Check for updates

RESEARCH ARTICLE

First draft genome assembly of the Argane tree (*Argania spinosa*) [version 1; referees: 1 approved, 1 approved with reservations]

Slimane Khayi [1*], Nour Elhouda Azza[2,3*], Fatima Gaboun[1], Stacy Pirro[4], Oussama Badad[3,5,6], M. Gonzalo Claros [7], David A. Lightfoot[5], Turgay Unver[8], Bouchra Chaouni[3,6], Redouane Merrouch[9], Bouchra Rahim[9], Soumaya Essayeh[10], Matika Ganoudi[1], Rabha Abdelwahd[1], Ghizlane Diria[1], Meriem Alaoui Mdarhi[1], Mustapha Labhilili[1], Driss Iraqi[1], Jamila Mouhaddab[2], Hayat Sedrati[11], Majid Memari[12], Noureddine Hamamouch[13], Juan de Dios Alché [14], Noureddine Boukhatem[3], Rachid Mrabet[1], Rachid Dahan[1], Adelkhaleq Legssyer[3], Mohamed Khalfaoui[9], Mohamed Badraoui[1], Yves Van de Peer[15-17], Tatiana Tatusova[18], Abdelhamid El Mousadik[2], Rachid Mentag[1*], Hassan Ghazal [2-4,9*]

[1]Biotechnology Unit, National Institute of Agricultural Research (INRA), Rabat, Morocco, Morocco
[2]Laboratory of Biotechnology and Valorization of Natural Resources (LBVRN), Faculty of Sciences, University Ibn Zohr, Agadir, Morocco
[3]Laboratory of Physiology, Genetics & Ethnopharmacology (LPGE), Faculty of Sciences, University Mohamed Premier, Oujda, Morocco
[4]Iridian Genomes, Inc., Bethesda, MD, 20817, USA
[5]Department of Plant, Soil and Agricultural Systems, Southern Illinois University, Carbondale, IL, 62901, USA
[6]Laboratory of Plant Physiology, Faculty of Sciences, University Mohamed V in Rabat, Rabat, 10000, Morocco
[7]Department of Molecular Biology and Biochemistry, and Plataforma Andaluza de Bioinformática, University of Malaga, Malaga, Spain
[8]International Biomedicine and Genome Institute (iBG-izmir), Dokuz Eylul University, Current address: Egitim Mah. Ekrem Guer Sok. 26/3 Balcova, Izmir, Turkey
[9]National Center for Scientific and Technological Research (CNRST), Rabat, Morocco
[10]Polydisciplinary Faculty of Nador, University Mohamed Premier, Nador, Morocco
[11]National School of Computer Sciences & Systems Analysis, University Mohammed V in Rabat, Rabat, Morocco
[12]Research Computing and Cyber infrastructure, Computer Science Department, Southern Illinois University, Carbondale, IL, 62901, USA
[13]Polydisciplinary Faculty, Sultan Moulay Slimane University, Beni-Mellal, Morocco
[14]Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas (CSIC), Granada, Spain
[15]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, B-9052 Ghent, Belgium, Belgium
[16]VIB Center for Plant Systems Biology, Technologiepark 927, Ghent, B-9052, Belgium
[17]Department of Genetics, Genomics Research Institute, University of Pretoria, Pretoria, 0028, South Africa
[18]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, 20817, USA

* Equal contributors

Open Peer Review

Referee Status: ？ ✔

Abstract

Background: The Argane tree (*Argania spinosa* L. Skeels) is an endemic tree of southwestern Morocco that plays an important socioeconomic and ecologic role for a dense human population in an arid zone. Several studies confirmed the importance of this species as a food and feed source and as a resource for both pharmaceutical and cosmetic compounds. Unfortunately, the argane tree ecosystem is facing significant threats from environmental changes (global warming, over-population) and over-exploitation. Limited research has been conducted, however, on argane tree genetics and genomics, which hinders its conservation and genetic improvement.
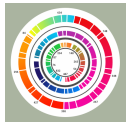
Methods: Here, we present a draft genome assembly of *A. spinosa*. A reliable reference genome of *A. spinosa* was created using a hybrid *de novo* assembly approach combining short and long sequencing reads.

Results: In total, 144 Gb Illumina HiSeq reads and 7.2 Gb PacBio reads were produced and assembled. The final draft genome comprises 75 327 scaffolds totaling 671 Mb with an N50 of 49 916 kb. The draft assembly is close to the genome size estimated by *k*-mers distribution and covers 89% of complete and 4.3 % of partial *Arabidopsis* orthologous groups in BUSCO.

Conclusion: The *A. spinosa* genome will be useful for assessing biodiversity leading to efficient conservation of this endangered endemic tree. Furthermore, the genome may enable genome-assisted cultivar breeding, and provide a better understanding of important metabolic pathways and their underlying genes for both cosmetic and pharmacological purposes.

Invited Referees

1 **2**

version 1
published
17 Aug 2018

report report

1 **Amit Sinha** iD , New England Biolabs Inc., USA

2 **Granger Sutton** iD , J. Craig Venter Institute, USA

Discuss this article

Comments (0)

This article is included in the Draft Genomes collection.

Introduction

Argania spinosa (L. Skeels) is a tree endemic to the South West of Morocco and occupying arid and semi-arid regions totaling up to around 900,000 ha[1]. The argane tree forest was recognized as biosphere reserve (Arganeraie Biosphere Reserve) by UNESCO in 1998[2]. It is the only unique member of the tropical Sapotaceae family in Morocco[3]. In addition to its ecological role in preventing soil erosion and desertification, the argane tree has great cultural and socio-economic importance. The oil extracted from the seed is considered the most expensive edible oil in the world with great cosmetic value and therapeutic potential[4–6]. Argane oil represents a significant source of dietary fatty acids, while the Argane fruit is used as livestock feed by the local population[7–9]. Phytochemical composition of Argane fruits reveals different classes of bioactive compounds, including essential oils, fatty acids, triacylglycerols, flavonoids and their acylglycosyl derivatives, monophenols, phenolic acids, cinnamic acids, saponins, triterpenes, phytosterols, ubiquinone, melatonin, new aminophenols, and vitamin E. Argane oil contains high levels of antioxidant compounds. The long-chain fatty acids in Argane oil are primarily represented by unsaturated oleic acid, then linoleic acid, palmitic acid and stearic acid[10].

The distribution area of the Argane forest decreased drastically during the 18th century. Furthermore, about 44 % of the forest was again lost between 1970 and 2007. While there are multiple causes, desertification and overgrazing form the main pressures on the Argane forest[11,12]. Therefore, the management and conservation of the remaining genetic resources of Argane forest are urgent priorities. In recent decades, several studies have been conducted to evaluate the genetic diversity of the Argane tree using morphological[13,14], chemical[10,15], biochemical[6,16] and standard molecular marker techniques, all with the aim of describing the genetic diversity of Argane trees and addressing ecological and conservation issues[17–25]. The karyotype of *A. spinosa* (L.) is constituted of ten pairs of chromosomes (2n =2x =20)[3]. Until now, no reference genome was available of the *A. spinosa* species. Here, we present the Argane tree genome assembled from short and long DNA reads using a hybrid assembly strategy.

Methods and results

Plant material

The Argane tree (*Argania spinosa*, taxid 85883, Sapotaceae, family, order Ericales), named *Argane AMGHAR*, to be sequenced was selected for its biological and ecological characteristics (Figure 1). This was a 9-year-old shrub, with weeping form (geotropic, unlike the erect have) with only one main trunk 3 m in height. The ripe fruits has a rounded shape. The plant had semi-evergreen dwarf leaves. The shrub is native to the valley of the plain of Sous, an arid climate with an annual average rainfall of around 220 mm, located between the hills of the Anti-Atlas towards the South East, the Western High Atlas towards the North-West and the Atlantic Ocean towards the West (9°32′ 00″N, 30°24′ 00″W; Altitude: 126 m).

DNA sequencing and data description

Genomic DNA was extracted from lyophilized leaf tissues of a single tree (*Argane AMGHAR*) using the Plant DNeasy mini kit (Qiagen, USA). The Argane tree genome was shotgun-sequenced using both Pacbio™ (Menlo Park, CA, USA) and Illumina™ (San Diego, CA, USA) sequencing technologies, generating 7.2 Gb and 144 Gb of data, respectively. A gel-based size selection of DNA was performed for fragments ≥ 2 kb. Paired-end libraries with average insert sizes of 600 bp were constructed with Nextera™ DNA Library Prep Kit for Illumina (New England Biolabs™, New Brunswick, MA, USA). These libraries were sequenced on an Illumina HiSeq XTen platform using the PE-150 module and yielded 957,451,810 reads (Table 1). These data was trimmed of adapters and low-quality sequences, yielding a clean set of 936,053,040 reads, representing 160× genome coverage, assuming a genome size of 573 Mb as estimated by the *k-mer* frequency analysis (described below). Raw reads were deposited at the NCBI Sequence Read Archive (SRA) under accession numbers: SRX3207155 and SRX3207156, corresponding to two independent runs from the same plant DNA sample. In addition, single-molecule long reads from the PacBio RS II platform (Pacific Biosciences, USA) were used to assist the subsequent *de novo* genome assembly using Illumina. Genomic sequencing libraries were constructed using the PacBio DNA template preparation kit 2.0 (Pacific Biosciences of California, Inc., Menlo Park, CA) for SMRT sequencing on the PacBio RS II machine (Pacific Biosciences of California, Inc.) according to the manufacturer's instructions, with a size range of 2-15 kb. The constructed libraries were sequenced on six SMRT cells on a PacBio RSII sequencer. The sequences of the 6 SMRT cell runs were deposited at the NCBI SRA under accession numbers: SRX1898029/SRX1898030/SRX1898031/SRX1898032/SRX1898033/SRX1898034. The sequencing runs produced about 7.2 Gb, consisting of 6,705,437 reads with an average read length of 2.5 kb and representing about 12× genome coverage, again assuming a genome size of 573 Mb (Table 1).

In silico genome size estimation

Quality-filtered reads from the Illumina platform were subjected to *k-mer* frequency distribution analysis with JELLYFISH v2.1.4 software[26,27]. Analysis parameters were set at -*k* 21 and 25, and the final result was plotted as a frequency graph (Figure 2). Two distinctive modes were observed from the distribution curve: the higher peak at a depth of 44 and reflecting the high heterozygosity of the Argane genome; the lower peak provided a peak depth of 87 for the estimation of the genome size[28]. Based on the total number of k-mers obtained, the Argane genome size was calculated to be approximately 573 Mb and 615 Mb, for 21- and 25-mers respectively, using the following formula: total number of *k*-mer / Peak depth. The double peak of *k*-mer distribution indicates heterozygosity whose rate is estimated to be 1.58 % (Figure 2). The estimated genome size seems to be credible compared to the ones of four other Sapotaceae family members. In fact, according to the Plant DNA c-values Database, the genome sizes of these four species ranged from 273 Mb in *Mimusops elengi* L. (c = 0.28 pg) to 2,513 Mb in *Isonandra villosa* L. (c = 2.57 pg). The other two species are *Planchonella eerwah* (c = 0.54 pg, 528 Mb) and *Madhuca longifolia* (c = 0.99 pg, 968 Mb).

Genome assembly and evaluation

Prior to assembly, Illumina and PacBio raw reads were trimmed for quality and adaptor removal using *bbduk.sh* from BBmap suite (https://github.com/BioInfoTools/BBMap). Short and long
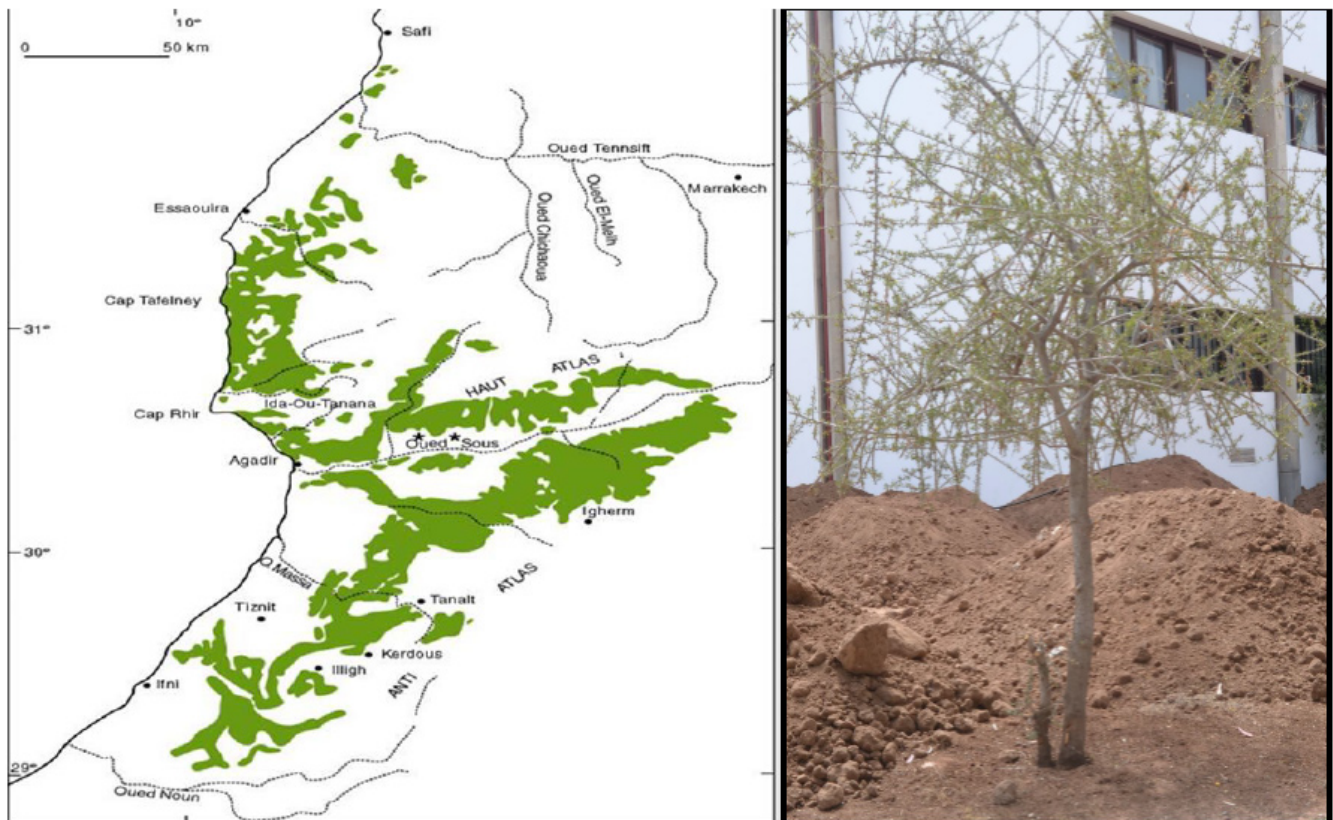
Figure 1. Information on the *Argania.spinosa* individual named *#Argane Amghar* whose DNA has been sequenced. (**A**) Map of Sous region with *Argania* species distribution. (**B**) Picture of the tree # *Argane Amghar*. (Photograph taken by A. El Mousadik).

Table 1. Summary of reads generated from genome sequencing and used in the assembly.

Library technology	Raw data		Trimmed data	
	Number of reads	Number of bases	Number of reads	Number of bases
Illumina HiSeq X Ten	957,451,810	144,575,223,310	936,053,040	135,539,587,270
PacBio RS II	6,705,437	7,649,825,228	1,910,887	7,078,584,472

reads were assembled following a hybrid approach using MaSuRCA assembler v.3.2.2[29]. The initial assembly consists of 671,690,540 bp composed of 82,183 contigs with the largest size being 422,848 bp and an N50 of 43,654 bp. The very few contigs (8) with length less than 200 bp were filtered out and the remaining contigs were scaffolded into 75,327 scaffolds totaling 670,096,797 bp; the N50 reached 49,916 bp and the assembly accounted for 2,982,868 Ns with 445.14 Ns per 100 kb (Table 2). The scaffolding was done using initial contigs and implemented in MaSuRCA v3.2.4 assembler script using Celera Assembler v8.3. The GC content was estimated to be 33%. The assembly was screened by VecScreen to look for and remove remaining vector contamination. Based on the VecSreen

report, contigs containing mitochondrial/chloroplast were also removed. Trimmed PE reads were mapped on the final assembly using CLC genomics (v11.0, CLCbio, Arhus, Denmark) with 0.8 in length and 0.9 in sequence similarity. In total, 94% of the reads were mapped against the Argane genome. The 6% reads that were unmapped may result from the stringency of mapping criteria used.

The difference between the genome size estimation and assembly size may be due to the use of parameters excluding extremely high frequency *k*-mers. They often represent organelle sequences, eventual contaminants inflating the genome size estimation[30], or the high-frequency of repetitive regions found in

Figure 2. Distribution of 21 and 25-mers using Jellyfish with PE data Argane whole genome sequencing.

Table 2. Statistics of the *Argania* spinosa genome assembly.
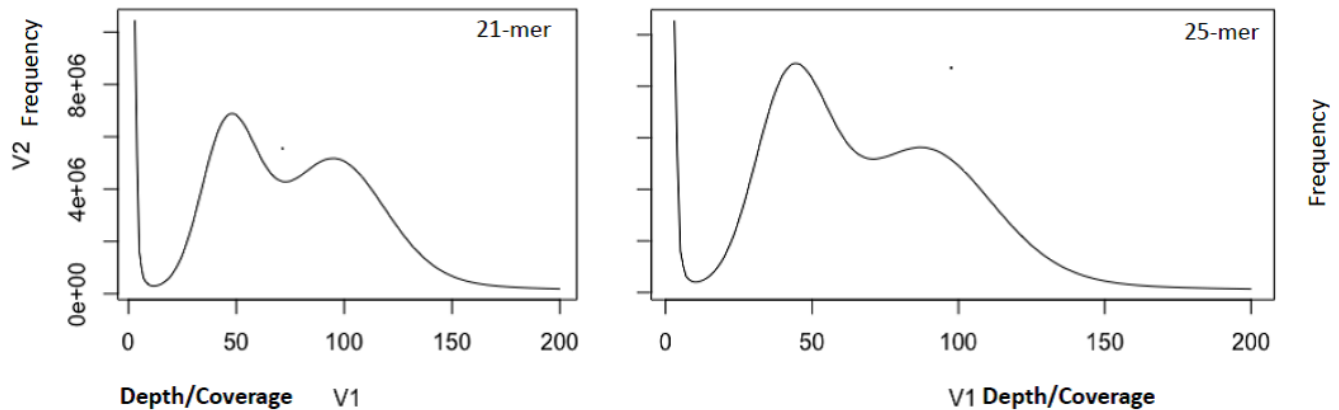
	Number	Total size (bp)	N50 (bp)	Largest (bp)
Contigs	82,183	671,690,540	43,654	422,848
Scaffolds ≥ 200 bp	75,327	670,096,797	49,916	422,848

N50 size defined as the value N such that at least 50% of the genome is covered by scaffolds of size N or larger.

plant genomes. Furthermore, the genome is highly heterozygous and different allelic regions would inflate assembly size. To assess the completeness of the final assembly, a Benchmarking Universal Single-Copy Orthologs (BUSCO) v3 software approach was used with Arabidopsis lineage-specific orthologous groups[31], which showed that the assembly contained 89% (1271 genes) of complete and 4.3% (62 genes) of partial sequences that were Arabidopsis orthologs.

Conclusions

This draft genome assembly is a first step towards a global and integrative omics strategy for exhaustive characterization of the Argane tree. In particular, future work will focus on structurally annotating the genome using predictive tools and transcriptome analysis. Other future work will focus on functional gene annotation, finding evidence for genome duplication and comparative genome evolution. A reliable annotation is highly dependent on transcriptomic research, and sequencing of Argane transcriptome analysis of different parts and developmental stages or the plant is ongoing. The metabolome, and analysis of Argane oil biosynthesis, as well as the tree's microbiome should also be analyzed. To this end, and in order to coordinate the strong interests of the Plant Genomics community for this precious tree, the International Argane Genome Consortium (IAGC) and a resource website has been created (www.arganome.org).

Data availability

All of the *A. spinosa* datasets can be retrieved under BioProject accession number PRJNA294096: http://identifiers.org/bioproject:PRJNA294096. The raw reads are available at NCBI Sequence Reads Archive under accession number SRP077839: http://identifiers.org/insdc.sra:SRP077839. The complete genome sequence assembly project has been deposited at GenBank under accession number QLOD00000000: http://identifiers.org/ncbigi/GI:1408199612. Data can also be retrieved via the International Argane Genome Consortium (IAGC) website: http://www.arganome.org.

Author information

Slimane Khayi and Nour Elhouda Azza are co-first authors; Rachid Mentag and Hassan Ghazal contributed equally as supervisors.

Competing interests

No competing interests were disclosed.

References

1. Lefhaili A: **FAO Forest Resources Assessment: Morocco Country Report.** Rome. FAO; 2010.
    Reference Source

2. **UNESCO - MAB Biosphere Reserves Directory.** 2018.
    Reference Source

3. Majourhat K, Jabbar Y, Araneda L, *et al.*: **Karyotype characterization of *Argania spinosa* (L.) Skeel (Sapotaceae).** *South Afr J Bot.* 2007; **73**(4): 661–3.
    Publisher Full Text

4. Monfalouti HE, Guillaume D, Denhez C, *et al.*: **Therapeutic potential of argan oil: a review.** *J Pharm Pharmacol.* 2010; **62**(12): 1669–1675.
    PubMed Abstract | **Publisher Full Text**

5. Khallouki F, Spiegelhalder B, Bartsch H, *et al.*: **Secondary metabolites of the argan tree (Morocco) may have disease prevention properties.** *Afr J Biotechnol.* 2005; **4**(5): 381–388.
    Reference Source

6. Khallouki F, Younos C, Soulimani R, *et al.*: **Consumption of argan oil (Morocco) with its unique profile of fatty acids, tocopherols, squalene, sterols and phenolic compounds should confer valuable cancer chemopreventive effects.** *Eur J Cancer Prev.* 2003; **12**(1): 67–75.
    PubMed Abstract | **Publisher Full Text**

7. Charrouf Z, Guillaume D: **Argan oil: Occurrence, composition and impact on human health.** *Eur J Lipid Sci Technol.* 2008; **110**(7): 632–6.
    Publisher Full Text

8. Lybbert TJ, Aboudrare A, Chaloud D, *et al.*: **Booming markets for Moroccan argan oil appear to benefit some rural households while threatening the endemic argan forest.** *Proc Natl Acad Sci.* 2011; **108**(34): 13963–8.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

9. Alba-Sánchez F, López-Sáez JA, Nieto-Lugilde D, *et al.*: **Long-term climate forcings to assess vulnerability in North Africa dry argan woodlands.** Rocchini D, editor. *Appl Veg Sci.* 2015; **18**: 283–96.
    Publisher Full Text

10. Khallouki F, Voggel J, Breuer A, *et al.*: **Comparison of the major polyphenols in mature Argan fruits from two regions of Morocco.** *Food Chem.* 2017; **221**: 1034–1040.
    PubMed Abstract | **Publisher Full Text**

11. le Polain de Waroux Y, Lambin EF: **Monitoring degradation in arid and semi-arid forests and woodlands: The case of the argan woodlands (Morocco).** *Appl Geogr.* 2012; **32**(2): 777–86.
    Publisher Full Text

12. McGregor HV, Dupont L, Stuut JBW, *et al.*: **Vegetation change, goats, and religion: a 2000-year history of land use in southern Morocco.** *Quat Sci Rev.* 2009; **28**(15–16): 1434–48.
    Publisher Full Text

13. Ait Aabd N, El Ayadi F, Ms F: **Evaluation of agromorphological variability of argan tree under different environmental conditions in Morocco: implication for selection.** *Int J Biodivers Conserv.* 2011; **3**(3): 73–82.
    Reference Source

14. Bani-Aameur F, Ferradous A: **Fruits and stone variability in three argan (*Argania spinosa* (L.) Skeels) populations.** *Int J For Genet.* 2001; **8**: 39–45.
    Reference Source

15. Khallouki F, Haubner R, Ricarte I, *et al.*: **Identification of polyphenolic compounds in the flesh of Argan (Morocco) fruits.** *Food Chem.* 2015; **179**: 191–198.
    PubMed Abstract | **Publisher Full Text**

16. Klika KD, Khallouki F, Owen RW: **Amino phenolics from the fruit of the argan tree *Argania spinosa* (Skeels L.).** *Z Naturforsch C.* 2014; **69**(9–10): 363–367.
    PubMed Abstract | **Publisher Full Text**

17. Pakhrou O, Medraoui L, Yatrib C, *et al.*: **Assessment of genetic diversity and population structure of an endemic Moroccan tree (*Argania spinosa* L.) based in IRAP and ISSR markers and implications for conservation.** *Physiol Mol Biol Plants.* 2017; **23**(3): 651–61.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

18. El Bahloul Y, Dauchot N, Machtoun I, *et al.*: **Development and characterization of microsatellite loci for the Moroccan endemic endangered species *Argania spinosa* (Sapotaceae).** *Appl Plant Sci.* 2014; **2**(4): pii: apps.1300071.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

19. Chakhchar A, Haworth M, El Modafar C, *et al.*: **An Assessment of Genetic Diversity and Drought Tolerance in Argan Tree (*Argania spinosa*) Populations: Potential for the Development of Improved Drought Tolerance.** *Front Plant Sci.* 2017; **8**: 276.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

20. Majourhat K, Jabbar Y, Hafidi A, *et al.*: **Molecular characterization and genetic relationships among most common identified morphotypes of critically endangered rare Moroccan species *Argania spinosa* (Sapotaceae) using RAPD and SSR markers.** *Ann For Sci.* 2008; **65**(8): 805–805.
    Publisher Full Text

21. El Mousadik A, Petit RJ: **Chloroplast DNA phylogeography of the argan tree of Morocco.** *Mol Ecol.* 1996a; **5**(4): 547–555.
    PubMed Abstract | **Publisher Full Text**

22. El Mousadik A, Petit RJ: **High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco.** *Theor Appl Genet.* 1996b; **92**(7): 832–839.
    PubMed Abstract | **Publisher Full Text**

23. Mouhaddab J, Ait Aabd N, Achtak H, *et al.*: **A Patterns of Genetic Diversity and Structure at Fine Scale of an Endangered Moroccan Endemic Tree (*Argania spinosa* L. Skeels) Based on ISSR Polymorphism.** *Notulae Botanicae Horti Agrobotanici Cluj- Napoca.* 2015; **43**(2): 528–535.
    Reference Source

24. Mouhaddab J, Ait Aabd N, Msanda F, *et al.*: **Assessing genetic diversity and constructing a core collection of an endangered Moroccan endemic tree [*Argania spinosa* (L.) Skeels].** *Moroccan J Biol.* 2017; **13**: 1–12.
    Reference Source

25. Mouhaddaba J, Msandaa F, Filali-Maltoufb A, *et al.*: **Using microsatellite markers to map genetic diversity and population structure of an endangered Moroccan endemic tree (*Argania spinosa* L. Skeels) and development of a core collection.** *Plant Gene.* 2017; **10**: 51–59.
    Publisher Full Text

26. Liu B, Shi Y, Yuan J, *et al.*: **Estimation of genomic characteristics by analyzing *k*-mer frequency in *de novo* genome projects.** *ArXiv13082012 Q-Bio.* 2013.
    Reference Source

27. Marçais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers.** *Bioinformatics.* 2011; **27**(6): 764–70.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

28. Kajitani R, Toshimoto K, Noguchi H, *et al.*: **Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads.** *Genome Res.* 2014; **24**(8): 1384–95.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

29. Zimin AV, Marçais G, Puiu D, *et al.*: **The MaSuRCA genome assembler.** *Bioinformatics.* 2013; **29**(21): 2669–77.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

30. Vurture GW, Sedlazeck FJ, Nattestad M, *et al.*: **GenomeScope: fast reference-free genome profiling from short reads.** *Bioinformatics.* 2017; **33**(14): 2202–4.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

31. Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–2.
    PubMed Abstract | **Publisher Full Text**

Open Peer Review

Current Referee Status: ❓ ✔️

Version 1

Referee Report 18 September 2018

✔️ **Granger Sutton** 🆔

J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA

In the article "First draft genome assembly of the Argane tree (*Argania spinosa*)", the authors present the draft assembly of the Argane tree using a combination of Illumina and PacBio reads. An argument for the importance of the Argane tree to the agriculture and economy of the endemic region is given. The assertion that "the genome may enable genome-assisted cultivar breeding, and provide a better understanding of important metabolic pathways and their underlying genes for both cosmetic and pharmacological purposes" seems reasonable based on the use of previous genome sequences for important agricultural resources. This article does not claim much beyond providing a reasonable draft genome for *A. spinosa*. The primary evidence for reasonable quality is three-fold: a k-mer analysis of genome size using Jellyfish, 94% of Illumina reads mapping back to the assembly, and a BUSCO analysis showing 89% full length and 4.3% partial length gene matches for these single copy genes. This is probably sufficient to claim reasonable quality and the usefulness of the genome for downstream research given the contiguity shown by the N50 contig size.

The authors do miss an opportunity to look more closely at the haplotype separation that may be occurring as they speculate possibly leading to a somewhat larger genome size than indicated by Jellyfish. I'm not sure why the number of BUSCO genes with more than one full or partial length match was not given – perhaps there were none? The implication that heterozygosity is uniform "The double peak of *k*-mer distribution indicates heterozygosity whose rate is estimated to be 1.58%" seems unfounded without more evidence nor is it clear that is the model resulting in this estimate.

Lastly the grammar would benefit from editing. For example, "A reliable annotation is highly dependent on transcriptomic research, and sequencing of Argane transcriptome analysis of different parts and developmental stages or the plant is ongoing." Might be better as "A reliable annotation is highly dependent on transcriptomic data, analysis and research. Sequencing of different parts and developmental stages of the Argane transcriptome is ongoing.".

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Whole genome shotgun assembly, pan-genome analysis

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

? Amit Sinha (iD)
New England Biolabs Inc., Ipswich, MA, USA

In the current manuscript, the authors present the first draft genome assembly for the argan tree *Argania spinosa*, a tree with significant ecological as well as commercial importance.
The authors have used a hybrid assembly approach, combing data from both Illumina short-reads and PacBio long-reads technology. Use of such hybrid approaches is welcome, as they can combine the advantages offered by different technological platforms.
Based on a k-mer analysis, the authors also provide a reasonable estimate of the genome-size and the level of heterozygosity. These are important metrics to report, as they can guide the methods to be used for genome assembly, as well as for interpretation of the assembled genome sequence. The draft genome has been satisfactorily assembled, using state of the art computational methods, and it contains 89% of the predicted core genes as defined in the BUSCO software.

It is hoped that in their future studies, the authors will follow up with more improvements to the draft genome, including gene annotations using transcriptomic data.

Overall, manuscript provides valuable information and sequence resources that will be useful to the scientific community. I have only 1 major and 6 minor comments/suggestions that I think the authors should be able to address.

Major reservation:

1. Both Illumina as well as PacBio reads have been processed and trimmed before being used as an input to the MaSuRCA software. However, the instructions for the MaSuRCA software explicitly ask for **NOT** using any third-party tools, or performing any trimming, cleaning or error correction on Illumina reads, as it will most likely lead to a deteriorated assembly (See https://github.com/alekseyzimin/masurca#overview).

It is therefore strongly suggested that the authors try an assembly with the raw Illumina reads directly, as it might lead to a better assembly. If such an analysis was already performed but the assembly was worse than the current version, please provide this information in the manuscript.

Minor comment/suggestion:

1. For PacBio data, please specify the form of processed reads that were input to the MaSuRCA software: were they the raw PacBio reads, or the Circular Consensus Sequence (CCS) reads?

2. The authors report that the genomic DNA was size-selected for fragments above 2kb. What was the approximate maximum size of input DNA for PacBio library preparation. Also, what was the maximum size of PacBio reads (e.g after the CCS step) ?

3. The raw coverage obtained from Illumina data is reported as 160X. But as per my calculations, based on number of trimmed bases in table 1 and a estimated genome size of 573 Mb, I obtain 135,539,587,270 / 573,000,000 = 236X. Please correct or explain further.

4. In addition to the observed heterozygosity, the repeat content of the genome could be an important reason for a fragmented assembly. The k-mer analysis can also be used to estimate the % of genome that is comprised of repeats. It will be great if this estimate is provided.

5. For the BUSCO analysis, the authors report using the Arabidopsis lineage-specific orthologous groups. However, from the BUSCO manual, the only valid value for the lineage parameter seems to be "embryophyta_odb9", which internally uses Arabidopsis hmms for Augustus-based gene predictions, but the member proteins are form multiple species. If this is the lineage parameter the authors used, please specifiy it as "embryophyta_odb9" rather than Arabidopsis, or provide more details on how the Arabidopsis BUSCO groups were determined.

6. The genome sequences of the organelles mitochondria and chloroplast are also very important in investigating the biology of any species, and also serve as important resource for evolutionary and phylogenetic studies. Therefore, it will be great if instead of completely removing the organellar genomes, the authors provide them as separate set of sequences, in separate files and ideally as separate GenBank entries.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias

- You can publish traditional articles, null/negative results, case reports, data notes and more

- The peer review process is transparent and collaborative

- Your article is indexed in PubMed after passing peer review

- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com