pathways. The combination of eggNOG-mapper, InterProScan and BLASTp against UniProtKB/Swiss-Prot provides complementary evidence for gene function. Nearly half of the predicted proteins (25,484; 49.9%) have strong UniProtKB/Swiss-Prot support, and 32,785 genes (64.2%) are assigned at least one GO term or domain annotation, indicating broad functional coverage of the gene space (Table S6).

GO term distributions for UniProt- and InterPro-supported genes are dominated by catalytic activity and binding in the molecular-function category, and by metabolic and cellular processes in the biological-process category, as expected for a plant nuclear genome. Mapping of enzyme-coding genes to the Plant Metabolic Network (PMN) links 7,044 genes to plant metabolic pathways, including those involved in lipid metabolism and tocopherol (vitamin E) biosynthesis. The consistent identification of enzymes in these economically important pathways across eggNOG, InterProScan and PMN provides additional confidence in the functional annotation of genes associated with oil quality and antioxidant content (Table S7).

## Usage Notes

**Software requirements.** The data are distributed in standard community formats (FASTA, GFF3, tab-delimited tables) and can be processed with widely used bioinformatics tools. Genome browsers and annotation utilities such as IGV, JBrowse and AGAT can be used to inspect and manipulate the GFF3 files. Sequence similarity searches and orthology analyses can be extended using BLAST+, eggNOG-mapper and InterProScan, and pathway-level exploration can be carried out with Plant Metabolic Network (PMN) tools or KEGG-based workflows. No proprietary software is required to reuse the dataset.

**Data integration considerations.** Gene identifiers are specific to this annotation release and may not map directly to loci from earlier A. spinosa draft annotations. When integrating this dataset with other genomic resources, users should rely on genomic coordinates on GCA_003260245.2 or on explicit sequence similarity (e.g. BLAST, orthology inference) rather than simple ID matching. All functional annotations (eggNOG, InterProScan, PMN, UniProt) are reported for the predicted proteome derived from this assembly and should be interpreted in that context.

The functional annotations represent computational predictions based on sequence homology, protein domain signatures and pathway databases. They provide useful hypotheses for gene function and metabolism, but experimental validation (e.g. expression profiling, biochemical assays or reverse genetics) is recommended before drawing strong functional conclusions, particularly for genes with no UniProtKB/Swiss-Prot support.

**Limitations and considerations.** The nuclear genome assembly remains at scaffold level, because it is based solely on Illumina short-read data without long-read or chromatin conformation (Hi-C) information. Although the assembly is highly contiguous and 11 large macro-scaffolds account for half of the assembled genome, these scaffolds should not be interpreted as fully resolved chromosomes, and some structural rearrangements or gaps in repeat-rich regions are likely to remain.

The structural annotation is based on ab initio and homology-guided gene prediction in the absence of *A. spinosa* RNA-seq or proteomics data, which means that some gene models may be incomplete or spurious and that alternative splice isoforms are likely under-represented. The extensive non-coding RNA catalog is derived from covariance-model and homology searches and should be considered putative until supported by expression or functional data.

The repeat annotation includes a substantial fraction of unclassified repetitive elements, reflecting the limited representation of Sapotaceae-specific repeats in current databases. Users focusing on repetitive regions may wish to perform additional species-specific repeat discovery and classification if finer resolution of repeat families is required.

Functional annotations and pathway assignments are inferred from sequence features and database matches and should therefore be treated as provisional. This is especially important for genes assigned to economically important traits such as argan oil composition, lipid metabolism or vitamin E biosynthesis, where follow-up validation is strongly encouraged.

**Recommended applications.** This dataset is well suited for comparative genomics within Sapotaceae and Ericales, where the annotated gene set and repeat landscape can be used to investigate genome evolution, gene family expansions and lineage-specific innovations. The integration of UniProt, InterProScan, eggNOG and PMN annotations makes the resource particularly useful for prioritising candidate genes involved in oil biosynthesis, vitamin E (tocopherol/tocotrienol) pathways and drought-related stress responses.

For breeding and biotechnology, the genome and annotation provide a foundation for marker development, candidate-gene discovery and pathway engineering aimed at improving argan oil yield and quality. In conservation genetics, the comprehensive gene catalog can support the design of population genetic markers and analyses of genetic diversity and structure across *A. spinosa* populations.

## Data availability

The *Argania spinosa* genome project is available in the NCBI BioProject database under accession PRJNA294096, and the biological sample used in this study is registered in the BioSample database under accession SAMN04014715. The nuclear genome assembly used for all analyses is available in the NCBI Genome/GenBank database under accession GCA_003260245.2. All annotation data and derived resources are provided through a single Zenodo repository (https://doi.org/10.5281/zenodo.17901083). This archive includes the soft-masked reference assembly, unified genome annotation GFF3, predicted proteome FASTA, repeat and ncRNA annotation files, and transcription factor list. When working with genomic coordinates, users should reference the NCBI genome assembly accession GCA_003260245.2, which defines the scaffold names and coordinate system used in all files. For most downstream analyses, the scaffold-level assembly and unified GFF3 are the recommended starting point, as they provide the most up-to-date and internally consistent view of the genome.

## Code availability

No custom code was developed specifically for this study. All analyses were performed using publicly available software tools described in the Methods section, with version numbers and key parameters reported to ensure reproducibility. Users can reproduce or adapt the annotation and validation steps using these tools together with the data hosted in the Zenodo repository (https://doi.org/10.5281/zenodo.17901083).

Received: 1 September 2025; Accepted: 8 January 2026;
Published online: 17 February 2026

## References


1. Ghazal, H. *et al*. Argane Genetics and Genomics. in *Oil Crop Genomics* (eds Tombuloglu, H., Unver, T., Tombuloglu, G. & Hakeem, K. R.) 123–134, https://doi.org/10.1007/978-3-030-70420-9_7 (Springer International Publishing, Cham, 2021).
2. Idrissi Azami, A. *et al*. Argan Tree Complete Chloroplast Genome Reconstruction via Hybrid Assembly. in *International Conference on Advanced Intelligent Systems for Sustainable Developent (AI2SD 2024)* (eds Ezziyyani, M., Kacprzyk, J. & Balas, V. E.) 86–94, https://doi.org/10.1007/978-3-031-91337-2_8 (Springer Nature Switzerland, Cham, 2025).
3. Khayi, S. *et al*. Complete Chloroplast Genome of Argania spinosa: Structural Organization and Phylogenetic Relationships in Sapotaceae. *Plants* **9**, 1354 (2020).
4. Idrissi Azami, A. *et al*. The complete mitochondrial genome data of *Argania spinosa* (L.) Skeels. *Data Brief* **57**, 110862 (2024).
5. Khayi, S. *et al*. First draft genome assembly of the Argane tree (Argania spinosa). *F1000Research* **7**, 1310 (2020).
6. Rupp, O. *et al*. Genome of Argania spinosa L.: insights into oil production and the tocopherol biosynthesis pathway. *Genet. Resour. Crop Evol.* **71**, 4027–4042 (2024).
7. 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. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
8. Blum, M. *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).
9. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
10. Johnson, M. *et al*. NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36**, W5–W9 (2008).
11. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. *Methods Mol. Biol. Clifton NJ* **406**, 89–112 (2007).
12. Hawkins, C. *et al*. Plant Metabolic Network 16: expansion of underrepresented plant groups and experimentally supported enzyme data. *Nucleic Acids Res.* **53**, D1606–D1613 (2025).
13. Jin, J. *et al*. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* **45**, D1040–D1045 (2017).
14. Argania spinosa (ID 294096) - BioProject - NCBI. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA294096/.
15. Ghazal, H. Plant sample from Sideroxylon spinosum - BioSample - NCBI. https://www.ncbi.nlm.nih.gov/biosample/SAMN04014715/.
16. Ghazal, H. SRP077839: Study: SRA Archive: NCBI. https://trace.ncbi.nlm.nih.gov/Traces/?view=study&acc=SRP077839.
17. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
18. Miller, J. R. *et al*. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824 (2008).
19. Boukhari, A. E. *et al*. Flow cytometry and chromosome numbers variation in argan tree Argania spinosa (L.) Skeels. *Not. Sci. Biol.* **15**, 11451–11451 (2023).
20. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
21. Bankevich, A. *et al*. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
22. Astashyn, A. *et al*. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol.* **25**, 60 (2024).
23. Kieras, M. Zanfona, a genome finishing process for use with paired-end short reads. (2021).
24. Mikheenko, A., Saveliev, V., Hirsch, P. & Gurevich, A. WebQUAST: online evaluation of genome assemblies. *Nucleic Acids Res.* **51**, W601–W606 (2023).
25. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
26. Tempel, S. Using and understanding RepeatMasker. *Methods Mol. Biol. Clifton NJ* **859**, 29–51 (2012).
27. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
28. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
29. Wei, C. *et al*. Draft genome sequence of Camellia sinensis var. sinensis provides insights into the evolution of the tea genome and tea quality. *Proc. Natl. Acad. Sci. USA.* **115**, E4151–E4158 (2018).
30. Swarbreck, D. *et al*. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**, D1009–D1014 (2008).
31. Unver, T. *et al*. Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci. USA.* **114**, E9413–E9422 (2017).
32. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics Bioinforma.* **2**, lqaa026 (2020).
33. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
34. Jacques, D. Another Gff Analysis Toolkit (AGAT). *Zenodo* https://doi.org/10.5281/zenodo.3549547 (2019).
35. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol. Biol. Clifton NJ* **1962**, 227–245 (2019).
36. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
37. Seemann, T. Barrnap: BAsic Rapid Ribosomal RNA Predictor. (2025).
38. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
39. Ontiveros-Palacios, N. *et al*. Rfam 15: RNA families database in 2025. *Nucleic Acids Res.* **53**, D258–D267 (2025).
40. Idrissi Azami, A. *et al*. Argania spinosa Genome Annotation. *Zenodo* https://doi.org/10.5281/zenodo.17901083 (2025).
41. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_003260245.2 (2022).


## Acknowledgements

Genome sequencing has been funded by Iridian Genome. The computational resources of HPC-MARWAN (www.marwan.ma/hpc) are provided by the National Center for Scientific and Technical Research (CNRST), Rabat, Morocco. HG is a recipient of an NIH grant through the h3abionet/h3africa consortium.

## Author contributions

Abdellah Idrissi Azami: Data curation, Formal analysis, Investigation, Methodology, Writing original draft. Stacy Pirro: Data curation, Formal analysis, Funding acquisition, Investigation, Resources, Writing review & editing. Nihal Habib: Data curation, Formal analysis, Visualization, Writing original draft. Turgay Unver: Investigation, Writing review & editing. M. Gonzalo CLAROS: Investigation, Writing review & editing. Juan de Dios Alche: Investigation, Writing review & editing. Sofia Sehli: Data curation, Formal analysis, Visualization, Writing review & editing. Zainab El Ouafi: Data curation, Formal analysis, Visualization, Writing review & editing. Douae El Ghoubali: Data curation, formal analysis, visualization, writing, review, and editing. Dalila Bousta: Investigation,

Methodology, Resources, Writing review & editing. Najib Al Idrissi: Funding acquisition, Methodology, Supervision, Writing review & editing. Fatima Gaboun: Data curation, Formal analysis, Investigation, Visualization, Writing review & editing. Abderrazak Rfaki: Investigation, Methodology, Resources, Writing review & editing. Abdelkhalek Legssyer: Conceptualization, Funding acquisition, Resources, Writing review & editing. Chakib Nejjari: Funding acquisition, Resources, Supervision, Writing review & editing. Saaid Amzazi: Methodology, Resources, Supervision, Writing review & editing. Lahcen Belyamani: Funding acquisition, Resources, Supervision, Writing review & editing. Abdelhamid El Mousadik: Conceptualization, Funding acquisition, Investigation, Resources, Supervision, Writing review & editing. Hassan Ghazal: Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Writing review & editing.

## Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-06596-7.

**Correspondence** and requests for materials should be addressed to H.G.

**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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/.

© The Author(s) 2026