




Article

Assembly and Analysis of the Complete Mitochondrial Genome of *Capsella bursa-pastoris*

Denis O. Omelchenko ^{1,*}, Maxim S. Makarenko ^{1,*}, Artem S. Kasianov ¹,
Mikhail I. Schelkunov ^{1,2}, Maria D. Logacheva ^{1,2} and Aleksey A. Penin ¹

¹ Institute for Information Transmission Problems of the Russian Academy of Sciences, 127051 Moscow, Russia; artem.kasianov@gmail.com (A.S.K.); shelkmike@gmail.com (M.I.S.); maria.log@gmail.com (M.D.L.); alekseypenin@gmail.com (A.A.P.)

² Skolkovo Institute of Science and Technology, 121205 Moscow, Russia

* Correspondence: omelchenkodo@iitp.ru (D.O.O.); mcmakarenko@yandex.ru (M.S.M.)

† These authors contributed equally to this work.

Received: 4 March 2020; Accepted: 4 April 2020; Published: 8 April 2020



Abstract: Shepherd's purse (*Capsella bursa-pastoris*) is a cosmopolitan annual weed and a promising model plant for studying allopolyploidization in the evolution of angiosperms. Though plant mitochondrial genomes are a valuable source of genetic information, they are hard to assemble. At present, only the complete mitogenome of *C. rubella* is available out of all species of the genus *Capsella*. In this work, we have assembled the complete mitogenome of *C. bursa-pastoris* using high-precision PacBio SMRT third-generation sequencing technology. It is 287,799 bp long and contains 32 protein-coding genes, 3 rRNAs, 25 tRNAs corresponding to 15 amino acids, and 8 open reading frames (ORFs) supported by RNAseq data. Though many repeat regions have been found, none of them is longer than 1 kbp, and the most frequent structural variant originated from these repeats is present in only 4% of the mitogenome copies. The mitochondrial DNA sequence of *C. bursa-pastoris* differs from *C. rubella*, but not from *C. orientalis*, by two long inversions, suggesting that *C. orientalis* could be its maternal progenitor species. In total, 377 C to U RNA editing sites have been detected. All genes except *cox1* and *atp8* contain RNA editing sites, and most of them lead to non-synonymous changes of amino acids. Most of the identified RNA editing sites are identical to corresponding RNA editing sites in *A. thaliana*.

Keywords: *Capsella bursa-pastoris*; complete mitochondrial genome; SMRT PacBio; structural variants; RNA editing

1. Introduction

Shepherd's purse (*Capsella bursa-pastoris*) is a small herbaceous plant of the mustard family (Brassicaceae) and one of the most common weeds growing in diverse habitats on almost every continent. Being a recent allotetraploid and a close relative of the well-known model plant *Arabidopsis thaliana*, it is a promising model plant for studying polyploidization and its role in the adaptation and evolution of the flowering plants [1,2].

Mitochondria are important organelles that provide energy conversion from the food fuel molecules into cell usable ATP energy storage molecules in eukaryotic cells. Mitochondrial genomes (mitogenomes) are a valuable source of genetic information for phylogenetic studies and the investigation of essential cellular processes. Plant mitogenomes are highly variable molecules in both size and structure, in contrast to chloroplast genomes that have highly conserved quadripartite structure among land plants [3]. Mitochondrial genomes vary greatly not only between species but sometimes even within the same species [4,5]. The size of known angiosperm mitogenomes varies

from 66 kbp in *Viscum scurruloideum* [6] to up to 11.3 Mbp in *Silene conica* [7]. Plant mitochondrial DNA (mtDNA) contains many repeats as well as inserts of nuclear and chloroplast origin, which makes mitogenome assembly difficult [8]. However, the development of long-read third-generation sequencing technologies (PacBio in particular) improves and simplifies assembly of such complex molecules, greatly facilitating the research of plant mitogenomes [9–13].

Due to the development of second- and third-generation sequencing technologies, the number of fully sequenced mitogenomes of a diverse spectrum of plants is growing fast, going beyond a set of economically important edible species [14]. Though to date, only 12 complete mitogenomes of species from the family Brassicaceae are presented in the GenBank genome database and only one of them from the genus *Capsella*—*C. rubella* (one of the *C. bursa-pastoris* progenitor species) [15]. In this work, we have assembled the complete mitogenome of *C. bursa-pastoris* using the single-molecule real-time (SMRT) PacBio sequencing technology. We have studied its gene profile and have analyzed its sequence and structure in comparison to currently available complete mitogenomes of the closely related species *C. rubella* and *A. thaliana*. We have also investigated RNA editing sites using RNAseq data obtained from rRNA depleted total RNA of *C. bursa-pastoris*, and have identified several long (more than 300 bp) open reading frames (ORFs), the expression of which is supported by the RNAseq data.

2. Results

2.1. Sequencing and Assembly of the Complete Mitogenome of *C. bursa-pastoris*

Capsella bursa-pastoris total DNA was sequenced, and raw data had been prepared for assembly, resulting in 1,687,990 circular consensus sequencing (CCS) reads with an average read length of 7948 bp (61–26,706 bp) and an average quality of 83. Reads were aligned to the *C. bursa-pastoris* cpDNA reference sequence, and those that mapped with less than 5% divergence had been removed to avoid the interfering of cpDNA reads in the mitogenome assembly. Due to the length of CCS reads (~8 kbp or longer on average), mitogenome reads that contain chloroplast inserts have also passed the filter. After filtration, a 10% subsample has been isolated (145,590 reads with the average read length of 7964 bp) and used to assemble mitogenome, as downsampling often improves organelle assembly (e.g., [16,17]). One contig was identified as of a mitochondrial origin based on BLASTn search. It was circularized by merging ~9 kbp repeats at its ends. Then all CCS reads were mapped using minimap2 on the circularized contig to check the correctness of the assembly, resulting in 90,104 reads mapped to the contig covering 100% of its bases with an average coverage of 392. Thus, 287,799 bp long circular contig with an overall GC content of 44.74% has been identified as *C. bursa-pastoris* mtDNA.

2.2. Repeats and Structural Variation Analysis

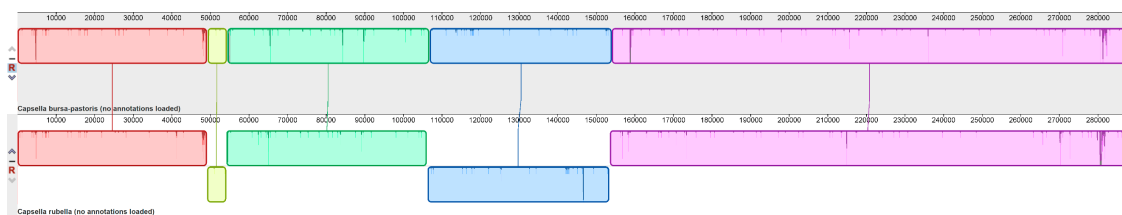
The sequence of the *C. bursa-pastoris* mitogenome contains 73 direct and 68 inverted repeats ranging from 28 to 854 bp in length with a minimum identity of 80% (Table S1). In total, repeats occupy 8.2% of the mitogenome. To identify structural variants possibly emerging from these repeats, all CCS reads were mapped to the mitogenome using NGMLR, and then Sniffles was used to identify structural variants in the alignment, and was restricted to search only variants supported by at least 2 reads. Despite a large number of repeats found in the mitogenome, only 9 supported structural variants, corresponding to 8 repeats, have been found (Table 1).

Table 1. The structural variants in the mitogenome of *C. bursa-pastoris* supported by at least two circular consensus sequencing (CCS) reads.

Repeat ID	Structural Variant Type	Number of Supporting CCS Reads	Repeat Length (bp)	First Repeat Unit Position (bp)	Second Repeat Unit Position (bp)	Repeat Units' Sequence Identity
Rep_1	inversion	13	854	54,219–55,072	49,506–48,653	98.9%
Rep_2	inversion	8	635	158,147–158,781	106,515–105,881	100.0%
Rep_3	duplication	6	538	134,019–134,556	116,143–116,679	99.8%
Rep_7	deletion	5	418	220,620–221,037	54,655–55,072	99.8%
Rep_10	inversion	4	420	220,623–221,042	49,067–48,648	97.9%
Rep_11	duplication	3	356	64,729–65,084	26,047–26,402	100.0%
Rep_12	duplication	3	327	93,763–94,089	62,446–62,772	100.0%
Rep_9	deletion	2	404	238,171–238,573	17,371–17,774	99.8%
Rep_11	deletion	2	356	64,729–65,084	26,047–26,402	100.0%

The number of reads that supports the structural variant existence is proportional to the Bit score of the MEGABLAST algorithm. Generally, the longer the repeat and the higher the nucleotide similarity between its units, the more often recombination will occur. Thus, three repeats with the highest Bit score (Rep_1, Rep_2, Rep_3) give the three most frequent structural variants. For those structural options that are represented by a small number of reads, the correlation with the Bit score is weaker, because it increases the probability that the number of supporting reads of those structural variants is random. In total, 12,019 reads were mapped using NGMLR software to the *C. bursa-pastoris* mitogenome with the average coverage of 332. It means that even the most frequent of the structural variants could be found in only ~4% of the mitogenome copies. Though, despite the low representation, some of the structural variants lead to extensive changes in the mitogenome structure. For example, Rep_2 inverts the ~50 kbp long fragment of the mitogenome, and Rep_7 deletion splits it into two ~166 and ~120 kbp long molecules.

The full-genome alignment of the *C. bursa-pastoris* and *C. rubella* mtDNA sequences has shown that their mitogenomes are mostly collinear, except for two 47,173 bp and 5021 bp long inversions (Figure 1). The inversions are localized at sites 106,923–154,095 bp and 49,308–54,328 bp in the *C. bursa-pastoris* mitogenome and their borders lie within repeats Rep_4 and Rep_1, correspondingly.

**Figure 1.** Full-genome alignment of mtDNA of *C. bursa-pastoris* (top) and *C. rubella* (bottom). Inversion block 47,173 bp is blue, and inversion block 5021 bp is yellow.

CCS reads were mapped to *C. rubella* and two *C. bursa-pastoris* mtDNA sequences—one with an originally assembled sequence and the second where inversion regions have been manually replaced by reverse-complementary ones. Alignment of CCS reads to the artificial inversion region would reveal the presence of another *C. bursa-pastoris* mtDNA isoform collinear to the *C. rubella* mtDNA if it exists in the mitochondria. Mapping has shown that the uniformity and depth of coverage decrease at the borders of inversions of the artificially created isoform of the *C. bursa-pastoris* mtDNA (Figure S1a–d). As for *C. rubella*, mapping has shown zero coverage at the borders of the inversions (Figure S1e,f), which indicates that the originally assembled version of the *C. bursa-pastoris* mtDNA is its primary form. To investigate the inversions further, we have checked their presence in another progenitor species of *C. bursa-pastoris*—*C. orientalis*. Illumina paired-end reads from the GenBank WGS data of *C. orientalis* were mapped to the *C. bursa-pastoris* and *C. rubella* mtDNA sequences. Mapping results have shown that *C. orientalis* mitochondrial reads better align to the *C. bursa-pastoris* than to

the *C. rubella* mtDNA across the entire sequence. *Capsella orientalis* reads aligned uniformly with high coverage at the borders of inversions to the *C. bursa-pastoris* mtDNA, while it failed to do so with the *C. rubella* mtDNA (Figure S2a–d). Thus, we suggest that *C. orientalis* is the maternal progenitor of *C. bursa-pastoris*, as mitochondria are primarily maternally inherited.

2.3. Gene Content of the *C. bursa-pastoris* Mitogenome and Comparison with *C. rubella* and *A. thaliana*

The mitochondrial genome of *C. bursa-pastoris* contains 32 protein-coding genes, 3 rRNAs, 8 ORFs of unknown function, and 25 tRNAs corresponding to 15 amino acids (Figure 2). Comparison of the *C. bursa-pastoris* mitogenome gene content with *C. rubella* and *A. thaliana* (Table S2) has shown that the set of protein-coding genes and rRNA in these plants is identical, except for two copies of the *atp6* gene in *A. thaliana*. However, some of the genes differ in both length and sequence between investigated species. *Arabidopsis thaliana* has two copies of the *atp6* gene, one of which is entirely different from the *atp6* genes of *C. rubella* and *C. bursa-pastoris*. The other copy differs from genes in *C. rubella* and *C. bursa-pastoris* by several non-synonymous substitutions and an indel spanning from 49 to 93 position in the gene nucleotide sequence. This indel makes the *atp6* gene of *A. thaliana* longer, and its nucleotide sequence encodes 15 amino acids instead of 5 amino acids in corresponding sequences of *C. bursa-pastoris* and *C. rubella*. *C. rubella* Current annotation of *C. rubella* states that the *atp9* gene is longer than its orthologs in *C. bursa-pastoris* and *A. thaliana* by 33 bp and starts from ATG codon upstream of the ATG start codon common for this gene among Brassicaceae. The *ccmFC* gene of *C. rubella* is 30 bp longer than its orthologs in *C. bursa-pastoris* and *A. thaliana*. RNA editing creates a TGA stop codon that ends the *ccmFC* gene sequence in *A. thaliana* and *C. bursa-pastoris*. However, in *C. rubella* it continues further downstream to the TAA stop codon. There is no information on RNA editing in *C. rubella*, so the difference in length of the *ccmFC* genes could be due to incorrect annotation of this gene in the GenBank record of the *C. rubella* mitogenome. The *ccmFN1* gene of *A. thaliana* is 6 bp longer than its orthologs in *C. rubella* and *C. bursa-pastoris* due to an insert at the beginning of the gene. The *matR* gene of *C. rubella* is 60 bp longer at the 5' end than its orthologs in *A. thaliana* and *C. bursa-pastoris*, and the annotation states that its start codon has not been determined. However, except for these additional nucleotides, gene sequences are similar for all Brassicaceae species. They have a common ATG start codon at the same position, including *C. rubella*, which allows us to suggest the incorrect annotation of this gene in the GenBank record of the *C. rubella* mitogenome. The *mttB* gene of *A. thaliana* is 27 bp longer than *mttB* in *C. rubella* and *A. thaliana* due to a frameshift caused by additional C nucleotides in the poly(C) sequence at the 3' end of the gene, changing the TAG stop codon into CCT and moving the termination signal further downstream to the TGA stop codon. The start codon of the *mttB* gene is not determined, and both variants of annotation, as in *A. thaliana*, or further upstream, as in *C. rubella*, could be found in the Brassicaceae annotations of this gene. The *rpl2* gene of *C. rubella* is longer due to 3 bp and 21 bp long inserts in its sequence compared to the *rpl2* genes of *A. thaliana* and *C. bursa-pastoris*.

Ribosomal RNAs 5S and 18S are identical in sequence between all three analyzed mitochondrial genomes, but sequence of the 26S rRNA of *A. thaliana* differs by three single nucleotide substitutions from its orthologs in *C. bursa-pastoris* and *C. rubella*.

The set of tRNA is different between species, mostly due to the duplication of some tRNA in repeats (e.g., *C. bursa-pastoris* has duplication of the *trnY-GUA* + *trnS-GCU* region, *trnY-GUA*, and *trnK-UUU*). However, despite the differences in the number of tRNA between species, all of them represent the same set of 15 amino acids. According to the BLASTn search of the mitochondrial DNA against chloroplast DNA of *C. bursa-pastoris*, sequences of six of these tRNAs align with more than 95% identity. Additionally, they have been identified by MITOFY and CPGAVAS2 annotation software as the same tRNAs in both genomes, which allowed us to suggest that these tRNAs are of chloroplast origin: *trnS-GGA*, *trnM-CAU*, *trnW-CCA*, *trnD-GUC*, *trnH-GUG*, and *trnN-GUU*. Besides tRNA, inserts with chloroplast gene sequences and intergenic regions have also been found: *psbA*, *rpoB*, *psbD-psbC*, *psaB-psaA*, *rbcl*, and *ycf1*.

and protein domains predicted by the InterProScan web-service, except *orf107* and *orf110*, which have no distinguishable protein features consistent with the results of the BLASTp search (Table 2).

Table 2. RNAseq supported open reading frames (ORFs) of the *C. bursa-pastoris* mitogenome.

Name *	InterProScan Predictions	BLASTp Similarity
<i>orf290</i>	Cytochrome c oxidase, subunit II (<i>cox2</i>) domain with 2 transmembrane regions within	<i>cox2</i>
<i>orf197</i>	1 transmembrane region	<i>atp6</i>
<i>orf107</i>	Nothing found	hypothetical protein
<i>orf161</i>	Member of Protein TIC214 (<i>ycf1</i>) InterPro family. Ycf1 domain and 4 transmembrane regions within the domain	<i>ycf1</i>
<i>orf284</i>	Signal peptide (located 1–17 aa) and 2 transmembrane regions	<i>atp9</i>
<i>orf230</i>	Member of Ribosomal protein L2 (<i>rpl2</i>) InterPro family. Ribosomal_L2 domain and 1 transmembrane region	<i>rpl2</i>
<i>orf263</i>	Member of ATP synthase, F0 complex, subunit C (<i>atp9</i>) InterPro family. ATP synthase, subunit C, isoform a domain and 3 transmembrane regions within the domain	<i>atp9</i>
<i>orf110</i>	Nothing found	hypothetical protein

*—ORFs are named by their amino acid length.

The BLASTn search of the ORF nucleotide sequences against the *C. rubella* and *A. thaliana* mtDNA sequences has shown that only *orf197*, *orf107*, and *orf110* fully align with 100% identity to the *C. rubella* mtDNA. The rest of the ORFs align with more than 90% identity to both *C. rubella* and *A. thaliana* sequences, with better alignment scores for the *C. rubella* mtDNA. Though similar nucleotide sequences of most of the ORFs could be found in mitogenomes of related species, the INDELS and SNPs in *orf284* and *orf290* changed their amino acid sequences significantly, resulting in a frameshift or premature stop codon. Thus, *orf284* and *orf290* could be considered as ORFs specific to *C. bursa-pastoris*.

2.4. RNA Editing

Ribosomal RNA depleted total RNA of *C. bursa-pastoris* was sequenced, generating 43,197,038 reads with an average length of 84 bp. Reads were mapped using the HISAT2 resulting in 28.8% of all reads mapped to the mitogenome sequence. Read alignment analysis has shown that there are 377 RNA editing sites (Table S3), and only C to U transition events have been found. Only 21 of the discovered substitutions are synonymous, while the predominant part of editing events causes amino acid changes in encoded proteins. Most frequently, amino acids have been replaced by the leucine (164 cases), and less by phenylalanine (60 cases). In 15 cases, two editing sites have been identified in the same codon forming non-synonymous amino acid substitution. Editing sites have been found in all protein-coding mitochondrial transcripts, except *cox1* and *atp8*. Ribosomal proteins (except *rps4*) and ATPase subunits have a relatively small number of RNA editing derived substitutions (1–8 sites), while the transcripts of NADH dehydrogenase subunits (*nad1*, *nad2*, *nad4*, *nad5*, *nad7*) and cytochrome c biogenesis genes (*ccmB*, *ccmC*) have been significantly edited (19–30 sites; Figure 3).

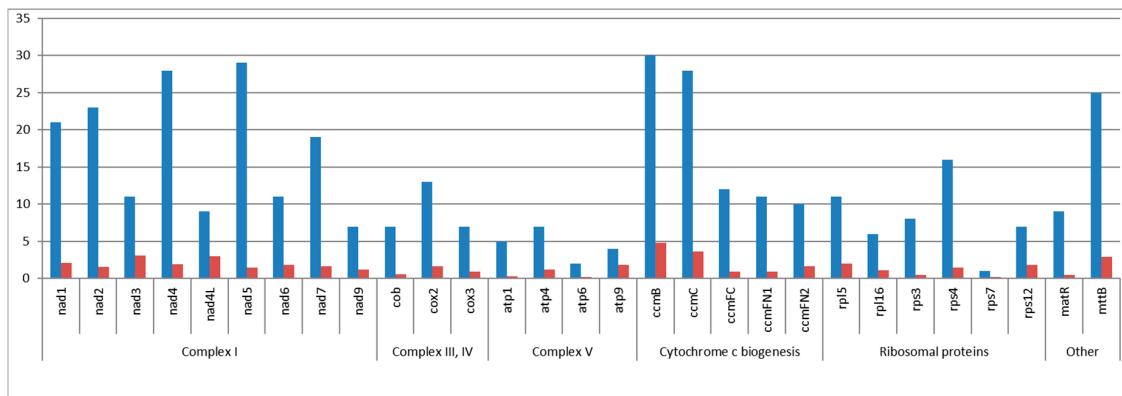


Figure 3. The absolute number of RNA editing substitutions per gene (blue bars) and the relative number of RNA editing substitutions by gene length normalized to 100 bp (red bars).

3. Discussion

Size and GC content of the *C. bursa-pastoris* mitogenome are consistent with the characteristics of the Brassicaceae family mitogenomes (220–368 kbp and 44.74–45.33%, respectively) and almost identical to its closest relative *C. rubella* (287,799 bp vs. 287,405 bp and 44.74% equal for both). The presence of multiple repeats of various lengths is a characteristic feature of the plant mitochondrial DNA. With rare exceptions (e.g., *Silene conica* [7]), the mutation rate in the mitochondrial DNA of Embryophyta is much lower than in nuclear or chloroplast DNA, and diversity in the organization of plant mitochondrial genomes is provided by frequent recombination events at the repeat sites [18,19]. Due to high recombination frequency, plant mitochondrial genomes have a dynamic structure and are represented in mitochondria in various configurations (master ring, sub-rings, linear molecules) and stoichiometry. Most angiosperms have both short (up to 1 kbp) and long (from 1 kbp to more than 10 kbp) repeats in mitochondrial DNA, and Brassicales plants have at least one repeat longer than 1 kbp (except *Batis maritima*) [20]. Interestingly, no repeats longer than 1 kbp have been found in *C. bursa-pastoris*, as well as in *C. rubella*. Recombination across long repeats usually leads to the multipartite organization of the plant mitochondrial genome, where mtDNA isoforms exist in approximately equal stoichiometry. Whereas, recombination between short repeats occurs sporadically, forming low-copy isoforms of rearranged plant mitochondrial DNA (“sublimons”) [5,21]. Given the lack of long repeats and low CCS long-read support for the found structural variants, it could be assumed that the *C. bursa-pastoris* mitogenome is not in a multipartite form and a ~288 kbp long circular molecule is its main conformation.

Two large inversions, the boundaries of which lie within 626 and 865 bp long inverted repeats, distinguish the mitochondrial genome of *C. bursa-pastoris* and *C. orientalis* from *C. rubella*. Due to the maternal inheritance of mitochondria, it could be assumed that *C. orientalis* is the maternal progenitor of the two progenitor species of *C. bursa-pastoris*. This is also indirectly confirmed by the higher coverage and better quality of the mapping of *C. orientalis* reads to the *C. bursa-pastoris* mitochondrial genome. This conclusion is consistent with the conclusions of the phylogenetic studies of the *C. bursa-pastoris* origin based on chloroplast DNA [2,22].

Plant mitochondrial genomes are redundantly large and consist of protein-coding genes, rRNA, and tRNA, interspersed with long intergenic regions, introns of cis- and trans-spliced genes, and many repeat sequences. A common ancestor set of genes of the angiosperms consisted of 41 genes: 24 core genes majorly involved in cellular respiration, 15 ribosomal genes, and succinate dehydrogenase complex subunits *sdh3* and *sdh4* [14,23]. According to the GenBank genome database, most of the flowering plants have preserved almost all the genes of this set in their mitogenomes. In the angiosperms, rRNA is usually represented by three genes—small 18S, large 26S, and 5S subunits, and tRNA varies significantly in number (from 0 to 44) and origin (mitochondrial or chloroplast). In the mitochondrial genome of *C. bursa-pastoris*, 32 protein-coding genes out of 41 have been detected,

which means that the missing ribosomal and succinate dehydrogenase genes most likely have been transferred to the nuclear genome as it frequently happened during the evolution of angiosperms [24]. *Capsella bursa-pastoris* set of genes completely coincides with the set of *C. rubella* and *A. thaliana*, except for the second copy of the *atp6* gene in *A. thaliana*. It is known that plant mitochondrial DNA transfer bits of its sequence to the nuclear DNA (and rarely to the chloroplast DNA), and incorporates some of the nuclear and chloroplast DNA sequences in return. Angiosperms on average have 3–6% of plastid DNA in their mitogenome, of which only tRNAs are functional after transfer [14]. Among 25 tRNA of *C. bursa-pastoris*, six were found to be of chloroplast origin. Many seed plants, including *A. thaliana* and *Brassica napus*, have *trnS-GGA*, *trnM-CAU*, *trnW-CCA*, *trnD-GUC*, *trnH-GUG*, and *trnN-GUU* of chloroplast origin [25]. Including fragments of genes and intergenic spaces, the plastid inserts make 1.3% of the *C. bursa-pastoris* mitochondrial genome, which is close to 1% in *A. thaliana* [14]. Six out of eight ORFs, expression of which supported by RNAseq, in the *C. bursa-pastoris* mitochondrial genome contain transmembrane regions, and four of them are chimeric sequences of *cox2*, *atp6*, and *atp9* genes. ORFs containing chimeric sequences of these genes and transmembrane regions are known to cause cytoplasmic male sterility (CMS) in a wide array of plant species [26,27]. Thus, we suggest that found *C. bursa-pastoris* ORFs could be CMS-associated under the control of nuclear restorer-of-fertility (Rf) genes. Though, this conclusion needs further investigation.

It is notable, with few exceptions, that all editing sites that have been discovered in *C. bursa-pastoris* are the same as in *A. thaliana* mitogenome. Somewhat conservative status of RNA editing in mitochondrial genes has been previously described for Brassicaceae species *B. napus* and *A. thaliana* [28]. While comparing editing conversion of *ccmFN*, *cob*, *mttB*, *nad2*, and *nad4*, several SNPs in *C. bursa-pastoris* have been found in the RNA editing sites of *A. thaliana*. Thus, non-edited triplets in *C. bursa-pastoris* are encoding the same amino acids as the edited ones in *A. thaliana*. These data emphasize the evidence that RNA editing could be considered as the mechanism for restoring conserved codon identities that have been lost on the DNA level [29]. Among the essential editing events, those resulting in the start and stop codons formation should be highlighted. The start codon ATG in *nad1* is formed by RNA editing, which is common for Brassicaceae [15,30,31]. As for the stop codons, alteration in *ccmFC* transcript creates a stop codon TGA as in *A. thaliana*, though for some Brassicaceae, including *C. rubella*, transcript stop is suggested further downstream at the TAA stop codon [15,30,31]. Also, an internal stop codon formation (Gln8Ter) at the N terminus of the protein encoded by *rpl16* has been identified. This C to U conversion assumedly leads to the shortage of a translated region of *rpl16*. In *B. napus*, as well as in *A. thaliana*, RNA editing creates a stop codon at the 21st amino acid from the start, leading to the suggestion of an alternate site of transcription initiation of *rpl16* or its pseudogenization [28].

4. Materials and Methods

4.1. DNA Extraction and Sequencing

Freshly harvested leaves of *C. bursa-pastoris* (the line 'msu-wt' [32]) were quick-frozen in liquid nitrogen and transferred on dry ice to the DNA Link laboratory (South Korea, Seoul) where DNA was extracted from the sample and sequenced using PacBio RS II. High-precision CCS reads were prepared from raw sequencing data by the DNA Link as well as by using the Circular Consensus Sequencing (CCS) application from the SMRTLink v8.0 software with default parameters.

4.2. Mitochondrial and Chloroplast Genome Assembly

Chloroplast reads were filtered out by mapping CCS reads to the *C. bursa-pastoris* chloroplast genome reference from the NCBI GenBank database (RefSeq NC_009270.1) using Minimap2 2.17-r954-dirty [33] with a set of predefined parameters for mapping PacBio CCS reads (" -ax asm5") and keeping only unmapped reads. Then a 10% subsample of the filtered reads was used for the assembly, using Canu v1.8 [34] without read correction (" -assemble -pacbio-corrected correctedErrorRate = 0.005 minOverlapLength = 500 minReadLength = 1000"). Candidate mitochondrial contigs were identified

by sequence identity to *C. rubella* mtDNA (Genbank accession: MH624151.1) using BLASTn 2.9.0 + [35] alignment. The contig with the highest identity and coverage was selected as the primary candidate mtDNA for further analysis.

Chloroplast reads that were filtered out previously were subsampled to 20,000 reads and assembled with Canu v1.9 [34] without read correction. The chloroplast contig was identified using BLASTn search against *C. bursa-pastoris* chloroplast genome reference (RefSeq accession: NC_009270.1). It contained a full long single copy (LSC), short single copy (SSC), and first inverted repeat (IRa) regions and border fragments of the second inverted repeat (IRb) at the ends of the contig. The chloroplast genome was completed and circularized by filling the unassembled IRb fragment between the contig's ends with the corresponding IRa sequence. The assembled chloroplast genome is longer than GenBank reference (RefSeq NC_009270.1) by 63 bp and differs by 84 SNPs and 33 INDELS.

4.3. Repeats and Structural Variation Analysis

The *C. bursa-pastoris* mtDNA was aligned against itself using the online version of MEGABLAST [35], with the e-value threshold set to 10^{-5} to locate repeats. Analysis of structural variants was carried out as described elsewhere [36] by mapping the CCS reads using NGMLR 0.2.7 with the default settings to the mtDNA, and finding the structural variants in the alignment with Sniffles 1.0.11. Structural variants had to be supported by at least 2 reads. The *C. bursa-pastoris* mtDNA was also aligned against assembled cpDNA (assigned GenBank accession MT040199) using BLASTn 2.9.0 + [35] with the default settings to locate inserts of chloroplast origin in the mitochondrial genome.

The sequence of *C. bursa-pastoris* mtDNA was compared with the *C. rubella* mtDNA (RefSeq: NC_042883.1) using Mauve snapshot_2015-02-13 build 0 [37] genome-wide alignment. Additionally, reads from the Russian population of *C. orientalis* (SRA SRR8904471) were mapped to both *C. rubella* and *C. bursa-pastoris* mtDNA to check the presence of identified structural variants in both progenitor species of *C. bursa-pastoris* by mapping quality assessment. Visual analysis of the alignments was performed using IGV 2.7.2 [38].

4.4. RNA Extraction, Sequencing, and Analysis

Total RNA was extracted from the fresh leaf material of *C. bursa-pastoris* (the line 'msu-wt' [32]) using the RNeasy Mini Kit (Qiagen, Hilden, Germany). RNAseq libraries were prepared with Zymo-Seq RiboFree Total RNA Library Kit (Zymo Research, Orange, CA, USA), and single-read sequencing was performed on Illumina NextSeq 500 using the NextSeq 500/550 High Output Kit v2 (75 cycles) (Illumina, Mountain View, CA, USA).

RNAseq analysis was conducted by mapping reads to the *C. bursa-pastoris* annotated mitogenome using HISAT2 2.1.0 [39] and the visual analysis of alignment with the IGV 2.7.2 [38]. RNA editing sites were identified using variant calling results of the mapped data generated by bcftools 1.9 (with settings "mpileup -I -B -d 8000" and then "call -m -V indels -Ov") [40,41] and the REDO script [42] with settings "-d 30 -c 10 -s 0 -a 0". To correctly identify the percent of reads supporting each editing site, all RNAseq reads were mapped to the original mtDNA sequence and modified the mtDNA with all predicted substitutions present in the sequence using CLC Genomics Workbench 9.5.4 (CLC bio, Aarhus, Denmark) with parameters allowing only unique mapping, no mismatches, and 100% of the read length mapped. To calculate read depth at each RNA editing position in both references, samtools 1.9 [43] with "depth" command was used. All editing sites were also checked manually using IGV 2.7.2 [38] and compared with *A. thaliana* editing sites.

4.5. Genome Annotation

Capsella bursa-pastoris mitogenome was annotated using a MITOFY web server [44], and the chloroplast genome was annotated using a CPGAVAS2 web server [45], with subsequent manual verification and, if necessary, correction of the found gene boundaries by comparison of nucleotide and amino acid sequences with corresponding ortholog sequences in *A. thaliana* and *C. rubella*. The annotated

mitochondrial and chloroplast genomes of *C. bursa-pastoris* were submitted to the NCBI GenBank database (assigned accessions MN746809 and MT040199, respectively). ORFfinder NCBI web-service was used to locate ORFs in the mitogenome with a length no less than 300 bp in regions where ORFs expression was supported by RNAseq data (read coverage ≥ 1000 within ORF, while the mitogenome median coverage is 73). The mitochondrial genome map was created with Circos v. 0.69–9 [46].

5. Conclusions

A single circular master molecule represents the mitogenome of *C. bursa-pastoris*. It is 287,799 bp long, contains 32 protein-coding genes, 3 rRNAs, and 25 tRNAs, which coincides with the average length and gene profile of the other known complete mitogenomes of the Brassicaceae species. Investigation of the RNA editing sites in coding regions of the mitogenome using RNAseq data has revealed 377 C to U transitions, most of which are the same editing sites as in *A. thaliana*. Nearly identical edited amino acid sequences of the genes, with both matched and mismatched RNA editing sites in *C. bursa-pastoris* and *A. thaliana*, provide additional evidence of the RNA editing conserved nature in the Brassicaceae species. We have also identified eight new long (more than 300 bp) ORFs with their expression confirmed by RNAseq data (orf107, orf110, orf161, orf197, orf230, orf263, orf284, orf290). Most of them contain transmembrane regions and chimeric sequences of mitochondrial and plastid genes, which resemble characteristics of the CMS-associated genes, though this connection requires further investigation. Analysis of the structural variants of the *C. bursa-pastoris* mitogenome has revealed that a single circular master molecule is its primary form, while other possible structural variants are almost absent in the mitochondria. We also suggest that *C. orientalis* is a maternal progenitor species of *C. bursa-pastoris* based on the presence of two large inversions in the mitogenome of *C. bursa-pastoris* when compared to *C. rubella*, and lack of thereof when compared to *C. orientalis*. Data obtained in the current research could be useful for future investigations associated with the organization of plant mitochondrial DNA and phylogenetic studies of angiosperms and the family Brassicaceae in particular.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2223-7747/9/4/469/s1>, Figure S1: Alignment of all CCS reads to the *C. bursa-pastoris* and *C. rubella* mtDNA sequences, Figure S2: Alignment of the *C. orientalis* Illumina pair-end reads to the *C. bursa-pastoris* and *C. rubella* mtDNA sequences, Table S1: Repeats in the *C. bursa-pastoris* mitogenome, Table S2: Comparison of the gene content between *C. bursa-pastoris*, *C. rubella*, and *A. thaliana* mitogenomes, Table S3: RNA editing analysis results.

Author Contributions: Conceptualization, D.O.O. and M.S.M.; software, D.O.O., M.S.M., A.S.K., and M.I.S.; validation, D.O.O. and M.S.M.; formal analysis, D.O.O., M.S.M., A.S.K., and M.I.S.; investigation, D.O.O., M.D.L., and A.A.P.; methodology, D.O.O., M.S.M., A.S.K., and M.I.S.; resources, M.D.L. and A.A.P.; data curation, D.O.O., M.S.M., and A.S.K.; writing—original draft preparation, D.O.O.; writing—review and editing, M.S.M., M.I.S., M.D.L., and A.A.P.; visualization, D.O.O.; supervision, A.A.P.; project administration, A.A.P.; funding acquisition, A.A.P. All authors have read and agreed to the published version of the manuscript.

Funding: The study was supported by a budgetary subsidy to IITP RAS (The Laboratory of Plant Genomics, Russian Academy of Sciences: 0053-2019-0005).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

- Hintz, M.; Bartholmes, C.; Nutt, P.; Ziermann, J.; Hameister, S.; Neuffer, B.; Theissen, G. Catching a “hopeful monster”: Shepherd’s purse (*Capsella bursa-pastoris*) as a model system to study the evolution of flower development. *J. Exp. Bot.* **2006**, *57*, 3531–3542. [[CrossRef](#)] [[PubMed](#)]
- Han, T.-S.; Wu, Q.; Hou, X.-H.; Li, Z.-W.; Zou, Y.-P.; Ge, S.; Guo, Y.-L. Frequent Introgressions from Diploid Species Contribute to the Adaptation of the Tetraploid Shepherd’s Purse (*Capsella bursa-pastoris*). *Mol. Plant* **2015**, *8*, 427–438. [[CrossRef](#)] [[PubMed](#)]
- Bock, R.; Knoop, V. (Eds.) *Genomics of Chloroplasts and Mitochondria*; Advances in Photosynthesis and Respiration; Springer: Dordrecht, The Netherlands, 2012; ISBN 978-94-007-2919-3.

4. Davila, J.I.; Arrieta-Montiel, M.P.; Wamboldt, Y.; Cao, J.; Hagmann, J.; Shedje, V.; Xu, Y.-Z.; Weigel, D.; Mackenzie, S.A. Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. *BMC Biol.* **2011**, *9*, 64. [[CrossRef](#)] [[PubMed](#)]
5. Maréchal, A.; Brisson, N. Recombination and the maintenance of plant organelle genome stability. *New Phytol.* **2010**, *186*, 299–317. [[CrossRef](#)]
6. Skippington, E.; Barkman, T.J.; Rice, D.W.; Palmer, J.D. Miniaturized mitogenome of the parasitic plant *Viscum scurruloideum* is extremely divergent and dynamic and has lost all nad genes. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E3515–E3524. [[CrossRef](#)]
7. Sloan, D.B.; Alverson, A.J.; Chuckalovcak, J.P.; Wu, M.; McCauley, D.E.; Palmer, J.D.; Taylor, D.R. Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates. *PLoS Biol.* **2012**, *10*, e1001241. [[CrossRef](#)]
8. Fauron, C.; Allen, J.; Clifton, S.; Newton, K. Plant Mitochondrial Genomes. In *Molecular Biology and Biotechnology of Plant Organelles: Chloroplasts and Mitochondria*; Daniell, H., Chase, C., Eds.; Springer: Dordrecht, The Netherlands, 2004; pp. 151–177, ISBN 978-1-4020-3166-3.
9. Kovar, L.; Nageswara-Rao, M.; Ortega-Rodriguez, S.; Dugas, D.V.; Straub, S.; Cronn, R.; Strickler, S.R.; Hughes, C.E.; Hanley, K.A.; Rodriguez, D.N.; et al. PacBio-Based Mitochondrial Genome Assembly of *Leucaena trichandra* (Leguminosae) and an Intrageneric Assessment of Mitochondrial RNA Editing. *Genome Biol. Evol.* **2018**, *10*, 2501–2517. [[CrossRef](#)]
10. Shearman, J.R.; Sonthirod, C.; Naktang, C.; Pootakham, W.; Yoocha, T.; Sangsrakru, D.; Jomchai, N.; Tragoonrung, S.; Tangphatsornruang, S. The two chromosomes of the mitochondrial genome of a sugarcane cultivar: Assembly and recombination analysis using long PacBio reads. *Sci. Rep.* **2016**, *6*, 1–7. [[CrossRef](#)]
11. Dong, S.; Zhao, C.; Chen, F.; Liu, Y.; Zhang, S.; Wu, H.; Zhang, L.; Liu, Y. The complete mitochondrial genome of the early flowering plant *Nymphaea colorata* is highly repetitive with low recombination. *BMC Genom.* **2018**, *19*, 614. [[CrossRef](#)]
12. Gui, S.; Wu, Z.; Zhang, H.; Zheng, Y.; Zhu, Z.; Liang, D.; Ding, Y. The mitochondrial genome map of *Nelumbo nucifera* reveals ancient evolutionary features. *Sci. Rep.* **2016**, *6*, 1–11. [[CrossRef](#)]
13. Liao, X.; Zhao, Y.; Kong, X.; Khan, A.; Zhou, B.; Liu, D.; Kashif, M.H.; Chen, P.; Wang, H.; Zhou, R. Complete sequence of kenaf (*Hibiscus cannabinus*) mitochondrial genome and comparative analysis with the mitochondrial genomes of other plants. *Sci. Rep.* **2018**, *8*, 1–13. [[CrossRef](#)] [[PubMed](#)]
14. Mower, J.P.; Sloan, D.B.; Alverson, A.J. Plant Mitochondrial Genome Diversity: The Genomics Revolution. In *Plant Genome Diversity Volume 1: Plant Genomes, their Residents, and their Evolutionary Dynamics*; Wendel, J.F., Greilhuber, J., Dolezel, J., Leitch, I.J., Eds.; Springer: Vienna, Austria, 2012; pp. 123–144, ISBN 978-3-7091-1130-7.
15. Lin, H.; Bai, D. The complete mitochondrial genome of a highly selfing species *Capsella rubella*. *Mitochondrial DNA Part B* **2019**, *4*, 1907–1908. [[CrossRef](#)]
16. Tao, Y.-T.; Suo, F.; Tusso, S.; Wang, Y.-K.; Huang, S.; Wolf, J.B.W.; Du, L.-L. Intraspecific Diversity of Fission Yeast Mitochondrial Genomes. *Genome Biol. Evol.* **2019**, *11*, 2312–2329. [[CrossRef](#)] [[PubMed](#)]
17. Twyford, A.D.; Ness, R.W. Strategies for complete plastid genome sequencing. *Mol. Ecol. Resour.* **2017**, *17*, 858–868. [[CrossRef](#)] [[PubMed](#)]
18. Palmer, J.D.; Herbon, L.A. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *J. Mol. Evol.* **1988**, *28*, 87–97. [[CrossRef](#)]
19. Wolfe, K.H.; Li, W.H.; Sharp, P.M. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 9054–9058. [[CrossRef](#)]
20. Wynn, E.L.; Christensen, A.C. Repeats of Unusual Size in Plant Mitochondrial Genomes: Identification, Incidence and Evolution. *G3 (Bethesda)* **2018**, *9*, 549–559. [[CrossRef](#)]
21. André, C.; Levy, A.; Walbot, V. Small repeated sequences and the structure of plant mitochondrial genomes. *Trends Genet.* **1992**, *8*, 128–132. [[CrossRef](#)]
22. Douglas, G.M.; Gos, G.; Steige, K.A.; Salcedo, A.; Holm, K.; Josephs, E.B.; Arunkumar, R.; Ågren, J.A.; Hazzouri, K.M.; Wang, W.; et al. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 2806–2811. [[CrossRef](#)]
23. Richardson, A.O.; Rice, D.W.; Young, G.J.; Alverson, A.J.; Palmer, J.D. The “fossilized” mitochondrial genome of *Liriodendron tulipifera*: Ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. *BMC Biol.* **2013**, *11*, 29. [[CrossRef](#)]

24. Adams, K.L.; Qiu, Y.-L.; Stoutemyer, M.; Palmer, J.D. Punctuated evolution of mitochondrial gene content: High and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 9905–9912. [[CrossRef](#)] [[PubMed](#)]
25. Sloan, D.B.; Alverson, A.J.; Štorchová, H.; Palmer, J.D.; Taylor, D.R. Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm *Silene latifolia*. *BMC Evol. Biol.* **2010**, *10*, 274. [[CrossRef](#)]
26. Chase, C.D.; Gabay-Laughnan, S. Cytoplasmic Male Sterility and Fertility Restoration by Nuclear Genes. In *Molecular Biology and Biotechnology of Plant Organelles*; Daniell, H., Chase, C., Eds.; Springer: Dordrecht, The Netherlands, 2004; pp. 593–621, ISBN 978-1-4020-2713-0.
27. Hanson, M.R.; Bentolila, S. Interactions of Mitochondrial and Nuclear Genes That Affect Male Gametophyte Development. *Plant Cell* **2004**, *16*, S154–S169. [[CrossRef](#)] [[PubMed](#)]
28. Handa, H. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* L.): Comparative analysis of the mitochondrial genomes of rapeseed and *Arabidopsis thaliana*. *Nucleic Acids Res.* **2003**, *31*, 5907–5916. [[CrossRef](#)] [[PubMed](#)]
29. Finster, S.; Legen, J.; Qu, Y.; Schmitz-Linneweber, C. Land Plant RNA Editing or: Don't Be Fooled by Plant Organellar DNA Sequences. In *Genomics of Chloroplasts and Mitochondria*; Bock, R., Knoop, V., Eds.; Advances in Photosynthesis and Respiration; Springer: Dordrecht, The Netherlands, 2012; pp. 293–321, ISBN 978-94-007-2920-9.
30. Li, J.; Bi, C.; Tu, J.; Lu, Z. The complete mitochondrial genome sequence of *Boechera stricta*. *Mitochondrial DNA Part B* **2018**, *3*, 896–897. [[CrossRef](#)]
31. Xu, Y.; Bi, C. The complete mitochondrial genome sequence of an alpine plant *Arabis alpina*. *Mitochondrial DNA Part B* **2018**, *3*, 725–727. [[CrossRef](#)]
32. Kasianov, A.S.; Klepikova, A.V.; Kulakovskiy, I.V.; Gerasimov, E.S.; Fedotova, A.V.; Besedina, E.G.; Kondrashov, A.S.; Logacheva, M.D.; Penin, A.A. High-quality genome assembly of *Capsella bursa-pastoris* reveals asymmetry of regulatory elements at early stages of polyploid genome evolution. *Plant J.* **2017**, *91*, 278–291. [[CrossRef](#)]
33. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094–3100. [[CrossRef](#)]
34. Koren, S.; Walenz, B.P.; Berlin, K.; Miller, J.R.; Bergman, N.H.; Phillippy, A.M. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **2017**, *27*, 722–736. [[CrossRef](#)]
35. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [[CrossRef](#)]
36. Sedlazeck, F.J.; Rescheneder, P.; Smolka, M.; Fang, H.; Nattestad, M.; von Haeseler, A.; Schatz, M.C. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **2018**, *15*, 461–468. [[CrossRef](#)] [[PubMed](#)]
37. Darling, A.E.; Mau, B.; Perna, N.T. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE* **2010**, *5*, e11147. [[CrossRef](#)] [[PubMed](#)]
38. Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **2013**, *14*, 178–192. [[CrossRef](#)] [[PubMed](#)]
39. Kim, D.; Paggi, J.M.; Park, C.; Bennett, C.; Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **2019**, *37*, 907–915. [[CrossRef](#)]
40. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **2011**, *27*, 2987–2993. [[CrossRef](#)]
41. Danecek, P.; Schiffels, S.; Durbin, R. Multi-allelic calling model in bcftools (-m). Available online: <https://samtools.github.io/bcftools/call-m.pdf> (accessed on 23 January 2020).
42. Wu, S.; Liu, W.; Aljohi, H.A.; Alromaih, S.A.; Alanazi, I.O.; Lin, Q.; Yu, J.; Hu, S. REDO: RNA Editing Detection in Plant Organelles Based on Variant Calling Results. *J. Comput. Biol.* **2018**, *25*, 509–516. [[CrossRef](#)]
43. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. 1000 Genome Project Data Processing Subgroup The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [[CrossRef](#)]
44. Alverson, A.J.; Wei, X.; Rice, D.W.; Stern, D.B.; Barry, K.; Palmer, J.D. Insights into the Evolution of Mitochondrial Genome Size from Complete Sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Mol. Biol. Evol.* **2010**, *27*, 1436–1448. [[CrossRef](#)]

45. Shi, L.; Chen, H.; Jiang, M.; Wang, L.; Wu, X.; Huang, L.; Liu, C. CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res.* **2019**, *47*, W65–W73. [[CrossRef](#)]
46. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S.J.; Marra, M.A. Circos: An information aesthetic for comparative genomics. *Genome Res.* **2009**, *19*, 1639–1645. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).