plants

MDPI

*Article*

# Complete Chloroplast Genome of *Argania spinosa*: Structural Organization and Phylogenetic Relationships in Sapotaceae

**Slimane Khayi [1,*], Fatima Gaboun [1], Stacy Pirro [2], Tatiana Tatusova [3], Abdelhamid El Mousadik [4], Hassan Ghazal [5] and Rachid Mentag [1,*]**

[1] CRRA-Rabat, National Institute for Agricultural Research (INRA), Rabat 10101, Morocco; fatima.gaboun@inra.ma

[2] Iridian Genomes, Inc., Bethesda, MD 20817, USA; info@iridiangenomes.org

[3] National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20817, USA; tatiana@ncbi.nlm.nih.gov

[4] Laboratory of Biotechnology and Valorization of Natural Resources (LBVRN), Faculty of Sciences, University Ibn Zohr, Agadir 80000, Morocco; a.elmousadik@uiz.ac.ma

[5] National Center for Scientific and Technological Research (CNRST), Rabat 10102, Morocco; hassan.ghazal@fulbrightmail.org

* Correspondence: rachid.mentag@inra.ma (R.M.); slimane.khayi@inra.ma (S.K.);
Tel.: +212-661-558-622 (R.M.); +212-664-141-380 (S.K.)

Received: 18 May 2020; Accepted: 25 June 2020; Published: 13 October 2020

**Abstract:** *Argania spinosa* (Sapotaceae), an important endemic Moroccan oil tree, is a primary source of argan oil, which has numerous dietary and medicinal proprieties. The plant species occupies the mid-western part of Morocco and provides great environmental and socioeconomic benefits. The complete chloroplast (cp) genome of *A. spinosa* was sequenced, assembled, and analyzed in comparison with those of two Sapotaceae members. The *A. spinosa* cp genome is 158,848 bp long, with an average GC content of 36.8%. The cp genome exhibits a typical quadripartite and circular structure consisting of a pair of inverted regions (IR) of 25,945 bp in length separating small single-copy (SSC) and large single-copy (LSC) regions of 18,591 and 88,367 bp, respectively. The annotation of *A. spinosa* cp genome predicted 130 genes, including 85 protein-coding genes (CDS), 8 ribosomal RNA (rRNA) genes, and 37 transfer RNA (tRNA) genes. A total of 44 long repeats and 88 simple sequence repeats (SSR) divided into mononucleotides (76), dinucleotides (7), trinucleotides (3), tetranucleotides (1), and hexanucleotides (1) were identified in the *A. spinosa* cp genome. Phylogenetic analyses using the maximum likelihood (ML) method were performed based on 69 protein-coding genes from 11 species of *Ericales*. The results confirmed the close position of *A. spinosa* to the *Sideroxylon* genus, supporting the revisiting of its taxonomic status. The complete chloroplast genome sequence will be valuable for further studies on the conservation and breeding of this medicinally and culinary important species and also contribute to clarifying the phylogenetic position of the species within Sapotaceae.

**Keywords:** chloroplast genome; sapotaceae; *Argania spinosa*; phylogenomic analysis; molecular marker

## 1. Introduction

The argan tree (*Argania spinosa L. Skeels*) is an endemic plant species of the middle west of Morocco and the unique member of the tropical Sapotaceae family in this Mediterranean country [1]. In 1999, UNESCO classified the argan tree as a world heritage. Extracted from seeds, argan oil is the worldwide precious product of the argan tree, used as edible or cosmetic oil [2]. Thus, this forest fruit and forage species is the backbone of a traditional arganian system that has hitherto served the needs of a dense population in an arid zone. Unfortunately, due to the pressure of several factors such as human overexploitation and climate change, the Argan forest was drastically deteriorated during the 18th century, and about 44% of the forest was lost between 1970 and 2007 [3,4]. Therefore, the management and conservation of the remaining genetic resources of this species are urgent priorities.

The Sapotaceae family is composed of about 50 genera and 1100 species which are distributed in the tropical regions with some exception, especially, *A. spinosa*, which occupies the mid-western part of Morocco [5]. Previous biogeography and phylogenetic analyses of Sapotaceae species based on different elements, such as several chloroplast genes [6–8], the chloroplast *ndhF* gene combined with morphological data [9], and nuclear ITS (Internal Transcribed Spacer) combined to the chloroplast *trnH–psbA* regions, have been reported [10]. These phylogenetic studies, inferred using few genetic makers, have strongly recommended that the satellite genus of *Argania* must be included into the genus of *Sideroxylon*, hence revisiting the phylogenetic status of *A. spinosa*. Thus, more genetic markers are needed to clarify this phylogenetic revision which still remains debatable.

During the last two decades, the advent of next-generation sequencing (NGS) technologies has accelerated the pace of deciphering the chloroplast (cp) genomes of many plant species; presently, 3949 land plant cp genomes have been deposited in GenBank Organelle Genome Resources (accessed on 25 June 2020). This photosynthetic organelle provides essential energy for plants and algae and represents a valuable resource for exploring intra- and inter-specific evolutionary histories of land plants [11–15]. In addition, due to their several characteristics such as small length, simple structure, maternal inheritance characters, conserved sequences, and very low level of recombination [16], the chloroplast genomes are commonly used for studies of plant evolution, phylogeny, and traceability [17,18]. To date, two full-length chloroplast sequences have been assembled within Sapotaceae, i.e., those of *Sideroxylon wightianum* (MG719834) and *Pouteria campechiana* (MH018545.1). A comprehensive phylogenetic analysis based on whole cp genomes should help to more accurately elucidate the phylogenetic status of *A. spinosa*. A first draft nuclear genome of *A. spinosa* has recently been published [19] and deposited in GenBank (QLOD00000000.1), but little is known about its cp genome structure, except from the work of El Mousadik and Petit [20] who studied the phylogeography of argan tree using universal chloroplast primers targeting specific chloroplast genes.

In the present study, we report the complete cp genome sequence of *A. spinosa*, assembled from Illumina short reads for the first time and compare the genomes of three Sapotaceae species to understand the variations among their cp genomes. The objectives of this study were: (i) assemble and depict the whole cp genome structure of *A. spinosa*, (ii) perform extensive comparative genomics with other Sapotaceae cp genomes, (iii) report the simple sequence repeats in cp genomes to provide tools for future genetic diversity and breeding studies, and lastly iv) assess the taxonomic positions of *A. spinosa* based on its complete cp genome.

## 2. Results and Discussion

### 2.1. Genome Assembly

Illumina sequencing of two libraries generated a total of 957,451,810 (SRR6062046, SRR6062045) raw reads. After a quality processing step, 936,053,040 reads were mapped against the cp reference genome to collect the *A. spinosa* cp-like reads. Paired-end reads were then extracted from the mapping, yielding 7,9724,230 sequences representing 8.5% of the total whole genome shotgun data. De novo assembly with CLC genomics (v11.0, CLCbio, Arhus, Denmark) generated 23 contigs with the maximum length of 39 kbp. The alignment of the contigs against the cp reference genome resulted in seven contigs, totaling 158,281 bp. The aligned contigs were assembled into one chromosome by the Genome Finishing Module (GFM) from CLC genomics using *S. wightianum* cp genome as a reference.

The *A. spinosa* genome consists of a circular molecule measuring 158,848 bp in length, with 36.9% GC content, which is consistent with other sequenced cp genomes of the Sapotaceae family, whose plastome

GC content was 36.8% and 38.9% for *P. campechiana* and *S. wightianum*, respectively. The whole genome alignment to the cp reference and the dot plot of the genome sequence confirmed the quadripartite structure found in most chloroplast genomes of plants [12,14,15,21,22] (Figure 1). The genome has an inverted repeat (IR) region 25,945 bp in length, a large single-copy (LSC) region of 88,367 bp, and a small single-copy (SSC) region of 18,591 bp. The GC content was 42% in the IR region and 34% and 30% in the LSC and SSC regions, respectively, at relatively the same level as in *S. wightianum* (IR 42%, LSC 34%, SSC 30%) and *P. campechiana* (IR: 42%, LSC: 34%, SCC: 30%). The high GC content registered in the IR regions is mainly due to the high GC contents of the four ribosomal RNA (rRNA) genes *rrn4.5*, *rrn5*, *rrn16*, *rrn23* that are located in the IR regions and display, respectively, 50%, 52%, 56%, and 54% of GC content.

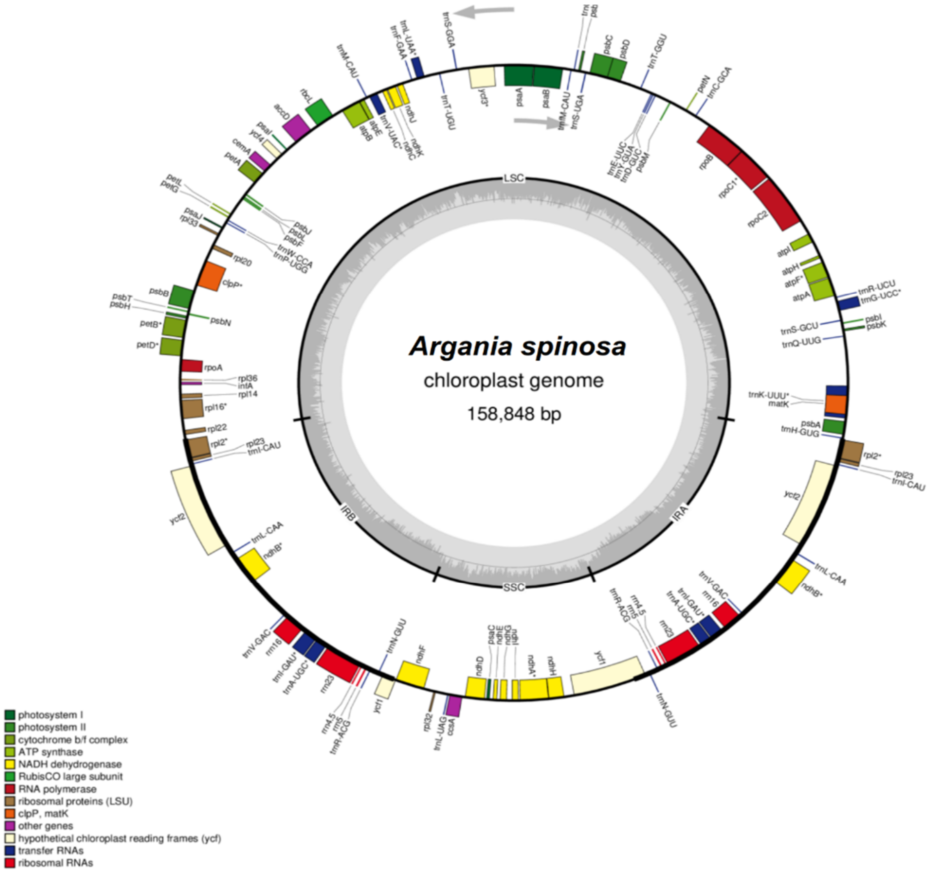

**Figure 1.** Gene map of *Argania spinosa* chloroplast genome. The thick lines colored in black in the inner circle indicate the extent of the inverted repeat regions (IRa and IRb; 25,946 bp), which separate the genome into small (SSC; 18,593 bp) and large (LSC; 88,367 bp) single-copy regions. Genes drawn inside the circle are transcribed clockwise, and those on the outer side are transcribed counter-clockwise. Genes belonging to different functional groups are color-coded. The dark grey in the inner circle corresponds to the GC content, and the light grey corresponds to the AT content.

The validation of the assembly was performed by PCR and Sanger sequencing using four couples of primers designed on the boundaries of the IR and single-copy regions (Table S1). PCRs were performed using the DNA extracts of four different individuals of *A. spinosa* (V1, V2, V3, and V4). The A and D couple of primers targeted the IRa/LSC and IRa/SSC junctions and amplified, respectively, 700 and 258 bp fragments, as shown in Figure S1. The B and C primers were designed to verify the IRb/SSC and IRb/LSC junctions, respectively, amplifying fragments of 699 and 361 bp in length (Figure S1).

### 2.2. Features of the *A. spinosa* Chloroplast Genome

The annotation process predicted a total of 130 functional genes representing 85 protein-coding genes, 37 tRNAs and 8 rRNAs. The coding domain sequences (CDSs) account for 80,967 bp in length, which represents 50.97% of *A. spinosa* cp genome. The gene proportion for tRNA is 1.75%, and that for *rRNA* is 5.69%. The proportion of non-coding regions, which contain intergenic spacers and introns, represents 49.02% of the cp genome. The protein-coding sequences include 6 duplicated genes (*rpl2*, *nhB*, *rpl23*, *rps7*, *ycf2*, and *rps12*), 1 pseudogene (Ψ*ycf1)*, 4 rRNAs in two copies, and 37 tRNAs with 7 duplicated genes (*trnA-UGC*, *trnI-CAU*, *trnI-GAU*, *trnL-CAA*, *trnN-GUU*, *trnV-GAC*, and *trnR-ACG*). The IR regions contains six CDS (*rpl2, rpl23*, *ycf2*, *ndhB*, *rps7*, *rps12*), four rRNAs and seven tRNAs. The SSC region contains 12 CDS and 1 tRNA, while the LSC region harbors 60 CDS and 22 tRNAs (Table 1, Figure 1).

**Table 1.** Features of *A. spinosa* chloroplast genome. NADH (Nicotinamide Adenine Dinucleotide Hydrogen), ORF (Open Reading Frame).

| Category | Group of Genes | Name of Genes |
|---|---|---|
| Self-replication | Ribosomal RNAs | *rrn16*, *rrn* 23, *rrn4.5*, *rrn5* |
| | Transfer RNAs | *trnA-UGC*, *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, *trnfM-CAU*, *trnG-UCC*, *trnH-GUG*, *trnI-CAU*, *trnI-GAU*, *trnK-UUU*, *trnL-CAA*, *trnL-UAA*, *trnL-UAG*, *trnM-CAU*, *trnN-GUU*, *trnS-GCU*, *trnP-UGG*, *trnQ-UUG*, *trnR-ACG*, *trnR-UCU trnW-CCA*, *trnY-GUA*, *trnS-GGA*, *trnS-UGA*, *trnT-GGU*, *trnT-UGU*, *trnV-GAC*, *trnV-UAC*, *trnG-GCC* |
| | Small subunit of ribosome | *rps2*, *rps3*, *rps4*, *rps7*, *rps8*, *rps11*, *rps12*, *rps14*, *rps15*, *rps16*, *rps18*, *rps19* |
| | Large subunit of ribosome | *rpl14*, *rpl16*, *rpl2*, *rpl20*, *rpl22*, *rpl23*, *rpl32*, *rpl33*, *rpl36* |
| | Translational initiation factor | *infA* |
| | DNA-dependent RNA polymerase | *rpoA*, *rpoB*, *rpoC1*, *rpoC2* |
| Genes for photosynthesis | NADH dehydrogenase | *ndhA*, *ndhB*, *ndhC* |
| | | *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK* |
| | PS1 | *psaA*, *psaB*, *psaC*, *psaI*, *psaJ* |
| | PS2 | *psaA*, *psaB*, *psaC*, *psaI*, *psaJ*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ* |
| | Cytochromeb/f complex | *petA*, *petB*, *petD*, *petG*, *petL*, *petN* |
| | ATP synthase | *atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI* |
| | rubisco large | *rbcL* |
| Other genes | maturase | *matK* |
| | protease | clpP |
| | Envelope membrane protein | cemA |
| | Subunit acetyl-co carboxylase | accD |
| | c-type cytochrome synthesis gene | *ccsA* |
| Genes with unknown function | ORF ycf | *Ycf1*, *ycf2*, *ycf3*, *ycf4* |

The *A. spinosa* cp genome was found to contain introns in some annotated genes, like other cp genomes of angiosperms [23,24]. A total of nine protein-coding genes (*rpl2, ndhB, rps12, ndhA, rps16*, *petD*, *petB*, *rpoC1*, *atpF*) and six tRNA genes contained a single intron, while three genes (*ycf*3, *clpP¡*, and *rps12*) contained two introns. The *rps*12 gene was predicted to be trans-spliced, with the 5′ end located in the LSC region and the duplicated 3′ end in the IR region. The *trnK-UUU* gene has the longest intron (2535 bp) that contains coding sequences of the *matK* gene, whereas the intron of *trnL-UAA* is the smallest (509 bp). The complete cp genome was deposited in Organelle Genome Resources (GenBank accession MK533159).

The sequences of tRNA and protein-coding genes were analyzed, and the codon-usage frequency was calculated for *A. spinosa* (Table S2). In total, 26,799 codons were identified for 83 protein-coding sequences in *A. spinosa* cp genome. The use of the codons ATG and TGG, which encode, respectively, Methionine and Tryptophan, exhibited no bias (Relative Synonymous Codon Usage, RSCU = 1). The maximum AUU (1078) and the minimum CGC (104) codons used coded for isoleucine and arginine, respectively, and the ending bases A and U were preferred in the synonymous codon (RSCU > 1). However, for the non-preferred synonymous codons, the ending bases were G or C. The same phenomenon was described in previous studies of cp genomes [14,25]. The frequencies of amino acids for the protein-coding sequence were calculated for *A. spinosa* (Figure 2). Leucine represents the most frequent amino acid in the *A. spinosa* cp genome, with 2586 codons (10.3%). Cysteine is the least frequent amino acid, with only 450 codons (1.3%). Similar ratios for amino acids were reported in previous studies [26,27].

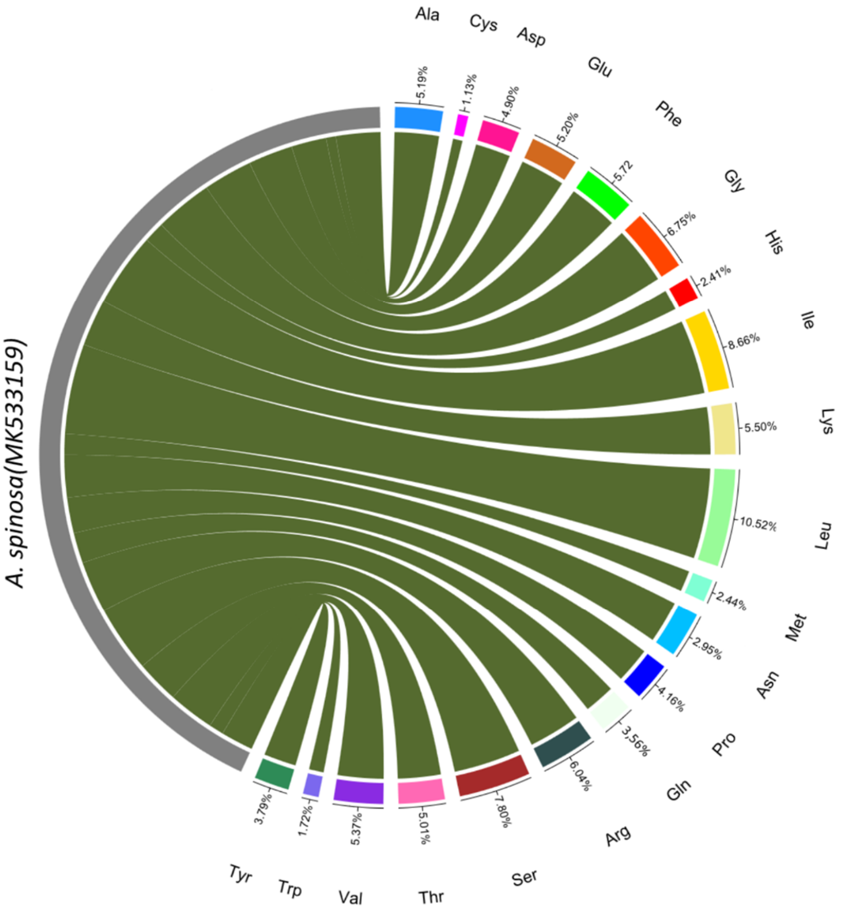

**Figure 2.** Amino acid frequencies of the *A. spinosa* chloroplast protein-coding sequences. The frequencies of amino acids were calculated for all 85 protein-coding genes from the start to the stop codons.

### 2.3. Comparative Analysis of cp Genome Structures

To understand the structural characteristics of the cp genomes of Sapotaceae sequenced to date, overall sequence alignment of the three cp genome sequences was conducted using the annotation of *A. spinosa* as a reference (Figure 3). The aligned sequences appeared to be relatively conserved, with a slight level of sequence divergence in some regions. The gene-coding regions are more highly conserved than those of their non-coding counterparts and intergenic regions, which is consistent with the pattern reported for several angiosperm cp genomes [22]. The most divergent sequences were found within the intergenic spacers and introns of these three sequences, including *rbcL–psaI*, *psaI–petA*, *psbE–petL*, *ycf2–trnL-CAA*, *rpl32–ndhF*, *ycf1–rpS15*, *psbA–trnH-GUG*, *trnQ-UUG–rpS16*, *psbI–psbK*, *psbM–petN*. In the coding regions, slight variations in sequences were observed in *petD*, *ndhA*, and *ycf2*. Most of these hotspot regions are located in the LSC regions, and only few regions are located in the SSC or IR regions, as shown in Figure 3.

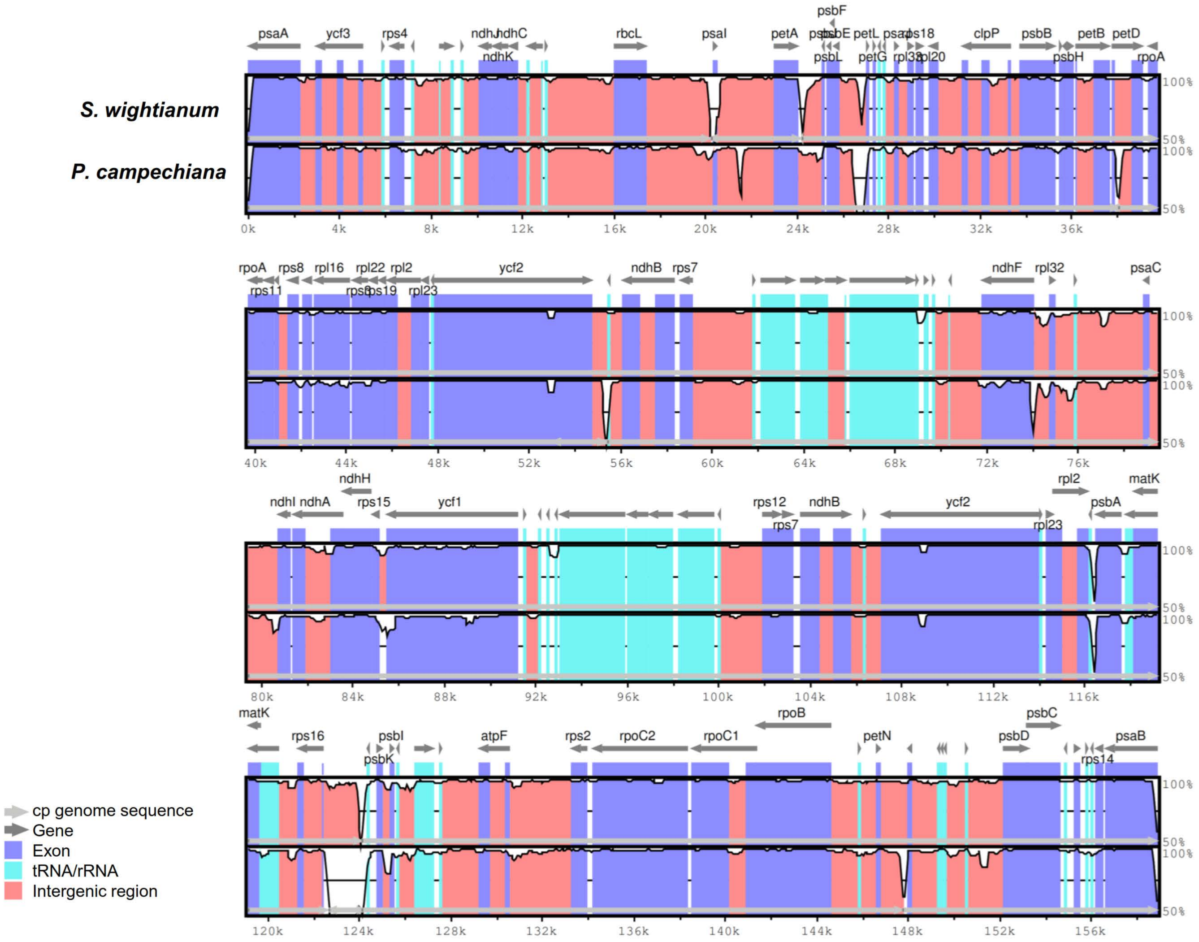

**Figure 3.** mVISTA percent identity plot comparing the three Sapotaceae plastid genomes with *A. spinosa* as a reference. The top line shows genes in order (transcriptional direction indicated by arrows). The y-axis represents the percent identity within 50–100%. The x-axis represents the coordinate in the chloroplast genome. Genome regions are color-coded as protein-coding (exon), tRNAs or rRNAs, and conserved noncoding sequences (intergenic region). The white block represents regions with sequence variation between the two species.

Inverted repeat regions are variable in land plants' cp genomes, ranging from a couple of hundred [28] to several thousand [13] base pairs in size. It is reported that large IR play a key role in genome stability of chloroplasts [29]. A detailed comparison of the four junctions LSC–IRa (JLA), LSC–IRb (JLB), SSC–IRa (JSA), and SSC–IRb (JSB) of three *Sapotaceae* cp genomes (*A. spinosa*, *S. wightianum*, and *P. campechiana*) is presented in Figure 4. Although the IR regions of the three cp genomes were highly conserved with slight variations. We found five genes, *rps19*, *ycf1*, *ndhF*, *rpl2*, and *psbA*, among the three species to be implicated in the four junctions. The *rps19* gene is located in the LSC region at 20, 4, and 6 bp from the JLB border, and the gene *psbA* is located in the same region at 340, 610, and 649 bp from JLA in these three *Sapotaceae* cp genomes, respectively. The *rpl2* gene is located in the IR region at 68, 65, and 67 bp from the JLA/JLB junctions, respectively, in the three cp genomes. The gene *ndhF*, located in the SSC, was found to end exactly at the JSB junction in *S. wightianum* genome and to cross the junction by 2 bp in *P. campechiana* genome. However, the distance between this junction and *ndhF* was of 143 bp for *A. spinosa*. The pseudogene *ycf1*, that was not predicted in *S. wightianum*, was found to cross the JSB junction by 143 bp in *A. spinosa* genome and to end at the limit of this junction in *P. campechiana* genome. The JSA junction was crossed by the gene *ycf1* in the three cp genomes, and the fragment located in the IRa region ranged from 928 to 1079 bp. These results showed the slight contraction of the IR region in *A. spinosa* cp genome.

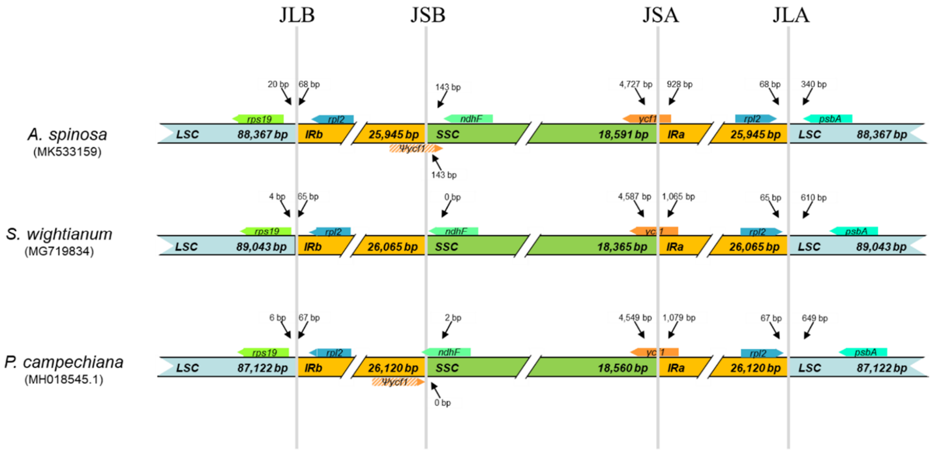

**Figure 4.** A comparison of the distance between adjacent genes and junctions in the SSC, LSC, and two IR regions for the chloroplast genomes of *A. spinosa*, *Sideroxylon wightianum*, and *Pouteria campechiana*. The figure shows relative changes at or near the IR/SC borders, with no scale to sequence length. The colored boxes indicate genes.

### 2.4. Nucleotide Diversity and Divergence of Coding Gene Sequences

Nucleotide diversity indices (Pi), determined using DnaSP and calculated for the three species using a window of 600 bp, showed Pi values ranged from 0 to 0.070, with an average of 0.007, indicating that the divergence between the genomes is small. As described in previous studies [13,25], the IR regions had a much lower nucleotide variability (Pi = 0.001785) than the SSC (Pi = 0.01362) and LSC (Pi = 0.010471) regions (Table S3, Figure 5). Based on this analysis, 14 midpoints of sliding windows showed high levels of nucleotide diversity, with Pi values > 0.025, corresponding to *petA–psbJ*, *psbJ*, *psbL*, *psbF*, *psbF–petL*, *rpl32–trnL-UAG*, *trnQ-UUG*, *ndhF*, *trnL-UAG–ccsA*, *ndhD*, *ycf1*, *trnQ-GCU–trnG-UCC*, *trnG-UCC–trnR-UCU*, *trnE-UUC–trnT-GGU*. Most of these highly variable regions are found in intergenic spacers in LSC and SSC regions. Therefore, these highly variables regions in cp genomes can be useful for phylogenetic reconstruction of the large family of Sapotaceace.

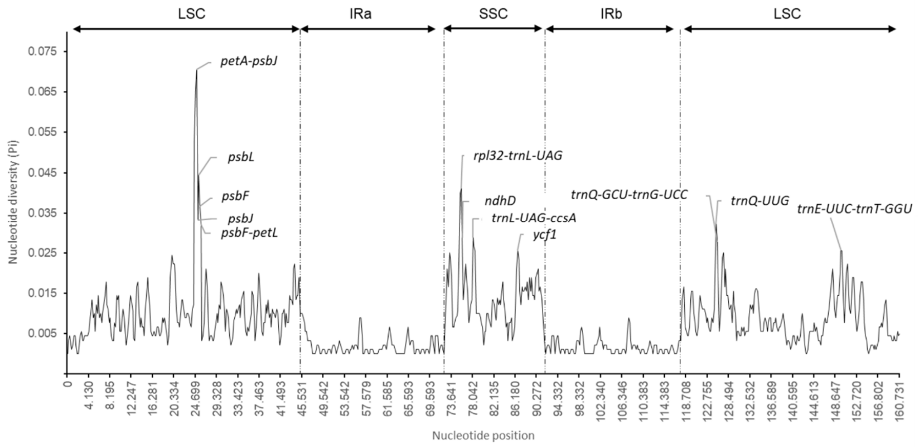

**Figure 5.** Nucleotide diversity (Pi) values for the whole chloroplast genomes of *A. spinosa*, *S. wightianum*, and *P. campechiana* species using a window length of 600 bp and a step size of 200 bp.

To elucidate the selective pressure on the 79 genes in common among the 3 cp genomes, the rates of synonymous (Ks) and nonsynonymous (Ka) substitutions and the Ka/Ks values were calculated (Table S4, Figure 6). The Ka/Ks values may indicate whether selective pressure occurred for plastid genes. Thus, Ka/Ks < 1 suggests that a cp DNA gene was under purifying selection, whereas Ka/Ks ≥ 1 indicates that the gene was affected by positive selection or neutral selection [30].

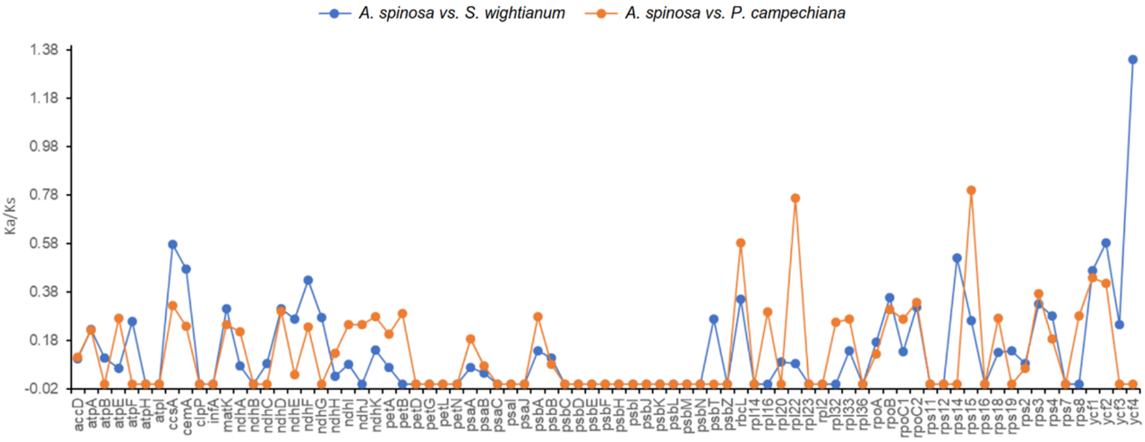

**Figure 6.** The Ka/Ks values of 79 protein-coding genes of the three Sapotaceae cp genomes. The calculation was performed for paired sequences of the species *A. spinosa* vs. *S. wightinaum* and *A. spinosa* vs. *P. campechiana*.

The lowest Ka/Ks ratio was observed for genes encoding NADH (Nicotinamide Adenine Dinucleotide Hydrogen) dehydrogenase, i.e., *ndhE* (Ka/Ks = 0.03) and *ndhD* (Ka/Ks = 0.3), for the paired species *A. spinosa/S. wightianum* and *A. spinosa/P. campechainana*, respectively. The highest Ka/Ks ratios were calculated for *rps15* (Ka/Ks = 0.8) and *ycf4* (Ka/Ks = 1.3) genes for *A. spinosa/S. wightianum* and *A. spinosa/P. campechainana*, respectively.

The Ka/Ks ratio was found to be 0 for 47 genes, the majority of which is located in the LSC region, for the two paired species *A. spinosa/S. wightianum* and *A. spinosa/P. campechainana* (Table S4). For these genes, the *Ka*/*Ks* values could not be calculated because *Ka* or *Ks* was extremely low or equal to 0 [31,32]. The remaining 38 genes, mostly located in the LSC region, showed Ka/Ks ratios below 1.00, indicating a purifying selection. The Ka/Ks ratio was found to indicate positive selection for only one gene, *ycf4* (Ka/Ks = 1.3), for the paired sequence species *A. spinosa/S. wightianum*. Similar results were reported for other cp genomes [33–35].

### *2.5. Long Repeat and Simple Sequence Repeats (SSR) Analysis*

The analysis of long repeats within the *A. spinosa* cp genome showed a total of 44 repeats represented by 18 forward repeats, 23 palindromic repeats, 2 reverse repeats, and 1 complement repeat. Out of 44 repeats within *A. spinosa* cp genome, 30 repeats (69%) were 30–39 bp long, 8 repeats (18%) were >50 bp long, and 6 repeats (13%) were 40–49 bp long (Table S5).

SSRs, also known as microsatellites, are short tandemly repeated sequences of typically 1 to 6 nucleotides repeat units [36]. They are widely distributed in cp genomes, are important for plants population studies because of their high level of polymorphism compared to neutral DNA regions, and are uniparentally inherited. SSRs have been widely used as molecular markers for variety/species identification, molecular breeding, and genetic diversity assessment [37,38]. In this study, SSRs distribution within the *A. spinosa* cp genome was determined. Using the MISA software, the analysis highlighted a total of 88 SSRs composed of 76 mononucleotides SSRs, 7 dinucleotides, and 3, 1, 1 tri-, tetra-, and hexanucleotide, respectively (Figure 7a). The highest number of mononucleotide SSRs were A (47.7%) and T (38.6%) motifs, and most dinucleotide SSRs were found to be AT/TA (6.8%) and TG (1,13) motifs (Figure 7b), which contribute to the A–T richness of *A. spinosa* cp genome. This phenomenon was previously observed in cp genomes of plants species [13,25]. The SSRs were mostly detected in the non-coding regions, containing about 61.36% of total SSRs, but we also found 27.27% of SSR distributed in coding regions, such as *rpoC2* ycF2, *ndhF/G*, and *matK* and 11.36% of SSRs located in tRNA sequences. Our results are comparable to those of several previous studies showing that SSRs in cp genomes are highly rich in polythymine (polyT) or polyadenine (polyA) repeats and infrequently contain tandem cytosine (C) and guanine (G) repeats [13,28,39]. The set of SSRs identified in *A. spinosa* cp genome can be evaluated for polymorphism at the intra-specific level and used as makers for evaluating the genetic diversity among and within populations of *A. spinosa*. These markers could be also used in order to assist the selection and characterization of elite genotypes suitable for the reconstruction and extension of this endangered species.

### *2.6. Phylogenetic Inference of A. spinosa*

In this study, 69 protein-coding genes shared by 11 members of Ericales and 1 species of Lamiales were utilized to depict the phylogenetic relationships of *A. spinosa*. Phylogenetic analyses were performed using the maximum likelihood (ML) method. The topology of the phylogenetic tree (Figure 8) separated the Sapotaceae family harboring the three species *A. spinosa, S. wightianum*, and *P. campechiana* from the neighboring families Theaceae, Primulaceae, and Acitinidiaceae. The obtained phylogenetic tree is not concordant with previous studies based on combined loci (chloroplast, mitochondrial, and nuclear), which placed the Sapotaceae close to the Ebenaceae and not to the Primulaceae [40,41]. Inconsistent topology between nuclear and plastome phylogenies has been reported for the Asterids clade and could be explained by considering several evolutionary processes such as hybridization, horizontal gene transfer, and gene duplication and loss [42]. Interestingly, the three Sapotaceae species were clustered in two clades: one clade grouping the genus *Sideroxylon* (*S. wightianum)* with *Argania* (*A. spinosa*), and the second harboring *Pouteria* (*P. campechiana*), with a strong bootstraps value (100%). In addition to the monophyletic character of the Sapotaceae members included in this study, the results highlight the close relationship of *A. spinosa* and *S. wightianum*, in accordance with the topologies inferred in precedent phylogenetic studies [7,8]. In a cladistic study of the largely tropical family Sapotaceae based on both morphological and molecular data (cp gene *ndhF*), the generated trees showed that *A. spinosa* and *Sideroxylon mascatense* attach as sisters to each other and belong to the genus *Sideroxylon* [9]. Moreover, phylogenetics inference based on 58 accessions of *trnH–psbA* and ITS sequences from *Sideroxylon* was congruent with this study [10]. Despite the indisputable fact, reported in previous studies, that *A. spinosa* is amended to *Sideroxylon*, autapomorphies characters distinguishing *Argania* from *Sideroxylon* [7,8,11] require a comprehensive exploration to confirm this relationship by including several cp genome sequences of *Sideroxylon* members close to *A. spinosa*, such as *S. mascatense*, *Sideroxylon canariense*, *Sideroxylon oxyacanthum*, and *Sideroxylon discolor*.

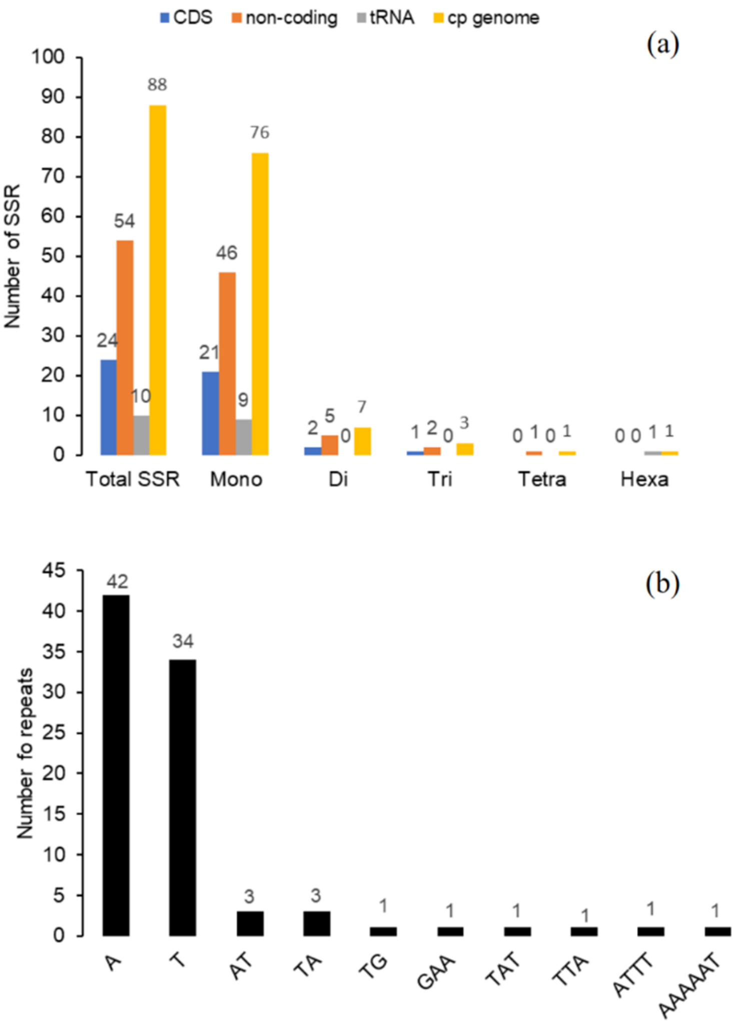

**Figure 7.** Analysis of simple sequence repeats (SSRs) in *A. spinosa* cp genome. (**a**) Number of SSR types in the cp complete genome, coding, and non-coding regions; (**b**) Number and type of SSR motifs identified in the cp genome. CDS, coding domain sequences.

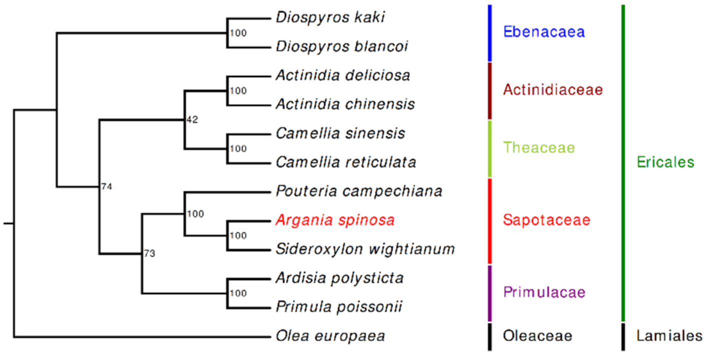

**Figure 8.** Phylogenetic tree reconstruction of 12 taxa using the maximum likelihood (ML) method based on 69 shared protein sequences. The number above each node indicates the bootstrap support values. *Olea europaea* was used as an outgroup.

## 3. Materials and Methods

### *3.1. Plant Material, DNA Extraction, and Sequencing*

A single argan tree named Amghar was selected to be sequenced based on its biological and ecological proprieties. The shrub was 9 years old, with one main trunk measuring 3 m in height, and was obtained from the valley of the plain of Souss (9°32′ 00″ N, 30°24′ 00″ W; Altitude: 126 m). Genomic DNA was extracted from lyophilized leaf tissues using the Plant DNeasy mini kit according to the manufacturer's recommendations (Qiagen, Germantown, MD, USA). Two paired-end libraries with an average insert size of 600 bp were constructed using the Nextera DNA Library Prep Kit for Illumina (New England Biolabs, New Brunswick, MA, USA) and then sequenced on the Illumina HiSeqXTen (San Diego, CA, USA) platform using 150-bp reads.

### *3.2. Chloroplast Genome Assembly*

To extract chloroplast-like reads, quality-filtered Illumina paired-end reads were mapped against the closest firstly available chloroplast genome of *P. campechiana* (taxid: 233737), using CLC Genomics (v11.0, CLCbio, Arhus, Denmark), with 0.9 and 0.95 in length and similarity, respectively. The extracted reads were de novo assembled using CLC Genomics (word size 24, bubble size 50). The generated contigs were blasted against the reference genome of *P. campechiana* (Costs: Match 2, Mismatch 3, Existence 5, Extension 2, Expectation value = $1.0 \times 10^{-15}$, Word size = 11) to assess the assembly and to retain only the contigs aligned to the reference. The retained contigs were than assembled using the Genome Finishing Module (GFM) from CLC genomics that uses the paired-end distance information and the reference genome to order the contigs and to fill the gaps. A quadripartite structure, including IR, SCC, and LSC regions, was detected by performing a dot plot of the *A. spinosa* chloroplast sequence using the Gepard (v1.30) software (word size = 10) [43]. The validation of the assembly was performed by PCR amplification and Sanger sequencing using four couples of primers designed on the boundaries of the IR and SC regions (Table S1).

### *3.3. Genome Annotation and Comparisons*

*A. spinosa* chloroplast genome was annotated through the DOGMA server [44]. The GenBank file produced was loaded into CLC genomics, and the gene list was processed manually gene by gene for the presence of start/stop codons and for internal stops in comparison to the closer cp genome of *P. campechiana*. Circular maps of the cp genome were generated using OGDraw v1.2 (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html) [45]. The codon usage percentage of protein-coding sequences was estimated using the MEGA7 software [46]. Comparative genomics of *A. spinosa*, *S. wightianum*, and *P. campechiana* cp genomes was conducted using the mVISTA program in the Shuffle-LAGAN mode [47]. To calculate nucleotide diversity (Pi) between *A. spinosa, S. wightianum*, and *P. campechiana* chloroplast genomes, sliding window analysis was performed using the DnaSP version 6 software [48] with window length of 600 bp and step size of 200 bp. To assess the selective pressure on the shared protein-coding genes across the three species, the rates of synonymous (Ks) and nonsynonymous (Ka) substitutions and the Ka/Ks ratio were calculated using a Ka/Ks calculator [49].

### *3.4. Long Repetitive Sequences and Simple Sequence Repeat Analysis*

Long repetitive repeat sequences, including forward, reverse, palindromic, and complement repeats, with repeat size ≥30 bp and sequence identity ≥90%, were identified using the REPuter software [50]. SSRs within the *A. spinosa* chloroplast genome were searched using the MISA software [51]. The criteria of SSR research were set to 10 repeat units as a minimum for mononucleotide repeats 5 repeat units for dinucleotides repeats, and 4 repeat units for tri- and tetranucleotides. For pentanucleotides and hexanucleotides, 3 repeats were used as the minimum.

### 3.5. Phylogenetic Analyses

To ascertain the phylogenetic position of *A. spinosa* within the Sapotaceae family, 11 chloroplast genomes were downloaded from NCBI: *Actinidia chinensis* (NC_026690), *Actinidia deliciosa* (NC_026691), *Ardisia polysticta* (NC_021121), *Camellia reticulata* (NC_024663), *Camellia sinensis* (NC_020019), *Diospyros blancoi* (KX426216); *Diospyros kaki* (NC_030789), *Pouteria campechiana* (KX426215), *Primula poissonii* (NC_024543), *Sideroxylon wightianum* (NC_041130), and *Olea europaea* (NC_013707). Phylogenetic analysis was based on 69 protein-coding genes shared among the 12 taxa including *A. spinosa*. Individual protein sequence alignments were performed using MAFFT with default parameters [52] and then were concatenated using SeqKit with defaults parameters [53]. The whole alignment was trimmed using TrimAl [54], and the phylogeny was inferred using Randomized Axelerated Maximum Likelihood (RAxML) [55] with 1000 bootstrap replications. *O. europaea* taxon was used as an outgroup in this analysis.

## 4. Conclusions

Comparative analyses of complete cp genomes contribute to the understanding of chloroplast structure and evolution, the identification of species, and the determination of phylogenetic relationships. In this study, we applied Illumina sequencing to determine, for the first time, the complete cp genome of the endemic species *A. spinosa*. The genome structure and genes order and content were found to be very conserved with respect to those of the close species *S. wightianum* and *P. campechiana*. Furthermore, the phylogenomic analyses based on whole cp genomes and 77 shared genes generated trees with the same topologies as previously reported, consolidating the taxonomical position of *A. spinosa* species within the Sapotaceae. To clarify the view of amending the genus *Argania* to *Sideroxylon*, it is appropriate to include more cp genomes of *Sideroxylon* members in future studies. The 44 long repeats and 88 SSRs identified here are a useful genetic resource that could be applied for population genetic studies and may also be useful for future breeding and cultivars identification. Finally, these genomic resources will certainly help in the management and conservation of this endangered species.

**Supplementary Materials:** The following are available online at http://www.mdpi.com/2223-7747/9/10/1354/s1, Figure S1: Gel electrophoresis images of PCRs performed on four different individuals of *A. spinosa* (V1, V2, V3, and V4). A and D couple of primers targeting the IRa/LSC and IRa/SSC junctions and amplifying, respectively, 700- and 258-bp fragments. B and C primers targeting the IRb/SSC and IRb/LSC junctions, amplifying, respectively, 699- and 361-bp fragments. The M lane represents the size marker. Table S1: Primers used for assembly validation of the IR/SC junction Table S2: Codon–anticodon recognition patterns and codon usage for the *A. spinosa* chloroplast genome. Table S3: Nucleotide diversity (Pi) values between the whole chloroplast genomes of *A. spinosa*, *S. wightianum*, and *P. campechiana* species using a window length of 600 bp and a step size of 200 bp, Table S4: Ka/Ks values of 79 protein-coding genes calculated for the paired sequences of the species *A. spinosa* vs. *S. wightinaum* and *A. spinosa* vs. *P. campechiana*. Table S5. Long repeats distribution within the *A. spinosa* cp genome

**Author Contributions:** Conceptualization, S.K. and R.M.; methodology, S.K., F.G., S.P., T.T., A.E.M., and R.M.; Z.Z.; investigation, S.K., F.G., S.P., H.G., and R.M.; resources, A.E.M., H.G.; S.K., F.G., S.P., T.T., A.E.M.; writing—original draft preparation, S.K. and R.M.; writing—review and editing, S.K., F.G., S.P., T.T., A.E.M., H.G., and R.M.; supervision, S.K., F.G., S.P., T.T., A.E.M., H.G., and R.M.; project administration, S.K., F.G., S.P., T.T., A.E.M., H.G., and R.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Mid-Term Program of the Research of National Institute of Agricultural Research of Morocco (INRA, PRMT 2017-2020).

**Acknowledgments:** Thanks are due to the KAFACI for supporting Moroccan mobility of researchers through the Young Scientist Pilot Research Project program in 2018. We thank Claude dePamphilis for his advices during the preparation of the manuscript. I would also like to thank the anonymous reviewers for their constructive comments that improved the manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Kenny, L. *Atlas De L'Arganier et de L'Arganeraie*; Institut Agronomique et Vétérinaire Hassan II, Complexe Horticole d'Agadir; Agropolis International: Agadir, Montpellier, France, 2007; ISBN 978-2-909613-00-0.
2. Charrouf, Z.; Guillaume, D. Argan oil: Occurrence, composition and impact on human health. *Eur. J. Lipid Sci. Technol.* **2008**, *110*, 632–636. [CrossRef]
3. McGregor, H.V.; Dupont, L.; Stuut, J.-B.W.; Kuhlmann, H. Vegetation change, goats, and religion: A 2000-year history of land use in southern Morocco. *Quat. Sci. Rev.* **2009**, *28*, 1434–1448. [CrossRef]
4. Le Polain de Waroux, Y.; Lambin, E.F. Monitoring degradation in arid and semi-arid forests and woodlands: The case of the argan woodlands (Morocco). *Appl. Geogr.* **2012**, *32*, 777–786. [CrossRef]
5. Vaghani, S.N. Fruits of Tropical Climates | Fruits of the Sapotaceae. In *Encyclopedia of Food Sciences and Nutrition*; Elsevier: New York, NY, USA, 2003; pp. 2790–2800, ISBN 978-0-12-227055-0.
6. Duangjai, S.; Wallnofer, B.; Samuel, R.; Munzinger, J.; Chase, M.W. Generic delimitation and relationships in Ebenaceae sensu lato: Evidence from six plastid DNA regions. *Am. J. Bot.* **2006**, *93*, 1808–1827. [CrossRef] [PubMed]
7. Smedmark, J.E.E.; Anderberg, A.A. Boreotropical migration explains hybridization between geographically distant lineages in the pantropical clade Sideroxyleae (Sapotaceae). *Am. J. Bot.* **2007**, *94*, 1491–1505. [CrossRef]
8. Smedmark, J.E.E.; Swenson, U.; Anderberg, A.A. Accounting for variation of substitution rates through time in Bayesian phylogeny reconstruction of Sapotoideae (Sapotaceae). *Mol. Phylogenet. Evol.* **2006**, *39*, 706–721. [CrossRef] [PubMed]
9. Swenson, U.; Anderberg, A.A. Phylogeny, character evolution, and classification of Sapotaceae (Ericales). *Cladistics* **2005**, *21*, 101–130. [CrossRef]
10. Stride, G.; Nylinder, S.; Swenson, U. Revisiting the biogeography of Sideroxylon (Sapotaceae) and an evaluation of the taxonomic status of Argania and Spiniluma. *Aust. Syst. Bot.* **2014**, *27*, 104. [CrossRef]
11. Anderberg, A.A.; Rydin, C.; Kallersjo, M. Phylogenetic relationships in the order Ericales s.l.: Analyses of molecular data from five genes from the plastid and mitochondrial genomes. *Am. J. Bot.* **2002**, *89*, 677–687. [CrossRef]
12. Jo, S.; Kim, H.-W.; Kim, Y.-K.; Cheon, S.-H.; Kim, K.-J. The first complete plastome sequence from the family Sapotaceae, *Pouteria campechiana* (Kunth) Baehni. *Mitochondrial DNA Part B* **2016**, *1*, 734–736. [CrossRef]
13. Kuang, D.-Y.; Wu, H.; Wang, Y.-L.; Gao, L.-M.; Zhang, S.-Z.; Lu, L. Complete chloroplast genome sequence of Magnolia kwangsiensis (Magnoliaceae): Implication for DNA barcoding and population genetics. *Genome* **2011**, *54*, 663–673. [CrossRef] [PubMed]
14. Liu, H.-Y.; Yu, Y.; Deng, Y.-Q.; Li, J.; Huang, Z.-X.; Zhou, S.-D.; Liu, H.-Y.; Yu, Y.; Deng, Y.-Q.; Li, J.; et al. The Chloroplast Genome of Lilium henrici: Genome Structure and Comparative Analysis. *Molecules* **2018**, *23*, 1276. [CrossRef] [PubMed]
15. Niu, Y.-F.; Ni, S.-B.; Liu, Z.-Y.; Zheng, C.; Mao, C.-L.; Shi, C.; Liu, J. The complete chloroplast genome of tropical and sub-tropical fruit tree Lucuma nervosa (Sapotaceae). *Mitochondrial DNA Part B* **2018**, *3*, 440–441. [CrossRef]
16. Ritland, K.; Clegg, M.T. Evolutionary Analysis of Plant DNA Sequences. *Am. Nat.* **1987**, *130*, S74–S100. [CrossRef]
17. Pérez-Jiménez, M.; Besnard, G.; Dorado, G.; Hernandez, P. Varietal Tracing of Virgin Olive Oils Based on Plastid DNA Variation Profiling. *PLoS ONE* **2013**, *8*, e70507. [CrossRef] [PubMed]
18. Santos, C.; Pereira, F. Identification of plant species using variable length chloroplast DNA sequences. *Forensic Sci. Int. Genet.* **2018**, *36*, 1–12. [CrossRef]
19. Khayi, S.; Azza, N.E.; Gaboun, F.; Pirro, S.; Badad, O.; Claros, M.G.; Lightfoot, D.A.; Unver, T.; Chaouni, B.; Merrouch, R.; et al. First draft genome assembly of the Argane tree (Argania spinosa). *F1000Research* **2018**, *7*, 1310. [CrossRef]
20. El Mousadik, A.; Petit, R.J. Chloroplast DNA phylogeography of the argan tree of Morocco. *Mol. Ecol.* **1996**, *5*, 547–555. [CrossRef] [PubMed]
21. Daniell, H.; Lin, C.-S.; Yu, M.; Chang, W.-J. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Biol.* **2016**, *17*, 134. [CrossRef] [PubMed]

22. Gitzendanner, M.A.; Soltis, P.S.; Yi, T.-S.; Li, D.-Z.; Soltis, D.E. Plastome Phylogenetics: 30 Years of Inferences Into Plant Evolution. In *Advances in Botanical Research*; Elsevier: New York, NY, USA, 2018; Volume 85, pp. 293–313, ISBN 978-0-12-813457-3.
23. Raman, G.; Park, S. The Complete Chloroplast Genome Sequence of Ampelopsis: Gene Organization, Comparative Analysis, and Phylogenetic Relationships to Other Angiosperms. *Front. Plant Sci.* **2016**, *7*, 341. [CrossRef] [PubMed]
24. Park, I.; Kim, W.; Yeo, S.-M.; Choi, G.; Kang, Y.-M.; Piao, R.; Moon, B. The Complete Chloroplast Genome Sequences of Fritillaria ussuriensis Maxim. and Fritillaria cirrhosa D. Don, and Comparative Analysis with Other Fritillaria Species. *Molecules* **2017**, *22*, 982. [CrossRef]
25. Asaf, S.; Waqas, M.; Khan, A.L.; Khan, M.A.; Kang, S.-M.; Imran, Q.M.; Shahzad, R.; Bilal, S.; Yun, B.-W.; Lee, I.-J. The Complete Chloroplast Genome of Wild Rice (Oryza minuta) and Its Comparison to Related Species. *Front. Plant Sci.* **2017**, *8*, 304. [CrossRef] [PubMed]
26. Chen, J.; Hao, Z.; Xu, H.; Yang, L.; Liu, G.; Sheng, Y.; Zheng, C.; Zheng, W.; Cheng, T.; Shi, J. The complete chloroplast genome sequence of the relict woody plant Metasequoia glyptostroboides Hu et Cheng. *Front. Plant Sci.* **2015**, *6*, 447. [CrossRef] [PubMed]
27. Qian, J.; Song, J.; Gao, H.; Zhu, Y.; Xu, J.; Pang, X.; Yao, H.; Sun, C.; Li, X.; Li, C.; et al. The complete chloroplast genome sequence of the medicinal plant Salvia miltiorrhiza. *PloS ONE* **2013**, *8*, e57607. [CrossRef] [PubMed]
28. Asaf, S.; Khan, A.L.; Khan, M.A.; Shahzad, R.; Lubna; Kang, S.M.; Al-Harrasi, A.; Al-Rawahi, A.; Lee, I.-J. Complete chloroplast genome sequence and comparative analysis of loblolly pine (Pinus taeda L.) with related species. *PLoS ONE* **2018**, *13*, e0192966. [CrossRef]
29. Palmer, J.D.; Thompson, W.F. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell* **1982**, *29*, 537–550. [CrossRef]
30. Nei, M.; Kumar, S. *Molecular Evolution and Phylogenetics*; Oxford University Press: Oxford, NY, USA, 2000; ISBN 978-0-19-513584-8.
31. Wang, D.; Liu, F.; Wang, L.; Huang, S.; Yu, J. Nonsynonymous substitution rate (Ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes. *Biol. Direct* **2011**, *6*, 13. [CrossRef]
32. Redwan, R.M.; Saidin, A.; Kumar, S.V. Complete chloroplast genome sequence of MD-2 pineapple and its comparative analysis among nine other plants from the subclass Commelinidae. *BMC Plant Biol.* **2015**, *15*. [CrossRef]
33. Rousseau-Gueutin, M.; Bellot, S.; Martin, G.E.; Boutte, J.; Chelaifa, H.; Lima, O.; Michon-Coudouel, S.; Naquin, D.; Salmon, A.; Ainouche, K.; et al. The chloroplast genome of the hexaploid Spartina maritima (Poaceae, Chloridoideae): Comparative analyses and molecular dating. *Mol. Phylogenet. Evol.* **2015**, *93*, 5–16. [CrossRef]
34. Xu, J.-H.; Liu, Q.; Hu, W.; Wang, T.; Xue, Q.; Messing, J. Dynamics of chloroplast genomes in green plants. *Genomics* **2015**, *106*, 221–231. [CrossRef]
35. Zhou, T.; Chen, C.; Wei, Y.; Chang, Y.; Bai, G.; Li, Z.; Kanwal, N.; Zhao, G. Comparative Transcriptome and Chloroplast Genome Analyses of Two Related Dipteronia Species. *Front. Plant Sci.* **2016**, *7*. [CrossRef] [PubMed]
36. Richards, R.I.; Sutherland, G.R. Simple repeat DNA is not replicated simply. *Nat. Genet.* **1994**, *6*, 114–116. [CrossRef] [PubMed]
37. Echt, C.S.; DeVerno, L.L.; Anzidei, M.; Vendramin, G.G. Chloroplast microsatellites reveal population genetic diversity in red pine, *Pinus resinosa* Ait. *Mol. Ecol.* **1998**, *7*, 307–316. [CrossRef]
38. Leclercq, S.; Rivals, E.; Jarne, P. Detecting microsatellites within genomes: Significant variation among algorithms. *BMC Bioinform.* **2007**, *8*, 125. [CrossRef] [PubMed]
39. Hao, Z.; Cheng, T.; Zheng, R.; Xu, H.; Zhou, Y.; Li, M.; Lu, F.; Dong, Y.; Liu, X.; Chen, J.; et al. The Complete Chloroplast Genome Sequence of a Relict Conifer Glyptostrobus pensilis: Comparative Analysis and Insights into Dynamics of Chloroplast Genome Rearrangement in Cupressophytes and Pinaceae. *PLoS ONE* **2016**, *11*, e0161809. [CrossRef]
40. Rose, J.P.; Kleist, T.J.; Löfstrand, S.D.; Drew, B.T.; Schönenberger, J.; Sytsma, K.J. Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient Neotropical and East Asian connections. *Mol. Phylogenet. Evol.* **2018**, *122*, 59–79. [CrossRef]

41. Larson, D.A.; Walker, J.F.; Vargas, O.M.; Smith, S.A. A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales. *Am. J. Bot.* **2020**, *107*, 773–789. [CrossRef]
42. Stull, G.W.; Soltis, P.S.; Soltis, D.E.; Gitzendanner, M.A.; Smith, S.A. Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel relationships among major lineages. *Am. J. Bot.* **2020**, *107*, 790–805. [CrossRef] [PubMed]
43. Krumsiek, J.; Arnold, R.; Rattei, T. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **2007**, *23*, 1026–1028. [CrossRef]
44. Wyman, S.K.; Jansen, R.K.; Boore, J.L. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **2004**, *20*, 3252–3255. [CrossRef]
45. Lohse, M.; Drechsel, O.; Kahlau, S.; Bock, R. OrganellarGenomeDRAW—A suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* **2013**, *41*, W575–W581. [CrossRef] [PubMed]
46. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **2016**, *33*, 1870–1874. [CrossRef] [PubMed]
47. Frazer, K.A.; Pachter, L.; Poliakov, A.; Rubin, E.M.; Dubchak, I. VISTA: Computational tools for comparative genomics. *Nucleic Acids Res.* **2004**, *32*, W273–W279. [CrossRef] [PubMed]
48. Rozas, J.; Ferrer-Mata, A.; Sánchez-DelBarrio, J.C.; Guirao-Rico, S.; Librado, P.; Ramos-Onsins, S.E.; Sánchez-Gracia, A. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* **2017**, *34*, 3299–3302. [CrossRef]
49. Wang, D.; Zhang, Y.; Zhang, Z.; Zhu, J.; Yu, J. KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. *Genom. Proteom. Bioinform.* **2010**, *8*, 77–80. [CrossRef]
50. Kurtz, S.; Choudhuri, J.V.; Ohlebusch, E.; Schleiermacher, C.; Stoye, J.; Giegerich, R. REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **2001**, *29*, 4633–4642. [CrossRef]
51. Beier, S.; Thiel, T.; Münch, T.; Scholz, U.; Mascher, M. MISA-web: A web server for microsatellite prediction. *Bioinformatics* **2017**, *33*, 2583–2585. [CrossRef] [PubMed]
52. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]
53. Shen, W.; Le, S.; Li, Y.; Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE* **2016**, *11*, e0163962. [CrossRef]
54. Capella-Gutiérrez, S.; Silla-Martínez, J.M.; Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinforma. Oxf. Engl.* **2009**, *25*, 1972–1973. [CrossRef]
55. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef]