Discovery of Four Novel ORFs Responsible for Cytoplasmic Male Sterility (CMS) in Cotton (*Gossypium Hirsutum* L.) through Comparative Analysis of the Mitochondrial Genomes of Four Isoplasmic Lines

Min Li 1, Li Chen 1,2, Danfeng Tang 3, Xiaofang Liao 4, Xiangjun Kong 1, Bin Li 1, Jingyi You 1 and Ruiyang Zhou 1,*

1 Key Laboratory of Plant Genetic and Breeding, College of Agriculture, Guangxi University, Nanning 530007, China; 1517402001@st.gxu.edu.cn (M.L.); nxyc@gxu.edu.cn (L.C.); xiangjunkong@gxu.edu.cn (X.K.); 1617303006@st.gxu.edu.cn (B.L.); 1817401016@st.gxu.edu.cn (J.Y.)
2 National Demonstration Center for Experimental Plant Science Education, College of Agriculture, Guangxi University, Nanning 530007, China
3 Guangxi Key Laboratory of Medicinal Resources Protection and Genetic Improvement, Guangxi Botanical Garden of Medicinal Plants, Nanning 530007, China; tangdf@gxyyzwy.com
4 Cash Crop Research Institute, Guangxi Academy of Agricultural Sciences, Nanning 530007, China; lixiaofang@gxaas.net
* Correspondence: kjczhxxk@gxu.edu.cn; Tel.: +86-18978969180
Received: 12 April 2020; Accepted: 25 May 2020; Published: 27 May 2020

**Abstract:** Cytoplasmic male sterility (CMS) is an important feature for achieving heterosis in the development of hybrid crops. Mitochondria contribute to CMS, especially via mitochondrial DNA (mtDNA) rearrangements and chimeric genes. However, the mechanisms of CMS have not been fully elucidated, and the isonuclear alloplasmic lines used in previous studies have limited utility in cotton CMS research. In this study, three CMS lines (J4A-1, J4A-2 and J4A-3) and their isoplasmic maintainer line (J4B) were analyzed for mtDNA structural differences via high-throughput sequencing. The results showed that mtDNA was conserved (with similarities higher than 99%) among the three CMS lines and their isoplasmic maintainer line. All lines harbored 36 known protein-coding genes, 3 rRNAs, and 15 tRNAs. The protein-coding genes with non-synonymous mutations mainly encoded two types of proteins: ATPase and ribosomal proteins. Four new open reading frames (ORFs) (orf116b, orf186a-1, orf186a-2 and orf305a) were identified as candidate ORFs responsible for CMS. Two of the ORFs (orf186a-1 and orf186a-2) were identified as orf4 and orf4-2 of the upland cotton CMS line 2074A (a line with *Gossypium harknessii* Brandegee CMS-D2-2 cytoplasm), respectively. These findings provide a reference for CMS research in cotton or other crops.

**Keywords:** cotton; cytoplasmic male sterility; mitochondrial genome; ORFs

1. Introduction

Cytoplasmic male sterility (CMS) refers to the loss of stamen function and pollination ability and is very important in the production of crop hybrids, such as rice, corn, and wheat [1–3]. Mitochondria...
are involved in CMS. Previous studies have revealed that CMS results from interactions between mitochondrial genes and nuclear genes [4], and the abnormal development of pollen is usually associated with defects in mitochondria. Mitochondria are essential, semi-autonomously replicating organelles that produce energy and encode for limited genetic information. As with any other organellar genome, plant mitochondrial genomes (mt genomes) depend on highly integrated functional coordination with the nucleus. During mitochondrial evolution, many essential genes are transferred to the host’s nuclear genome or are lost, although the mt genome remains complete [5]. The mt genomes of animals are small, ranging from 15–18 kb, whereas the mt genomes of plants have evolved to vary greatly in size since the initial symbiosis between eukaryotic cells and α-proteobacteria [6]. The size of angiosperm mitochondrial DNA (mtDNA) ranges from 208 kb (Brassica hirta) to 11.3 Mb (Silene conica) [5,7,8]. The large mt genomes in plants contain numerous genes and putative open reading frames (ORFs), many of which are of unknown function [9]. Additionally, higher plants have larger mt genomes sizes, with larger and more complex non-coding and repeat regions, than other organisms [10,11], and the repeated sequences can result in mtDNA rearrangement [12–15].

Chimeric ORFs result from rearranged mt genomes. In flowering plants, chimeric ORFs are thought to block mitochondrial function during critical periods of anther development, leading to the occurrence of CMS [16]. A chimeric gene is a combination of an ORF and a mitochondrial protein-coding gene that usually results in CMS. Previous studies have shown that these coding genes mainly code for ribosomal, cytochrome c oxidase or ATPase [17]. These ORF-encoded proteins usually have the same structure as proteins encoded by mitochondrial genes. In this way, they reduce ATP synthesis or interfere with energy metabolism by acting on complexes or producing toxic proteins, thus affecting pollen fertility. The first identified gene related to CMS was T-urfl3 in maize, with a CMS-T-type cytoplasm; it was formed after 7 recombinations and is composed of a fragment of partial rRM26 and atp6 [18]. The mtDNA expression region associated with CMS in the rape nap-CMS line contains a pol CMS-related orf224 gene. The orf79 gene in the HL-CMS line of rice, which is located downstream of atp6, is composed of a partial fragment of coxl and a fragment of unknown origin; it was formed by the recombination of multiple genes in mtDNA [19]. In a sugar beet sterile line and maintainer line [20], several chimeric ORFs were found to be unique to one of the lines. Such differences may be due to normal differentiation between lines, yet some unique ORFs are thought to cause CMS, and studies have shown that one of these ORFs may be the CMS-related gene of sugar beet [21].

Cotton is one of the world’s leading natural fiber crops and an important oil-producing plant. CMS in cotton was first described in CMS-D2, which contains Gossypium harknessii (G. harknessii) cytoplasm and was bred through distant hybridization and backcrossing. Both upland cotton and island cotton can maintain its sterility [22]. Another CMS line, CMS-D8, was developed by crossing Gossypium trilobum (G. trilobum) with Gossypium hirsutum (G. hirsutum) [23]. A new CMS line, 104-7A, which possesses the cytoplasm of the commercial G. hirsutum variety ‘Shiduan 5’ was bred by interspecific hybridization between G. hirsutum and Gossypium barbadense (G. barbadense) [24]. However, the use of heterosis in cotton breeding is hindered by the limited availability of cotton germplasms with CMS [25,26]. Similar to the cytoplasm of the CMS line 104-7A, the cytoplasm of the materials (J4A-1, J4A-2 and J4A-3) used in the present study is from another upland cotton maintainer line, J4B (a major commercial cultivar developed by Hubei Gushen Science and Technology Co., Ltd., Wuhan, China, in 2010). Although cotton mt genomes were recently sequenced, the genomes of the materials used were alloplastic. A previous analysis of the mt genomes of four lines (the CMS line 2074A (G. harknessii Brandegee, CMS-D2 cytoplasm), 2074S (G. hirsutum L, CMS-AD1 cytoplasm), their maintainer line 2074B (a cultivar of upland cotton, “Sumian No. 20” cytoplasm) and the restorer line E5903 (normal male-fertile G. harknessii Brandegee cytoplasm) revealed 4 ORFs related to CMS [26]. Here, we used the four isoplasmic lines as materials to compare the mtDNA among three CMS lines (J4A-1, J4A-2 and J4A-3) and the maintainer line J4B to determine candidate CMS factors. The
results provide insight into the intricate relationships between mtDNA and CMS generation and may stimulate further research.

2. Materials and Methods

2.1. Plant Materials

The CMS lines J4A-1, J4A-2 and J4A-3 and their maintainer line J4B were provided by the Key Laboratory of Plant Genetics and Breeding, College of Agriculture, Guangxi University. The three CMS lines are mutants of their maintainer line J4B, so the CMS lines and their maintainer line J4B possess almost identical nuclear and cytoplasmic genomes. Plants were cultivated at the Agricultural College of Guangxi University, Nanning city (longitude: 108°22’ N, latitude: 22°48’ N), Guangxi Province, China, in July 2016.

2.2. Sequencing and Assembly of Mitochondrial (mt) Genomes

Approximately 5 g fresh cotton leaf was collected and mtDNA was extracted using an improved method [27]. Then, 1 µg purified mtDNA was fragmented, and a short (350 bp) insertion library was constructed following Illumina’s standard procedure. The cotton mt genomes were sequenced using a high-throughput sequencing platform (Illumina HiSeq 4000, Shanghai Biozeron Co., Ltd., Shanghai, China). Before assembly, SPAdes 3.10.1 [28] was used to estimate genome size based on K-mer statistical analysis. SOAP aligner software (with the default parameters -m 200 -x 600 -i 8 -p 8) was used to compare the reads of the samples to the mt reference genome of upland cotton [29] (GenBank ID: JX065074.1) and to summarize the coverage of the reads on the mt reference genome, the coverage statistics and all additional sequencing data. Then, SOAPdenovo [30] (version 2.04) software with the default parameters (-K 121 -F -p 8) was used to assemble the sequences. Comparing the reads back to the contig obtained by the assembly and according to the relationship of the paired-end and overlap, the results were assembled and optimized locally, and then the order and direction of the contig were determined. Finally, GapCloser (version 1.12) software was used to repair the gaps in the assembly and remove the redundant segment sequences. Employing OGDRAW v1.2 [31] software, mt genome maps (as circular molecules) of the four lines were constructed.

2.3. Analysis and Annotations of the mt Genomes

The mitochondrial genes were annotated using Mitofy and AUGUSTUS software [32], and Evidence Modeler v1.1.1 was used to integrate the gene set [33]. Ribosomal RNA (rRNA) and transfer RNA (tRNA) genes were analyzed using tRNAscan-SE and Glimmer 1.2, respectively [34–37]. Repeats were assessed by RepeatMasker, and the screening criteria for long repeats were as follows: length ≥100 bp and identity greater than or equal to 97%. Using the chloroplast genome and nuclear genome of G. hirsutum as the reference, Nucleotide BLAST (BLASTN) was used to search the chloroplast source sequence and nucleus-derived insertions (identity ≥ 97% and length ≥ 100 bp). ORFs were predicted using ORFfinder and EMBOSS (6.3.1: getorf) [38,39]. Transmembrane domains of ORFs were identified by TMHMM Server version 2.0 [40].

2.4. Detection and Annotation of Single Nucleotide Polymorphisms (SNPs)

Single nucleotide polymorphisms (SNPs) were evaluated by SAMtools software [41] using two screening criteria: mapping quality >20 and variant position depth >4. SNPs in repeat regions were not included. The locations of SNPs were based on the annotation of gene models from the reference genome database [42]. SNPs were classified as exonic SNPs and intronic SNPs depending on their location. SNPs located in coding sequence regions were further divided into synonymous and non-synonymous mutations by GeneWise [43].
3. Results

3.1. Sequencing Data Statistics and Assembly

The original amount of data (Supplementary Table S1) obtained for the CMS lines (J4A-1, J4A-2 and J4A-3) and their maintainer line J4B was 1116 Mb, 1383 Mb, 1427 Mb and 1417 Mb, respectively. The GC content of the sequences of J4A-1, J4A-2, J4A-3 and J4B after filtering was 40%–42%, and the Q20 values were >95%. The K-mer number of the four lines was 19,316 kb–21,549 kb. These values indicate that the constructed database and sequences were suitable for subsequent mt genome assembly and bioinformatics analysis.

3.2. Structures and Contents of the mt Genomes

The mt genomes of J4A-1 National Center for Biotechnology Information (NCBI) accession: MN149527, J4A-2 (NCBI accession: MN149528), J4A-3 (NCBI accession: MN149526) and J4B (NCBI accession: MN149525) (the accession numbers in GenBank are not yet available) were each assembled into a single circular molecule with a total length of 623,067 bp, 623,343 bp, 622,848 bp and 621,841 bp, respectively (Figure 1 and Supplementary Figures S1–S3), and the mt genome sizes of the CMS lines were larger than the genome size of their maintainer line (Table 1). The number of predicted genes differed among J4A-1, J4A-2, J4A-3 and J4B, being 191, 189, 191 and 186, respectively. Additionally, the total length of genes in each CMS line was greater than that in the maintainer line J4B, and intergenic sequences accounted for approximately 83% of the genome sequence in all lines.

![Figure 1](image-url)  
**Figure 1.** Physical map of the mt genomes of the cytoplasmic male sterility (CMS) line J4A-3. Note: The outer circle represents the physical map scaled in kilobase pairs. Different colors indicate different genes as defined in the legend shown in bottom left of the figure: yellow, NADH dehydrogenase; green, succinate dehydrogenase; light green, ubiquinol cytochrome c reductase; pink, cytochrome c oxidase; chartreuse, ATP synthase; tan and brown, ribosomal proteins; orange, maturase; purple,
other genes; turquoise, open reading frames (ORFs); dark blue, transfer RNAs; dark red, ribosomal RNAs.

Table 1. The assembled features of the mt genomes.

| Genome Characteristic          | J4A-1  | J4A-2  | J4A-3  | J4B   |
|--------------------------------|--------|--------|--------|-------|
| Genome size (bp)               | 623,067| 623,343| 622,848| 621,841|
| Gene number (#)                | 191    | 191    | 191    | 186   |
| Gene total length (bp)         | 102,957| 101,838| 103,416| 101,091|
| Gene length/Genome (%)         | 16.52  | 16.34  | 16.6   | 16.3  |
| Intergenic region length       | 520,110| 521,505| 519,432| 520,750|

A total of 36 protein-coding genes, 3 rRNAs, and 15 tRNAs were predicted in the four lines (Table 2). Among the 36 protein-coding genes, multiple copies were detected for 3 (nad1 with 4 copies, nad2 with 2 copies, and nad5 with 3 copies), and 20 genes were identified as being involved in the oxidative phosphorylation system and electron transport. The genes involved in the electron transport system comprised 9 NADH-ubiquinone oxidoreductase (complex I) genes (nad1, nad2, nad3, nad4, nad4L, nad5, nad6, nad7 and nad9), two succinate dehydrogenase (complex II) genes (sdh3 and sdh4), one cytochrome bc1 complex (complex III) gene (cob), three complex IV cytochrome c oxidase genes (cox1, cox2 and cox3), and five ATP synthase genes (atp1, atp4, atp6, atp8, and atp9). Moreover, 10 genes were identified that encode ribosomal proteins, including 6 ribosomal small subunit protein genes (rps3, rps4, rps7, rps10, rps12 and rps14) and four ribosomal large subunit protein genes (rpl2, rpl5, rpl10 and rpl16). Among the ribosomal proteins, only rps3 and rps10 had a single intron. In addition, four cytochrome c biological origin-related genes (ccmB, ccmC, ccmFC and ccmFN) were identified. One maturase gene (matR) and one mttB gene coding for a transporter were identified.

Table 2. Gene contents of cotton mitotypes.

| Product Group | Gene | J4A-1 | J4A-2 | J4A-3 | J4B |
|---------------|------|-------|-------|-------|-----|
| Complex I     | nad1 | +4    | +4    | +4    | +4  |
|               | nad2 | +2    | +2    | +2    | +2  |
|               | nad3 | +     | +     | +     | +   |
|               | nad4 | +     | +     | +     | +   |
|               | nad5 | +3    | +3    | +3    | +3  |
|               | nad6 | +     | +     | +     | +   |
|               | nad7 | +     | +     | +     | +   |
|               | nad9 | +     | +     | +     | +   |
| Complex II    | sdh3 | +     | +     | +     | +   |
|               | sdh4 | +     | +     | +     | +   |
| Complex III   | cob  | +     | +     | +     | +   |
|                      |   |   |   |   |
|----------------------|---|---|---|---|
| **Complex IV**       | cox1 | + | + | + | + |
|                      | cox2 | + | + | + | + |
|                      | cox3 | + | + | + | + |
| **Complex V**        | atp1 | + | + | + | + |
|                      | atp4 | + | + | + | + |
| **Cytochrome C**     | ccmB | + | + | + | + |
|                      | ccmC | + | + | + | + |
|                      | ccmFN| + | + | + | + |
|                      | ccmFC| + | + | + | + |
| **Other gene**       | mttB | + | + | + | + |
|                      | matR | + | + | + | + |
| **Ribosome**         | rps3 | + | + | + | + |
|                      | rps4 | + | + | + | + |
|                      | rps7 | + | + | + | + |
|                      | rps10| + | + | + | + |
|                      | rps12| + | + | + | + |
|                      | rps14| + | + | + | + |
|                      | rpl2 | + | + | + | + |
|                      | rpl5 | + | + | + | + |
|                      | rpl10| + | + | + | + |
|                      | rpl16| + | + | + | + |
| **Total protein-coding genes** | 36 | 36 | 36 | 36 |
| trnC(GCA)-cp         | + | + | + | + |
| trnD(GLC)-cp         | +2 | +2 | +2 | +2 |
| trnE(UUC)            | + | + | + | + |
| **tRNA**             | trnF(GAA) | +2 | +2 | +2 | +2 |
|                      | trnFM(CAU)-cp | +4 | +4 | +4 | +4 |
|                      | trnG(GCC)    | + | + | + | + |
|                      | trnH(GUG)-cp | + | + | + | + |
3.3. Repeat Sequences in the mt Genome

Repeats are DNA sequences with multiple copies in the genome. Previous studies showed that some repeat sequences were involved in rearrangements of the mt genome, which can also result in CMS [2]. The lengths of the repeats in J4A-1, J4A-2, J4A-3 and J4B (length ≥100 bp, identity ≥97%) varied from 100 bp to 27,501 bp (Table 3 and Supplementary Tables S2–S5), and no significant differences in the number of repeat sequences among the four lines were observed. The total length of the repeat sequences in J4A-1, J4A-2, J4A-3 and J4B accounted for 10.60%, 10.77%, 12.60% and 10.90%, respectively, of the total genome length. Among the large (>500 bp) repeats, a 9284 bp repeat sequence (A3R5) was found only in J4A-3. In addition, A3R5 is a copy of A3R4 with a 464 bp deletion at the 5' end. The mtDNA of J4A-3 contains the 9748 bp repeat sequence A3R4, whereas in the other three lines, this sequence is separated into two sequences (8898 bp and 906 bp, 8926 bp and 906 bp, and 8915 bp and 906 bp) by an unknown sequence. The total repeat sequence lengths of J4A-1, J4A-2, J4A-3 and J4B were 66 kb, 67 kb, 75 kb, and 67 kb, respectively, with copy numbers varying from 2 to 4. There were three large repeats that exceeded 10 kb in all lines. Other small repeats were distributed throughout the genome and variable in copy number (Figure 2). The distributions of repeat sequence sizes of J4A-1, J4A-2, J4A-3 and J4B were approximately the same, and fragments 100–200 bp in length were the most common, followed by those with a length of 200–300 bp.

| tRNA genes      | J4A-1 | J4A-2 | J4A-3 | J4B | Total |
|-----------------|-------|-------|-------|-----|-------|
| trnK(UUU)       | +     | +     | +     | +   | +     |
| trnM(CAU)       | +     | +     | +     | +   | +     |
| trnL(UAU)       | +     | +     | +     | +   | +     |
| trnN(GUU)-cp    | +     | +     | +     | +   | +     |
| trnP(UGG)       | +3 a  | +3 a  | +3 a  | +3 a| +3 a  |
| trnQ(UUG)       | +     | +     | +     | +   | +     |
| trnS(GGA)-cp    | +     | +     | +     | +   | +     |
| trnS(GCU)       | +2 a  | +2 a  | +2 a  | +2 a| +2 a  |
| trnS(UGA)       | +     | +     | +     | +   | +     |
| trnSup(UUA)     | +     | +     | +     | +   | +     |
| trnV(GAC)       | +     | +     | +     | +   | +     |
| trnW(CCA)-cp    | +2 a  | +2 a  | +2 a  | +2 a| +2 a  |
| trnY(GLIA)      | +     | +     | +     | +   | +     |

| rRNA genes      | J4A-1 | J4A-2 | J4A-3 | J4B | Total |
|-----------------|-------|-------|-------|-----|-------|
| rRNA_5S         | +     | +     | +     | +   | +     |
| rRNA_18S        | +     | +     | +     | +   | +     |
| rRNA_26S        | +2 a  | +2 a  | +2 a  | +2 a| +2 a  |

Note: + denotes presence; * multiple copies of the gene.

| Total rRNA genes | 4 | 4 | 4 | 4 |          |

---

3.3. Repeat Sequences in the mt Genome

Repeats are DNA sequences with multiple copies in the genome. Previous studies showed that some repeat sequences were involved in rearrangements of the mt genome, which can also result in CMS [2]. The lengths of the repeats in J4A-1, J4A-2, J4A-3 and J4B (length ≥100 bp, identity ≥97%) varied from 100 bp to 27,501 bp (Table 3 and Supplementary Tables S2–S5), and no significant differences in the number of repeat sequences among the four lines were observed. The total length of the repeat sequences in J4A-1, J4A-2, J4A-3 and J4B accounted for 10.60%, 10.77%, 12.60% and 10.90%, respectively, of the total genome length. Among the large (>500 bp) repeats, a 9284 bp repeat sequence (A3R5) was found only in J4A-3. In addition, A3R5 is a copy of A3R4 with a 464 bp deletion at the 5' end. The mtDNA of J4A-3 contains the 9748 bp repeat sequence A3R4, whereas in the other three lines, this sequence is separated into two sequences (8898 bp and 906 bp, 8926 bp and 906 bp, and 8915 bp and 906 bp) by an unknown sequence. The total repeat sequence lengths of J4A-1, J4A-2, J4A-3 and J4B were 66 kb, 67 kb, 75 kb, and 67 kb, respectively, with copy numbers varying from 2 to 4. There were three large repeats that exceeded 10 kb in all lines. Other small repeats were distributed throughout the genome and variable in copy number (Figure 2). The distributions of repeat sequence sizes of J4A-1, J4A-2, J4A-3 and J4B were approximately the same, and fragments 100–200 bp in length were the most common, followed by those with a length of 200–300 bp.
Table 3. Lengths and percentages of repeats (≥100 bp).

| Genome  | Ge-Len (bp) a | Re-Len (bp) b | % of Genome | Min-Len (bp) c | Max-Len (bp) d | No. e | Ge-len-without Dup f (bp) (%) |
|---------|---------------|---------------|-------------|----------------|----------------|-------|-----------------------------|
| J4A-1   | 623,067       | 66,058        | 10.60       | 100            | 27,501         | 39    | 557,009 (89.40)             |
| J4A-2   | 623,343       | 67,156        | 10.77       | 104            | 27,501         | 38    | 556,18 (89.23)              |
| J4A-3   | 622,848       | 75,134        | 12.60       | 100            | 27,501         | 37    | 547,59 (87.92)              |
| J4B     | 621,841       | 67,514        | 10.90       | 105            | 27,362         | 37    | 554,19 (89.12)              |

Note: a, Genome length; b, Repeat sequence length; c, Minimal length; d, Maximal length; e, Number of repeat sequences; f, Genome length without repeat sequences; Per, percentage.

Figure 2. The number of repeat sequences of different sizes. Note: The X axis shows the repeat size category: more than 10 kb, 1–10 kb, 0.5–1 kb, 300–500 bp, 200–300 bp, or 100–200 bp. The Y axis (primary axis) shows the number of repeat pairs.

3.4. Similarity of mtDNA with the Cotton Chloroplast and Nuclear Genomes

Insertions of DNA from other genomes, i.e., the chloroplast and nuclear genomes, into mtDNA have been reported for almost all completely sequenced plant genomes. In the present study, no difference in the chloroplast-like sequences was observed among J4A-1, J4A-2, J4A-3 and J4B (Supplementary Table S6). The total length of three sequence fragments (cp1, cp2, and cp3) was 7116 bp, accounting for 1.14% of the total genome size. Two chloroplast-like sequences, namely, cp1 and cp2, were similar to a hypothetical protein of G. hirsutum. In contrast, cp3 had no similar sequences in the NCBI database.

In addition to containing the plastid genome, plant mtDNA includes large amounts of fragments that potentially derive from the nuclear genome. The total length of nucleus-derived sequences in J4A-1, J4A-2, J4A-3 and J4B was 495,108 bp, 493,822 bp, 495,002 bp and 498,095 bp, accounting for
79.46%, 79.22%, 79.47% and 80.10% of the total genome size, respectively (Supplementary Table S7). Among the four cotton lines, the distribution of nuclear-like sequences among the 26 chromosomes was uneven, and the chromosome to which longest sequence aligned was chromosome 15, with a sequence length of 271.4–274.0 kb, accounting for approximately 43.57%–44.06% of the total chromosome length. In contrast, the shortest sequence aligned to chromosome 14, with a length of 4.53–4.63 kb, accounting for 0.73%–0.74% of the total chromosome length (Figure 3).

**Figure 3.** Distribution of nucleus-derived sequences on the chromosome. Note: The X axis shows the chromosomes of cotton. The Y axis shows the lengths of nucleus-derived sequences.

### 3.5. Unique Open Reading Frames (ORFs) of J4A mtDNA

A total of 155, 153, 155 and 150 ORFs with a length ≥300 bp were found in J4A-1, J4A-2, J4A-3 and J4B, respectively. Among these ORFs, we focused on the unique ORFs of the CMS lines compared with the maintainer line J4B (Supplementary Tables S8 and S9 and Table 4). The number of ORFs detected relative to the CMS lines J4A-1, J4A-2 and J4A-3 was 18, 15 and 18, respectively. Thirteen of these unique ORFs (orf106e, orf111f, orf116b, orf118b, orf123c, orf123f, orf129a, orf156a, orf208a-1, orf208a-2, orf270b and orf592a) were shared among the three CMS lines. J4A-1 and J4A-2 shared orf109b. The ORFs orf119d, orf138b, orf317a and orf456a were unique to J4A-1, orf115b was unique to J4A-2, and orf126b, orf141a, orf186a-1, orf186a-2 and orf305a were unique to J4A-3. Ten ORFs (orf116b, orf126b, orf138b, orf186a-1, orf186a-2, orf208a-1, orf208a-2, orf228a, orf305a and orf317a) were situated in 25 specific mtDNA repeated regions (AR4, AR6, AR8, AR9, AR10, AR12, AR13, AR15, AR18, AR20 and AR37) of the three CMS lines. The ORF orf111f in J4A-3 mtDNA, homologous to Arabidopsis thaliana chloroplast DNA, was unique. This indicated that orf111f may have derived from an exogenic chloroplast genome. The orf129a of J4A-3 exhibited homology to the orf174 of Batis maritima mtDNA.
## Table 4. Unique ORFs (>300 bp) in the mt genome of the CMS line J4A-3.

| ORF     | Len a (bp) | Location          | Tra-Dom b | ORF-Seq c | Homologous Sequence d |
|---------|------------|-------------------|-----------|-----------|-----------------------|
| orf106e | 318        | 111193:111510     | 0         | -         | -                     |
| orf11f  | 333        | 188666:188998     | 0         | -         | -                     |
| orf116b | 348        | 534858:535205     | 2         | 193 bp upstream of rpl5; 41 bp overlap with rpl2; partial A1R22, A1R39 |
| orf118b | 354        | 429729:430082     | 0         | -         | -                     |
| orf123c | 369        | 21393:21761       | 0         | -         | -                     |
| orf123f | 369        | 403911:404279     | 0         | -         | -                     |
| orf126b | 378        | 279972:280349     | 0         | A3R8      | -                     |
| orf129a | 378        | 566270:566656     | 0         | -         | 76–366, 84%, orf174 (mitochondrion), Batis maritima |
| orf141a | 423        | 259437:259859     | 1         | -         | 396 bp, sdl4; 69 bp overlap with cox3 |
| orf156a | 468        | 499754:500221     | 0         | -         | -                     |
| orf186a-1 | 558      | 226860:227417     | 1         | 508 bp downstream of atp8; A3R9, A3R10, A3R12, partial A3R4, A3R6, A3R20, 1–57 bp, 92%, rps3; 69–171 bp, 82%, sdl3 |
| orf186a-2 | 558      | 532727:533284     | 1         | 389 bp upstream of rpl2; 56 bp downstream of nad1 (CDS5), A3R9; partial A3R6, A3R20, 1–57 bp, 92%, rps3; 69–171 bp, 82%, sdl3 |
| orf208a-1 | 624      | 613583:614206     | 3         | A1R7, A1R8, partial A1R1, A1R15, - |
| orf208a-2 | 624      | 456201:456824     | 3         | A1R30     | -                     |
| orf228a | 684        | 261023:261706     | 0         | 304 bp downstream of cox3, 433 bp upstream of cox1, A1R27, - |
| orf270b | 810        | 387038:387847     | 0         | -         | -                     |
| orf305a | 915        | 279431:280345     | 3         | 130 bp downstream of atp1, A3R8, 4–45 bp, 88%, rpl2 |
| orf592a | 1776       | 508363:510138     | 4         | 1734 bp overlap with atp1 | 1734 bp, ccmFN |

Note: Len a, ORF length; Tra-dom b, Transmembrane domain; ORF/R-seq c, Relation of ORFs or repeat sequences; Homologous sequence d, Homologous sequence containing the fragment from cotton and mtDNA from other plants. The ORFs were named after the number of amino acids: except for orf174 of Batis maritima, names follow the conventions orfxx a/b/c, where xx represents the amino acid
number and a/b/c represents different genes with the same length, or orfxx-1/2/3, where xx represents the amino acid sequence length of the ORF and -1/2/3 represents repeat genes with the same length.

ORFs close to known genes and consistently transcribed in the direction with up- or downstream genes may be cotranscripts associated with CMS. Since these new ORFs may be related to CMS, we investigated their location, origin, conservation and function, and categorized the unique ORFs into four groups. The first group included ORFs located up- or downstream of an adjacent gene that were likely cotranscripts (orf115b, orf116b, orf129a, orf138b, orf186a-1, orf186a-2, orf228a, orf305a and orf317a). The second group included unique ORFs with partial homology to fragments of chloroplast DNA and mtDNA from other plants, such as orf111f and orf129a. The third group included ORFs composed of homologous sequences of protein-coding genes, such as orf141a, orf186a-1, orf186a-2 and orf592a. The fourth group included eight ORFs with transmembrane domains: orf116b, orf141a, orf186a-1, orf186a-1, orf208a-1, orf208a-2, orf305a and orf392a.

Orf116b was found 193 bp downstream of rpl5, to have a 41 bp overlap with rpl2, and to contain partial repeats and two transmembrane domains (Figures 4 and 5a,b). Orf186a-1 was found to be 508 bp downstream of atp8 and contained sequences homologous to rps3 (1–57 bp, 92%) and sdh3 (69–171 bp, 82%), and it had one transmembrane domain. Additionally, three repeat sequences (A3R9, A3R10 and A3R12) and three incomplete repeat sequences (A3R4, A3R6 and A3R20) were found in orf186a-1 (Figures 4 and 5c). Orf186a-2 was located 389 bp upstream of rpl2 and 56 bp downstream of nad1 (CDS5) and contained partial sequences of AR9, AR6 and AR19 (Figures 4 and 5d). Orf305a, with five transmembrane domains and AR8, was located 130 bp downstream of atp1 and contained a sequence homologous to rpl2 (4–45 bp, 88% homology) (Figures 4 and 5e).

Figure 4. The probability of transmembrane domains of orf116b, orf186a-1/-2 (the sequences of orf186a-1 and orf186a-2 were the same and had the same transmembrane domain) and orf305a.
3.6. Statistical Analysis and Annotation of SNPs

The SNP variant sites of J4A-1, J4A-2, J4A-3 and J4B mtDNA were obtained by comparison with the reference sequence of the mt genome of upland cotton. In this study, 15 SNPs in 13 protein-coding genes were identified in the three CMS lines (with the SNPs being identical among the three CMS lines) relative to their maintainer line J4B. Five genes (apt4, cox1, nad2, nad7 and rpl16) contained synonymous SNPs, and 8 genes (atp8, cox3, matR, rpl2, rpl5, rpl10, rps4 and sdh3) contained non-synonymous SNPs (Table 5). The SNPs of rpl2 were the most variable and included two non-synonymous mutations. However, the other genes had only one SNP. Moreover, all of the SNPs were transversions and most were A–C conversions, accounting for 80% of the total SNPs. Interestingly, compared with their maintainer line, the CMS lines contained more variations in ribosomal protein-coding genes, such as rpl2, rpl5, rpl10, rpl16 and rps4. In addition, evolutionary rate analysis revealed that the non-synonymous mutation (dN) rate for 6 protein-coding genes (atp4, cox1, cox2, ccmFC, nad4 and rps4).
and **nad7** was lower than the synonymous mutation (dS) rate, making dS/dN greater than 1 and indicating positive selection on these genes [44]. Additionally, the dS/dN ratios of four genes (**cox3**, **rpl2**, **matR** and **nad3**) were less than 1, indicating purifying selection (Supplementary Table S10).

**Table 5.** The protein variation in the two mitogenomes.

| Gene | Len | Var | Loc | J4A-1 | J4A-2 | J4A-3 | J4B | NSM | SM | aa-Var | SNP Type |
|------|-----|-----|-----|-------|-------|-------|-----|-----|----|--------|---------|---------|
| atp4 | 585 | 1   | 222 | ttT   | ttT   | ttT   | ttC | 0   | 1  |        |         |
| atp8 | 465 | 1   | 171 | agC   | agA   | agA   | agA | 1   | 0  |        | Ser-Arg |
| cox1 | 1593| 2   | 960 | atA   | atA   | atA   | atC | 0   | 2  |        |         |
| cox3 | 798 | 1   | 157 | Atc   | Atc   | Atc   | Ctc | 1   | 0  |        | Leu-Ile |
| matR | 1967| 1   | 1858| Aaa   | Aaa   | Aaa   | Caa | 1   | 0  |        | Gln-Lys |
| nad2 | 1467| 1   | 1227| atC   | atC   | atC   | atA | 0   | 1  |        |         |
| nad7 | 1185| 1   | 24  | atC   | atC   | atC   | atA | 0   | 1  |        |         |
| rpl2 | 1005| 2   | 45  | ttG   | ttG   | ttG   | ttT | 2   | 0  |        | Phe-Leu |
| rpl5 | 582 | 1   | 139 | Caa   | Caa   | Caa   | Aaa | 1   | 0  |        | Lys-Gln |
| rpl10| 489 | 1   | 361 | Aaa   | Aaa   | Aaa   | Gaa | 1   | 0  |        | Glu-Lys |
| rpl16| 435 | 1   | 270 | gtC   | gtC   | gtC   | gtA | 0   | 1  |        |         |
| rps4 | 1098| 1   | 535 | Caa   | Caa   | Caa   | Aaa | 1   | 0  |        | Lys-Gln |
| sdh3 | 435 | 1   | 33  | ttC   | ttC   | ttC   | ttA | 1   | 0  |        | Leu-Phe |

| Total number of non-synonymous mutations | 9 |
| Total number of synonymous mutations    | 5 |

Note: Len, Length of the gene coding sequence; Var, Variant sites in two mitogenomes; Loc, Location of variant sites; NSM, Non-synonymous mutation; SM, Synonymous mutation; aa-Var, Amino acid variation.
4. Discussion

Previous studies of CMS in cotton utilized genetic backgrounds with different cytoplasms at the mt genome level, such as 2074A and 2074S [25], for comparative analyses. Although these studies identified some meaningful differences, the materials used were alloplasmic lines. In the present study, the CMS lines and their maintainer line were isogenic, with almost the same mtDNA and nuclear genome. Therefore, the four lines used herein are ideal materials for investigating candidate CMS mitochondrial genes in cotton.

4.1. Characteristics of Plant Mitochondrial Genes

Unlike the chloroplast genome, which is very stable, the mt genome varies greatly, even at the subspecies level [20, 45–47]. However, the mt genomes of J4A-1, J4A-2, J4A-3 and J4B were almost identical, with similarities higher than 99%, confirming the highly consistent genetic background of these four lines. The difference in mt genome size among the four lines was only approximately 1 kb, and all four mt genomes contained 36 genes.

Repeated sequences are identical to symmetrical fragments that occur at different locations in the genome and include forward repeat sequences and palindromic repeat sequences. Repeats are ubiquitous in plant mitochondria and show strong polymorphism. Moreover, mt genome sizes in plants are large because of the presence of long repeats [37–39]. Generally, larger genomes consist of larger numbers of repeat sequences. However, this is not always the case. For example, the total mt genome size Cucurbita is 371 kb, but 38% of the genome comprises repeat sequences [48], while Vitis has a relatively large mt genome (approx. 773 kb), only 7% of which comprises repeat sequences [49]. Here, the total size of repeats in J4A-1, J4A-2, J4A-3 and J4B was 66,058 bp, 67,156 bp, 75,134 bp and 67,514 bp, representing 10.60%, 10.77%, 12.60% and 10.90%, respectively, of the genome. Among the four lines, J4A-3 had the largest proportion of repetitive sequences. In addition, the repeat sequences of the four lines were highly homologous (having similarities greater than 95%), with higher homology than that of non-homologous cytoplasmic materials (2074A, 2074S and E5903) [26].

Genes of the chloroplast genome and nuclear genome, even from other plants, can be transferred horizontally to the mt genome, and horizontal gene migration is the major route for obtaining foreign genes [49–51]. However, most genes that migrate to the mt genome are non-functional. In the present study, three chloroplast-like sequences (cp1, cp2 and cp3) in all four lines were found to be non-functional.

4.2. ORFs Associated with Cytoplasmic Male Sterility (CMS)

In many plants, CMS is associated with abnormal ORFs in the mitochondria, and the occurrence of CMS in cotton plays an important role in cotton breeding. In some plants, a common feature of ORFs associated with CMS is a location upstream or downstream of a protein-coding gene and cotranscription with the gene, which leads to improper functional transcription and translation [3, 52–56]. The location, origin, conservation, and function of predicted ORFs were analyzed in this study because new ORFs may be related to CMS. Our analysis revealed four novel chimeric ORFs (orf116b, orf186a-1, orf186a-2 and orf305a), that contain transmembrane domains and are located near known genes. These ORFs are thus candidate ORFs involved in CMS and are similar to the CMS-related ORFs in other CMS lines, such as S-orf355/orf77 [5], orf224 of rape [57–59], orf256 of wheat [60], orf125 of radish [52], and orf4 and orf4-2 of cotton [26]. In this study, two of four CMS-associated ORFs (orf186a-1 and orf186a-2) showed 99% identity with orf4 and orf4-2 of cotton CMS line 2074A (Supplementary Figure S4), which further supports a potential CMS-related role of orf186a-1 and orf186a-2. Orf116b is common and specific to the three sterile lines; therefore, it may also play a role in CMS. Although chimeric ORFs are common in plant mt genomes, their potential functions at specific locations in specific plants have not been explored. Therefore, to determine whether the products of these genes cause CMS, further characterization is necessary. The current results are
based only on genomic data, and further analysis, such as overexpression and knockout of these genes, should be performed.

4.3. SNPs and Plant Mitochondrial DNA Evolve

Although plant mt genomes change rapidly in structure, the mutation rate is very low [61]. SNPs are mutations involving single nucleotides in the genome and are highly polymorphic. Compared with that of nuclear genes, the mutation rate of coding sequences (CDSSs) in plant mitochondrial genes is relatively low [62]. Essential genes in the plant mt genome are highly conserved and exhibit a very low evolutionary rate. The present study identified 15 non-synonymous SNPs in 13 protein-coding genes of the CMS lines, which is generally consistent with the findings of previous research [26] and implies that these non-synonymous substitutions play important roles in CMS.

5. Conclusions

In summary, the materials used in this study are new CMS lines. Through the analysis of mt genomes, we found 36 protein-coding genes, 3 rRNAs, and 15 tRNAs in the four lines, which were largely identical to those in the reference mt genome of G. hirsutum (JX065074.1). SNP analysis revealed 15 SNPs in 13 protein-coding genes, including 9 non-synonymous mutations and 6 synonymous mutations. Most importantly, there were 18, 15 and 18 ORFs in the J4A-1, J4A-2 and J4A-3 CMS lines, respectively, which were absent in J4B. Among these ORFs, 4 showed characteristics related to CMS, and 2 were consistent with the results of previous studies. These results can inform future research on CMS in cotton and other species. However, whether these four ORFs are truly related to CMS and their mechanisms of action require further study.

Supplementary Materials: The following are available online at www.mdpi.com/2073-4395/10/6/765/s1: Figures S1-S3: Physical maps of the mt genomes of the CMS lines J4A-1, J4A-2 and J4B. Figure S4: Sequence alignment of orf186a-1, orf186a-2, orf4 and orf4-2. Table S1: Statistics of mitochondrial sequencing. Table S2: Repeats (≥100 bp) found in J4A-1 mtDNA. Table S3: Repeats (≥100 bp) found in J4A-2 mtDNA. Table S4: Repeats (≥100 bp) found in J4A-3 mtDNA. Table S5: Repeats (≥100 bp) found in J4B mtDNA. Table S6: Distribution of chloroplast sequences in the cotton genome. Table S7: Distribution of nuclear sequences in the mt genome (identity >95%). Table S8: Unique ORFs (>300 bp) present in the J4A-1 mt genome. Table S9: Unique ORFs (>300 bp) present in the J4A-2 mt genome. Table S10: Evolutionary rates of protein-coding genes.

Author Contributions: Conceptualization, M.L. and R.Z.; Data curation, M.L. and L.C.; Formal analysis, D.T.; Methodology, L.C., X.L. and X.K.; Resources, M.L. and R.Z.; Software, L.C.; Validation, B.L.; Visualization, J.Y.; Writing—original draft, M.L.; Writing—reviewing and editing, R.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (31571719), Innovation Project of Guangxi Graduate Education (YCBZ2017014) and Weng Hongwu Original Research Fund of Peking University of China (WHW201809).

Acknowledgments: We thank the Analytical and Testing Center of the Agricultural College of Guangxi University for sampling assistance and providing large scientific facilities.

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

References

1. Clifton, S.W.; Minx, P.; Fauron, C.M.; Gibson, M.; Allen, J.O.; Sun, H.; Thompson, M.; Barbazuk, W.B.; Kanuganti, S.; Tayloe, C.; et al. Sequence and comparative analysis of the maize NB mitochondrial genome. *Plant Physiol.* 2004, 136, 3486–3503. doi:10.1104/pp.104.044602.

2. Liu, H.; Cui, P.; Zhan, K.; Lin, Q.; Zhuo, G.; Guo, X.; Ding, F.; Yang, W.; Liu, D.; Hu, S.; et al. Comparative analysis of mitochondrial genomes between a wheat K-type cytoplasmic male sterility (CMS) line and its maintainer line. *BMC Genom.* 2011, 12, 163. doi:10.1186/1471-2164-12-163.
3. Luo, D.; Xu, H.; Liu, Z.; Guo, J.; Li, H.; Chen, L.; Fang, C.; Zhang, Q.; Bai, M.; Yao, N.; et al. A detrimental mitochondrial-nuclear interaction causes cytoplasmic male sterility in rice. Nat. Genet. 2013, 45, 573–577. doi:10.1038/ng.2570.

4. Chen, L.; Liu, Y.G. Male sterility and fertility restoration in crops. Annu. Rev. Plant Biol. 2014, 65, 579–606. doi:10.1146/annurev-arplant-050213-040119.

5. Sloan, D.B.; Alverson, A.J.; Storchova, 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. doi:10.1186/1471-2148-10-274.

6. Archibald, J.M. Endosymbiosis and eukaryotic cell evolution. Curr. Biol. 2015, 25, R911–R921. doi:10.1016/j.cub.2015.07.055.

7. 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, 12714. doi:10.1038/s41598-018-30297-w.

8. Wang, M.; Tu, L.; Yuan, D.; Zhu, D.; Shen, C.; Li, J.; Liu, F.; Pei, L.; Wang, P.; Zhao, G.; et al. Reference genome sequences of two cultivated allotetraploid cottons, Gossypium hirsutum and Gossypium barbadense. Nat. Genet. 2019, 51, 224–229. doi:10.1038/s41588-018-0282-x.

9. Tang, H.; Zheng, X.; Li, C.; Xie, X.; Chen, Y.; Chen, L.; Zhao, X.; Zheng, H.; Zhou, J.; Ye, S.; et al. Multi-step formation, evolution, and functionalization of new cytoplasmic male sterility genes in the plant mitochondrial genomes. Cell Res. 2017, 27, 130–146. doi:10.1038/cr.2016.115.

10. Horn, R. Molecular diversity of male sterility inducing and male-fertile cytoplasm in the genus Helianthus. Theor. Appl. Genet. 2002, 104, 562–570. doi:10.1007/s00122-001-0771-6.

11. Touzet, P.; Meyer, E.H. Cytoplasmic male sterility and mitochondrial metabolism in plants. Mitochondrion 2014, 19 P.1–B, 166–171. doi:10.1016/j.mito.2014.04.009.

12. Small, I.; Suffolk, R.; Leaver, C.J. Evolution of plant mitochondrial genomes via substoichiometric intermediates. Cell 1989, 58, 69–76. doi:10.1016/0092-8674(89)90403-0.

13. Albert, B.; Godelle, B.; Gouyon, P. Evolution of the plant mitochondrial genome: Dynamics of duplication and deletion of sequences. J. Mol. Evol. 1998, 46, 155–158. doi:10.1007/p10006290.

14. Kmiec, B.; Woloszynska, M.; Janska, H. Heteroplasmy as a common state of mitochondrial genetic information in plants and animals. Curr. Genet. 2006, 50, 149–159. doi:10.1007/s00294-006-0082-1.

15. Woloszynska, M.; Trojanowski, D. Counting mtDNA molecules in Phaseolus vulgaris: Sublimons are constantly produced by recombination via short repeats and undergo rigorous selection during substoichiometric shifting. Plant Mol. Biol. 2009, 70, 511–521. doi:10.1007/s11103-009-9488-8.

16. Gabay-Laughnan, S.; Newton, K. Plant mitochondrial mutations. In Genomics of Chloroplasts and Mitochondria. Advances in Photosynthesis and Respiration (Including Bioenergy and Related Processes); Bock, R., Knoop, V., Eds.; Springer: Dordrecht, The Netherlands, 2012; pp. 267–292.

17. Arrieta-Montiel, M.P.; Mackenzie, S.A. Plant mitochondrial genomes and recombination. In Plant Mitochondria; Kempken, F., Ed.; Springer: New York, NY, USA, 2011; pp. 65–82.

18. Dewey, R.E.; Timothy, D.H.; Levings, C.S. A mitochondrial protein associated with cytoplasmic male sterility in the T cytoplasm of maize. Proc. Natl. Acad. Sci. USA 1987, 84, 5374–5378. doi:10.1073/pnas.84.15.5374.

19. Yi, P. Discovery of mitochondrial chimeric gene associated with cytoplasmic male sterility of HL-rice. Chin. Sci. Bull. 2002, 47, 744. doi:10.1360/02tb9168.

20. Satoh, M.; Kubo, T.; Nishizawa, S.; Estiati, A.; Ichoda, N.; Mikami, T. The cytoplasmic male-sterile type and normal type mitochondrial genomes of sugar beet share the same complement of genes of known function but differ in the content of expressed ORFs. Mol. Genet. Genom. 2004, 272, 247–256. doi:10.1007/s00438-004-1058-9.

21. Yamamoto, M.P.; Kubo, T.; Mikami, T. The 5'-leader sequence of sugar beet mitochondrial atp6 encodes a novel polypeptide that is characteristic of Osman cytoplasmic male sterility. Mol. Genet. Genom. 2005, 273, 342–349. doi:10.1007/s00438-005-1140-y.

22. Meyer, V.G. Male sterility from gossypium harknessii. J. Hered. 1975, 66, 23–27. doi:10.1093/oxfordjournals.jhered.a108566.

23. Stewart, J.M. A new male sterility from G. trilobum. In Proceedings of the Beltwide Cotton Conference (National Cotton Council), Memphis, TN, USA, 1992; p. 610.

24. Zhang, X.; Meng, Z.; Zhou, T.; Sun, G.; Shi, J.; Yu, Y.; Zhang, R.; Guo, S. Mitochondrial SCAR and SSR Markers for distinguishing cytoplasmic male sterile lines from their isogenic maintainer lines in cotton. Plant Breed. 2012, 131, 563–570. doi:10.1111/j.1439-0523.2012.01971.x.
25. Chen, Z.; Nie, H.; Wang, Y.; Pei, H.; Li, S.; Zhang, L.; Hua, J. Rapid evolutionary divergence of diploid and allotetraploid Gossypium mitochondrial genomes. BMC Genom. 2017, 18, 876. doi:10.1186/s12864-017-4282-5.

26. Li, S.; Chen, Z.; Zhao, N.; Wang, Y.; Nie, H.; Hua, J. The comparison of four mitochondrial genomes reveals cytoplasmic male sterility candidate genes in cotton. BMC Genom. 2018, 19, 775. doi:10.1186/s12864-018-5122-y.

27. Ling, X.; Zhou, P.; Guan, H.; Jing, R.; Zhu, Y. Isolation of the mtDNA bands associated with CMS by AFLP technique. Hereditas 1999, 21, 33–36.

28. Nurk, S.; Anton, B.; Dmitry, A.; Alexey, G.; Anton, K.; Alla, L.; Andrey, P.; Alexey, P.; Alexander, S.; Yakov, S.; et al. Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. In Research in Computational Molecular Biology; Deng, M., Jiang, R., Sun, F., Zhang, X., Eds.; RECOMB, 2013, Lecture Notes in Computer Science, Springer: Berlin/Heidelberg, Germany, 2013; Volume 7821, pp. 158–170. doi:https://doi.org/10.1007/978-3-642-37195-0_13.

29. Liu, G.Z.; Cao, D.; Li, S.S.; Su, A.G.; Geng, J.n.; Grover, C.E.; Hu, S.N.; Hua, J.p. The complete mitochondrial genome of Gossypium hirsutum and evolutionary analysis of higher plant mitochondrial genomes. PLoS ONE 2013, 8, e69476. doi:10.1371/journal.pone.0069476.

30. Luo, R.B.; Liu, B.H.; Xie, Y.L.; Li, Z.Y.; Huang, W.H.; Yuan, J.Y.; He, G.Z.; Chen, Y.X.; Pan, Q.; Liu, Y.J.; et al. Soapdenovo2: An empirically improved memory-efficient short-read de novo assembler. Gigascience 2012, 1, 1–18. doi:https://doi.org/10.1186/2047-217X-1-18.

31. Lohse, M.; Drechsel, O.; Bock, R. OrganellarGenomeDRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr. Genet. 2007, 52, 267–274. doi:10.1007/s00294-007-0161-y.

32. Stanke, M.; Steinkamp, R.; Waack, S.; Morgenstern, B. AUGUSTUS: A web server for gene finding in eukaryotes. Nucleic Acids Res. 2004, 32, W309–W312. doi:10.1093/nar/gkh379.

33. Haas, B.J.; Salzberg, S.L.; Zhu, W.; Pertea, M.; Allen, J.E.; Orvis, J.; White, O.; Buell, C.R.; Wortman, J.R. Automated eukaryotic gene structure annotation using EViidenceModeler and the program to assemble spliced alignments. Genome Biol. 2008, 9, R7. doi:10.1186/gb-2008-9-1-r7.

34. Lowe, T.M.; Eddy, S.R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997, 25, 955–964. doi:10.1093/nar/25.5.955.

35. Ibba, M.; Soll, D. Aminoacyl-tRNA synthesis. Annu. Rev. Biochem. 2000, 69, 617–650. doi:10.1146/annurev.biochem.69.1.617.

36. Cole, J.R.; Wang, Q.; Cardenas, E.; Fish, J.; Chai, B.; Farris, R.J.; Kulam-Syed-Mohideen, A.S.; McGarrell, D.M.; Marsh, T.; Garrity, G.M.; et al. The ribosomal database project: Improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 2009, 37, D141–D145. doi:10.1093/nar/gkn879.

37. Lowe, T.M.; Chan, P.P. tRNAscan-SE On-line: Integrating search and context for transfer RNA genes. Nucleic Acids Res. 2016, 44, W54–W57. doi:10.1093/nar/gkw413.

38. Olson, S.A. EMBOSS opens up sequence analysis. European molecular biology open software suite. Brief. Bioinform. 2002, 3, 87–91. doi:10.1093/bib/3.1.87.

39. Rombel, I.T.; Sykes, K.F.; Rayner, S.; Johnston, S.A. OR-FINDER: A vector for high-throughput gene identification. Gene 2002, 282, 33–41. doi:10.1016/s0378-1119(01)00819-8.

40. Song, J.; He, Q.-F. Bioinformatics analysis of the structure and linear B-cell epitopes of aquaporin–3 from Schistosoma japonicum. Asian Pac. J. Trop. Med. 2012, 5, 107–109. doi:10.1016/s1995-7645(12)60005-4.

41. Kurtz, S.; Phillippy, A.; Delcher, A.L.; Smoot, M.; Shumway, M.; Antonescu, C.; Salzberg, S.L. Versatile and open software for comparing large genomes. Genome Biol. 2004, 5, R12. doi:10.1186/gb-2004-5-2-r12.

42. Hulse-Kemp, A.M.; Ashrafi, H.; Zheng, X.; Wang, F.; Hoogenauer, K.A.; Maeda, A.B.; Yang, S.S.; Stoffel, K.; Matvienko, M.; Clemons, K.; et al. Development and bin mapping of gene-associated interspecific SNPs for cotton (Gossypium hirsutum L.) introgression breeding efforts. BMC Genom. 2014, 15, 945. doi:10.1186/1471-2164-15-945.

43. Birney, E.; Clamp, M.; Durbin, R. GeneWise and genomewise. Genome Res. 2004, 14, 988–995. doi:10.1101/gr.186550.4.

44. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 1998, 15, 568–573. doi:10.1093/oxfordjournals.molbev.a025957.

45. Allen, J.O.; Fauron, C.M.; Minx, P.; Roark, L.; Oddiraju, S.; Lin, G.N.; Meyer, L.; Sun, H.; Kim, K.; Wang, C.; et al. Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. Genetics 2007, 177, 1173–1192. doi:10.1534/genetics.107.073312.

46. Park, J.Y.; Lee, Y.P.; Lee, J.; Choi, B.S.; Kim, S.; Yang, T.J. Complete mitochondrial genome sequence and identification of a candidate gene responsible for cytoplasmic male sterility in radish (Raphanus sativus L.) containing DCGMS cytoplasm. Theor. Appl. Genet. 2013, 126, 1763–1774. doi:10.1007/s00122-013-2090-0.
47. Alversion, 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. doi:10.1093/molbev/msq029.
48. Goremykin, V.V.; Salamini, F.; Velasco, R.; Viola, R. Mitochondrial DNA of Vitis vinifera and the issue of rampant horizontal gene transfer. Mol. Biol. Evol. 2009, 26, 99–110. doi:10.1093/molbev/msn226.
49. Berghothorssen, U.; Adams, K.L.; Thomason, B.; Palmer, J.D. Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature 2003, 424, 197–201. doi:10.1038/nature01743.
50. Rice, D.W.; Alversion, A.J.; Richardson, A.O.; Young, G.J.; Sanchez-Puerta, M.V.; Munzinger, J.; Barry, K.; Boore, J.L.; Zhang, Y.; dePamphilis, C.W.; et al. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm Amborella. Science 2013, 342, 1468–1473. doi:10.1126/science.1246275.
51. Straub, S.C.; Cronn, R.C.; Edwards, C.; Fishbein, M.; Liston, A. Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (apocynaceae). Genome Biol. Evol. 2013, 5, 1872–1885. doi:10.1093/gbe/evt140.
52. Koizuka, N.; Imai, R.; Iwabuchi, M.; Sakai, T.; Imamura, J. Genetic analysis of fertility restoration and accumulation of ORF125 mitochondrial protein in the kosena radish (Raphanus sativus cv. Kosena) and a Brassica napus restorer line. Theor. Appl. Genet. 2000, 100, 949–955. doi:10.1007/s001220051375.
53. Kubo, T.; Newton, K.J. Angiosperm mitochondrial genomes and mutations. Mitochondrion 2008, 8, 5–14. doi:10.1016/j.mito.2007.10.006.
54. Yang, J.H.; Huai, Y.; Zhang, M.F. Mitochondrial atpA gene is altered in a new orf220-type cytoplasmic male-sterile line of stem mustard (Brassica juncea). Mol. Biol. Rep. 2009, 36, 273–280. doi:10.1007/s11033-007-9176-1.
55. Yang, J.H.; Zhang, M.F.; Yu, J.Q. Mitochondrial nad2 gene is co-transcribed with CMS-associated orfB gene in cytoplasmic male-sterile stem mustard (Brassica juncea). Mol. Biol. Rep. 2009, 36, 345–351. doi:10.1007/s11033-007-9185-0.
56. Heng, S.; Wei, C.; Jing, B.; Wan, Z.; Wen, J.; Yi, B.; Ma, C.; Tu, J.; Fu, T.; Shen, J. Comparative analysis of mitochondrial genomes between the hau cytoplasmic male sterility (CMS) line and its iso-nuclear maintainer line in Brassica juncea to reveal the origin of the CMS-associated gene orf288. BMC Genom. 2014, 15, 322. doi:10.1186/1471-2164-15-322.
57. L’Homme, Y.; Stahl, R.J.; Li, X.Q.; Hameed, A.; Brown, G.G. Brassica nap cytoplasmic male sterility is associated with expression of a mtDNA region containing a chimeric gene similar to the pol CMS-associated orf224 gene. Curr. Genet. 1997, 31, 325–335. doi:10.1007/s002940050212.
58. Menassa, R.; L’Homme, Y.; Brown, G.G. Post-transcriptional and developmental regulation of a CMS-associated mitochondrial gene region by a nuclear restorer gene. Plant J. 1999, 17, 491–499. doi:10.1046/j.1365-313x.1999.00397.x.
59. Gallagher, L.J.; Betz, S.K.; Chase, C.D. Mitochondrial RNA editing truncates a chimeric open reading frame associated with S male-sterility in maize. Curr. Genet. 2002, 42, 179–184. doi:10.1007/s00294-002-0344-5.
60. Song, J.; Hedgcock, C. A chimeric gene (orf256) is expressed as protein only in cytoplasmic male-sterile lines of wheat. Plant Mol. Biol. 1994, 26, 533–539. doi:10.1007/bf00039566.
61. Palmer, J.D.; Herbon, L.A. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J. Mol. Evol. 1988, 28, 87–97. doi:10.1007/bf02143500.
62. Kubo, T.; Mikami, T. Organization and variation of angiosperm mitochondrial genome. Physiol. Plant. 2007, 129, 6–13. doi:10.1111/j.1399-3054.2006.00768.x.

© 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/).