The complete chloroplast genome and characteristics analysis of *Musa basjoo* Siebold

Fenxiang Liu¹ · Ali Movahedi²  · Wenguo Yang³ · Dezhi Xu¹ · Chuanbei Jiang⁴ · Jigang Xie¹ · Yu Zhang¹

Received: 25 June 2021 / Accepted: 7 September 2021 / Published online: 19 September 2021
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract

**Background** An ornamental plant often seen in gardens and farmhouses, *Musa basjoo* Siebold can also be used as Chinese herbal medicine. Its pseudostem and leaves are diuretic; its root can be decocted together with ginger and licorice to cure gonorrhea and diabetes; the decoct soup of its pseudostem can help relieve heat, and the decoct soup of its dried flower can treat cerebral hemorrhage. There have not been many chloroplast genome studies on *M. basjoo* Siebold.

**Methods and results** We characterized its complete chloroplast genome using Novaseq 6000 sequencing. This paper shows that the length of the chloroplast genome *M. basjoo* Siebold is 172,322 bp, with 36.45% GC content. *M. basjoo* Siebold includes a large single-copy region of 90,160 bp, a small single-copy region of 11,668 bp, and a pair of inverted repeats of 35,247 bp. Comparing the genomic structure and sequence data of closely related species, we have revealed the conserved gene order of the IR and LSC/SSC regions, which has provided a very inspiring discovery for future phylogenetic research.

**Conclusions** Overall, this study has constructed an evolutionary tree of the genus *Musa* species with the complete chloroplast genome sequence for the first time. As can be seen, there is no obvious multi-branching in the genus, and *M. basjoo* Siebold and *Musa itinerans* are the closest relatives.

**Keywords** *Musa basjoo* Siebold · Chloroplast genome · Comparative analysis · Phylogeny analysis

**Abbreviations**

| Abbreviation | Description                          |
|--------------|--------------------------------------|
| PCR          | Polymerase chain reaction quantitative |
| PE           | Paired-end                            |
| RSCU         | Relative synonymous codon usage       |
| ML           | Maximum likelihood                    |
| LSC          | Large single copy                     |
| SSC          | Short single copy                     |
| IRs          | Inverted repeats                      |

Fenxiang Liu and Ali Movahedi have contributed equally to this work.

Research Square has already preprinted this manuscript with a Doi: 10.21203/rs.3.rs-531360/v1 and License: This work is licensed under a CCBY 4.0 License.

---

¹ School of Business and Trade, Nanjing Vocational University of Industry Technology, Nanjing 210046, China
² Co-Innovation Center for Sustainable Forestry in Southern China, Key Laboratory of Forest Genetics & Biotechnology, Ministry of Education, College of Biology and the Environment, Nanjing Forestry University, Nanjing 210037, China
³ Department of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing 210023, China
⁴ Genepioneer Biotechnologies Inc., Nanjing 210023, China
nalenone, triterpenoids, xanthones, and alkaloids [10–12].

M. acuminate, M. Nana, and M. sapientum include phenylphe-

nalenone, triterpenoids, xanthones, and alkaloids [10–12].

Seventy-six reported main compounds of als, selenium, other trace elements, and various compounds.

dS Synonymous
dN Non Synonymous

ITS Internal transcribed spacer region

IRa Inverted repeats a

IRb Inverted repeats b

JLB Junction line between LSC and IRb

JSB Junction line between IRb and SSC

JSA Junction line between SSC and IRA

Ka/Ks The ratio between the rate of synonymous (Ks), and non-synonymous (Ka) substitutions

CBOL The Consortium for the barcode of life

GTR General time reversible

RAxML Random accelerated maximum likelihood

Introduction

Musa basjoo Siebold, a perennial herb of the family Musaceae belonging to the genus Musa, is one of the main tropical plants. It is mainly distributed in subtropical areas in China, such as Guizhou, Guangdong, Guangxi, Hainan, Sichuan, Yunnan, and Taiwan [1, 2]. More than 40 species of genus Musa can be found in the southeast of India and Thailand, slightly fewer in Indonesia [3]. Since ancient times, the genus Musa has been widely spread in China as a good food and medicine product. Its sweet and refreshing pulp makes it an excellent fruit with appetizing and digesting functions. Also, its flowers, leaves, and roots have high medicinal values. They are mainly used to treat rheumatism and other diseases in cardiovascular, cerebrovascular, digestive, and circulatory systems [4–9]. The pulp, flower, leaf, and root of genus Musa are rich in sugar, amino acid, cellulose, minerals, selenium, other trace elements, and various compounds. Seventy-six reported main compounds of M. acuminata, M. balbisiana, M. sapientum, and M. Nana include phenylphenalenone, triterpenoids, xanthones, and alkaloids [10–12].

Most family Musaceae species are similar in shape, but their origin, evolution, and phylogeny could remain controversial. Different scholars have identified Musa species based on morphological characteristics, physical and chemical analysis, tissue anatomy, and molecular markers, which are significant for Musa species classification. In recent years, DNA barcoding gains popularity as it is not affected by the external environment and can help quickly and accurately identify species [13, 14]. At present, Internal transcribed spacer region (ITS) 1 and 2 sequences, chloroplast matK, rbcL, rpoB, trnH-psbA, psbK-psbl, and atpB-rbcL DNA barcodes have been reported for molecular identification of species [15–18].

Chloroplast genome is the second-largest genome containing rich genetic information. The intron with a relatively fast mutation rate is mainly divided into a highly conserved chloroplast coding sequence. Barcode fragments with small molecular weight, simple structure, horotelic evolution, low mutation rate, and stable heredity have apparent advantages in determining phylogenetic, genetic, and homologous relationships among species. They are also used in species identification, molecular geology, and species origin research [19]. At present, chloroplast DNA barcoding has been widely used in plants. The combination of rbcL + psbA trnH is a universal DNA barcode for terrestrial plants [20]. CBOL Plant Working Group has proposed the rbcL + matK combination as the core barcode for terrestrial plant identification [21]. Plant chloroplast genome has incomparable advantages in the studies of phylogeny, population dynamics, and species evolution. It is suitable for plant taxonomy and adaptive evolution studies, especially those regarding interspecific identification and related species phylogeny [16, 22]. With the development of high-throughput DNA sequencing technology, an increasing number of chloroplast genome sequences are available, which provide essential references for the chloroplast genome research of M. basjoo Siebold.

In this study, the whole chloroplast genome sequence of M. basjoo Siebold was obtained by sequencing. The chloroplast genome of M. basjoo Siebold was assembled, annotated, and analyzed. The chloroplast genome sequences of M. basjoo Siebold were compared with other published chloroplast genomes of genus Musa.

The fundamental characteristics and variation patterns of M. basjoo Siebold chloroplast genome were investigated to compare the interspecific and intraspecies variation of the sequences and select the high variation sequence among species. The chloroplast phylogenetic analysis of representative medicinal plants of genus Musa could provide a reference for the classification and identification, conservation genetics, resource development, and utilization of M. basjoo Siebold.

This study used Illumina sequencing technology to display the whole chloroplast genome of M. basjoo Siebold and explore its relationship with other genus species. The generated results will shed light on the study of the genetic structure and phylogenetic process of the natural population of this species. They will also contribute to the understanding of the structural diversity of plastids and the phylogeny of Musaceae.

Materials and methods

Sampling, filtering of raw reads, and sequencing

We collected fresh leaves from one M. basjoo Siebold tree, which grew in the Botanical Garden of Medicinal Plants,
Xianlin Campus of Nanjing University of Chinese Medicine, China. Total genomic DNA was extracted by the modified CTAB extraction method [23]. DNA was sheared by compressed nitrogen atomization to produce a fragment with a length of 300 bp. The fragmented quality was checked on the Bioanalyzer 2100 (Agilent Technology). The short-insert libraries preparation and sequencing were performed by Genepioneer (Nanjing, China). Genomic DNA was sequenced on the Illumina Novaseq PE150 platform.

The fastp (version 0.20.0, https://github.com/OpenGene/fastp) software was used to filter the original data. The filtering criteria are as follows: (1) The sequencing adaptor and primer sequences in reads were removed; (2) The reads with an average mass value less than Q5 were filtered out; (3) The reads with the N number greater than five were weeded out.

After a series of quality controls like this, 16,108,597 high-quality raw reads with 150-bp paired-end (PE) were obtained. This method of connecting pair-end readings is adopted in order to improve assembly accuracy.

### Assembly, annotation, and analysis of the plastid genome sequences

After quality filtering, reads were mapped to the database of Musaceae species with a chloroplast genome available (NCBI download), using Bowtie2 v.2.2.6 to exclude reads of nuclear and mitochondrial origins. All putative chloroplast reads were mapped to the above reference sequence and then used for de novo assembly.

In reference to published Musaceae species, we used the chloroplast-like reads to assemble genomic sequence with NOVOPlasty [24], assembled parts of reads, and stretched it as much as possible until a circular genome was formed. The assembled chloroplast sequence was annotated by the CpGA-VAS [25]. In our study, we checked the annotation results by DOGMA and BLAST [26]. The circular gene maps of the *Musa basjoo* Siebold plastid genome were formed by OGDRAW [27]. The codon usage distribution was detected by the software CodonW (University of Texas, Houston, TX, USA) with the relative synonymous codon usage (RSCU) ratio. Ka/Ks value for each gene was calculated using the KaKs_calculator [28] with the following settings: genetic code Table 11 (bacterial and plant plastid code); method of calculation: MLWL. Microsatellite (Mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats) detection was performed by using MISA with thresholds of eight repeat units for mononucleotide SSRs, five repeat units for di- and trinucleotide SSRs, and three repeat units for tetra-, penta- and hexanucleotide SSRs. The identified long repeat regions and corresponding melt values were compared in Musaceae. The plastid genome structure of *M. basjoo* Siebold shows a typical quadripartite circular molecule (Fig. 1, including a large single copy (LSC; 90,160 bp and a single copy SSC; 11,668 bp), which is divided into a pair of inverted repeats (IRs; 35,247 bp) regions (Fig. 1, including a large single copy (LSC; 90,160 bp and a single copy SSC; 11,668 bp), which is divided into a pair of inverted repeats (IRs; 35,247 bp) regions (Fig. 1, including a large single copy (LSC; 90,160 bp and a single copy SSC; 11,668 bp), which is divided into a pair of inverted repeats (IRs; 35,247 bp) regions (Fig. 1, including a large single copy (LSC; 90,160 bp and a single copy SSC; 11,668 bp), which is divided into a pair of inverted repeats (IRs; 35,247 bp) regions (Fig. 1, including a large single copy (LSC; 90,160 bp and a single copy SSC; 11,668 bp), which is divided into a pair of inverted repeats (IRs; 35,247 bp) regions (Fig. 1). The plastid genome features and guanine-cytosine content of *M. basjoo* Siebold

The complete cp genome sequence of *M. basjoo* Siebold was compared with that of its related species of *M. acuminata* subsp (HF677508.1), *M. beccarii* (MK012089.1), *M. banksii* (MK012089.2), *M. itinerans* (MK012089.3), *M. textilis* (MK012089.4), *M. balbisiana* (MK012089.5), *M. itinerans* (MK012089.5), *M. balbisiana* var. balbisiana (MK012089.7), and *M. ornata* (MK012089.8) using Mauve [30, 31]. With *M. basjoo* Siebold as a reference, ten plastids were compared in Musaceae.

### Phylogenetic position analysis

We constructed phylogenies by the Maximum Likelihood method (ML) using 10 Musaceae species to study the coding region evolution. The default whole-genome analysis of the evolutionary tree was adopted by setting the same starting point for the ring sequence. Multiple sequence alignment was done with MAFFT software [32] across species. The data results could help construct the largest likelihood evolutionary tree, using trimAl, the RAxML v8.2.10 (https://cme.h-its.org/exelixis/software.html) software, the GTRGAMMA model, and rapid Bootstrap analysis (Bootstrap = 1000).

### Results

### Chloroplast genome features and guanine-cytosine content of *M. basjoo* Siebold

The complete cp genome sequence of *M. basjoo* Siebold is 172,322 bp in length. Nucleotide sequences in other species of *M. basjoo* Siebold range narrowly from 161,347 bp (M. textilis (NC_022926.1)) to 171,815 bp (M. itinerans) (Table S1) [33]. The plastid genome structure of *M. basjoo* Siebold shows a typical quadripartite circular molecule Fig. 1, including a large single copy (LSC; 90,160 bp and a short single copy SSC; 11,668 bp), which is divided into a pair of inverted repeats (IRs; 35,247 bp) regions (Fig. 1 and Table S1).

We annotated 139 different genes with the same arrangement order in *M. basjoo* Siebold, including 89 messenger RNAs (mRNA), 8 ribosomal RNAs (rRNA), 38 transfer RNAs (tRNA), and 4 pseudogenes (Table S1). The LSC
The region contains 60 protein-coding and 21 tRNA genes, while the SSC region contains 9 protein-coding genes and 1 tRNA gene. The IRa and IRb regions include 10 protein-coding genes, 8 tRNA genes, and 4 rRNA genes. In the anti-clockwise direction, the intermediate region from LSC to SSC is defined as IRa, and the intermediate region from SSC to LSC is IRb (Table 1 and Fig. 1).

The GC content accounts for 39.72% in the IRa and IRb regions, while the LSC and SSC regions GC content 34.70% and 30.22%, respectively. The total GC content of the chloroplast DNA sequences is 36.45%, almost consistent with other Musaceae species (Table 1 and Table S1). In contrast with the LSC and SSC regions, IR regions have higher percents of GC content, and the skewness proves to be an indicator of DNA leading chains, lagging chains, replication starting, and terminal points [34, 35]. The chloroplast genome of *M. basjoo* Siebold is AT-rich (63.55%), similar to other Musaceae species (Tables S1, S2); for example, *M.
itinerans (63.49%), M. balbisiana (63.24%) [36], M. balbisiana var. balbisiana (63.23%) [37], M. banksii (62.99%) [38], M. acuminata subsp. malaccensis (63.18%) [39], and M. ornata (63.18%) [37].

The genes of chloroplast mainly work for photosynthetic pathways and self-replication. In addition, there are some genes with other functions or unknown functions, which have been shown in Table 2. We found that four genes are duplicated in the IRa and IRb regions, including rpl2, ndhB, trnI-UGA, and trnA-UGC (Table 3), identical to those found in M. basjoo Siebold. M. basjoo Siebold plastid genome contains 23 gene introns, with each gene seeing one intron, except for the ycf3, rps12, and clpP genes containing two introns (Table 3). It was found that ycf3 is an essential substance for the stable accumulation of photosystem I complexes [40].

**Codon usage bias**

Each amino acid includes one to six codons because of the degeneracy of codons. Codon usage plays a key role in chloroplast evolution [41]. Studies have shown very different codon utilization rates via different species and organisms. Relative Synonymous Codon Use (RSCU) refers to the inequality of synonymous codon use. Moreover, the relative usage is considered a combination of natural selection, mutation, and genetic drift. We then calculated the RSCU frequency of the M. basjoo Siebold plastid genome (Table S3). All protein-coding genes have 29,249 codons, among which leucine contains 2931 codons, making up 10.02% of the total, as the most abundant amino acid in the M. basjoo Siebold plastid genome.

### Table 1

| Region     | Number of CDS | Number of tRNA | Number of rRNA | Total |
|------------|---------------|----------------|----------------|-------|
| LSC region | 60            | 21             | 0              | 81    |
| IRb region | 10            | 8              | 4              | 22    |
| SSC region | 9             | 1              | 0              | 10    |
| IRa region | 10            | 8              | 4              | 22    |

### Table 2

| Category          | Gene group | Gene name                                                                 |
|-------------------|------------|---------------------------------------------------------------------------|
| Photosynthesis    | Subunits of photosystem I | psaA, psaB, psaC, psaL, psaJ                                              |
|                   | Subunits of photosystem II | psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbK, psbL, psbM, psbN, psbT, psbZ |
|                   | Subunits of NADH dehydrogenase | ndhA*, ndhB*(2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH(2), ndhL, ndhJ, ndhK |
|                   | Subunits of cytochrome b/f complex | petA, petB*, petD*, petG, petL, petN                                      |
|                   | Subunits of ATP synthase | atpA, atpB, atpF*, atpH, atpI                                           |
|                   | Large subunit of rubisco | rbcL                                                                       |
|                   | Subunits photosynthetic reductase | –                                                                          |
| Self-replication  | Proteins of large ribosomal subunit | rpl14, rpl16*, rpl22(2), rpl23(2), rpl32, rpl33, rpl36                 |
|                   | Proteins of small ribosomal subunit | rps11, rps12**, rps14, rps15(2), rps16*, rps18, rps19(2), rps2, rps3, rps4, rps7(2), rps8 |
|                   | Subunits of RNA polymerase | rpoA, rpoB, rpoC1*, rpoC2                                                 |
|                   | Ribosomal RNAs | rnm16(2), rnm23(2), rnm4.5(2), rnm5(2)                                     |
|                   | Transfer RNAs | trnA-UGC*, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-UCC, trnG-UGU*, trnH-GUG(2), trnI-CAU (2), trnL-GAU*, trnK-UUU*, trnL-CAA (2), trnL-UAU*, trnL-UAG, trnM-CAU, trnN-GUU (2), trnP-UGU, trnQ-UUG, trnR-ACG (2), trnR-UCU, trnS-GCU, trnS-GGA, trnT-GGU, trnT-UGC, trnV-GAc (2), trnV-UAC*, trnW-CCA, trnY-GUA, trnM-CAU |
| Other genes       | Maturase | matK                                                                       |
|                   | Protease | clpP**                                                                     |
|                   | Envelope membrane protein | cemA                                                                       |
|                   | Acetyl-CoA carboxylase | accD                                                                       |
|                   | c-type cytochrome synthesis gene | ccsA                                                                       |
|                   | Translation initiation factor | infA                                                                       |
|                   | Other | –                                                                           |
| Genes of unknown function | Conserved hypothetical chloroplast ORF | #ycf15(2), #ycf68(2), ycf1(2), ycf2(2), ycf3**, ycf4                      |

Gene*: Gene with one intron, Gene**: Gene with two introns, #Gene: Pseudogene, Gene (2): Number of copies of multi-copy genes
with 2524 codons (8.63%), while cysteine accounts for only 1.16%. In addition, almost all RSCU values are greater than one in the A/U-ending codons, while all RSCU values less than one in the C/G-ending codons. Met (AUC) is encoded by only one codon, with no codon preference (Fig. 2 and Figure S1).

SSR and long repeat analysis

Simple sequence repeats (SSRs) have high intraspecific variability in the chloroplast genome and are often used as genetic markers in population genetics and evolutionary research. Our study analyzed the simple sequence repeats (SSRs) in the cp genomes of *M. basjoo* Siebold (Table S4, Fig. 3). We detected 246 SSRs and 6 kinds of SSRs: mono-nucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide in *M. basjoo* Siebold. The mono-nucleotide repeat is the most abundant (51.82%), contributing more to genetic variation than other SSRs. The mono-nucleotide A and T repeats account for 96.09%, the highest proportion in all SSRs identified. It was found that mono-nucleotide, dinucleotide, trinucleotide, tetranucleotide, and pentanucleotide repeats comprise higher levels of poly-A or poly-T, consistent with the overall A-T abundance of the chloroplast genome (82.59%). We found that plastid SSRs are usually made up of A and T and rarely contain tandem guanine (G) or cytosine (C) repeats [42]. This might be attributed to the fact that A-T transformation is more effortless than G-C transformation in the chloroplast genome [43]. Long repeats, unlike simple sequence repeats, are dispersed in the genome. In the research, we identified 372 long repeats in the *M. basjoo* Siebold chloroplast genome (Table S5, Fig. 4), including 219 forward (F) repeats, 139 palindromic (P) repeats, 12 reverse (R) repeats, and 2 complement (C) repeats. Most of the long repeats range from 30 to 214 bp, and intergenic spacer regions (IGS) have 110 repeats at most. *M. basjoo* Siebold has fewer SSRs and long repeats than other species.

KaKs Analysis of *M. basjoo* Siebold and *M. itinerans*

Previous studies have revealed that non-synonymous and synonymous nucleotide substitution patterns serve as critical markers in gene evolution. We compared the gene sequences with mafftv7.310 software and calculated the gene Ka/Ks value with the KaKs calculator [28]. In this research, the non-synonymous (Ka)/synonymous (Ks) substitution ratios of *rpoc2, rpoC1, rpoB, rpa* within *Musa basjoo* Siebold and *Musa invenans* (NC 035723.1) were calculated to study the functional limitation of four copies at the DNA sequence level (Table S6). The results indicated that the gene with the

| Table 3 | The length of exons and introns in genes with introns in the *Musa basjoo* Siebold plastid genome |
|---------|--------------------------------------------------------------------------------------------------|
|Gene     | Location | Exon I (bp) | Intron II (bp) | Exon II (bp) | Intron II (bp) | Exon III (bp) |
|---------|----------|-------------|----------------|--------------|----------------|----------------|
|trnK-UUU | LSC      | 37          | 2690           | 35           |                |
|rps16    | LSC      | 40          | 828            | 212          |                |
|trnG-UCG | LSC      | 23          | 698            | 48           |                |
|atpF     | LSC      | 144         | 843            | 411          |                |
|rpoC1    | LSC      | 435         | 701            | 1638         |                |
|ycf3     | LSC      | 132         | 751            | 228          | 736            | 153           |
|trnL-UAA | LSC      | 35          | 545            | 50           |                |
|trnV-UAC | LSC      | 38          | 588            | 37           |                |
|rps12    | IRa      | 114         | –              | 229          | 543            | 26             |
|clpP     | LSC      | 69          | 648            | 291          | 809            | 252           |
|petB     | LSC      | 6           | 811            | 648          |                |
|petD     | LSC      | 8           | 729            | 475          |                |
|rpl16    | LSC      | 9           | 1067           | 402          |                |
|rpl2     | IRb      | 387         | 659            | 432          |                |
|ndhB     | IRb      | 777         | 700            | 756          |                |
|rps12    | IRb      | 229         | –              | 26           | 543            | 114           |
|trnL-GAU | IRb      | 42          | 958            | 35           |                |
|trnA-UGC | IRb      | 38          | 801            | 35           |                |
|ndhA     | IRa      | 552         | 1085           | 540          |                |
|trnA-UGC | IRa      | 38          | 801            | 35           |                |
|trnL-GAU | IRa      | 42          | 958            | 35           |                |
|ndhB     | IRa      | 777         | 700            | 756          |                |
|rpl2     | IRa      | 387         | 659            | 432          |                |
Fig. 2 RSCU histogram of *Musa basjoo* Siebold. The blocks underneath represent all codons encoding each amino acid. The columns on the top represent the sums of RSCU values of all codons.

Fig. 3 Distribution of SSRs in the chloroplast genomes of *Musa basjoo* Siebold. Horizontal coordinates represent SSR duplicate units, and longitudinal coordinates represent the number of SSR types.
value of Ka/Ks less than 1 (Ka/Ks < 1) has a strong purification choice.

IRscope analysis of Musaceae

In this study, the contraction and expansion of the IR boundaries in 10 Musaceae genomes were visualized between four chloroplast genome regions (LSC/IRb/SSC/IRa). The plastid genome shows a circular structure. IR has four boundaries with LSC-IRb, namely IRb-SSC, SSC-IRa, IRa-LSC, and LSC-SSC. We compared IR/LSC and IR/SSC binding sites among Musaceae in detail. Our analysis unveiled that the lengths of LSC, IRa, IRb, and SSC regions in the plastids of ten species are similar. JLA (IRa-LSC) border lies between rps19 and psbA genes of Musaceae, similar to that of Musa balbisiana var. balbisiana and Musa ornata (Fig. 5). The boundary of JLB (LSC-IRb) is situated at the coding region between rpl22 and rps19 genes. The NdhA gene generally spans JSA (SSC-IRa) region. In the previously studied species, IR boundary displacement was relatively slight and reported only a minor number of genes. The NdhA gene was found at the boundary of JSA (SSC-IRa) and only showed irregular translocation. In this study, the ndhA sequence of Musa itinerans is the longest with 1191 bp, and the ndhH and ndhF genes deviate from JSB (IRb-SSC) region.

Pi analysis of nucleotide diversity

The nucleotide diversity (π) of the chloroplast genome shows the difference in nucleic acid sequences among different species. Highly variable regions provide potential molecular markers in population genetics (Fig. 6).

The pi analysis makes evident gene variation, and the SNP and indel are the global comparison points. Our results revealed the Pi values of matK, rps16, psaC, rpl16, ndhF, of the scattered repeat sequence. F represents forward repetition, P represents palindrome repetition, R represents reverse repetition, and C represents complementary repetition

Comparative analysis of chloroplasts in Musaceae

This research compared the M. basjoo Siebold chloroplast genome with nine other Musaceae species (Fig. 7). The mauve program was adopted to compare the whole genome of ten Musaceae species [44, 45]. The ten plastid lengths vary from 161,347 bp (M. textilis) to 172,322 bp (M. basjoo Siebold). M. basjoo Siebold cp genome is similar in gene order to the other Musaceae species analyzed (Fig. 7). As a result, Musaceae plants are highly conserved in plastid genome content, gene sequence, and genetic structure, without inversion or translocation in the species we analyzed.

Phylogenetic relationship analysis

Phylogenetic analysis in the light of chloroplast genome sequence is fundamental in tracing many plant species lineages [46, 47]. Our study selected the complete plastid genome sequences from ten Musaceae species to study the phylogenetic position of M. basjoo Siebold. The ten selected complete plastome sequences were aligned using MAFFT software [48]. ML analysis was carried out using the RAxML software [49], and most of the nodes had 100% boot support in our ML tree. GTR model and hill clipping algorithm were used to construct the evolutionary tree (Fig. 8).

We obtained a complete genome of M. basjoo Siebold chloroplast, which studied M. basjoo Siebold phylogeny in Musaceae. We used the complete chloroplast genomes of ten Musaceae species for multiple sequence alignment. The results showed that this evolutionary tree could be divided into three branches. M. basjoo Siebold and M. itinerans are the closest relatives within one sister group. M. balbisiana
and *M. textilis* belong to another sister group. In addition, *M. banksii*, *M. ornata*, *M. acuminata*, and *M. beccarii* find themselves in a third group.

**Discussion**

In this study, we used Illumina sequencing technology to sequence the plastid sequence of *M. basjoo* Siebold. Chloroplast genome analysis showed that the genome has a pair of inverted repeat regions (IRa and IRb) with quadrilateral structure, separated by a large single-copy region (LSC) and small single-copy region (SSC). The structure and organization of *M. basjoo* Siebold chloroplast genome are similar to other sequenced *M. chloroplast genomes*. The content of GC in the chloroplast genome of *M. basjoo* Siebold is 36.45%. In addition, *rps12* is a trans-spliced gene, which has also been reported in other species. The gene content and arrangement of the chloroplast genome of *M. basjoo* Siebold are similar to other chloroplast genome sequences of *Musa*. Some genes in *M. basjoo* Siebold chloroplast genome begin with ATC, ATA, and ATG codon reported in *Musa* chloroplast genome.

The cp microsatellites (cpSSRs) have often been used as molecular markers in evolutionary studies such as genetic

![Fig. 5](image_url) Comparison of the LSC, SSC, and IR regions in the plastomes of eight Musaceae species. The colored boxes denote the genes. The gaps between the genes and the boundaries are indicated by the base lengths (bp). The thin lines represent each area connection point, and the figures give the information of the genes near the connection points.
diversity, short repeats in a cp genome inherited by a single parent. The cp microsatellite analysis indicated 98 SSRs in the chloroplast genome of *M. basjoo* Siebold, among which mononucleotides A and T are the most. Poly A and T are the most abundant repetitive sequences in chloroplast genome of plants. Most cpSSRs are located in the non-coding regions.

Fig. 6 Line graph of the PI values of *Musa basjoo* Siebold genes. Horizontal coordinate represents gene names and longitudinal coordinates representing PI value.

Fig. 7 Mauve alignment of plastid genomes of 8 species of Musaceae. The *Musa basjoo* Siebold genome is put at the top as the reference genome. Local collinear blocks within each of the alignments are represented by blocks of the same color connected by lines. The rectangular blocks represent the similarity between genomes. The lines between the rectangular blocks represent the collinear relationship. The short squares represent the gene locations of the genomes. Among them, white represents CDs, green represents tRNA, and red represents rRNA. (Color figure online)
but few in the protein-coding gene regions. In this study, the detected microsatellite will contribute to the study of the evolution of *Musa* and the protection and identification of the genus.

The variation of chloroplast genome size is due to the contraction and expansion of inverted repeats (IRs). We observed the contraction and expansion of IRs region in the chloroplast genome of *M. basjoo* Siebold and other sequenced Musaceae. Among these 10 species, the boundaries of IR-SC region are different. According to the positions of rps19, rpl22, ndhH, ndhF, ndhA, psbA, and trnH, we identified some types of connections caused by the contraction and expansion of the inverted repeat region.

The dN/dS ratio, which is Synonymous (dS) and non-synonymous (dN) substitution rate, was used to evaluate the purification selection of protein-coding genes and the sequence difference. Our results showed that most of the gene sequences make little difference (dS < 0.1). The dN/dS analysis revealed that most protein-coding genes are under negative selection, with only a few under positive selection (dN/dS > 1). Other plastids display similar findings.

Plastid genome is an excellent resource to infer the relationship between evolution and phylogeny. Many studies have used plastid sequences to analyze phylogenetic relationships at different taxonomic levels. Previously, only a few genes were used to evaluate the phylogenetic relationship and tribe classification of *Musa*, but tribe classification still needs to be explained. This paper used the maximum parsimony method to reconstruct the phylogenetic relationships of chloroplast genomes of 10 species representing the four major tribes in *Musa*. Our phylogenetic tree shows the same topological structure with high-resolution values at the branches. In this study, ten *Musa* species were selected as research objects, which has confirmed that *M. basjoo* Siebold and *M. itinerans* are the closest relatives within one sister group.

### Conclusions

Chloroplast organelles are typical in green plants and other organisms. Simple, conservative, and easy to be rearranged, the genome is mainly used to analyze species origin and evolution [50]. In this study, we obtained and analyzed for the first time the complete chloroplast genome sequence of a Chinese traditional medicinal plant (*M. basjoo* Siebold) by using Illumina high-throughput sequencing technology. The genome sequencing, assembly, annotation, and comparative analysis revealed that *M. basjoo* Siebold cp genomes have a typical quadruple structure with a conserved arrangement. Their GC content, arrangement, and codon usage features are similar to those found in the cp genome of other Musaceae species. In addition, the analytical result (Ka/Ks > 1, as shown by matK, ycf2) indicated more substantial natural selection effects. The *M. basjoo* Siebold cp genome analysis revealed some exciting features, which can pave the way for a better understanding of the plant’s anti-resurrection ability. For the first time, we attempted to analyze genetic diversity using whole-genome information. We found these genes with high nucleotide diversity: matK, rps16, psaC, rpl16, ndhF, rpl36, rpl32, ccsA, accD, rps15, ycf1, and trnG-UCC, which can set the barcode for subsequent species identification analysis.

In conclusion, the current study has demystified the complete chloroplast genome of *M. basjoo* Siebold for the first time. As genus *Musa* and *zingiberales* plants generally grow in tropical areas, they share a close relationship. Previous research used *Musa* and *Zingiberales* species to construct the evolutionary tree to illustrate the relationship between the two genera. It is the first time that we have used *Musa* species alone, and by far the most genus *Musa* species, to conduct an evolutionary tree as a whole to determine whether there exist multiple sister groups among *Musa* species. The analysis found out mainly three such groups. *M. basjoo* and *M. itinerans* are the closest relatives within one sister group. *M. balbisiana* and *M. textilis* belong to another sister group. In addition, *M. banksii*, *M. ornata*, *M. acuminate*, and *M. beccarii* find themselves in a third group. This finding may well provide a relatively complete reference for future studies.

### Supplementary Information

The online version contains supplementary material available at [https://doi.org/10.1007/s11033-021-06702-5](https://doi.org/10.1007/s11033-021-06702-5).
Author contributions FL: Conceptualization, software, formal analysis, writing—original draft, visualization, and project administration; AM: Writing—review & editing, visualization, supervision, and funding acquisition; WT: Validation, writing—review & editing, visualization, supervision, and funding acquisition; DX: Review & editing, and data curation; CJ: Review & editing, and data curation; JX: Review & editing, and data curation; WT: Validation, writing—review & editing, visualization, supervision, and funding acquisition; AM: Review & editing, and data curation.

Funding The study was supported by the Start-up Foundation of Introducing Talents for Scientific Research, Nanjing Institute of Industry Technology (Grant No. 201050619YK701), and Nanjing Forestry University foundation (Grant No. 163108059).

Data availability The identified long repeat regions [29] are deposited in GenBank (login No. BankIt2410783 Musa basjoo_Siebold MW376865). The following supplementary files are available online: Table S1: Summary of the plastid genome features of the 10 Musaceae studied; Table S2: Simple sequence repeats (SSRs) in the Musa basjoo Siebold plastid genome; Table S3: Long repeat sequences in the Musa basjoo Siebold plastid genome.

Code availability Not applicable.

Declarations

Conflict of interest The authors have no conflict of interest.

References

1. Amano M, Sawada Y, Motohashi T, Miyata M, Yoshizawa T, Masuda S (1991) A consideration to the original home of Musa basjoo sieb. et zucc. J Agricul Sci 36:66
2. Yang L, Qun LU, Zhiyuan Z, Jiechun C (2013) Advances in studies on Musa basjoo Sieb. Et Zucc. Journal of Guangdong Pharmaceutical University
3. Pollux ÉK (2012) Musa (genus). Chrono Press, Cambridge
4. Balthasar S, Michaelis K, Dinauer N, von Briesen H, Kreuter J, Langer K (2005) Preparation and characterisation of antibody modified gelatin nanoparticles as drug carrier system for uptake in lymphocytes. Biomaterials 26:2723–2732
5. Gupta AK, Gupta M, Yarwood SJ, Curtis AS (2004) Effect of cellular uptake of gelatin nanoparticles on adhesion, morphology and cytoskeleton organisation of human fibroblasts. J Controll Release 95:197–207
6. Han S, Li M, Liu X, Gao H, Wu Y (2013) Construction of amphiphilic copolymer nanoparticles based on gelatin as drug carriers for doxorubicin delivery. Colloids Surf B 102:833–841
7. Kim DW, Kang JH, Oh DH, Yong CS, Choi H-G (2012) Development of novel surfactin-protein-loaded solid self-microemulsifying drug delivery system using gelatin as solid carrier. J Microencap-sul 29:323–330
8. Morimoto K, Chono S, Kosai T, Seki T, Tabata Y (2008) Design of cationic microspheres based on aminated gelatin for controlled release of peptide and protein drugs. Drug Delivery 15:113–117
9. Xu F, Wu H, Wang X, Yang Y, Wang Y, Qian H, Zhang Y (2014) RP-HPLC characterization of lupenone and β-sitosterol in rhizome Musae and evaluation of the anti-diabetic activity of lupenone in diabetic Sprague-Dawley rats. Molecules (Basel, Switzerland) 19:14114–14127
10. Otálvaro F, Görls H, Hölischer D, Schmitt B, Echeverri F, Quiñones W, Schneider B (2002) Dimeric phenylphenalenones from Musa acuminata and various haemodoraceae species crystal structure of anigorootin. Phytochemistry 60:61–66
11. Pascual-Villalobos MJ, Rodríguez B (2007) Constituents of Musa balbisiana seeds and their activity against Cryptolestes pusillus. Biochem Syst Ecol 35:11–16
12. Tamura KT (1998) Cycloartane triterpenes from the fruit peel of Musa sapientum. Phytochemistry (Oxford) 47:1107
13. Hsiao S-T, Chuang S-C, Chen K-S, Ho P-H, Wu C-L, Chen CA (2016) DNA barcoding reveals that the common cupped oyster in Taiwan is the Portuguese oyster Crassostrea angulata (Ostreoida; Ostreidae), not C. gigas. Sci Rep 6:34057
14. Jiao J, Huang W, Bai Z, Liu F, Ma C, Liang Z (2018) DNA barcoding for the efficient and accurate identification of medicinal polygons rhizoma in China. PLoS ONE 13:e0201015
15. Battini MA, Bourguiba H, Trifi-Farah N, Krichen L (2019) Molecular diversity and phylogeny of Tunisian Prunus armeniaca L. by evaluating three candidate barcodes of the chloroplast genome. Sci Hortic 245:99–106
16. He P, Ma Q, Dong M, Yang Z, Liu L (2019) The complete chloroplast genome of Leoncotea inerita and phylogeny of Berberidaceae. Mitochondrial DNA Part B 4:101–102
17. Terakami S, Matsumura Y, Kurita K, Kanamori H, Katayose Y, Yamamoto T, Katayama H (2012) Complete sequence of the chloroplast genome from peat (Pyrus pyrifolia): genome structure and comparative analysis. Tree Genet Genomes 8:841–854
18. Xue S, Shi T, Luo W, Ni X, Iqbal S, Ni Z, Huang X, Yao D, Shen Z, Gao Z (2019) Comparative analysis of the complete chloroplast genome among Prunus mume, P. armeniaca, and P. salicina. Hortic Res 6:89
19. Xu DH, Abe J, Kanazawa A, Gai JY, Shimamoto Y (2001) Identification of sequence variations by PCR-RFLP and its application to the evaluation of cpDNA diversity in wild and cultivated soybeans. Theor Appl Genet 102:683–688
20. Kress W, Erickson D (2007) A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. PLoS ONE 2:e508
21. Hollingsworth ML, Andra Clark A, Forrest LL, Richardson J, Pennington RT, Long DG, Cowan R, Chase MW, Gaudeul M, Hollingsworth PM (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Mol Ecol Resour 9:439–457
22. Henriquez CL, Abdullah AL, Carlsen MM, Zuluaga A, Croat TB, McKain MR (2020) Evolutionary dynamics of chloroplast genomes in subfamily Aroideae (Araceae). Genomics 112:2349–2360
23. Doyle J, Doyle JL (1986) A rapid DNA isolation procedure from small quantities of fresh leaf tissues. Phytochem Bull 19
24. Dierckxsens N, Mardulyn P, Smits G (2016) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw955
25. Chang L, Linchun S, Yingjie Z, Haimei C (2012) CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics 13:1–7
26. Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics (Oxford, England) 20:3252–3255
27. Lohse M, Drechsel O, Kahlau S, Bock R (2013) OrganelarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. https://doi.org/10.1093/nar/gkt289
28. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J (2010) KaKs Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genom Proteom Bioinform 8:77–80
29. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29:4633–4642
30. Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S (2003) Glocal alignment: finding rearrangements during alignment. Bioinformatics (Oxford, England) 19(Suppl 1):i54–i62
31. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. Nucleic Acids Res 32:W273–279
32. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274
33. Zhang Y-Y, Liu F, Tian N, Che J-R, Sun X-L, Lai Z-X, Cheng C-Z (2019) Characterization of the complete chloroplast genome of sanming wild banana (Musa itinerans) and phylogenetic relationships. Mitochondrial DNA Part B 4:2614–2616
34. Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13:660–665
35. Saina JK, Gichira AW, Li Z-Z, Hu G-W, Wang Q-F, Liao K (2018) The complete chloroplast genome sequence of Dodonaea viscosa: comparative and phylogenetic analyses. Genetica 146:101–113
36. Wang Z, Miao H, Liu J, Xu B, Yao X, Xu C, Zhao S, Fang X, Jia C, Wang J, Zhang J, Li J, Xu Y, Wang J, Ma W, Wu Z, Yu L, Yang Y, Liu C, Guo Y, Sun S, Baurens F-C, Martin G, Salman F, Garsmeur O, Yahiaoui N, Hervouet C, Rouard M, Laboureau N, Habas R, Ricci S, Peng M, Guo A, Xie J, Li Y, Ding Z, Yan Y, Tie W, D’Hont A, Hu W, Jin Z (2019) Musa balbisiana genome reveals subgenome evolution and functional divergence. Nature Plants 5:810–821
37. Niu Y-F, Gao C-W, Liu J (2018) The complete chloroplast genome sequence of wild banana, Musa balbisiana variety ’Pisang Klutuk Wulung’. Mitochondrial DNA Part B 3:460–461
38. Liu J, Gao C-W, Niu Y-F (2019) Complete chloroplast genome sequence of Musa banksii and its phylogenetic analysis. Mitochondrial DNA Part B 4:789–790
39. Munir A, Mehmood A, Azam S (2016) Structural and Function Prediction of Musa acuminata subsp. Malaccensis protein. Int J Bioautomation 20:19–30
40. Boudreau E, Takahashi Y, Lemieux C, Turmel M, Rochaix JD (1997) The chloroplast ycf3 and ycf4 open reading frames of Chlamydomonas reinhardtii are required for the accumulation of the photosystem I complex. EMBO J 16:6095–6104
41. Ivanova Z, Sahlok G, Daskalova E, Zahanova G, Apostolova E, Yahubyan G, Baev V (2017) Chloroplast genome analysis of resurrection tertiary relict Haberlea rhodopenis highlights genes important for desiccation stress response. Front Plant Sci. https://doi.org/10.3389/fpls.2017.00204
42. Shen X, Wu M, Liao B, Liu Z, Bai R, Xiao S, Li X, Zhang B, Xu J, Chen S (2017) Complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant Artemisia annua. Molecules (Basel, Switzerland). https://doi.org/10.3390/molecules22081330
43. Xie DF, Yu Y, Deng YQ, Li J, Liu HY, Zhou SD, He XJ (2018) Comparative analysis of the chloroplast genomes of the Chinese endemic genus Urophylla and their contribution to chloroplast phylogeny and adaptive evolution. Int J Mol Sci. https://doi.org/10.3390/ijms19071847
44. Darling AC, Mau B, Blattner FR, Perna NT (2004) MAUve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403
45. Doose D, Grand C, Lesire C (2017) MAUVE runtime: a component-based middleware to reconfigure software architectures in real-time. In 2017 First IEEE International Conference on Robotic Computing (IRC), 10 Apr 2017 pp 208–211. IEEE
46. Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Muller KF, Gisgener-Bellian M, Haberle RC, Hansen AK, Chumley TW, Lee S-B, Peery R, McNeal JR, Kuehl JV, Boore JL (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104:19369–19374
47. Xu C, Dong W, Li W, Lu Y, Xie X, Jin X, Shi J, He K, Suo Z (2017) Comparative analysis of six lagerstroemia complete chloroplast genomes. Front Plant Sci 8:15
48. Katoh K, Standley D (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780
49. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313
50. Liu L-X, Li R, Worth JRP, Li X, Li P, Cameron KM, Fu C-X (2017) The complete chloroplast genome of Chinese bayberry (Morella rubra, Myricaceae) implications for understanding the evolution of fagales. Front Plant Sci. https://doi.org/10.3389/fpls.2017.00968

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.