Genet Resour Crop Evol (2024) 71:4027–4042
https://doi.org/10.1007/s10722-024-01931-6

SHORT COMMUNICATION

Genome of *Argania spinosa* L.: insights into oil production and the tocopherol biosynthesis pathway

Oliver Rupp · Clemens Roessner · Naemi Lederer-Ponzer · Tassilo Erik Wollenweber · Annette Becker · Mouna Lamaoui

Received: 29 September 2023 / Accepted: 19 February 2024 / Published online: 17 March 2024
© The Author(s) 2024

Abstract Argan (*Argania spinosa* L.) is a highly valued tree for its multiple uses as food and feed and for being linked to a broad range of benefits of pharmaceutical and cosmetic relevance. This multipurpose resource is becoming increasingly overused, which may disrupt the whole ecosystem's sustainability. Due to the high socio-economic status of this tree, research interventions are needed to reverse the forest regressive trend, restore the disturbed ecosystem, and conserve genetic diversity. However, research on argan is restricted by the lack of accessible information on the genetic and genomic bases of the species, specifically a functional annotated genome. Herein, we report the reference transcriptome aided annotation of the argan tree genome using de novo gene prediction programs aided by homology information from different plants. The results of the genome annotation using AUGUSTUS were subsequently improved by performing RNA sequencing. A total of 62,590 gene loci could be identified with 82,286 isoforms and a BUSCO completeness of 91.7%. To gain insight into the agronomically important compounds in argan oil, a comparative genome analysis and ortholog identification was performed, followed by phylogenetic tree construction of the main biosynthesis genes. Among those are fatty acids and tocopherols, the latter being the main factor behind the increasing demands for argan oil. Our analysis is the initial step to provide breeders, geneticists, and the industries with adequate genomic information, facilitate improvement of economically important traits and to selectively adapt the tree to the increasing impact of climate change.

Keywords Argan tree · Genome annotation · Argan oil · Tocopherols · Fatty acid · Triacylglycerol

O. Rupp
Bioinformatics and Systems Biology, Justus-Liebig-University, Giessen, Germany

C. Roessner · N. Lederer-Ponzer · A. Becker · M. Lamaoui (✉)
Developmental Biology of Plants Group, Institute of Botany, Justus-Liebig-University, Giessen, Germany
e-mail: mounalamaoui@gmail.com

T. E. Wollenweber
Biological Medical Research Center (BMFZ), Medical Faculty and University Hospital Düsseldorf, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

M. Lamaoui
Department of Biology, Polyvalent Laboratory of Research and Development, Polydisciplinary Faculty, Sultan Moulay Sliman University, Beni-Mellal, Morocco

Introduction

Argania spinosa L. (Fig. 1), belonging to the family Sapotaceae in the order Ericales, is an endemic species distributed in more than 800,000 Hectares in the southwestern area of Morocco. Argan oil production has tremendously eased the poverty of around 2.5 million members of the rural population living nearby

Fig. 1 Argan (*Argania spinosa* L.,) tree, in its natural habitat in southwestern area of Morocco

(Lybbert et al. 2002, 2011; Guillaume and Charrouf 2011; Mateille et al. 2016). With the ever-growing population and the booming market, the demand for the species has amplified greatly in the latest years owing to the oil extracted from its almonds, which is marketed worldwide and highly appreciated due to its high medicinal, cosmetic, and pharmacological value (Gonzálvez et al. 2010; Guillaume and Charrouf 2011; Cabrera-Vique et al. 2012). In addition to its ecological and socio-economic value, many investigations have revealed that the species has a high level of stress tolerance (Chakhchar et al. 2015; Ain-Lhout et al. 2016), which makes it the best candidate for an environmentally friendly agriculture (Polain and Waroux 2013; Mateille et al. 2016). However, despite the tree's well-known tolerance, there is an evidence of an alarming regression in argan forest, which is altering the ecosystem to a large extent. Morocco, as the main argan oil producer, is also highly vulnerable to land degradation due to its geographic location within a drought-prone spot of the Mediterranean area (Gonzálvez et al. 2010; Cabrera-Vique et al. 2012). This situation has led to a decline in the green forest landscape and to the genetic diversity erosion (Alados and Aich 2008; Díaz-Barradas et al. 2010; Alba Sánchez et al. 2015; Zhao et al. 2019). Since the late 1900s, the forest has experienced significant shrinking in milder areas with a subsequent decrease in fruit production and oil yielding (Díaz-Barradas et al. 2010; Zunzunegui et al. 2010; Ait Aabd et al. 2014).

Many investigations have designated climate changes along with other anthropogenic factors as the key elements leading to the perceived loss (Alados and Aich 2008; Charrouf et al. 2008; Alba Sánchez et al. 2015). Therefore, under the incessant threat of global warming and the thriving oil market (Le Polain and Waroux 2013), interventions are required to restore the natural forest, which is in danger of extinction, specifically as additional contractions are predicted under future climate change scenarios (Alba Sánchez et al. 2015). The implementation of new strategies that could address the issues related to forest loss is urgently needed for the sustainable development of the agroforestry system.

Many gaps remain to be filled in terms of research in this area and further inputs are required regarding the argan genomic basis to improve our knowledge of the bottlenecks that limit argan domestication and breeding. Over the past several years, efforts have mainly been oriented to the use of DNA-marker-based technology to characterize the high level of intra-species genetic diversity across a broad geographical range of forests (Ait Aabd et al. 2019).

Understanding this diversity, will allow for the incorporation of superior economic traits through breeding programs. Moreover, research trials on the argan genetic and genomic bases will greatly boost these efforts. Prior to 2018, these efforts were impeded by the lack of availability of nucleotide sequences in the public databases. Recently, a draft genome of *A. spinosa* assembly was generated by a hybrid de novo assembly method that combines short- and long sequencing reads (Khayi et al. 2018). The 671 Mb assembly was estimated to cover 89% of the genome of the highly heterozygous diploid species, 2n = 2x = 20, with x = ploidy number and n = gametic number of chromosomes (Majourhat et al. 2007; Khayi et al. 2018; El Boukhari et al. 2023). However, an annotation was lacking and the genome was not easily accessible. Here, we provide an annotation for the generated raw genome and an easy access for scientific community to perform further research. We generated a reference transcriptome for argan and used it to annotate the argan genome (Version May 2019). Further, we identified the argan homologs of biosynthetic pathways relevant to its agronomical value with special focus on the tocopherol biosynthesis.

Materials and methods

Gene and transposable element prediction and genome annotation

To annotate the *A. spinosa* genome, the draft genome generated in a previous project (Khayi et al. 2018) was used as a resource to execute a homology-based gene prediction program that combines gene models from multiple reference organisms. The obtained sequences and annotations from the evolutionarily related reference species kiwi (*Actinidia chinensis*; GCA_003024255.1), tea (*Camellia sinensis* L.; GCF_004153795.1), *Rhododendron williamsianum* (GCA_009746105.1), and blueberry (*Vaccinium corymbosum* L.) (www.vaccinium.org, Draper v1.0) were integrated to build preliminary gene models using Gene Model Mapper (GeMoMa) v1.6.3 (Keilwagen et al. 2016). The models were subsequently combined and filtered to generate a set of complete gene models. From these models, 2000 genes were randomly selected and split into two sets.

Genome annotation was subsequently performed using AUGUSTUS software (v. 3.4.0). The selected subset of predicted genes was used during the training step and the evaluation of the AUGUSTUS gene prediction tool (Stanke et al. 2006) according to Hoff and Stanke (2019). All unfiltered GeMoMa models were subsequently converted into AUGUSTUS "hints" as "CDSpart" and "intron".

RepeatModeler (Smit and Hubley 2015) was used to identify repeats and transposable elements in the genome. The resulting sequences were applied to the genome using RepeatMasker (Smit et al. 2015). The RepeatMasker results were converted into AUGUSTUS "hints" as "nonexonpart". The final gene model was built using AUGUSTUS with default parameters and allowing for alternative splice variants based on "hints".

Genome completeness and gene function prediction

The gene prediction quality was evaluated exploiting the software Benchmarking Universal Single-Copy Orthologs (BUSCO) v.4 (Seppey et al. 2019). BUSCO analyses the coverage of Single-Copy Orthologs that are designated as BUSCO versions of the gene models used to train the AUGUSTUS (v3.4.0) gene prediction tool (Stanke et al. 2004). The completeness of the model was evaluated using the "viridiplantae" and "eudicots" datasets.

Functional descriptions were assigned to the predicted models based on homology. The predicted genes were blasted alongside the UniProt/Swiss-Prot (http://beta.uniprot.org/uniprot) and UniProt/TrEMBL (http://www.uniprot.org/) databases. The functional annotation was derived from the reference protein with the best blast hit (the highest score). The final genome assembly results were uploaded to the GenDBE annotation platform (Meyer et al. 2003).

RNA extraction and PacBio IsoSeq library construction

In order to improve the annotated *A. spinosa* genome assembly with full-length transcripts, an IsoSeq analysis of *A. spinosa* RNA was facilitated. The gene prediction was subsequently improved by performing RNA sequencing.

Sample material and RNA extraction

Plant material for RNA-seq was collected from Bonn University botanic gardens at Germany (IPEN numbers: XX-0-BONN-19646 procured in 1980, and MA-0-BONN-24662 and MA-0-BONN-40479 acquired in 2005 from Agadir, Morocco). Due to the high and disturbing amounts of phenolic compounds in the leave tissues, 100–500 mg of root and flower tissues was separately collected and used for our experiment. Extraction of the total RNA was performed with the NucleoSpin® RNA Plant and Fungi Kit (Macherey–Nagel GmbH & Co. KG, Düren, Germany) according to the manufacturer's instructions. The cDNA libraries were prepared, and strand-specific sequencing was conducted by the Biological and Medical Research Center (BMFZ), Düsseldorf, Germany.

To avoid bias toward specific transcripts, the RNAs from the different tissues were used at equal amount and sequenced using PacBio's Iso-Seq method. The *A. spinosa* RNA pool, of root and flower tissues, was assessed using the Qubit RNA HS Assay (Thermo Fisher Scientific Massachusetts, U.S.A.) and the NanoDrop (Thermo Fisher Scientific, Massachusetts, U.S.A.) to check the concentration, while the FragmentAnalyzer (Agilent Technologies, California, U.S.A.) was used to check the RNA integrity. Subsequently, the intact poly(A) RNA was captured and purified using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England BioLabs, Massachusetts, U.S.A.) following the manufacturer's instructions. Next, the cDNA was synthesized according to the TeloPrime Kit Version 2 (Lexogen, Vienna, Austria). The optimal amount of cycles was determined by qPCR (QuantStudio 3, Thermo Fisher Scientific, Massachusetts, U.S.A.) with 0.1 × SYBR Green (Merck Millipore, Darmstadt, Germany), TeloPrime kit chemistry and 10% of the cDNA as input material. Based on the qPCR, the amount of cycles was adjusted to yield appropriate amounts of amplified cDNA (500 ng to 2 µg) within the Endpoint-PCR. The quality of the mass-amplified full-length cDNA was assessed according to the manufacturer's instructions and column-based purified (TeloPrime Kit Version 2, Lexogen, Vienna, Austria) followed by the IsoSeq PacBio library preparation, using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, California, U.S.A.). Subsequently, the quality (FragmentAnalyzer, Agilent Technologies, California, U.S.A.) and quantity (Qubit, Thermo Fisher Scientific, Massachusetts, U.S.A.) was determined. The long-read sequencing was executed on a Sequel IIe system with a single SMRT cell (8 M), the Sequel II Binding kit 2.1, and the Sequel II Sequencing kit 2.0 (Pacific Biosciences, California, U.S.A.). The SMRT cell was loaded with 120 pM of material. The movie time of 24 h started after a 2 h of immobilization step and a 2 h of pre-extension step.

Processing of raw PacBio reads and Iso-Seq analysis

The raw subread dataset was processed with SMRT Link v10.1 (Pacific Biosciences, California, U.S.A.) to generate the circular consensus sequences (CCS) reads, internally using ccs v.6.0.0. cDNA barcodes introduced with the TeloPrime Kit Version 2 (Lexogen, Vienna, Austria) were trimmed via lima (v 2.1.0). Subsequently, the high-quality full-length transcripts were processed with isoseq3 (v 3.4.0).

The RNA-seq reads were reallocated to the genome assembly in the GenDBE annotation platform (Meyer et al. 2003).

Comparative genome analysis and ortholog identification

Protein sequences from *Arabidopsis thaliana* (https://www.arabidopsis.org), *A. chinensis* (https://kiwifruitgenome.org/ftp/A_chinensis/Red5/v1.0/), *A. spinosa, Brassica napus* (http://www.genoscope.cns.fr/brassicanapus), *Helianthus annuus* (https://sunflowergenome.org), *C. sinensis* (https://www.ncbi.nlm.nih.gov/assembly/GCF_004153795.1), *Medicago truncatula* (http://www.medicagogenome.org/), and *Triticum aestivum* (https://wheat-urgi.versailles.inra.fr/Tools) were used to perform a quantitative analysis of the different protein sets, their intersections, and aggregates of intersections through the visualization of intersections among multiple sets (Lex et al. 2014). Only the longest splice-variant of each gene was used; the sequences with internal stop-codons or internal gaps were removed. The number of orthologous groups, both shared and unique, is presented for each genome.

To proceed with the analysis, we used a set of the main genes related to argan oil production, namely tocopherol and phytosterol-pathway-related genes and implicated in fatty acid (FA) and triacylglycerol (TAG) biosynthesis. Tocopherol, phytosterol, FA, and TAG are the four main components of argan oil. A manual examination of the genes present in the argan genome was performed using the *A. thaliana* protein sequence and the GenDBE blast tool. To investigate the contraction of the gene families in argan and to determine lineage-specific genes, data on the genome of *A. spinosa*, *C. sinensis, A. chinensis, M. truncatula*, *H. annuus*, *B. napus* and *A. thaliana* were used to perform a comparative analysis using OrthoFinder (v.2.4.0) to identify the orthogroups (Emms and Kelly 2015).

Sequence analysis and construction of phylogenetic trees

Sequences were retrieved using BLAST (Altschul et al. 1990) searches in JGI Phytozome (Goodstein et al. 2012) (*Oryza sativa, Z. mays, Hordeum vulgare, T. aestivum, Solanum lycopersicum, H. annuus, Glycine max, A. thaliana, Capsella rubella, Brassica rapa and Amborella trichopoda*), NCBI (Sayers et al. 2022) (*A. chinensis* and *C. sinensis*) and an in house database for *A. spinosa*. Full-length protein sequences were aligned using MAFFT (Katoh and Standley 2013) (--localpair --maxiterate 10,000 --reorder), resulting alignments were then used to generate consensus trees (maximum-likelihood, 100 bootstraps) with iqtree (Minh et al. 2020) (-T AUTO -m TEST -b 100 -allnni -con).

Results

Gene prediction, transposable elements and genome completeness

Herein, we present an annotated *A. spinosa* genome assembly that represents 99.76% of the predicted 671 Mb genome. A total of 287,123 genes were obtained from four reference genomes, kiwi fruit (*A. chinensis*), tea (*C. sinensis* L.), *R. williamsianum*, and blueberry (*V. corymbosum* L.). 62,590 predicted genes, which are encoding 82,286 transcripts were identified by the AUGUSTUS software (Stanke et al. 2006). Of these transcripts, 49,575 transcripts were completely supported by homologous proteins from at least one of the reference annotations and 10,295 were novel gene models. The detailed characterization of these predicted genes suggests that the mean number of exons per gene is 4, with an average of 220 bp and 776 bp for exon and intron lengths respectively. The transcripts lengths (mRNA) varying from 9 to 136,337 bp, with an average of 3717 bp. Furthermore, based on the RepeatModeler analysis used to process the genomic sequence and identify repeated sequences, a total of 2390 repeat family sequences were identified in the complete genome and 56.33% of all sequences were masked by RepeatMasker.

To ensure that the predicted gene set is complete and accurate, the annotation was evaluated by BUSCO (Simao et al. 2015). The results from the analysis reveal that the adopted pipeline was able to identify 92%, with 0.2% missing, and 91.7%, with 2.4% missing, of the complete data for “viridiplantae” and “eudicots”, respectively. The functional annotation resulted in a reliable assignment for 25,436 transcripts, with 14,292 transcripts not having a significant hit based on their homology to the used protein databases.

Reference transcriptome

The argan reference transcriptome comprises 128,978 Pac Bio IsoSeq HQ transcripts extracted from 2,907,823 PacBio IsoSeq reads. Of those, 21.21% fully mapped the genome, while 59.35% were unmapped. Of the latter one, a large fraction mapped to fungi (21.27%), suggesting strong fungal association of the roots with fungi or putative fungal infections of flower material collected from the botanical garden. Around 6% of the transcripts were identified as putative transposons.

Comparative genome analysis and gene identification

Comparative analyses of the whole protein sets (Table 1) from *A. thaliana*, *A. chinensis* (kiwi), *A. spinosa, B. napus* (canola), *H. annuus* (sunflower), *C. sinensis* (tea), *M. truncatula* (barrel clover), and *T. aestivum* (wheat) were performed. The argan genome includes 1,140 ortholog groups not identified in the other species (Fig. 2). 7,753 groups were shared

Table 1 Comparative analysis of orthologous groups involved in oil production and vitamin E biosynthesis identified in *Argania spinosa*, *Arabidopsis thanliana, Actinidia chinensis, Brassica napus, Helianthus annuus* and *Camellia sinensis*

Gene name	Arabidopsis ID	Number of genes				
		Argania spinosa	*Brassica nappus*	*Helianthus annus*	*Actinidia chinensis*	*Camellia sinensis*
Tocopherol pathway-related genes						
PDS1/HPPD	AT1G06570	QLOD01006489.1.g882.t1	XP_013695640.1 XP_013695641.1 XP_013716383.1 XP_013723237.1 XP_013723238.1 XP_022563345.1	XP_022005727.1	PSS05627.1	XP_028059699.1 XP_028059700.1
VTE5	AT5G04490	QLOD01018272.1.g582.t1 QLOD01018272.1.g582.t2 QLOD01018272.1.g582.t3 QLOD01018272.1.g582.t4 QLOD01065300.1.g177.t1	XP_013668182.1 XP_013724551.1 XP_013729436.1 XP_022554205.1 XP_022569215.1	XP_021976875.1 XP_022006148.1 XP_035831086.1	PSR98861.1 PSS05050.1	XP_028072315.1
VTE6	AT1G78620	QLOD01071129.1.g42.t1	XP_013650142.1 XP_022560444.1	XP_021999954.1 XP_022009533.1	PSS05930.1	XP_028053793.1 XP_028053794.1 XP_028053795.1 XP_028053796.1 XP_028102961.1
VTE2	AT2G18950	QLOD01009084.1.g1166.t1	XP_013652757.1 XP_013652758.1 XP_013731670.1	XP_021989602.1 XP_021996863.1 XP_035836649.1	PSS20757.1 PSS36395.1	XP_028063573.1
VTE3	AT3G63410	QLOD01013891.1.g18.t1 QLOD01061624.1.g654.t3	XP_013648530.1 XP_013706563.1 XP_022567384.1 XP_022567385.1 XP_022567386.1	XP_021970310.1 XP_022035105.1	PSR86098.1 PSR92770.1 PSS06349.1	XP_028063270.1 XP_028120485.1
VTE1	AT4G32770	QLOD01029813.1.g420.t1	XP_013707373.1 XP_013719549.1 XP_013719625.1	XP_021978512.1	PSS01197.1	XP_028095320.1
VTE4	AT1G64970	QLOD01065222.1.g1059.t1	NP_001302948.1 NP_001303012.1 NP_001303158.1 XP_013686352.1 XP_013702364.1 XP_022553058.1 XP_022553059.1	XP_021976819.1 XP_021976820.1 XP_021976821.1 XP_021993981.1	PSS11874.1 PSS33003.1	XP_028068374.1 XP_028119337.1
GGPS	AT2G18620.1 AT4G36810.1	QLOD01009755.1.g457.t1 QLOD01009755.1.g457.t2 QLOD01009755.1.g457.t3 QLOD01013488.1.g1001.t1 QLOD01013488.1.g1001.t2 QLOD01048430.1.g823.t1	XP_013685469.1 XP_013722211.1 XP_013738757.1 XP_022570754.1	XP_021998516.1 XP_022009289.1	PSR88026.1 PSR99708.1 PSS30802.1	XP_028079465.1 XP_028108146.1
MCT	AT2G02500	QLOD01014832.1.g834.t1 QLOD01014832.1.g834.t2 QLOD01014832.1.g834.t3 QLOD01018920.1.g1087.t1 QLOD01018920.1.g1087.t2 QLOD01018920.1.g1087.t3	XP_013692059.1 XP_022552730.1 XP_022552731.1	XP_021981854.1 XP_022038869.1	PSS21671.1 PSS29099.1	XP_028127240.1
HMBPP	AT5G60600	QLOD01014508.1.g525.t1 QLOD01014508.1.g525.t3	XP_013660691.1 XP_013660692.1 XP_013667146.1 XP_022565353.1	XP_021973422.1 XP_022005842.1 XP_035831158.1	PSR95187.1 PSS14396.1	XP_028070816.1 XP_028109511.1
TAT	PSR96377.1	QLOD01074171.1.g911.t1 QLOD01074171.1.g911.t2	XP_013676250.1 XP_022551697.1 XP_022551698.1	XP_021989100.1 XP_021999922.1	PSR96377.1	XP_028055318.1 XP_028055320.1
Fatty acid biosynthesis related genes						
BCCP	AT5G16390.1	QLOD01061605.1.g537.t1 QLOD01061605.1.g537.t2 QLOD01061605.1.g537.t3	XP_013649844.1 XP_013663620.1 XP_013663621.1 XP_013719861.1 XP_022567554.1		PSR92923.1 PSR98117.1	XP_028054409.1
	AT5G15530	QLOD01004145.1.g656.t1 QLOD01071143.1.g118.t1 QLOD01071143.1.g118.t2 QLOD01071143.1.g118.t3	XP_013667772.1 XP_013720889.1 XP_013720931.1 XP_013723061.1 XP_013728269.1 XP_013728908.1 XP_022554431.1	XP_021983849.1 XP_022016303.1 XP_022017547.1	PSR95777.1 PSR96615.1	XP_028112272.1
α-CT	AT2G38040	QLOD01011086.1.g295.t1 QLOD01011086.1.g295.t2 QLOD01016262.1.g567.t1 QLOD01016262.1.g567.t2	NP_001302844.1 XP_013723918.1 XP_013723919.1 XP_013723921.1 XP_013734703.1 XP_013739095.2 XP_013748367.1 XP_013748368.1 XP_022568514.1	XP_021987858.1XP_021988212.1 XP_022010319.1	PSS32670.1 PSS34625.1	XP_028064549.1 XP_028064550.1 XP_028067376.1 XP_028067384.1

Table 1 (continued)

Gene name	Arabidopsis ID	Number of genes				
		Argania spinosa	*Brassica nappus*	*Helianthus annus*	*Actinidia chinensis*	*Camellia sinensis*
MCMT	AT2G30200	QLOD01062612.1.g1242.t1 QLOD01062612.1.g1242.t2 QLOD01062612.1.g1242.t3	XP_013707428.1 XP_013748818.1 XP_022551600.1	XP_021974192.1 XP_022011299.2	PSS09521.1	XP_028057891.1
ACP5	AT5G27200 AT4G25050.2 AT3G05020.1 AT1G54630.1 AT1G54580.1	N.D	XP_013684429.1 XP_013684436.1 XP_013732111.1 XP_013732113.1 XP_013732415.1 XP_013732416.1 XP_013735795.1 XP_013750277.1 XP_013750679.1 XP_013750751.1 XP_022548030.1 XP_022575553.1	N.D	N.D	N.D
KASII	AT1G74960	QLOD01001226.1.g500.t1 QLOD01062040.1.g286.t1 QLOD01062040.1.g286.t2	XP_022556216.1	XP_021980867.1 XP_021980868.1	PSS02051.1 PSS15439.1 PSS20760.1 PSS36397.1	XP_028056275.1
KAR	AT1G24360	QLOD01004627.1.g1207.t1 QLOD01004627.1.g1207.t2 QLOD01004627.1.g1207. t3QLOD01065352.1.g325.t1	NP_001302983.1 NP_001303056.1 XP_013653269.1 XP_013662543.1 XP_013694584.1 XP_013721653.1	XP_021990467.1 XP_022025737.1	PSR84786.1 PSS07695.1 PSS13657.1 PSS26936.1 PSS28781.1	XP_028098462.1 XP_028098463.1 XP_028113377.1 XP_028113455.1 XP_028118573.1
HAD	AT2G22230.1, AT5G10160.1	QLOD01036103.1.g983.t1 QLOD01065532.1.g692.t1 QLOD01065532.1.g692.t2	NP_001302840.1 XP_013680285.1 XP_013712370.1 XP_013726624.1	XP_021988579.1 XP_022027749.1	PSS26215.1	XP_028081821.1 XP_028122483.1 XP_028122487.1
ER	AT2G05990	QLOD01017438.1.g1409.t1 QLOD01074999.1.g283.t1 QLOD01074999.1.g283.t2	NP_001303180.1 NP_001303196.1 XP_013685121.1 XP_013685122.1 XP_013714468.1 XP_013714469.1	XP_022018336.1 XP_022020754.1 XP_022026461.1 XP_035841570.1	PSR86864.1 PSR90228.1 PSS21133.1 PSS36309.1	XP_028116071.1 XP_028116072.1 XP_028116073.1
FAB2	AT2G43710.2 AT3G02630.1	QLOD01062571.1.g1121. t1QLOD01074982.1.g208.t1 QLOD01074982.1.g208.t2 QLOD01075084.1.g728.t1	NP_001302885.1 XP_013656022.1 XP_013692042.1 XP_013735719.1 XP_013744476.1 XP_022554884.1	XP_021982111.1 XP_021994071.1	PSR98141.1 PSS32813.1 PSS34835.1 PSS35739.1 PSS35740.1	XP_028062579.1 XP_028081174.1 XP_028081175.1 XP_028081176.1 XP_028086173.1 XP_028110737.1
FatA-1	AT3G25110.1 AT4G13050.1	QLOD01013551.1.g1131.t1 QLOD01013551.1.g1131.t2 QLOD01013551.1.g1131. t3QLOD01027318.1.g630.t1 QLOD01027318.1.g630.t2 QLOD01027318.1.g630.t3	NP_001302598.1 XP_013676282.1 XP_013689229.1 XP_013724176.1 XP_013732603.1 XP_013742950.1	XP_021976166.1	PSS14097.1 PSS19368.1	XP_028079988.1
FatB	AT1G08510	QLOD01022658.1.g659.t1 QLOD01065424.1.g513.t1	NP_001302816.1 NP_001303052.1 XP_013641077.1 XP_013657614.1 XP_013657615.1 XP_013657646.1 XP_013657647.1 XP_013672632.1 XP_013672633.1 XP_013672634.1 XP_013672635.1 XP_013678692.1 XP_013678693.1 XP_013678714.1 XP_013678716.1 XP_013711736.1 XP_013716474.1 XP_022545386.1 XP_022546165.1 XP_022546166.1 XP_022547306.1	XP_021970581.1 XP_021970582.1 XP_021981073.1 XP_022037542.1	PSS01852.1 PSS15617.1 PSS20889.1 PSS20891.1	XP_028060544.1 XP_028091757.1
LACS8	AT2G04350	QLOD01074157.1.g874.t1 QLOD01074157.1.g874.t2 QLOD01074157.1.g874.t3 QLOD01074157.1.g874.t4	XP_013657081.1 XP_013686799.1	XP_022036256.1	PSS06471.1 PSS16071.1	XP_028121908.1
LACS9	AT1G77590	QLOD01031498.1.g1199. t1QLOD01031498.1.g1199. t2QLOD01031498.1.g1199.t3	NP_001303207.1 XP_013648691.1 XP_013648692.1 XP_013686440.2 XP_013686441.2 XP_022568668.1 XP_022569031.1 XP_022569032.1	XP_021982865.1 XP_021992725.1 XP_035835469.1	PSS19913.1	XP_028089154.1 XP_028089155.1

Table 1 (continued)

Gene name	Arabidopsis ID	Number of genes				
		Argania spinosa	*Brassica nappus*	*Helianthus annus*	*Actinidia chinensis*	*Camellia sinensis*
FAD2	AT3G12120	QLOD01000980.1.g135.t1 QLOD01000980.1.g135.t2 QLOD01000980.1.g135.t3 QLOD01014567.1.g641.t1 QLOD01014567.1.g641.t2, QLOD01063726.1.g1081.t1 QLOD01063726.1.g1081.t2	XP_013750122.1 XP_013750123.1 XP_022568458.1 XP_022575480.1	XP_021972383.1 XP_022000024.1 XP_022000025.1 XP_022005851.1	PSR86868.1 PSR90232.1 PSS21132.1 PSS36308.1	XP_028064735.1 XP_028080752.1 XP_028080753.1
FAD6	AT4G30950	QLOD01003815.1.g162.t1 QLOD01003815.1.g162.t2	XP_013671715.1 XP_013720391.1 XP_013730775.1	XP_021990098.1	PSS13938.1	XP_028063331.1
FAD7	AT3G11170	QLOD01003562.1.g1184.t1	XP_013648169.1	XP_021973951.1	PSR93207.1	XP_028091727.1
FAD8	AT5G05580.1	QLOD01071352.1.g14.t1	XP_013718998.1	XP_021977282.1	PSS23830.1	XP_028121231.1
FAD3-1	AT2G29980.1	QLOD01013184.1.g751.t1 QLOD01024743.1.g664.t1 QLOD01024743.1.g664.t2 QLOD01024743.1.g664.t3, QLOD01071667.1.g1215.t4, QLOD01071667.1.g1215.t5, QLOD01071667.1.g1215.t6	XP_022575534.1 XP_022575535.1 XP_022575536.1 NP_001302640.1 NP_001302922.1 NP_001302953.1 XP_013683288.1 XP_022556026.1 XP_022571486.1	XP_021988417.1 XP_022022379.2 XP_021973565.1 XP_021976548.1 XP_021982993.1 XP_022007955.2	PSS33314.1 PSR86174.1 PSR95055.1 PSS09495.1 PSS14199.1	XP_028121232.1 XP_028082631.1 XP_028086292.1
KAS III	AT1G62640.1	QLOD01063552.1.g789.t1 QLOD01063552.1.g789.t2	XP_013644350.1 XP_013732147.1	XP_022022677.1	PSR91570.1 PSS04068.1 PSS11513.1	XP_028061555.1 XP_028080918.1
Triacylglycerol (TAG) biosynthesis related genes						
GPAT9	AT5G60620	QLOD01010876.1.g77.t1 QLOD01010876.1.g77.t2	XP_013660694.1 XP_022544647.1, XP_022544664.1	XP_021993396.1, XP_022006398.1, XP_022006399.1	PSS12140.1, PSS21806.1, PSS31548.1	XP_028093899.1, XP_028093900.1, XP_028093901.1
LPAT2	AT3G57650	QLOD01013530.1.g1089.t1	NP_001302955.1 NP_001302981.1 NP_001303095.1 XP_013648394.1	XP_021971464.1, XP_021988122.1, XP_035834470.1	PSS01819.1, PSS15655.1	XP_028114840.1, XP_028114841.1, XP_028114842.1
LPAT4/5	AT1G75020 AT3G18850	QLOD01017911.1.g296.t1, QLOD01023620.1.g1275.t1 QLOD01023620.1.g1275.t2	NP_001303230.1 XP_013648887.1 XP_013682530.1 XP_013713895.1 XP_013721420.1 XP_013721425.1 XP_013744108.1 XP_013744109.1 XP_013749364.1 XP_013749365.1 XP_013749366.1 XP_022566545.1 XP_022568253.1 XP_022568705.1	XP_021983651.1 XP_021983652.1 XP_021983655.1 XP_021983656.1 XP_021983657.1 XP_021989513.1 XP_022026113.1 XP_035833738.1 XP_035833739.1 XP_035833740.1 XP_035833741.1 XP_035833742.1 XP_035833743.1 XP_035833744.1	PSR99704.1 PSS19167.1 PSS26065.1	XP_028116956.1 XP_028116957.1 XP_028116958.1 XP_028120683.1
LPAT3	AT1G51260	N.D	XP_013648815.1	N.D	N.D	N.D
DGAT1-1	AT2G19450	QLOD01019286.1.g1331.t1 QLOD01019286.1.g1331. t2QLOD01019286.1.g1331. t3QLOD01019286.1.g1331.t4	NP_001302732.1 NP_001302852.1 NP_001303201.1 XP_013677909.1 XP_013698240.1 XP_013703160.1 XP_022546587.1	XP_021990560.1 XP_021996532.1 XP_022027170.1 XP_022027171.1	PSS01992.1	XP_028074203.1 XP_028074204.1
DGAT3	AT1G48300	QLOD01061660.1.g863.t1	XP_013655610.1 XP_013655611.1 XP_013711644.1 XP_013711650.1	XP_022038542.1	PSR92444.1 PSR97838.1	XP_028086598.1
LPCAT1 *LPCAT2*	AT1G12640 AT1G63050.1	QLOD01014043.1.g171.t1	XP_013647131.1 XP_013653836.1 XP_013653837.1 XP_013653838.1 XP_013661594.1 XP_013742603.1	XP_022002186.1 XP_022022257.1	PSS04135.1 PSS09786.1 PSS11440.1	XP_028070630.1
PDAT	AT5G13640	QLOD01071155.1.g209.t4 QLOD01071155.1.g209.t5 QLOD01074301.1.g246.t2	XP_013702678.2 XP_022568870.1	XP_021989957.1 XP_021996485.1 XP_022009772.1 XP_035843204.1	PSR96190.1 PSS20785.1 PSS36413.1	XP_028068140.1 XP_028068141.1 XP_028096748.1 XP_028096749.1
PDAT2	AT3G44830	QLOD01061562.1.g199.t1 QLOD01061562.1.g199.t2	XP_022555556.1	XP_022005723.1	PSR86522.1	XP_028090929.1
PDCT	AT3G15820 AT3G15830	QLOD01011046.1.g223.t1 QLOD01011046.1.g223.t2	XP_013647218.1 XP_013676101.1 XP_013682366.1 XP_013695400.1 XP_013700801.1 XP_013702685.1 XP_013702686.1 XP_013702687.1 XP_013736317.1	XP_022016096.1 XP_022026633.1 XP_022026634.1 XP_022026635.1	PSS07512.1 PSS17841.1	XP_028082991.1

Table 1 (continued)

Gene name	Arabidopsis ID	Number of genes				
		Argania spinosa	*Brassica nappus*	*Helianthus annus*	*Actinidia chinensis*	*Camellia sinensis*
DGAT2	AT3G51520	QLOD01000196.1.g200.t1 QLOD01014890.1.g885.t1 QLOD01017589.1.g60.t1 QLOD01075147.1.g1010.t1 QLOD01075147.1.g1010.t2 QLOD01075147.1.g1010. t3QLOD01075147.1.g1010.t4	XP_013644392.1 XP_013734399.2 XP_013737142.1 XP_013745595.1	XP_022009919.1 XP_022009922.1 XP_022009923.1 XP_022012849.1 XP_035838706.1 XP_035838707.1	PSS00351.1 PSS30104.1	XP_028062730.1 XP_028062731.1 XP_028076783.1
Phytosterol pathway-related genes						
SQE1	AT1G58440 AT2G22830.1 AT4G37760.1	QLOD01010950.1.g152.t1 QLOD01010950.1.g152.t2, QLOD01021269.1.g1303.t1 QLOD01021269.1.g1303.t2 QLOD01023617.1.g1264. t1QLOD01062204.1.g983.t1 QLOD01072090.1.g1010. t1QLOD01072090.1.g1010.t2	XP_013661694.1 XP_013661703.1 XP_013723029.1 XP_013723045.1 XP_013728814.1 XP_022548287.1 XP_022550124.1 XP_022556881.1 XP_022556882.1 XP_022567860.1 XP_022572297.1	XP_021987080.1 XP_021988698.1 XP_022013209.1 XP_022034985.1 XP_022037494.1	PSR87906.1 PSR99943.1 PSS21202.1 PSS26160.1 PSS31511.1	XP_028082948.1 XP_028098982.1 XP_028099341.1 XP_028101446.1 XP_028126688.1
OC	AT1G16600	QLOD01061933.1.g1122.t1	XP_022020442.1, XP_035830391.1	XP_028057583.1	PSR84602.1, PSS13730.1	XP_013641199.1
CYP51G1	AT1G11680	QLOD01061933.1.g1122.t1	XP_022020442.1, XP_035830391.1	XP_028057583.1	PSR84602.1, PSS13730.1	XP_013641199.1
STE1	AT3G02580 AT3G02590.1	QLOD01062284.1.g21.t1 QLOD01072430.1.g247.t1	XP_013661925.1 XP_013678388.1 XP_013735702.1 XP_013740870.1 XP_022544189.1	XP_022009441.1	PSR87830.1 PSS30593.1	XP_028057465.1
SMT2	AT1G20330 AT1G76090.1	QLOD01019647.1.g91.t1 QLOD01063444.1.g550.t1	XP_013642036.1 XP_013642037.1 XP_013702663.1 XP_013717974.1 XP_013728777.1 XP_013730004.1	XP_022011826.1 XP_022022330.1	PSR99550.1PSS21774.1	XP_028056461.1 XP_028115152.1

among all species and 603 groups are shared with its closest relatives, *C. sinensis.* Among those groups, 4,534 genes were identified as a specific singleton to argan compared with other species.

Further analysis of the conserved functional domains was performed through Pfam database (http://pfam.xfam.org/) to further characterize the hypothetical argan-specific genes. Among the other plant genomes of members of Ericales that we used in our study, 244 orthologue groups were shared in the intersection with the *A. chinensis* genome, while 603 orthologue groups were shared with the tea genome, which has the closest evolutionary relationship to the argan genome (Table 1).

A total of 664 genes, which consists mostly of transcription factors, were designated as specific to the *A. chinensis* genome, which is estimated to be 637.99 Mb (Wu et al. 2019). Regarding the genome of *C. sinensis*, more than 8,000 singletons were assigned to the 3.8–4.0 Gb genome (Sharma et al. 2018) (Fig. 2). The argan genome shares 3,510 and 831 orthologs with the tea genome and the kiwi genome, respectively (Fig. 2).

Sequence analysis and construction of phylogenetic trees

To gain insight into the agronomically important components of the argan genetic inventory, genes related to the major traits in oil crops, oil content and quality, the copy number of genes involved in these traits was analyzed. We identified the argan orthologs of oil biosynthesis related genes and their regulators known from other oil seed plants from the Ericales and Asterids using the OrthoFinder program. Biosynthetic pathways of specific interest were the tocopherol, fatty acid, triacylglycerol (TAG), and the phytosterol biosynthesis pathways. Members of these pathways were identified in the genomes of *A. spinosa, A. chinensis, A. thaliana, B. napus, C. sinensis,* and *H. annuus* and are summarized (Table 1). Due to its alloploid nature of its genome (Lu et al. 2019), *B. napus* encodes for often more than double the number of genes than the other plants species included in this analysis. Generally, gene numbers of those families are quite similar between *A. thaliana* and *A. spinosa* and we mention here only those genes that deviated in number from *A. thaliana*. By analyzing the

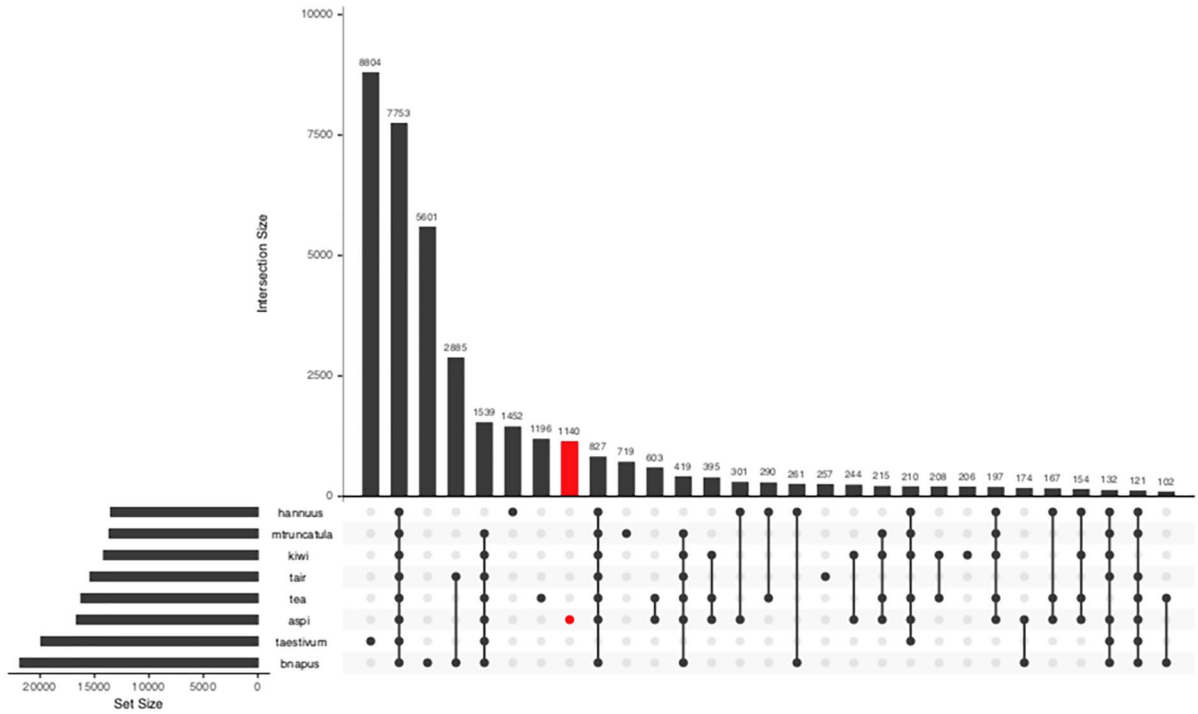

Fig. 2 Analysis of orthogroups in *Arabidopsis thaliana, Actinidia chinensis, Argania spinosa, Brassica napus, Helianthus annuus, Camellia sinensis, Medical truncatula,* and *Triticum aestivum*) created with UpSet R (Lex et al. 2014). The bars represent the number of orthologous groups (intersection size). The dots symbolize the species included in the intersection. The plot was restricted to intersections with more than 100 groups of orthologous

phytosterol pathway-related genes we found for the most important candidate gene, coding for the sterol methyltransferase II (*SMT2*) enzyme, three orthologs in *A. spinosa* and only two in *A. thaliana*. Furthermore, three *A. thaliana* genes encode for SQE1, but seven in *A. spinosa*.

In the TAG biosynthesis pathway, the genes encoding for LPAT2, DGAT3, and PDAT are single copy in *A. thaliana* and have two orthologs in *A. spinosa*. A single gene in *A. thaliana* also encodes DGAT3 but five orthologs are found in *A. spinosa*. However, LPCAT1 and LPCAT2 together have only a single ortholog in *A. spinosa*.

The fatty acid biosynthesis pathway includes a larger number of enzymes, most of the encoded by several genes. The enzymes α-CT, FatB, LACS9, and KASIII are encoded by a single gene in *A. thaliana* and by two in *A. spinosa*. The enzymes KASII, KAR, and FAD2 are also encoded by a single gene in *A. thaliana* but three in *A. spinosa*. The enzymes, BCCP, HAD and FAB2, are encoded by several genes in *A. thaliana* but *A. spinosa* encodes these genes by more copies.

Tocopherol biosynthesis (Table 1) includes mainly single copy genes in *A. thaliana* with the exception of GGPS (11 genes) and TAT (2 genes). Most of these single genes encoding for enzymes have two orthologs in *A. spinosa* (*VTE3*, *VTE5*, *MCT* and *HMBPP*). However, two genes encode TAT in *A. thaliana* and three in *A. spinosa* and the GGPS enzyme in *A. spinosa* is encoded by only 5 genes. Taken together, the gene numbers of the oil crop *A. spinosa* does not deviate substantially from the non-oil plant *A. thaliana*.

Evolutionary relationships of the tocopherol biosynthesis gene families

One of the main factors for the thriving demands for argan oil is its high content in total tocopherol in comparison to other oil seed crops. Tocopherols, vitamin E family substances, are lipophilic

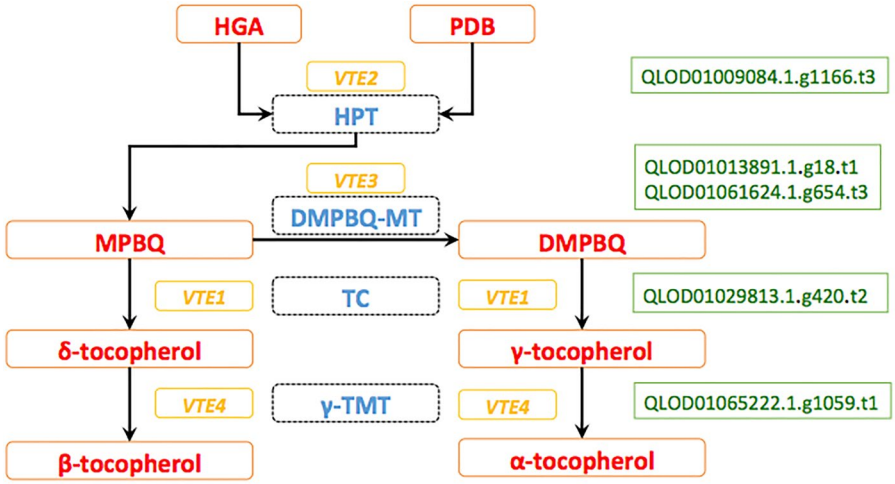

Fig. 3 The tocopherol biosynthetic pathway. Enzymes are shown in blue and metabolites in red. The initial prenylation reaction of HGA and PDP is catalyzed by HPT enzyme to form MPBQ. MPBQ is then methylated by MPBQ methyltransferase to form DMPBQ, which can be cyclized by TC and further methylated by γ-TMT to form α-Tocopherol (Savidge et al. 2002). *Arabidopsis thaliana* genes, encoding the enzymes, are emphasized in yellow. The key biosynthetic genes identified in *Argania spinosa* are highlighted in green. PDP: phytyldiphosphate; HGA: homogentisic acid; MPBQ: 2-methyl-6-phytylbenzoquinol; DMPBQ: 2,3-dimethyl-6-phytyl-1,4-benzoquinol; HPT: homogentisate phytyltransferase; MPBQ MT: MPBQ methyltransferase; TC: tocopherol cyclase; γ-TMT: γ-tocopherol methyltransferase

antioxidants that prevent the oxidation of unsaturated fatty acids. They come in four isoforms of differing biological activity with α-Tocopherol (α-Toc) showing the strongest antioxidant properties. The key genes implicated in the downstream stages of the biosynthesis pathway are encoding for homogentisate phytyltransferase (HPT, encoded by *VTE2* gene), 2-methyl-6-phenyl benzoquinone methyltransferase (MPBQ MT, encoded by *VTE3* gene), tocopherol cyclase (TC, encoded by *VTE1* gene), and γ-tocopherol-methyltransferase (γ-TMT, encoded by *VTE4* gene) (Fig. 3).

We identified four key genes involved in tocopherol biosynthesis in argan genome encoding for tocopherol cyclase (TC, encoded by *VTE1* gene), homogentisate phytyltransferase (HPT, encoded by *VTE2*), 2-methyl-6-phytylbenzoquinol methyltransferase (MPBQ MT, encoded by *VTE3*), and γ-tocopherol methyltransferase (TMT, encoded by *VTE4*). Their evolutionary histories were analyzed separately by phylogeny reconstructions (Fig. 4).

The Maximum likelihood (ML) phylogeny of *VTE1*, encoding for TC enzyme, shows the *A. trichopoda VTE1* sequence as sister group to all other *VTE1*-like proteins. Monocot sequences form a clade with 100% bootstrap support and all dicot sequences form a group with 86% support. Except for *G. max* with two *VTE1* genes and *T. aestivum* with three, all species included in the analysis have only a single *VTE1* gene. Ericales genes including the argan sequence form a subclade with 95% bootstrap support, and argan codes for a single *VTE1* ortholog.

The ML phylogeny of *VTE2*, encoding for HPT enzymes, supports the monocot representatives with 100% bootstrap value and the dicot sequences with 83% bootstrap value. Several independent gene duplications occurred within the *VTE2*-like genes, resulting three *VTE2* homologs in *T. aestivum*, two in *G. max*, two in *H. annuus* and two in *A. chinensis*, and a single one in argan.

The *VTE3*-like genes cluster in a monocot clade supported by 100% bootstrap support and a dicot clade with 79% bootstrap support. Interestingly, a dicot specific gene duplication leading to two monocot subclades, one supported by 95%, the other one by 98% bootstrap support. Thus, most species analyzed encode at least two homologs of *VTE3*, except for *Z. mays*, with only a single gene. Independently, gene duplication occurred within the dicots, leading a more complex situation. The lineage including the *A. thaliana VTE3* gene is supported by a bootstrap value

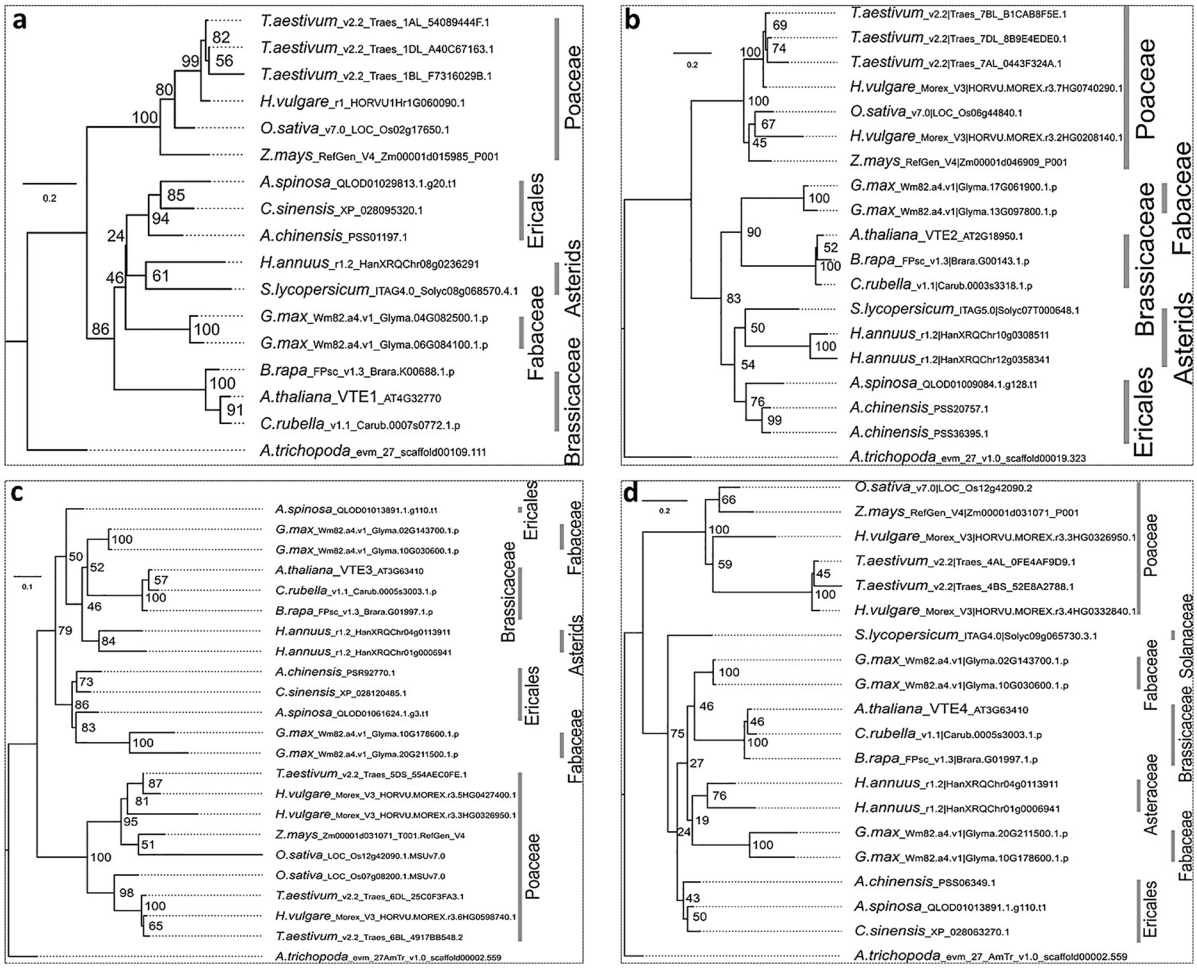

Fig. 4 Maximum likelihood phylogeny of *VTE* 1 (**a**), *VTE* 2 (**b**), *VTE* 3 (**c**), *VTE* 4 (**d**) -like genes including angiosperm representative with sequenced genomes. Numbers next to branches denote boostrap support of the respective lineage split

of 50% and includes sequences from argan, *B. napus*, *A. thaliana* and other Brassicaceae and *H. annuus*. The other subfamily is supported by 86% bootstrap support and includes sequences from all analyzed Ericales and *B. napus*. The argan genome encodes two *VTE3* homologs.

The *VTE4*, encoding for TMT enzyme, protein family shows support of 100% for the monocot sequences, and 75% for the dicot members. Within the dicots, the best-supported subclade includes Brassicaceae members with a support of 100%, all other subclades show support below 50%, suggesting an extensive duplication history and rapid sequence evolution, unlike observed in the other *VTE* gene families. A single *VTE4* copy was identified in argan.

Discussion

Owing to the lack of annotated genomic data on *A. spinosa*, little is known about this economically important plant. In this work, we annotated a draft assembly using a model-based prediction approach to generate a data resource, available in a public database to advance research on the argan tree. In this work, the draft genome assembly generated by Khayi et al. (2018) was used for gene prediction and functional annotation. Ultimately, the draft genome was predicted to encode for 42,095 genes. The number of genes predicted for *A. spinosa* is high compared with the number of genes predicted for the closely related and annotated species *Primula veris*

(19,507 predicted genes; Nowak et al. 2015), *Primula vulgaris* (24,599 predicted genes; Cocker et al. 2018), but similar to *A. chinensis* (kiwi, 39,040 genes; Huang et al. 2013), and is higher than the genes number covered in the well-annotated species *A. thaliana* (27,029 genes; Swarbreck et al. 2007). This suggests extensive genome evolution in the Ericales.

The number of repeated genes in the genome of argan (56.33% of the assembly) appears to be higher than that reported in the related species kiwi with 36% of the assembly (Huang et al. 2013) which mainly made up by retrotransposons including the long terminal repeat (LTR) family as the most abundant. For comparison, the *A. thaliana* genome includes only 14% repetitive sequences suggesting that the many genes predicted from the argan genome may be transposons or retrotransposons.

To the best of our knowledge, this is the sixth whole genome in the order Ericales to be sequenced and annotated, after *A. chinensis*, *P. vulgaris*, *P. veris, C. sinensis, R. williamsianum*, and *V. corymbosum*. The reports presented in this study, represents the first whole genome-annotation of species member of under-explored Sapotaceae family including other economically important species such as *Vitellaria paradoxa* (shea tree) and *Chrysophyllum cainito* (star apple or cainito) (Vallée et al. 2016).

In this study, a total of 45 putative sequences that can be phylogenetically clustered were identified in the argan genome, including the membrane-bound FA desaturases (FADs). FADs play primordial role in FA composition regulation in plants, which sprightly affects oil quality and is involved in tolerance to abiotic stresses. Therefore, FADs should be added to the list of the main candidate genes for crop improvement (Xu et al. 2019). Moreover, one of the main factors underlying the thriving demand for argan oil is its high total tocopherol content compared with other oil seed crops. Beyond its antioxidant activity and implications for human health (Stampfer et al. 1993; Brigelius-Flohe and Traber 1999), analyses in *A. thaliana* indicate, that genes involved in tocopherol biosynthesis are important for biotic/abiotic stress response (Falk and Munné-Bosch 2010). An example is homogentisate phytyltransferase (*HPT*), which has an impact on seed longevity and adjustment to low-temperature degrees (Hunter and Cahoon 2007; Maeda et al. 2008; Ren et al. 2011; Ji et al. 2016). Their characterization in highly drought resistant species such as argan may pave the way to improve crop plant responses to abiotic/biotic stresses (Munné-Bosch 2005; Ji et al. 2016). However, more effort has yet to be deployed to investigate their functions in biochemical assays.

The genetic diversity of the tocopherol pathway genes, and the intra- and inter-species variations in the α-Toc content, in particular between different wild argan genotypes, is also unknown. Based on the sequence homology, the enzymes TC (*VTE1*), and TMT (*VTE4*) are the product of single copy genes in argan, kiwi and tea, while two genes in argan genome encode the enzyme MPBQ-MT (*VTE3*), but only single copy genes in tea and kiwi (Fig. 4). This finding contradicts previous reports claiming the tocopherol biosynthetic pathway enzymes are represented with single-copy genes only (DellaPenna 2005).

Interestingly, the gene encoding the HPT (*VTE2*) is represented by two genes in kiwi, a single gene in argan, while being deficient in tea genome. This might be the reason for its richness in the new tocopherol analogues, such as δ-tocomonoenol, in kiwi fruit (Fiorentino et al. 2009).

The various gene duplication scenarios revealed in oilseed plants (Table 1), in particular wheat and soybean, could be the explaining factor leading to the reported variations in their tocopherol contents. The final step of the pathway, which consists in the conversion of the large γ-Toc pool to α-Toc, is variable and not always effective in all plants. This could be the limiting step leading to intra- and inter-species variations in particularly α-Toc content (Grusak and DellaPenna 1999; Kumari et al. 2019).

The difference in gene copy numbers is revealed as well among the species for the *VTE4* encoding the γ-TMT enzyme, which is catalyzing this critical step of the pathway. These inter-species variations could explain the variable conversion efficiencies to α-Toc, which is not always effective in plants. A single copy in argan genome represents the *VTE4* while gene duplications are found *H. annuus* and *B. napus.* The latter two contain predominantly α-Toc, with more than 95% and 56.4% of the total tocopherol amount, respectively (Cao et al. 2015), while γ-Toc represents the major part for argan seeds with 85% (Marfil et al. 2011).

Conclusion

In summary, we present the reference transcriptome annotated argan genome, together with an analysis of homologs of the important biosynthesis genes for oil production and tocopherol biosynthesis genes, in selected species, including oil crops and with a special focus on argan. The annotated genome generated in this work could be used as a reference for comprehensive genome-assisted cultivar breeding and genome editing improvements and to revisit any trait in order to elucidate the molecular mechanisms that trigger the relevant metabolic pathways, including the genes involved in cosmetic and pharmacological products. Such analyses could contribute to the improvement of argan oil traits and facilitate the development of strategies to breed and domesticate the argan tree through the manipulation of the identified genes. The genes identified that are associated with pathways of the main oil components could be used as candidate genes for improvement of both the seed content and quality and to improve plant responses to abiotic/biotic stresses to help them cope with climate change.

Acknowledgements This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI). We thank Thorsten Wachtmeister (Düsseldorf) for continuous discussion on sequencing strategies and and the DFG Research Infrastructure West German Genome Center (407493903) as part of the Next Generation Sequencing Competence Network (project 423957469) for the sequencing support. Computational support of the Zentrum für Informations- und Medientechnologie, especially the HPC team (High Performance Computing) at the Heinrich-Heine University is acknowledged.

Author contributions CR carried out the phylogeny reconstructions and orthology group searches with the help of NLP. OR did all the bioinformatics analysis related to argan genome annotation and gene family analysis. TEW optimized the library preparation and sequencing methodology. ML and AB planned the experiments, wrote and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding Open Access funding enabled and organized by Projekt DEAL. M.L. received funding from the Alexander von Humboldt Foundation, Germany, for this work.

Data availability The genome annotation results of this study can accessed at the argan genome database, https://gendbe.computational.bio.uni-giessen.de/Argania_spinosa.html. The nucleotide (*A. thaliana*) and protein sequences (all other species) used for this work were obtained from the following databases: *Arabidposis thaliana*: https://www.arabidopsis.org; INRA (https://www.heliagene.org/); Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html); National Center for Biotechnology Information (NCBI): https://www.ncbi.nlm.nih.gov; Maize: https://www.maizegdb.org/.

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

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/.

References

Ain-Lhout F, Boutaleb S, Diaz-Barradas MC, Jauregui J, Zunzunegui M (2016) Monitoring the evolution of soil moisture in root zone system of *Argania spinosa* using electrical resistivity imaging. Agric Water Manag. https://doi.org/10.1016/j.agwat.2015.08.007

Ait Aabd N, Msanda F, Abdelhamid EM (2014) Evaluation of variability in argan oil content through different environments and preselection of elite genotypes. Euphytica. https://doi.org/10.1007/s10681-013-0969-0

Ait Aabd N, Bouharroud R, Tahiri A, Wifaya A, Mimouni A, El Mousadik A (2019) Genetic diversity and breeding of argan tree (*Argania spinosa* L. Skeels). In: Al-Khayri J, Jain S, Johnson D (eds) Advances in plant breeding strategies, nut and beverage crops. Springer, Cham, pp 31–35. https://doi.org/10.1007/978-3-030-23112-5_2

Alados CL, El Aich A (2008) Stress assessment of argan (*Argania spinosa* (L.) Skeels) in response to land uses across an aridity gradient: translational asymmetry and branch fractal dimension. J Arid Environ. https://doi.org/10.1016/j.jaridenv.2007.06.015

Alba Sánchez F, López Sáez JA, Nieto Lugilde D, Svenning JC (2015) Long-term climate forcings to assess vulnerability in North Africa dry argan woodlands. Appl Veg Sci. https://doi.org/10.1111/avsc.12133

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol. https://doi.org/10.1016/S0022-2836(05)80360-2

Brigelius-Flohe R, Traber MG (1999) Vitamin E: function and metabolism. FASEB J. https://doi.org/10.1096/fasebj.13.10.1145

Cabrera-Vique C, Marfil R, Giménez R, Martínez-Augustin O (2012) Bioactive compounds and nutritional significance of virgin argan oil-an edible oil with potential as a functional food. Nutr Rev. https://doi.org/10.1111/j.1753-4887.2012.00478.x

Cao J, Li H, Xia X, Zou XG, Li J, Zhu XM, Deng ZY (2015) Effect of fatty acid and tocopherol on oxidative stability of vegetable oils with limited air. Int J Food Prop. https://doi.org/10.1080/10942912.2013.864674

Chakhchar A, Wahbi S, Lamaoui M, Ferradous A, El Mousadik A, Ibnsouda-Koraichi S, Filali-Maltouf A, El Modafar C (2015) Physiological and biochemical traits of drought tolerance in *Argania spinosa*. J Plant Interact. https://doi.org/10.1080/17429145.2015.1068386

Charrouf Z, Harhar H, Gharby S, Guillaume D (2008) Enhancing the value of argan oil is the best mean to sustain the argan grove economy and biodiversity, so far. Ol Corps Gras Lipides. https://doi.org/10.1051/ocl.2008.0205

Cocker JM, Wright J, Li J, Swarbreck D, Dyer S, Caccamo M, Gilmartin PM (2018) *Primula vulgaris* (primrose) genome assembly, annotation and gene expression, with comparative genomics on the heterostyly supergene. Sci Rep. https://doi.org/10.1038/s41598-018-36304-4

DellaPenna D (2005) Progress in the dissection and manipulation of vitamin E synthesis. Trends Plant Sci. https://doi.org/10.1016/j.tplants.2005.10.007

Díaz-Barradas MC, Zunzunegui M, Ain-Lhout F, Jáuregui J, Boutaleb S, Álvarez-Cansino L, Esquivias MP (2010) Seasonal physiological responses of *Argania spinosa* tree from Mediterranean to semi-arid climate. Plant Soil. https://doi.org/10.1007/s11104-010-0518-8

El Boukhari A, Tabi S, El Mousadik A, El Boullani R, Tahiri A, Koufan M, Benyahia H, Bouharroud R, Ait Aabd N (2023) Flow cytometry and chromosome numbers variation in argan tree *Argania spinosa* (L.) Skeels. Not Sci Biol. https://doi.org/10.55779/nsb15111451

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. https://doi.org/10.1186/s13059-015-0721-2

Falk J, Munné-Bosch S (2010) Tocochromanol functions in plants: antioxidation and beyond. J Exp Bot. https://doi.org/10.1093/jxb/erq030

Fiorentino A, Mastellone C, D'Abrosca B, Pacifico S, Scognamiglio M, Cefarelli G et al (2009) δ-Tocomonoenol: a new vitamin E from kiwi (*Actinidia chinensis*) fruits. Food Chem. https://doi.org/10.1016/j.foodchem.2008.11.094

Gonzálvez A, Armenta S, De la Guardia M (2010) Adulteration detection of argan oil by inductively coupled plasma optical emission spectrometry. Food Chem. https://doi.org/10.1016/j.foodchem.2009.11.091

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. https://doi.org/10.1093/nar/gkr944

Grusak MA, DellaPenna D (1999) Improving the nutrient composition of plants to enhance human nutrition and health. Annu Rev Plant Physiol Plant Mol Biol. https://doi.org/10.1146/annurev.arplant.50.1.133

Guillaume D, Charrouf Z (2011) Argan oil and other argan products: use in dermocosmetology. Eur J Lipid Sci Technol. https://doi.org/10.1002/ejlt.201000417

Hoff KJ, Stanke M (2019) Predicting genes in single genomes with AUGUSTUS. Curr Protoc Bioinform. https://doi.org/10.1002/cpbi.57

Huang S, Ding J, Deng D, Tang W, Sun H, Liu D et al (2013) Draft genome of the kiwifruit *Actinidia chinensis*. Nat Commun. https://doi.org/10.1038/ncomms3640

Hunter SC, Cahoon EB (2007) Enhancing vitamin E in oilseeds: unraveling tocopherol and tocotrienol biosynthesis. Lipids. https://doi.org/10.1007/s11745-007-3028-6

Ji CY, Kim YH, Kim HS, Ke Q, Kim GW, Park SC et al (2016) Molecular characterization of tocopherol biosynthetic genes in sweetpotato that respond to stress and activate the tocopherol production in tobacco. Plant Physiol Bioch. https://doi.org/10.1016/j.plaphy.2016.04.037

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. https://doi.org/10.1093/molbev/mst010

Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F (2016) Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw092

Khayi S, Azza NE, Gaboun F, Pirro S, Badad O, Claros MG et al (2018) First draft genome assembly of the Argane tree (*Argania spinosa*). F1000Research. https://doi.org/10.12688/f1000research.15719.2

Kumari K, Rai MP, Bansal N, Prashat GR, Kumari S, Krishnan V et al (2019) Analysis of γ-tocopherol-methyltransferase3 promoter activity and study of methylation patterns of the promoter and its gene body. Plant Physiol Biochem. https://doi.org/10.1016/j.plaphy.2019.09.044

Le Polain Y, Waroux Y (2013) Dégradation environnementale et développement économique dans l'arganeraie d'Aoulouz (Maroc). Sécheresse. https://doi.org/10.1684/sec.2013.0373

Lex A, Nils G, Hendrik S, Romain V, Hanspeter P (2014) UpSet: visualization of intersecting sets. IEEE Trans vis Comput Graph. https://doi.org/10.1109/TVCG.2014.2346248,2014

Lu K, Wei L, Li X et al (2019) Whole-genome resequencing reveals *Brassica napus* origin and genetic loci involved in its improvement. Nat Commun. https://doi.org/10.1038/s41467-019-09134-9

Lybbert T, Barrett C, Narjisse H (2002) Market-based conservation and local benefits: the case of argan oil in Morocco. Ecol Econ. https://doi.org/10.1016/S0921-8009(02)00020-4

Lybbert TJ, Aboudrare A, Chaloud D, Magnan N, Nash M (2011) Booming markets for Moroccan argan oil appear to benefit some rural households while threatening the endemic argan forest. Proc Natl Acad Sci. https://doi.org/10.1073/pnas.1106382108

Maeda H, Sage TL, Welti R, DellaPenna D (2008) Tocopherols modulate extraplastidic polyunsaturated fatty acid

metabolism in *Arabidopsis* at low temperature. Plant Cell. https://doi.org/10.1105/tpc.107.054718

Majourhat K, Jabbar Y, Araneda L, Zeinalabedini M, Hafidi A, Martínez-Gómez P (2007) Karyotype characterization of *Argania spinosa* (L.) Skeel (Sapotaceae). S Afr J Bot. https://doi.org/10.1016/j.sajb.2007.06.007

Marfil R, Giménez R, Martínez O, Bouzas PR, Rufián Henares JA, Mesías M, Cabrera Vique C (2011) Determination of polyphenols, tocopherols, and antioxidant capacity in virgin argan oil (*Argania spinose*, Skeels). Eur J Lipid Sci Technol. https://doi.org/10.1002/ejlt.201000503

Mateille T, Tavoillot J, Martiny B, Dmowska E, Winiszewska G, Ferji Z, Msanda F, El Mousadik A (2016) Aridity or low temperatures: What affects the diversity of plant-parasitic nematode communities in the Moroccan argan relic forest? Appl Soil Ecol. https://doi.org/10.1016/j.apsoil.2015.11.026

Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Pühler A (2003) GenDBE an open source genome annotation system for prokaryote genomes. Nucleic Acids Res. https://doi.org/10.1093/nar/gkg312

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015

Munné-Bosch S (2005) The role of α-tocopherol in plant stress tolerance. J Plant Physiol. https://doi.org/10.1016/j.jplph.2005.04.022

Nowak MD, Russo G, Schlapbach R, Huu CN, Lenhard M, Conti E (2015) The draft genome of *Primula veris* yields insights into the molecular basis of heterostyly. Genome Biol. https://doi.org/10.1186/s13059-014-0567-z

Ren W, Zhao L, Zhang L, Wang Y, Cui L, Tang Y, Sun X, Tang K (2011) Molecular analysis of a homogentisate phytyltransferase gene from *Lactuca sativa* L. Mol Biol Rep. https://doi.org/10.1007/s11033-010-0297-6

Savidge B, Weiss JD, Wong YH, Lassner MW, Mitsky TA, Shewmaker CK, Post-Beittenmiller D, Valentin HE (2002) Isolation and characterization of homogentisate phytyltransferase genes from *Synechocystis* sp. PCC 6803 and *Arabidopsis*. Plant Physiol. https://doi.org/10.1104/pp.010747

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC et al (2022) Database resources of the national center for biotechnology information. Nucleic Acids Res. https://doi.org/10.1093/nar/gkab1112

Seppey M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M (ed) Gene prediction. Methods in molecular biology. Humana, New York, pp 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14

Sharma S, Kaushik S, Raina S (2018) Estimation of nuclear DNA content and its variation among Indian tea accessions by flow cytometry. Physiol Mol Biol Plants. https://doi.org/10.1007/s12298-018-0587-3

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv351

Smit AFA, Hubley R (2015) RepeatModeler open-1.0. 2008–2015. http://www.repeatmasker.org

Smit AFA, Hubley R, Green P (2015) RepeatMasker open-4.0. 2013–2015, pp 289–300

Stampfer MJ, Hennekens CH, Manson JE, Colditz GA, Rosner B, Willett WC (1993) Vitamin E consumption and the risk of coronary disease in women. N Engl J Med. https://doi.org/10.1056/NEJM199305203282003

Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. https://doi.org/10.1093/nar/gkh379

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. https://doi.org/10.1093/nar/gkl200

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H et al (2007) The Arabidopsis information resource (TAIR): gene structure and function annotation. Nucleic Acids Res. https://doi.org/10.1093/nar/gkm965

Vallée GC, Muñoz DS, Sankoff D (2016) Economic importance, taxonomic representation and scientific priority as drivers of genome sequencing projects. BMC Genom. https://doi.org/10.1186/s12864-016-3100-9

Wu H, Ma T, Kang M, Ai F, Zhang J, Dong G, Liu J (2019) A high-quality *Actinidia chinensis* (kiwifruit) genome. Hortic Res. https://doi.org/10.1038/s41438-019-0202-y

Xu L, Zeng W, Li J, Liu H, Yan G, Si P et al (2019) Characteristics of membrane-bound fatty acid desaturase (FAD) genes in *Brassica napus* L. and their expressions under different cadmium and salinity stresses. Environ Exp Bot. https://doi.org/10.1016/j.envexpbot.2019.02.016

Zhao X, Dupont L, Cheddadi R, Kölling M, Reddad H, Groeneveld J et al (2019) Recent climatic and anthropogenic impacts on endemic species in southwestern Morocco. Quat Sci Rev. https://doi.org/10.1016/j.quascirev.2019.105889

Zunzunegui M, Ain-Lhout F, Jáuregui J, Barradas MD, Boutaleb S, Álvarez-Cansino L, Esquivias MP (2010) Fruit production under different environmental and management conditions of argan, *Argania spinosa* (L.). J Arid Environ. https://doi.org/10.1016/j.jaridenv.2010.03.016

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