DRAFT GENOME ASSEMBLY, ORGANELLE GENOME SEQUENCING AND DIVERSITY ANALYSIS OF MARAMA BEAN (TYLOSEMA ESCULENTUM), THE GREEN GOLD OF AFRICA

by

JIN LI

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Dissertation Advisor: Dr. Christopher A. Cullis

Department of Biology

CASE WESTERN RESERVE UNIVERSITY

May, 2023

CASE WESTERN RESERVE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

We hereby approve the dissertation of

Jin Li

Candidate for the degree of **Doctor of Philosophy***.

Committee Chair

Dr. Hillel J. Chiel

Committee Member

Dr. Christopher A. Cullis

Committee Member

Dr. Jean H. Burns

Committee Member

Dr. Sarah C. Bagby

Committee Member

Dr. Peter A. Zimmerman

Date of Defense

March 20, 2023

*We also certify that written approval has been obtained for any proprietary material contained therein.

Table of Contents

List of Tables

Chapter 2

Chapter 3

Chapter 4

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 5

Acknowledgements

First of all, I would like to thank my advisor, Dr. Christopher Cullis. Since we first met at DeGrace to discuss course selection, I have been deeply impressed by your knowledge and wisdom. You always have clear goals and detailed plans for everything you do. This is what I have always wanted to learn. Thank you so much for your patient teaching and constant encouragement over the years. You have always been very nice to me, even in the beginning, I was not prepared to be a researcher, doing things passively and making mistakes all the time. Under your guidance, I have gradually developed a strong interest in the research I am doing, and strive to become a qualified researcher. I thoroughly enjoyed my time with you and would like to express my sincerest gratitude for your hard work over the years. I would also like to thank your family, Margaret and other members, for their care and hospitality.

I would like to thank Dr. Jean Burns, Dr. Sarah Bagby, Dr. Peter Zimmerman, and Dr. Hillel Chiel for your willingness to serve on the committee. Your enthusiasm, rigor, and kindness deeply infected me. Thank you for your contributions over the years and for all the advice you have given. These are very precious gifts to me.

I would like to thank Dr. Ellyn Evans, Daniela Dulworth, Miles Lanicca and all other lab members for your help in my research. You guys make the lab full of joy and feel like home.

I would like to thank all my family members and friends for your encouragement and support over the years. It is a blessing to be able to pursue my dreams and I know I would have been lost without your dedication.

I would also like to thank Dr. Kyle Logue for helping with the genome assembly, to Dr. Percy Chimwamurombe, Dr. Mutsa Takundwa, Dr. Juan Vorster, and Dr. Karl Kunert for providing marama samples, and to the EHI Finish Line Fund of the College of Arts and Sciences for covering part of the publication fee. This work would not have been possible without the generous support of everyone, for which I would like to express my heartfelt thanks!

Draft Genome Assembly, Organelle Genome Sequencing and Diversity Analysis of Marama Bean (*Tylosema esculentum*), the Green Gold of Africa

Abstract

by

JIN LI

Tylosema esculentum (marama bean) is an underutilized legume, long considered as a local potential crop due to its rich nutritional value. The reference plastome and mitogenome were assembled using a hybrid method with both Illumina and PacBio data. The diversity was explored with the WGS data of 84 samples from various geographic locations in Namibia and Pretoria. Phylogenetic analysis revealed two cytotypes with distinct plastomes and mitogenomes with differing levels of variability. Deep sequencing has identified heteroplasmy with both types of organellar genomes present, albeit one at a very low frequency. The inheritance of this complex of organellar genomes appears to be fairly constant, providing a conundrum of how the two genomes co-exist and are propagated through generations.

The type 1 mitogenome has two autonomous rings with a total length of 399,572 bp, which can be restructured into five smaller circular molecules through recombination on 3 pairs of long direct repeats. The type 2 mitogenome contains a unique 2,108 bp sequence, which connects distant segments to form a new structure consisting of three circular molecules and one linear chromosome. This increased the copy number of *nad9*,

rrns, *rrn5*, *trnC*, and *trnfM*. The two mitogenomes differed at another 230 loci, with only one nonsynonymous substitution in *matR*. cpDNA insertions were concentrated in one subgenomic ring of the mitogenome, including a 9,798 bp long fragment that contains potential *psbC*, *rps14*, *psaA*, and *psaB* pseudogenes. The two types of plastomes range in length from 161,537 bp to 161,580 bp, differing at 122 loci and at a 230 bp inversion. The chloroplast genes *rpoC2*, *rpoB*, and *ndhD* were found to be more diverse than other genes in marama plastome.

21.6 Gb PacBio HiFi data was assembled using Canu v2.2 into an unphased assembly of 1.24 Gb. k-mer analysis indicated that marama may be ancient tetraploid with an estimated genome size of only 277 Mb. The generated assembly has an N50 value of 1.28 Mb, with some contigs nearly half the length of the *Bauhinia* chromosome. The BUSCO completeness was 99.5% and repetitive sequences accounted for 27.35% of the assembly.

Chapter 1. Introduction

1.1 Introduction of Marama Bean

Tylosema esculentum, also known as marama bean, gemsbok bean, or tamani berry, is a long-lived perennial legume native to the arid and semi-arid areas of Kalahari Desert and neighboring Botswana, Namibia, and South Africa (National Research Council, 1979; Bower et al., 1988) (Figure 1.1). Marama grows naturally in environments with prolonged drought, low rainfall (100-900 mm), and high temperatures (daily maximum of 37 ℃ in the growing season) (Maesen, 2006). One reason it can survive such harsh conditions is because marama has developed a unique drought avoidance strategy of growing giant tubers weighing up to 500 pounds to store water for emergency use (Figure 1.2) (Cullis et al., 2018; Lawlor, 2018; Cullis et al., 2022).

Figure 1. 1 Morphology of marama (*Tylosema esculentum*) plants from Namibia. (A) Double-lobed leaves, soft and red-brown when young, becoming leathery and gray-green

over time (NRC, 2006). (B) Brownish-black seeds, edible when roasted, rich in protein, lipids, and micronutrients (Jackson et al., 2010). (C) Yellow-orange flowers that bloom every midsummer, starting generally from the 3rd or 4th year after planting (Bower et al., 1988). (D) Prostrate vines, can grow up to six meters or longer (Vietmeyer, 1986).

Figure 1. 2 Photo of a giant marama tuber weighing over 500 pounds (BIOL 309/409: Biology Field Studies Class; Photo courtesy of Dr. Cullis).

Marama is often referred to as “the green gold of Africa” because its seeds and tubers are edible and nutritious. The protein content of marama seeds is approximately 30-39% dry matter (dm), comparable to that of the commercial crop soybean (Belitz et al., 2004; Holse et al., 2010). Its lipid content is between 35% and 48% dm, which is close to that of peanut (Amarteifio, 1998; Belitz et al., 2004). Marama also provides many micronutrients, including calcium, iron, zinc, phosphate, magnesium, and B vitamins (folate), as well as various phytonutrients, such as phenolics, flavonoids, saponins, and phytosterols, which help strength the immune system and help reduce the risk of cancer, diabetes, and cardiovascular disease (Jackson et al., 2010; Khan, 2018; Omotayo and Aremu, 2021). Despite its socioeconomic value, marama remains

underutilized. The main way of acquisition for local villagers is picking from wild plants. The domestication and breeding of marama is thought to have the potential to improve food security threatened by global warming in Southern Africa, where traditional crops unsuitable for cultivation.

Marama is a non-nodulating legume from the subfamily Cercidoideae, which contains over 400 species in 14 genera (LPWG, 2017; Sinou et al., 2020). The genus *Tylosema,* in the tribe Cercideae subtribe Bauhiniinae, contains three other species, *Tylosema argentea*, *Tylosema fassoglense* and *Tylosema humifusa* (Coetzer and Ross, 1977). They are distributed throughout Africa, but marama is endemic to southern Africa (Jackson et al., 2010; Wunderlin, 2010). However, even marama plants grown in the same environment and of the same age still show very large differences in phenotypes such as plant size and vegetative growth rate (Cullis et al., 2019). In addition, marama usually does not bloom until two to four years after planting, making traditional breeding time-consuming. Therefore, the selection of improved marama individuals and the study of the molecular genetic basis behind the phenotypes of interest are of great importance for the breeding of the bean.

The total genome size of marama was estimated to be 1 Gb, calculated by dividing the size of the WGS data by the average genome coverage (Cullis, 2019). The chromosome number of marama was determined to be 44 by Feulgen staining, and marama is considered an ancient tetraploid plant ($2n = 4x = 44$) (Takundwa et al., 2012). The reference chloroplast genome of marama was assembled and found to be 161,537 bp in length with a unique inversion of 7,479 bp, containing six genes *rbcL, accD, psaI, ycf4, cemA* and *petA,* and this has not been found in legumes

other than *Tylosema* (Kim and Cullis, 2017). Substantial intra-population genetic diversity was found in Namibian individuals, but little variation between different populations, as assessed by SSR markers. These are thought to be related to plant stress responses to extreme environments and lack of cross-population pollinators, respectively (Nepolo, 2010).

1.2 Overview of Research Topics

Chapter 2 is about the mitochondrial genome assembly of a single marama individual. Mitochondria are not only well-known energy production factories, but also are considered to be related to many important traits, such as cytoplasmic male sterility (CMS) and self-incompatibility in angiosperms (Sloan et al., 2012b). With the development of sequencing technology, the understanding of its function is gradually deepening. The next-generation sequencing technology Illumina has many advantages such as fast, cheap, and high-throughput, which has greatly promoted genome sequencing, but one of its main drawbacks is that the generated reads are short, with a length of only 100 to 300 bp (Zavodna et al., 2014; Ari and Arikan, 2015). Plant mitochondrial genomes are very complex and are typically much larger than animal mitogenomes (Mower et al., 2012; Wu et al., 2015). They often contain many long repeats, and recombination between the repeats can form a large number of subgenomic structures that coexist in the same individual (Kozik et al., 2019; Zardoya, 2020). The assembly of plant mitogenomes was made possible by PacBio, a third-generation sequencing technology that generates reads longer than 10 kb to span the repeats in the genomes. However, the accuracy of PacBio long reads was low at the first, even below 90% (Ono et al., 2012). Moreover, the direct assembly of such long reads requires a lot of

memory, and there were not many tools available at that time could do the job. Therefore, in this study, a hybrid assembly approach was performed alternatively using data from both Illumina and PacBio platforms. First, the mitochondrial sequence of marama was extracted from the whole genome sequencing (WGS) data by alignment with the publicly available reference mitogenome of the related species *Millettia pinnata*. Next, contigs constructed on the Illumina data were connected using the PacBio long reads to fill the gaps. Sequencing error correction was then performed by mapping Illumina short reads to the assembled reference sequence. Finally, the reference mitogenome of marama was obtained, which contained two autonomous circular chromosomes (Li and Cullis, 2021). The recombination that occurred on three pairs of long direct repeats and one pair of long inverted repeats led to the coexistence of different subgenomic structures in the studied individual and they were found to be in close molar concentrations. Gene annotation found that *atp1* and *atp8* were on the long repeats with doubled copy number.

Chapter 3 is about the diversity analysis of marama chloroplast genome. Chloroplasts are important organelles responsible for photosynthesis. Compared with the mitogenome, the frequency of sequence rearrangement of the chloroplast genome is much lower, but the rate of point mutations in the gene sequence is higher (Palmer and Herbon, 1988; Drouin et al., 2008). These make chloroplast genomes ideal candidates for evolutionary studies. The reference chloroplast genome of marama was previously published but refined in this study (Kim and Cullis, 2017). Chloroplast genome variations, including structural alterations, SNPs, and indels, were analyzed by mapping the Illumina WGS data from 84 marama individuals to the reference chloroplast genome. These included 44 samples collected from various geographical locations in Namibia and

South Africa to study plastome genetic diversity under selection in different environments. Another 40 samples were related progeny plants grown from seeds of 7 accessions for the study of cytoplasmic inheritance. Two distinct germplasms were found in this study, differing at 122 loci and a 230 bp inversion in the plastome. One type of chloroplast genome appeared to have greater variability than the other, and geographical patterns could be seen in the distribution of the variations, although this needs to be verified by studies with larger sample sizes. Furthermore, heteroplasmy (co-existence of both genotypes in the same individual) was seen in the plastomes of many individuals at the differential loci, although one type was usually at a low frequency. Substoichiometric shifting, which occasionally occurs under environmental selection, is thought to result in the replacement of the major genome by massive amplification of the minor genome in plant organelles (Arrieta-Montiel et al., 2001; Woloszynska, 2010). However, this was not seen in our study, and the inheritance of heteroplasmy appeared to be fairly stable in the chloroplast genome of marama, except for one individual that exhibited higher heteroplasmy.

Chapter 4 is about the genetic diversity analysis of marama mitogenome. The whole-genome sequencing data of the 84 individuals described in Chapter 3 were mapped to the reference mitogenome of marama (assembled in Chapter 2) to explore the genetic diversity. Significant differences were also found between the mitogenomes of the two germplasms mentioned in Chapter 3. Different from the reference, the newly assembled type 2 mitogenome contains three circular molecules and one long linear chromosome. This structural variation resulted in increased copy number of genes including *nad9*, *rrns*, *rrn5*, *trnC*, and *trnfM*. The type 2 mitogenome also contained a unique 2,108 bp fragment

on which primers for mitogenome type identification were designed. The two types of mitogenomes also differed from each other at another 230 loci. Evolutionary analysis based on all identified differential loci suggested that soil moisture levels may play an important role in the divergence of marama mitogenomes. The phylogenetic tree constructed on 24 conserved mitochondrial genes of different legumes was consistent with previously published trees built on the chloroplast genes, but the similarity between mitogenomes of even closely related legumes was found to be low (Kim and Cullis, 2017; Wang et al., 2018). Heteroplasmy in the mitogenome of marama was not as prevalent as in the chloroplast genome, but was found to be concentrated at a few loci, suggesting they may have been critical in marama evolution. All mitogenomes of marama were found to contain an over 9 kb cpDNA insertion. The study of the polymorphism on it speculated that the originally conserved chloroplast sequence began to accumulate numerous mutations only after being inserted into the mitogenome, resulting in the loss of function of the genes on the fragment.

Chapter 5 is about the assembly of marama draft nuclear genome. A high-quality genome assembly lays an important foundation for deciphering the genetic mechanisms of important traits and genetic improvement of crops (Li et al., 2021). However, as a polyploid plant, marama's genome is considered to be highly heterozygous and contains a large number of repetitive sequences, which makes the genome sequencing very difficult. The third-generation sequencing technologies, especially the recent high-precision PacBio HiFi sequencing technology, have greatly facilitated the sequencing of complex genomes (Hon et al., 2020). In this study, high molecular weight DNA samples were prepared and used to generate 21.6 Gb of PacBio HiFi data. The k-mer analysis of

the reads indicated that marama might be an ancient tetraploid plant, and the haplotype genome size was estimated to be about 277.4 Mb. An unphased genome assembly of 1.24 Gb was obtained with an N50 value of 1.28 Mb, a significant increase from the previous 3 kb. The longest contig has reached half the length of the *Bauhinia* chromosome. The BUSCO completeness was over 99%, and the proportion of repetitive sequences in the assembly was 27.35%. This is considered an ongoing project and will continual to be improved in the future.

Chapter 2. The Multipartite Mitochondrial Genome of Marama (*Tylosema esculentum*)

Published in *Frontiers in Plant Science* Vol. 12, No. 787443, 2021.

DOI: 10.3389/fpls.2021.787443

Authors: Jin Li and Christopher Cullis*

Department of Biology, Case Western Reserve University, Cleveland, OH 44106, USA

* Correspondence: cac5@case.edu

(Received 30 September 2021, Accepted 15 November 2021, Published 08 December 2021)

2.1 Abstract

Tylosema esculentum (marama bean), a wild legume from tropical Africa, has long been considered as a potential crop for local farmers due to its rich nutritional value. Genomics research of marama is indispensable for the domestication and varietal improvement of the bean. The chloroplast genome of marama has been sequenced and assembled previously using a hybrid approach based on both Illumina and PacBio data. In this study, a similar method was used to assemble the mitochondrial genome of marama. The mitochondrial genome of the experimental individual has been confirmed to have two large circles OK638188 and OK638189, which do not recombine according to the data. However, they may be able to restructure into five smaller circles through recombination on the 4 pairs of long repeats (>1kb). The total length of marama mitogenome is 399,572 bp. A 9,798 bp DNA fragment has been found that is homologous to the chloroplast genome of marama, accounting for 2.5% of the mitogenome. In the Fabaceae family, the mitogenome of *Millettia pinnata* is highly

similar to marama, including for both the genes present and the total size. Some genes including *cox2, rpl10, rps1* and *sdh4* have been lost during the evolution of angiosperms and are absent in the mitogenomes of some legumes. However, these remain intact and functional in marama. Another set of genes, *rpl2, rps2*, *rps7, rps11*, *rps13* and *rps19* are either absent, or present as pseudogenes, in the mitogenome of marama.

Keywords: *Tylosema esculentum*, marama, Fabaceae, mitochondrial genome sequence, hybrid genome assembly, genome sequencing, homologous recombination, multipartite genomic conformations

2.2 Introduction

As the powerhouse of the cell, mitochondria are where oxidative phosphorylation takes place and adenosine triphosphate (ATP), the cellular energy source, is produced (Siekevitz, 1957). Mitochondria play important roles in plant growth and development by regulating physiological metabolism, programmed cell death (PCD), and other intracellular signaling pathways (Epstein et al., 2001; Galluzzi et al., 2012). Mitochondrial defects are associated with cytoplasmic male sterility (CMS) in plants, preventing them from producing functional pollen. This can be utilized by breeders to ensure crossing of self-pollinating plants and to obtain hybrid seeds (Kempken and Pring, 1999). Mitochondrial function is also believed to be involved in the adaptation of organisms to the environment and their responses to abiotic stresses. For example, during the development of mung bean (*Vigna radiata*) cotyledons, *in vivo* freeze-thaw treatment converted the mitochondrial rosette genome structure to a longer linear DNA shown by electron microscopy (Cheng et al., 2016). Knowing whether different mtDNA structures coexist in marama individuals and the transition between them may provide information

to understand better why marama can perform well in harsh environments. In addition, since plant mitochondrial genes are highly conserved with a low nucleotide mutation rate, they can also be used to study the evolution of marama (Palmer and Herbon, 1988).

Self-incompatibility (SI) is a common genetically determined mechanism in flowering plants that promotes outcrossing and prevents inbreeding to increase genetic diversity. Marama is self-incompatible. This is one major obstacle to its cultivation and breeding, since SI not only reduces the yield of seeds and/or fruits for crops (Miller & Gross, 2011), but also makes it difficult to combine desirable traits of two incompatible parents through simple cross-pollination (Claessen et al., 2019). The interaction between NaSIPP, a mitochondrial protein with phosphate transporter activity, exclusively transcribed in mature pollen, and stigma protein NaStEP has been found to be critical to SI in *Nicotiana* spp., and this process is thought to destabilize mitochondria and stop pollen tube growth (Garc á-Valencia et al., 2017). Since most self-incompatible Fabaceae plants are believed to apply a similar gametophytic SI system, the sequencing of marama mtDNA may be crucial for studying its self-incompatibility (Aguiar et al., 2015).

Plant mitochondrial genomes are very complex and diverse in size, sequence arrangement, repeat content and structure (Mower et al., 2012). Unlike animal mitochondrial genomes, which are typically only 16-20 kb, those of higher plants vary in size from 200 to 2,000 kb (Fauron et al., 1995). Some newly sequenced plant mitogenomes exceed this range, even reaching 11.3 Mb in the flowering plant *Silene conica* (Sloan et al., 2012b). Although many plant mitochondrial genomes are displayed as a master ring, in reality, their physical structure can be very complex, consisting of multiple circular, linear, and branched molecules as revealed by electron microscopy

(Backeret et al., 1997). A large number of repeats exist in plant mitochondrial genomes, including long repeats (>1 kb), short repeats (<100 nt), and tandem repeats. Repeat-mediated recombination is believed to be the main cause of structural variations in mtDNA and plays an important role in mitochondrial replication and repair (Maréchal and Brisson, 2010). Although the size of plant mitogenomes is usually large, their gene pool (24 core genes with 17 variable genes) is usually small, since many genes have been either lost or transferred to nucleus during angiosperm evolution, but the coding sequences of the remaining genes are highly conserved (Adams et al., 2002).

The existence of many repetitive sequences and multipartite subgenomes makes the assembly of plant mitochondrial genomes difficult. A hybrid assembly approach using PacBio long reads to fill gaps and high coverage Illumina short reads to correct sequencing errors has been shown to be effective in accomplishing this task (Figure 2.1) (Kozik et al., 2019). In this study, a similar method was used to unveil the mitogenome of marama.

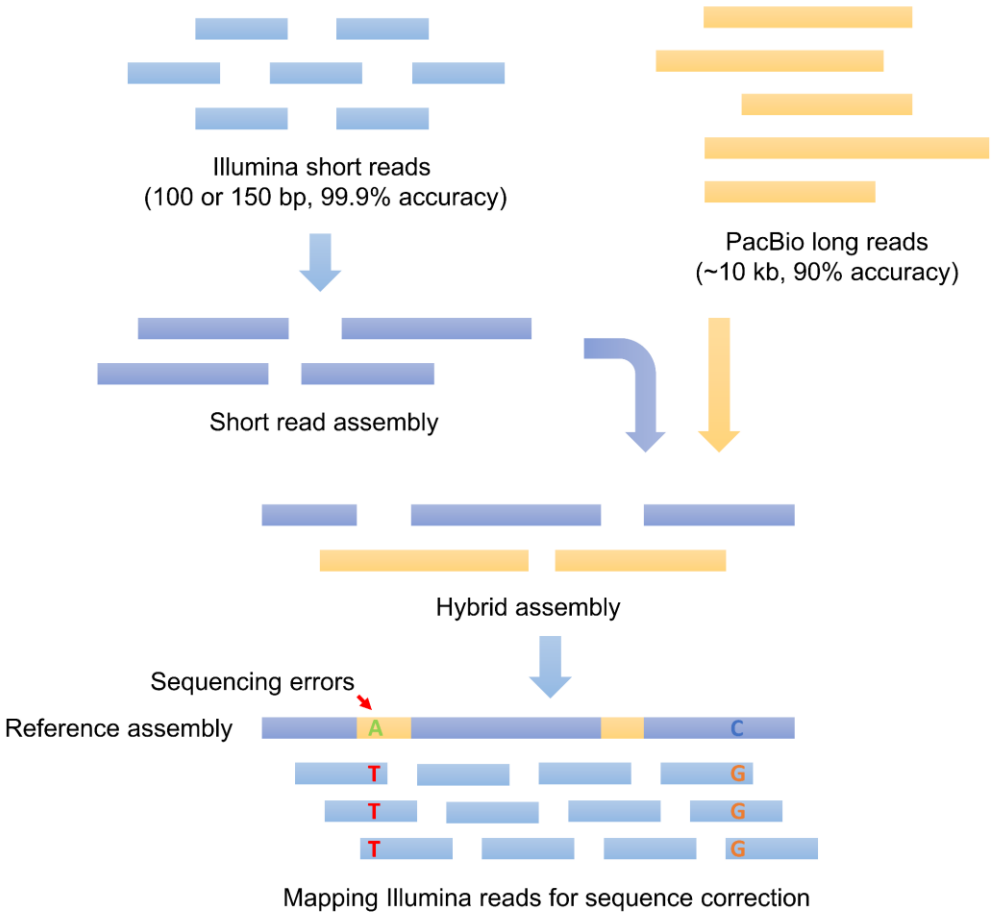

Figure 2. 1 Overview of Illumina and PacBio hybrid assembly.

2.3 Materials and Methods

2.3.1 Plant Materials Sources and Genome Sequencing

The marama samples were collected from plants growing at the University of Pretoria Farm in South Africa, and at various locations in Namibia. Total DNA was extracted from fresh young leaves using a Qiagen Plant miniprep kit following the manufacturer's protocol. The purity of DNA samples was assessed using a NanoDrop spectrophotometer to measure the ratio of 260/280 and 260/230 and was confirmed by

electrophoresing 20 µL of the sample on a 1.0% agarose TBE gel at 100 V running for 50 minutes, using GelRed® for DNA staining. The samples were then sent to the Génome Québec Innovation Centre, CWRU Genomics Core and Novogene Corporation for Illumina sequencing and long-read sequencing data were also generated using the PacBio RSII SMRT/Québec platform. 179,470,509 reads covering 26.9 Gb were obtained from the Illumina HiSeq 2000 platform for an individual collected in Namibia and the data was used for the first round assembly due to the high coverage. 21,373,859 reads, 37,816,777 reads and 46,425,865 reads were obtained from the same Illumina platform for one individual growing at the University of Pretoria Farm and two other plants grown from seeds collected in Namibia. The sample from the University of Pretoria Farm was also sequenced in five PacBio SMRT cells, producing 1.78 Gb of sequence with an average coverage of 15x-20x for the mitogenome. All raw Illumina and PacBio reads were submitted to the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra) under the accession number PRJNA779273.

2.3.2 Mitochondrial Genome Assembly

The paired end Illumina reads were assembled *de novo* by ABySS (Simpson et al., 2009), which were further elongated using DBG2OLC (Ye et al., 2016) to get the preliminary genome assembly. The mitochondrial contigs were extracted based on the BLAST alignment (Johnson et al., 2008) with the mitogenomes of other species (mainly Fabaceae). Ten mitochondrial genomes from 7 species of legumes and 3 other model plants including *Arabidopsis thaliana, Glycine max, Lotus japonicas, Medicago truncatula, Millettia pinnata, Vicia faba, Vigna angularis, Vigna radiata*, and *Zea mays (2)* were retrieved from the NCBI Organelle Genome Resource. The Illumina reads of the

studied individual were aligned to the 10 mitogenomes by Bowtie 2 (Langmead and Salzberg, 2012) using the i-Plant Discovery Environment (Cyverse).

The mitochondrial genome of *Millettia pinnata* (NC_016742.1) (Kazakoff et al., 2012) showed the highest alignment rate and was used as the reference to restore the marama mitochondrial genome. The coding sequences of *Millettia pinnata* mitochondrial genes including *atp1, atp4, atp6, atp8, atp9, ccmB, ccmC, ccmFc, ccmFn, cob, cox1, cox2, cox3, matR, mttB, nad1, nad2, nad3, nad4, nad4L, nad5, nad6, nad7, nad9, rpl2, rpl5, rpl16, rps1, rps2, rps3, rps4, rps7, rps10, rps11, rps12, rps14, rps19, sdh3,* and *sdh4* were obtained from NCBI GenBank and used to search for corresponding homologs in the marama genome assembly using BLAST. The acquired homologous sequences were then extended and assembled manually based on both the Illumina and PacBio sequencing data. The 500 bp sequences at both ends of each contig was BLASTed against the PacBio database on Geneious 9, and all obtained results were BLASTed back to the genome assembly to show all possible connections. The ID and number of PacBio reads going across each contig junction were recorded. Two long sequences were generated from these steps.

The Illumina reads of the studied individual were aligned to the two assembled sequences using Bowtie 2. A total of 16 primary scaffolds were obtained, which were then used for the second round of assembly. The process was repeated until the complete mitochondrial genome of marama is recovered. The results were validated by mapping Illumina reads from other marama individuals to the assembly by Bowtie 2 and the alignments were visualized in IGV (Robinson et al., 2011). The obtained mitochondrial

genomic structures were also verified by directly assembling the PacBio data using Canu v2.2 (Koren et al., 2017) (correctedErrorRate=0.15, genomeSize=20m).

2.3.3 Gene Annotation

The protein coding and rRNA genes were annotated using MITOFY (Alverson et al., 2010; https://dogma.ccbb.utexas.edu/mitofy/mitofy.cgi) and AGORA (Jung et al., 2018; https://bigdata.dongguk.edu/gene_project/AGORA/). The two programs gave similar annotation results, including gene sets, positions and matching rates, which were summarized and recorded in one spreadsheet. MITOFY BLASTN was used to annotate tRNA genes in the assembled genome. The reference sequences of the two chromosomes of the mitogenome of marama with gene annotation records were deposited in GenBank under the accession numbers OK638188 and OK638189. The circular visualization of the gene annotation was generated using OGDRAW (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html).

2.4 Results and Discussion

The Bowtie2 alignments resulted in 0.48% of marama Illumina whole genome sequencing reads mapping to the mitogenome of *Millettia pinnata*, 0.43% to *Lotus japonicus*, 0.31% to *Glycine max* and 0.30% to *Vigna radiata*. Therefore, the homologous contigs of *Millettia pinnata* mitochondrial genes in the marama genome assembly were used as the starting point for the assembly.

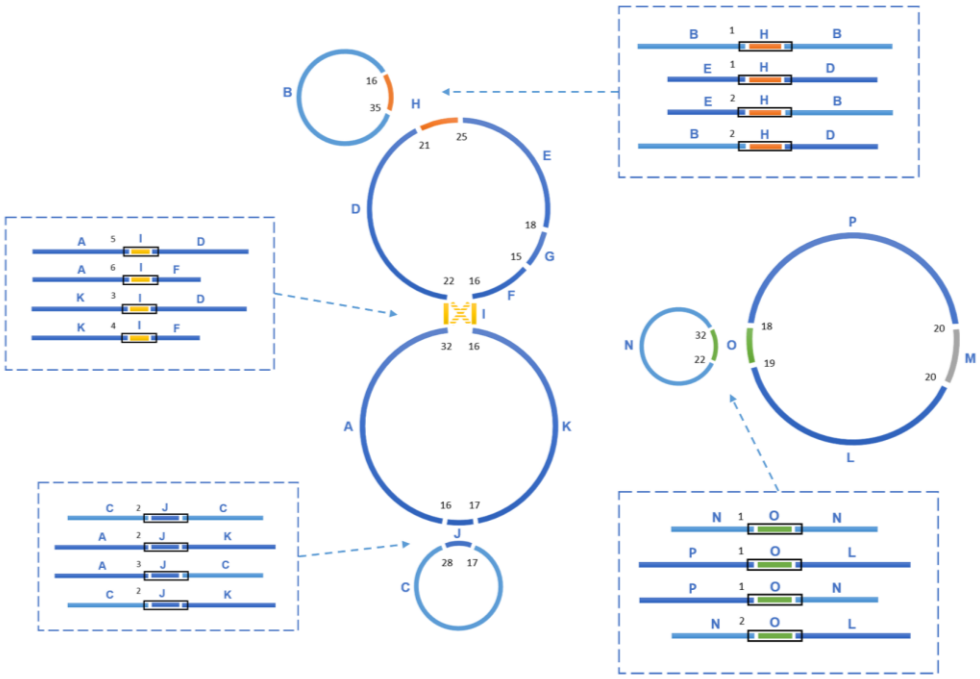

Figure 2. 2 The assembly graph of the multipartite *T. esculentum* mitochondrial genome. The marama mitochondrial genome is composed of 5 circular molecules, which were built based on 16 primary long scaffolds. 4 of them shown here including H, I, J, O are long repeats, which are present in two copies in the marama mitogenome. The scaffolds were assembled together according the PacBio long reads across the gaps. Connections between two adjacent scaffolds were quantified by PacBio reads and the counts are shown as numbers in black. Recombination on repeats causes the existence of different connections between the same scaffolds, which are shown (within black boxes) in the dashed boxes and also quantified by PacBio reads. The numbers next to the black boxes represent their counts. According to the connections shown in the dashed boxes, the 16 scaffolds can also be assembled into two large rings, as shown in Figure 2.3. The PacBio long reads data confirmed that these two structures co-exist in the experimental individual and they are close in ratio.

Table 2. 1 lengths of primary scaffolds for the assembly of *T. esculentum* mitochondrial genome.

Unit number	Length (bp)	Unit number	Length (bp)
A	39,443	N	27,635
J	3,908	O	4,926
K	39,470	P	56,817
C	35,384	L	42,335
B	39,313	M	9,760
D	34,472		
H	5,212		
E	27,177		
G	5,443		
F	8,311		
I	2,351		

The sequencing data were initially assembled into 16 primary scaffolds, ranging in size from 2,351 to 56,817 bp with a median of 27,406 bp (Table 2.1). According to the PacBio long reads at the junctions, these scaffolds were connected to form five possible circular molecules with sizes of 169,330 bp, 44,455 bp, 39,474 bp, 32,520 bp and 113,793 bp (Figure 2.2 and Table 2.2). The Illumina short reads were remapped to the final assembly using Bowtie 2 and the proportion of mitochondrial reads in the whole genome sequencing data varied from 2-3% between individuals. Four long repeats of 2,351 bp, 5,212 bp, 3,908 bp and 4,926 bp were found in these molecules (H, I, J and O on Figure 2.2), accounting for 8.2% of the mitogenome (Table 2.1 and Figure 2.2). The sequencing coverages of the regions where the four repeats are located are shown to be doubled (Figure S2.1-2.4). Four long fragments with lengths of 9,798 bp (M on Figure 2.2), 859 bp, 342 bp and 273 bp in marama mtDNA were identified as homologous to marama chloroplast genome, with a similarity ranging from 74% to 98%, covering a total of 7.2% of the chloroplast genome and 2.8% of the mitochondrial genome. These

fragments are all located on one circular molecule (M5 on Figure 2.3) and contain some chloroplast genes including *psaA*, *psaB* and part of *psbC* (Figure S2.5). When these chloroplast insertions occurred in the evolution and their role in the mitogenome remains unknown.

Table 2. 2 Summary of *T. esculentum* mitochondrial sub-genomic features.

Molecule	A (%)	C (%)	G (%)	T (%)	G~C (%)	Length (bp)
M1	27.88	22.14	22.26	27.72	44.40	169,330
M2	27.67	21.68	23.18	27.47	44.86	44,455
M3	27.87	22.23	23.06	26.84	45.29	39,474
M4	27.76	21.65	22.83	27.76	44.48	32,520
M5	27.29	22.43	22.57	27.71	45.00	113,793
LS1	27.81	22.34	22.28	27.58	44.62	253,259
LS2	27.39	22.52	22.36	27.72	44.88	146,313

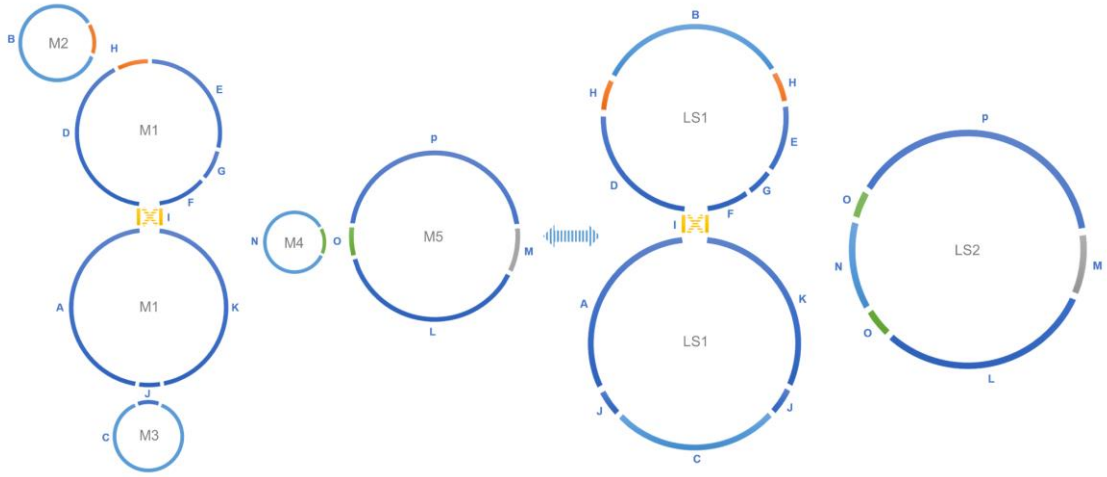

Figure 2. 3 Recombination between repeats forms alternative mitochondrial genomic conformations. Recombination across the direct repeats H (orange), J (blue), and O (green) merges the five small circular molecules into two large rings. M2 and the upper

ring of M1 are merged into the upper ring of LS1, the lower ring of M1 and M3 form the lower ring of LS1, and the M4 and M5 are merged into LS2, as shown in the figure. This transformation is believed to be reversible. Recombination between a single pair of inverted repeats I (yellow) changes the connection between the upper M1/LS1 ring and the lower M1/LS1 ring (AID, KIF to AIF, KID). The connection shown by the yellow solid line and the connection shown by the yellow dashed line are interchangeable through recombination. All structures mentioned above are experimentally verified in the data.

Homologous recombination between long repeats can reorganize the plant mitochondrial genome structure causing the formation of different subgenomic molecules. Two large circular molecules have been confirmed to exist in the marama mitochondrial genome, and they may be able to reversibly transform into five smaller molecules through repeat mediated recombination (Figure 2.3). No connection was found between the two large circular molecules in the mitogenome of the experimental individual indicating they are autonomous. 12 long contigs were obtained from the direct assembly of PacBio long reads using Canu v2.2 and the contigs were aligned to the structure units of LS1 and LS2 using BLAST to verify the connections as shown in Figure S2.6-S2.17. However, whether the two autonomous rings in other marama individuals can further recombine to form one master molecule, as described in the mitogenomes of other legumes such as *Millettia pinnata*, *Vigna radiata* and *Glycine max* (Chang et al., 2013; Alverson et al., 2011; Kazakoff et al., 2012) still needs to be determined. In addition, based on the coverage of the PacBio long reads, the ratio of the five small rings to the two large molecules was about 1:1 (Figure 2.2), but the sample size

is still small and more PacBio reads or qPCR amplifications at the junctions are required to verify the ratio.

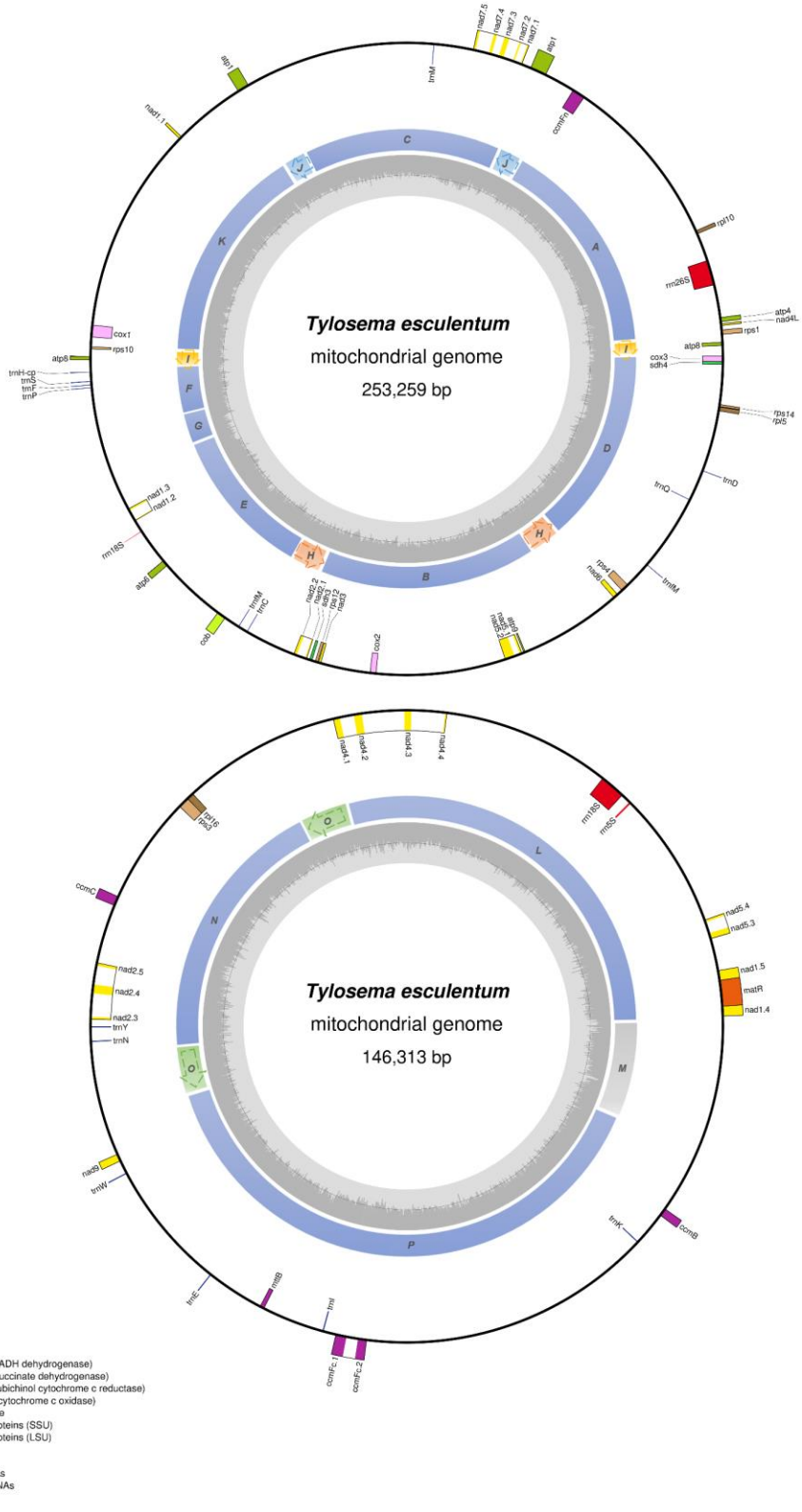
Tylosema esculentum
mitochondrial genome
253,259 bp
Tylosema esculentum
mitochondrial genome
146,313 bp
complex I (NADH dehydrogenase)
complex II (succinate dehydrogenase)
complex III (ubichinol cytochrome c reductase)
complex IV (cytochrome c oxidase)
ATP synthase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
maturases
other genes
transfer RNAs
ribosomal RNAs
introns

Figure 2. 4 Circular gene map of the two large molecules, LS1 (top) and LS2 (bottom) in the mitochondrial genome of *T. esculentum.* Genes inside the circle are transcribed clockwise, while genes outside are transcribed counterclockwise. The sequential order of exons is displayed by the decimal after the gene name abbreviation. Genes belonging to different functional groups are color-coded as shown at bottom left. Colors in the inner circle indicate the position of the structural units and repeats listed in Figure 2.3. The dashed arrows represent the direction of the repeats. GC content is represented by the grey shade in the inner circle.

Table 2. 3 Annotated genes in the mitochondrial genomes of *T. esculentum.*

Category	Names of Genes
Complex I (NADH dehydrogenase)	*nad1-7*, *nad4L*, and *nad9*
Complex II (succinate dehydrogenase)	*sdh3*, *sdh4*
Complex III (ubiquinol cytochrome-c reductase)	*cob*
Complex IV (cytochrome-c oxidase)	*cox1-3*
Complex V (ATP synthase)	*atp1*, *atp4*, *atp6*, *atp8*, *atp9*
Cytochrome c biogenesis	*ccmB*, *ccmC*, *ccmFc*, *ccmFn*
Large subunit ribosomal proteins	*rpl5*, *rpl10*, *rpl16*
Small subunit ribosomal proteins	*rps1*, *rps3*, *rps4*, *rps10*, *rps12*, *rps14*
Maturases	*matR*
Transport membrane protein	*mttB*
Ribosomal RNAs	*rrn5, rrn26, rrn18*
Transfer RNAs	*trnC-GCA*, *trnD-GTC*, *trnE-TTC*, *trnF-GAA*, *trnH-cp*, *trnI*, *trnK-TTT*, *trnM*, *trnfM-CAT*, *trnN-GTT*, *trnP-TGG*, *trnQ-TTG*, *trnS-GCT*, *trnW-CCA*, *trnY-GTA*

The total length of the marama mitochondrial genome is 399,572 bp, which is slightly smaller than the mitogenomes of some other Fabaceae, 402,558 bp in *Glycine max*, 401,262 bp in *Vigna radiata* and 425,718 bp in *Millettia pinnata* (Chang et al.,

2013; Alverson et al., 2011; Kazakoff et al., 2012), but larger than that of *Arabidopsis thaliana* (366,924 bp) (Unseld et al., 1997). The GC content is 44.71%, close to 44.8% of *Arabidopsis thaliana* (Unseld et al., 1997), 45.4% of *Millettia pinnata* (Kazakoff et al., 2012) and 45.1% of *Vigna radiata* (Alverson et al., 2011). 35 unique protein coding genes, 3 rRNA genes, and 16 tRNA genes were identified from the assembly using MITOFY and AGORA organelle genome annotation platforms (Table 2.3), as shown in Figure 2.4. Four protein coding genes *rpl2, rps2, rps11* and *rps13* are completely missing from the mitochondrial genome of marama (Table 2.4). Another two protein coding genes *rps7* and *rps19* are present as pseudogenes since each contains an internal stop codon. The genes *atp1, atp8* and one open reading frame (ORF) of *nad6* have two copies since they are located on long repeats H, J and I separately. The two copies of *atp1* have ORFs of different lengths since both stop codons are outside the repeat, and one of them may be a pseudogene (Figure S2.18). Whether marama mtDNA recombination and structural variation have altered the expression of these genes remains to be determined. Genes including *nad1, nad2, nad4, nad5, nad7,* and *ccmFc* contain introns as shown in Figure 2.4. The tRNA gene *trnfM-CAT* has two copies in the marama mitochondrial genome, while it has four copies in the mitogenomes of soybean (Chang et al., 2013). Some mitochondrial genes have been lost during the evolution of legumes (Table 2.4). For example, *rpl10* is absent from the mitochondrial genomes of *Millettia pinnata* and *Vigna radiata,* however, it is present and functional in marama (Kazakoff et al., 2012; Alverson et al., 2011). Genes *cox2* and *rps1* are missing in the mtDNA of *Vigna radiata* and *Lotus japonicus*, respectively, but remain intact in marama, *Millettia pinnata* and *Glycine max* (Kazakoff et al., 2012; Chang et al., 2013). *sdh4* exists as a pseudogene in

the four legumes, but it was found to be intact in the mitogenome of marama (Chang et al., 2013; Kazakoff et al., 2012). A phylogenetic tree was drawn based on 29 chloroplast protein coding genes to study the evolutionary relationship between marama and other related plants (Kim and Cullis, 2016). A similar phylogenetic study can be carried out in the future on the mitochondrial genes mentioned above to further verify the taxonomic relationship between marama and other Fabaceae species, which will provide insight into how marama evolved.

Table 2. 4 Comparison of some protein coding genes known to be variable between Fabaceae species.

Gene	*Tylosema esculentum*	*Millettia pinnata*	*Vigna radiata*	*Lotus japonicus*	*Glycine max*
sdh3	+	+	-	-*	-
sdh4	+	-*	-*	-*	-*
cox2	+	+	-	+	+
rpl2	-	-*	-	-	-
rpl10	+	-	-	-*	-
rps1	+	+	+	-	+
rps2	-	-	-	-	-
rps7	-*	-*	-*	-*	-*
rps10	+	+	+	+	+
rps11	-	-	-	-	-
rps13	-	-	-	-	-
rps19	-*	-*	-*	-*	-*

+ exist and functional, - completely missing, -* present as pseudogene

2.5 Supplementary materials

Figure S2.1 IGV visualization of the long repeat owned by circle 28 (M2) and circle LS1a2 (M1) shows the coverage is doubled. The Illumina reads of individual A11 were aligned to Circle 28 using Bowtie 2 and the result was visualized in IGV after converting the sam file to a sorted bam file by Samtools 1.7. The number of reads increased from 700 to 1600 after 39,200 bp where the repeat is located, as this region is owned by both the molecules circle 28 and LS1a2.

Figure S2.2 IGV visualization of the long repeat shared by circle LS1a2 (M1) and circle 313 (M3) displays an increase of the coverage at this region. The Illumina raw reads of individual A11 were aligned to Circle LS1a2 using Bowtie 2 and the result was visualized in IGV after converting the sam file to a sorted bam file by Samtools 1.7. The read count increased from 750 to 1500 for the sequence between 42,500 bp and 46,500 bp, where the overlap between circle LS1a2 and circle 313 is located.

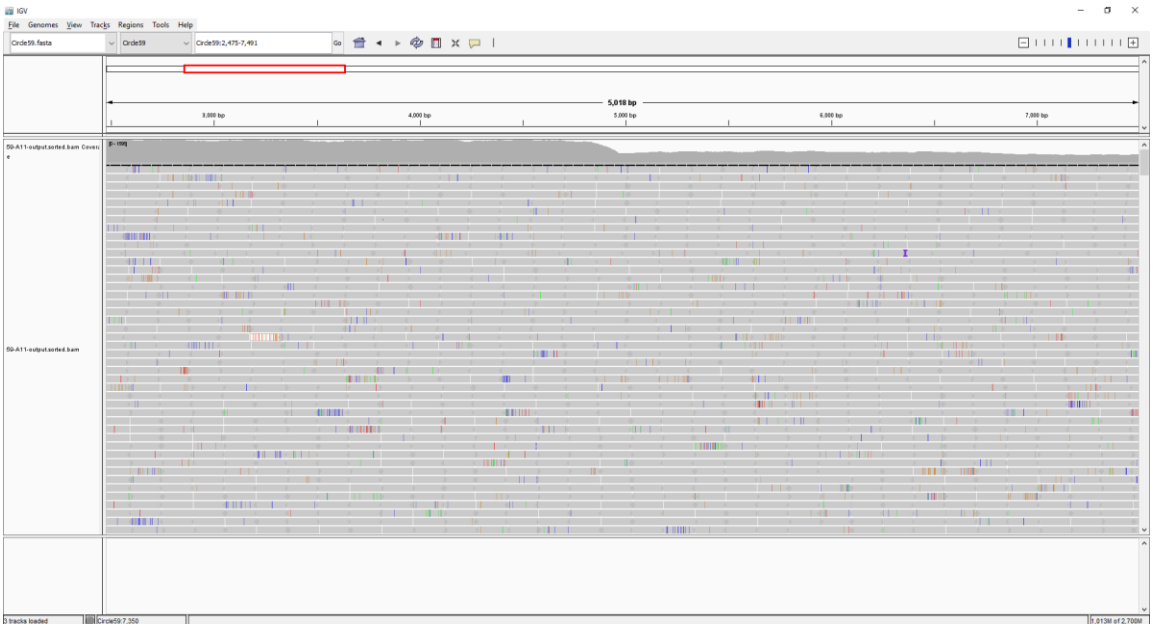

Figure S2.3 IGV visualization of the long repeat shared by circle 59 (M4) and circle 192 (M5) shows an increase of the coverage at the overlap. The Illumina raw reads of individual A11 were aligned to Circle 59 using Bowtie 2 and the result was visualized in IGV after converting the sam file to a sorted bam file by Samtools 1.7. The sequencing depth is about 1,500 (two copies) for the sequence before 5,000 bp, since this piece represents the repeat owned by both the molecules circle 59 and circle 192. However, the sequence after 5,000 bp is exclusive to circle 59, so the number of reads is decreased to one copy 750.

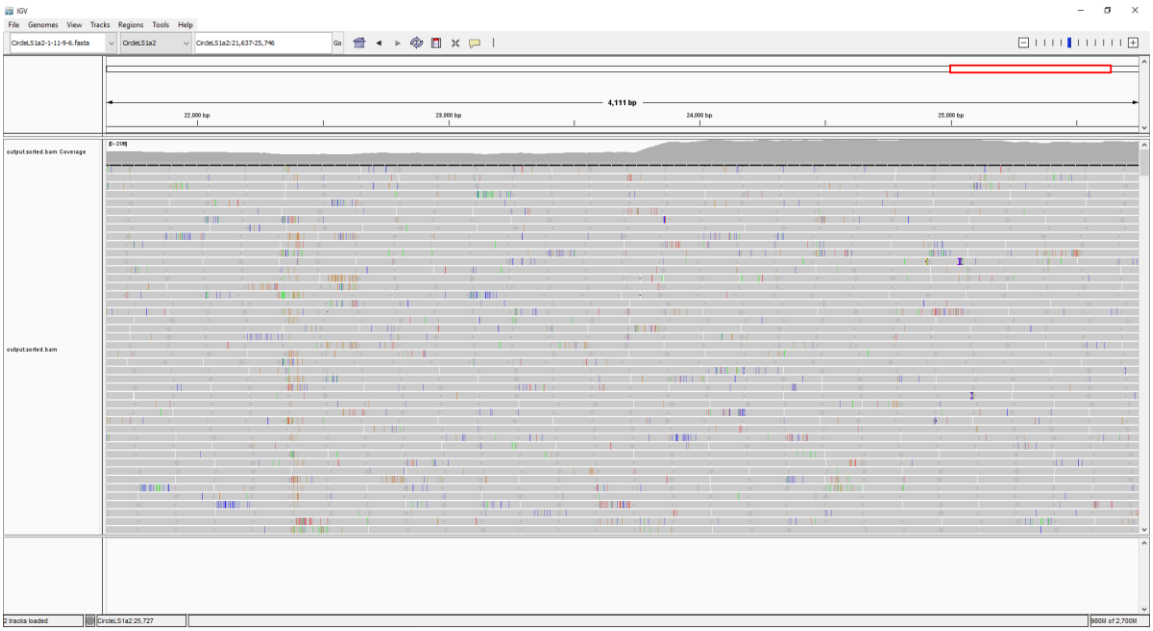

Figure S2.4 IGV visualization of a part of the 2,351bp inverted repeat on LS1a2 (M1) showing a doubled coverage. The Illumina raw reads of individual A11 were aligned to a part of the circle LS1a2 (26 kb fragment) using Bowtie 2 and the result was visualized in IGV after converting the sam file to a sorted bam file by Samtools 1.7. The sequencing depth of the sequence after 23,750 bp (where the repeat is located) is increased from 800 to 1,800.

Name	Accession	Description	Interval	E-value
Photo_RC super family	cl08220	D1, D2 subunits of photosystem II (PSII); M, L subunits of bacterial photosynthetic reaction ...	4-183	8.71e-37
PSII super family	cl08223	Photosystem II protein;	134-568	1.13e-89
PSII super family	cl08223	Photosystem II protein;	570-1526	0e+00
psbZ	CHL00082	photosystem II protein Z	2199-2384	6.42e-19
psaB	CHL00054	photosystem I P700 chlorophyll a apoprotein A2	3715-5907	0e+00
psaA	CHL00056	photosystem I P700 chlorophyll a apoprotein A1	5941-8115	0e+00
ycf3 super family	cl31773	photosystem I assembly protein Ycf3	8899-9057	1.24e-26
rps14	CHL00074	ribosomal protein S14	3309-3608	1.94e-53

Figure S2.5 Conserved domains found in the 9,798 bp chloroplast DNA insertion in the mitogenome of marama. Conserved domains were searched on the long cpDNA insertion on NCBI (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Full length chloroplast genes *psaA*, *psaB* and part of *psbC* and *rps14* were found on this fragment.

12 contigs were got from the assembly of PacBio reads using Canu v2.2. All the contigs were aligned to the structure units of LS1 and LS2 independently to confirm the existence of the structures. The complete sequence of the subgenomic structure could not be obtained directly from Canu due to the insufficient read coverage at some places, where the connections have been verified by the assembly of Illumina reads.

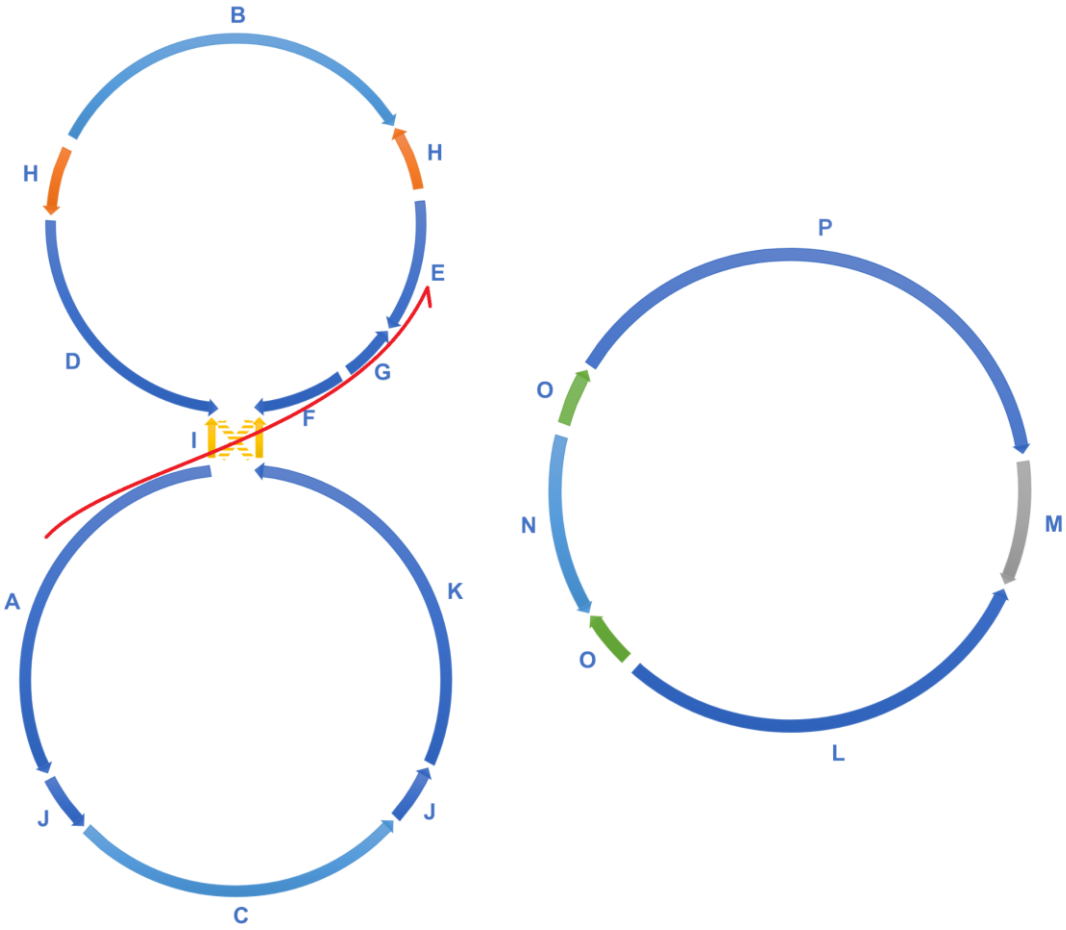

Figure S2.6 The mitochondrial contig obtained from Canu verified the connection of some structural units. Read 78 with a length of 21,293 bp was got from the direct assembly of the 1.78 Gb PacBio reads by Canu 2.2 (GenomeSize =20m, correctedErrorRate=0.15). The sequence was BLASTed with the structural units on NCBI, and a red curve on the mtDNA structure diagram was used to show its position.

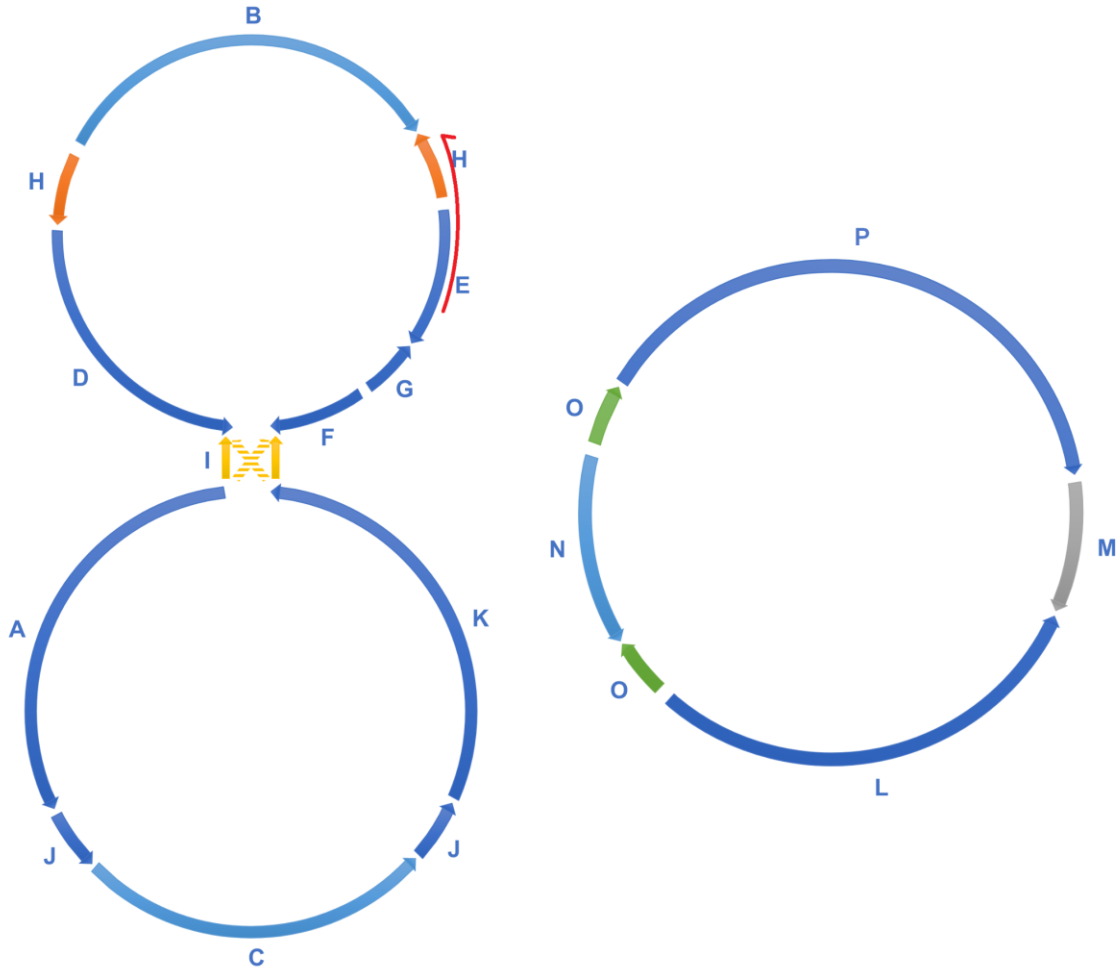

Figure S2.7 The mitochondrial contig obtained from Canu verified the connection of some structural units. Read 117 with a length of 25,771 bp was got from the direct assembly of the 1.78 Gb PacBio reads by Canu 2.2 (GenomeSize =20m, correctedErrorRate=0.15). The sequence was BLASTed with the structural units on NCBI, and a red curve on the mtDNA structure diagram was used to show its position.

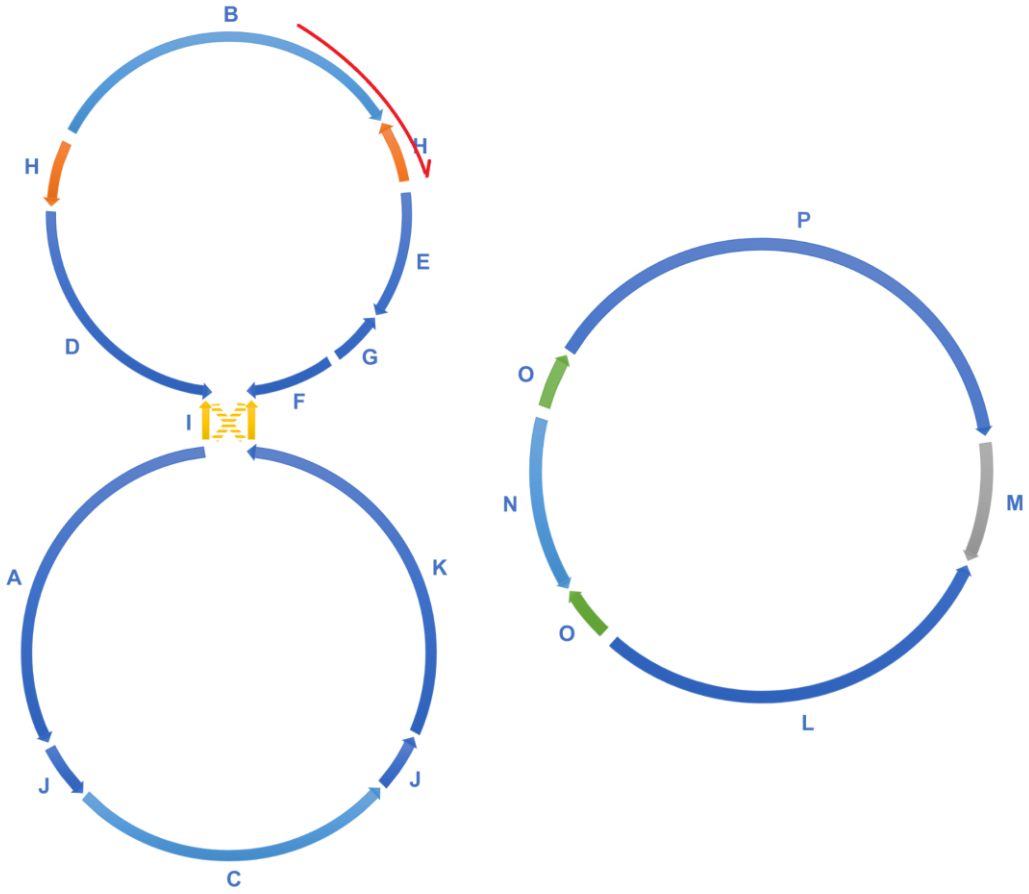

Figure S2.8 The mitochondrial contig obtained from Canu verified the connection of some structural units. Read 57 with a length of 18,913 bp was got from the direct assembly of the 1.78 Gb PacBio reads by Canu 2.2 (GenomeSize =20m, correctedErrorRate=0.15). The sequence was BLASTed with the structural units on NCBI, and a red curve on the mtDNA structure diagram was used to show its position.

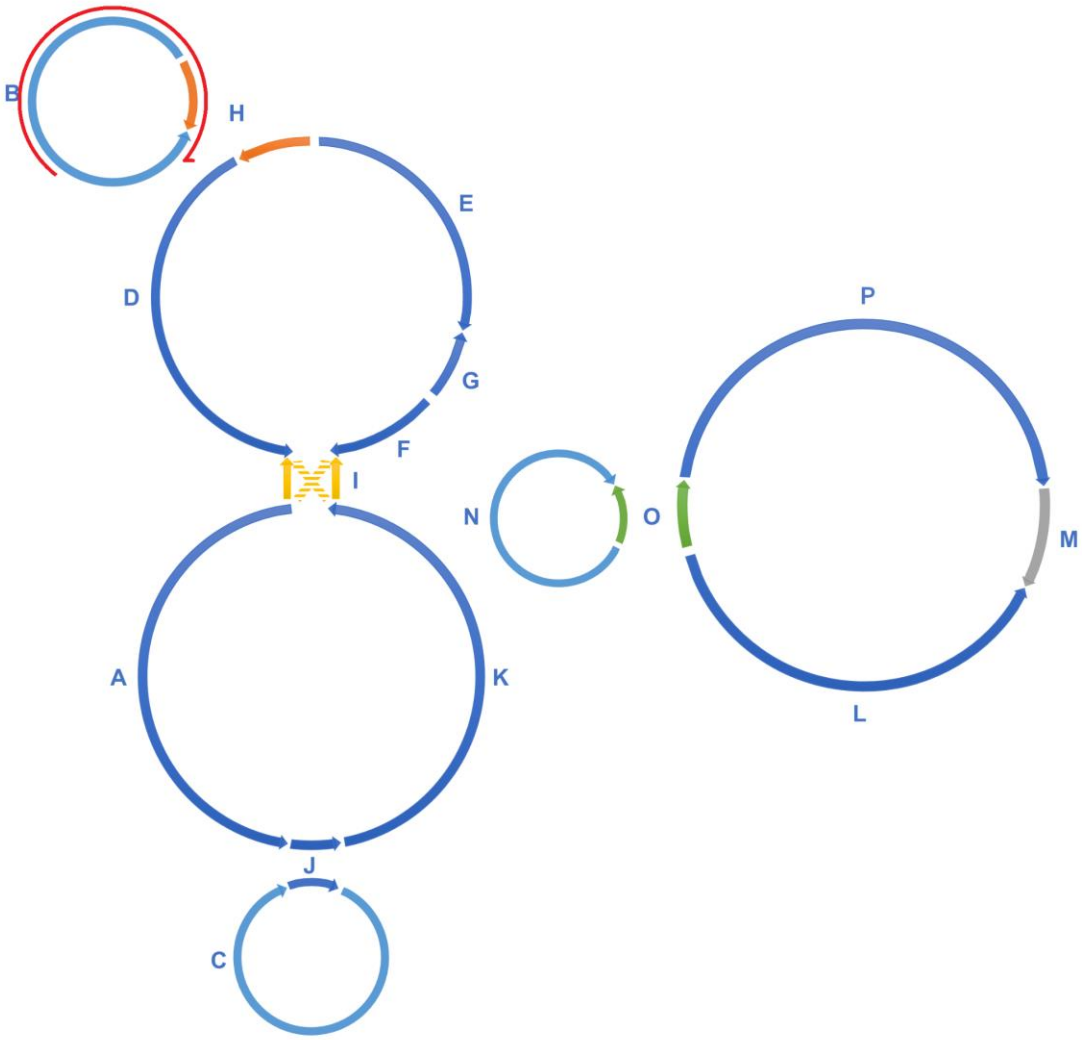

Figure S2.9.1 The mitochondrial contig obtained from Canu verified the connection of some structural units. Read 111 with a length of 29,553 bp was got from the direct assembly of the 1.78 Gb PacBio reads by Canu 2.2 (GenomeSize =20m, correctedErrorRate=0.15). The sequence was BLASTed with the structural units on NCBI, and a red curve on the mtDNA structure diagram was used to show its position.

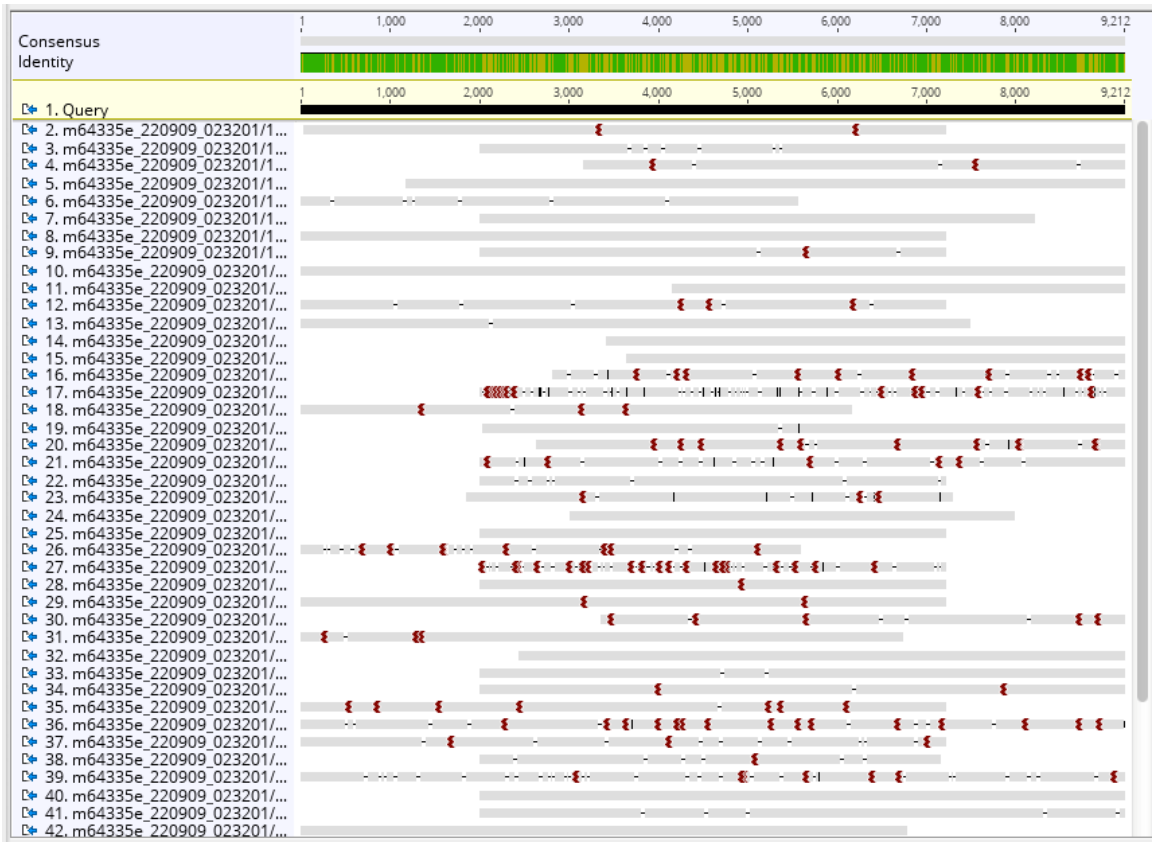

Figure S2.9.2 Blast results of sequence B-H-B to the database built based on sample 32 PacBio HiFi reads in Geneious 9. The obtained reads confirmed the four connection types of this region, BH(D), (E)HB, (E)H(D), and BHB. B-H-B verified the existence of independent BH ring.

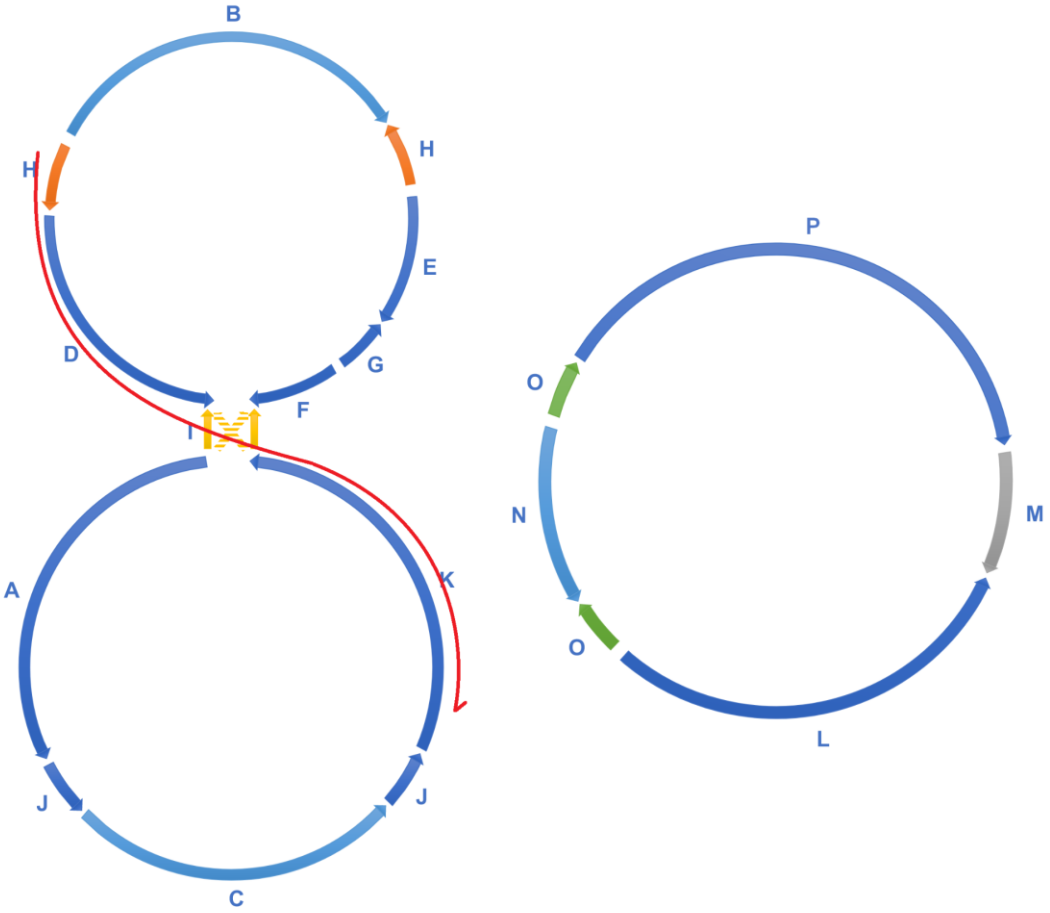

Figure S2.10 The mitochondrial contig obtained from Canu verified the connection of some structural units. Read 297 with a length of 70,055 bp was got from the direct assembly of the 1.78 Gb PacBio reads by Canu 2.2 (GenomeSize =20m, correctedErrorRate=0.15). The sequence was BLASTed with the structural units on NCBI, and a red curve on the mtDNA structure diagram was used to show its position.

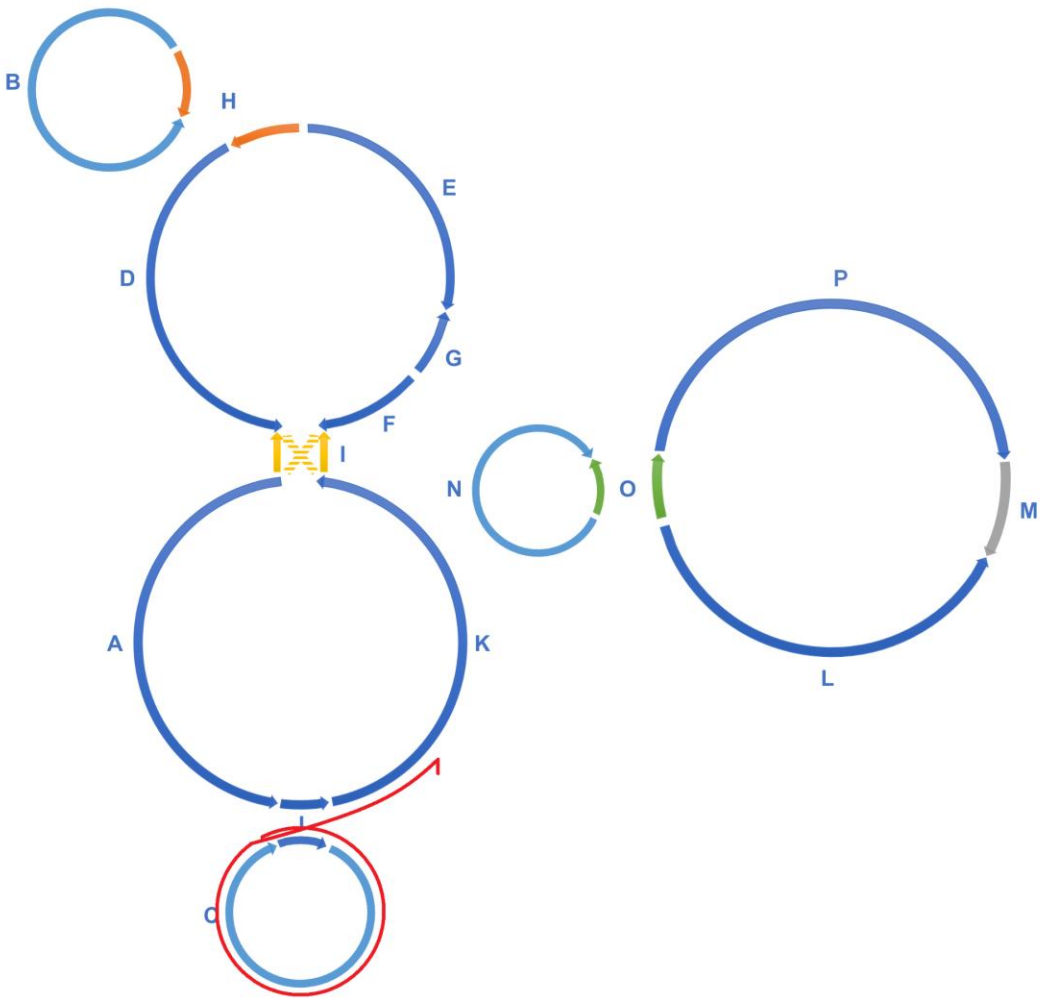

Figure S2.11 The mitochondrial contig obtained from Canu verified the connection of some structural units. Read 186 with a length of 55,037 bp was got from the direct assembly of the 1.78 Gb PacBio reads by Canu 2.2 (GenomeSize =20m, correctedErrorRate=0.15). The sequence was BLASTed with the structural units on NCBI, and a red curve on the mtDNA structure diagram was used to show its position.

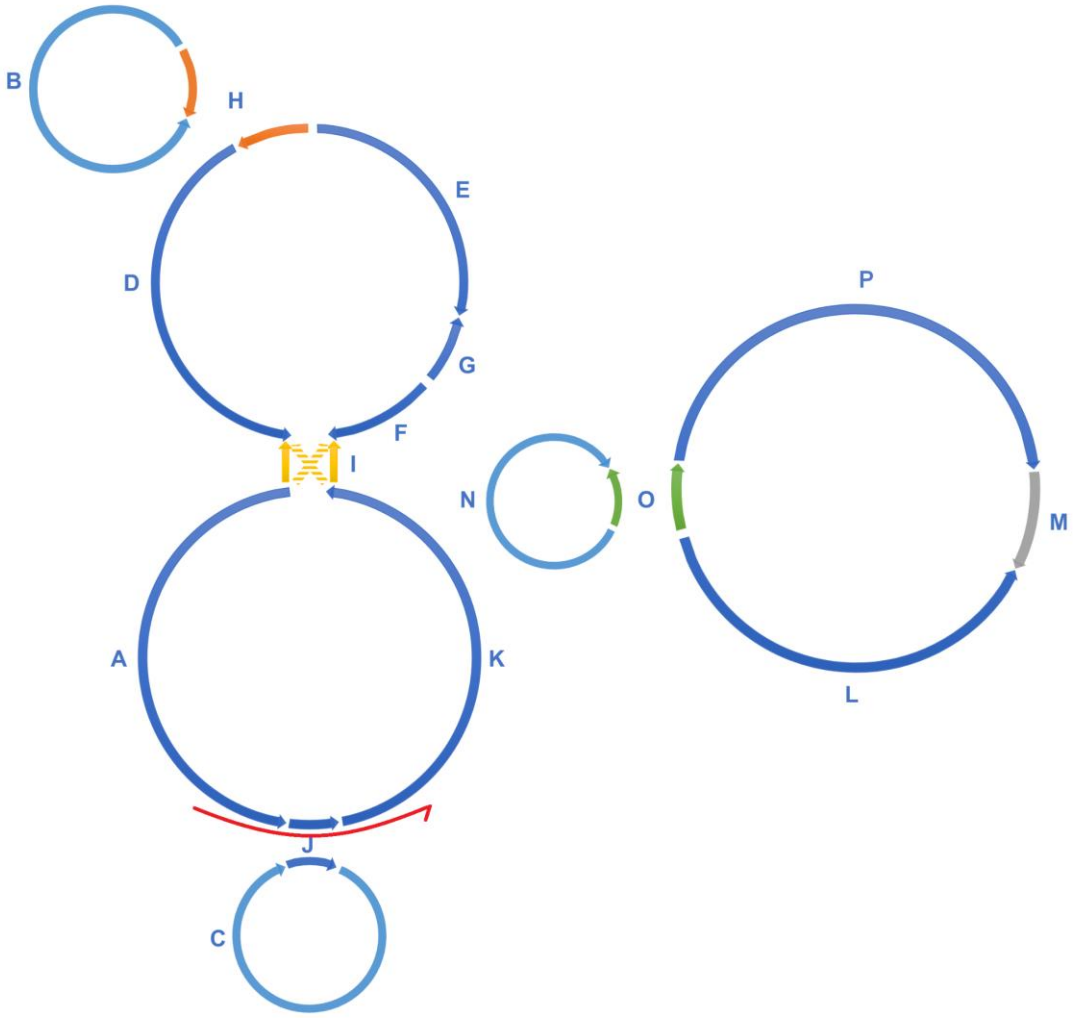

Figure S2.12.1 The mitochondrial contig obtained from Canu verified the connection of some structural units. Read 22 with a length of 14,304 bp was got from the direct assembly of the 1.78 Gb PacBio reads by Canu 2.2 (GenomeSize =20m, correctedErrorRate=0.15). The sequence was BLASTed with the structural units on NCBI, and a red curve on the mtDNA structure diagram was used to show its position.

Figure S2.12.2 Blast results of sequence C-J-C to the database built based on sample 32 PacBio HiFi reads in Geneious 9. The obtained reads confirmed the four connection types of this region, CJ(K), (A)JC, (A)J(K), and CJC. C-J-C verified the existence of independent CJ ring.

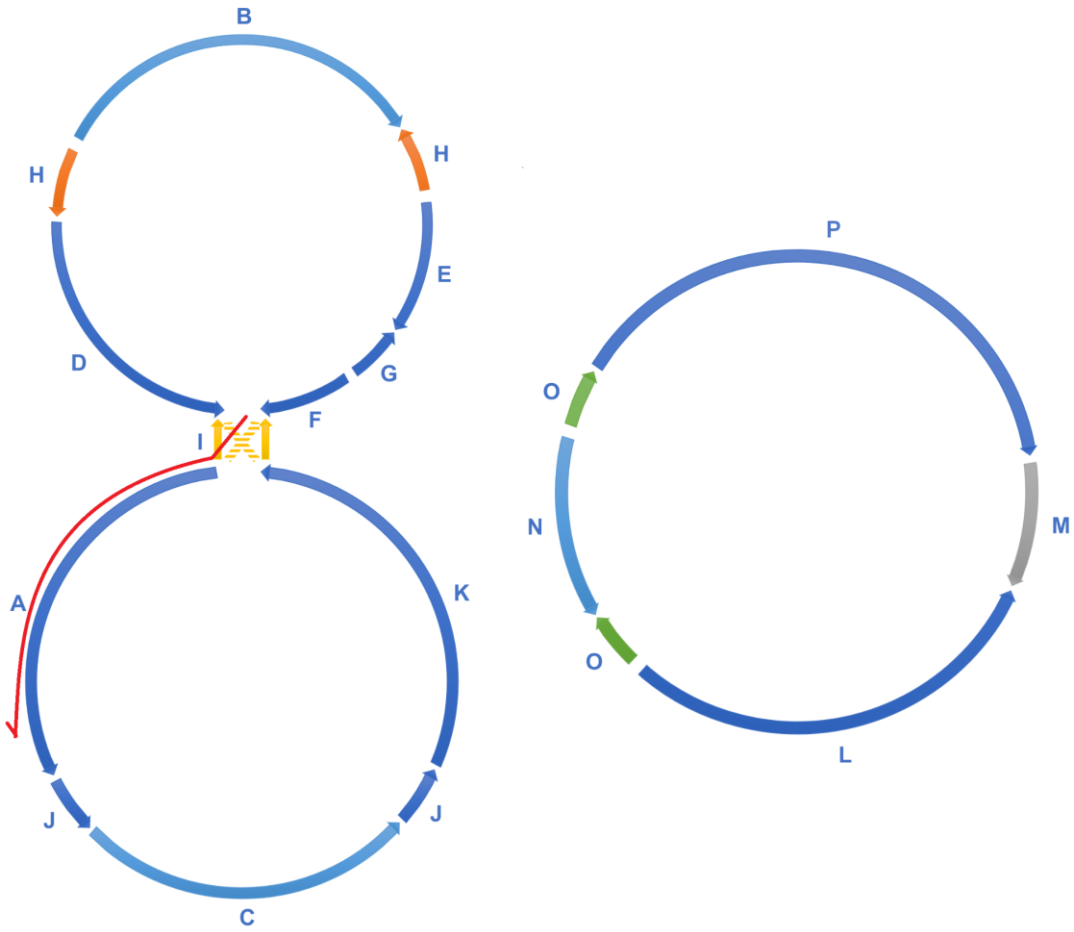

Figure S2.13 The mitochondrial contig obtained from Canu verified the connection of some structural units. Read 130 with a length of 31,102 bp was got from the direct assembly of the 1.78 Gb PacBio reads by Canu 2.2 (GenomeSize =20m, correctedErrorRate=0.15). The sequence was BLASTed with the structural units on NCBI, and a red curve on the mtDNA structure diagram was used to show its position.

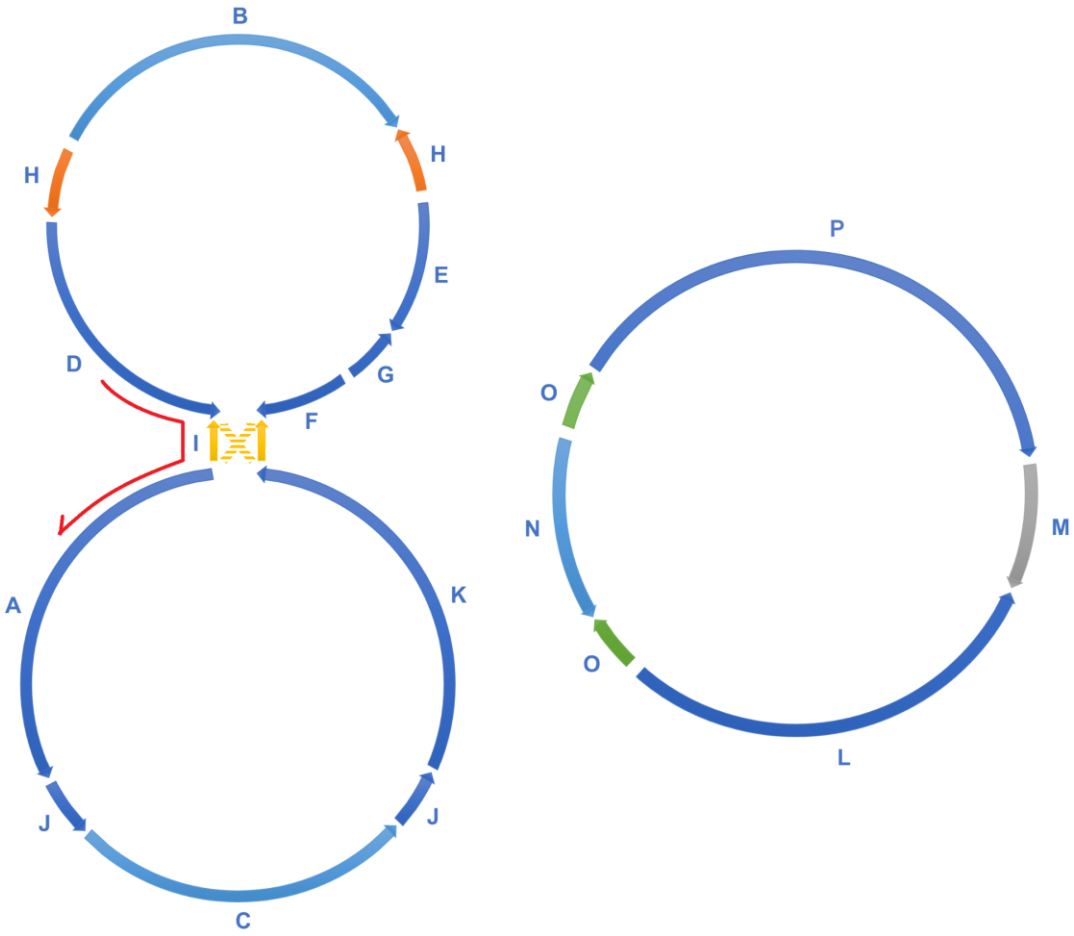

Figure S2.14 The mitochondrial contig obtained from Canu verified the connection of some structural units. Read 4 with a length of 7,143 bp was got from the direct assembly of the 1.78 Gb PacBio reads by Canu 2.2 (GenomeSize =20m, correctedErrorRate=0.15). The sequence was BLASTed with the structural units on NCBI, and a red curve on the mtDNA structure diagram was used to show its position.

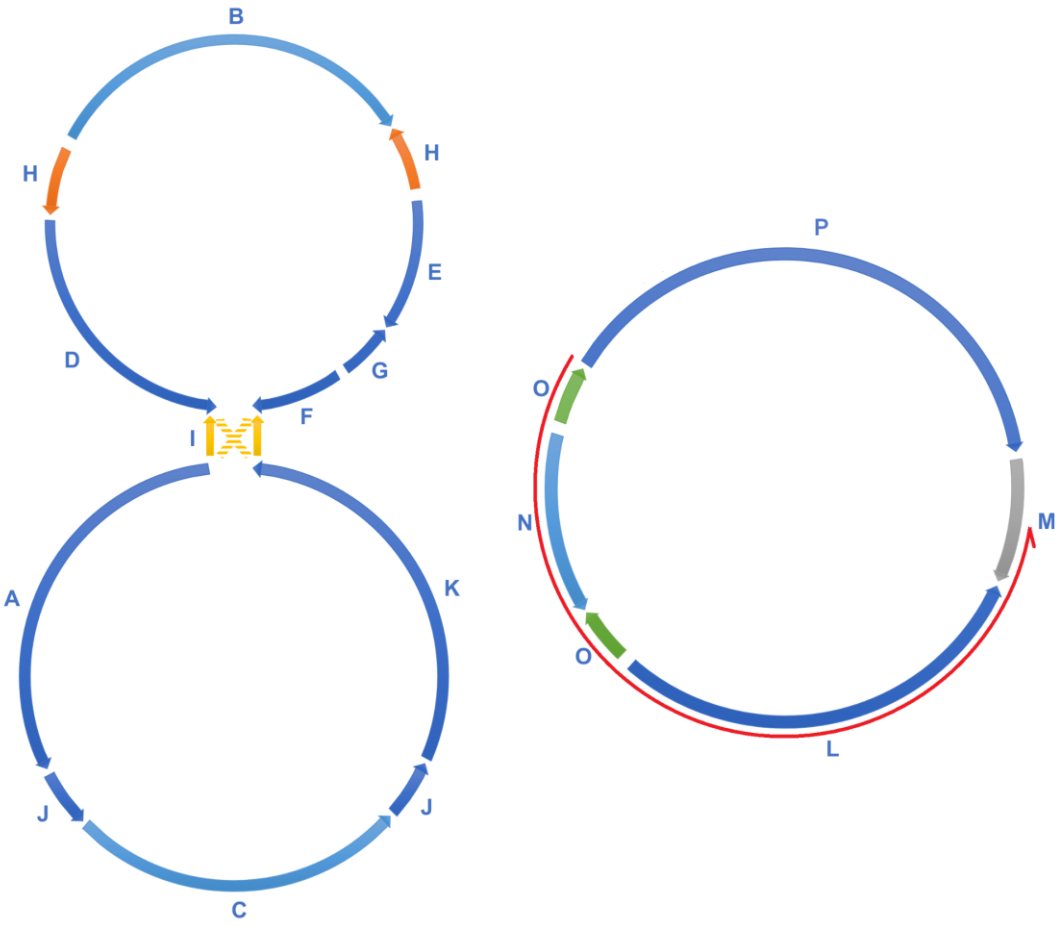

Figure S2.15.1 The mitochondrial contig obtained from Canu verified the connection of some structural units. Read 371 with a length of 84,053 bp was got from the direct assembly of the 1.78 Gb PacBio reads by Canu 2.2 (GenomeSize =20m, correctedErrorRate=0.15). The sequence was BLASTed with the structural units on NCBI, and a red curve on the mtDNA structure diagram was used to show its position.

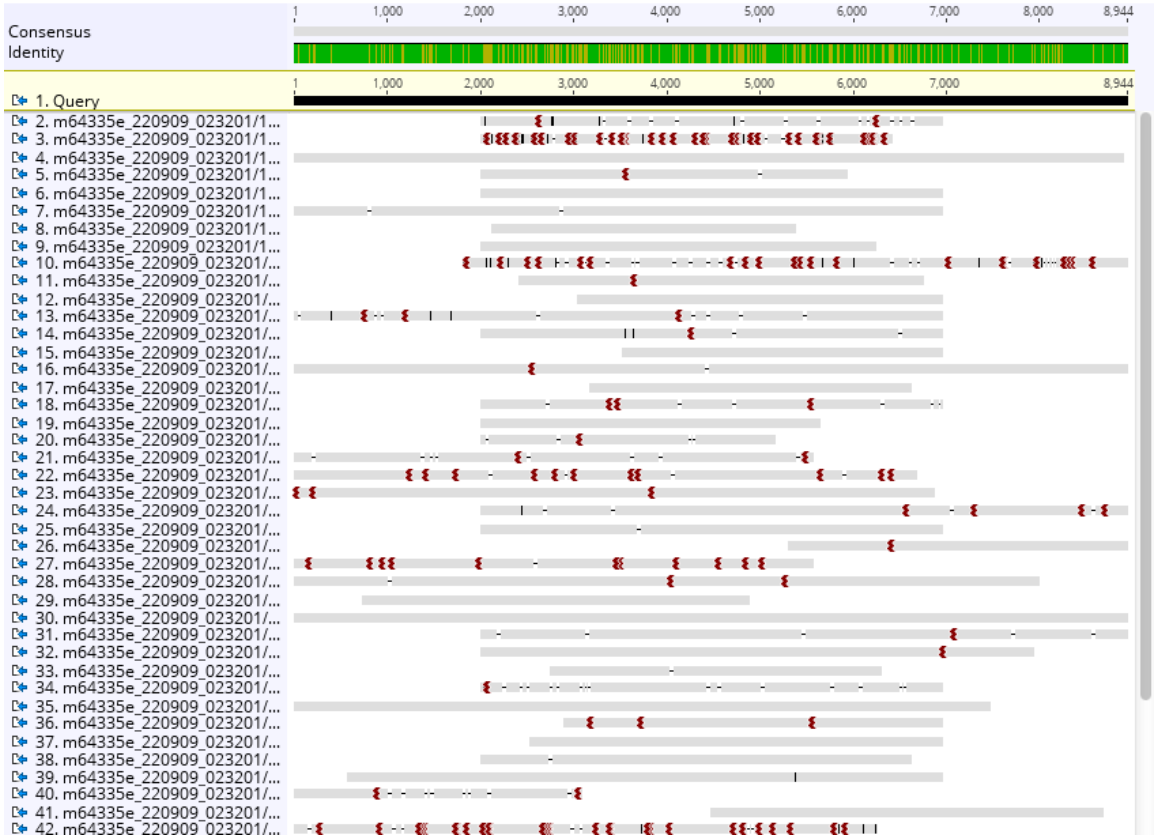

Figure S2.15.2 Blast results of sequence N-O-N to the database built based on sample 32 PacBio HiFi reads in Geneious 9. The obtained reads confirmed the four connection types of this region, NO(P), (L)ON, (L)O(P), and NON. N-O-N verified the existence of independent ON ring.

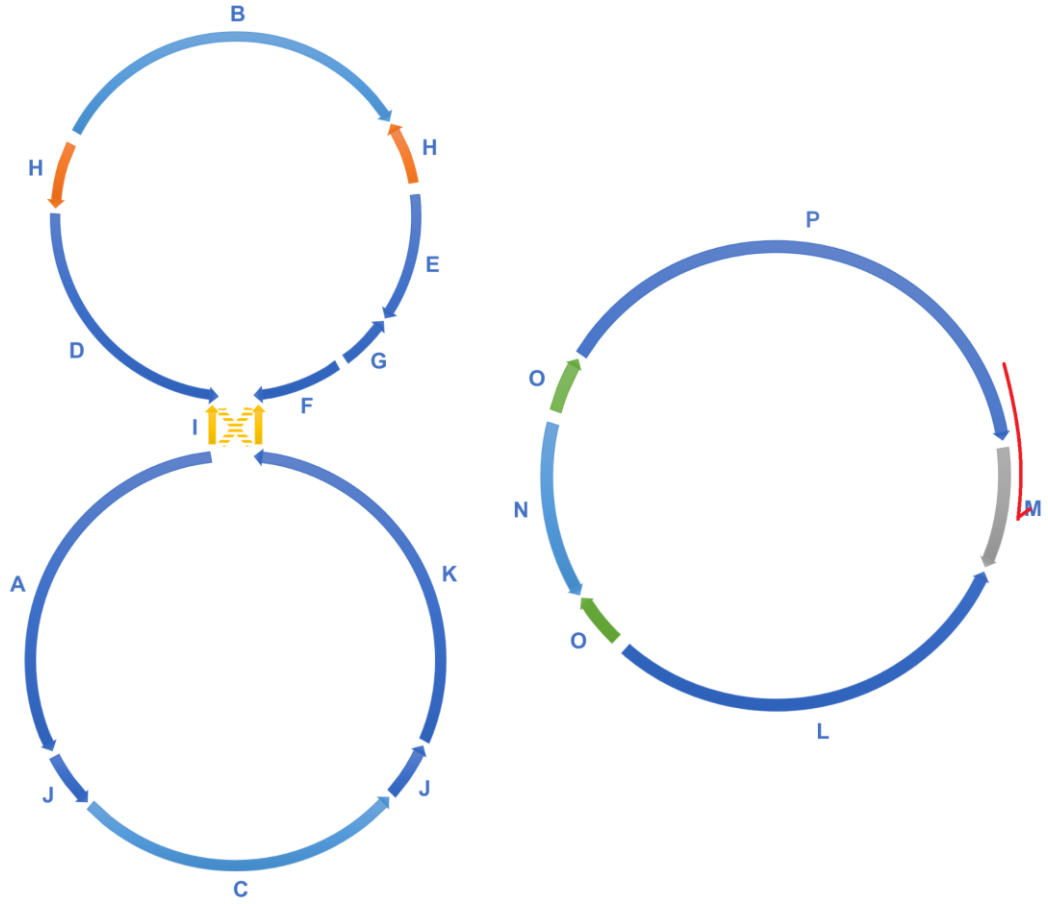

Figure S2.16 The mitochondrial contig obtained from Canu verified the connection of some structural units. Read 53 with a length of 11,701 bp was got from the direct assembly of the 1.78 Gb PacBio reads by Canu 2.2 (GenomeSize =20m, correctedErrorRate=0.15). The sequence was BLASTed with the structural units on NCBI, and a red curve on the mtDNA structure diagram was used to show its position.

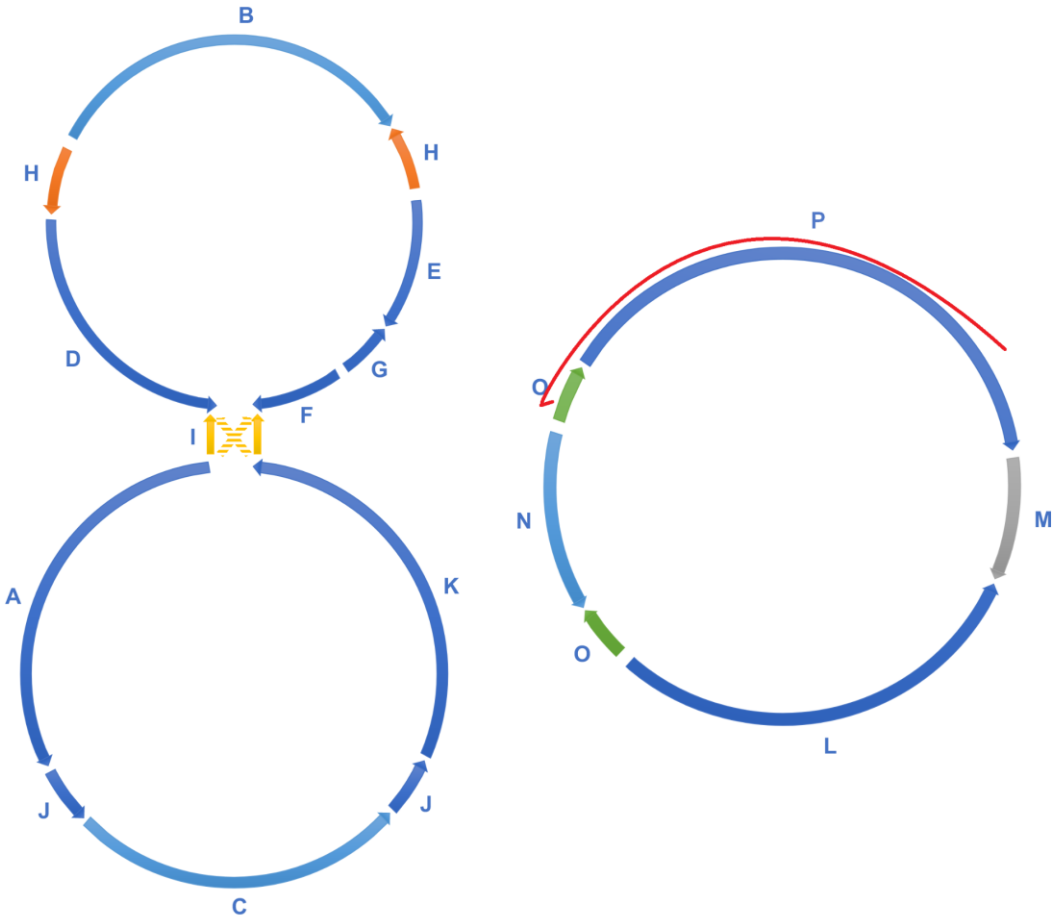

Figure S2.17 The mitochondrial contig obtained from Canu verified the connection of some structural units. Read 231 with a length of 53,705 bp was got from the direct assembly of the 1.78 Gb PacBio reads by Canu 2.2 (GenomeSize =20m, correctedErrorRate=0.15). The sequence was BLASTed with the structural units on NCBI, and a red curve on the mtDNA structure diagram was used to show its position.

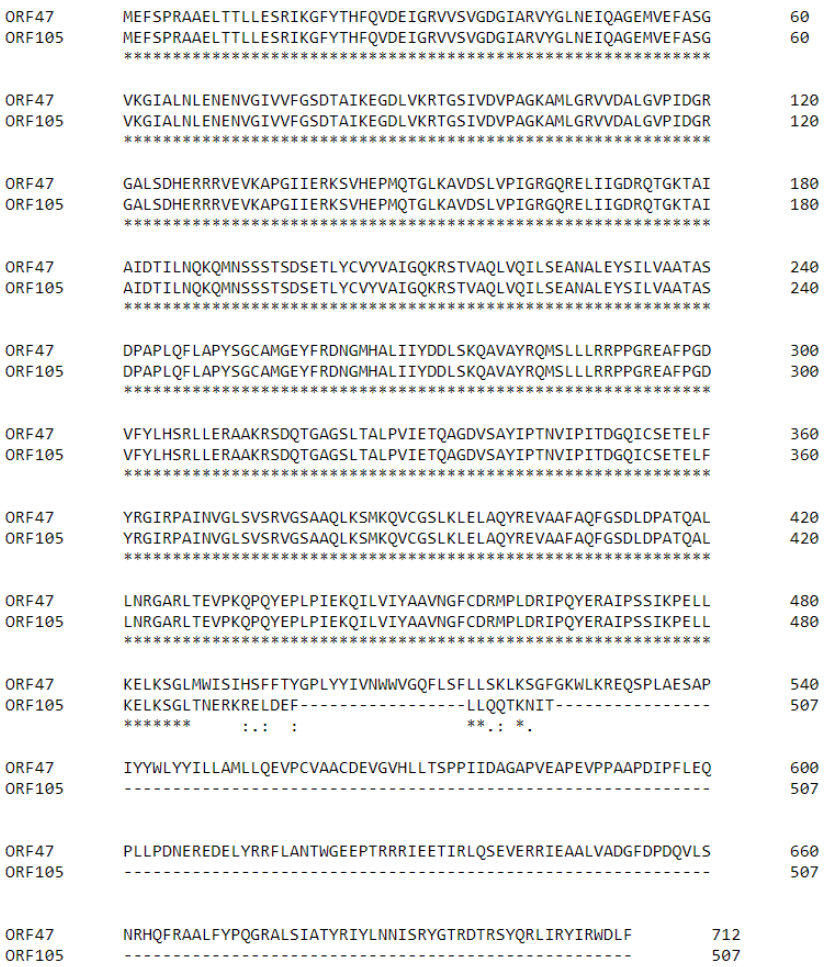

Figure S2.18 Protein alignment between the two open reading frames ORF47 (M3) and ORF105 (M1) of marama mitochondrial gene *atp1*. The ORF of *atp1* on circle 313 (M3) is 205 amino acids longer than the one on circle LS1a2 (M1). The query coverage is 68% with an identity of 96.65%.

Chapter 3. Comparative Analysis of 84 Chloroplast Genomes of *Tylosema esculentum* Reveals Two Distinct Cytotypes

Published in *Frontiers in Plant Science* Vol. 13, No. 1025408, 2022.
DOI: 10.3389/fpls.2022.1025408
Authors: Jin Li and Christopher Cullis*
Department of Biology, Case Western Reserve University, Cleveland, OH 44106, USA
* Correspondence: cac5@case.edu
(Received 22 August 2022, Accepted 21 December 2022, Published 31 January 2023)

3.1 Abstract

Tylosema esculentum (marama bean) is an important orphan legume from southern Africa that has long been considered to have the potential to be domesticated as a crop. The chloroplast genomes of 84 marama samples collected from various geographical locations in Namibia and Pretoria were compared in this study. The cp genomes were analyzed for diversity, including SNPs, indels, structural alterations, and heteroplasmy. The marama cp genomes ranged in length from 161,537 bp to 161,580 bp and contained the same sets of genes, including 84 protein-coding genes, 37 tRNA genes, and 8 rRNA genes. The genes *rpoC2*, *rpoB*, and *ndhD*, and the intergenic spacers *trnT-trnL* and *ndhG-ndhI* were found to be more diverse than other regions of the marama plastome. 15 haplotypes were found to be divided into two groups, differing at 122 loci and at a 230 bp inversion. One type appears to have greater variability within the major genome present, and variations amongst individuals with this type of chloroplast genome seems to be distributed within specific geographic regions but with very limited sampling for some regions. However, deep sequencing has identified that within most of the

individuals, both types of chloroplast genomes are present, albeit one is generally at a very low frequency. The inheritance of this complex of chloroplast genomes appears to be fairly constant, providing a conundrum of how the two genomes co-exist and are propagated through generations. The possible consequences for adaptation to the harsh environment in which *T. esculentum* survives are considered. The results pave the way for marama variety identification, as well as for understanding the origin and evolution of the bean.

Keywords: *Tylosema esculentum*, legume, chloroplast genome, genetic diversity, genomic structure, phylogenetic analysis, heteroplasmy, plastome evolution

3.2 Introduction

Chloroplasts are important organelles in plants and are thought to have originated from the endosymbiosis of cyanobacteria in eukaryotes 1.2-1.5 billion years ago (Dyall et al., 2004). The functions of chloroplasts not only include photosynthesis, but also participate in the regulation of plant responses to stress conditions such as heat and drought (Estavillo et al., 2011; Song et al., 2021). The chloroplast genome is commonly configured as a double-stranded circular molecule. It consists of four parts, a small single-copy region (SSC) and a large single-copy region (LSC), separated by a pair of inverted repeats (IRa and IRb) with a small genomic size ranging from 120 to 170 kb (Palmer, 1991; Smith, 2017). The inverted repeats are thought to play an important role in maintaining chloroplast genome stability and conserved sequence arrangement (Palmer and Thompson, 1982). Chloroplast genome size is largely determined by the expansion, contraction and even loss of IR, which is particularly pronounced in legumes (Wang et al., 2018). Compared with nuclear DNA, plant chloroplast genomes are highly conserved

with low recombination and substitution rate (Banks and Birky, 1985; Twyford and Ness, 2016). They generally contain 110-130 common genes (Jansen et al., 2005), so they are often used for phylogenetic analysis and species identification (Manos et al., 2008). Chloroplasts are semi-autonomous organelles (Dobrogojski et al., 2020), in which a large number of proteins are encoded by the nucleus. During the evolution of angiosperms, many chloroplast genes have been transferred into the nucleus, and in the original organelle genome, they are either lost or become pseudogenes (Martin, 2003). However, the remaining genes in cpDNA are thought to still play important roles in cellular activity (Drescher et al., 2000; Zhang et al., 2020).

Heteroplasmy, which refers to the coexistence of different types of organellar DNA in the same cell or individual, has previously been reported in Fabaceae (Johnson and Palmer, 1989; Lei et al., 2016; Lee et al., 2021). This could be caused by the accumulation of mutations in organellar DNA over time, or occasional non-Mendelian biparental cytoplasmic inheritance (Kondo et al., 1990; Carbonell-Caballero et al., 2015; Ramsey and Mandel, 2019; Iannello et al., 2021). In theory, this gives plants stronger adaptability allowing genotypes favored by natural selection to be retained. The organelle genetic bottleneck hypothesis suggests that allele frequencies may change rapidly during transmission from one generation to the next, possibly as a result of environmental influences (Ashley et al., 1989; Zhang et al., 2018). In other words, cells tend to pass on healthier, more adaptable organelles to offspring (Marlow, 2017). Occasionally, offspring with only mutant organellar DNA can be produced by a heteroplasmic mother. The gene *MSH1* has been found to play an important role in sorting chloroplast heteroplasmy, but this is mainly to correct for *de novo* mutations (Broz et al., 2022). Heteroplasmy arising

from paternal leakage may have different fates. The study of heteroplasmy will contribute to a better understanding of organelle maintenance and inheritance in marama.

The genome size of marama bean was found to be small in legumes, about 1 Gb, calculated based on the size of the WGS data and the corresponding genome coverage (Cullis et al., 2019). Feulgen staining indicated that marama may be an ancient hexaploid plant with 44 chromosomes (Takundwa et al., 2012) although it could also be an ancient tetraploid since there are legumes with a haploid chromosome number of 11 (Bandel 1974). An rDNA marker-based study has shown that marama has low inter-population diversity but high intra-population diversity, possibly due to the lack of gene flow between populations in the environment where marama grows, but the bean itself is a predominantly outcrossing plant (Nepolo et al., 2010). The chloroplast genome of marama was previously sequenced and assembled using a hybrid method based on both Illumina and PacBio datasets, with a total length of 161,537 bp (KX792933.1) (Kim and Cullis, 2017). Compared with the plastid genome of *Cercis canadensis*, a 7479 bp sequence containing the genes *rbcL*, *accD*, *psaI*, *ycf4*, *cemA*, and *petA* was inverted in the cp genome of marama, and this inversion was not seen in other legumes (Kim and Cullis, 2017). The same approach was also used to assemble the mitochondrial genome of marama, which contains a 9,798 bp long insertion of cpDNA with potentially non-functional chloroplast genes *psaA*, *psaB*, and *psbC* (Li and Cullis, 2021). This insertion has a large number of mutations compared to the original chloroplast sequence, and studying the base ratios at these loci helps to distinguish heteroplasmy in the chloroplast genome from differences in homologous sequences between different organelles in alignment.

A comparative genomics study was performed in this study on the chloroplast genomes of 84 marama individuals collected from various geographic regions (Figure 3.1; Table S3.1) to identify polymorphisms including structural, single nucleotide and indels variants. This facilitates the study of phylogenetic relationships between the different marama populations and also between the individuals, which may help us better understand the origin of marama and the impact of extreme environments on plant genome evolution. In addition, SSRs in the plastid genome were analyzed and mutations that altered the coding sequence of the gene were reported, which could be useful in the breeding of the bean. Furthermore, plastid heteroplasmy and associated inheritance were investigated by analyzing chloroplast DNA allele frequencies and by comparing the chloroplast genomes of related plants.

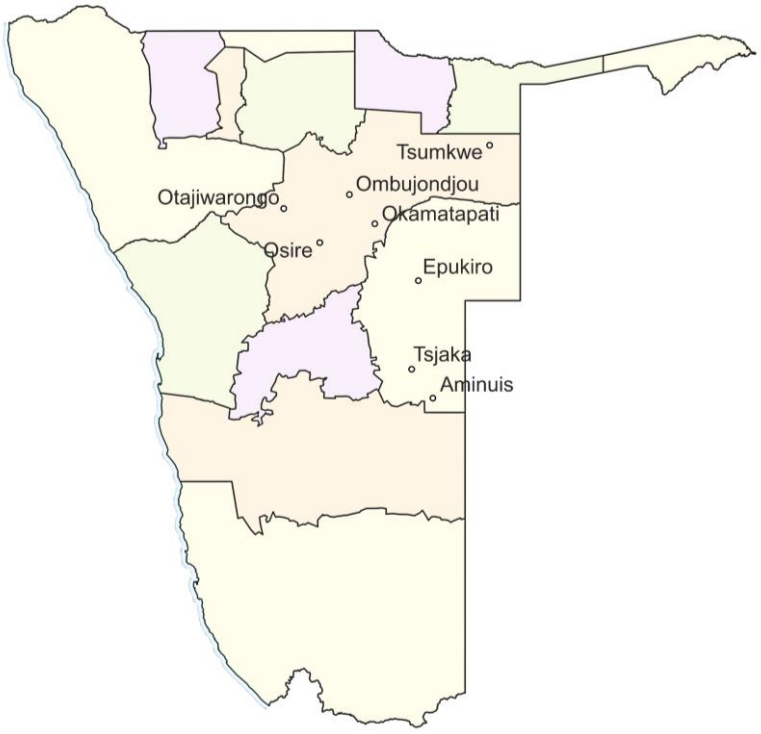

Figure 3. 1 A map of Namibia showing the eight different locations where the wild marama samples were collected.

3.3 Materials and Methods

3.3.1 Plant Materials

A total of 84 marama plants that were growing in South Africa and Namibia were sampled. Of these, 6 were plants that are growing on the University of Pretoria (UP) Farm (S25 45.490 E28 11.368) and 38 wild individuals from different geographical locations in Namibia, including Tsjaka (S22 75.039 E19 20.712), Okamatapati (S20 40.233 E18 21.59), Aminuis (S23 38.000 E19 22.00), Osire (S21 02.031 E17 21.244), Tsumkwe (S19 21.000 E20 16.000), Ombujondjou (S20 18.600 E17 58.525), Epukiro (S21 39.642 E19 25.092) and Otjiwarongo (S20 46.092 E16 65.123) (Figure 3.1). The

remaining 40 were progeny plants grown from seeds collected from plants from Aminuis, a small holding in Namibia designated UNAM Farm and University of Pretoria Farm (Table S3.1). Note that although these two areas are designated as farms, marama is not being cultivated as a crop, but being deliberately maintained in these two curated locations as long-term perennial plants.

The 6 plants that are currently growing on the University of Pretoria (UP) farm were collected from North-West South Africa in the 1990's but exact original location is not known. These plants differ phenotypically in a number of characteristics, including leaf size, internode length, stamen length and overall vigor (rate of stem elongation each growing season). DNA was extracted from leaves of the individual marama plants that were sampled from the various regions but not maintained as specimens. The progeny from individuals from the various regions do not have maternal data since at collection the pods were mature and all the foliage was dead, since the vegetative growth senesces each year. The plants from the Namibia small holding (designated UNAM Farm) were planted in 2010 from seed supplied by Professor P. Chimwamurombe but no identification of the origin of this seed was recorded. The DNA for the M samples (except M40) were extracted from immature seeds from a single plant for which DNA, leaf and flower stored material is available.

Therefore, the material sampled includes individual plants from various geographical locations as well as progeny from 7 different maternal individuals from various regions. Since it is expected that the chloroplast is maternally inherited, only one sample from each family is included in the analyzed data (43 samples), although full data is available for all 84 samples.

3.3.2 DNA Extraction and High-throughput Sequencing

DNA was extracted from fresh young leaves or the embryonic axis of germinating seeds. The plant material was frozen in liquid nitrogen and ground with a mortar and pestle. Total genomic DNA was extracted using the QIAGEN DNeasy® Plant Kit following the manufacturer's protocol. The quality and quantity of the produced DNA were estimated by a NanoDrop™ 8000 Spectrophotometer and by electrophoresing on a 1.0% agarose TBE gel. The DNA was also quantified by the Qubit™ 3.0 Fluorometer after mixing 5 µL of DNA with 195 µL of Qubit® working solution.

The DNA samples were sent in batches to the Génome Québec Innovation Centre, CWRU Genomics Core, and Novogene Corporation for sequencing, as described in the previous studies (Kim and Cullis, 2017; Li and Cullis, 2021). The Illumina HiSeq® 2000 PE100 and HiSeq® 2500 PE150 platforms were used to generate 10,358,444 to 358,941,018 reads equivalent to 1.0 Gb to 35.9 Gb of raw data for the 84 samples. An average of 10.7% of the reads were aligned to the chloroplast genome with a coverage of 8177×. All raw reads were uploaded to the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra) under accession number PRJNA779273.

3.3.3 Chloroplast Genome Assembly and Annotation

The reference chloroplast genome of *T. esculentum* was assembled *de novo* previously using a hybrid method (Kim and Cullis, 2017). ABySS (Simpson et al., 2009; https://www.bcgsc.ca/resources/software/abyss) was first used to get contigs from the NGS short reads, then the contigs were mapped to the 3GS PacBio long reads by DBG2OLC (Ye et al., 2016; https://github.com/yechengxi/DBG2OLC) and further

assembled according to sequence overlaps. The chloroplast reference genome was re-annotated in this study using BLAST, Expasy (Gasteiger, 2003; https://web.expasy.org/translate), and CPGAVAS2 (Shi et al., 2019; http://47.96.249.172:16019/analyzer/annotate), and the results were uploaded to NCBI GenBank with accession ID KX792933.1.

3.3.4 Comparative Analysis of Chloroplast Genomes

The paired-end Illumina reads of the 84 individuals were aligned to the *T. esculentum* chloroplast reference genome using Bowtie 2.2.1 (Langmead and Salzberg, 2012) on CyVerse Discovery Environment platform (https://de.cyverse.org). SAM_to_Sorted_BAM-0.1.19 was used for format changes and indexing the BAM files. SNP and indels were obtained using Calling SNPs INDELs with SAMtools BCFtools and BAM-to-SHOREmap3.8, available on CyVerse. The variation data (including reference/variant alleles and corresponding genomic positions and frequencies) from all generated VCF files were transferred into the same spreadsheet. The identified variations were further verified by visualizing the alignments in IGV. Only variants which had Phred scores above 20 and were present in strands in both directions were retained to avoid interference from low-quality reads and strand bias. Low-frequency heteroplasmic polymorphisms were manually identified in IGV with a cutoff frequency of 2%.

The complete chloroplast genomes of the 84 marama individuals were obtained by correcting the reference cp genome according to the alignment result with the help of contigs from directly assembling the Illumina reads using ABySS.

A sliding window analysis was performed in DnaSP6 (Rozas et al., 2017; http://www.ub.edu/dnasp), with a window length of 1200 bp and a step size of 400 bp, to calculate the nucleotide diversity (Pi) of the cp genomes from the 43 independent marama individuals collected in various geographic regions (Table S3.1).

Intraspecific and interspecific divergent regions in the chloroplast genome were detected by mVISTA (Frazer et al., 2004; https://genome.lbl.gov/vista/mvista/about.shtml) using the Shuffle-LAGAN pairwise alignment algorithm. The chloroplast genomes of four selected marama samples and the related species *T. fassoglense* (NC_037767.1) were used as the input with the *T. esculentum* cp genome (KX792933.1) as the reference.

3.3.5 Simple Sequence Repeat Analysis and Phylogenetic Tree Construction

The two types of *T. esculentum* chloroplast genomes were used separately to identify microsatellites by MISA (Beier et al., 2017; https://webblast.ipk-gatersleben.de/misa). The minimum number of repetitions was set to 10, 6, 5, 5, 5, 5 for the searches of mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide repeats, respectively.

The complete chloroplast genomes of 43 independent marama individuals were aligned with the cp genomes of three outgroup species, *T. fassoglense* (NC 037767.1), *Glycine max* (NC_007942.1), and *Millettia pinnata* (NC_016708.2) by Muscle 3.8.31 (Edgar, 2004; https://www.drive5.com/muscle) with two iterations (maxiters=2). The result was used to draw a Maximum Likelihood (ML) phylogenetic tree using the Jukes-Cantor model and the Tamura-Nei model with 1000 bootstrap replicates in MEGA 11

(Kumar et al., 2008; https://www.megasoftware.net). The overall topology was validated by the Maximum Parsimony (MP) method using the Subtree Pruning Re-grafting (SPR) algorithm with 1000 bootstrap replicates in MEGA 11.

3.4 Results

3.4.1 Chloroplast Genome Characteristics

An average of 10.7% of the WGS reads mapped to the marama reference cp genome in the 84 individuals. The average read depth of the chloroplast DNA was about 8175x (Table S3.1). The estimated chloroplast genome lengths ranged from 161,537 bp to 161,580 bp, mainly due to the different lengths of LSC regions, as the remaining regions were of the same length in the 84 individuals.

Two germplasms, Type 1 and Type 2, were found in the studied marama plants with distinct chloroplast genomes. All marama plants collected from UP Farm and UNAM Farm had the exactly the same chloroplast genome (Type 2), which was different from the previously assembled marama reference chloroplast genome (KX792933.1) at 122 loci and at a 230 bp inversion. The chloroplast genomes of the remaining marama samples collected from different geographic locations in Namibia were similar but not completely identical to the reference, termed Type 1.

The LSC region of Type 2 cpDNA is 25 bp longer than the Type 1 reference cp genome built up on a Namibian individual in the previous study (Table 3.1) (Kim and Cullis, 2017). There was no difference in the number of genes contained in the cpDNA of the 84 individuals.

Table 3. 1 Comparison of genomic features of the two types *T. esculentum* chloroplasts.

	Type 1 (KX792933.1)	**Type 2** (OP271860)
Size (bp)	161,537	161,562
LSC (bp)	86,113	86,138
SSC (bp)	13,632	13,632
IR (bp)	30,896	30,896
Total genes	129 (18)	129 (18)
Protein coding genes	84 (7)	84 (7)
tRNA genes	37 (7)	37 (7)
rRNA genes	8 (4)	8 (4)
GC content (%)	36.03	36.03

The numbers in parentheses indicate the number of genes with two copies in the *T. esculentum* chloroplast genome. The chloroplast genomes of two individuals are compared here, with annotation deposited in GenBank under accession numbers KX792933.1 and OP271860, respectively.

The annotation of *T. esculentum* cpDNA has been done previously but has required corrections (Kim and Cullis, 2017). The boundaries of some previously annotated genes and newly identified tRNA genes were updated (Table S3.2-3.3). Two copies of *ycf15* should be pseudogenes since they lack valid start codons. The genes *psbL* and *ndhD* have an ACG start codon, and plant plastids have been reported to have a C~U editing mechanism to convert ACG to AUG (Hoch et al., 1991). There are two copies of *rps19*, the one on IRb is a pseudogene due to lack of complete 3' end sequence, and the other copy, starting with an alternative atypical GTG start codon on IRa and ending on LSC, is complete and functional. The chloroplast genes *rps19, psbC, ycf15,* and *infA* possess a GTG initiation codon, which was also reported before (Hirose et al., 1999).

3.4.2 Chloroplast Genome Variant Analysis

As can be seen from the distribution of variation shown in the inner circle in Figure 3.2, numerous variations are located in the LSC and SSC regions. However, the region where the two large inverted repeats are located has barely no variation except that each contains a locus with three consecutive base substitutions. All orange line segments have the same length, indicating that all Type 2 plants are different from the reference cp genome at these loci.

Figure 3. 2 Circular gene map of the plastid genomes of *T. esculentum* drawn by OGDRAW. Genes inside the circle are transcribed clockwise, while genes outside the circle are transcribed counter clockwise. Genes are colored according to their function. Genes with introns are marked with an asterisk*. GC content is indicated by the dark grey shading of the inner circle. The two inverted repeats are highlighted by gradient colors. The loci where Type 2 plants and Type 1 plants differ are represented by the orange lines in the inner circle and loci with differences in Type 1 plants from different geographical

regions are represented by the purple lines. The length of the lines reflects the number of individuals with the alternative alleles.

As shown in Figure 3.3, a total of 147 variants were distributed among 15 haplotypes (Table S3.4). Based on these variations, the chloroplast genomes of these plants can be clearly divided into two categories. The two types differ from each other at 122 loci (Table 3.2). No differences were found between the 7 Type 2 plants, but the 36 Type 1 individuals could then be divided into subgroups based on the other 21 loci. Individuals from the same geographic location are likely to have the same cpDNA. Some substitutions appeared only in plants from the same geographic location.

Table 3. 2 Total counts of variation in the chloroplast genomes by Calling SNPs with Samtools when aligning the cpDNA of Type 1 and Type 2 plants to the previously published marama reference chloroplast genome (KX792933.1).

Variation Type	**Type 2 vs. Ref.**	**Type 1 vs. Ref.**
Deletion	17	3
Insertion	19	5
SNP	90	17
Total	126 (122exclusive)	25

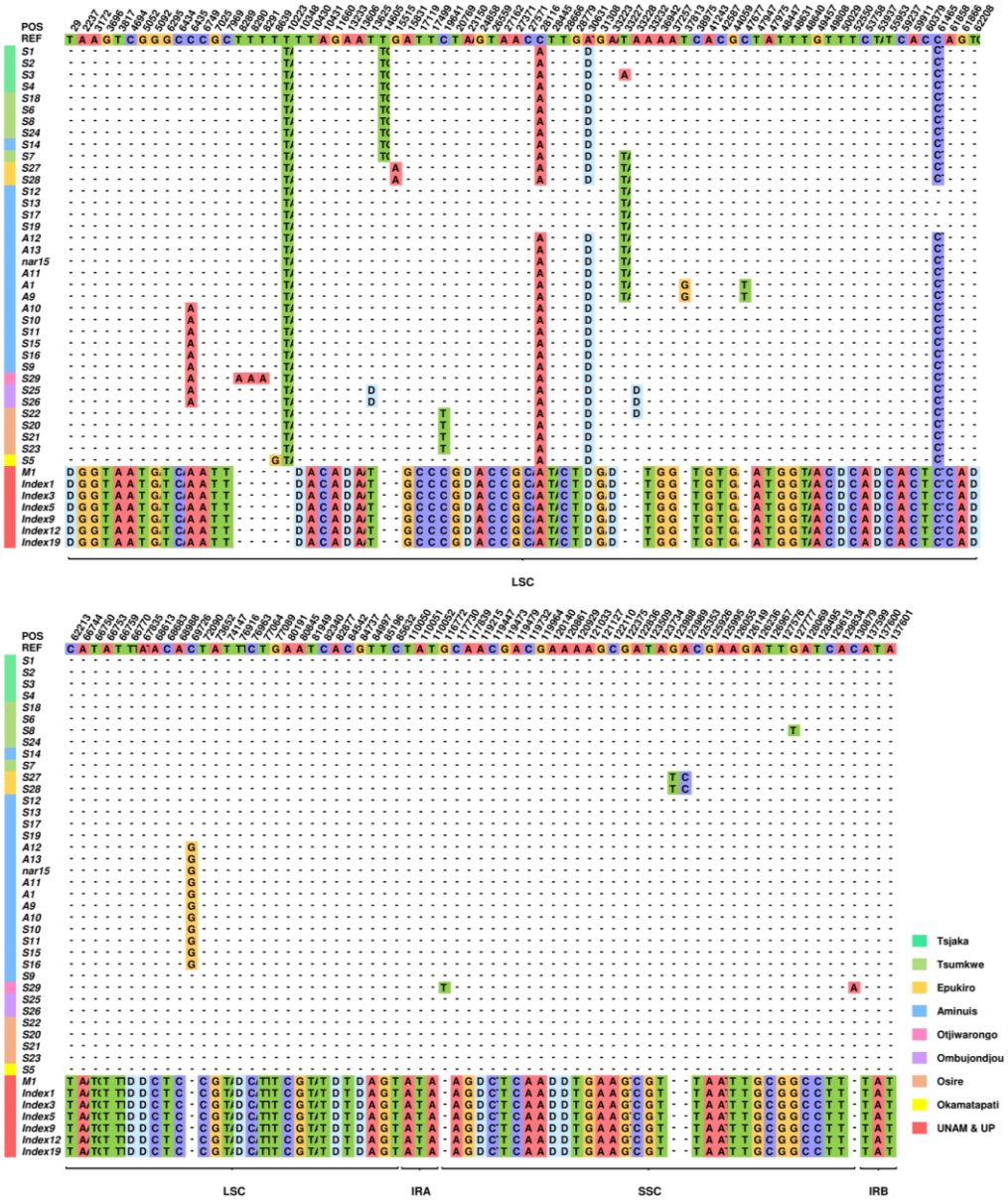

Figure 3. 3 Distribution map of all variations in 43 independent *T. esculentum* individuals. This figure includes only loci that differ in these plants, and the top row shows the sequence of the reference genome. Only bases different from the reference are exhibited, others are represented by dashes. To save space, only the first two bases of insertions are displayed. The letter D means deletions. The color bar to the left of the plant ID shows the source of the sample.

There were 105 SNPs in total, accounting for 71.4% of all variations. 31 SNPs but no indels were found in the coding sequence, 17 of which were silent mutations and 14 were nonsynonymous. Variations that altered the resulting amino acid sequence accounted for 45.2% of the total variation found in coding sequences and 9.5% of all variation in the cpDNA of the 84 individuals. The specific positions and effects of these SNPs are shown in Table 3.3. Genes including *rpoC2, rpoB, rpoA, ndhF, ndhD, ndhH, rps3* contain multiple SNPs in CDS. The introns of genes *ndhA, petB* and *trnK* have three or more variation sites (Table 3.4).

Table 3. 3 Variation positions found in the marama chloroplast coding sequence and their effect on the resulting amino acid sequence.

Gene Abbreviation	**Position**	**Reference**	**Mutation**	**Amino acid change**
matK	3172	A	G	synonymous
atpA	11660	G	A	synonymous
atpI	15515	G	A	P58S
rpoC2	17119	T	C	synonymous
rpoC2	17499	T	C	S1177G
rpoC2	19641	C	T	G463S
rpoC2	20769	T	G	N87H
rpoB	24858	G	A	synonymous
rpoB	26559	T	C	synonymous
rpoB	27162	A	C	I48M
psaB	41243	A	G	synonymous
psaA	42587	C	T	synonymous
ndhC	52550	T	C	T24A
atpB	55953	T	C	synonymous
accD	60379	C	T	synonymous
rps18	69726	C	G	T117S
rpoA	80191	A	C	N264K
rpoA	80845	A	G	synonymous
rps3	85196	T	G	synonymous
rps3	85632	C	T	synonymous
ndhF	117730	C	A	A495S

Gene Abbreviation	Position	Reference	Mutation	Amino acid change
ndhF	117839	A	G	synonymous
ccsA	122110	C	A	L294I
ndhD	122836	A	C	S251A
ndhD	123509	T	G	L26F
ndhD	123734	A	T	synonymous
ndhG	125353	C	T	synonymous
ndhI	126236	A	G	synonymous
ndhA	126967	T	C	synonymous
ndhH	129615	C	T	R230H
ndhH	129834	A	T	F157Y

For nonsynonymous substitutions, for example, P58S indicates a change of the 58th amino acid from proline to serine.

Table 3. 4 Number of variations found in introns of *T. esculentum* chloroplast genes.

Gene	Product	SNP CTS	Indel CTS	Indel Type
ndhA	NADH-plastoquinone oxidoreductase subunit 1	4		
rps16	Ribosomal protein S16	2	1	1 bp INS
atpF	ATP synthase subunit b		1	1 bp DEL
rpoC1	RNA polymerase beta subunit		1	5 bp DEL
clpP	ATP-dependent Clp protease proteolytic subunit	2		
petB	Cytochrome b6	1	3	5 bp DEL/1 bp INS/10 bp INS
rpl16	ribosomal protein L16	1	1	1 bp DEL
trnL	tRNA-Leu	1		
trnK	tRNA-Lys	3		
trnV	tRNA-Val	1		

CTS = Counts

3.4.3 Identification of Variable Regions

The average nucleotide diversity per site is 0.00032. However, a high π value (0.02070 in a 1200 bp window) was detected in the region where the *psbM-trnD* intergenic spacer is located. A 230 bp inversion was found between *psbM* and 4 closely located tRNA genes (Figure 3.4 and S3.1). This 230 bp was surrounded by a pair of 20 bp inverted repeats "ATTAGTAATTGAAATTAGTA" and "TACTAATTTCAATTACTAAT". It was speculated that a rare recombination occurred on the inverted repeats that flipped the sequence in the middle. Sequence alignment indicated that all Type 1 plants were identical to the reference genome in this region, but this 230 bp sequence was inverted in all Type 2 samples (Figure S3.3.1, S3.3.2, and S3.3.3).

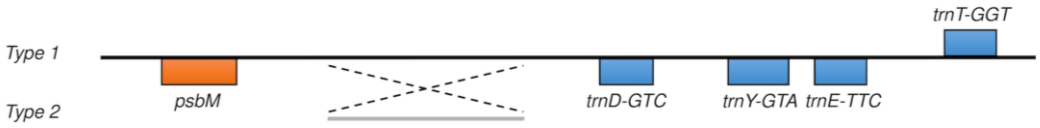

Figure 3. 4 Schematic diagram of the 230 bp inversion in the *psbM-trnD* intergenic spacer. A 230 bp sequence (30,663-31,480 in the marama reference cpDNA) is reversed in all Type 2 individuals. However, all Type 1 plants are consistent with the reference genome in this area. The surrounding region contains a set of closely located tRNA genes.

Nucleotide diversity within each 1200 bp window of the rest of the chloroplast genome ranged from 0 to 0.00116 (Figure 3.5). The most variable regions included *trnT-trnL*, *ndhG-ndhI* intergenic spacers, and the intron of *rps16* (all with a π value above

0.001). Although the marama cp genome was diverged into two clades, and multiple subgroups exist within them, overall, the cpDNA of marama remains highly conserved.

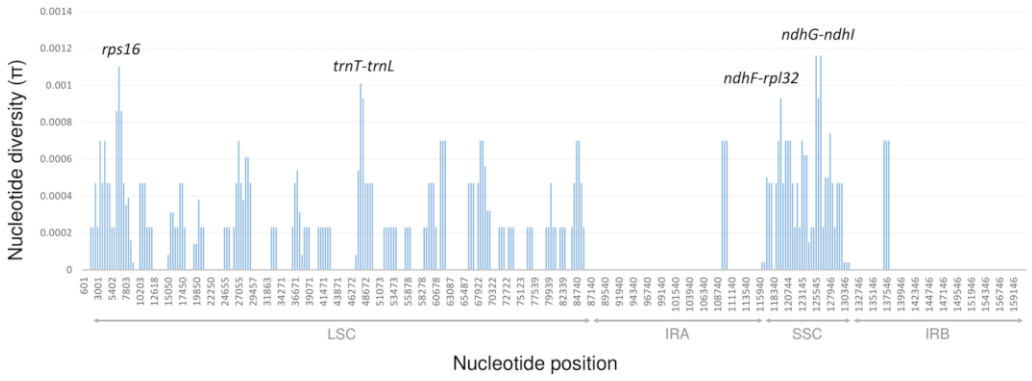

Figure 3. 5 Sliding window analysis of the chloroplast genomes of the 43 independent *T. esculentum* individuals (window length: 1200 bp, step size: 400 bp). The *x*-axis shows the midpoint position of each window. 30,949-31,218, where the 230 bp inversion is located was excluded from this analysis (Figure S3.1-3.2).

Pairwise comparisons of cpDNA of Type 1 and Type 2 marama samples and the related species *T. fassoglense* by mVISTA alignment revealed low levels of sequence divergences between the chloroplast genomes (Figure 3.6). In the coding sequence of *rps4* and the positions of the genes *trnG* and *trnM*, *T. fassoglense* showed clear differences from *T. esculentum*. Furthermore, at the 5' end of *accD*, the two Type 2 plants differed from the Type 1 plants but were more consistent with *T. fassoglense*. Type 2 plants and Type 1 plants also diverged between *psbM* and *trnD*, due to the presence of a 230 bp inversion. Another divergence was found at 48,640, where the Type 2 plants contained a 48 bp insertion that made the sequence "taattagaattaagtaattataaa" triplicated in this region and shown as a tandem repeat.

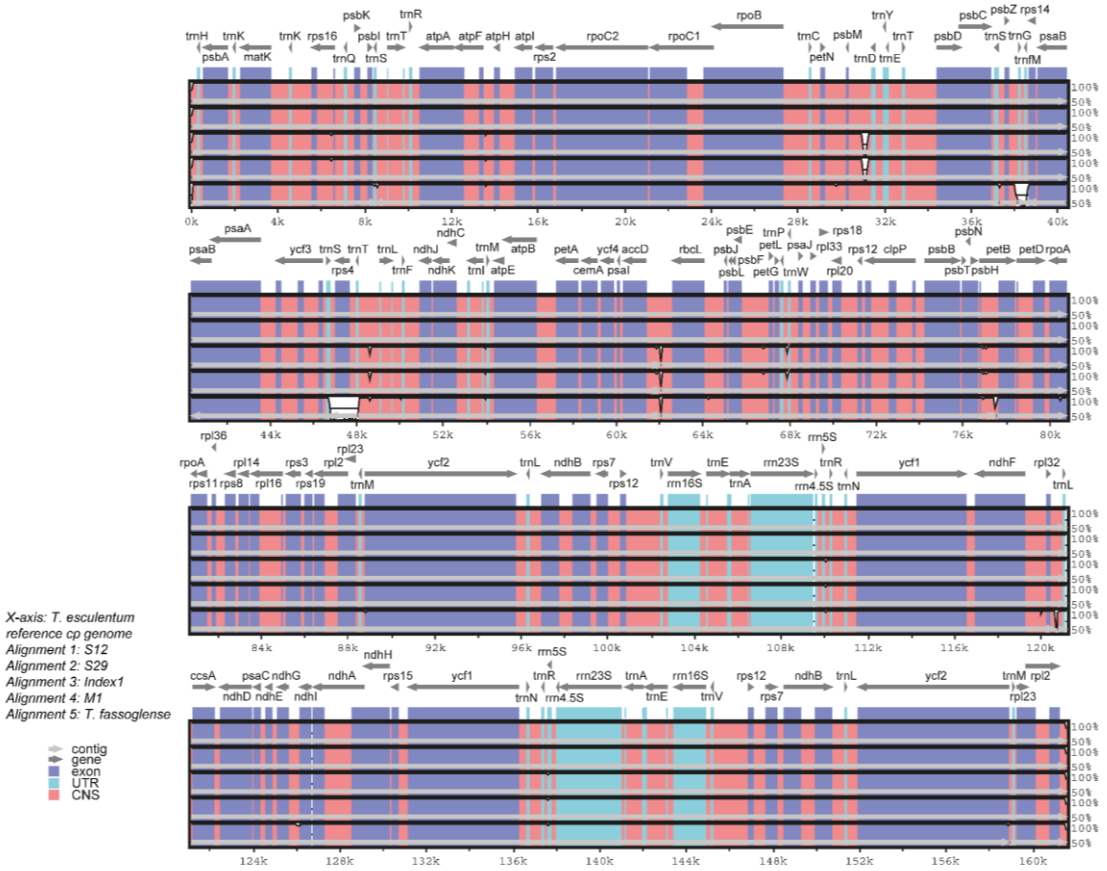

Figure 3. 6 Comparison of cpDNA from four *T. esculentum* individuals (Type 1: S12 and S29, and Type 2: Index1 and M1) and the related species *T. fassoglense* (NC_037767.1) by mVISTA Shuffle-LAGAN alignment. The plastome of *T. esculentum* (KX792933.1) was used as the reference. White blanks indicate regions of sequence divergence. The *x*-axis represents the position in the chloroplast DNA. The *y*-axis shows the similarity of sequence alignment, ranging from 50% to 100%. The location and transcription direction of the cp gene are labeled at the top of the block. Conserved exons, introns and noncoding regions are marked on the graph with different colors.

Blastn alignment of the cpDNAs of *T. esculentum* and *T. fassoglense* revealed a 38,314 bp long inversion in the LSC region (marama reference cpDNA: 8,427 to 46,740)

(Figure 3.7 and S3.4), which was also reported before by Wang et al., 2018. This inversion accounted for 44.5% of the length of the LSC region and 23.7% of the total length of the cpDNA, containing 17 protein-coding genes and 12 tRNA genes. Both boundaries are the 3' ends of the two *trnS*. The sequences of the two *trnS* genes share 78% (69/88) similarity. Therefore, it was speculated that the recombination that occurred at the two *trnS* genes reversed the intermediate sequence.

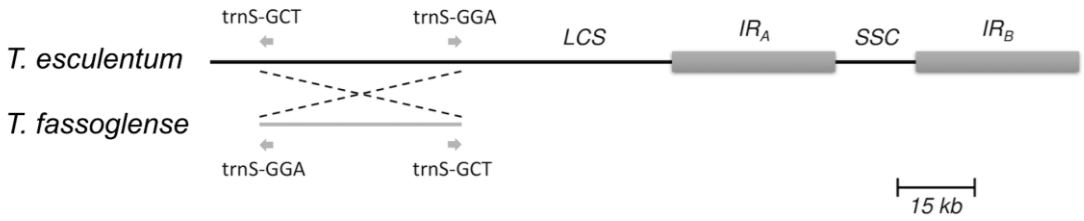

Figure 3. 7 Diagram showing a 38,314 bp inversion between the two *trnS* genes in the LSC region of *T. esculentum* and *T. fassoglense* cpDNA. This was also reported by Wang et al., 2018.

3.4.4 Phylogenetic Construction

The phylogenetic study indicated that the cpDNA of the marama samples were clearly divided into two clades (Figure 3.8). The cpDNAs of the Type 2 samples were basically the same and very conserved, but Type 1 samples could be further divided into multiple subgroups. The differences found between Type 1 plants were potentially related to the geographical origin of the samples, however, this needs to be proved by experiments with larger sample sizes. Diverged cpDNAs also appeared in plants from the

same geographic location. For example, at least four cpDNA subgroups were found in the marama plants from Aminuis. One individual of them, S14, was unexpectedly found to be highly similar to the samples collected in the geographically distant Tsumkwe region. More samples collected in these two regions are needed for comparison.

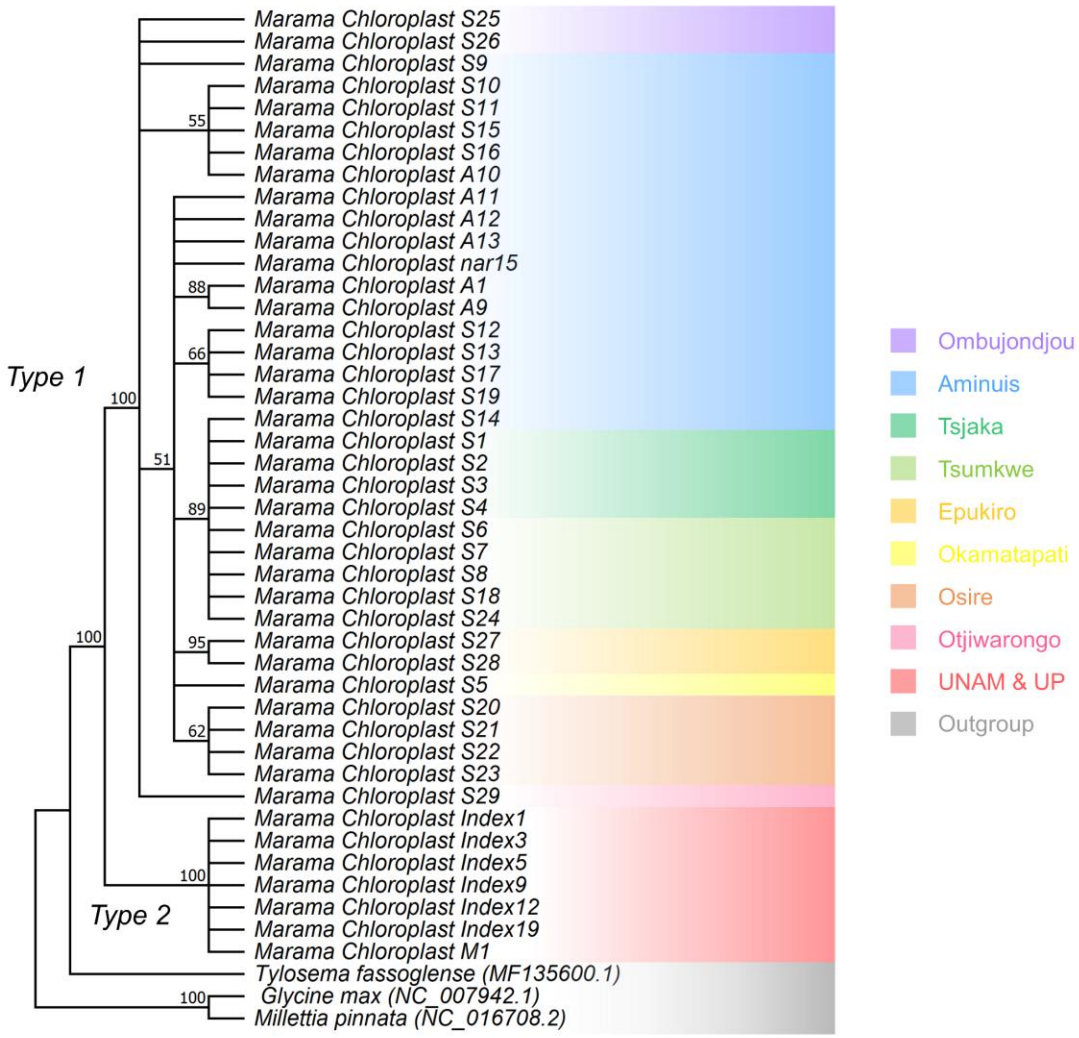

Figure 3. 8 Maximum Likelihood (ML) phylogenetic tree based on the Jukes-Cantor model and the Tamura-Nei model showing the relationship of the chloroplast genomes of 43 independent *T. esculentum* individuals and three other Fabaceae species. *T.*

fassoglense (MF135600.1), *Glycine max* (NC_007942.1), and *Millettia pinnata* (NC_016708.2) were used as outgroups. Bootstrap values from 1000 replicates were marked on the branches with 50% as cutoff. This was validated by the Maximum Parsimony (MP) method using the Subtree Pruning Re-grafting (SPR) algorithm with 1000 bootstrap replicates in Mega 11. Background colors indicate the geographic origin of the samples.

3.4.5 SSRs and Heteroplasmy Analysis

MISA analysis found a total of 79 SSRs, most of which were A or T mononucleotide repeats accounting for 79.7% (Figure 3.9). There were 5 AT or TA dinucleotide repeats, only one trinucleotide repeat ATT, and 10 compound SSRs. 60 SSRs were in noncoding intergenic regions, 11 were in introns, and 8 were in the coding sequences of genes including *rpoC2, rpoB, atpB, accD, rps18, clpP, ndhF,* and *rps19*. The analytical results for Type 1 samples and Type 2 samples were consistent.

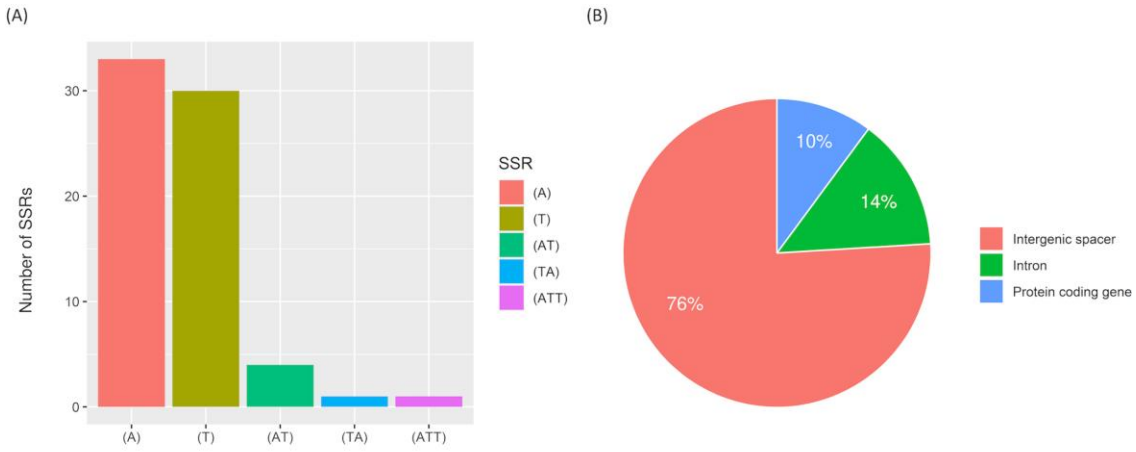

Figure 3. 9 Statistical analysis of SSRs in *T. esculentum* chloroplast genome by MISA. (A) Number of identified SSR motifs in different repeat types. (B) Location distribution of SSR repeats.

By examining allele frequencies in IGV, we found that three individuals contained both Type 1 and Type 2 cpDNA (Figure 3.10). Index8, as a Type 1 individual, also contained the Type 2 alleles among the 122 inter-type difference loci, the frequency of which remained consistent around 2%, and vice versa for S27. The minor alleles are easily ignored because their frequencies are well below the default cutoff for Calling SNPs with Samtools, and they are also difficult to distinguish from sequencing and alignment errors. With one exception, as a Type 1 individual from Aminuis, A11 also contained the Type 2 alleles in its cp genome with a frequency of approximately 11% among these 122 loci. The read depth of these minor alleles is about 450x, much higher than the nuclear DNA coverage (35-50x) indicating the existence of heteroplamsy in the marama chloroplast genome. In the cp genomes of other individuals, heterogeneity only existed at some loci but not others. If the cutoff value is further reduced from 2% to

lower, it is expected to see more heteroplasmy. It is speculated that paternal leakage resulted in the presence of both types of chloroplast DNAs in the same individual. The frequency of the minor alleles is too low to be detected effectively, and it is uncertain whether there is selection in favor of certain gene alleles or DNA segments over others.

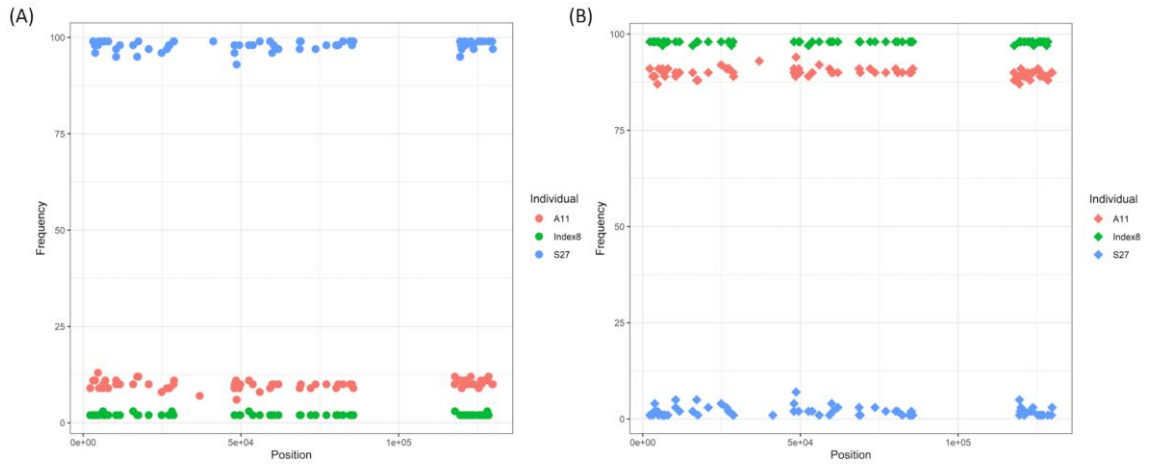

Figure 3. 10 Diagram of allele frequencies of the 105 SNP loci different in Type 1 and Type 2 plants. (A) Frequency of the Type 2 alleles at these 105 loci. (B) Frequency of the Type 1 alleles at these loci. Only the frequencies of SNPs were recorded and compared here, as accurate indel frequencies were hard to obtain.

3.5 Discussion

As an orphan species, marama can not only survive in extreme environments of Africa, but it also provides edible and nutritious seeds, making the domestication and genetic research of the bean very valuable. Molecular studies of the marama chloroplast genome can help us better understand this species from an evolutionary perspective, as

well as identify existing polymorphism that could be associated with phenotypes of interest, which will aid marama domestication and breeding.

By conducting DNA sequencing on a large number of individuals, this study aimed to compare the chloroplast genomes of marama plants at the individual and population level, and to investigate intraspecific variations that exist among them. A total of 84 samples (43 independent individuals) were collected from different geographic locations in Namibia and Pretoria. The cp genomes of these plants were compared to identify polymorphisms, including SNPs, indels and genomic structural variations. Highly variable cp genes and genomic regions were discussed. Phylogenetic analysis indicated the existence of two distinct germplasms in marama, with chloroplast genomes distinguished from each other at 122 loci. There was also a structural difference, a 230 bp inversion, found between the two types of chloroplast genomes, which provides valuable information for studying germplasm evolution (Palmer et al., 2000b). Our study also confirmed a previous finding by Wang et al., in 2018 that a large inversion of 38,314 bp exists between the two *trnS* genes in the LSC region of *T. esculentum* and *T. fassoglense* plastomes. Large inversions of cpDNA within the same genus are uncommon, although they have been reported previously in genera such as *Artemisia* and *Astragalus* (Johansson, 1999; Charboneau et al., 2021). This large inversion is not fully reflected in our unscaled tree (Figure 3.8). In the scaled tree, long branch lengths were expected to be seen between the two species, however, all marama individuals of the same type would be clustered together, thus obscuring intra-type differences. One type of marama cpDNA was found to be relatively conserved with very low diversity. The other type appeared to have more variability within the major genome present, with a total of 25 intra-type

variations and these variations seemed to be distributed within specific geographic regions.

Of all the variants detected, 31 SNPs were located in the coding sequence of marama cpDNA, 14 of which were nonsynonymous, altering the synthesized protein sequence. Both *rpo* genes and *ndh* genes appeared to be less conserved. The coding sequences of the genes *rpoC2*, *rpoB*, and *ndhD*, and the introns of the genes *ndhA*, *petB*, *trnK*, and *rps16*, and the intergenic spacers *trnT-trnL*, *ndhG-ndhI* were found to be more variable compared to the other regions of the marama plastome. Whether these genomic variations and the formation of distinct chloroplast genomes have anything to do with the adaptation of marama to harsh environments will be an interesting question to explore. Understanding how these polymorphisms are associated with any particular phenotype, like plant stress responses, still requires further large-sample statistical analysis to link phenotypic data to the different genotypes (Members of the Complex Trait Consortium, 2003). Therefore, SSRs and long sequence repeats were also analyzed, which could be used to establish molecular markers to aid in rapid genotyping of large numbers of samples.

In addition, it was interesting to find heteroplasmy in the cpDNA of some marama individuals. Both types of cpDNA co-existed in those individuals, but one predominated and the other had a very low frequency between 0% and 2%. However, there was one exception. In one individual, the minor cpDNA frequency reached 11%. Previous studies speculated that the occasional leakage of paternal cytoplasmic information led to heteroplasmy (McCauley et al., 2005), but the effect of the long-term accumulation of spontaneous *de novo* mutations in cpDNA could not be ruled out. Why plant cells can

possess both organelle genomes at the same time and maintain the ratio of the two to a certain extent, and whether this ratio changes during cell development as described by the animal mitochondrial bottleneck hypothesis remains unknown (Cao et al., 2007). Understanding the genetic mechanisms behind this could help us better understand the function and inheritance of the organelles.

In general, this in-depth study of the chloroplast genome of marama not only contributes to marama variety identification, but it also helps us better understand the origin and evolution of the bean. All these are conducive to the domestication and breeding of marama in the future.

3.6 Supplementary materials

Table S3.1 Information on the source and sequencing details of the 84 samples.

Sample	Plant source	Raw reads	Raw data	CP (%)	CP (bp)
M17	UNAM Farm Progenies	96750044	14512506600	8.37	161562
S_35	UNAM Farm Progenies	86626894	12994034100	19.2	161562
S_19	UNAM Farm Progenies	55375806	8306370900	14.81	161562
S_4	UNAM Farm Progenies	73435564	11015334600	12.32	161562
S_13	UNAM Farm Progenies	217730028	32659504200	14.62	161562
S_27*	UNAM Farm Progenies	107264134	16089620100	0.29	161562
S_20	UNAM Farm Progenies	55371622	8305743300	9.15	161562
S_30	UNAM Farm Progenies	106501502	15975225300	17.64	161562
S_33	UNAM Farm Progenies	130811500	19621725000	17.54	161562
M7	UNAM Farm Progenies	172383630	25857544500	17.78	161562
M8	UNAM Farm Progenies	123934796	18590219400	17.54	161562
M1 #	UNAM Farm Progenies	170816890	25622533500	12.84	161562
M2	UNAM Farm Progenies	194335620	29150343000		161562
M11	UNAM Farm Progenies	125005082	18750762300	14.32	161562
M12	UNAM Farm Progenies	128360580	19254087000		161562
M15	UNAM Farm Progenies	134439570	20165935500	15.99	161562
M16	UNAM Farm Progenies	132448260	19867239000	18.9	161562
M23	UNAM Farm Progenies	134475978	20171396700	15.07	161562
M22	UNAM Farm Progenies	111447592	16717138800	16.11	161562

Sample	Plant source	Raw reads	Raw data	CP (%)	CP (bp)
M24	UNAM Farm Progenies	127392478	19108871700	19.48	161562
M26	UNAM Farm Progenies	225925198	33888779700	1.18	161562
M28	UNAM Farm Progenies	193804308	29070646200	11.85	161562
M25	UNAM Farm Progenies	145937916	21890687400	13.13	161562
N29	UNAM Farm Progenies	147710832	22156624800	17.32	161562
M31	UNAM Farm Progenies	177899436	26684915400	10.75	161562
M34	UNAM Farm Progenies	124221640	18633246000	6	161562
M36	UNAM Farm Progenies	125524008	18828601200	17.31	161562
M37	UNAM Farm Progenies	185151336	27772700400	15.11	161562
M38	UNAM Farm Progenies	168052326	25207848900	13.35	161562
M40	Namibia Unknown	135702768	20355415200	18.56	161562
Index1 #	UP Farm	41499124	4149912400	7.25	161562
Index10	UP Farm Progenies	36382172	3638217200	12.1	161562
Index11	UP Farm Progenies	34631932	3463193200	11.62	161562
Index12 #	UP Farm	39339004	3933900400	13.83	161562
Index19 #	UP Farm	35994474	3599447400	17.91	161562
Index3 #	UP Farm	35991990	3599199000	16.5	161562
Index5 #	UP Farm	34349602	3434960200	38.46	161562
Index8	Namibia Unknown	42747718	4274771800	6.73	161562
Index9 #	UP Farm	34312880	3431288000	13	161562
R1R2	Unknown	358941018	35894101800	6.25	
A1 #	Aminuis Progenies	93324222	13998633300	8.31	161554
A2	Aminuis Progenies	93445492	14016823800	6.2	161554
A3	Aminuis Progenies	84081976	12612296400	11.28	161554
A4	Aminuis Progenies	60965538	9144830700	13.29	161554
A5	Aminuis Progenies	91044746	13656711900	13.57	161554
A6	Aminuis Progenies	89785050	13467757500	17.46	161554
A7	Aminuis Progenies	86000118	12900017700	14.26	161554
A8	Aminuis Progenies	57087632	8563144800	12.53	161554
A9 #	Aminuis	62849376	9427406400	13.37	161554
A10 #	Aminuis	75633554	11345033100	19.93	161553
A11 #	Aminuis	92851730	13927759500	4.74	161554
A12 #	Aminuis	84374620	12656193000	10.76	161554
A13 #	Aminuis	83450100	12517515000	4.43	161554
nar15 #	Aminuis	81618952	12242842800	14.2	161554
nar16	UP Farm Progenies	96595344	14489301600	10.25	161562
S1* #	Tsjaka	10358444	1035844400	0.29	161579
S2 #	Tsjaka	20343934	2034393400	3.82	161579
S3 #	Tsjaka	33100100	3310010000	5.37	161579
S4 #	Tsjaka	25428338	2542833800	4.36	161579

Sample	Plant source	Raw reads	Raw data	CP (%)	CP (bp)
S5 #	Okamatapati	27183834	2718383400	6.6	161553
S6 #	Tsumkwe	24185808	2418580800	7.95	161579
S7* #	Tsumkwe	22075112	2207511200	1.79	161580
S8* #	Tsumkwe	16337544	1633754400	3.23	161579
S9 #	Aminuis	25274640	2527464000	5.24	161553
S10 #	Aminuis	28329944	2832994400	7.21	161553
S11 #	Aminuis	26741342	2674134200	7.52	161553
S12 #	Aminuis	29520092	2952009200	5.73	161552
S13 #	Aminuis	NA	NA	NA	161552
S14 #	Aminuis	22928100	2292810000	3.81	161579
S15 #	Aminuis	22856344	2285634400	7	161553
S16 #	Aminuis	25333752	2533375200	6.33	161553
S17 #	Aminuis	23418786	2341878600	7.07	161552
S18 #	Tsumkwe	25764102	2576410200	4.14	161579
S19 #	Aminuis	24407080	2440708000	6.92	161552
S20 #	Osire	25692398	2569239800	5.48	161553
S21 #	Osire	25940358	2594035800	8.04	161552
S22 #	Osire	32107310	3210731000	8.01	161552
S23 #	Osire	35844166	3584416600	7.09	161553
S24 #	Tsumkwe	31973008	3197300800	6.58	161579
S25 #	Ombujondjou	24401422	2440142200	6.91	161551
S26 #	Ombujondjou	32758062	3275806200	5.89	161551
S27 #	Epukiro	28122544	2812254400	7.98	161555
S28 #	Epukiro	31273940	3127394000	7.95	161555
S29 #	Otjiwarongo	25274640	2527464000	5.24	161553

* samples with low chloroplast coverage, # 43 independent samples for sequence diversity and phylogenetic analysis, CP (%): plastome alignment rate, CP (bp): plastome length.

Table S3.2 Correction of *T. esculentum* chloroplast gene positons.

Gene Abbrv.	Former Position	Corrected Position	Region	Function
matK	3812..2215	3612..2269	LSC	Other genes
rps16	5803..5567	6602..6561, 5794..5567	LSC	Self replication
psbK	7559..7753	7568..7753	LSC	Photosynthesis
psbI	8123..8275	8165..8275	LSC	Photosynthesis
atpA	12049..10517	12049..10544	LSC	Photosynthesis

Gene Abbrv.	Former Position	Corrected Position	Region	Function
atpF	13437..12117	13437..13294, 12521..12114	LSC	Photosynthesis
rpoC2	19131..16888	21027..16888	LSC	Self replication
rpoC1	24065..21205	24065..23634, 22818..21214	LSC	Self replication
petN	29088..29186	29097..29186	LSC	Photosynthesis
psbC	35402..36862	35441..36862	LSC	Photosynthesis
psaA	43468..41324	43558..41306	LSC	Photosynthesis
ycf3	46381..44327	46396..46256, 45536..45309, 44479..44327	LSC	Unknown
ndhK	52215..51583 52377..51592	52215..51556	LSC	Photosynthesis
accD	61305..60364	61401..60349	LSC	Other genes
rbcL	64051..62603	64030..62603	LSC	Photosynthesis
psaJ	68411..68545	68411..68539	LSC	Photosynthesis
rps18	69377..69694	69377..69754	LSC	Self replication
rps12	71291..71163 100874..100566	71291..71178, 100808..100566	LSC	Self replication
clpP	72885..71530	73748..73680, 72882..72589, 71748..71482	LSC	Other genes
psbT	75988..76104	75997..76104	LSC	Photosynthesis
psbN	76378..76148	76302..76171	LSC	Photosynthesis
psbH	76398..76628	76407..76628	LSC	Photosynthesis
petB	77650..78336	76758..76763, 77695..78336	LSC	Photosynthesis
petD	79220..79744	78528..78536, 79271..79744	LSC	Photosynthesis
rpoA	80982..79999	80982..79981	LSC	Self replication
rps11	81442..81053	81484..81053	LSC	Self replication
rpl16	83894..83484	84959..84951, 83882..83484	LSC	Self replication
rps3	85802..85146	85802..85137	LSC	Self replication
rpl2	87955..86460	87955..87566, 86894..86460	IRA	Self replication
rpl23	88344..88066	88344..87901	IRA	Self replication
ycf2	90160..95700	88822..95700	IRA	Unknown
ycf15	95879..95968 101763..101572	95968..95879 (pseudo)	IRA	Unknown
ndhB	99163..96952	99163..98387, 97707..96952	IRA	Photosynthesis
rrn23S	106037..109461	106651..109462	IRA	Self replication
rpl32	120263..120415	120263..120430	SSC	Self replication

Gene Abbrv.	Former Position	Corrected Position	Region	Function
ccsA	121231..122184	121231..122190	SSC	Other genes
ndhD	123955..122441	123937..122441	SSC	Photosynthesis
ndhE	124879..124574	124876..124574	SSC	Photosynthesis
ndhI	126664..126161	126664..126167	SSC	Photosynthesis
ndhA	129120..126745	129120..128584, 127266..126745	SSC	Photosynthesis
rrn23S	141614..138190	141000..138189	IRB	Self replication
rps12	146777..147805	71291..71178, 146843..147085	IRB	Self replication
ndhB	148488..150699	148488..149264, 149944..150699	IRB	Photosynthesis
ycf15	151772..151683	151772..151683 (pseudo)	IRB	Unknown
ycf2	157491..151951	158829..151951	IRB	Unknown
rpl23	159307..159585	159307..159750	IRB	Self replication
rpl2	159696..161191	159696..160085, 160757..161191	IRB	Self replication
rps19	161360..161537	161360..161536 (pseudo)	IRB	Self replication

Table S3.3 Newly identified tRNA genes in *T. esculentum* chloroplast genome by CPGAVAS2.

Gene Name	Position
trnK-UUU	4583..4547, 1970..1936
trnT-CGU	9043..9077, 9784..9827
trnL-UAA	49078..49112, 49626..49675
trnI-AAU	53845..53815, 53222..53162
trnE-UUC	104535..104566, 105520..105559
trnA-UGC	105624..105660, 106463..106498
trnA-UGC	142027..141991, 141188..141153
trnE-UUC	143116..143085, 142131..142092

Table S3.4 Population haplotype analysis of the 43 independent marama individuals.

Population	n	Hn	Hd	π
Pretoria	7	1	0	0
Epukiro	2	1	0	0
Aminuis	17	6	0.83088	0.00001
Osire	4	1	0	0

Population	n	Hn	Hd	π
Tsumkwe	5	2	0.4	0
Tsjaka	4	1	0	0
Ombujondjou	2	1	0	0
Okamatapati	1	1	0	0
Otjiwarongo	1	1	0	0

n = number of sequences, Hn = number of haplotypes, Hd = haplotype diversity, π = nucleotide diversity

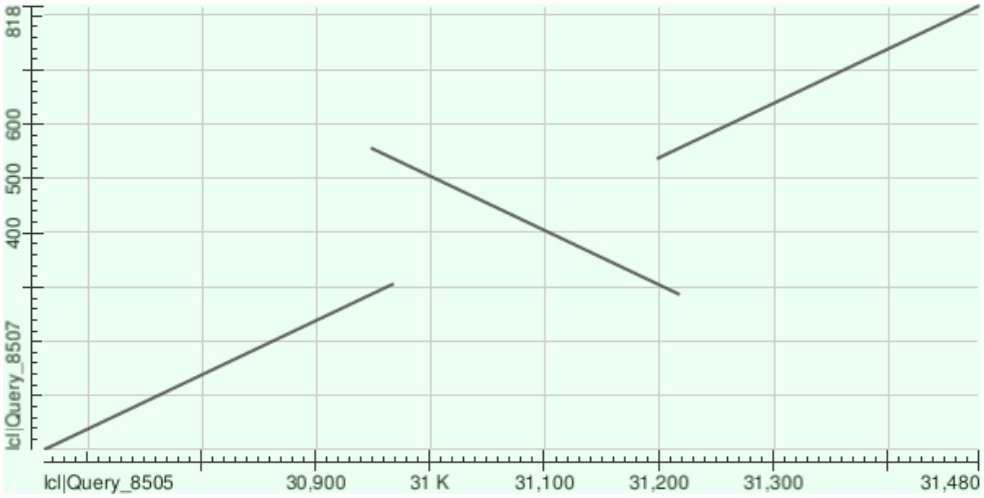

Figure S3.1 Alignment of cpDNA sequences of Type 1 and Type 2 marama samples revealed the existence of a 230 bp inversion. A portion of the marama reference cpDNA was blasted to the corresponding region of the Type 2 sample and displayed as the dot plot above. The inversion is located at the intergenic region (30,949-31,218) between *psbM* and four closely located tRNA genes, *trnD-GTC*, *trnY-GTA*, *trnE-TTC*, and *trnT-GGT* in the chloroplast genomes of marama.

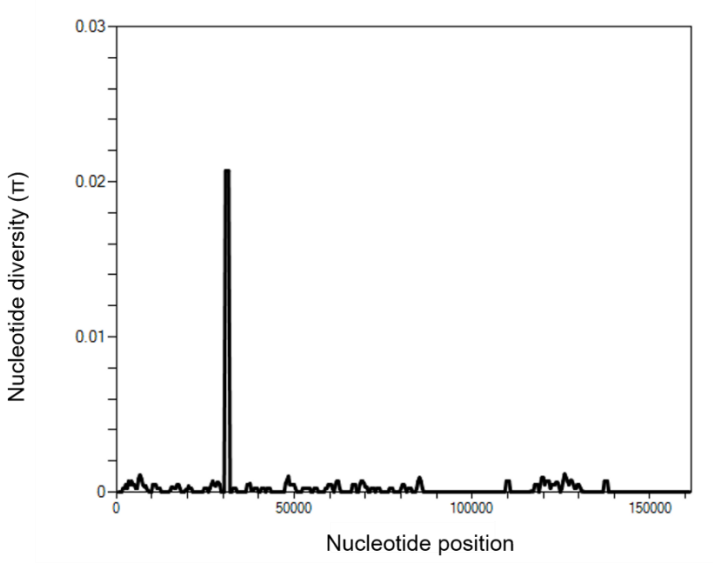

Figure S3.2 Sliding window analysis of the chloroplast genomes of the 43 independent *T. esculentum* individuals (window length: 1200 bp, step size: 400 bp) The *x*-axis shows the midpoint position of each window.

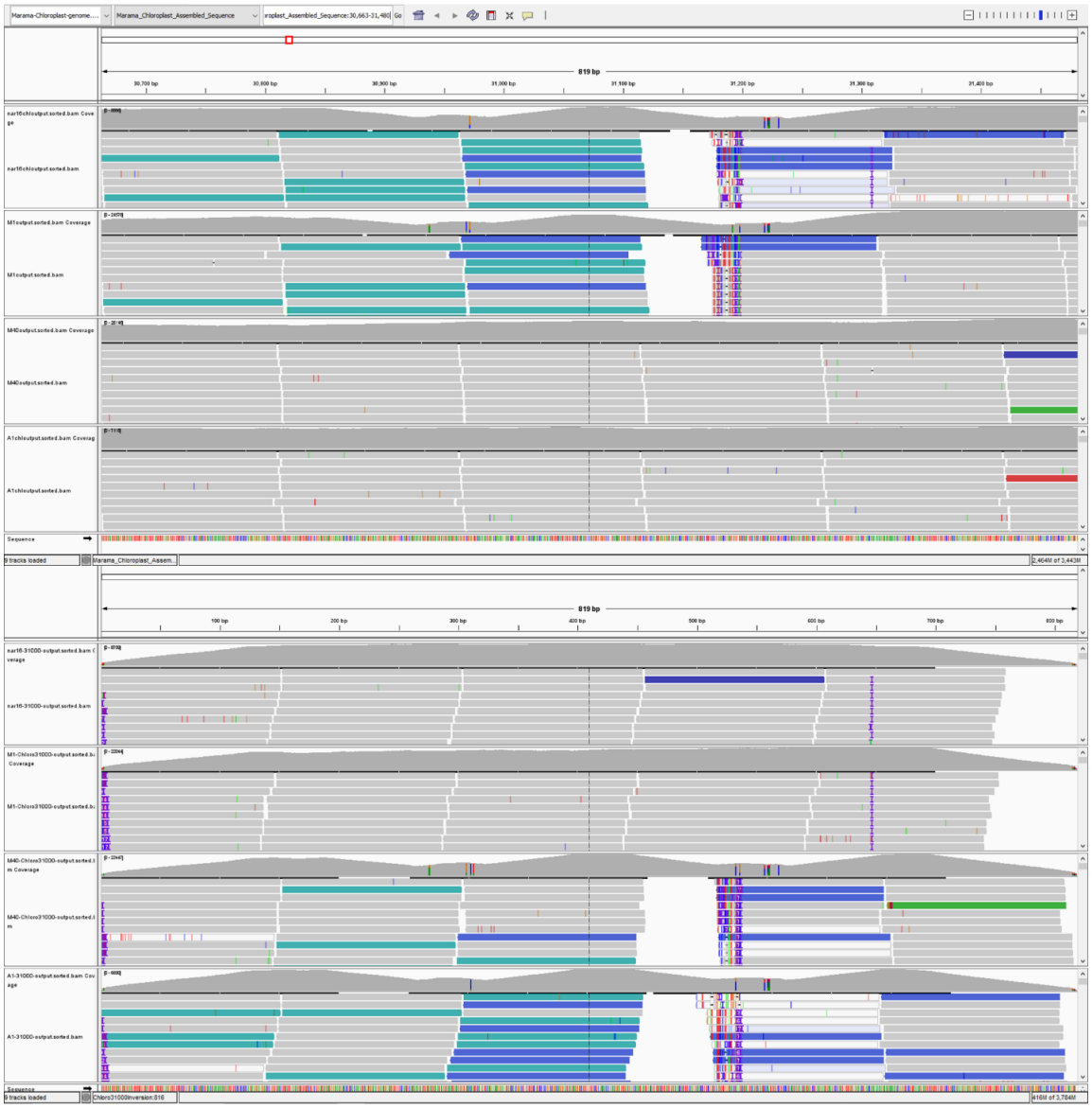

Figure S3.3.1 Visualization of sequence alignment in IGV confirmed a 230 bp inversion between 30,949 and 31,218. Top: The Illumina reads from 4 individuals (two Type 2 plants: nar16 and M1, and two Type 1 plants: M40 and A1) were aligned with the reference marama cp genome and distinct gaps could be seen in both Type 2 plants. Bottom: After inverting the 230 bp between 30,949 and 31,218 in the reference cp genome and redid the alignment, the gaps were missing in the two Type 2 samples but

appeared in the two Type 1 samples. This was consistent with the rest of Type 1 and Type 2 individuals.

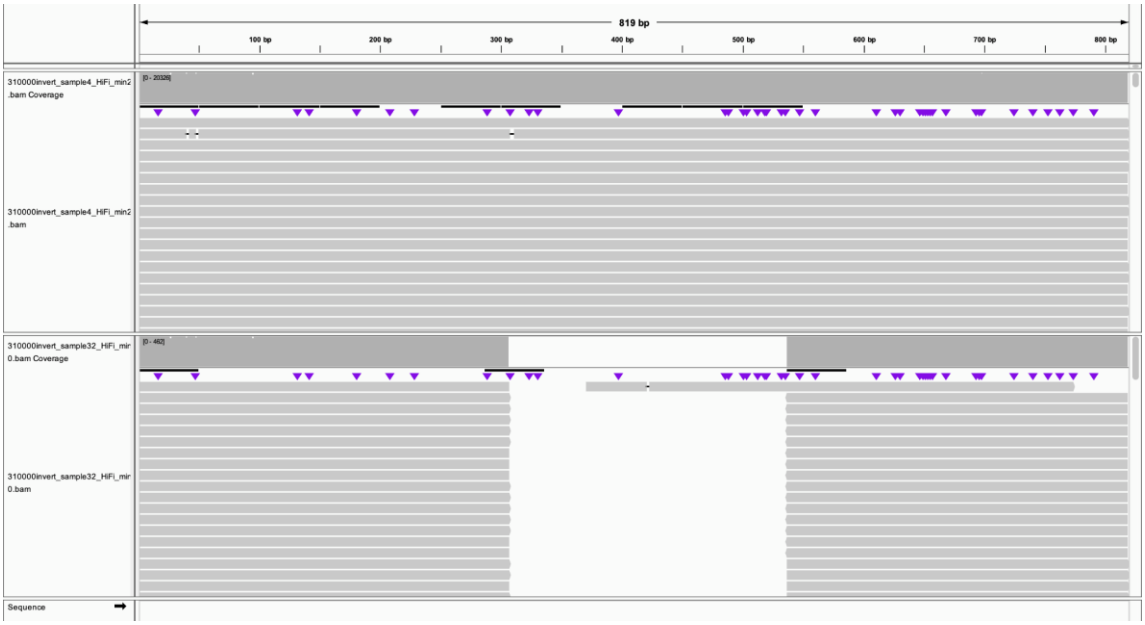

Figure S3.3.2 IGV visualization of SMRT pbmm2 alignments of PacBio HiFi reads to the marama reference chloroplast genome 30,663 bp to 31,480 bp, with the 230 bp in the middle inverted. Sample 4 and sample 32, which have type 2 and type 1 cp genomes, respectively, are two plants growing in the greenhouse of the Department of Biology, Case Western Reserve University. The minimum alignment length was set to 250 bp to filter out the interference of homologous sequences.

Figure S3.3.3 IGV visualization of SMRT pbmm2 alignments of PacBio HiFi reads to the marama reference chloroplast genome 30,663 bp to 31,480 bp. Sample 4 and sample 32 have type 2 and type 1 cp genomes respectively and they are two plants growing in the greenhouse of the Department of Biology at Case Western Reserve University. The minimum alignment length was set to 250 bp filter out the interference of homologous sequences.

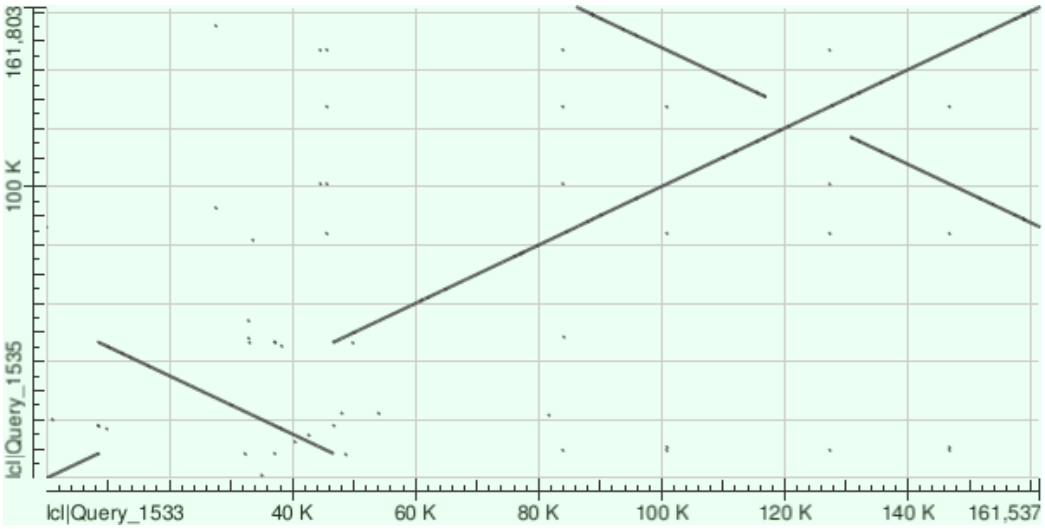

Figure S3.4 Alignment of the cpDNAs of *T. esculentum* and *T. fassoglense* revealed a 38314 bp long inversion in the LSC region. The *T. esculentum* reference cpDNA was aligned with the *T. fassoglense* cp genome (NC_037767.1) available in NCBI GeneBank using Blastn and shown as a dot plot. The *x*-axis shows the coordinates of the reference cpDNA of *T. esculentum*. The *y*-axis represents the cpDNA of *T. fassoglense*.

Chapter 4. Population study of *Tylosema esculentum* mtDNA diversity indicates the existence of two distinct mitogenome structures

4.1 Introduction

Mitochondria in eukaryotic cells are generally considered to have originated from the endosymbiosis of alpha-proteobacteria, although a number of changes have occurred since then, including the loss of many genes and their transfer to the nuclear genome (Andersson et al., 2002). Among them, the mitochondrial genomes of animals and plants have been found to vary greatly. Animal mitogenomes are usually small, only about 16 kb in size, and contain 37 genes (Boore, 1999). Plant mitogenomes are commonly larger with expanded intergenic non-coding regions that result from DNA transfer from other cellular compartments or even from different organisms (Kazuyoshi and Kubo, 2010). In land plants, they range in size from 66 kb in *Viscum scurruloideum* to 11.3 Mb in *Silene conica* (Sloan et al., 2012b; Skippington et al., 2015).

Plant mitochondrial genes are very conserved, are considered to play important roles in ATP synthesis, and are also related to plant fertility and environmental adaptation (Hanson, 1990; Budar and Roux, 2011; Heng et al., 2014). The base substitution rate of mitochondrial genes is lower than that of chloroplast genes and far lower than that of nuclear genes, only about one-tenth (Wolfe et al., 1987; Drouin et al., 2008). Plant mitogenomes evolve even up to 100 times slower than animal mitogenomes (Palmer and Herbon; 1988). However, the structure of plant mitochondrial genome is very dynamic, with a large number of sequence rearrangements, and repeat-mediated homologous

recombination plays an important role in its structural variations (Gualberto and Newton, 2017; Cole et al., 2018).

In the mitochondrial genomes of many angiosperms, repetitive sequences account for 5-10% of the total genome size, and in a few plants such as *S. conica* and *Nymphaea colorata*, the proportion can exceed 40% with repetitive fragments up to 80 kb in length (Sloan et al., 2012; Dong et al., 2018). Recombination mediated by short or intermediate length repeats is considered to be less frequent, but recombination on long repeats (>1 kb) is thought to occur more frequently and usually generates equimolar recombined molecules in the plant mitogenomes (Arrieta-Montiel, 2009; Guo et al., 2016; Li and Cullis, 2021).

The third-generation sequencing technology, such as PacBio, provides long reads with an average length of 10-25 kb, spanning the long repeat in the plant mitogenome. This makes the study of structural variations caused by the long repeat mediated recombination possible (Kozik et al., 2019; Hon et al., 2020). High sequencing coverage is also important, not only to make the genome assembly more reliable, but also to tell us other information like the proportion of different chromosomes structures. It is also indispensable for the accurate assessment of nucleotide polymorphisms (Telenti et al., 2016). The latest PacBio HiFi sequencing with an extremely high accuracy of 99.9%, which needs little correction by the data generated from other sequencing platforms, further promotes the genome assembly (Miga et al., 2020; Naish et al., 2021).

Although plant mitochondrial genomes are often reported as one master circular chromosome, in reality, they often exist as multipartite structures. This includes a combination of linear molecules, branched molecules, and subgenomic circular

molecules (Oldenburg and Bendich, 1996; Manchekar et al., 2006). For example, the mitogenome assembly of *Solanum tuberosum* was found to contain at least three autonomous chromosomes, including two small circular molecules and a long linear chromosome of 312,491 bp in length (Varré et al., 2019).

In the previously published study on the mitogenome of *T. esculentum*, two different equimolar structures were found to coexist in the same individual. There is two autonomous rings with a total length of 399,572 bp and five smaller circular molecules (Li and Cullis, 2021). These two structures are believed to be interchangeable by recombination on three pairs of long direct repeats (3-5 kb in length). As described in the study of *Brassica campestris* mitogenome, recombination on a pair of 2 kb repeats was postulated to split the 218 kb master chromosome into two subgenomic circular molecules of 135 kb and 83 kb in length (Palmer and Shields, 1984).

The previous comparative analysis of 84 *T. esculentum* chloroplast genomes has found two distinct germplasms. The two types of chloroplast genomes are different from each other at 122 loci and at a 230 bp inversion (Li and Cullis, 2023). Among many of these loci, heteroplasmy, the existence of two or more different alleles, could be seen, albeit one generally at a low frequency below 2%. The occasional paternal leakage is considered to cause this phenomenon (Kvist et al., 2003; Luo et al., 2018). The reason for its stable inheritance is thought to be related to the developmental genetic bottleneck, but the specific mechanism is still unclear (Floros et al., 2018).

Research on heteroplasmy should avoid interference from homologous sequences of mitochondrial DNA in other organelles. In fact, the horizontal gene transfer of DNA from chloroplast genome to mitogenome and nuclear genome, or between mitogenome

and nuclear genome are very common (Woloszynska et al., 2004; Keeling and Palmer, 2008). The transfer of DNA from mitogenome to chloroplast genome is very rare, but it has been reported in a few studies (Goremykin et al., 2008; Straub et al., 2013). Many chloroplast genes have been found to become pseudogenes and lost function after being inserted into the mitogenome, but the reason behind it is still unclear (Li et al., 2022).

Although the mechanisms underlying the effects of cytoplasmic activities on agronomic traits are not well understood, previous studies on potato cytoplasmic diversity have found that cytoplasmic types are directly related to traits such as tuber yield, tuber starch content, disease resistance, and cytoplasmic male sterility (Sanetomo and Gebhardt, 2015; Anisimova and Gavrilenko, 2017; Smyda-Dajmund et al., 2020). Furthermore, certain mtDNA types and certain chloroplast DNA types were found to be linked (Lössl et al., 1999). A comparative genomics analysis based on the chloroplast genomes of 3,018 modern domesticated rice cultivars found that their genotypes fall into two distinct clades, suggesting that the domestication of these cultivars may have followed two distinct evolutionary paths (Moner et al., 2020). These studies have the potential to reveal important selections occurring in organelle genomes, help us better understand plant adaptation to different environments, and provide a basis for crop breeding to increase yield in corresponding environments.

The chloroplast genome is widely used in research on plant evolution, but the comparative analysis based on plant mitochondrial genome is not so extensive, and there is even less research on the mitochondrial diversity within the same species (Nikiforova et al., 2013; Carbonell-Caballero et al., 2015). In this study, the diversity of marama mitochondrial genome was analyzed by mapping the WGS reads of 84 *T. esculentum*

individuals to the previously assembled reference mitogenome aiming to: (1) discover possible mitogenome structural diversity and the impact of structural variations on gene sequence and copy numbers; (2) compare the differential loci in mitogenomes of 43 independent marama individuals collected from different geographical locations in Namibia and South Africa to explore the divergences that have occurred and possible decisive environmental factors behind them; (3) look at heteroplasmy, the co-existence of multiple types of mitogenomes within the same individuals, and compare allele frequencies in related individuals to better understand the underlying cytoplasmic inheritance; (4) track polymorphisms accumulated in the chloroplast DNA insertion and interpret the fate of the inserted gene residues. In addition, the conserved protein-coding genes from the mitogenomes of *T. esculentum* and other Fabaceae species were compared to explore the evolutionary relationship between them.

4.2 Materials and Methods

4.2.1 Plant materials and DNA extraction

Samples 4 and 32 were two individuals in the greenhouse of Case Western Reserve University grown from seeds collected in Namibia at undocumented locations. They were identified by PCR amplification to have type 2 and type 1 germplasms, respectively. 1 g fresh young leaves were collected from the two plants respectively and ground thoroughly with a pestle in a mortar containing liquid nitrogen. DNA was then extracted using a Quick-DNA HMW MagBead kit (Zymo Research) following the protocol. Then, double-stranded DNA was quantified by the Invitrogen™ Qubit™ 3.0 Fluorometer after mixing 5 μl DNA with 195 μl working solution, and 200 ng DNA was

electrophoresed on a 1.5% agarose TBE gel at 40 V for 24 hours. The plant materials included another 84 marama individuals, 44 of which were plants grown in different geographical locations in Namibia or South Africa, and the remaining 40 were progeny plants grown from seeds collected there, as described in the previous study (Li and Cullis, 2023). The WGS Illumina reads of these 84 individuals are available and stored in the NCBI SRA database (PRJNA779273).

4.2.2 High-throughput sequencing

DNA extracted from Samples 4 and 32 (10 micrograms or more per sample) was sent to the Genomics Core Facility at the Icahn School of Medicine at Mount Sinai for sequencing. The HiFi sequencing libraries were prepared using the SMRTbell® express template prep kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA). SMRT sequencing was performed on four 8M SMRT® Cells (two per sample) on the Sequel® II system. 2,184,632 PacBio HiFi reads with a total length of 21.6 G bases were generated for Sample 4 and 498 Mb for Sample 32. In addition, whole-genome sequencing was performed with the Illumina platform on the DNA from fresh young leaves or embryonic axis of germinating seeds from the 84 samples collected in Namibia and South Africa, as described in the previous study (Li and Cullis, 2023).

4.2.3 Mitogenome assembly and annotation

The PacBio HiFi long reads were assembled using the HiCanu assembler (Nurk et al., 2020; https://canu.readthedocs.io/en/latest/quick-start.html#assembling-pacbio-hifi-with-hicanu). The input genome size was set to 2 Mb to obtain more complete organelle genome contigs. The PacBio long reads spanning the ends of the contigs were used to

further scaffold the assembly. The assembled mitogenome was annotated using MITOFY (Alverson et al., 2010; https://dogma.ccbb.utexas.edu/mitofy/), BLAST (Johnson et al., 2008; https://blast.ncbi.nlm.nih.gov/Blast.cgi), and tRNAscan-se 2.0 (Chan et al., 2021; http://lowelab.ucsc.edu/tRNAscan-SE/). The assembly was verified by mapping Illumina reads and PacBio HiFi reads from different individuals to the generated mitogenome sequences using Bowtie 2 v2.4.4 (Langmead and Salzberg, 2012; https://github.com/BenLangmead/bowtie2) and pbmm2 v1.10.0 (https://github.com/PacificBiosciences/pbmm2), respectively. The alignment was visualized and checked in IGV (Robinson et al., 2011; https://software.broadinstitute.org/software/igv/), showing no ambiguity. The mitochondrial gene arrangement maps of *T. esculentum* were drawn by OGDRAW (Greiner et al., 2019; https://chlorobox.mpimp-golm.mpg.de/OGDraw.html).

4.2.4 Mitogenome polymorphism

The program NUCmer in MUMmer4 (Marçais et al., 2018; https://mummer4.github.io/index.html) was used to locate highly conserved regions in the two types of mitochondrial genomes of *T. esculentum*. The alignment was visualized via a synteny plot drawn by the RIdeogram (Hao et al., 2020; https://github.com/TickingClock1992/RIdeogram) package in R.

The 2,108 bp type 2 unique fragment was blasted in the NCBI database for potential origin. Two pairs of primers were designed for mitogenome typing, using NCBI Primer-BLAST (Ye et al., 2012; https://www.ncbi.nlm.nih.gov/tools/primer-blast/), to amplify across the two ends of this fragment by PCR. The primers to amplify across the left end included the left forward primer (GAGACCGAGCGCAAGAACTA) and the left

reverse primer (TCAGATGGCTAAACAGGCGG), and the product size was 990 bp. The primers to amplify across the right end included the right forward primer (CGCTCGTGACTCATTGAGGA) and the right reverse primer (TTGGTAAGCGGATGCTCTGG), and the product size was 289 bp.

20 uL of mixtures were prepared by separately mixing DNA from six randomly selected *T. esculentum* samples and Promega GoTaq Green Master Mix. The amplifications started with denaturation at 95 °C for 5 min, followed by 30 cycles of 95 ° C for 45 s, 54 °C for 45 s, and 72 °C for 1 min, and a final 72 °C for 5 min.

The Illumina reads from the 84 individuals were mapped to the type 1 reference mitogenome of *T. esculentum* (OK638188 and OK638189) using Bowtie 2 v2.4.4. The alignments were searched for SNPs and indels using SAMtools 1.7 mpileup (Li, 2011; http://www.htslib.org/) and BCFtools 1.8 call (Li, 2011; http://www.htslib.org/doc/1.8/bcftools.html), and visualized in IGV to find heteroplasmy manually. To minimize the interference of sequencing errors and strand bias, only alleles with a frequency of at least 2%, a Phred score above 20, and presence in strands in both orientations were recorded. Alleles in the mtDNA homologous reads in the nuclear genome were excluded. The alleles from the mitochondrial or chloroplast genomes were distinguished by their frequency at the differential loci on the 9,798 bp chloroplast DNA insertion.

4.2.5 Mitogenome divergence

A pairwise comparison was performed on the divergent mitogenomes of *T. esculentum* (OK638188 and OK638189) and six other Fabaceae species, including *Cercis*

canadensis (MN017226.1), *Lotus japonicus* (NC_016743.2), *Medicago sativa* (ON782580.1), *Millettia pinnata* (NC_016742.1), *Glycine max* (NC_020455.1), and *Vigna radiata* (NC_015121.1) using the program PROmer (Delcher et al., 2002) in MUMmer4 to detect the syntenic regions. The alignments were then visualized by the RIdeogram package in R. A synteny block diagram between these seven mitogenomes was draw by Mauve 2.4.0 (Darling et al., 2004; https://darlinglab.org/mauve/mauve.html) and the genes contained in each block were marked on the plot.

4.2.6 SSR analyses

Microsatellites were analyzed by MISA (Beier et al., 2017; https://webblast.ipk-gatersleben.de/misa) on the type 1 reference mitogenome of *T. esculentum*, looking for repeats of motifs one to six base pairs long. The minimum number of repetitions were set to 10, 6, 5, 5, 5, 5 for mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats, respectively.

4.2.7 Phylogenetic tree construction

Two phylogenetic trees were constructed separately, one based on the mitogenome conserved gene sequences to explore the evolutionary relationship between *T. esculentum* and several other selected legumes, and the other tree was built on all differential loci found in the mitogenomes of *T. esculentum* to explore the inter-population and intra-population relationship among the 43 independent samples.

The 24 conserved mitochondrial genes, *atp1*, *atp4*, *atp6*, *atp8*, *atp9*, *nad3*, *nad4*, *nad4L*, *nad6*, *nad7*, *nad9*, *mttB*, *matR*, *cox1*, *cox3*, *cob*, *ccmFn*, *ccmFc*, *ccmC*, *ccmB*, *rps3*, *rps4*, *rps12*, and *rpl16* from the mitogenomes of *T. esculentum* (OK638188 and

OK638189), *Arabidopsis thaliana* (NC_037304.1), *C. canadensis* (MN017226.1), *L. japonicus* (NC_016743.2), *M. sativa* (ON782580.1), *M. pinnata* (NC_016742.1), *G. max* (NC_020455.1), and *V. radiata* (NC_015121.1) were concatenated to make artificial chromosomes. A Maximum Likelihood (ML) phylogenetic tree using the Jukes-Cantor model was built in Mega 11 (Tamura et al., 2021; https://www.megasoftware.net/) after the chromosomes were aligned by Muscle v5 (Edgar, 2022; https://www.drive5.com/muscle/). The topology was validated by a Bayesian inference phylogenetic tree drawn by BEAST v1.10.4 (Suchard et al., 2017; https://beast.community/) and FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Artificial chromosomes concatenated by 40 bp segments at each of the 254 differential loci found in the mitogenomes of *T. esculentum* were prepared for the 43 independent individuals and aligned by Muscle v5. A Maximum Likelihood (ML) phylogenetic tree using the Jukes-Cantor model was drawn on the 43 chromosomes. Frequencies from 1000 bootstrap replicates were labeled on the branches with 40% as cutoff. The topology was verified by a neighbor-joining tree in Mega 11.

4.2.8 Genetic information exchange between organelles

The reference chloroplast genome sequence of *T. esculentum* (KX792933.1) was blasted to its type 1 reference mitogenome (OK638188 and OK638189) and visualized by the Advanced Circos function in TBtools (Krzywinski et al., 2009; Chen et al., 2020; https://github.com/CJ-Chen/TBtools). Primers were designed to verify the presence of the 9,798bp long homologous fragment in both the mitochondrial and chloroplast genomes. Four pairs of primers were designed, using NCBI Primer-BLAST, to amplify the products spanning the two ends of the 9,798 bp cpDNA insertion in the mitochondrial

and chloroplast genomes, respectively. The two pairs of primers designed based on the mitogenome sequence included: MitoLL Forward (ACGCAGAAAAGAGGCCGAA) and MitoLR Reverse (CCTTCGTTTAAGAGAATGTTTTTGG), and the product size was 117 bp. MitoRL Forward (TCTTTGCTACAGCTGATAAAAATCG) and MitoRR Reverse (CCTATGTTCGTTTTCGCCCTG), and the product size was 120 bp. The two pairs of primers designed based on the plastome sequence included: ChlLL Forward (CGTAGTCGGTCTGGCCC) and MitoLR Reverse (CCTTCGTTTAAGAGAATGTTTTTGG), and the product size was 117 bp. MitoRL Forward (TCTTTGCTACAGCTGATAAAAATCG) and ChlRR Reverse (GCTTTTAATAATATGGCCGTGATCT), and the product size was 120 bp.

20 uL of mixtures were prepared by separately mixing DNA from two randomly selected *T. esculentum* samples and Promega GoTaq Green Master Mix. The amplifications started with denaturation at 95 °C for 5 min, followed by 32 cycles of 95 ° C for 30 s, 55 °C for 30 s, and 72 °C for 30 s, and a final 72 °C for 5 min.

4.3 Results

4.3.1 Genome structure and rearrangement

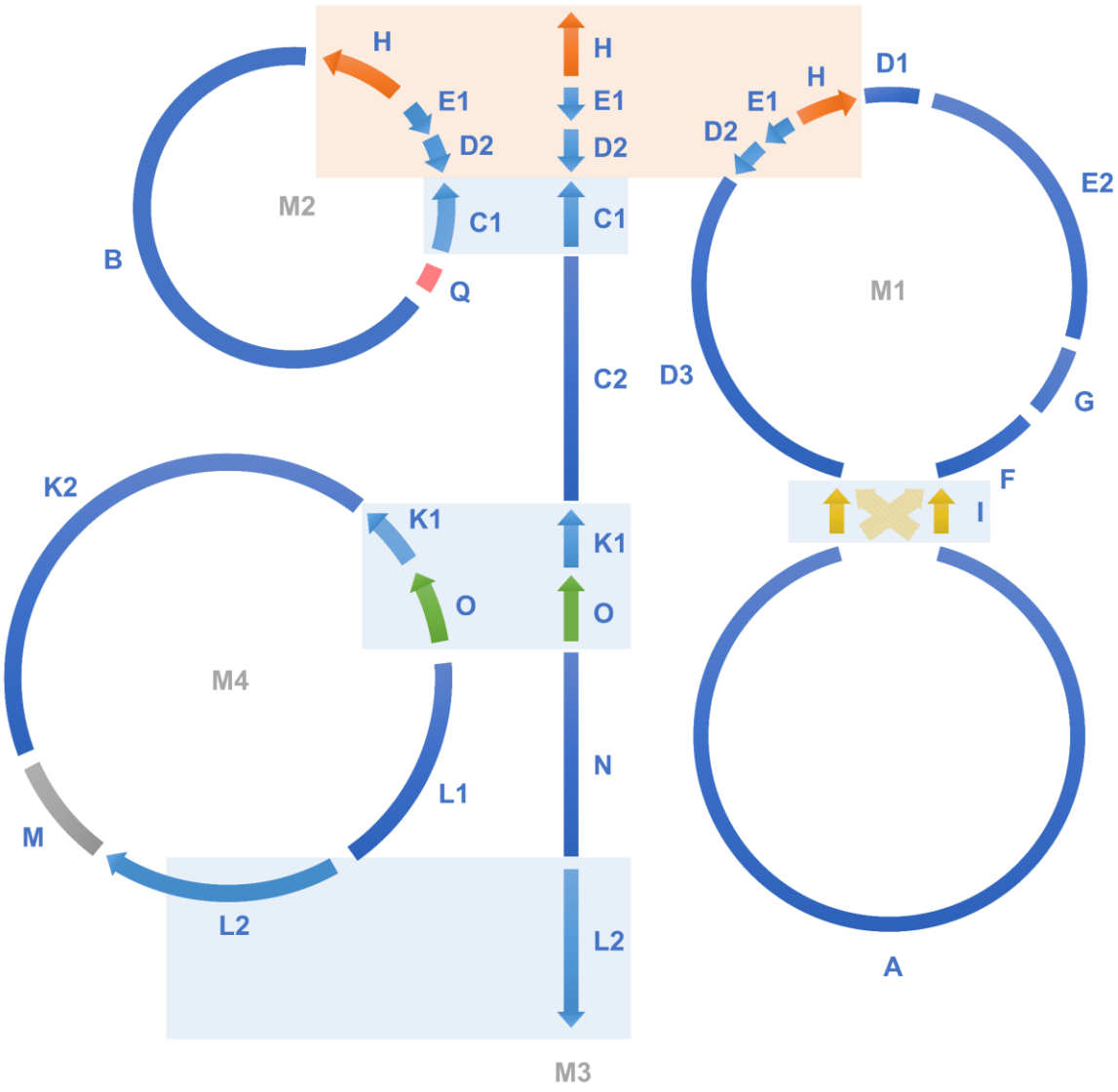

Figure 4. 1 The assembly graph of the type 2 mitogenome of *T. esculentum*. The type 2 mitogenome consists of three circular molecules M1, M2, and M4, and one linear molecule M3, and they were built on 21 long scaffolds (Table 4.1). Blue blocks show double-copy regions that are identical between two chromosomes, and orange blocks indicate triple-copy regions owned by three chromosomes. Close sequencing coverage

was found for single-copy regions of the four chromosomes. Recombination on a pair of long inverted repeats I can change the junction of the upper and lower halves of M1 to that indicated by the yellow dashed arrows. The two structures before and after recombination have been confirmed by PacBio long reads, and their frequencies were close in the same individual (Figure S4.6-4.9). A long chloroplast insertion was found at the position of the gray segment M, about 9,798 bp, and its length varied slightly among different individuals. This chloroplast insertion and long repeats H, I, and O are also present in the type 1 mitogenome of *T. esculentum*. The type 1 mitogenome also has two rings, very similar to the M1 and M4 here, but other molecules have undergone dramatic changes.

When the WGS Illumina reads from the 84 individuals were mapped to the two chromosomes of the reference mitogenome of *T. esculentum*, LS1 (OK638188) and LS2 (OK638189), two distinct mitogenomes were found. The mitogenomes of 45 individuals were similar to the reference mitogenome, with only a few substitutions and indels seen, termed type 1. However, the mitogenomes of the other 39 individuals were very different from the reference, with a large number of sequence and structural differences, termed type 2 (Table S4.1). This is consistent with the previously published study of *T. esculentum* chloroplast genomes, where these 84 individuals were found to contain two distinct germplasms (Li and Cullis, 2023). These two cytotypes actually differ not only in the chloroplast genome but also in the mitochondrial genome. The PacBio HiFi reads from the type 2 individual Sample 4 were assembled by Canu to generate three circular molecules M1 (OP795449), M2 (OP795450), and M4 (OP795447), and one linear chromosome M3 (OP795448), with a total length of 436,568 bp and a GC content of

44.8% (Table 4.1) (Figures S4.1-4.5). The four chromosomes consisted of 21 contigs assembled directly from Illumina reads of type 2 individuals, containing four double-copy regions and one triple-copy region (Figure 4.1 and Table 4.2). Among these multi-copy regions, homologous sequences of contigs H and I were also doubled in coverage in the type 1 mitogenome of *T. esculentum*, but the rest were present as single-copy sequences in the type 1 mitogenome (Li and Cullis, 2021).

Table 4. 1 Chromosome base composition of the type 2 *T. esculentum* mitogenome.

Molecule	A (%)	C (%)	G (%)	T (%)	G~C (%)	Length (bp)
M1	27.90	22.18	22.22	27.70	44.40	169,406
M2	27.60	21.63	23.26	27.52	44.90	56,355
M3	27.46	22.25	22.73	27.55	45.00	97,120
M4	27.70	22.58	22.44	27.29	45.00	113,687

Table 4. 2 Length of the primary scaffolds constituting the type 2 *T. esculentum* mitogenome.

Unit number	Length (bp)	Unit number	Length (bp)
A	82,874	K1	4,052
B	39,273	K2	52,730
C1	30,017	L1	20,069
C2	5,272	L2	22,301
D1	5,866	N	27,581
D2	2,722	O	4,881
D3	26,671	M	9,654
H	5,212	Q	2,108
E1	1,768		
E2	25,384		
F	8,590		
G	5,798		
I	2,265		

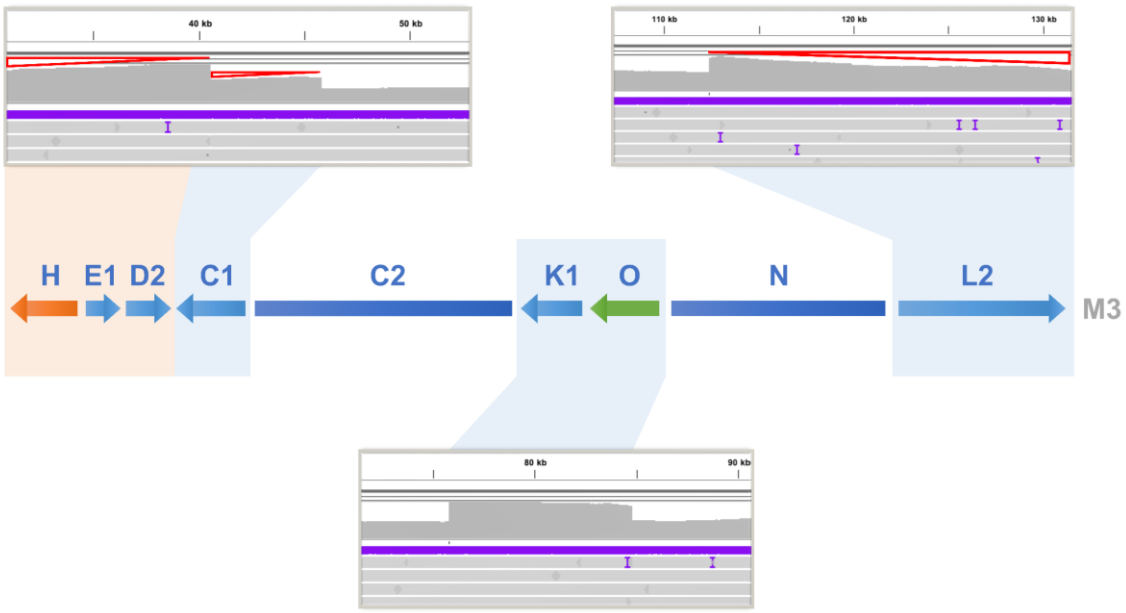

Figure 4. 2 Changes in sequencing coverage on the type 2 mitogenome chromosome M3 of *T. esculentum*. PacBio HiFi reads from the type 2 individual Sample 4 were aligned to the multicopy regions of the chromosome M3 using pbmm2. Chromosome M3 contains some long repeats that are identical to parts of other chromosomes, thus increasing the sequencing coverage of these regions in the alignment (Figures S4.10-4.12). At the positions of scaffolds C1, K1-O, and L2, with blue shading, the read depth was found to be doubled. At the location of scaffold H shaded in orange, the sequence depth was increased to 3-fold. However, a progressive decrease in coverage indicated by red triangles was also seen at both ends of M3, as some linear M3 chromosomes had degenerated at both ends without telomere protection.

Both ends of the chromosome M3 were found to be long repeats that were homologous to parts of other chromosomes (Figure 4.2). In addition, in a very long range of 15 to 20 kb, the sequencing depth gradually decreased towards both ends, and many reads were found to stop in this range. This furthers confirms that this is a linear

chromosome that exists in different lengths in cells because of the lack of telomere protection. Linear chromosomes have been found to stably exist in eukaryotic cells even in the absence of telomeres, through strand-invasion between terminal sequences and their homologous internal sequences to form t-loops to protect the chromosomes from degradation (de Lange, 2015). Because the repeats at both ends of M3 are very long, the PacBio reads we obtained cannot span them to verify whether this linear molecule recombines with other chromosomes. Long range PCR amplification that can amplify sequences above 20 kb can be considered here to answer this question.

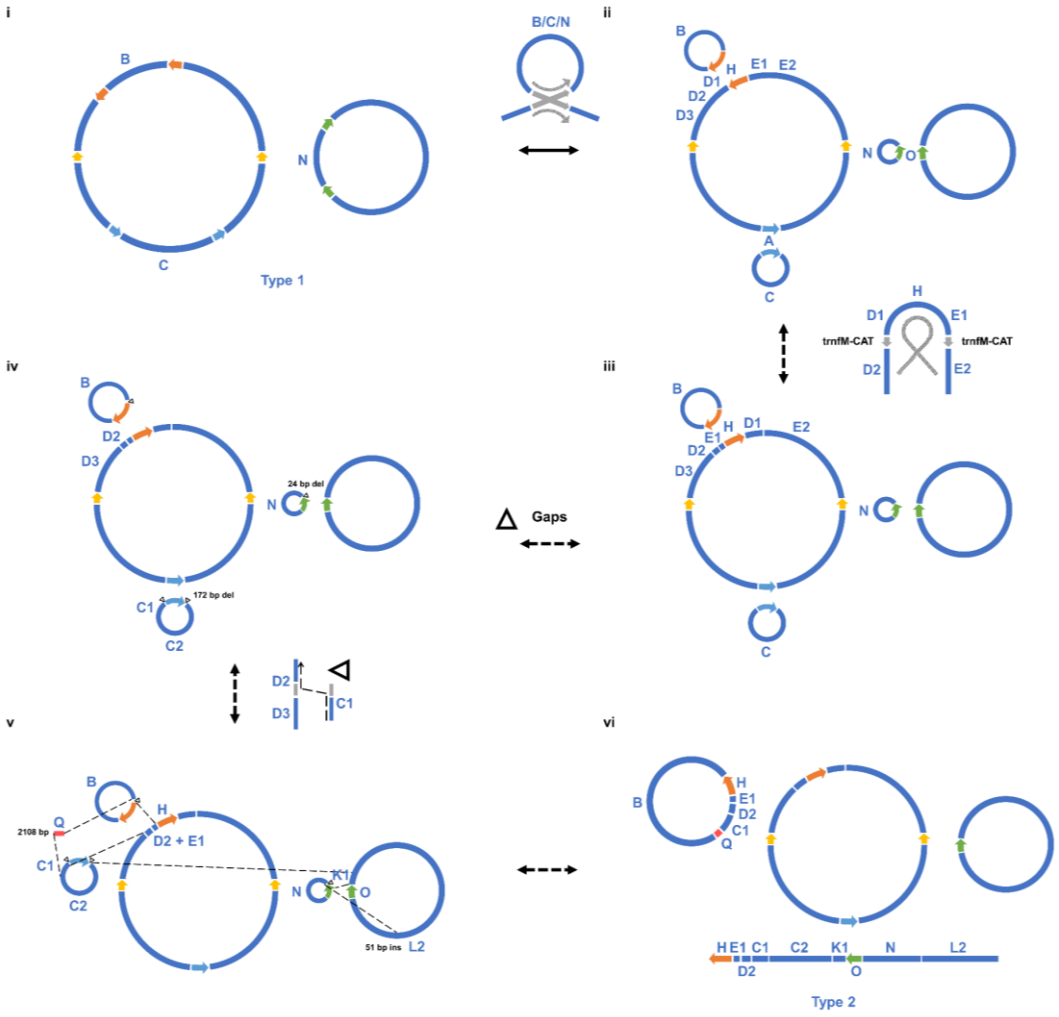

Figure 4. 3 Step-by-step analysis of the structural differences between the two types of *T. esculentum* mitogenomes. i. The two autonomous circular chromosomes of type 1 *T. esculentum* mitogenome, LS1 (OK638188) (left) and LS2 (OK638189) (right). Colored arrows indicate the four pairs of long repeats (>1 kb). ii. Recombination on the direct repeats split the two large rings into five small circular molecules. Both conformations before and after recombination have been confirmed by PacBio reads to exist in type 1 individuals. iii. A rare recombination on a pair of *trnfM* genes at the junctions of D1 and

D2, and E1 and E2, inverted the sequence D1-H-E1 in between. iv. Gaps with or without sequence deletions resulted in the three small circular chromosomes of the type 1 mitogenome present as different forms in type 2 individuals (Figures S4.13-4.18). v. New DNA fragments, including a 2,108 bp contig Q unique to type 2 individuals, joined originally remotely located sequences to form new structures. vi. The final type 2 mitogenome of *T. esculentum* with three circular and one linear chromosomes.

Numerous differences were found between the type 1 and type 2 mitogenome structures (Figure 4.3). One is a rare recombination on a pair of *trnfM* genes on the large circular molecule M1, which inverts the 12,846 bp sequence in between. The sequence before inversion also appeared in type 2 plants, with a frequency less than 2%. Furthermore, the three small rings of the type 1 mitogenome exist in different forms in the type 2 mitogenome, and four gaps have been found on them. New type 2 exclusive fragments were discovered, including a 2,108 bp segment, which connected originally distantly located contigs C1 and B. In addition, the recombination on a pair of 35 bp direct repeats joined contigs C1 and D2 to form a new circular molecule M2. C2 was found to connect with K1 and then further extended to N and L2 to form a linear chromosome, but the mechanism behind this is unclear. It can be seen that the deletion and insertion of the entire DNA segment, alongside repeat mediated recombination, can lead to dramatic changes in the mitogenome structure.

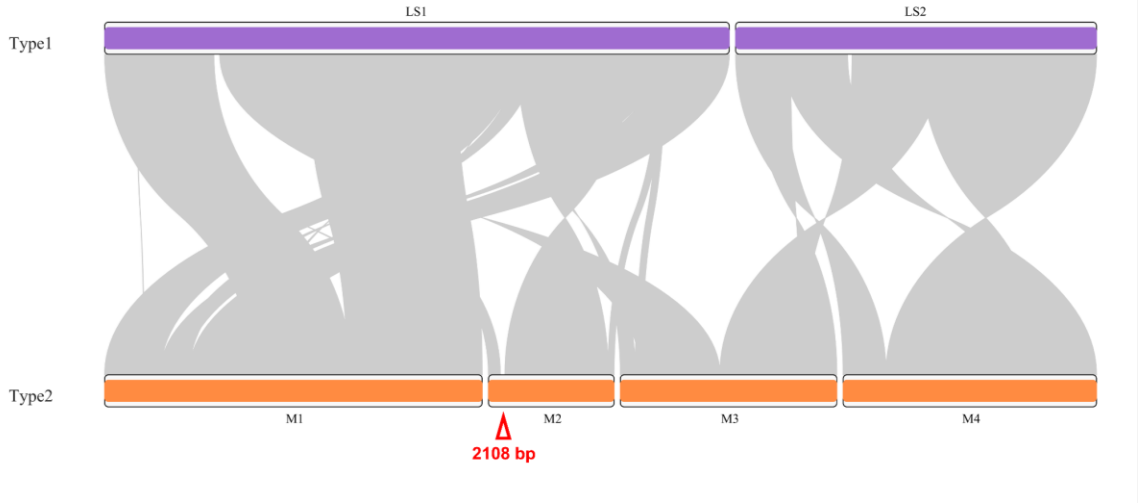

Figure 4. 4 Synteny visualization of the two types of mitogenomes of *T. esculentum* by the R package RIdeogram after NUCmer alignment. The red triangle indicates the 2,108 bp type 2 mitogenome exclusive fragment of *T. esculentum*.

The two types of *T. esculentum* mitogenomes were compared by NUCmer alignment and then visualized by a synteny diagram, showing a high degree of similarity (Figure 4.4). The basic blocks making up the two mitogenomes are highly similar, except for the 2,108 bp type 2 unique fragment and some other short insertions and deletions, but the order of these blocks has been changed by recombination. When the WGS Illumina reads from the 84 individuals were mapped to the region where the 2,108 bp type 2 unique fragment resides, 39 samples were found to contain this fragment while 45 plants did not (Figures S4.19-4.28). Furthermore, in type 1 mitogenome, LS1 and LS2 are two autonomous circular chromosomes that do not recombine into one master circle, but in type 2 mitogenome, a linear chromosome M3 was found to contain homologous sequences from both LS1 and LS2, suggesting that these two molecules may have been related in evolution.

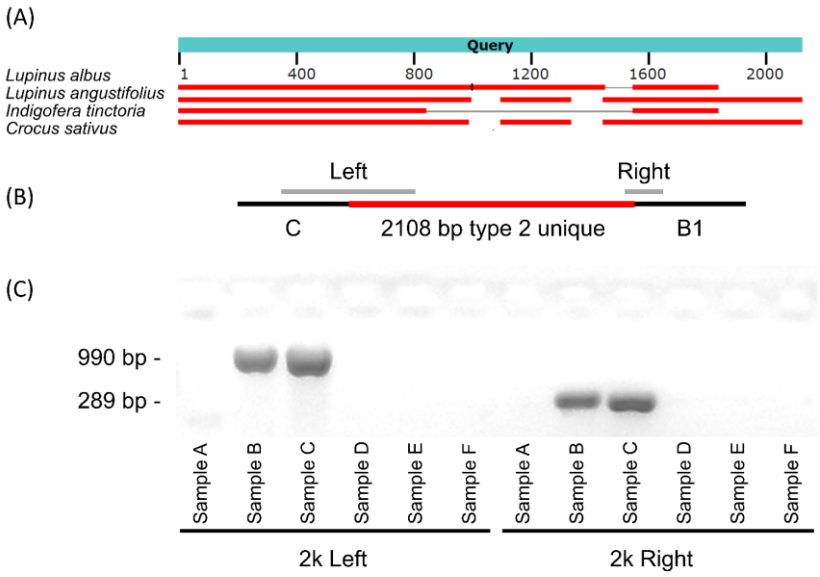

Figure 4. 5 Homology analysis of the 2,108 bp fragment unique to type 2 *T. esculentum* mitogenome and design of primers for its PCR identification. (A) The 2,108 bp type 2 mitogenome exclusive sequence was blasted as a query in the NCBI database. Red horizontal bars indicate where database sequences are aligned, and separately aligned regions from the same database subject are connected by thin gray lines. (B) Two pairs of primers were designed to amplify products across both ends of the 2,108 bp fragment. The estimated size of the left end product is 990 bp, and the right end product is 289 bp. (C) Gel image of PCR amplification of DNA from six randomly selected samples with the two pairs of primers designed separately. The PCR products were electrophoresed on a 1.5% agarose gel at 80V for 1 hour.

Blast results showed that this 2,108 bp type 2 mitogenome exclusive fragment was highly similar to the mitochondrial sequences of Fabaceae species *Lupinus albus* and *Indigofera tinctoria*, suggesting that this fragment was possibly derived in the evolution from *Lupinus* or *Indigofera* (Figure 4.5A). Two pairs of primers were designed and found to effectively identify the 2,108 bp fragment. As shown in Figure 4.5B and 4.5C, sample B and C, out of the six randomly selected samples, contained both the 990 bp left end band and the 289 bp right end band after amplification, indicating that only these two of the six had type 2 mitogenomes.

4.3.2 Gene annotation

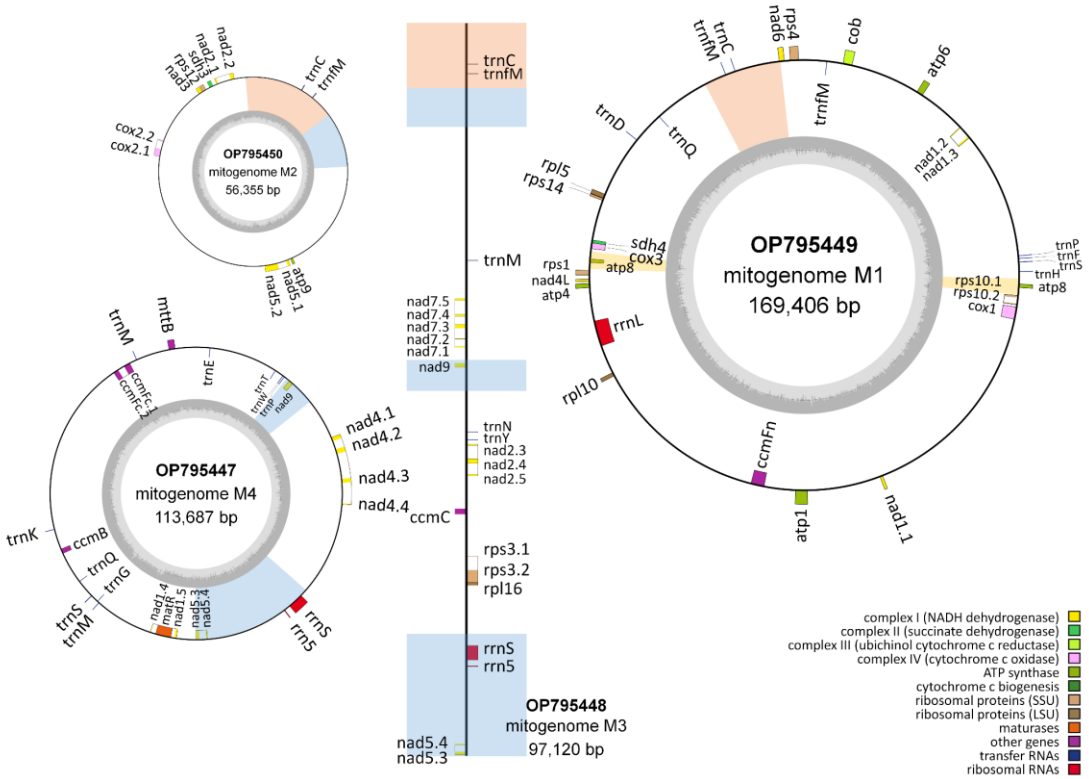

Figure 4. 6 The map of type 2 *T. esculentum* mitogenome gene arrangement drawn by OGDRAW. The annotation was performed by MITOFY and BLAST on the mtDNA of

individual Index1 from UP Farm and deposited in GenBank under accession numbers OP795447-OP795450. All type 2 plants were found to have a similar gene arrangement. The dark gray pattern in the inner circle indicates GC content. Genes are colored according to their function. The decimal part after the gene name indicates the order of the exons. Genes inside the circle are transcribed clockwise, while those outside the circle are transcribed counterclockwise. Blue blocks represent two-copy regions that are identical between two chromosomes, and orange blocks show three-copy regions that are the same across three chromosomes.

Table 4. 3 Gene annotation of the type 2 mitogenome of *T. esculentum*.

Category	Gene Name			
	M1	M2	M3	M4
Complex I (NADH dehydrogenase)	*nad1#, nad4L, nad6*	*nad2#, nad3, nad5#*	*nad2#, nad5#, nad7#, nad9*	*nad1#, nad4#, nad5#, nad9*
Complex II (succinate dehydrogenase)	*sdh4*	*sdh3*		
Complex III (ubiquinol cytochrome-c reductase)	*cob*			
Complex IV (cytochrome-c oxidase)	*cox1, cox3*	*cox2#*		
Complex V (ATP synthase)	*atp1, atp4, atp6, atp8*(2)*	*atp9*		
Cytochrome c biogenesis	*ccmFn*		*ccmC*	*ccmB, ccmFc#*
Large subunit ribosomal proteins	*rpl5, rpl10*		*rpl16*	
Small subunit ribosomal proteins	*rps1, rps4, rps10#, rps14*	*rps12*	*rps3#*	
Maturases				*matR*
Transport membrane protein				*mttB*
Ribosomal RNAs	*rrnL*		*rrn5, rrnS*	*rrn5, rrnS*
Transfer RNAs	*trnD-GTC, trnC-GCA, trnQ-TTG,*	*trnC-GCA,*	*trnN-GTT, trnC-GCA, trnfM-CAT,*	*trnQ-TTG, trnE-TTC, trnG-GCC ,*

Category	Gene Name			
	trnH-GTG, trnfM-CAT(2), trnF-GAA, trnP-TGG, trnS-GCT*	*trnfM-CAT*	*trnM-CAT, trnY-GTA*	*trnK-TTT, trnM-CAT*(2), trnP-TGG, trnS-TGA, trnT-TGT#, trnW-CCA*

Genes with introns are marked with #. Genes with multiple copy numbers on the same chromosome are labeled with *, and the numbers in parentheses indicate the corresponding copy numbers.

Both type 1 and type 2 mitogenomes of *T. esculentum* were found to contain 35 unique protein-coding genes, 3 unique rRNA genes, and 16 different tRNA genes (Figure 4.6; Table 4.3) (Li and Cullis, 2021). The type 2 mitogenomes have two copies of *nad9* and *atp8*. The gene *nad9* is located on contig K1, a long repeat possessed by both chromosomes M3 and M4 in type 2 mitogenomes. Whereas, there is only one copy of K1 in type 1 mitogenomes. The gene *atp8* is located on a pair of long repeats J, so its copy number is doubled as is the case in both types of mitogenomes. The copy number of exon 3 and 4 of gene *nad5* is also doubled in type 2 mitogenomes but not in type 1, and it is not known whether this affects its expression level. In addition, there are two copies of *rrn5* and *rrnS* in type 2 mitogenomes but only one copy in type 1. A total of 26 tRNA genes were found in type 2 mitogenomes, including four copies of *trnfM-CAT*, three copies of *trnM-CAT*, three copies of *trnC-GCA*, two copies of *trnP* and *trnQ*, and 12 single-copy tRNA genes.

The atypical start codon ACG was used by three genes *nad1*, *nad4L*, and *rps10*, and ATT was used by gene *mttB*. This is consistent with the research on the mitogenome of common beans from which these four genes were all reported to use an alternative initiation codon ACG (Bi et al., 2020). C-to-U editing was found to be widely used in mitochondrial and chloroplast genes in land plants (Takenaka et al., 2013). ATT is also usually used as an alternative start codon in the mitogenome. For example, *mttB* in *Salix purpurea* was reported to use an ATT start codon as well (Wei et al., 2016).

4.3.3 Mitogenome divergence

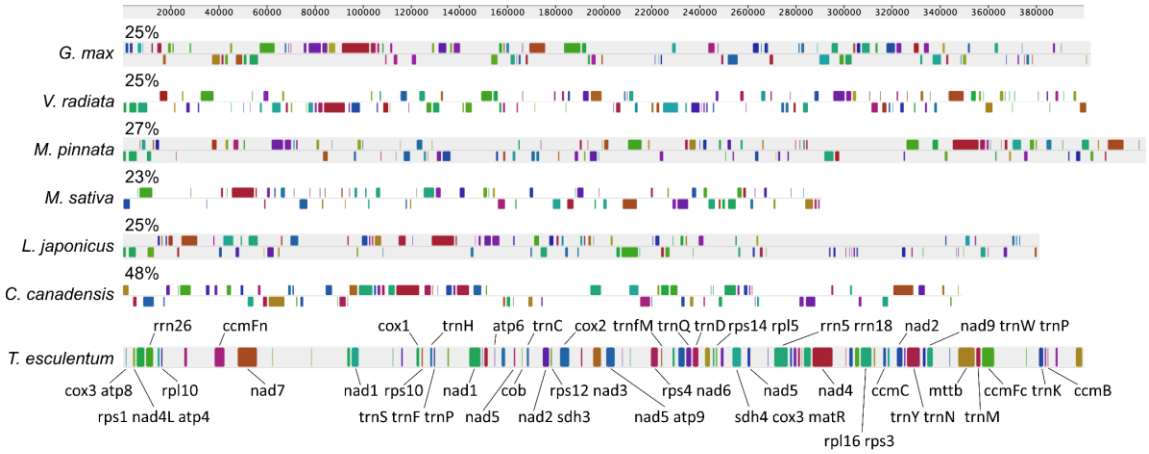

Figure 4. 7 Synteny block diagram of the Mauve alignment between the mitogenomes of *T. esculentum* and six other Fabaceae species, *Cercis canadensis* (MN017226.1), *Lotus japonicus* (NC_016743.2), *Medicago sativa* (ON782580.1), *Millettia pinnata* (NC_016742.1), *Glycine max* (NC_020455.1), and *Vigna radiata* (NC_015121.1). Similarities (percentage of marama mitogenome covered) are labeled above the bars. Genes contained in the synteny blocks are marked by gene symbols.

The mitogenome of *T. esculentum* was highly divergent from those of the six selected legume species, which covered 23% to 48% of the marama mitogenome, ranging in length from 91.9-191.8 kb (Figure 4.7). Of these species, *C. canadensis* was most closely related to marama, while *M. sativa* was the least similar to marama. The mitogenomes of *G. max*, *L. japonicus*, and *V. radiata* all contain homologous sequences covering 25% of the marama mitogenome, equal to 99.89 kb in length. *Bauhinia variegata* is closer to marama than *C. canadensis* in the phylogenetic tree, but its mitogenome sequence is not available in NCBI GenBank (Wunderlin, 2010).

Table 4. 4 List of mitochondrial protein-coding genes lost during the evolution of some Fabaceae species.

Gene	*T. esculentum*	*C. canadensis*	*L. japonicus*	*M. sativa*	*M. pinnata*	*V. radiata*	*G. max*
sdh3	+	+	-	-	+	-	-
sdh4	+	+	-	-	-	-	-
cox2	+	+	+	+	+	-	+
rpl2	-	-	-	-	-	-	-
rpl5	+	+	+	+	+	+	-
rpl10	+	+	-	-	-	-	-
rps1	+	+	-	+	+	+	+
rps2	-	-	-	-	-	-	-
rps7	-	-	-	-	-	-	-
rps8	-	-	-	-	-	-	-
rps11	-	-	-	-	-	-	-
rps13	-	-	-	-	-	-	-
rps19	-	-	-	-	-	-	-

The loss of mitochondrial protein-coding genes and the functional transfer of these genes from the organelle genome to the nuclear genome are common in the evolution of angiosperms, but some plants tend to retain a more complete set of

mitochondrial genes (Palmer et al., 2000). The mitogenomes of Cercidoideae species *T. esculentum* and *C. canadensis* contain functional protein-coding genes *sdh3*, *sdh4*, and *rpl10*, which have been lost in many other legumes (Table 4.4). Rare gene losses were also seen in the mitogenomes of legumes, such as *rpl5* in *G. max*, *cox2* in *V. radiata* and *rps1* in *L. japonicus*, but these genes all remain intact and functional in *T. esculentum* and *C. canadensis* (Alverson et al., 2011; Kazakoff et al., 2012; Chang et al., 2013).

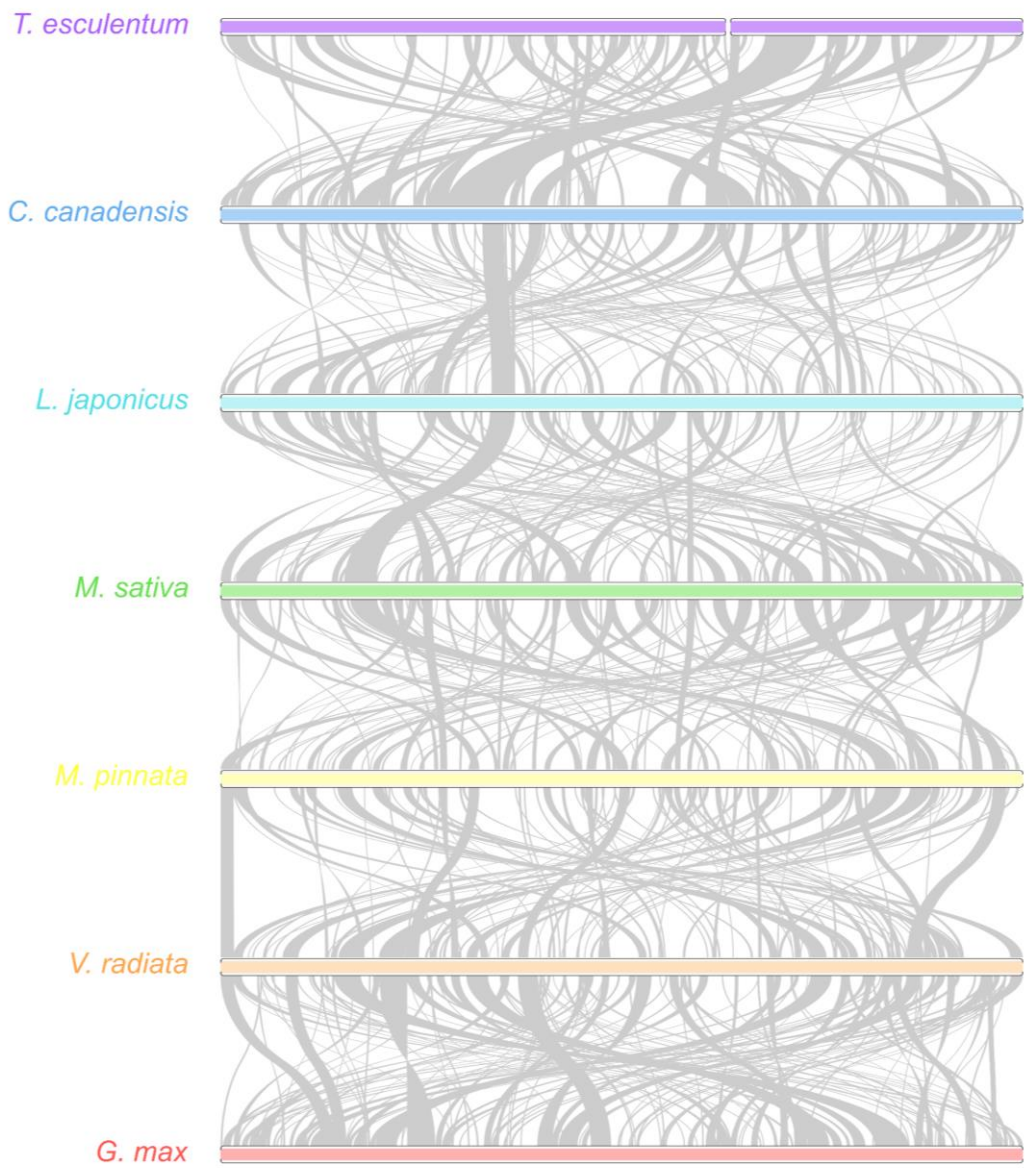

Figure 4. 8 Synteny maps of the mitochondrial genomes of seven legume species. The colored bars represent the mitochondrial chromosomes of *C. canadensis* (MN017226.1), *L. japonicus* (NC_016743.2), *M. sativa* (ON782580.1), *M. pinnata* (NC_016742.1), *V. radiata* (NC_015121.1), *G. max* (NC_020455.1), and *T. esculentum*, which contains two chromosomes, LS1 (OK638188) and LS2 (OK638189). The gray ribbons indicate homologous sequences between the two neighboring species. The species were ordered

according to the phylogenetic tree in Figure 4.9. Promer was used to detect syntenic regions between highly divergent genomes, which were then visualized by RIdeogram package in R.

In a pairwise comparison of the mitogenomes of seven legume species, the synteny plot revealed numerous rearrangements and a high degree of divergence among the mitogenomes of even closely related species (Figure 4.8). The mitogenomes of *C. canadensis* and *T. esculentum* contain many distinct regions but also some long homologous segments.

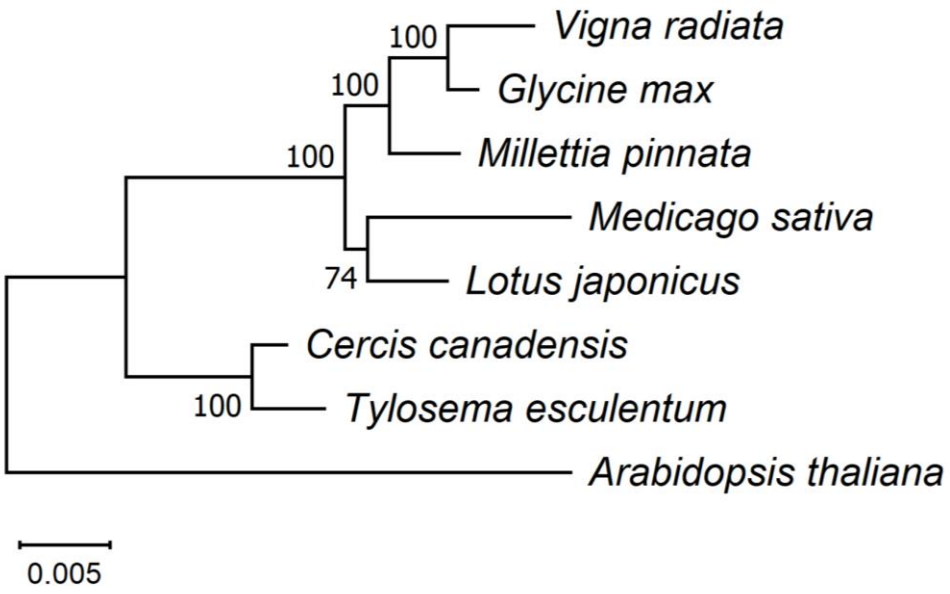

Figure 4. 9 Maximum Likelihood (ML) phylogenetic tree with the Jukes-Cantor model based on artificial chromosomes concatenated by 24 conserved mitochondrial genes, *atp1*, *atp4*, *atp6*, *atp8*, *atp9*, *nad3*, *nad4*, *nad4L*, *nad6*, *nad7*, *nad9*, *mttB*, *matR*, *cox1*, *cox3*, *cob*, *ccmFn*, *ccmFc*, *ccmC*, *ccmB*, *rps3*, *rps4*, *rps12*, and *rpl16* from *Arabidopsis thaliana* (NC_037304.1), *Cercis canadensis* (MN017226.1), *Lotus japonicus*

(NC_016743.2), *Medicago sativa* (ON782580.1), *Millettia pinnata* (NC_016742.1), *Glycine max* (NC_020455.1), and *Vigna radiata* (NC_015121.1) in NCBI. The tree was drawn in Mega 11 after sequence alignment with Muscle v5. Percentage probabilities based on 1000 bootstrap replications are labeled on the branches. The topology was validated by the Bayesian inference phylogenetic tree drawn by BEAST (Figure S4.29).

The phylogenetic tree shown in Figure 4.9 was built on the 24 conserved mitochondrial protein-coding genes *atp1*, *atp4*, *atp6*, *atp8*, *atp9*, *nad3*, *nad4*, *nad4L*, *nad6*, *nad7*, *nad9*, *mttB*, *matR*, *cox1*, *cox3*, *cob*, *ccmFn*, *ccmFc*, *ccmC*, *ccmB*, *rps3*, *rps4*, *rps12*, and *rpl16*, which are present in all these eight species. This tree is consistent with previously published phylogenetic trees constructed on chloroplast protein-coding genes (Kim and Cullis, 2017; Wang et al., 2018). As another species of Cercidoideae, *C. canadensis* was expected to be the closest relative of these plants to *T. esculentum*. Among Faboideae species, *M. sativa* and *L. japonicus* are closely related, and *V. radiata*, *G. max*, and *M. pinnata* belong another clade.

4.3.4 Nucleotide polymorphism

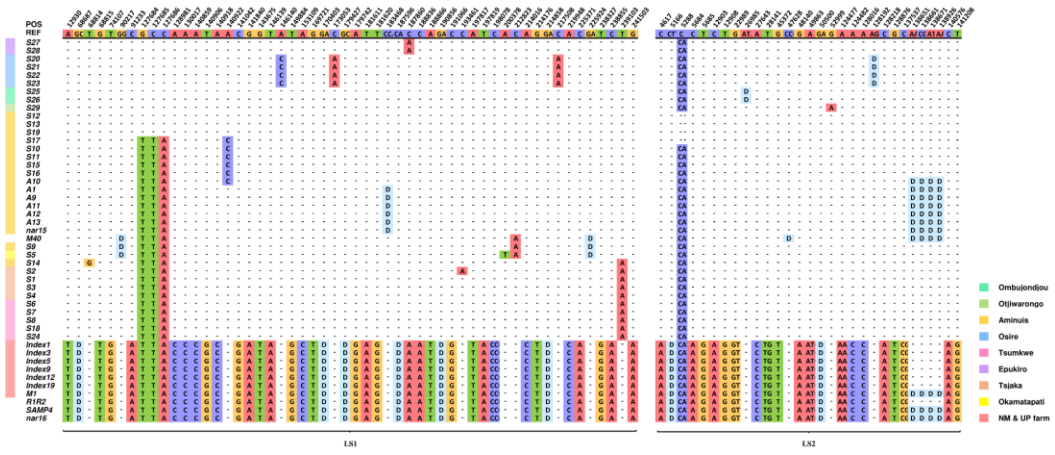

Figure 4. 10 Nucleotide matrices showing the distribution of mitochondrial genome variations in the 43 independent individuals and 4 additional samples of unknown origin. The first row indicates the alleles of the type 1 reference mitogenome of *T. esculentum* (LS1:OK638188 and LS2: OK638189). From the second row onwards, only bases different from the reference are shown, and bases identical to the reference are replaced by dashes. The two types of mitogenomes also differ from each other at another 170 loci, not shown here to save space (no within-type differences were found at those 170 loci). The full variation distribution is shown in Figure S4.30 and S4.31. All insertions are represented by the first two bases and deletions by the letter "D". The color bar to the left of the plant ID shows the source of the sample and is left blank for unknown sources.

17 haplotypes were found in these 47 plants, which could be clearly divided into two groups: namely the type 2 plants from the Namibian farm and the UP farm and the remaining type 1 plants (Figure 4.10). The mitogenomes of type 2 plants are relative conserved, which may be caused by sampling errors, and a larger sample size is needed to verify. The only differences between type 2 plants were four deletions at four closely

located loci on the chromosome LS2. However, the type 1 plants can be divided into many groups according to the variations. Some geographical patterns can be seen in the distribution of variation. For example, in the four plants from Osire, there were three substitutions on the chromosome LS1, A>C at 146,140 bp, C>A at 173,053 bp, and C>A at 217,508 bp, and a deletion at 128,192 bp on the chromosome LS2. These are variations unique to Osire plants. Similar geographic-specific variation can be seen in plants elsewhere. However, our data may still have sampling issues. For example, only a single sample was collected in some regions including Otjiwarongo and Okamatapati. In addition, although distant wild individuals in each geographic region were intentionally selected, there is no guarantee that they are not related. The findings here still need to be validated by sequencing more samples and studied alongside the phenotypic performance of the plants to determine whether any of these variations are the result of plants evolving to better adapt to different environments.

Table 4. 5 Total number of variations found when mapping the WGS reads of all 84 individuals to the type 1 *T. esculentum* reference mitochondrial genomes OK638188 and OK638189.

Variation Type	Type 2 vs. Ref.	Type 1 vs. Ref.	Total
LS1			
Deletion	29 (29)	3 (3)	32
Insertion	34 (34)	0	34
SNP	79 (75)	13 (9)	88
LS2			
Deletion	25 (21)	7 (3)	27
Insertion	18 (17)	1 (0)	18
SNP	54 (54)	1 (1)	55
	239 (230)	24 (16)	254

Type 2 refers to 7 Index plants excluding Index8 from the Pretoria Farm, 29 M descendent plants originally from the Namibia Farm excluding M40, and two individuals R1R2 and nar16 of unknown origin. Type 1 represents all remaining plants, including A plants, S plants, Index8, and M40. Numbers in parentheses indicate counts of exclusive variants of this type. LS1 and LS2 are type 1 marama reference mitochondrial chromosomes in GenBank with accession numbers OK638188 and OK638189.

A total of 254 differential loci were found in the mitogenomes of the 84 *T. esculentum* individuals, including 143 SNPs, 52 insertions, and 59 deletions (Table 4.5). Type 1 and type 2 mitogenomes differed at 230 loci, including 129 substitutions, 50 deletions, and 51 insertions. The mitogenomes of type 2 plants differed at only 4 loci, whereas that of type 1 plants differed at 24 loci.

The mitochondrial gene sequence of *T. esculentum* is very conserved. A total of 11 variations were found in the mitochondrial gene sequence, and only 1 of them was in the coding sequence, which was a 2368A>G substitution and resulted in a N303D change in the gene *matR* (Table 4.6). Furthermore, 10 of the 11 variations were found on one subgenomic ring LS2 of the reference mitogenome of *T. esculentum*. Whether the chromosome LS1 is more conserved than the chromosome LS2 is unknown. Although the intergenic spacer of LS1 contained more variations than LS2, the gene sequence on LS1 appeared to be more conserved, and cpDNA insertions were also found rarely in LS1, but abundantly in chromosome LS2.

Table 4. 6 Variations found in *T. esculentum* mitochondrial gene sequences of the 84 individuals.

Position	Variant	Gene		Product
LS1 54113	Indel	*nad7*	Intron	NADH dehydrogenase subunit 7
LS2 2368	SNP	*matR*	Exon	maturase
LS2 35483	Indel	*nad4*	Intron	NADH dehydrogenase subunit 4
LS2 37645	SNP	*nad4*	Intron	NADH dehydrogenase subunit 4
LS2 39597	Indel	*nad4*	Intron	NADH dehydrogenase subunit 4
LS2 40927	SNP	*nad4*	Intron	NADH dehydrogenase subunit 4
LS2 41027	Indel	*nad4*	Intron	NADH dehydrogenase subunit 4
LS2 56879	SNP	*rps3*	Intron	ribosomal protein S3
LS2 69748	SNP	*nad2*	Intron	NADH dehydrogenase subunit 2
LS2 71388	Indel	*nad2*	Intron	NADH dehydrogenase subunit 2
LS2 71633	Indel	*nad2*	Intron	NADH dehydrogenase subunit 2

LS1 = OK638188, LS2 = OK638189

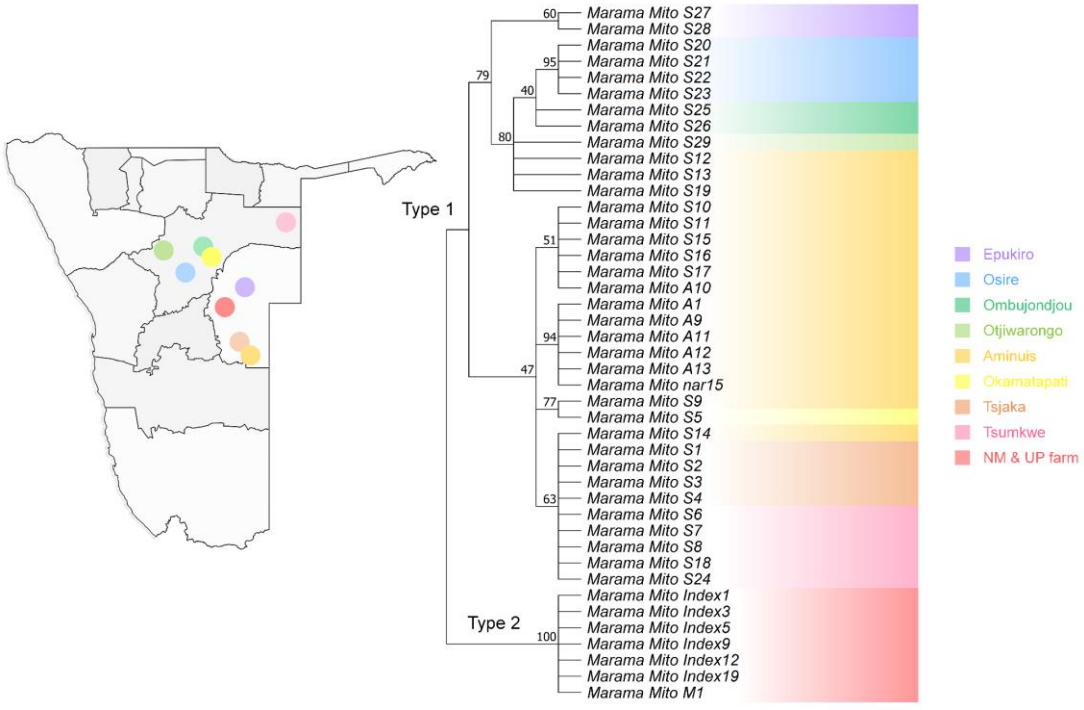

Figure 4. 11 Maximum Likelihood (ML) phylogenetic tree with the Jukes-Cantor model built on artificial chromosomes concatenated by 40 bp fragments at each of the 254

differential loci in the mitogenomes of *T. esculentum* according to the mitogenome sequences of the 43 independent individuals. Frequencies from 1000 bootstrap replicates were labeled on the branches with 40% as cutoff. The topology was verified by the neighbor-joining method in Mega 11. Individuals with the same background color came from the same geographic location, and the sampling points were marked on the map of Namibia.

In the phylogenetic tree constructed on the differential loci of the mitogenomes of *T. esculentum*, the two germplasms fell into two clusters as expected (Figure 4.11). Furthermore, the type 1 plants were then divided into groups. Tsumkwe and Tsjaka were two distant sampling sites, but the plants from these two locations were clustered in one clade, which was also seen in the phylogenic tree built on the complete chloroplast genome. One wondered whether there were factors other than geographical distance that determined the grouping of these plants. On the other hand, plants from Epukiro, Osire, Ombujondjou, and Otjiwarongo belonged to one clade, and these sites were geographically close together and were all arid areas with soil moisture anomalies (SMA) typically below -.04 m^3/m^3 (NASA, n.d.). In contrast, plants from another clade seemed to grow in less arid regions. Notable differences between the mitogenomes of plants from the two clades included three consecutive substitutions from GCC to TTA at positions 127,684 to 127,686 on chromosome LS1 (OK638188), although the function is unknown. Of course, the sample size of this study is still relatively small, and more plants need to be collected in areas with different soil moisture for verification.

4.3.5 SSRs and heteroplasmy analyses

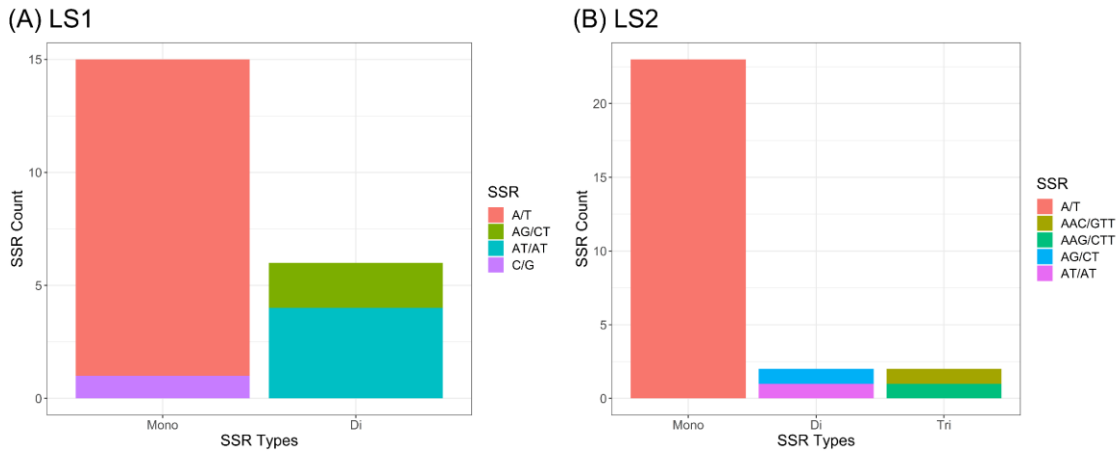

Figure 4. 12 Distribution of SSR motifs of different repeat types in the type 1 reference mitogenome of *T. esculentum* analyzed by MISA. (A) Chromosome LS1 (OK638188). (B). Chromosome LS2 (OK638189).

A total of 48 SSR motifs were found by MISA in the reference mitogenome of *T. esculentum*, LS1 (OK638188) and LS2 (OK638189), of which 38 were simple mononucleotide microsatellites, accounting for 79.2% of all discovered SSR motifs (Figure 4.12). Among them, 37 are A/T mononucleotide repeats, and only one is a G/C repeat. There are 8 dinucleotide repeats and 2 trinucleotide repeats. No simple sequence repeats with core motifs of four nucleotides or longer were found. There are three microsatellites in the coding sequence of the gene, including two 10 bp A/T repeats, one located at the boundary of the coding sequence of the gene *sdh3*, the other located in the exon 4 of the gene *nad1*, and a 12 bp AT repeat siting in the coding sequence of *mttB*.

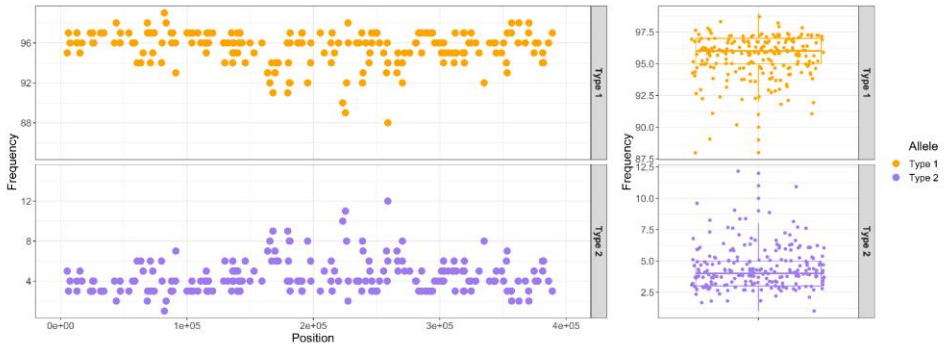

Figure 4. 13 Allele frequency plot of all differential loci between the two types of mitogenomes of *T. esculentum* in Aminuis individual A11. The x-axis of the left panel indicates the position on the chromosome concatenated by the reference marama mitogenomes LS1 (OK638188) and LS2 (OK638189).

As an individual collected from Aminuis, A11 had approximately 96% type 1 alleles and 4% type 2 alleles at all differential loci between the two types of marama mitogenomes (Figure 4.13). It has been confirmed that these minor alleles are not from homologous sequences in the nuclear genome or the chloroplast genome, suggesting the presence of two different mitogenomes in the same individual. It was previously reported that heteroplasmy also exists in the chloroplast genome of A11, with a minor genome frequency reaching 11% (Jin and Cullis, 2023). This means that in the chloroplasts and mitochondria of A11, both major and minor genomes exist, and the frequency of the minor genome is higher than that of the other studied plants. This is most likely to be caused by an accidentally occurred paternal leakage. The proportions of mitochondrial and chloroplast minor genomes differed in A11, indicating that this may be true heterogeneity rather than due to accidental mixing of samples. However, being the only

sample with high overall organelle genome heterogeneity is not convincing, and more plants from Aminuis need to be sequenced and studied.

In the mitogenomes of other individuals, heteroplasmy was found to be less common than in their chloroplast genomes, and even low proportions of base substitutions below 2% were relatively rare, and even if present, many were found from mitochondrial homologous segments in the nuclear genome. However, at a few differential loci, including the three consecutive substitutions at positions 127,684 to 127,686 and another substitution at 140,922 on chromosome LS1 (OK638188), obvious heteroplasmy could be seen in multiple individuals, with the proportion of minor alleles even up to 35%. These loci may have played important roles in the evolution of these individuals under environmental selection.

4.3.6 Sequence transfer between chloroplast and mitochondrial genomes

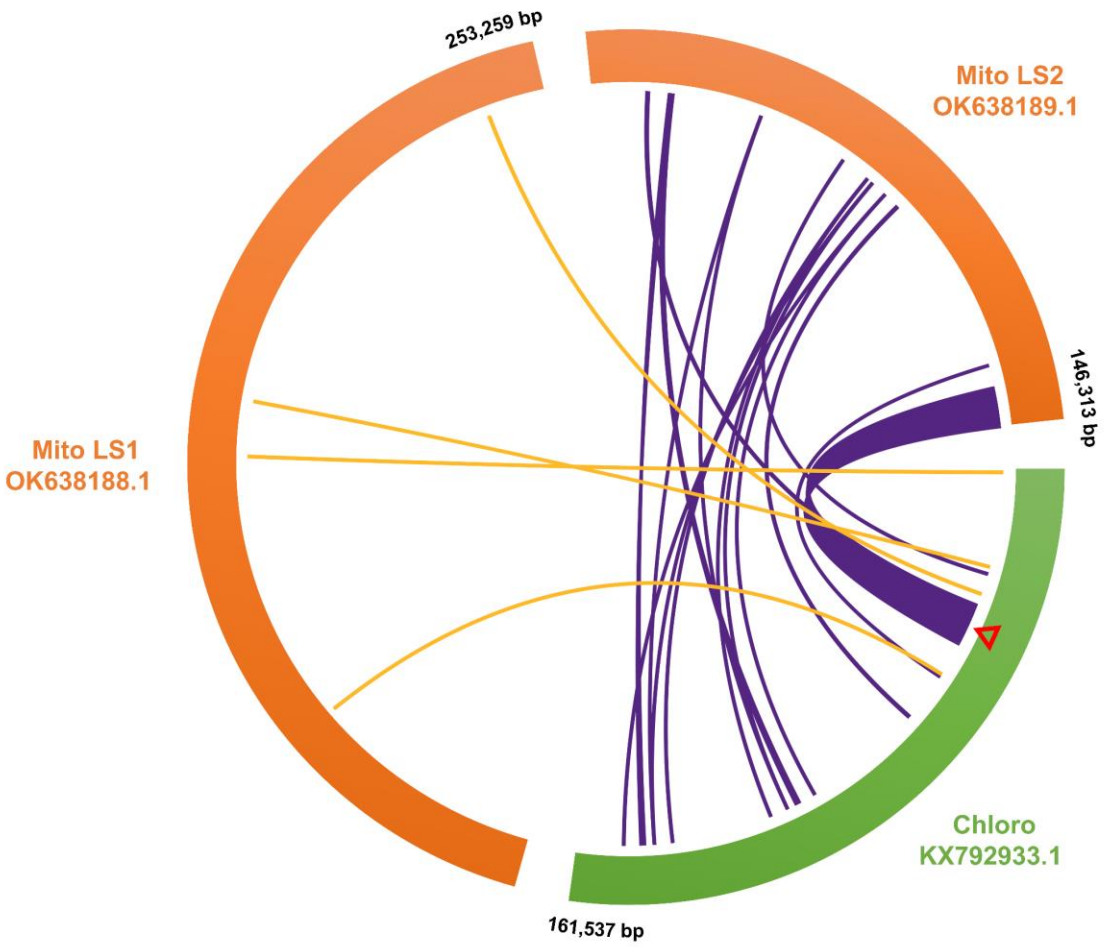

Figure 4. 14 Map of chloroplast DNA insertions in the mitochondrial genome of *T. esculentum* drawn by TBtools Advanced Circos. The reference mitogenome chromosomes LS1 (OK638188) and LS2 (OK638189) were aligned with the chloroplast genome (KX792933) of *T. esculentum* using BLAST (Figure S4.32; Table S4.2). The curves in the middle connect the homologous sequences of the plastome and mitogenome. The cpDNA insertions in the two mitochondrial chromosomes are colored orange and purple, respectively. cpDNA insertions were concentrated on the mitochondrial chromosome LS2, with only four short fragments on LS1. All the three

chromosomes are circular but are shown here as linear, with numbers next to them indicating their length and orientation. The red triangle marks the position of the 9,798 bp long homologous fragment.

It was interesting to find that the chloroplast DNA insertions concentrated in a subgenomic ring of the mitogenome (Figure 4.14). A low collinearity between the two genomes was seen, and the inserted cpDNA fragments seemed to be completely rearranged. These contained a 9,798 bp fragment, in which a large number of variations were observed. Primers were designed to amplify across the two ends of this fragment in both plastome and mitogenome to verify its presence (Figure 4.15).

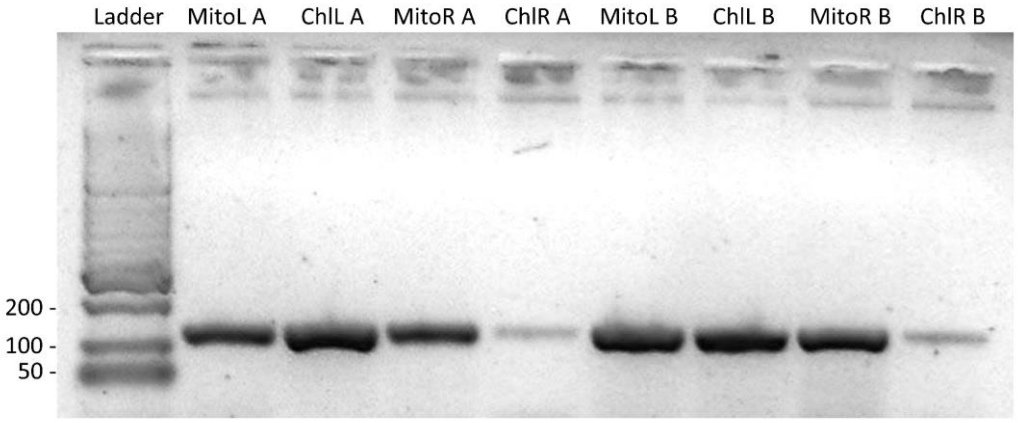

Figure 4. 15 Amplification across the two ends of the 9,798 bp homologous fragment of the mitochondrial and chloroplast genomes in two random samples A and B. Lane1, HyperLadder II; Lanes 2-5, DNA from sample A; Lanes 6-9, DNA from sample B. The products in lanes 2 and 6 were amplified with primers MitoLL and MitoLR, in lane 3 and 7 were amplified with primers ChlLL and MitoLR, in lanes 4 and 8 were amplified with primers MitoRL and MitoRR, and in lanes 5 and 9 were amplified with primers MitoRL and ChlRR. The gel was run on a 1.5% agarose gel for 40 min at 100V. The annealing

temperature of 55 °C was a bit high for the primer ChlRR, resulting in low yields and faint bands in lanes 5 and 9. Decreasing the annealing temperature by 1 ℃ and repeating the PCR resulted in normal amplification.

Table 4. 7 Variations found in the 9,798 bp homologous segment within the organelle genomes of the 84 *T. esculentum* individuals.

Position	Variation	Type 1 Chloro	Type 1 Mito	Type 2 Chloro	Type 2 Mito	Localization
35570*	SNP	Ref	**Alt**	Ref	Ref	Mito
36347*	DEL	Ref	**Alt**	Ref	Ref	Mito
36942	SNP	Ref	Ref	**Alt**	Ref	Chloro
37257	SNP	**Ref**	Alt	Alt	Alt	Chloro
37813	SNP	**Alt** (some)	Ref	Ref	Ref	Chloro
38753*	DEL	Ref	Ref	Ref	**Alt**	Mito
38975	SNP	Ref	Ref	**Alt**	Ref	Chloro
39429*	SNP	Ref	**Alt**	Ref	Ref	Mito
40061*	SNP	Ref	**Alt**	Ref	Ref	Mito
40544*	SNP	Ref	**Alt**	Ref	Ref	Mito
41243*	SNP	**Ref**	Alt	Alt	Alt	Chloro
41716*	SNP	Ref	Ref	Ref	**Alt**	Mito
41718*	SNP	Ref	Ref	Ref	**Alt**	Mito
42587*	SNP	Ref	Ref	**Alt**	Ref	Chloro
43559	SNP	Ref	Ref	Ref	**Alt**	Mito
44059	INS	**Ref**	Alt	Alt	Alt	Chloro
44235	DEL	Ref	**Alt**	Ref	Ref	Mito
44302	SNP	Ref	**Alt**	Ref	Ref	Mito
44552	SNP	Ref	Ref	Ref	**Alt**	Mito
44762	DEL	Ref	Ref	Ref	**Alt**	Mito

Refs refer to alleles identical to the type 1 chloroplast reference genome sequence. Alt refers to an allele that differs from the type 1 chloroplast reference genome sequence. Only two alleles, Ref or Alt, were seen at each locus. Comparing alleles in the same row, the locations where the mutations occurred can be estimated and written in the last column. This table does not include differences between plastome and mitogenome at

another 72 loci on this fragment (no inter-type differences were seen at these loci). Positions in the gene sequence are marked with an asterisk*.

20 variations were found in this long homologous DNA fragment, of which 13 mutations occurred in the mitogenome and 7 occurred in the chloroplast genome (Table 4.7). These do not include the differences between the two organelle genomes at another 72 loci in this segment (this included 22 deletions, 6 insertions, and 44 SNPs.), which are the same for both germplasms, making it difficult to tell where the mutation occurred. 10 variations were found in the gene sequence on this segment, of which only 2 synonymous substitutions occurred in the chloroplast genome, and the remaining 8 were in the mitogenome, which had a great impact on transcription, including the introduction of early stop codons (Table S4.3).

The mitogenome of *T. esculentum* was 399,572 bp (type 1) in length, and a total of 254 variations were found in the mitogenomes of the 84 individuals. The length of the chloroplast genome was 161,537 bp (type 1), and a total of 147 variations were found in the plastomes of these plants. The value of variation per nucleotide of the chloroplast genome is even higher than that of the mitogenome, so it is speculated that these chloroplast genes *psbC*, *rps14*, *psaB*, and *psaA* are protected by some mechanism in the chloroplast genome, but after being transferred to the mitogenome, the protection mechanism disappears and these genes begin to accumulate mutations that render them nonfunctional and pseudogenes.

4.4 Conclusion

The comparative analysis of the organelle genomes of 84 *T. esculentum* individuals revealed two germplasms with distinct mitochondrial and chloroplast genomes. The type 1 mitogenome contains two autonomous rings, or five smaller subgenomic circular molecules. These two equimolar structures are thought to be interchangeable through recombination on three pairs of long direct repeats (Li and Cullis, 2021). The type 2 mitogenome contains three circular molecules and one linear chromosome. It also has a unique fragment of 2,108 bp in length, likely derived from the mitogenome of *Lupinus*. Primers were designed to amplify on this fragment for germplasm typing. Both ends of the linear chromosome are repetitive sequences, also present in other molecules, on which recombination may occur to protect the linear chromosome from degradation.

The structural variation resulted in increased copy number of the genes *atp8, nad5* (exon3 and exon4), *nad9*, *rrnS, rrn5*, *trnC*, and *trnfM* in the type 2 mitogenome, but it is unknown whether this is reflected in the gene expression level. The genes *nad1*, *nad4L*, *rps10* and *mttB* were found to use alternative start codons ACG and ATT, similar to the mitogenome of common bean (Bi et al., 2020). A total of 254 differential loci were found in the mitogenomes of the 84 *T. esculentum* individuals. Type 1 and type 2 mitogenomes differ from each other at 230 of these loci. Only one of these 254 variations was found in the mitochondrial gene coding sequence, which altered the amino acid sequence synthesized by *matR*.

The evolutionary study of the differential loci in the mitogenomes of *T. esculentum* found that the two types of plants fell into two clusters, as expected. The type

1 plants can be further divided into two clades, and the divergence is thought to be possibly related to soil moisture content, although this still needs to be verified by studies with larger sample sizes. Three consecutive substitutions at positions 127,684 to 127,686 on chromosome LS1 (OK638188) may have played an important role here. The phylogenetic tree constructed based on the mitochondrial genes of *T. esculentum* and other related Fabaceae species was consistent with the previously published one built on chloroplast genes. *C. canadensis* and *T. esculentum* were found to be closely related, and they have more complete sets of mitochondrial genes than the other legumes.

Heteroplasmy in the mitochondrial genome of *T. esculentum* is not as prevalent as in its chloroplast genome, but higher levels do exist at certain loci, such as loci 127,684 to 127,686 on chromosome LS1. Among all samples, only one individual A11 from Aminuis had a generally high degree of heteroplasmy at most differential loci, consistent with what was seen in its chloroplast genome. Whether it is because of some of its own characteristics that A11 escaped the genetic bottleneck at the developmental stage is still unclear.

Chloroplast insertions are thought to be concentrated in one of the subgenomic rings of the mitogenome of *T. esculentum*. The mitogenomes of all marama individuals were found to contain a long chloroplast DNA insertion over 9 kb in length. The study of the polymorphisms on this segment infers that the sequence is protected by some mechanism in the chloroplast genome, so only a small amount of synonymous substitutions is retained, but after being inserted into the mitogenome, a large number of mutations have accumulated on it, making the genes on it lose their function.

4.5 Supplementary Materials

Table S4.1 Summary of WGS Illumina data and sources of the 84 samples

Sample	Plant source	Raw reads	Raw data	Mito Cvrg	Cytotype
M17	NM Farm Seeds	96750044	14512506600	2480	Type 2
S_35	NM Farm Seeds	86626894	12994034100	1928	Type 2
S_19	NM Farm Seeds	55375806	8306370900	1493	Type 2
S_4	NM Farm Seeds	73435564	11015334600	2817	Type 2
S_13	NM Farm Seeds	217730028	32659504200	5368	Type 2
S_27	NM Farm Seeds	107264134	16089620100	349	Type 2
S_20	NM Farm Seeds	55371622	8305743300	1455	Type 2
S_30	NM Farm Seeds	106501502	15975225300	2963	Type 2
S_33	NM Farm Seeds	130811500	19621725000	4788	Type 2
M7	NM Farm Seeds	172383630	25857544500	4279	Type 2
M8	NM Farm Seeds	123934796	18590219400	3713	Type 2
M1 #	NM Farm Seeds	170816890	25622533500	5240	Type 2
M2	NM Farm Seeds	194335620	29150343000	933	Type 2
M11	NM Farm Seeds	125005082	18750762300	4213	Type 2
M12	NM Farm Seeds	128360580	19254087000	1500	Type 2
M15	NM Farm Seeds	134439570	20165935500	3781	Type 2
M16	NM Farm Seeds	132448260	19867239000	3789	Type 2
M23	NM Farm Seeds	134475978	20171396700	3642	Type 2
M22	NM Farm Seeds	111447592	16717138800	3891	Type 2
M24	NM Farm Seeds	127392478	19108871700	3546	Type 2
M26	NM Farm Seeds	225925198	33888779700	699	Type 2
M28	NM Farm Seeds	193804308	29070646200	4799	Type 2
M25	NM Farm Seeds	145937916	21890687400	3860	Type 2
N29	NM Farm Seeds	147710832	22156624800	4863	Type 2
M31	NM Farm Seeds	177899436	26684915400	3647	Type 2
M34	NM Farm Seeds	124221640	18633246000	1060	Type 2
M36	NM Farm Seeds	125524008	18828601200	2242	Type 2
M37	NM Farm Seeds	185151336	27772700400	6847	Type 2
M38	NM Farm Seeds	168052326	25207848900	4617	Type 2
M40	Namibia Unknown	135702768	20355415200	1720	Type 1
Index1 #	UP Farm	41499124	4149912400	269	Type 2
Index10	UP Farm Seeds	36382172	3638217200	209	Type 2
Index11	UP Farm Seeds	34631932	3463193200	295	Type 2
Index12 #	UP Farm	39339004	3933900400	804	Type 2
Index19 #	UP Farm	35994474	3599447400	258	Type 2
Index3 #	UP Farm	35991990	3599199000	395	Type 2

Sample	Plant source	Raw reads	Raw data	Mito Cvrg	Cytotype
Index5 #	UP Farm	34349602	3434960200	814	Type 2
Index8	Namibia Unknown	42747718	4274771800	359	Type 1
Index9 #	UP Farm	34312880	3431288000	793	Type 2
R1R2	Unknown	358941018	35894101800	1550	Type 2
A1 #	Aminuis Seeds	93324222	13998633300	707	Type 1
A2	Aminuis Seeds	93445492	14016823800	437	Type 1
A3	Aminuis Seeds	84081976	12612296400	855	Type 1
A4	Aminuis Seeds	60965538	9144830700	607	Type 1
A5	Aminuis Seeds	91044746	13656711900	587	Type 1
A6	Aminuis Seeds	89785050	13467757500	677	Type 1
A7	Aminuis Seeds	86000118	12900017700	528	Type 1
A8	Aminuis Seeds	57087632	8563144800	549	Type 1
A9 #	Aminuis	62849376	9427406400	559	Type 1
A10 #	Aminuis	75633554	11345033100	541	Type 1
A11 #	Aminuis	92851730	13927759500	923	Type 1
A12 #	Aminuis	84374620	12656193000	561	Type 1
A13 #	Aminuis	83450100	12517515000	864	Type 1
nar15 #	Aminuis	81618952	12242842800	389	Type 1
nar16	UP Farm Seeds	96595344	14489301600	807	Type 2
S1 #	Tsjaka	10358444	1035844400	25	Type 1
S2 #	Tsjaka	20343934	2034393400	198	Type 1
S3 #	Tsjaka	33100100	3310010000	597	Type 1
S4 #	Tsjaka	25428338	2542833800	373	Type 1
S5 #	Okamatapati	27183834	2718383400	396	Type 1
S6 #	Tsumkwe	24185808	2418580800	419	Type 1
S7 #	Tsumkwe	22075112	2207511200	162	Type 1
S8 #	Tsumkwe	16337544	1633754400	163	Type 1
S9 #	Aminuis	25274640	2527464000	270	Type 1
S10 #	Aminuis	28329944	2832994400	311	Type 1
S11 #	Aminuis	26741342	2674134200	288	Type 1
S12 #	Aminuis	29520092	2952009200	423	Type 1
S13 #	Aminuis	NA	NA	236	Type 1
S14 #	Aminuis	22928100	2292810000	283	Type 1
S15 #	Aminuis	22856344	2285634400	486	Type 1
S16 #	Aminuis	25333752	2533375200	338	Type 1
S17 #	Aminuis	23418786	2341878600	258	Type 1
S18 #	Tsumkwe	25764102	2576410200	234	Type 1
S19 #	Aminuis	24407080	2440708000	421	Type 1
S20 #	Osire	25692398	2569239800	407	Type 1
S21 #	Osire	25940358	2594035800	381	Type 1

Sample	Plant source	Raw reads	Raw data	Mito Cvrg	Cytotype
S22 #	Osire	32107310	3210731000	609	Type 1
S23 #	Osire	35844166	3584416600	697	Type 1
S24 #	Tsumkwe	31973008	3197300800	411	Type 1
S25 #	Ombujondjou	24401422	2440142200	539	Type 1
S26 #	Ombujondjou	32758062	3275806200	433	Type 1
S27 #	Epukiro	28122544	2812254400	573	Type 1
S28 #	Epukiro	31273940	3127394000	699	Type 1
S29 #	Otjiwarongo	25274640	2527464000	498	Type 1

43 independent samples for sequence diversity and phylogenetic analysis; Mito Cvrg = approximate read depth for single-copy regions of the mitogenome

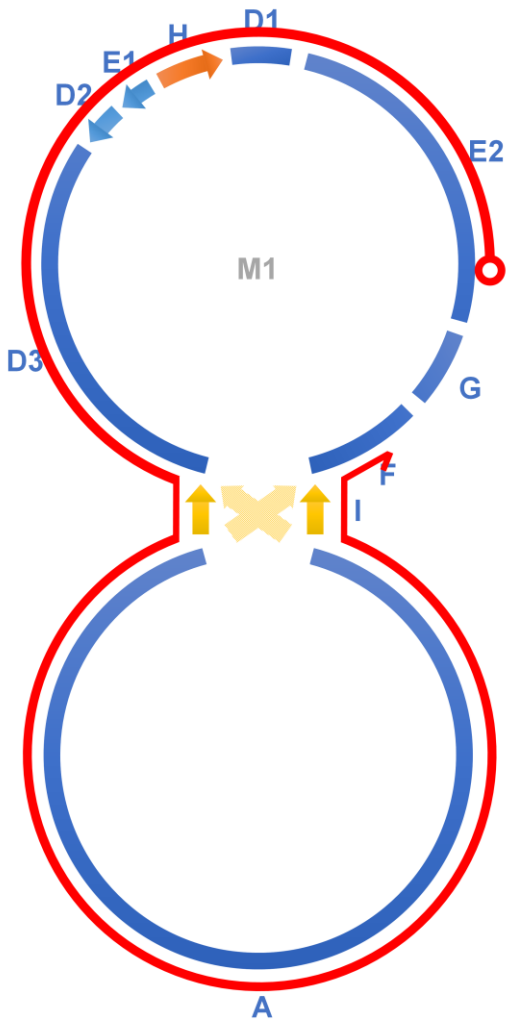

Figure S4.1 Alignment of contig tig00000040 of 159,605 bp, assembled by HiCanu on the Sample 4 PacBio HiFi reads, to the type 2 marama subgenomic chromosome M1. The Canu input genome size was set to 2M to better capture complete organelle genome sequences. The red curve indicates where the contig is aligned, starting from the red circle and ending with the red arrow.

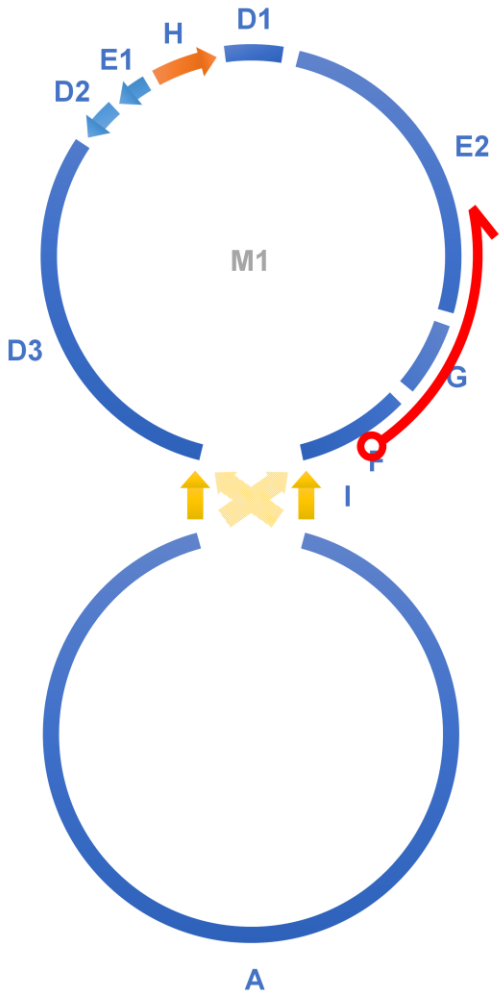

Figure S4.2 Alignment of contig tig00000050 of 13,828 bp, assembled by HiCanu on the Sample 4 PacBio HiFi reads, to the type 2 marama subgenomic chromosome M1. The Canu input genome size was set to 2M to better capture complete organelle genome

sequences. The red curve indicates where the contig is aligned, starting from the red circle and ending with the red arrow. This together with tig00000040 (Figure S4.1) verifies the structure of M1. Discontinuous assembly results from fluctuating in coverage at some locations, and the complete structure has been verified by manual extension.

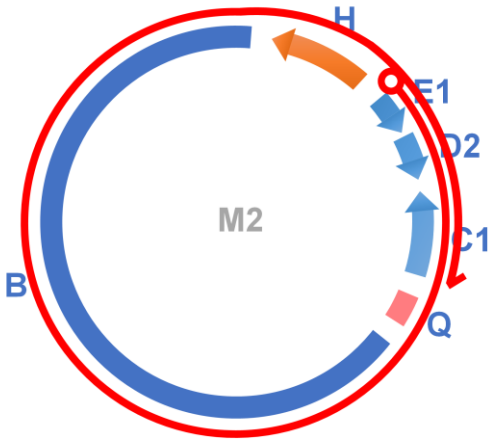

Figure S4.3 Alignment of contig tig00000055 of 64,325 bp, assembled by HiCanu on the Sample 4 PacBio HiFi reads, to the type 2 marama subgenomic chromosome M2. The Canu input genome size was set to 2M to better capture complete organelle genome sequences. The red curve indicates where the contig is aligned, starting from the red circle and ending with the red arrow. This contig validates the circular structure of M2.

Figure S4.4 Alignment of contig tig00000057 of 87,648 bp, assembled by HiCanu on the Sample 4 PacBio HiFi reads, to the type 2 marama subgenomic chromosome M3. The Canu input genome size was set to 2M to better capture complete organelle genome sequences. The red line indicates where the contig is aligned, starting from the red circle and ending with the red arrow.

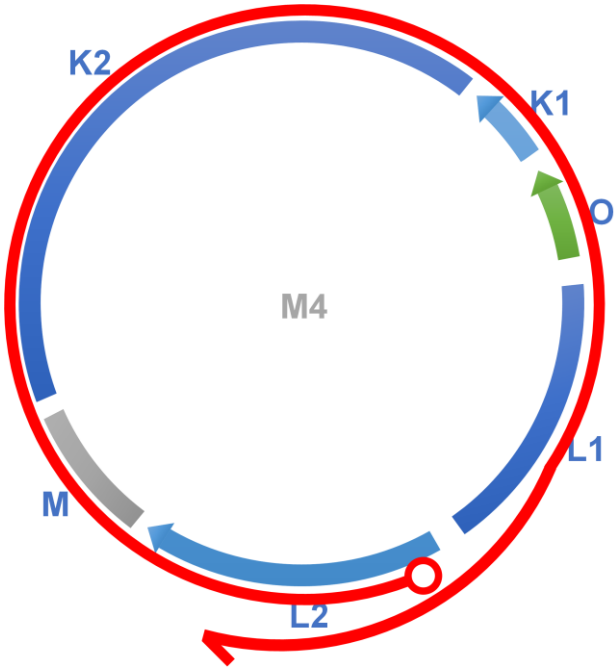

Figure S4.5 Alignment of contig tig00000058 of 123,793 bp, assembled by HiCanu on the Sample 4 PacBio HiFi reads, to the type 2 marama subgenomic chromosome M4. The Canu input genome size was set to 2M to better capture complete organelle genome sequences. The red curve indicates where the contig is aligned, starting from the red circle and ending with the red arrow. This verifies the circular structure of M4.

Two different connections were seen at the pair of inverted repeats I, which formed a normal circular molecule, or an 8-shaped ring resulting from recombination on the repeats (A^{1}-I-D3 and $A^{82,874}$-I-F became $A^{82,874}$-I-D3 and A^{1}-I-F). Both molecules were confirmed by the PacBio HiFi reads of Sample 4 and found to be in very close proportions, as shown in Figure S4.6-4.9.

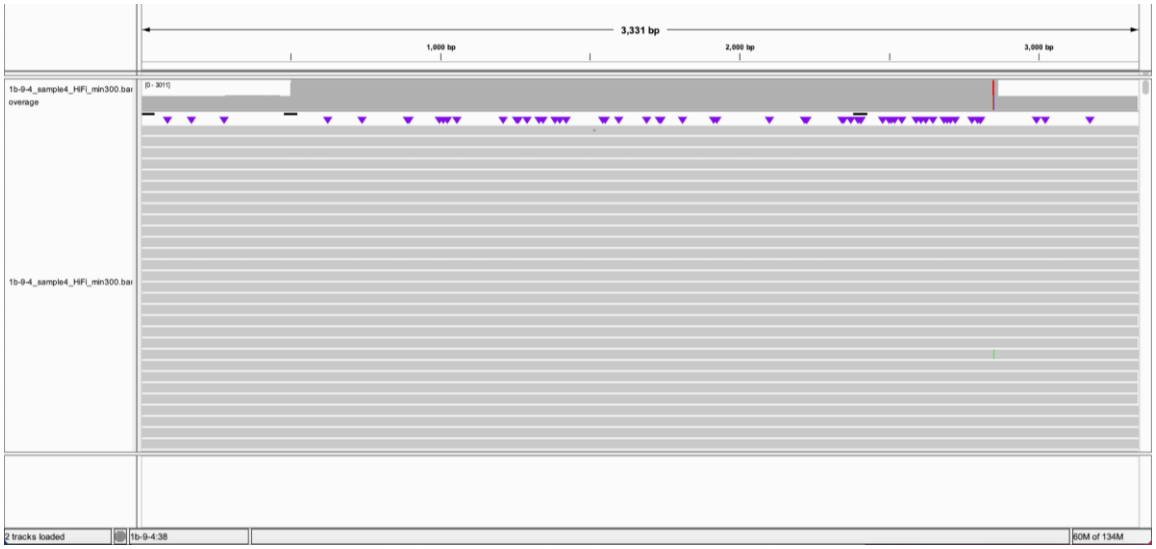

Figure S4.6 Alignment of Sample 4 PacBio HiFi reads to the artificial chromosome concatenated by node A (82,375 - 82,874 bp), node I, and node D3 (26,671-26,172 bp) using pbmm2. The minimum length was set to 300 to avoid interference from homologous DNA fragments. The coverage of the inverted repeat node I was doubled as expected.

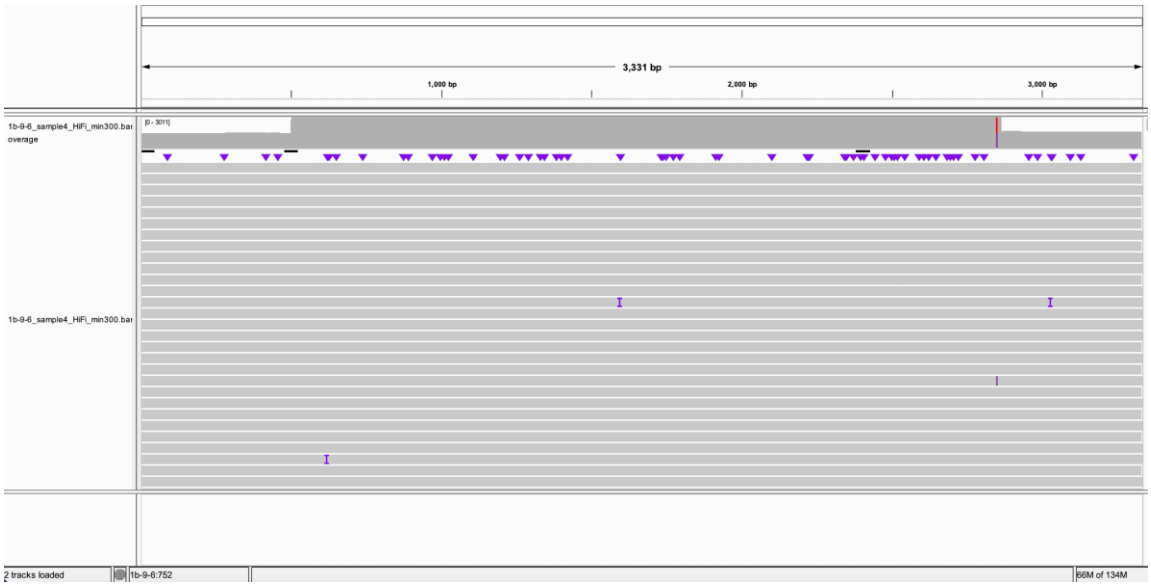

Figure S4.7 Alignment of Sample 4 PacBio HiFi reads to the artificial chromosome concatenated by node A (82,375 - 82,874 bp), node I, and node F (8,590 – 8,091 bp) using pbmm2. The minimum length was set to 300 to avoid interference from homologous DNA fragments. The coverage of the inverted repeat node I was doubled as expected.

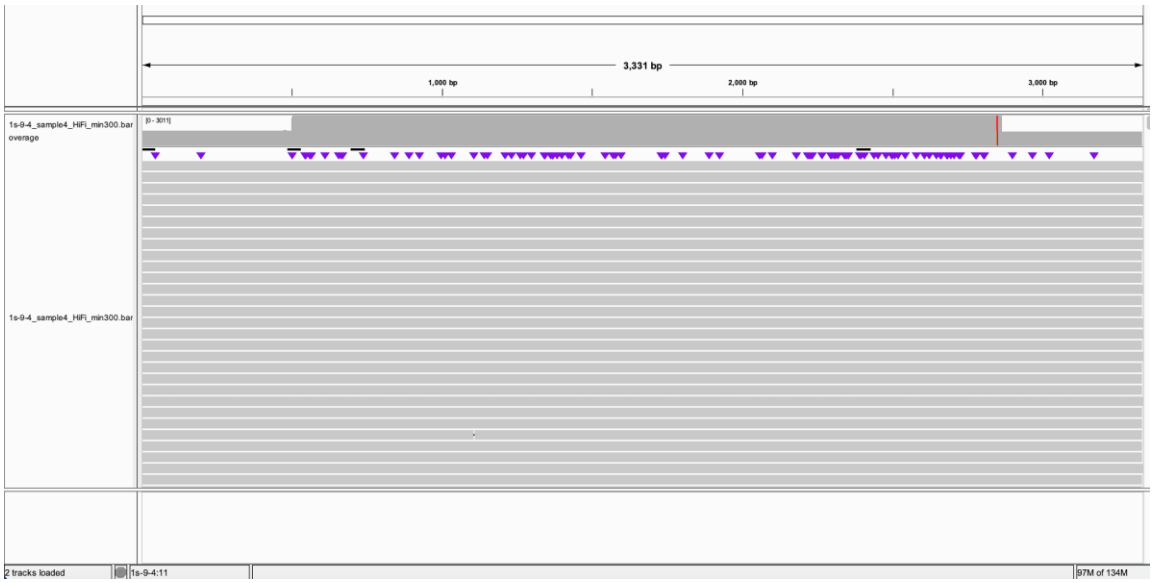

Figure S4.8 Alignment of Sample 4 PacBio HiFi reads to the artificial chromosome concatenated by node A (500 - 1 bp), node I, and node D3 (26,671-26,172 bp) using pbmm2. The minimum length was set to 300 to avoid interference from homologous DNA fragments. The coverage of the inverted repeat node I was doubled as expected.

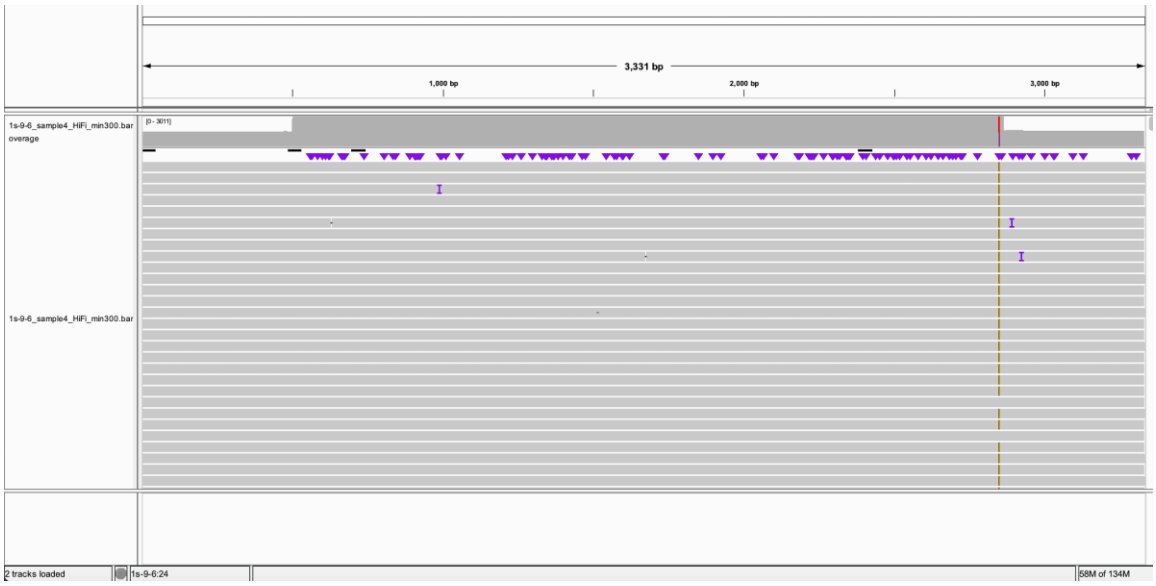

Figure S4.9 Alignment of Sample 4 PacBio HiFi reads to the artificial chromosome concatenated by node A (500 - 1 bp), node I, and node F (8,590 – 8,091 bp) using pbmm2. The minimum length was set to 300 to avoid interference from homologous DNA fragments. The coverage of the inverted repeat node I was doubled as expected.

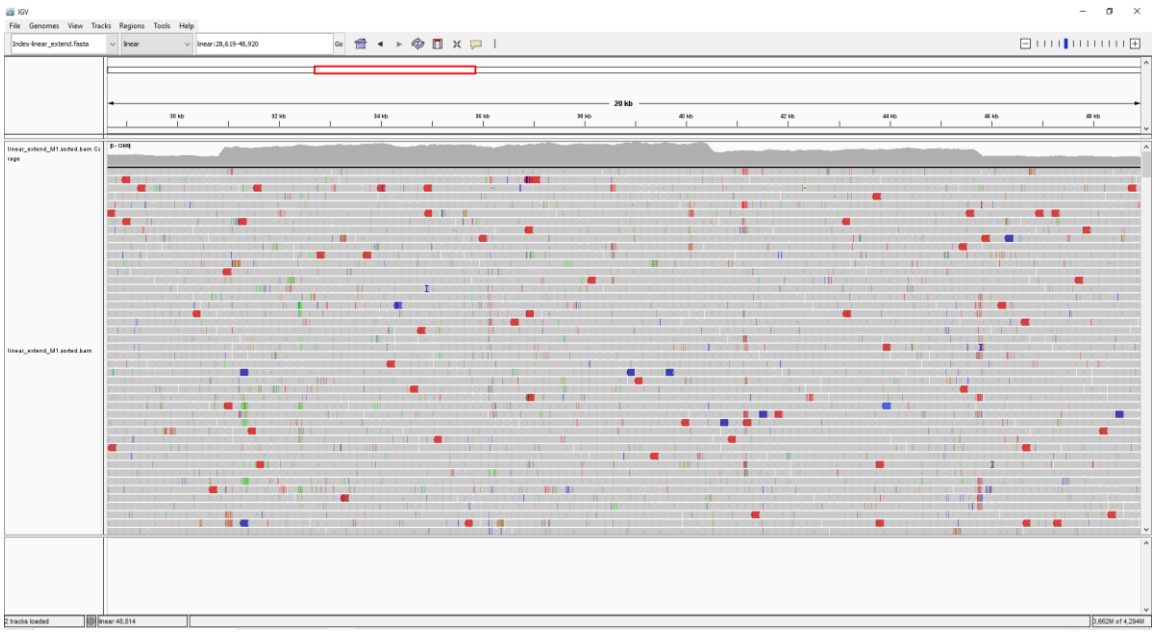

Figure S4.10 IGV visualization of WGS Illumina reads of individual M1 aligned to the one end of the linear chromosome M3 (extended by the sequence of D1, thus arranged as D1-H-E1-D2-C1-C2). A clear increase of coverage from one copy to two, and then to three can be seen from right to left. The two ends, C2 and D1 have only one copy. C1 has two copies, as it is owned by both chromosomes M2 and M3. H-E1-D2 is a long repeat present in three chromosomes, M1, M2, and M3, with tripled depth as shown.

Figure S4.11 IGV visualization of WGS Illumina reads of individual M1 aligned to the mitochondrial genome fragment C2-K1-O-N via Bowtie2. The coverage of K1 and O in the middle is doubled because it is a long repeat owned by two chromosomes, M3 and M4.

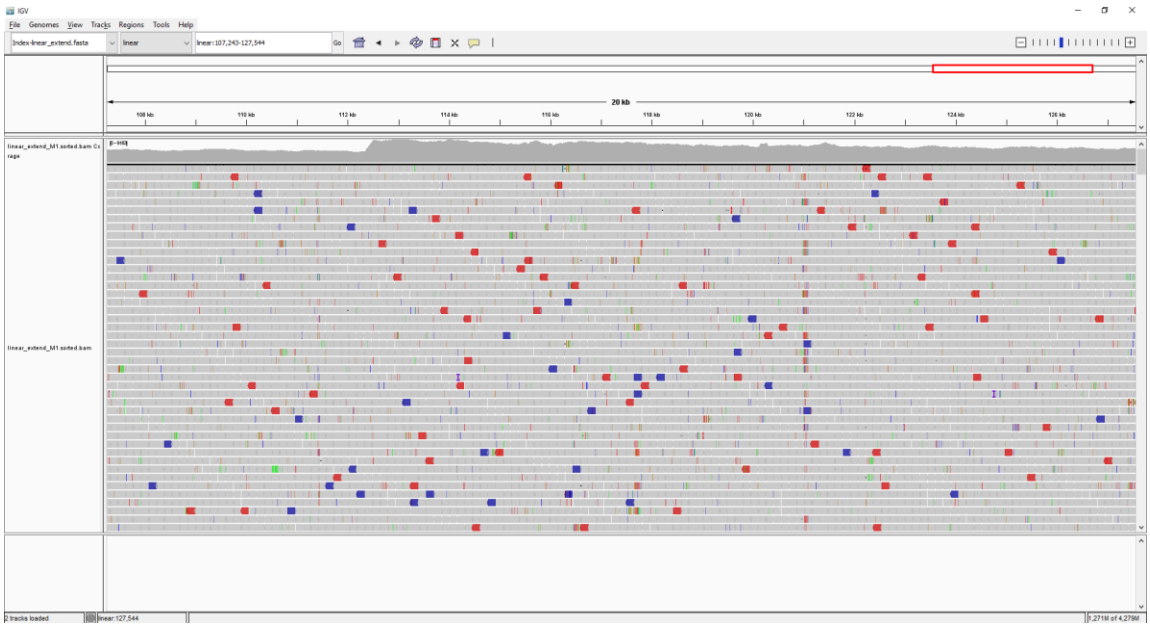

Figure S4.12 IGV visualization of WGS Illumina reads of marama individual M1 aligned to the mitochondrial genome fragment L1-L2 by Bowtie2. From L2 onwards, the read depth is doubled because L2 is a segment owned by both chromosomes M3 and M4. However, the coverage gradually decreases from left to right after doubling, as the end of the linear chromosome M3 is reached.

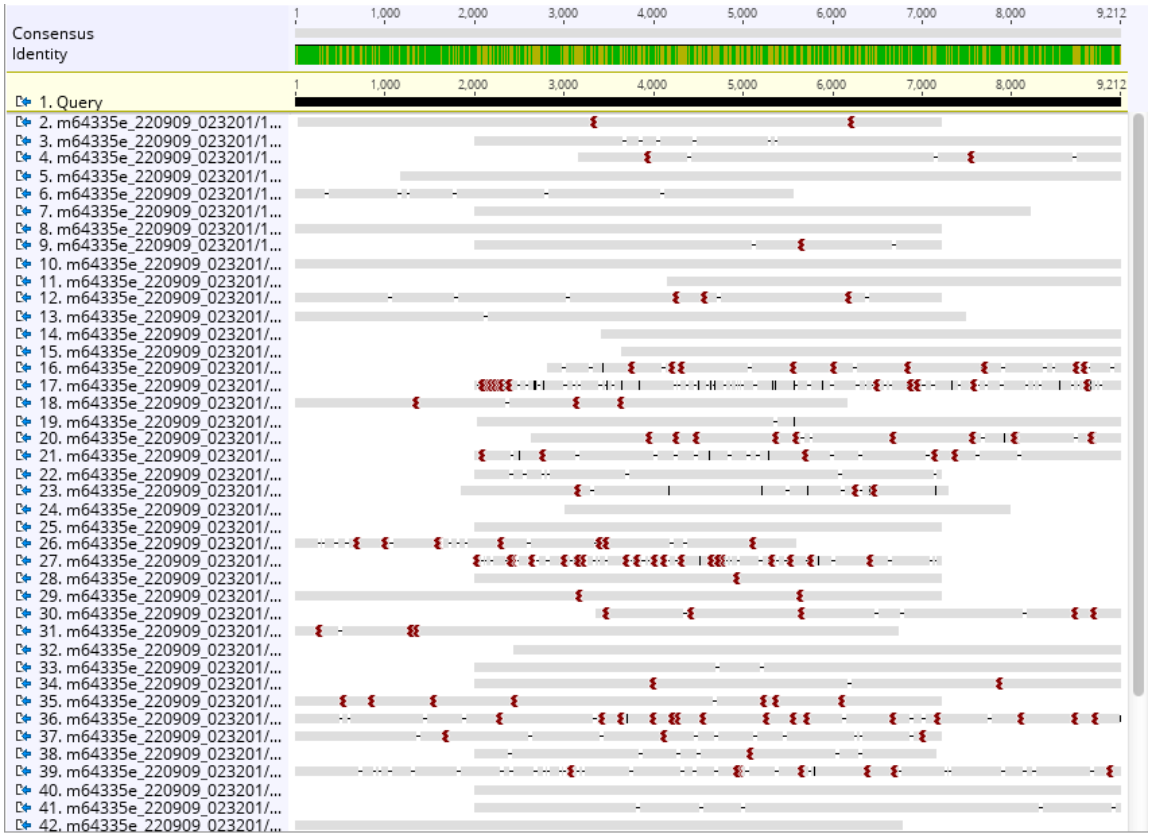

Figure S4.13 The PacBio HiFi reads of the type 1 individual Sample 32 were mapped to the fragment B-H-B (2000 bp sequences were taken from both ends of node B, with node H in the middle) by BLAST in Geneious 9. Reads going across the entire fragment B-H-B can be seen, indicating the presence of a closed ring consisting only of nodes B and H in type 1 individual Sample 32. In addition, reads from chromosome M1, mapped only to node H, were also observed. Reads, mapped to node H and one end of node B (the other end entered chromosome M1, not shown here), were seen, representing a combination of the two subgenomic structures.

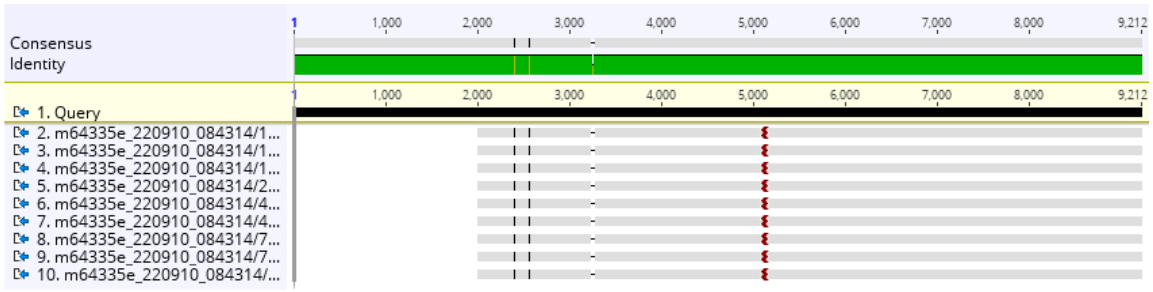

Figure S4.14 The PacBio HiFi reads of the type 2 individual Sample 4 were mapped to the fragment B-H-B (2000 bp sequences were taken from both ends of node B, with node H in the middle) by BLAST in Geneious 9. Node H was only connected to one end of node B but not to the other. The closed ring consisting only of node B and H does not exist in type 2 individual Sample 4.

Figure S4.15 The PacBio HiFi reads of the type 1 individual Sample 32 were mapped to the fragment C-A-C (node A, 39,444-43,351 bp, surrounded by 2000 bp sequences at both ends of node C) by BLAST in Geneious 9. Reads spanning the entire fragment C-A-C can be seen, indicating the existence of a closed subgenomic ring consisting only of nodes C and A in type 1 individual Sample 32. Furthermore, reads from chromosome M1 were observed to map only to node A in the middle, and also reads were seen to map to node A and one end of node C (the other end entered chromosome M1, not shown here), representing a combination of the two subgenomic structures.

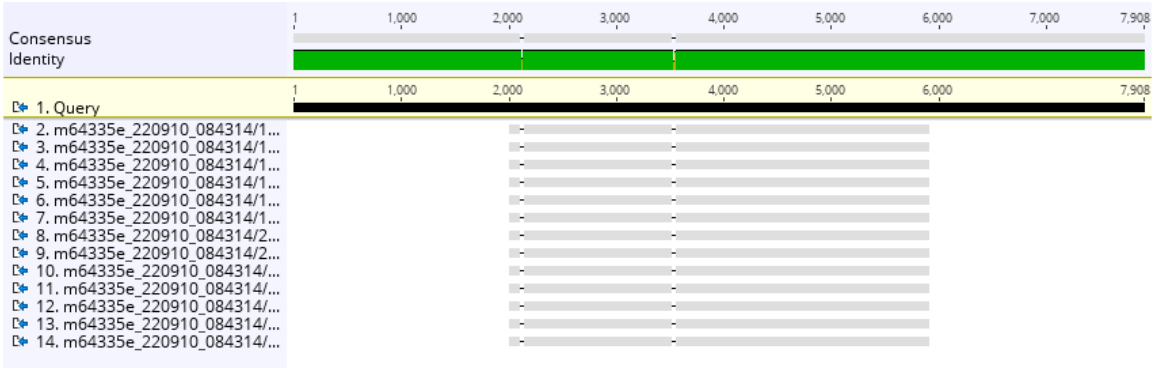

Figure S4.16 The PacBio HiFi reads of the type 2 individual Sample 4 were mapped to the fragment C-A-C (node A, 39,444-43,351 bp, surrounded by 2000 bp sequences at both ends of node C) by BLAST in Geneious 9. Node A (39,444-43,351 bp) was not connected to either end of node C in Sample 4, suggesting that the two types of mitochondrial genomes have distinct structures in this region.

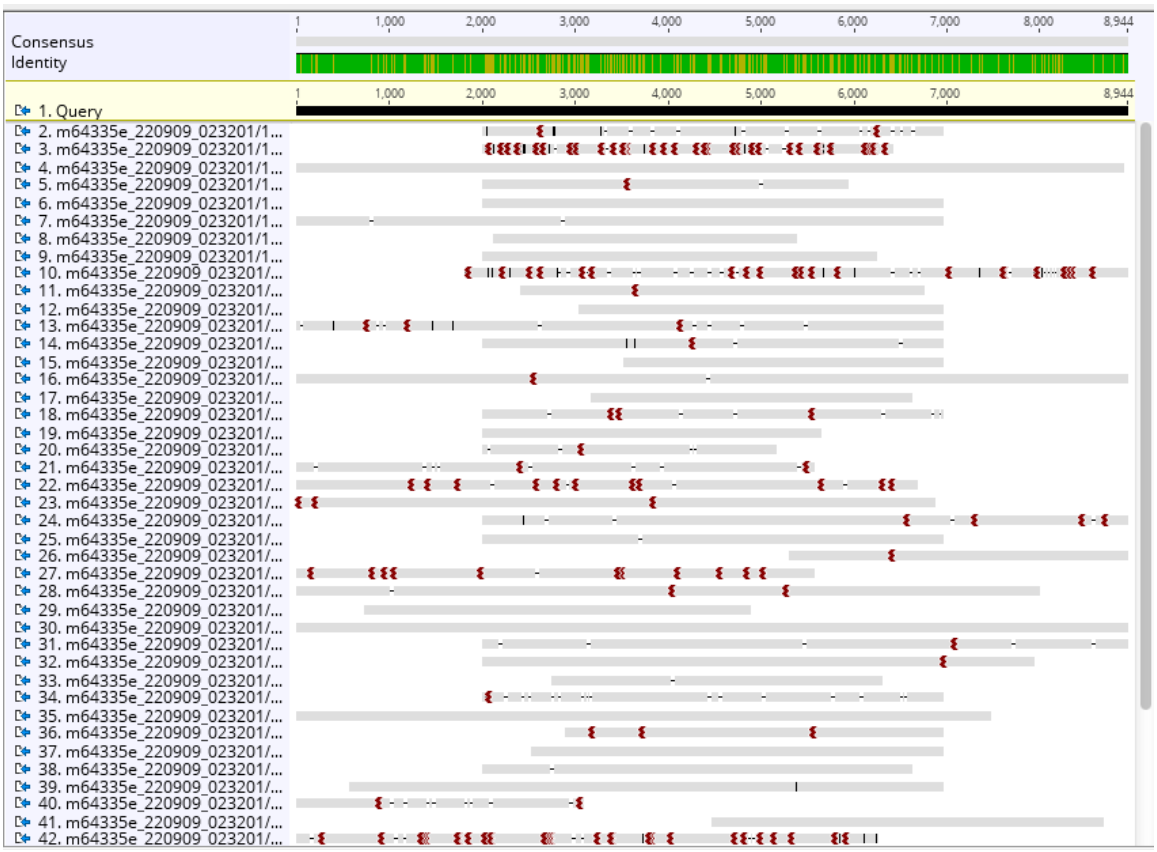

Figure S4.17 The PacBio HiFi reads of the type 1 individual Sample 32 were mapped to the fragment N-O-N (2000 bp sequences were taken from both ends of node N, with node O in the middle) by BLAST in Geneious 9. Reads spanning the entire fragment N-O-N can be seen, indicating the existence of a closed subgenomic ring consisting only of nodes N and O in type 1 individual Sample 32. Furthermore, reads from chromosome M1 were observed to map only to node O in the middle, and also reads were seen to map to node O and one end of node N (the other end entered chromosome M1, not shown here), representing a combination of the two subgenomic structures.

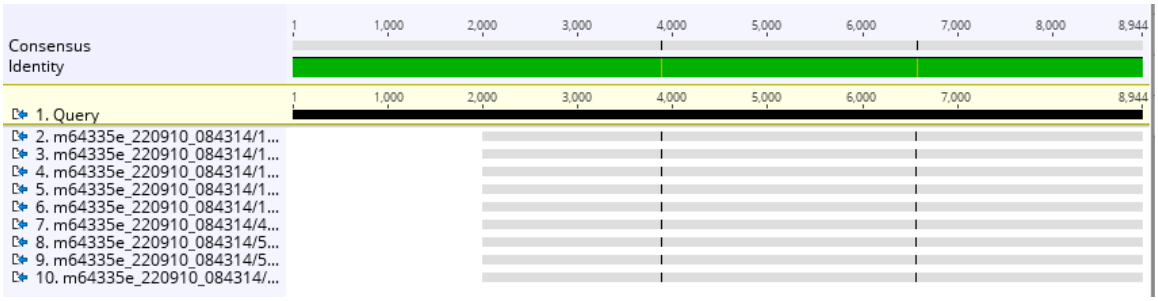

Figure S4.18 The PacBio HiFi reads of the type 2 individual Sample 4 were mapped to the fragment N-O-N (2000 bp sequences were taken from both ends of node N, with node O in the middle) by BLAST in Geneious 9. Node O was only connected to one end of node N but not to the other. The closed ring consisting only of node N and O does not exist in type 2 individual Sample 4.

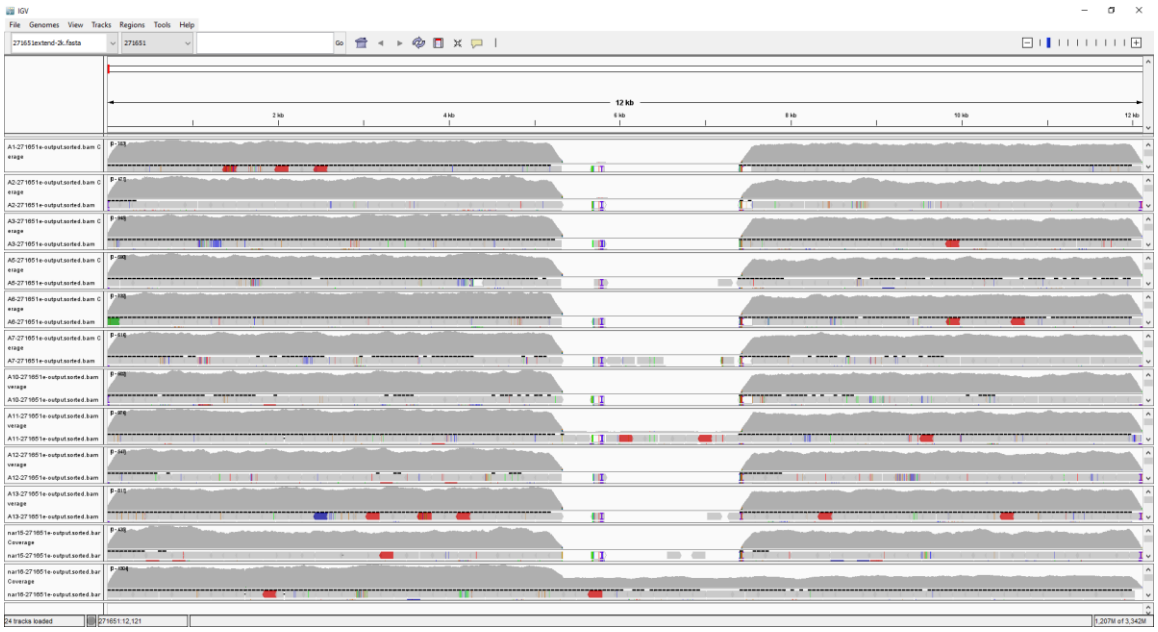

Figure S4.19 Bowtie 2 alignment of WGS Illumina reads from 12 marama samples to the 12,113 bp fragment from the type 2 mitochondrial genome chromosome M2, visualized in IGV. The reference sequence starts at C1 and ends at B (Figure 4.1), with the 2,108 bp type 2 exclusive sequence in between, from 5,325 bp to 7,432 bp. The 10 A plants and

nar15, all originally from Aminuis, do not contain this 2,108 bp fragment, indicating that they have a type 1 mitochondrial genome. Furthermore, in these plants, the coverage of C1 is close to that of B, as expected for the type 1 structure. nar16 is a descendent of plants grown on Pretoria Farm and has a type 2 mitochondrial genome. Not only does it contain this 2,108 bp fragment, but it also has a depth-doubled C1 because that's a repeat sequence that both M2 and M3 have, as shown in Figure 4.1.

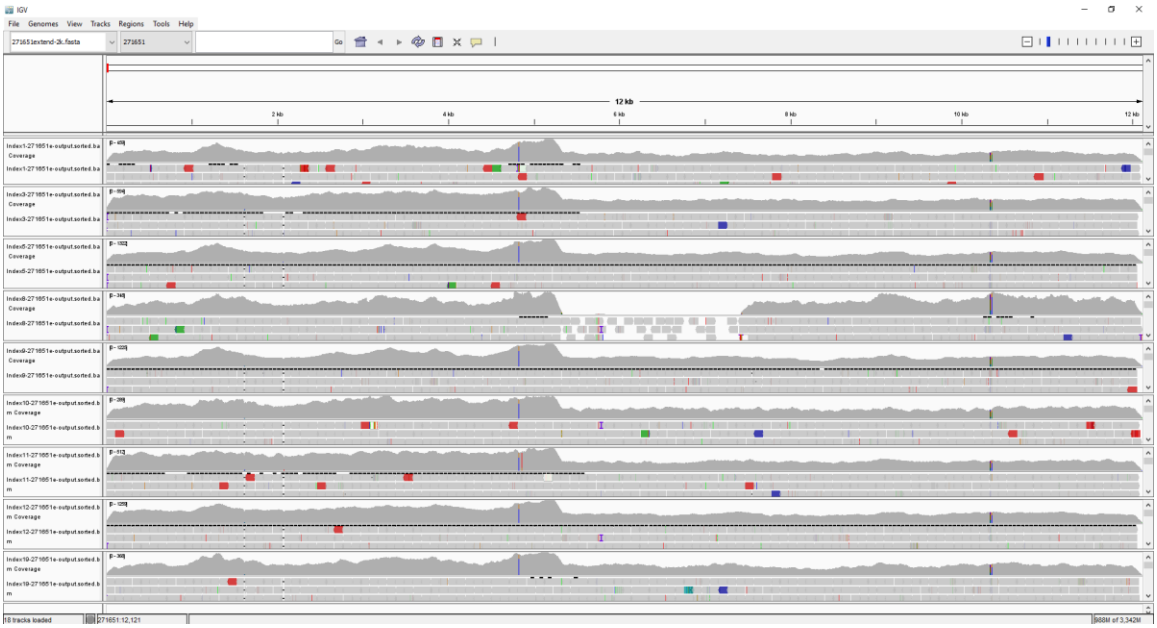

Figure S4.20 Bowtie 2 alignment of WGS Illumina reads from 9 Index plants to the 12,113 bp fragment from the type 2 mitochondrial genome chromosome M2, visualized in IGV. The reference sequence starts at C1 and ends at B (Figure 4.1), with the 2,108 bp type 2 specific sequence in between, from 5,325 bp to 7,432 bp. Among these plants, Index8, as the only one not originating from the Pretoria Farm, does not contain this 2,108 bp fragment, indicating that it has a type 1 mitochondrial genome. All remaining Index plants were either collected from the Pretoria Farm or were grown from seeds collected there. They all have type 2 mitochondrial genomes and they all contain this

2,108 bp fragment. Besides, the depth of C1 is doubled in all these plants except Index8 because C1 is a repeat sequence that both M2 and M3 have in the type 2 mitogenome, as shown in Figure 4.1.

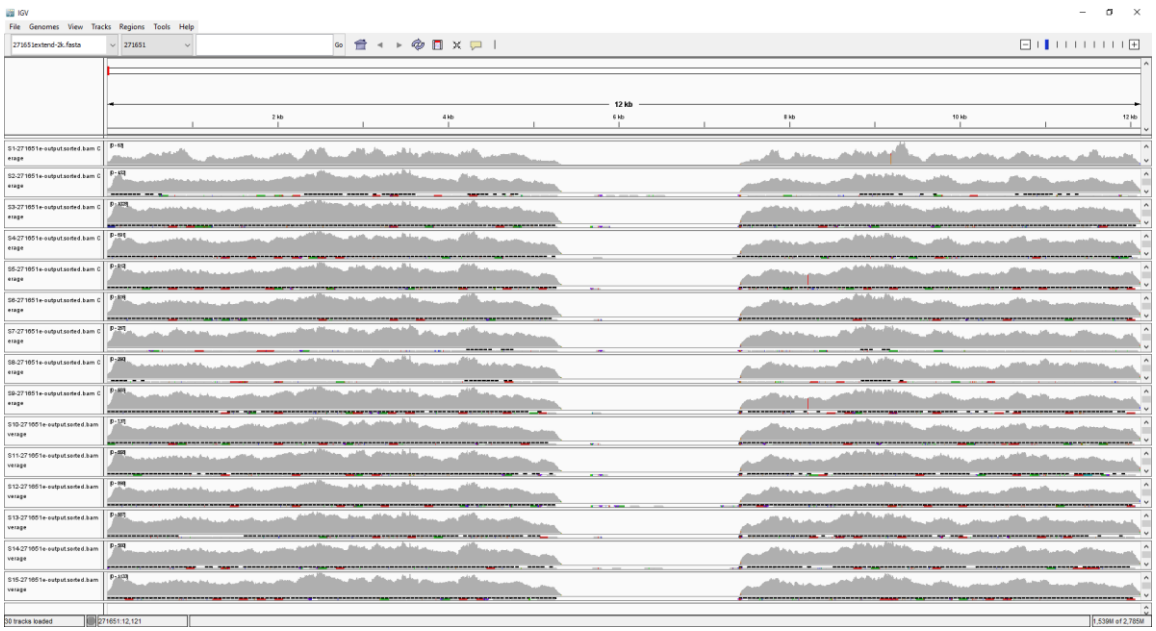

Figure S4.21 Bowtie 2 alignment of WGS Illumina reads from 15 S plants (S1-S15) to the 12,113 bp fragment from the type 2 mitochondrial genome chromosome M2, visualized in IGV. The reference sequence starts at C1 and ends at B (Figure 4.1), with the 2,108 bp type 2 specific sequence in between, from 5,325 bp to 7,432 bp. S plants are wild plants collected from 8 different geographical locations in Namibia. None of them contain this 2,108 bp fragment, indicating that they all have the type 1 mitochondrial genome.

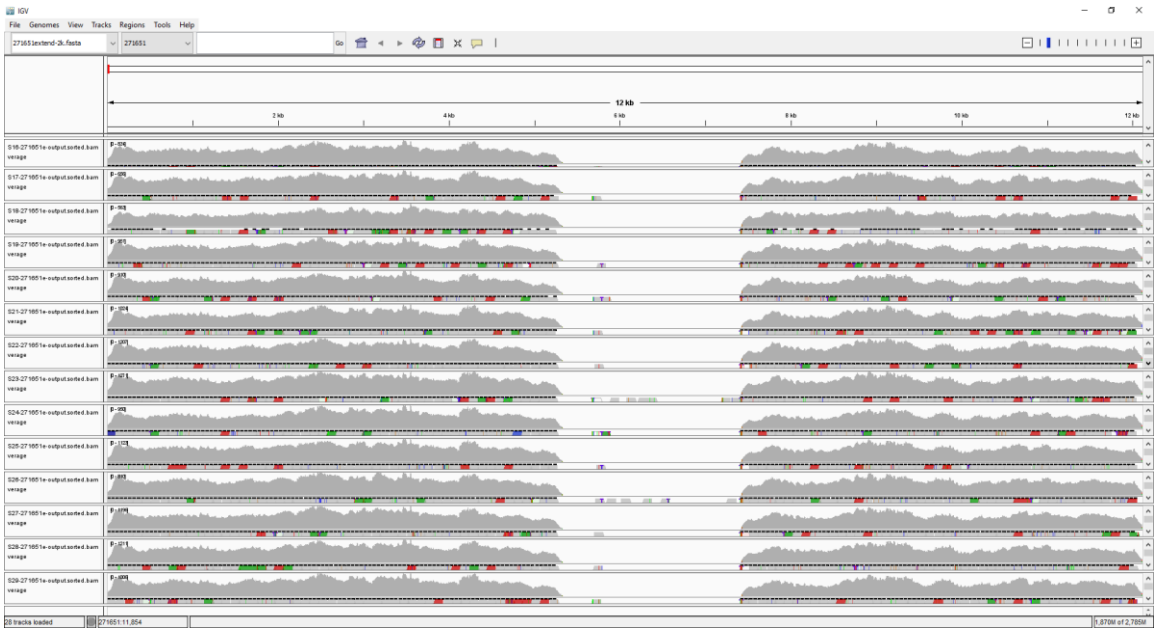

Figure S4.22 Bowtie 2 alignment of WGS Illumina reads from 14 S plants (S16-S29) to the 12,113 bp fragment from the type 2 mitochondrial genome chromosome M2, visualized in IGV. The reference sequence starts at C1 and ends at B (Figure 4.1), with the 2,108 bp type 2 specific sequence in between, from 5,325 bp to 7,432 bp. S plants are wild plants collected from 8 different geographical locations in Namibia. None of them contain this 2,108 bp fragment, indicating that they all have the type 1 mitochondrial genome.

Figure S4.23 Bowtie 2 alignment of WGS Illumina reads from 5 M plants (M1, M2, M7, M8, and M11) to the 12,113 bp fragment from the type 2 mitochondrial genome chromosome M2, visualized in IGV. The reference sequence starts at C1 and ends at B (Figure 4.1), with the 2,108 bp type 2 specific sequence in between, from 5,325 bp to 7,432 bp. M plants (except M40) were grown from seeds collected from Namibia Farm. They all contain this 2,108 bp fragment, indicating that they all have the type 2 mitochondrial genome. Furthermore, the coverage of the C1 region is doubled in these plants because it is a repeat sequence that both chromosomes M2 and M3 have in the type 2 mitogenome, as shown in Figure 4.1.

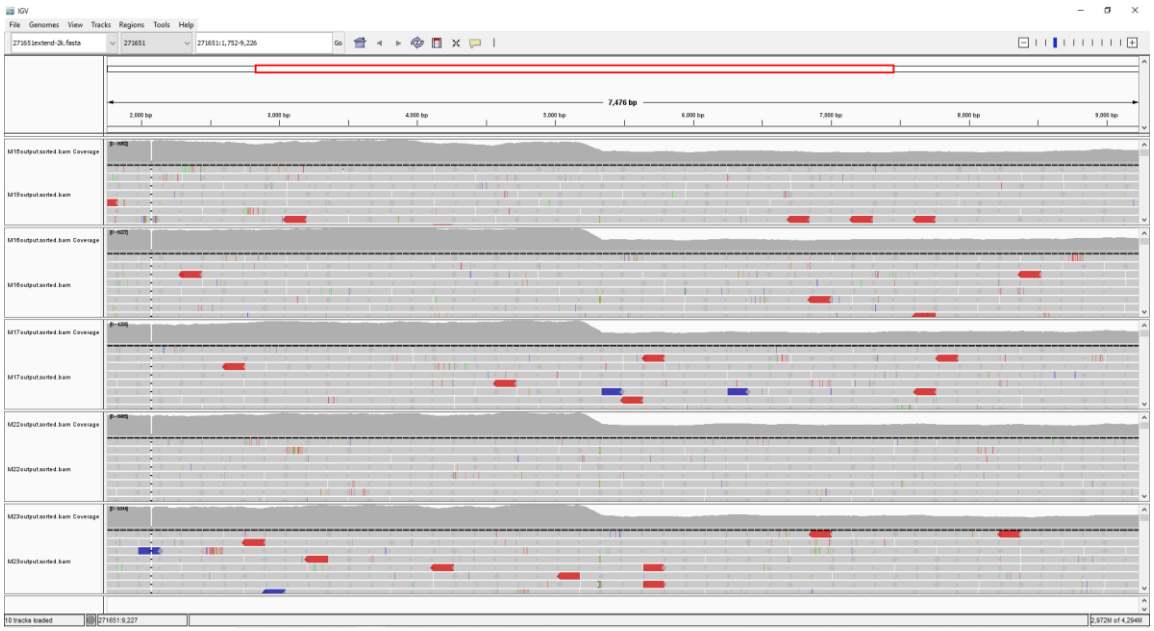

Figure S4.24 Bowtie 2 alignment of WGS Illumina reads from 5 M plants (M15, M16, M17, M22, and M23) to the 12,113 bp fragment from the type 2 mitochondrial genome chromosome M2, visualized in IGV. The reference sequence starts at C1 and ends at B (Figure 4.1), with the 2,108 bp type 2 specific sequence in between, from 5,325 bp to 7,432 bp. M plants (except M40) were grown from seeds collected from Namibia Farm. They all contain this 2,108 bp fragment, indicating that they all have type 2 mitochondrial genomes. Furthermore, the coverage of the C1 region is doubled in these 5 plants because it is a repeat sequence that both chromosomes M2 and M3 have in the type 2 mitogenome, as shown in Figure 4.1.

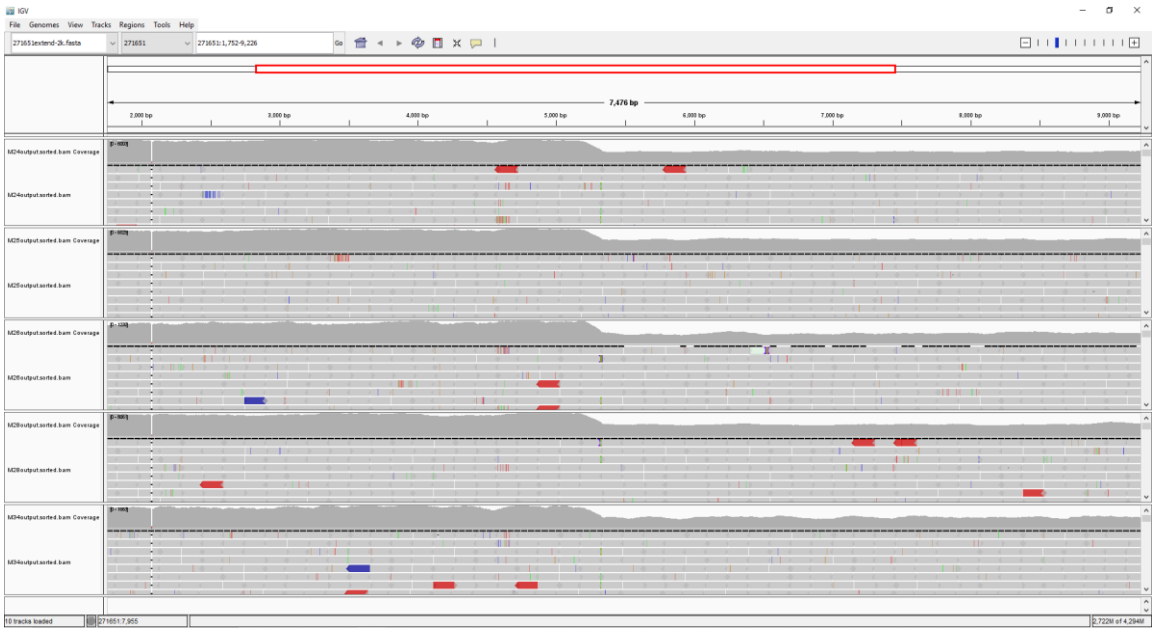

Figure S4.25 Bowtie 2 alignment of WGS Illumina reads from 5 M plants (M24, M25, M26, M28, and M34) to the 12,113 bp fragment from the type 2 mitochondrial genome chromosome M2, visualized in IGV. The reference sequence starts at C1 and ends at B (Figure 4.1), with the 2,108 bp type 2 specific sequence in between, from 5,325 bp to 7,432 bp. M plants (except M40) were grown from seeds collected from Namibia Farm. They all contain this 2,108 bp fragment, indicating that they all have type 2 mitochondrial genomes. Furthermore, the coverage of the C1 region is doubled in these plants because it is a repeat sequence that both chromosomes M2 and M3 have in the type 2 mitogenome, as shown in Figure 4.1.

Figure S4.26 Bowtie 2 alignment of WGS Illumina reads from 5 M plants (M31, M36, M37, M38, and M40) to the 12,113 bp fragment from the type 2 mitochondrial genome chromosome M2, visualized in IGV. The reference sequence starts at C1 and ends at B (Figure 4.1), with the 2,108 bp type 2 specific sequence in between, from 5,325 bp to 7,432 bp. All M plants (except M40) were grown from seeds collected from Namibia Farm. The origin of sample M40 is unknown. All plants except M40 contain this 2,108 bp fragment, indicating that they have type 2 mitochondrial genomes. Furthermore, the coverage of the C1 region is doubled in these 4 plants because it is a repeat sequence that both chromosomes M2 and M3 have in the type 2 mitogenome, as shown in Figure 4.1. M40 doesn't have this 2,108 bp fragment, and its C1 and B have close sequencing coverage, suggesting that M40 has a type 1 mitogenome.

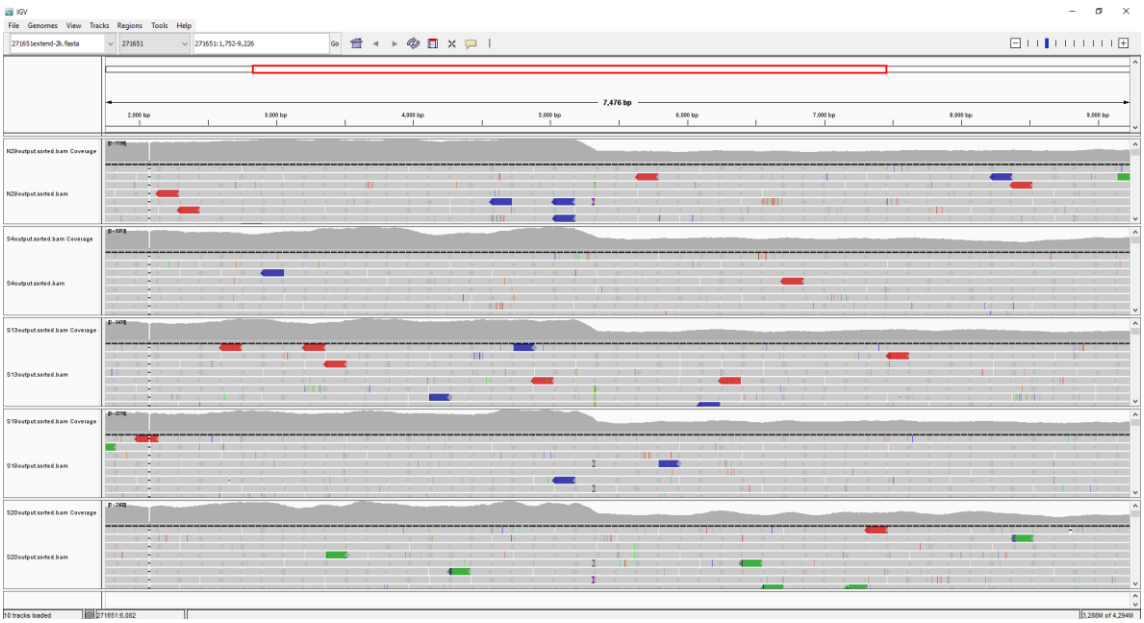

Figure S4.27 Bowtie 2 alignment of WGS Illumina reads from 5 M individuals (N29, S_4, S_13, S_19, and S_20) to the 12,113 bp fragment from the type 2 mitochondrial genome chromosome M2, visualized in IGV. The reference sequence starts at C1 and ends at B (Figure 4.1), with the 2,108 bp type 2 specific sequence in between, from 5,325 bp to 7,432 bp. M plants (except M40) were grown from seeds collected from Namibia Farm. They all contain this 2,108 bp fragment, indicating that they all have type 2 mitochondrial genomes. Furthermore, the coverage of the C1 region is doubled in these plants because it is a repeat sequence that both chromosomes M2 and M3 have in the type 2 mitogenome, as shown in Figure 4.1.

Figure S4.28 Bowtie 2 alignment of WGS Illumina reads from 4 M plants (S_27, S_30, S_33, and S_35) and R1R2 to the 12,113 bp fragment from the type 2 mitochondrial genome chromosome M2, visualized in IGV. The reference sequence starts at C1 and ends at B (Figure 4.1), with the 2,108 bp type 2 specific sequence in between, from 5,325 bp to 7,432 bp. M plants (except M40) were grown from seeds collected from Namibia Farm. The origin of sample R1R2 is unknown. However, they all contain this 2,108 bp fragment, indicating that they all have type 2 mitochondrial genomes. Furthermore, the coverage of the C1 region is doubled in all these 5 individuals, which consistent with the type 2 mitogenome structure, as C1 is a repeat sequence contained in both chromosomes M2 and M3 in the type 2 mitogenome, as shown in Figure 4.1.

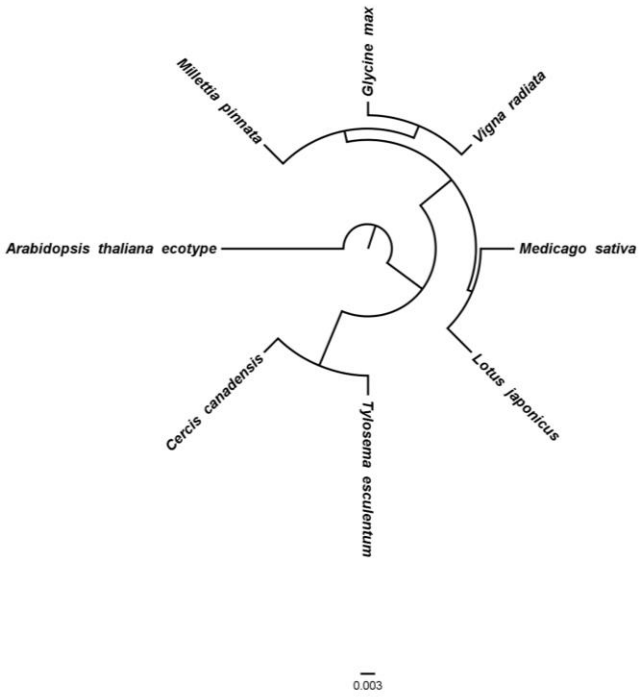

Figure S4.29 Bayesian inference tree on artificial chromosomes concatenated by 24 conserved mitochondrial genes, *atp1*, *atp4*, *atp6*, *atp8*, *atp9*, *nad3*, *nad4*, *nad4L*, *nad6*, *nad7*, *nad9*, *mttB*, *matR*, *cox1*, *cox3*, *cob*, *ccmFn*, *ccmFc*, *ccmC*, *ccmB*, *rps3*, *rps4*, *rps12*, and *rpl16* from the mitochondrial genomes of *Arabidopsis thaliana* (NC_037304.1), *Cercis canadensis* (MN017226.1), *Lotus japonicus* (NC_016743.2), *Medicago sativa* (ON782580.1), *Millettia pinnata* (NC_016742.1), *Glycine max* (NC_020455.1), and *Vigna radiata* (NC_015121.1) in NCBI drawn by BEAST and FigTree.

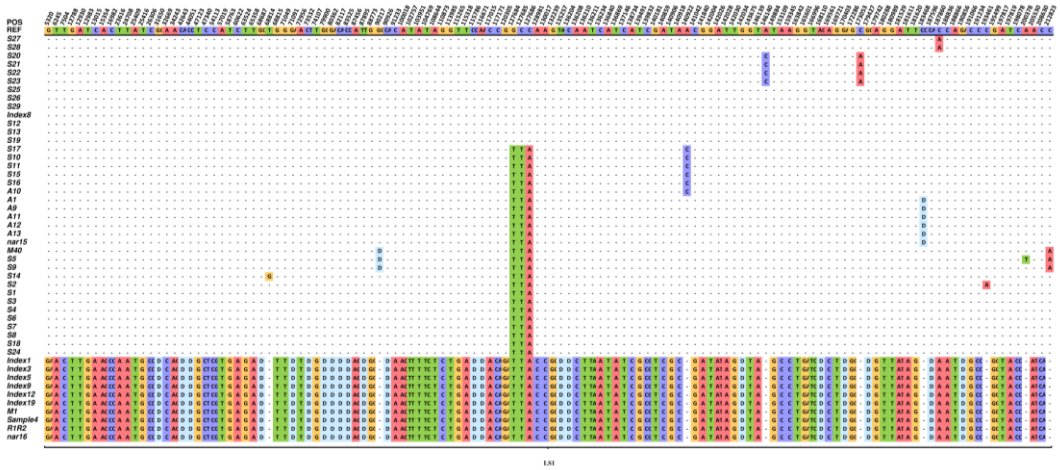

Figure S4.30 Distribution of variations discovered when aligning the WGS reads of 38 type 1 individuals and 10 type 2 plants to the first 212,823 bp of the marama reference mitogenome chromosome LS1, OK638188. The first row shows the bases in the reference sequence, from the second row onwards only the bases that differ from the reference are shown, and the bases that are the same as the reference are replaced by dashes. All insertions are represented by the first two bases and deletions by the letter "D".

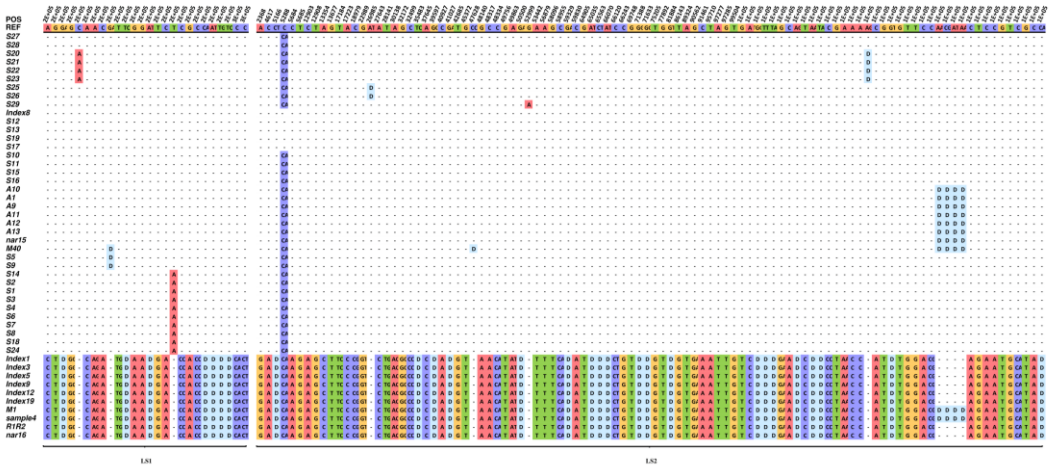

Figure S4.31 Distribution of variations discovered when aligning the WGS reads of 38 type 1 individuals and 10 type 2 plants to the marama reference mitogenome chromosomes LS1 (OK638188, 212,823 bp to 253,259 bp) and LS2 (OK638189). The first row shows the bases in the reference sequence, from the second row onwards only the bases that differ from the reference are shown, and the bases that are the same as the reference are indicated by dashes. All insertions are represented by the first two bases and deletions by the letter "D".

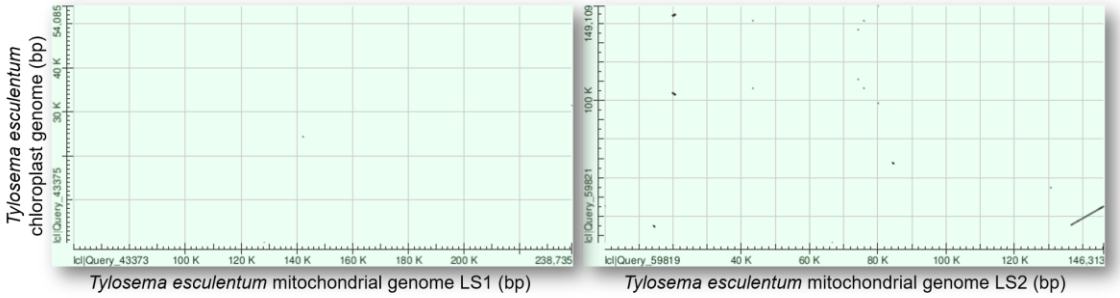

Figure S4.32 A dot plot view showing the alignment of marama mitochondrial chromosomes to the chloroplast genome. The two reference mitochondrial chromosomes of marama (OK638188 and OK638189) were blasted against the reference chloroplast genome of marama (KX792933.1), respectively in NCBI.

Table S4.2 List of homologous fragments between the mitochondrial and chloroplast genomes of *Tylosema esculentum*.

Plastome		Mitogenome			Alignment		
Start	End	Chr	Start	End	Length	%Identity	E-value
260	339	LS1	128344	128265	80	98.75	4.66E-33
24312	24387	LS1	142158	142233	76	98.684	7.80E-31
26169	26237	LS2	66608	66676	69	100	1.31E-28
31434	31539	LS1	238735	238628	108	93.519	4.63E-38
34586	34943	LS2	14640	14299	360	83.056	4.34E-83
35308	45222	LS2	136516	146313	9926	98.146	0
38063	38123	LS2	139272	139214	61	88.525	6.20E-12
45221	45257	LS2	146312	35	37	100	8.02E-11
54011	54085	LS1	59870	59796	75	94.667	2.83E-25
54818	54845	LS2	130707	130734	28	100	8.08E-06
67414	67689	LS2	84625	84353	282	90.426	1.53E-97
98539	98610	LS2	80134	80205	72	97.222	6.07E-27
102961	103824	LS2	20757	19899	891	74.074	1.21E-83
106303	106383	LS2	43447	43375	81	90.123	1.02E-19
106303	106383	LS2	75967	75895	81	90.123	1.02E-19
110921	111003	LS2	74229	74312	84	97.619	4.66E-33
136648	136730	LS2	74312	74229	84	97.619	4.66E-33
141268	141348	LS2	43375	43447	81	90.123	1.02E-19
141268	141348	LS2	75895	75967	81	90.123	1.02E-19
143827	144690	LS2	19899	20757	891	74.074	1.21E-83
149041	149112	LS2	80205	80134	72	97.222	6.07E-27

Table S4.3 Potential effect of variations found in the gene sequences on the 9,798 bp cpDNA insertion in the 84 individuals.

Position	**Reference**	**Variation**	**Gene**	**AA Substitution**
35570	A	C	*psbC*	N44H
35634	G	T	*psbC*	G65V
35774	TTTGTGTCT	DEL	*psbC*	FVS112-114Δ
35884	TTATGTAT	DEL	*psbC*	Y149fs*
36347	GGACCTACT	DEL	*psbC*	GPT303-305Δ
38753	CGATCTTC	DEL	*rps14*	G67fs*
38867	G	T	*rps14*	synonymous
38873	TTTTTTTGAGGATCGGCGAA	DEL	*rps14*	I23fs*
38893	T	G	*rps14*	I23L
38922	CTTTTCTTCT	DEL	*rps14*	E10fs*

Position	Reference	Variation	Gene	AA Substitution
39144	A	G	*psaB*	V711A
39264	CAATATCCA	DEL	*psaB*	GYW 669-671Δ
39412	C	A	*psaB*	D622Y
39414	C	A	*psaB*	R621I
39429	A	C	*psaB*	L616W
40061	G	T	*psaB*	D405E
40517	A	G	*psaB*	synonymous
40544	A	C	*psaB*	F244L
41030	A	G	*psaB*	synonymous
41243	A	G	*psaB*	synonymous
41716	C	A	*psaA*	D615Y
41718	G	T	*psaA*	S614Δ*
42060	GTTGCACCAGGGGCC	DEL	*psaA*	APGAT491-495Δ
42385	C	T	*psaA*	V392M
42587	C	T	*psaA*	synonymous
42793	G	C	*psaA*	Q256E
42817	CCAAAAGAT	DEL	*psaA*	DLL245-247Δ
42838	G	T	*psaA*	L241I
43508	ATCCCTAT	DEL	*psaA*	A15fs*

Genetic info represented by symbols in the table: AA, amino acids; Δ, deletion; fs, frameshift; *, nonsense mutation.

Chapter 5. The First High Quality Draft Genome Assembly of *Tylosema esculentum*

5.1 Introduction

Tylosema esculentum (marama bean), an underutilized orphan legume from southern Africa, has long been considered to have the potential to be domesticated as a crop (Jackson et al., 2009). Marama has a unique drought avoidance strategy by growing tubers weighing up to 500 pounds to store water, enabling it to survive the prolonged hot and dry conditions of the Kalahari Desert (Keith and Renew, 1975; Cullis et al., 2018). The seeds of marama are edible and nutritious, comparable to some commercial crops, and its domestication is thought to have the potential to improve local food security (Dakora, 2013; Omotayo and Aremu, 2021). A major obstacle to marama breeding is that it does not flower until at least the second year after planting, making traditional breeding inefficient. Studying the genetic diversity that exists in nature and utilizing molecular marker-assisted breeding strategies are considered good alternatives (Cullis et al., 2019; Hasan et al., 2019). The main goals of marama breeding include developing plants with an erect habit, which would facilitate seed harvesting in the field, and overcoming self-incompatibility, which would allow the development of inbred lines to accelerate the production of new varieties with favorable allelic combinations (Enciso-Rodr ǵuez et al., 2019; Cullis et al., 2022). Having a high-quality genome assembly will undoubtedly provide a reference for these studies.

The total genome size of *T. esculentum* was estimated to be 1 Gb with 44 chromosomes (2n=4x=44), according to the next-generation sequencing data and Feulgen staining (Takundwa et al., 2012; Cullis et al., 2019). Currently, Illumina whole genome

sequencing data of more than 80 marama individuals collected from various geographical locations in Namibia and South Africa, as well as PacBio long reads from a few individuals are available and deposited under PRJNA779273. These have been successfully used in the assembly of marama chloroplast and mitochondrial reference genomes (Kim and Cullis, 2017; Li and Cullis, 2021). Comparative genomic analyses were also performed to investigate the genetic diversity present in the marama organelle genome (Li and Cullis, 2023). However, the assembly of the nuclear genome is still at a very rudimentary level, with an N50 value of only 3 kb, by Dr. Kyle Logue using only the Illumina reads of marama.

Genome assembly has been greatly facilitated as next-generation sequencing has become cheaper, faster, and with higher throughput (Von Bubnoff, 2008). However, for the assembly of complex genomes, including polyploid genomes and repeat-rich genomes, short reads generated by the next-generation sequencing cannot fulfill these tasks. As a third-generation sequencing technology, PacBio provides longer reads, with an average length over 10 kb and up to 25 kb, making up for the previous shortcomings. The latest PacBio HiFi sequencing improves the accuracy rate to over 99.9% on the basis of retaining the length of reads (Hon et al., 2020). In this study, the first high-quality genome assembly of marama was accomplished using only data from the PacBio HiFi platform assembled by the tools HiCanu and Hifiasm. This has particular significance in the projected molecular breeding research work on marama.

5.2 Materials and Methods

The 20.11 GB PacBio HiFi reads generated for *T. esculentum* Sample 4 were assembled using HiCanu (Nurk et al., 2020; https://canu.readthedocs.io/en/latest/quick-

start.html#assembling-pacbio-hifi-with-hicanu) with input genome size set to 1 Gb according to the previous estimate (Cullis et al., 2019). Jellyfish 2.3.0 (Marçais and Kingsford, 2011; https://github.com/gmarcais/Jellyfish) was used to count k-mer for the PacBio HiFi reads of Sample 4 with the k-mer length set to 21 and to generate a k-mer count histogram, which was then used to draw k-mer spectra on GenomeScope 2.0 (Ranallo-Benavidez et al., 2020; http://qb.cshl.edu/genomescope/genomescope2.0/). The assembly quality was evaluated by QUAST 5.2.0. (Mikheenko et al, 2018; https://github.com/ablab/quast) and the results were visualized by Matplotlib v.1.3.1. (Hunter, 2007). The genome completeness was assessed by comparison with the Embryophyta ortholog database (embryophyta_odb10) containing 1614 genes and the Fabales ortholog database (fabales_odb10) containing 5366 genes using BUSCO v5.4.4 (Simão et al., 2015; https://busco.ezlab.org/). Genome repeat annotation was conducted by RepeatMasker 4.1.4 (Tarailo-Graovac and Chen, 2009; https://www.repeatmasker.org/) with the Repbase library (Bao et al., 2015). The marama genome assembly was mapped to the haplotype genome assembly of *Bauhinia variegata* ASM2237911v2 (Zhong et al., 2022) via minimap2 v2.24 (Li, 2018; https://github.com/lh3/minimap2). The resulting pairwise mapping format (PAF) data was visualized by a dot plot drawn by the R package pafr (https://github.com/dwinter/pafr).

5.3 Results

21.58 Gbp PacBio HiFi reads were obtained, based on which GenomeScope 2.0 was used to construct the k-mer distribution map, showing three peaks at 1-fold, 2-fold, and 4-fold coverage respectively (Figure 5.1). The PacBio data was found to fit the

tetraploid model best despite marama was initially thought to be an ancient hexaploid plant. In addition, the k-mer spectra showed high heterozygosity, with 2.2% of the genome being heterozygous. The frequency of aaab (1.410%) was found to be greater than that of aabb (0.498%), indicating that the individual studied was possibly an autotetraploid (Ranallo-Benavidez et al., 2020). The estimated genome size of *T. esculentum* was only 277.4 Mb, suggesting that marama has a compact genome, as reported for the legume *Amphicarpaea edgeworthii*. It has a genome size of 298.1 Mb and also has 11 chromosomes (*2n* = 22) (Liang et al., 2009; Liu et al., 2021).

Table 5. 1 *T. esculentum* sequencing and draft genome assembly statistics.

Genome Sequence	PacBio HiFi reads	2,184,632 reads 21.6 G bases
Genome Assembly Assessment	HiFi Read coverage*	17.97x
	Numbers of contigs	9,532
	N50	1,282,156 bp
	L50	252
	L90	3235
	Longest contigs	9,804,478 bp
	Assembly size	1,239,227,260 bp
	GC (%)	36.06
BUSCO Completeness	embryophyta_odb10	C: 99.5% [S: 0.8%, D: 98.7%], F: 0.2%, M: 0.3%, n: 1614
	fabales_odb10	C: 94.1% [S: 0.5%, D: 93.6%], F: 0.3%, M: 5.6%, n: 5366

BUSCO notation (C: complete, S: single, D: duplicated, F: fragmented, M: missing, and n: number of genes. * Read coverage was calculated based on a genome size of 1.2 Gb.

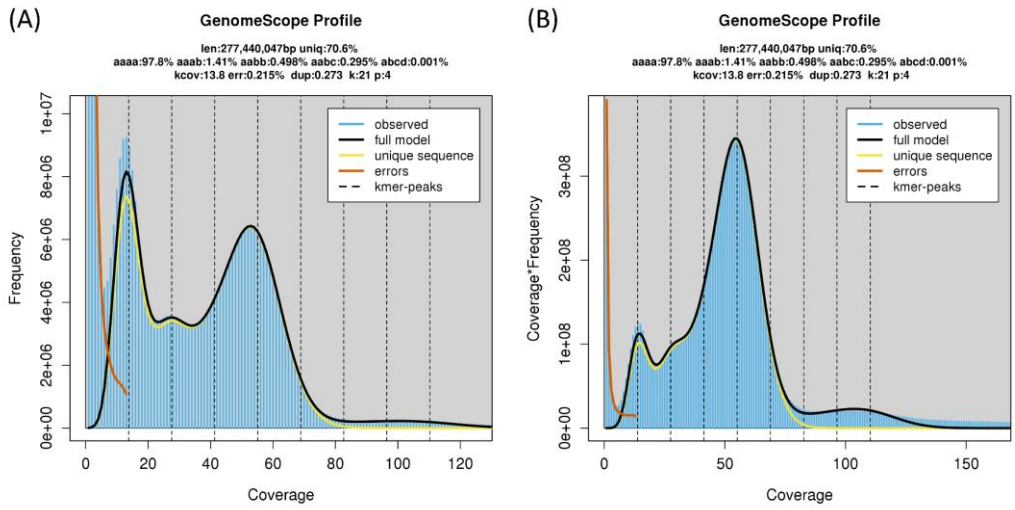

Figure 5. 1 K-mer spectra built on the PacBio HiFi reads of Sample 4 using GenomeScope 2.0. (A). Frequency-coverage k-mer spectrum. (B). Coverage*frequency-coverage k-mer spectrum.

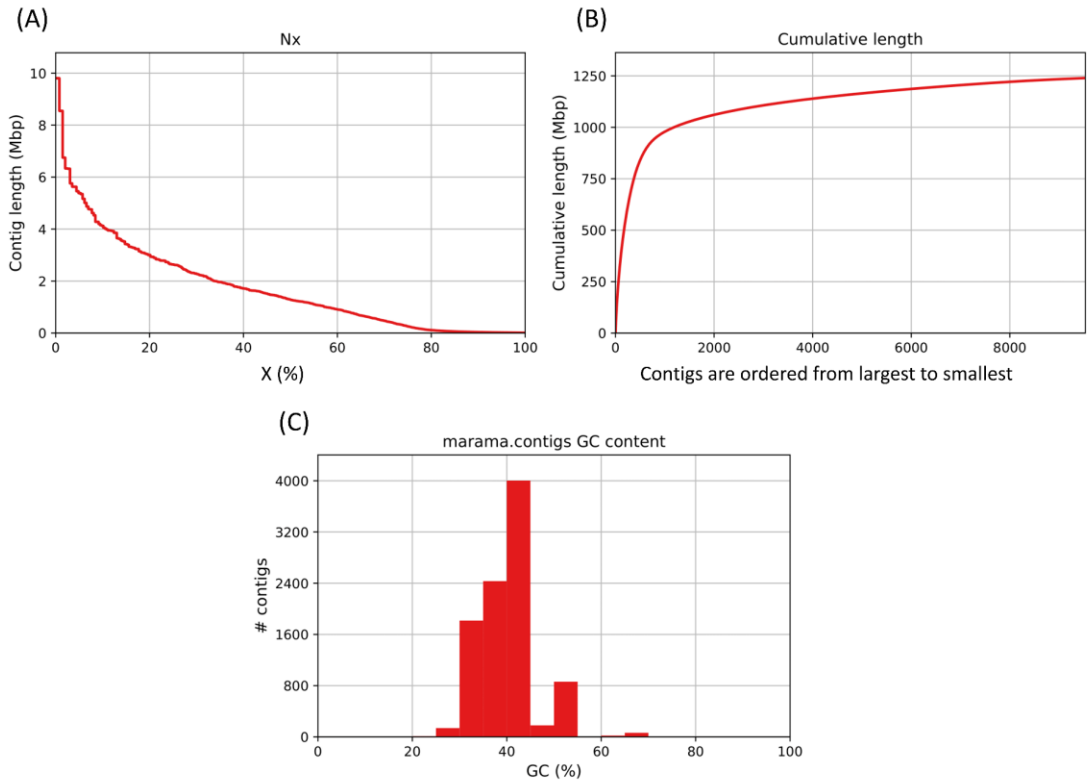

Figure 5. 2 Genome assembly quality assessment plots drawn by QUAST 5.2.0. (A) Nx plot showing the distribution of contig lengths as x varies from 0 to 100%. (B) Cumulative length plot. The contigs were sorted from largest to smallest. (C) GC plot showing the distribution of GC content in the contigs.

The 2,184,632 PacBio HiFi reads were assembled by Canu to generate a genome assembly of 1.24 Gb consisting of 9532 contigs, with an N50 value of 1.28 Mb and an L50 of 252 (minimum number of contigs with a total length equal to half the genome size) (Table 5.1). The current genome assembly is considered to contain contigs from the four haplotype genomes in preparation for future haplotype phasing. The average length of the obtained contigs was 1.24 Mb, and the longest contig was 9.80 Mb (Figure 5.2A). The L90 value of the genome assembly was 3,235, which means that the total length of

the top 3,235 contigs accounted for 90% of the genome size (Figure 5.2B). The average guanine-cytosine (GC) of all contigs was 36.06% (Table 5.1 and Figure 5.2C). BUSCO was used to evaluate the completeness of the *T. esculentum* genome assembly based on the comparison with 5366 genes from the Fabales ortholog database. 94.1% of these genes were found in the *T. esculentum* genome assembly. Furthermore, 99.5% of the genes from the Embryophyta ortholog database were detected in our assembly, indicating that it is highly complete. As expected, the proportions of duplicated genes were 93.6% and 98.7%, respectively, since most genes should have four copies in tetraploid plants. 27.35% of the *T. esculentum* genome assembly was annotated as repeats by RepeatMasker (Table 5.2). Long-terminal repeat (LTR) retroelements accounted for 13.21% of the genome, of which Ty1/Copia made up to 3.55% of the genome, and Gypsy/DIRS1 accounted for 9.14% of the genome. Low-complexity regions (LCRs) and simple sequence repeats (SSRs) were found to account for 6.47% and 4.05% of the *T. esculentum* genome size, respectively.

Table 5. 2 Summary of repeat elements in the *T. esculentum* genome assembly by RepeatMasker.

	Number of elements*	Length occupied (bp)	Percentage of sequence (%)
Retroelements	239,152	176,219,041 bp	14.22%
SINEs:	2,428	151,942 bp	0.01%
Penelope	4,697	257,720 bp	0.02%
LINEs:	56,298	12,306,565 bp	0.99%
CRE/SLACS	148	13,819 bp	0.00%
L2/CR1/Rex	3,208	171,686 bp	0.01%
R1/LOA/Jockey	1,800	96,061 bp	0.01%
R2/R4/NeSL	433	22,878 bp	0.00%
RTE/Bov-B	3,045	560,899 bp	0.05%
L1/CIN4	39,173	10,918,951 bp	0.88%
LTR elements	180,426	163,760,534 bp	13.21%
BEL/Pao	2,400	286,196 bp	0.02%
Ty1/Copia	45,240	43,947,485 bp	3.55%
Gypsy/DIRS1	111,682	113,301,860 bp	9.14%
Retroviral	5,724	287,611 bp	0.02%
DNA transposons	58,669	6,960,572 bp	0.56%
hobo-Activator	10,674	1,367,473 bp	0.11%
Tc1-IS630-Pogo	2,621	148,140 bp	0.01%
En-Spm	0	0 bp	0.00%
MULE-MuDR	10,709	1,298,115 bp	0.10%
PiggyBac	316	16,841 bp	0.00%
Tourist/Harbinger	3,996	537,089 bp	0.04%
Other (Mirage, P-element, Transib)	891	39,450 bp	0.00%
Rolling-circles	6,352	1,190,800 bp	0.10%
Unclassified:	3,289	278,950 bp	0.02%
Total interspersed repeats:		183,458,563 bp	14.80%
Small RNA:	14,055	14,349,721 bp	1.16%
Satellites:	19,870	9,567,228 bp	0.77%
Simple repeats:	421,704	50,213,770 bp	4.05%
Low complexity:	114,847	80,219,159 bp	6.47%

* Most repeats fragmented by insertions or deletions have been counted as one element

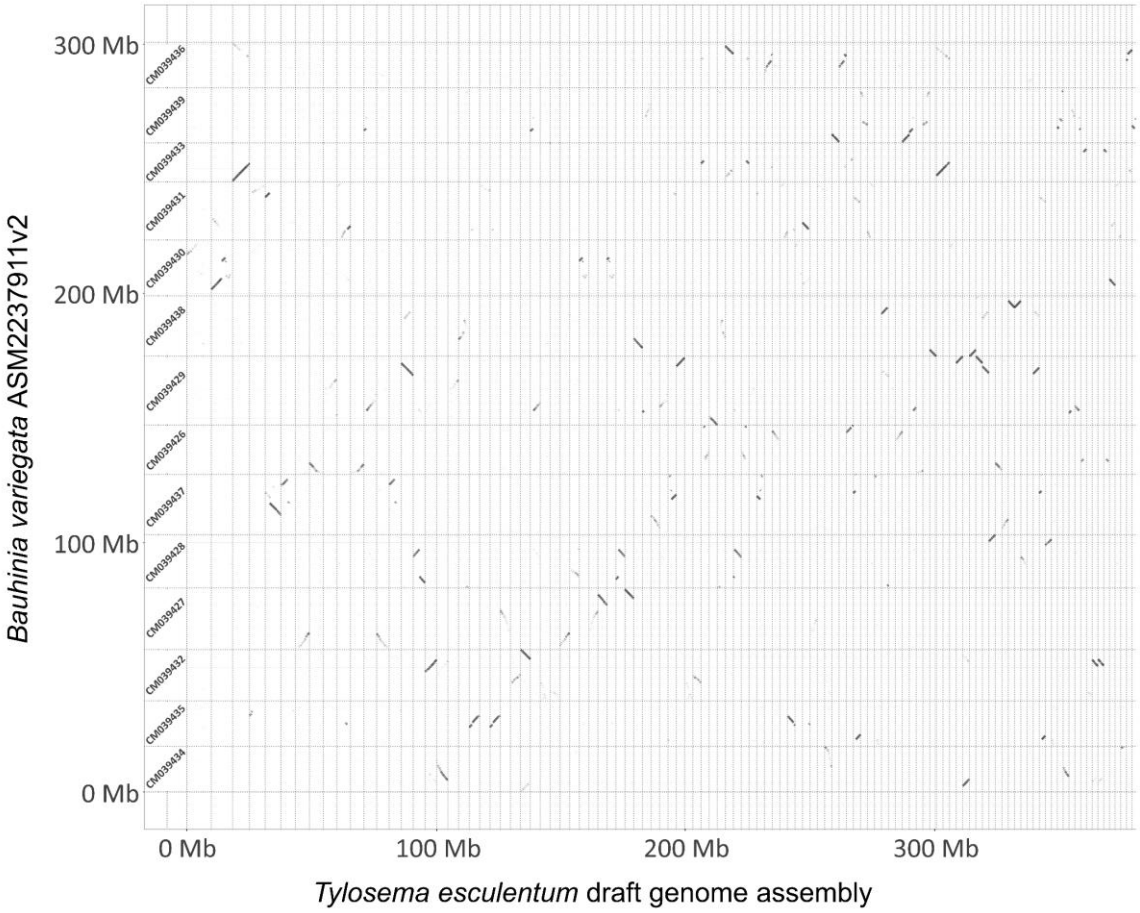

Figure 5. 3 A dot plot of alignment of partial *T. esculentum* assembled contigs against the 14 chromosomes of *B. variegata* genome assembly ASM2237911v2. This figure was drawn by the R package pafr on the pairwise mapping format (PAF) document generated by minimap2. Each row represents one chromosome of the *B. variegata* genome with the chromosome GenBank ID labeled at the beginning of the row. Each column represents one contig from the *T. esculentum* genome assembly, sorted by size. Only the first 370 Mb of the 1.2 Gb assembly are included here, with highly fragmented contigs not shown. The black dotted lines show where the two genomes align. The ticks on both axes indicate the genomic scales in base pairs.

Genomes of only a few plants from the Cercidoideae subfamily have been assembled, of which *B. variegata* is the evolutionarily closest to *T. esculentum* (Wunderlin, 2010). The haplotype genome assembly of *B. variegata* ASM2237911v2 has a size of 326.4 Mb, and contains 14 chromosomes ($2n = 28$) ranging in length from 18,256,449 bp to 27,622,603 bp (Zhong et al., 2022). The genome assembly of *T. esculentum* was mapped to the *B. variegata* genome by minimap2 and the result was visualized as a dot plot using R package pafr (Figure 5.3). Some contigs from the *T. esculentum* genome assembly reached half the chromosome length of *B. variegata* and exhibited a high degree of collinearity, confirming the reliability of our assembly. When Illumina reads from randomly selected samples (M1, M40, Index1) were mapped to the *B. variegata* genome via Bowtie2 v2.4.4 (Langmead and Salzberg, 2012; https://github.com/BenLangmead/bowtie2), the overall alignment rate was only around 20.36%, and the alignment rate with the genome of *Vigna radiata* PRJNA301363 was only 2.7% (Kang et al., 2014), indicating that *T. esculentum* has a highly divergent genome from these legumes.

5.4 Discussion

The is the first reported high-quality draft genome assembly of *T. esculentum* with an N50 value of 1.28 Mb, which has been dramatically improved from the 3 kb of the previous assembly done by Dr. Kyle Logue on Illumina reads. Although the current genome assembly of *T. esculentum* still contains numerous highly fragmented contigs, this is considered an ongoing project that will continual to be optimized in the future. Moreover, many of the obtained contigs are long enough to be used in the study of genes of interest, providing an important reference for marama breeding. The genomic resource

lays a foundation for the study of genetic diversity existing in nature, which can be used to explore the genetic mechanism of self-incompatibility in marama and study the adaptation of plants to harsh environments, etc.

Another assembler Hifiasm 0.18.5 (Cheng et al., 2021; https://github.com/chhylp123/hifiasm) was used for haplotype assembly of *T. esculentum* and generated a partially phased assembly of 564.8 Mb. This contained 4,123 contigs with an N50 value of 2.75 Mb and an L50 of 35. The BUSCO score was 99.1% (S:62.0%, D:37.1%, F:0.4%, M:0.5%, n=1614) compared to the Embryophyta ortholog database (embryophyta_odb10) and 93.5% (S:45.5%, D:48.0%, F:0.5%, M:6.0%, n=5366) to the Fabales ortholog database (fabales_odb10). The contigs generated by Hifiasm have longer N50 and better continuity, but it is still a partially phased genome assembly that contains a large number of duplications, which need to be purged by third-party tools. However, this may also collapse repeats or segmental duplications that should have been included. Although the assembly from HiCanu is more fragmented, it is considered to have more complete genetic information. In the future, by using data from other sequencing platforms like Hi-C, contigs from the same chromosome can be grouped and then further scaffolded to the near-chromosome level (Aiden et al., 2009; Mascher et al., 2017).

Chapter 6. Discussion

T. esculentum, as an underutilized legume, is considered to be of high research value for the several reasons. First, the seeds and tubers of marama are edible and nutritious. Its domestication has long been considered important for improving local food security. Second, as a special plant species, marama can grow in harsh environments with long-term high temperature and low rainfall. The genetic study on marama may help us better understand plant adaptions and stress responses to extreme conditions. This is not only of great significance to the breeding of the bean, but also provides guidance for the improvement of traditional crops. In this study, the mitogenome, chloroplast genome, and draft nuclear genome of marama were assembled using data from multiple sequencing platforms, including Illumina, PacBio, and PacBio HiFi. Meanwhile, the genetic diversity was explored by sequencing and comparing a large number of individuals collected from many different geographical locations. This laid an important foundation for the improvement of marama in the future.

The diversity analysis revealed two distinct cytotypes that differed significantly from each other in both mitogenomes and chloroplast genomes. The two cytotypes appear to have differing levels of variability, with one being found to be more conserved than the other, although this may be due to sampling errors. Various degrees of heteroplasmy were observed, with alleles from different types of organelle genomes co-existing in the same individual. Heteroplasmy was found to be more prevalent in the chloroplast genome of marama, although the frequency of minor alleles was generally below 2%. Heteroplasmy within the mitogenome was found to be concentrated at several differential loci and to be present at higher levels, suggesting that these loci may have

played important roles in the divergence of marama mitogenome. Comparing the levels of organelle heteroplasmy in the related individuals suggests that its inheritance is fairly stable, providing a conundrum of how the two genomes co-exist and are propagated through generations. The nuclear genes *MSH1*, *OSB1*, and *RECA* homologs were found to be associated with the prevalence of heteroplasmy, but mainly by sorting *de novo* mutations through homologous recombination (Zaegel et al., 2007; Miller-Messmer et al., 2012; Amanda et al., 2022). It is unclear whether the heteroplasmy caused by paternal leakage is also regulated by these genes. In this study, higher levels of heteroplasmy were found in both organelle genomes of an individual, and it is unclear whether this was due to accidental mixing of samples or the individual had traits that led to escape from genetic bottlenecks (Floros et al., 2018). Sequencing and comparing plant cells under different environmental selections and at different developmental stages can reveal whether the level of heteroplasmy changes during plant cell development and whether there are bottlenecks that preferentially amplify one genome over another. This will help us better understand cytoplasmic inheritance and its role in plant stress responses.

The type 1 mitogenome was found to have two autonomous rings with a total length of 399,572 bp, and recombination on the 3 pairs of long direct repeats (>1kb) was thought to be able to restructure them into five smaller circular molecules. Recombination between a pair of long inverted repeats can invert the sequence in between. All detected structures were found to have very close molar concentrations, suggesting that these mentioned recombinations occur with high frequency in cells. The function of the structural changes is unclear so far. Based on our data, they were not found to be involved in the alteration of gene coding sequences, but may still play a role

in chromosomal sequence repair. The drivers behind the recombination are unclear, but the genes *OSB1*, *RECA*, and *MSH1* have been reported to affect homologous recombination in mtDNA, which is thought to be required for mtDNA repair (Zaegel et al., 2007; Miller-Messmer et al., 2012; Amanda et al., 2022). In addition, changes in chromosome structure can have many other potential effects, such as altering regulatory regions or affecting the binding of DNA strands to key proteins to regulate gene expression. These require further investigation of transcriptome data. The type 2 mitogenome contains a unique sequence of 2,108 bp, likely derived from the mitogenome of *Lupinus*. This sequence doesn't contain any genes, but is thought to play an important role in the variation of marama mitogenome structure. It serves as a key unit that connects distant fragments to form a new ring. Along with several other minor changes, it results in three circular molecules and one linear chromosome in the type 2 mitogenome. The structural change was found to result in increased copy number of the genes *nad9*, *rrn5*, *rrnS*, *trnC*, and *trnfM*. Whether this is reflected in the level of gene expression needs to be verified by transcriptome data.

The cpDNA insertions were concentrated in one subgenomic ring of the mitogenome of marama, suggesting that this ring is more active than others in exchanging genetic information with the chloroplast genome, although the reason remains unknown. A 9,798 bp long cpDNA insertion containing potential *psbC*, *rps14*, *psaA*, and *psaB* pseudogenes was found in the mitogenomes of all marama individuals. The study of the polymorphism on this segment indicated that only synonymous substitutions may be retained when this segment is located in the chloroplast genome. However, after being inserted into the mitogenome, this fragment began to accumulate a

large number of mutations, making the genes on it lose function. Mitochondrial genes themselves are actually quite conserved with a lower mutation rate than that of other organelle genomes, suggesting that organelle genomes may have mechanisms to protect their own genes rather than externally inserted ones (Wolfe et al., 1987; Drouin et al., 2008).

The two types of chloroplast genomes range in length from 161,537 bp to 161,580 bp, differing at 122 loci and at a 230 bp inversion. This inversion has not been reported before in any other species, and it may be of some use for evolutionary studies of marama. More variations were found in the coding sequences of marama chloroplast genes than mitochondrial genes. The genes *rpoC2*, *rpoB*, and *ndhD* were found to be more diverse than other chloroplast genes.

The analysis of mitogenome diversity found that soil moisture levels may play an important role in the divergence of the type 1 mitogenome, although this still needs to be verified by studies with larger sample sizes. The comparative analysis of mitochondrial conserved genes in marama and other legumes found that, among the selected species, Cercis was most closely related to marama, and both of them tended to have more complete sets of mitochondrial genes. However, the similarity between even the closest mitogenomes was very low. This is mainly due to the fact that plant mitochondrial genomes, despite their highly conserved gene sequences, frequently acquire non-coding DNA segments from other organelles and even other species for unknown reasons (Kazuyoshi and Kubo, 2010).

In this study, a draft genome of marama was assembled based on 21.6 G bases PacBio HiFi data generated from the prepared high molecular weight DNA samples. A

1.24 Gb unphased assembly was obtained using Canu v2.2 and a partially phased assembly of 564.8 Mb was got from Hifiasm 0.18.5. The k-mer analysis based on the PacBio data indicates that marama is likely to be ancient tetraploid species, although it has long been considered an ancient hexaploid plant. However, different levels of polyploidy may exist in marama, as has been described in many other species (Duchoslav et al., 2010). The genome size estimated from the k-mer spectra is very small, only 277 Mb. This suggests that marama has one of the smallest genomes in legumes (Liu et al., 2021). Our main assembly (from Canu) is considered to have all the information from the four haplotype genomes, which will need to be phased in the future. The assembly has a high continuity with an N50 value of 1.28 Mb, and the length of some conitgs reaches half the length of the *Bauhinia* chromosome. This is a significant increase from the 3 kb N50 of the previous assembly based on Illumina reads. This makes us the best genome assembly for marama to date. The BUSCO completeness of the assemblies from both tools exceeded 99%. Repeats were found to account for 27.35% of the genome of marama.

The nuclear genome assembly of marama is an ongoing project. In the future, the obtained contigs will be further scaffolded into the chromosome level and genome phasing will be performed with the aim of extracting a haplotype-resolved assembly using data from other sequencing platforms such as Hi-C. Genome phasing that separates alleles from different chromosomal copies to facilitate the construction of precise chromosomal structures is thought to be important. For example, CsSRC2 and CsGGPS1 are two important genes in plant physiology involved in the cold stress response and terpenoid backbone biosynthesis, respectively (Zhang et al., 2021). Both were found to

have haplotype-specific expression and contain haplotype-specific variants, which could not be clearly explained by traditional unphased or pseudo-haplotype genome assemblies.

References

Adams, K. L., Qiu, Y. L., Stoutemyer, M., and Palmer, J. D. (2002). Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proceedings of the National Academy of Sciences, 99(15), 9905-9912. doi: 10.1073/pnas.042694899

Aguiar, B., Vieira, J., Cunha, A. E., and Vieira, C. P. (2015). No evidence for Fabaceae gametophytic self-incompatibility being determined by Rosaceae, Solanaceae, and Plantaginaceae *S-RNase* lineage genes. BMC Plant Biology, 15, 129. doi: 10.1186/s12870-015-0497-2

Aiden, E. L., Van Berkum, N. L., Williams, L. H., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326(5950), 289-293. doi: 10.1126/science.1181369

Alverson, A. J., Wei, X., Rice, D. W., Stern, D. B., Barry, K., and Palmer, J. D. (2010). Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Molecular Biology and Evolution, 27(6), 1436-1448. doi: 10.1093/molbev/msq029

Alverson, A. J., Zhuo, S., Rice, D. W., Sloan, D. B., and Palmer, J. D. (2011). The mitochondrial genome of the legume *Vigna radiata* and the analysis of recombination across short mitochondrial repeats. PLoS ONE, 6(1), e16404. doi: 10.1371/journal.pone.0016404

Amarteifio, J. O., and Moholo, D. (1998). The chemical composition of four legumes consumed in Botswana. Journal of Food Composition and Analysis, 11(4), 329-332. doi: 10.1006/jfca.1998.0595

Andersson, S. G. E., Karlberg, O., Canbäck, B., and Kurland, C. G. (2003). On the origin of mitochondria: a genomics perspective. Philosophical Transactions of the Royal Society B, 358(1429), 165-179. doi: 10.1098/rstb.2002.1193

Anisimova, I. N., and Gavrilenko, T. (2017). Cytoplasmic male sterility and prospects for its utilization in potato breeding, genetic studies and hybrid seed production. Russian Journal of Genetics: Applied Research, 7(7), 721-735. doi: 10.1134/s2079059717070024

Ari, Ş., and Arikan, M. (2015). Next-generation sequencing: advantages, disadvantages, and future. Springer International Publishing EBooks, 109-135. doi: 10.1007/978-3-319-31703-8_5

Arrieta-Montiel, M. P., Lyznik, A., Woloszynska, M., Janska, H., Tohme, J., and Mackenzie, S. A. (2001). Tracing evolutionary and developmental implications of mitochondrial stoichiometric shifting in the common bean. Genetics, 158(2), 851-864. doi: 10.1093/genetics/158.2.851

Arrieta-Montiel, M. P., Shedge, V., Davila, J. I., Christensen, A. J., and Mackenzie, S. A. (2009). Diversity of the *Arabidopsis* mitochondrial genome occurs via nuclear-controlled recombination activity. Genetics, 183(4), 1261-1268. doi: 10.1534/genetics.109.108514

Ashley, M. V., Laipis, P. J., and Hauswirth, W. W. (1989). Rapid segregation of heteroplasmic bovine mitochondria. Nucleic Acids Research, 17, 7325-7331. doi: 10.1093/nar/17.18.7325

Backert, S., Nielsen, B. L., and Börner, T. (1997). The mystery of the rings: structure and replication of mitochondrial genomes from higher plants. Trends in Plant Science, 2(12), 477-483. doi: 10.1016/s1360-1385(97)01148-5

Bandel, G. (1974). Chromosome numbers and evolution in the Leguminosae. Caryologia, 27, 17-32. doi: 10.1080/00087114.1974.10796558

Banks, J. A., and Birky, C. W. (1985). Chloroplast DNA diversity is low in a wild plant, *Lupinus texensis*. Proceedings of the National Academy of Sciences, 82, 6950-6954. doi: 10.1073/pnas.82.20.6950

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA, 6(1). doi: 10.1186/s13100-015-0041-9

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). Misa-web: a web server for microsatellite prediction. Bioinformatics, 33, 2583-2585. doi: 10.1093/bioinformatics/btx198

Belitz, H. D., Grosch, W., and Schieberle, P. (2004). Food Chemistry. Berlin: Springer. doi: 10.1007/978-3-662-07279-0

Bi, C., Lu, N., Xu, Y., He, C., and Lu, Z. (2020). Characterization and analysis of the mitochondrial genome of common bean (*Phaseolus vulgaris*) by comparative

genomic approaches. International Journal of Molecular Sciences, 21(11), 3778. doi: 10.3390/ijms21113778

Boore, J. L. (1999). Animal mitochondrial genomes. Nucleic Acids Research, 27(8), 1767-1780. doi: 10.1093/nar/27.8.1767

Bower, N., Hertel, K., Oh, J., and Storey, R. (1988). Nutritional evaluation of marama bean (*Tylosema esculentum*, Fabaceae): analysis of the seed. Economic Botany, 42(4), 533-540. doi: 10.1007/bf02862798

Broz, A. K., Keene, A., Gyorfy, M. F., Hodous, M., Johnston, I. G., and Sloan, D. B. (2022). Sorting of mitochondrial and plastid heteroplasmy in *Arabidopsis* is extremely rapid and depends on *MSH1* activity. Proceedings of the National Academy of Sciences, 119. doi: 10.1073/pnas.2206973119

Budar, F., and Roux, F. (2011). The role of organelle genomes in plant adaptation. Plant Signaling and Behavior, 6(5), 635-639. doi: 10.4161/psb.6.5.14524

Cao, L., Shitara, H., Horii, T., Nagao, Y., Imai, H., Abe, K., et al. (2007). The mitochondrial bottleneck occurs without reduction of mtDNA content in female mouse germ cells. Nature Genetics, 39, 386-390. doi: 10.1038/ng1970

Carbonell-Caballero, J., Alonso, R., Ibañez, V., Terol, J., Talon, M., and Dopazo, J. (2015). A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus *Citrus*. Molecular Biology and Evolution, 32, 2015-2035. doi: 10.1093/molbev/msv082

Chan, P., Lin, B., Mak, A., and Lowe, T. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Research, 49(16), 9077-9096. doi: 10.1093/nar/gkab688

Chang, S., Wang, Y., Lu, J., Gai, J., Li, J., Chu, P., et al. (2013). The mitochondrial genome of soybean reveals complex genome structures and gene evolution at intercellular and phylogenetic levels. PLoS ONE, 8(2), e56502. doi: 10.1371/annotation/5bf22546-6983-42c9-9cb5-1a6459b29a79

Charboneau, J. L. M., Cronn, R. C., Liston, A., Wojciechowski, M. F., and Sanderson, M. J. (2021). Plastome structural evolution and homoplastic inversions in Neo-Astragalus (Fabaceae). Genome Biology and Evolution, 13. doi:10.1093/gbe/evab215

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., and Xia, R. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. Molecular Plant, 13(8), 1194-1202. doi: 10.1016/j.molp.2020.06.009

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods, 18(2), 170-175. doi: 10.1038/s41592-020-01056-5

Cheng, N., Lo, Y., Ansari, M. I., Ho, K., Jeng, S., Lin, N., et al. (2016). Correlation between mtDNA complexity and mtDNA replication mode in developing cotyledon mitochondria during mung bean seed germination. New Phytologist, 213(2), 751-763. doi: 10.1111/nph.14158

Claessen, H., Keulemans, W., Poel, B. V., and Storme, N. D. (2019). Finding a compatible partner: self-incompatibility in European pear (*Pyrus communis*); molecular control, genetic determination, and impact on fertilization and fruit set. Frontiers in Plant Science, 10, 407. doi: 10.3389/fpls.2019.00407

Cole, L. W., Guo, W., Mower, J. P., and Palmer, J. D. (2018). High and variable rates of repeat-mediated mitochondrial genome rearrangement in a genus of plants. Molecular Biology and Evolution. doi: 10.1093/molbev/msy176

Cullis, C., Chimwamurombe, P., Barker, N., Kunert, K., and Vorster, J. (2018). Orphan legumes growing in dry environments: marama bean as a case study. Frontiers in Plant Science, 9. doi: 10.3389/fpls.2018.01199

Cullis, C., Chimwamurombe, P., Kunert, K., and Vorster, J. (2022). Perspective on the present state and future usefulness of marama bean (*Tylosema esculentum*). Food and Energy Security. doi: 10.1002/fes3.422

Cullis, C., Lawlor, D. W., Chimwamurombe, P., Bbebe, N., Kunert, K., and Vorster, J. (2019). Development of marama bean, an orphan legume, as a crop. Food and Energy Security, 8. doi: 10.1002/fes3.164

Dakora, F. D. (2013). Biogeographic distribution, nodulation and nutritional attributes of underutilized indigenous African legumes. Acta Horticulturae, 979(3), 53-64. doi: 10.17660/actahortic.2013.979.3

Darling, A. E., Mau, B., Blattner, F. R., and Perna, N. T. (2004b). Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Research, 14(7), 1394-1403. doi: 10.1101/gr.2289704

de Lange, T. (2015). A loopy view of telomere evolution. *Frontiers in Genetics*, 6. doi: 10.3389/fgene.2015.00321

Delcher, A. L., Phillippy, A., Carlton, J., and Salzberg, S.L. (2002). Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Research, 30(11), 2478-2483. doi: 10.1093/nar/30.11.2478

Dobrogojski, J., Adamiec, M., and Luciński, R. (2020). The chloroplast genome: a review. Acta Physiologiae Plantarum, 42. doi: 10.1007/s11738-020-03089-x

Dong, S., Zhao, C., Chen, F., Liu, Q. H., Zhang, S., Wu, H., Zhang, L., and Liu, Y. (2018). The complete mitochondrial genome of the early flowering plant *Nymphaea colorata* is highly repetitive with low recombination. BMC Genomics, 19(1). doi: 10.1186/s12864-018-4991-4

Drescher, A., Ruf, S., Calsa, T., Carrer, H., and Bock, R. (2000). The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. The Plant Journal, 22, 97-104. doi: 10.1046/j.1365-313x.2000.00722.x

Drouin, G., Daoud, H., and Xia, J. (2008). Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Molecular Phylogenetics and Evolution, 49(3), 827-831. doi: 10.1016/j.ympev.2008.09.009

Duchoslav, M., Šafářová, L., and Krahulec, F. (2010). Complex distribution patterns, ecology and coexistence of ploidy levels of *Allium oleraceum* (Alliaceae) in the Czech Republic. Annals of Botany, 105(5), 719-735. doi: 10.1093/aob/mcq035

Dyall, S. D., Brown, M., and Johnson, P. J. (2004). Ancient invasions: from endosymbionts to organelles. Science, 304(5668), 253-257. doi: 10.1126/science.1094884

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32, 1792-1797. doi: 10.1093/nar/gkh340

Edgar, R. C. (2022). Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nature Communications, 13(1). doi: 10.1038/s41467-022-34630-w

Epstein, C. B., Waddle, J. A., Hale, W., Davé, V., Thornton, J., Macatee, T. L., et al. (2001). Genome-wide responses to mitochondrial dysfunction. Molecular Biology of the Cell, 12(2), 297-308. doi: 10.1091/mbc.12.2.297

Enciso-Rodríguez, F. E., Manrique-Carpintero, N. C., Nadakuduti, S. S., Buell, C. R., Zarka, D., and Douches, D. S. (2019). Overcoming self-incompatibility in diploid potato using CRISPR-Cas9. Frontiers in Plant Science, 10. doi: 10.3389/fpls.2019.00376

Estavillo, G. M., Crisp, P. A., Pornsiriwong, W., Wirtz, M., Collinge, D., Carrie, C., et al. (2011). Evidence for a SAL1-PAP chloroplast retrograde pathway that functions in drought and high light signaling in *Arabidopsis*. Plant Cell, 23, 3992-4012. doi: 10.1105/tpc.111.091033

Fauron, C., Casper, M., Gao, Y., and Moore, B. (1995). The maize mitochondrial genome: dynamic, yet functional. Trends in Genetics, 11(6), 228-235. doi: 10.1016/s0168-9525(00)89056-3

Floros, V. I., Pyle, A., Dietmann, S., Wei, W., Tang, W. W. C., Irie, N., Payne, B. a. I., Capalbo, A., Noli, L., Coxhead, J., Hudson, G., Crosier, M., Strahl, H., Khalaf, Y., Saitou, M., Ilic, D., Surani, M. A., and Chinnery, P. F. (2018). Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. Nature Cell Biology, 20(2), 144-151. doi: 10.1038/s41556-017-0017-8

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. Nucleic Acids Research, 32, W273-W279. doi: 10.1093/nar/gkh458

Galluzzi, L., Kepp, O., Trojel-Hansen, C., and Kroemer, G. (2012). Mitochondrial control of cellular life, stress, and death. Circulation Research, 111(9), 1198-1207. doi: 10.1161/circresaha.112.268946

Garc á-Valencia, L. E., Bravo-Alberto, C. E., Wu, H. M., Rodr ́guez-Sotres, R., Cheung, A. Y., and Cruz-Garc á, F. (2017). SIPP, a novel mitochondrial phosphate carrier, mediates in self-incompatibility. Plant Physiology, 175(3), 1105-1120. doi: 10.1104/pp.16.01884

Gasteiger, E. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research, 31, 3784-3788. doi: 10.1093/nar/gkg563

Goremykin, V. V., Salamini, F., Velasco, R., and Viola, R. (2008). Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. Molecular Biology and Evolution, 26(1), 99-110. doi: 10.1093/molbev/msn226

Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Research, 47(W1), W59-W64. doi: 10.1093/nar/gkz238

Gualberto, J. M., and Newton, K. J. (2017). Plant mitochondrial genomes: dynamics and mechanisms of mutation. Annual Review of Plant Biology, 68(1), 225-252. doi: 10.1146/annurev-arplant-043015-112232

Guo, W., Grewe, F., Fan, W., Young, G. J., Knoop, V., Palmer, J. D., and Mower, J. P. (2016). *Ginkgo* and *Welwitschia* mitogenomes reveal extreme contrasts in Gymnosperm mitochondrial evolution. Molecular Biology and Evolution, 33(6), 1448-1460. doi: 10.1093/molbev/msw024

Hanson, M. R. (1990). Plant mitochondrial mutations and male sterility. Annual Review of Genetics, 25(1), 461-486. doi: 10.1146/annurev.ge.25.120191.002333

Hao, Z., Lv, D., Ge, Y., Shi, J., Weijers, D., Yu, G., and Chen, J. (2020). RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. PeerJ Computer Science, 6, e251. doi: 10.7717/peerj-cs.251

Hasan, N., Choudhary, S., Naaz, N., Sharma, N., and Laskar, R. A. (2021). Recent advancements in molecular marker-assisted selection and applications in plant breeding programmes. Journal of Genetic Engineering and Biotechnology, 19(1). doi: 10.1186/s43141-021-00231-1

Heng, S., Wei, C., Jing, B., Wan, Z., Wen, J., Yi, B., Ma, C., Tu, J., and Fu, T. (2014). 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 Genomics, 15(1). doi: 10.1186/1471-2164-15-322

Hirose, T., Ideue, T., Wakasugi, T., and Sugiura, M. (1999). The chloroplast *infA* gene with a functional UUG initiation codon. FEBS Letters, 445, 169-172. doi:10.1016/s0014-5793(99)00123-4

Hoch, B., Maier, R. M., Appel, K., Igloi, G. L., and Kössel, H. (1991). Editing of a chloroplast mRNA by creation of an initiation codon. Nature, 353, 178-180. doi: 10.1038/353178a0

Holse, M., Husted, S., and Hansen, S. (2010). Chemical composition of marama bean (*Tylosema esculentum*) - a wild African bean with unexploited potential. Journal of Food Composition and Analysis, 23, 648-657. doi: 10.1016/j.jfca.2010.03.006

Hon, T., Mars, K., Young, G., Tsai, Y., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., Knapp, S. J., Ware, D., Shapiro, B., Peluso, P. R., and Kingan, S. B. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. Scientific Data, 7(1). doi: 10.1038/s41597-020-00743-4

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science and Engineering, 9(3), 90-95. doi: 10.1109/mcse.2007.55

Iannello, M., Bettinazzi, S., Breton, S., Ghiselli, F., and Milani, L. (2021). A naturally heteroplasmic clam provides clues about the effects of genetic bottleneck on paternal mtDNA. Genome Biology and Evolution, 13. doi: 10.1093/gbe/evab022

Jackson, J. C., Duodu, K. G., Holse, M., Lima de Faria, M. D., Jordaan, D., Chingwaru, W., et al. (2010). The morama bean (*Tylosema esculentum*): a potential crop for southern Africa. Advances in Food and Nutrition Research, 61, 187-246. doi: 10.1016/b978-0-12-374468-5.00005-2

Jansen, R. K., Raubeson, L. A., Boore, J. L., Depamphilis, C. W., Chumley, T. W., Haberle, R. C., et al. (2005). Methods for obtaining and analyzing whole chloroplast genome sequences. Methods in Enzymology, 395, 348-384. doi: 10.1016/s0076-6879(05)95020-9

Johansson, J. T. (1999). There large inversions in the chloroplast genomes and one loss of the chloroplast gene *rps16* suggest an early evolutionary split in the genus *Adonis (Ranunculaceae)*. Plant Systematics and Evolution, 218, 133-143. doi:10.1007/bf01087041

Johnson, L. B., and Palmer, J. D. (1989). Heteroplasmy of chloroplast DNA in *Medicago*. Plant Molecular Biology, 12, 3-11. doi: 10.1007/bf00017442

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T. L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Research, 36(Web Server), W5-W9. doi: 10.1093/nar/gkn201

Jung, J., Kim, J. M., Jeong, Y., and Yi, G. (2018). AGORA: organellar genome annotation from the amino acid and nucleotide references. Bioinformatics, 34(15), 2661-2663. doi: 10.1093/bioinformatics/bty196

Kang, Y. J., Kim, S. K., Kim, M. Y., Lestari, P., Kim, K. H., Ha, B. K., Jun, T. H., et al. (2014). Genome sequence of mungbean and insights into evolution within *Vigna* species. Nature Communications, 5(1). doi: 10.1038/ncomms6443

Kazakoff, S. H., Imelfort, M., Edwards, D., Koehorst, J., Biswas, B., Batley, J., et al. (2012). Capturing the biofuel wellhead and powerhouse: the chloroplast and mitochondrial genomes of the leguminous feedstock tree *Pongamia pinnata*. PLoS ONE, 7(12), e51687. doi: 10.1371/journal.pone.0051687

Keeling, P. J., and Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics, 9(8), 605-618. doi: 10.1038/nrg2386

Keith, M., and Renew, A. (1975). Notes on some Edible wild plants found in the Kalahari. Koedoe, 18(1). doi: 10.4102/koedoe.v18i1.911

Kempken, F., and Pring, D. R. (1999). Male sterility in higher plants - fundamentals and applications. Progress in Botany, 60, 139-166.

Khan, A., Suleman, M., Baqi, A., Samiullah, and Ayub, M. (2018). Phytochemicals and their role in curing fatal diseases: a review. Pure and Applied Biology, 7, 343-354. doi: 10.19045/bspab.2018.700193

Kim, Y., and Cullis, C. (2017). A novel inversion in the chloroplast genome of marama (*Tylosema esculentum*). Journal of Experimental Botany, 68, 2065-2072. doi: 10.1093/jxb/erw500

Kitazaki, K., and Kubo, T. (2010). Cost of having the largest mitochondrial genome: evolutionary mechanism of plant mitochondrial genome. Journal of Botany, 2010, 1-12. doi: 10.1155/2010/620137

Kondo, R., Satta, Y., Matsuura, E. T., Ishiwa, H., Takahata, N., and Chigusa, S. I. (1990). Incomplete maternal transmission of mitochondrial DNA in *Drosophila*. Genetics, 126, 657-663. doi: 10.1093/genetics/126.3.657

Koren, S., Walenz, B. P., Berlin, K., Miller, J., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research, 27(5), 722-736. doi: 10.1101/gr.215087.116

Kozik, A., Rowan, B. A., Lavelle, D., Berke, L., Schranz, M. E., Michelmore, R. W., et al. (2019). The alternative reality of plant mitochondrial DNA: one ring does not rule them all. PLOS Genetics, 15(8), e1008373. doi: 10.1371/journal.pgen.1008373

Krzywinski, M., Schein, J., Birol, N., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645. doi: 10.1101/gr.092759.109

Kubo, T., and Mikami, T. (2007). Organization and variation of angiosperm mitochondrial genome. Physiologia Plantarum, 129(1), 6-13. doi: 10.1111/j.1399-3054.2006.00768.x

Kumar, S., Nei, M., Dudley, J., and Tamura, K. (2008). MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings in Bioinformatics, 9, 299-306. doi: 10.1093/bib/bbn017

Kvist, L., Martens, J., Nazarenko, A. Y., and Orell, M. (2003). Paternal leakage of mitochondrial DNA in the great tit (*Parus major*). Molecular Biology and Evolution, 20(2), 243-247. doi: 10.1093/molbev/msg025

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9, 357-359. doi: 10.1038/nmeth.1923

Lawlor, D. W., (2018). "*Marama bean (Tylosema esculentum): A review of morphological and physiological adaptations to environment, and crop potential" in A praise of Demeter*. *Studies in Honour of Professor A. J. Karamanos*, Vol. 412. Eds. E. Eleftherohorinos, Paplomatas, and G. Economou-Antonaka (Athens, Greece: Editions Papazissi).

Lee, C., Choi, I. S., Cardoso, D., de Lima, H. C., de Queiroz, L. P., Wojciechowski, M. F., et al. (2021). The chicken or the egg? Plastome evolution and an independent loss of the inverted repeat in papilionoid legumes. The Plant Journal, 107, 861-875. doi: 10.1101/2021.02.04.429812

Lei, W., Ni, D., Wang, Y., Shao, J., Wang, X., Yang, D., et al. (2016). Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of *Astragalus membranaceus*. Scientific Reports, 6. doi: 10.1038/srep21669

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27(21), 2987-2993. doi: 10.1093/bioinformatics/btr509

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, J., and Cullis, C. (2021). The multipartite mitochondrial genome of marama (*Tylosema esculentum*). Frontiers in Plant Science, 12. doi: 10.3389/fpls.2021.787443

Li, J., and Cullis, C. (2023). Comparative analysis of 84 chloroplast genomes of *Tylosema esculentum* reveals two distinct cytotypes. Frontiers in Plant Science, 13:1025408. doi: 10.3389/fpls.2022.1025408

Li, J., Li, J., Ma, Y., Kou, L., Wei, J., and Wang, W. (2022). The complete mitochondrial genome of okra (*Abelmoschus esculentus*): using nanopore long reads to investigate gene transfer from chloroplast genomes and rearrangements of mitochondrial DNA molecules. BMC Genomics, 23(1). doi: 10.1186/s12864-022-08706-2

Li, G., Wang, L., Yang, J., He, H., Jin, H., Li, X., Ren, T., et al. (2021). A high-quality genome assembly highlights rye genomic characteristics and agronomically

important genes. Nature Genetics, 53(4), 574-584. doi: 10.1038/s41588-021-00808-z

Liang, Z., Huang, P., Yang, J., and Rao, G. (2009). Population divergence in the amphicarpic species *Amphicarpaea edgeworthii* Benth. (Fabaceae): microsatellite markers and leaf morphology. Biological Journal of the Linnean Society. doi: 10.1111/j.1095-8312.2008.01154.x

Liu, Y., Zhang, X., Han, K., Li, R., Xu, G., Han, Y., Cui, F., Fan, S., Seim, I., Fan, G., Li, G., and Wan, S. (2021). Insights into amphicarpy from the compact genome of the legume *Amphicarpaea edgeworthii*. Plant Biotechnology Journal, 19(5), 952-965. doi: 10.1111/pbi.13520

Lössl, A. G., Adler, N., Horn, R., Frei, U., and Wenzel, G. (1999). Chondriome-type characterization of potato: Mt α, β, γ, δ, ε and novel plastid-mitochondrial configurations in somatic hybrids. Theoretical and Applied Genetics, 99(1-2), 1-10. doi: 10.1007/s001220051202

Luo, S., Valencia, C. A., Zhang, J., Lee, N., Slone, J., Gui, B., Wang, X., Li, Z., Dell, S., Brown, J., Chen, S. X., Chien, Y., Hwu, W., Fan, P., Wong, L. C., Atwal, P. S., and Huang, T. (2018). Biparental inheritance of mitochondrial DNA in humans. Proceedings of the National Academy of Sciences, 115(51), 13039-13044. doi: 10.1073/pnas.1810946115

The Legume Phylogeny Working Group (LPWG). (2017). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. TAXON, 66, 44-77. doi: 10.12705/661.3

Manchekar, M., Scissum-Gunn, K., Song, D., Khazi, F. R., McLean, S. L., and Nielsen, B. L. (2006). DNA recombination activity in soybean mitochondria. Journal of Molecular Biology, 356(2), 288-299. doi: 10.1016/j.jmb.2005.11.070

Manos, P. S., Cannon, C. H., and Oh, S. H. (2008). Phylogenetic relationships and taxonomic status of the paleoendemic Fagaceae of western North America: recognition of a new genus, *Notholithocarpus*. Madroño, 55, 181-190. doi: 10.3120/0024-9637-55.3.181

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). MUMmer4: a fast and versatile genome alignment system. PLOS Computational Biology, 14(1), e1005944. doi: 10.1371/journal.pcbi.1005944

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 27(6), 764-770. doi: 10.1093/bioinformatics/btr011

Maréchal, A., and Brisson, N. (2010). Recombination and the maintenance of plant organelle genome stability. New Phytologist, 186(2), 299-317. doi: 10.1111/j.1469-8137.2010.03195.x

Marlow, F. L. (2017). Mitochondrial matters: mitochondrial bottlenecks, self-assembling structures, and entrapment in the female germline. Stem Cell Research, 21, 178-186. doi: 10.1016/j.scr.2017.03.004

Martin, W. (2003). Gene transfer from organelles to the nucleus: frequent and in big chunks. Proceedings of the National Academy of Sciences, 100, 8612-8614. doi: 10.1073/pnas.1633606100

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S., Wicker, T., Radchuk, V., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. Nature, 544(7651), 427-433. doi: 10.1038/nature22043

McCauley, D. E., Bailey, M. F., Sherman, N. A., and Darnell, M. Z. (2005). Evidence for paternal transmission and heteroplasmy in the mitochondrial genome of *Silene vulgaris*, a *gynodioecious* plant. Heredity, 95, 50-58. doi: 10.1038/sj.hdy.6800676

Members of the Complex Trait Consortium. (2003). The nature and identification of quantitative trait loci: a community's view. Nature Reviews Genetics, 4, 911-916. doi: 10.1038/nrg1206

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature, 585(7823), 79-84. doi: 10.1038/s41586-020-2547-7

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. Bioinformatics, 34(13), i142–i150. doi: 10.1093/bioinformatics/bty266

Miller, A. J., and Gross, B. L. (2011). From forest to field: perennial fruit crop domestication. American Journal of Botany, 98(9), 1389-1414. doi: 10.3732/ajb.1000522

Miller-Messmer, M., Kühn, K., Bichara, M., Le, M., Ret, Imbault, P., and Gualberto, J. M. (2012). RecA-dependent DNA repair results in increased heteroplasmy of the

Arabidopsis mitochondrial genome. Plant Physiology, 159(1), 211-226. doi: 10.1104/pp.112.194720

Moner, A. M., Furtado, A., and Henry, R. J. (2020). Two divergent chloroplast genome sequence clades captured in the domesticated rice gene pool may have significance for rice production. BMC Plant Biology, 20(1). doi: 10.1186/s12870-020-02689-6

Mower, J. P., Sloan, D. B., and Alverson, A. J. (2012). Plant mitochondrial genome diversity: the genomics revolution. Plant Genome Diversity, 1, 123-144. doi:10.1007/978-3-7091-1130-7_9

Naish, M., Alonge, M., Wlodzimierz, P., Tock, A. J., Abramson, B. W., Schmücker, A., Mandáková, et al. (2021). The genetic and epigenetic landscape of the *Arabidopsis* centromeres. Science, 374(6569). doi: 10.1126/science.abi7489

NASA. (n.d.). *Drought harms corn crops in southern Africa*. NASA. Retrieved February 11, 2023, from https://www.earthobservatory.nasa.gov/images/144704/drought-harms-corn-crops-in-southern-africa

National Research Council. (1979). "Tropical legumes: resources for the future," In: *The National Research Council in 1979* (Washington, DC: The National Academies Press). doi: 10.17226/19836

National Research Council. (2006). *Lost Crops of Africa: Volume II: Vegetables.* National Academies Press.

Nepolo, E. (2010). *Assessment of genetic variations within and between populations of marama bean (Tylosema Esculentum (Burchell) Schreiber) based on microsatellites (SSRs) and intergenic spacer length variation markers in the Namibian germplasm.* Windhoek: University of Namibia.

Nepolo, E., Chimwamurombe, P. M., Cullis, C. A., and Kandawa-Schulz, M. A. (2010). Determining genetic diversity based on ribosomal intergenic spacer length variation in marama bean (*Tylosema esculentum*) from the Omipanda area, Eastern Namibia. African Journal of Plant Science, 4, 368-373. doi: 10.5897/AJPS.9000060

Nikiforova, S. V., Cavalieri, D., Velasco, R., and Goremykin, V. V. (2013). Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line. Molecular Biology and Evolution, 30(8), 1751-1760. doi: 10.1093/molbev/mst092

Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., Miga, K. H., Eichler, E. E., Phillippy, A. M., and Koren, S. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Research, 30(9), 1291-1305. doi: 10.1101/gr.263566.120

Oldenburg, D. J., and Bendich, A. J. (1996). Size and structure of replicating mitochondrial DNA in cultured tobacco cells. The Plant Cell, 447-461. doi: 10.1105/tpc.8.3.447

Omotayo, A. O., and Aremu, A. O. (2021). Marama bean [*Tylosema esculentum* (Burch.) A. Schreib.]: an indigenous plant with potential for food, nutrition, and economic sustainability. Food and Function, 12, 2389-2403. doi: 10.1039/d0fo01937b

Ono, Y., Asai, K., and Hamada, M. (2012). PBSIM: PacBio reads simulator-toward accurate genome assembly. Bioinformatics, 29(1), 119-121. doi: 10.1093/bioinformatics/bts649

Palmer, J. D. (1991). "Plastid chromosomes: structure and evolution" in *The molecular biology of plastids*. Eds. L. Bogorad, and I. K. Vasil (San Diego, CA: Academic Press), 5-53. doi: 10.1016/b978-0-12-715007-9.50009-8

Palmer, J. D., and Herbon, L. A. (1988). Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. Journal of Molecular Evolution, 28(1-2), 87-97. doi: 10.1007/bf02143500

Palmer, J. D., and Shields, C. R. (1984). Tripartite structure of the *Brassica campestris* mitochondrial genome. Nature, 307(5950), 437-440. doi: 10.1038/307437a0

Palmer J. D., and Thompson W.F. (1982). Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. Cell, 29, 537-550. doi: 10.1016/0092-8674(82)90170-2

Palmer, J. D., Adams, K. L., Cho, Y., Parkinson, C. L., Qiu, Y. L., and Song, K. (2000). Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proceedings of the National Academy of Sciences, 97(13), 6960-6966. doi: 10.1073/pnas.97.13.6960

Palmer, R., Sun, H., and Zhao, L. (2000b). Genetics and cytology of chromosome inversions in soybean germplasm. Crop Science, 40, 683-687. doi: 10.2135/cropsci2000.403683x

Ramsey, A. J., and Mandel, J. R. (2019). When one genome is not enough: organellar heteroplasmy in plants. Annual Review of Plant Biology, 2, 1-40. doi: 10.1002/9781119312994.apr0616

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nature Communications, 11(1). doi: 10.1038/s41467-020-14998-3

Robinson, J. A., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. Nature Biotechnology, 29(1), 24-26. doi: 10.1038/nbt.1754

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. Molecular Biology and Evolution, 34, 3299-3302. doi: 10.1093/molbev/msx248

Sanetomo, R., and Gebhardt, C. (2015). Cytoplasmic genome types of European potatoes and their effects on complex agronomic traits. BMC Plant Biology, 15(1). doi: 10.1186/s12870-015-0545-y

Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X., Huang, L., et al. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. Nucleic Acids Research, 47, W65-W73. doi: 10.1093/nar/gkz345

Siekevitz, P. (1957). Powerhouse of the cell. Scientific American, 197(1), 131-140. doi: 10.1038/scientificamerican0757-131

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210-3212. doi: 10.1093/bioinformatics/btv351

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, İ. (2009). ABySS: a parallel assembler for short read sequence data. Genome Research, 19, 1117-1123. doi: 10.1101/gr.089532.108

Sinou, C., Cardinal-McTeague, W., and Bruneau, A. (2020). Testing generic limits in Cercidoideae (Leguminosae): insights from plastid and duplicated nuclear gene sequences. TAXON, 69, 67-86. doi: 10.1002/tax.12207

Skippington, E., Barkman, T. J., Rice, D. W., and Palmer, J. D. (2015). Miniaturized mitogenome of the parasitic plant *Viscum scurruloideum* is extremely divergent and dynamic and has lost all *nad* genes. Proceedings of the National Academy of Sciences, 112(27). doi: 10.1073/pnas.1504491112

Sloan, D. B., Alverson, A. J., Chuckalovcak, J. P., Wu, M., McCauley, D. E., Palmer, J. D., et al. (2012). Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biology, 10(1), e1001241. doi: 10.1371/journal.pbio.1001241

Sloan, D. B., Müller, K., McCauley, D. E., Taylor, D., and Štorchová, H. (2012b). Intraspecific variation in mitochondrial genome sequence, structure, and gene

content in *Silene vulgaris*, an angiosperm with pervasive cytoplasmic male sterility. New Phytologist, 196(4), 1228-1239. doi: 10.1111/j.1469-8137.2012.04340.x

Smith, D. R. (2017). Does cell size impact chloroplast genome size? Frontiers in Plant Science, 8. doi: 10.3389/fpls.2017.02116

Smyda-Dajmund, P., Śliwka, J., Janiszewska, M., and Zimnoch-Guzowska, E. (2020). Cytoplasmic diversity of potato relatives preserved at Plant Breeding and Acclimatization Institute in Poland. Molecular Biology Reports, 47(5), 3929-3935. doi: 10.1007/s11033-020-05486-4

Song, Y., Feng, L., Alyafei, M. A. M., Jaleel, A., and Ren, M. (2021). Function of chloroplasts in plant stress responses. International Journal of Molecular Sciences, 22, 13464. doi: 10.3390/ijms222413464

Straub, S. C. K., Cronn, R., Edwards, C., Fishbein, M., and Liston, A. (2013). Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae). Genome Biology and Evolution, 5(10), 1872-1885. doi: 10.1093/gbe/evt140

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2017). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution, 4(1). doi: 10.1093/ve/vey016

Takenaka, M., Zehrmann, A., Verbitskiy, D., Härtel, B., and Brennicke, A. (2013). RNA editing in plants and its evolution. Annual Review of Genetics, 47(1), 335-352. doi: 10.1146/annurev-genet-111212-133519

Takundwa, M., Chimwamurombe, P. M., and Cullis, C. A. (2012). A chromosome count in marama bean (*Tylosema esculentum*) by Feulgen staining using garden pea (*Pisum sativum* L.) as a standard. Research Journal of Biology, 2, 177-181.

Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: molecular evolutionary genetics analysis version 11. Molecular Biology and Evolution, 38(7), 3022-3027. doi: 10.1093/molbev/msab120

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics, 25(1). doi: 10.1002/0471250953.bi0410s25

Telenti, A., Pierce, L. C. T., Biggs, W. H., Di Iulio, J., Wong, E. B., Fabani, M. M., Kirkness, E. F., Moustafa, A. A., Shah, N., Xie, C., Brewerton, S. C., Bulsara, N., Garner, C., Metzker, G., Sandoval, E., Perkins, B. A., Och, F. J., Turpaz, Y., and Venter, J. C. (2016). Deep sequencing of 10,000 human genomes. Proceedings of the National Academy of Sciences, 113(42), 11901-11906. doi: 10.1073/pnas.1613365113

Twyford, A. D., and Ness, R. W. (2016). Strategies for complete plastid genome sequencing. Molecular Ecology Resources, 17, 858-868. doi: 10.1111/1755-0998.12626

Unseld, M., Marienfeld, J. R., Brandt, P., and Brennicke, A. (1997). The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. Nature Genetics, 15, 57-61. doi: 10.1038/ng0197-57

van der Maesen, L. J. G. (2006). "*Tylosema esculentum*" in PROTA, eds. M. Brink and G. Belay (Netherlands: Earthprint Ltd).

Varré, J., D'Agostino, N., Touzet, P., Gallina, S., Tamburino, R., Cantarella, C., Ubrig, E., Cardi, T., Drouard, L., Gualberto, J. M., and Scotti, N. (2019). Complete sequence, multichromosomal architecture and transcriptome analysis of the *Solanum tuberosum* mitochondrial genome. International Journal of Molecular Sciences, 20(19), 4788. doi: 10.3390/ijms20194788

Vietmeyer, N. D. (1986). Lesser-known plants of potential use in agriculture and forestry. Science, 232(4756), 1379-1384. doi: 10.1126/science.232.4756.1379

Von Bubnoff, A. (2008). Next-generation sequencing: The race is on. Cell, 132(5), 721-723. doi: 10.1016/j.cell.2008.02.028

Wallace, D. C. (2005). A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary biology. Annual Review of Genetics, 39(1), 359-407. doi: 10.1146/annurev.genet.39.110304.095751

Wang, Y. H., Wicke, S., Wang, H., Jin, J. J., Chen, S. Y., Zhang, S. D., et al. (2018). Plastid genome evolution in the early-diverging legume subfamily Cercidoideae (Fabaceae). Frontiers in Plant Science, 9. doi: 10.3389/fpls.2018.00138

Ward, B. J., Anderson, R. H., and Bendich, A. J. (1981). The mitochondrial genome is large and variable in a family of plants (Cucurbitaceae). Cell, 25(3), 793-803. doi: 10.1016/0092-8674(81)90187-2

Wei, S., Wang, X., Bi, C., Xu, Y., Wu, D., and Ye, N. (2016). Assembly and analysis of the complete *Salix purpurea* L. (Salicaceae) mitochondrial genome sequence. SpringerPlus, 5(1), 1894. doi: 10.1186/s40064-016-3521-6

Wolfe, K. H., Li, W., and Sharp, P. M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proceedings of the National Academy of Sciences, 84(24), 9054-9058. doi: 10.1073/pnas.84.24.9054

Woloszynska, M. (2010). Heteroplasmy and stoichiometric complexity of plant mitochondrial genomes-though this be madness, yet there's method in't. Journal of Experimental Botany, 61(3), 657-671. doi: 10.1093/jxb/erp361

Woloszynska, M., Bocer, T., Mackiewicz, P., and Janska, H. (2004). A fragment of chloroplast DNA was transferred horizontally, probably from non-eudicots, to mitochondrial genome of *Phaseolus*. Plant Molecular Biology, 56(5), 811-820. doi: 10.1007/s11103-004-5183-y

Wu, Z., Cuthbert, J., Taylor, D., and Sloan, D. B. (2015). The massive mitochondrial genome of the angiosperm *Silene noctiflora* is evolving by gain or loss of entire chromosomes. Proceedings of the National Academy of Sciences, 112(33), 10185-10191. doi: 10.1073/pnas.1421397112

Wunderlin, R. P. (2010). Reorganization of the Cercideae (Fabaceae: Caesalpinioideae). Phytoneuron, 48, 1-5.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics, 13(1). doi: 10.1186/1471-2105-13-134

Ye, C., Hill, C. M., Wu, S., Ruan, J., and Ma, Z. (2016). DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Scientific Reports, 6. doi: 10.1038/srep31900

Zaegel, V., Guermann, B., Le, M., Ret, Andrés, C., Meyer, D., Erhardt, M., Canaday, J., Gualberto, J. M., and Imbault, P. (2007). The plant-specific ssDNA binding protein *OSB1* is involved in the stoichiometric transmission of mitochondrial DNA in *Arabidopsis*. The Plant Cell, 18(12), 3548-3563. Doi: 10.1105/tpc.106.042028

Zardoya, R. (2020). Recent advances in understanding mitochondrial genome diversity. F1000Research, 9, 270. doi: 10.12688/f1000research.21490.1

Zavodna, M., Bagshaw, A. P., Brauning, R., and Gemmell, N. J. (2014). The accuracy, feasibility and challenges of sequencing short tandem repeats using next-generation sequencing platforms. PLOS ONE, 9(12), e113862. doi: 10.1371/journal.pone.0113862

Zhang, H., Burr, S. P., and Chinnery, P. F. (2018). The mitochondrial DNA genetic bottleneck: inheritance and beyond. Essays in Biochemistry, 62, 225-234. doi: 10.1042/ebc20170096

Zhang, X., Chen, S., Shi, L., Gong, D., Zhang, S., Zhao, Q., Zhan, D., et al. (2021). Haplotype-resolved genome assembly provides insights into evolutionary history

of the tea plant *Camellia sinensis*. Nature Genetics, 53(8), 1250-1259. doi: 10.1038/s41588-021-00895-y

Zhang, Y., Zhang, A., Li, X., and Lu, C. (2020). The role of chloroplast gene expression in plant responses to environmental stress. International Journal of Molecular Sciences, 21, 6082. doi: 10.3390/ijms21176082

Zhong, Y., Chen, Y., Zheng, D., Pang, J., Liu, Y., Luo, S., Meng, S., Qian, L., Wei, D., Dai, S., and Zhou, R. (2022). Chromosomal-level genome assembly of the orchid tree *Bauhinia variegata* (Leguminosae; Cercidoideae) supports the allotetraploid origin hypothesis of *Bauhinia*. DNA Research, 29(2). doi: 10.1093/dnares/dsac012