The Complete Chloroplast Genome and Characteristics Analysis of Musa basjoo Siebold

Fenxiang Liu
Nanjing Institute of Industry Technology
Ali Movahedi (ali_movahedi@njfu.edu.cn)
Nanjing Forestry University https://orcid.org/0000-0001-5062-504X
Wenguo Yang
Nanjing University of Chinese Medicine
Dezhi Xu
Nanjing Institute of Industry Technology
Chuanbei Jiang
Genepioneeer Biotechnologies
Jigang Xie
Nanjing Institute of Industry Technology
Yu Zhang
Nanjing Institute of Industry Technology

Research Article

Keywords: Musa basjoo Siebold, Chloroplast genome, Comparative analysis, Phylogeny analysis

Posted Date: June 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-531360/v1

Abstract

Musa basjoo Siebold is an ornamental plant often seen in gardens and farmhouses. It can also be used as a kind of Chinese herbal medicine. The decoct soup of its pseudo stem can help relieve the heat. Its pseudo stem and leaves are diuretic. The decoct soup of its dried flower can treat cerebral hemorrhage. Its root can be decocted together with ginger and licorice to cure gonorrhea and diabetes.

The chloroplast genome study on *Musa basjoo* Siebold is rarely seen. This paper showed that the length of the chloroplast genome *Musa basjoo* Siebold was 172,322 bp, with 36.45% GC content. *Musa basjoo* Siebold included 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 revealed the conserved gene order of the IR and LSC/SSC regions, which provided a very inspiring discovery for future phylogenetic research.

Overall, this study is the first time an evolutionary tree of the genus *Musa* species has been constructed with the complete chloroplast genome sequence. We can see that there is no obvious multi-branching in the genus, and *Musa basjoo* Siebold, and *Musa itinerans* are the closest relatives.

Key Message

We analyzed the whole chloroplast genome of *Musa basjoo* Siebold to achieve the IR and LSC/SSC regions, providing a very inspiring discovery for genetic researches of the natural population of this species.

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, Guizhou, Guangdong, Guangxi, Hainan, Sichuan, Yunnan, and Taiwan (Yang et al. 2013; Amano et al. 1991). In addition, more than 40 species of genus *Musa* can be found in the southeast of India and Thailand, followed by Indonesia (Pollux 2012). Since ancient times, the genus *Musa* has been widely spread in China as a good food and medicine product. Its pulp is an excellent fruit, sweet and refreshing, bearing the appetizing and digesting function. Also, its flowers, leaves, and roots have high medicinal values. They are mainly used to treat rheumatism and other cardiovascular, cerebrovascular, digestive, and circulatory systems (Xu et al. 2014; Morimoto et al. 2008; Balthasar et al. 2005; Gupta et al. 2004; Han et al. 2013; Kim et al. 2012). 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. acuminate*, *M. balbisiana*, *M. sapientum*, and *M. Nana* are including phenylphenalenone, triterpenoids, xanthones, and alkaloids (Otálvaro et al. 2002; Pascual-Villalobos and Rodríguez 2007; Tamura 1998).

Most family Musaceae species are similar in shape, but their origin, evolution, and phylogeny have been controversial. Therefore, different scholars have identified *Musa* species based on morphological

characteristics, physical and chemical analysis, tissue anatomy, and molecular markers, which are of great significance for *Musa* species classification. In recent years, DNA barcoding is quite popular as it is not affected by the external environment and can help quickly and accurately identify species (Hsiao et al. 2016; Jiao et al. 2018). At present, Internal transcribed spacer region (ITS) 1 and 2 sequences, chloroplast *matK*, *rbcL*, *rpoB*, *trnH-psbA*, *psbK-psbI*, and *atpB-rbcL* DNA barcodes have been reported for molecular identification of species (Xue et al. 2019; He et al. 2019; Batnini et al. 2019; Terakami et al. 2012).

Chloroplast genome is the second-largest genome, which contains rich genetic information. It is mainly divided into highly conserved chloroplast coding sequence, intron with relatively fast mutation rate. Barcode fragments, with small molecular weight, simple structure, horotelic evolution, low mutation rate, and stable heredity, have obvious advantages in determining phylogenetic, genetic, and homologous relationships among species. They are also used in species identification, molecular geology, and species origin research (Xu et al. 2001). 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 (Kress and Erickson 2007). CBOL Plant Working Group proposed the rbcL + matK combination as the core barcode for terrestrial plant identification (Hollingsworth et al. 2009). 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 (Henriquez et al. 2020; He et al. 2019). 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 *Musa basjoo* Siebold.

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

The fundamental characteristics and variation patterns of *Musa 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* reveals the reference for the classification and identification, conservation genetics, resource development, and utilization of *Musa basjoo* Siebold.

This study used Illumina sequencing technology to display the whole chloroplast genome of *Musa basjoo* Siebold and explore its relationship with other genus species. The generated results will help study 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.

Methods

Sampling, filtering of raw reads, and sequencing

We collected fresh leaves from one *Musa basjoo* Siebold tree, which grew in the Botanical Garden of Medicinal Plants, Xianlin Campus of Nanjing University of Chinese Medicine, China. Total chloroplast DNA was extracted by the modified CTAB extraction method (Doyle and Doyle 1986). After the genomic DNA was tested, the DNA was fragmented by mechanical interruption (Ultrasonic). Then fragments were purified, terminal repaired, 3' end plus A, and the sequencing pair. The fragment size was selected by agarose gel electrophoresis, and the PCR was amplified to form a sequencing library. The library was first constructed for quality inspection, and Illumina then checked the qualified library with quality inspection. After that, the DNA was used for quality control. Finally, the Novaseq platform was used for sequencing, and the reading length was PE150 reads.

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 first removed. (2) The reads with an average mass value less than Q5 were then filtered out. (3) The reads with the N numbers greater than five were finally 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. In order to improve the assembly accuracy, this method of connecting pair-end readings is adopted.

Assembly, annotation, and analysis of the plastid genome sequences

This article assembled the chloroplast genome by SPAdes (Bankevich et al. 2012) (http://cab.spbu.ru/software/spades/) and used 55, 87 121 kmer, respectively. Thus, the assembly did not depend on the reference genome. However, given the second-generation sequencing features and the genome-specific structure, the complete circular genome sequence cannot be directly obtained by one-time splicing. Thus, some other strategies would be used to obtain a whole circular genome sequence.

We referred to published Musaceae species and used the chloroplast-like reads to assemble genomic sequence with NOVOPlasty (Dierckxsens et al. 2016), assembled parts of reads, and stretched as much as possible until a circular genome was formed. The assembled chloroplast sequence was annotated by the CpGAVAS (Chang et al. 2012). Our study checked the results of the annotation by DOGMA and BLAST (Wyman et al. 2004). The circular gene maps of the *Musa basjoo* Siebold plastid genome were formed by OGDRAW (Lohse et al. 2013). The codon usage, GC content, and relative synonymous codon usage (RSCU) of the complete chloroplast genome were analyzed. The identified long repeat regions and corresponding genome coordinates were checked and annotated with the REPUTER tool (Kurtz et al. 2001) and deposited in GenBank (login No. BankIt2410783 *Musa_basjoo*_Siebold MW376865).

Genome Comparison

The complete cp genome of *Musa basjoo* Siebold was compared with that of its related species of *Musa acuminata* subsp (HF677508.1), *Musa beccarii* (MK012089.1), *Musa banksii* (MK012089.2), *Musa itinerans* (MK012089.3), *Musa textilis* (MK012089.4), *Musa balbisiana* (MK012089.5), *Musa itinerans* (MK012089.6), *Musa balbisiana* var. balbisiana (MK012089.7), and *Musa ornata* (MK012089.8) using Mauve (Frazer et al. 2004; Brudno et al. 2003). With *Musa 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. First, the default whole-genome analysis of the evolutionary tree was adopted by setting the same starting point for the ring sequence. Then, multiple sequence alignment was done with MAFFT software (Nguyen et al. 2015) 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 of Musa basjoo Siebold

The complete cp genome sequence of *Musa basjoo* Siebold is 172,322 bp in length. Nucleotide sequences in other species of *Musa basjoo* Siebold range narrowly from 161,347 bp (*Musa textilis* (NC_022926.1)) to 171,815 bp (*Musa itinerans*) (Table S1) (Zhang et al. 2019). The plastid genome structure of *Musa 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 *Musa basjoo* Siebold, including 89 messenger RNAs (mRNA), 8 ribosomal RNAs (rRNA), 38 transfer RNAs (tRNA), and 4 pseudogenes (Table S1). The LSC 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 accounted for 39.72% in both the IRa and IRb regions, while the LSC and SSC regions GC contents accounted for 34.70% and 30.22%, respectively. The total GC content of the chloroplast DNA sequences was 36.45%, almost consistent with the other Musaceae species (Table 1 and Table S1). In contrast with the LSC and SSC regions, IR regions had higher percents of GC content, and the skewness proved to be an indicator of DNA leading chains, lagging chains, replication starting, and terminal points (Saina et al. 2018; Lobry 1996). The chloroplast genome of *Musa basjoo* Siebold was AT-rich (63.55%), which was also similar to the other Musaceae species (Table 2 and Table S1); for example, *Musa*

itinerans (63.49%), *Musa balbisiana* (63.24%) (Wang et al. 2019), *Musa balbisiana* var. *balbisiana* (63.23%) (Niu et al. 2018), *Musa banksii* (62.99%) (Liu et al. 2019), *Musa acuminata* subsp. *malaccensis* (63.18%) (Munir et al. 2016), and *Musa ornata* (63.18%) (Niu et al. 2018).

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 3. We found that four genes are duplicated in the IRa and IRb regions, including *rpl2*, *ndhB*, *trnI-GAU*, and *trnA-UGC* (Table 4), which were identical to those found in *Musa basjoo* Siebold. *Musa basjoo* Siebold plastid genome contained 23 gene introns, while each gene includes one intron, the *ycf3*, *rps12*, and *clpP* genes containing two introns (Table 4). It was found that *ycf3* was an essential substance for the stable accumulation of photosystem I complexes (Boudreau et al. 1997).

Codon Usage Bias

Each amino acid includes one to six codons because of the degeneracy of codons. Therefore, codon usage plays a key role in chloroplast evolution (Ivanova et al. 2017). It has been 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. Using the coding sequence, we calculated the RSCU frequency of the *Musa basjoo* Siebold plastid genome (Table 5). All protein-coding genes had 29,249 codons, of which leucine had 2,931 codons, making up 10.02% of the total, as the most abundant amino acid in the *Musa basjoo* Siebold plastid genome. The next was isoleucine, with 2,524 codons (8.63%), while cysteine accounted for only 1.16%. In addition, almost all RSCU values greater than one were in the A/U-ending codons, while all RSCU values less than one were in the C/G-ending codons. As Met (AUC) was encoded by only one codon, it had no codon preference (Fig. 2 and Fig. 3).

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 *Musa basjoo* Siebold (Table S2, Fig. 4). We detected 246 SSRs and 6 kinds of SSRs, namely, mono-nucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide in *Musa basjoo* Siebold. The mono-nucleotide repeat was the most abundant (51.82%), contributing more to genetic variation than other SSRs. The mono-nucleotide A and T repeats were accounted 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, in consistence with the overall A-T abundance of chloroplast genome (82.59%). We found that plastid SSRs were usually made up of A and T and rarely contain tandem guanine (G) or cytosine (C) repeats (Shen et al. 2017). This might be attributed to the fact that A-T transformation is more effortless than G-C transformation in the chloroplast genome (Xie et al. 2018). Long repeats, unlike simple sequence repeats, are dispersed in the genome. In

the research, we identified 372 long repeats in the *Musa basjoo* Siebold chloroplast genome (Table S3, Fig. 5), including 219 forward (F) repeats, 139 palindromic (P) repeats, 12 reverse (R) repeats, and 2 complement (C) repeats. Most of the long repeats ranged from 30 bp to 214 bp, and intergenic spacer regions (IGS) had 110 repeats at most. *Musa basjoo* Siebold had fewer SSRs and long repeats than other species.

KaKs Analysis of Musa basjoo **Siebold and** Musa 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 (Wang et al. 2010). In this research, the non-synonymous (Ka)/synonymous (Ks) substitution ratios of *rpoc2*, *rpoC1*, *rpoB*, *ropA* within *Musa basjoo* Siebold and *Musa invenans* (NC 035723.1) were calculated to study the functional limitation of the four copies at the DNA sequence level (Table 6). The results indicated that the gene with the value of Ka/Ks less than 1 (Ka/Ks < 1) had a strong purification choice.

IRscope analysis of Musaceae

In this study, The contraction and expansion of the IR boundaries in 10 Musaceae genomes visualized between the four regions of the chloroplast genome (LSC/IRb/SSC/IRa). The plastid genome had a circular structure. IR had 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 showed that the lengths of LSC, IRa, IRb, and SSC regions in the plastids of ten species were similar. JLA (IRa-LSC) border lay between *rps19* and *psbA* genes of Musaceae, similar to that of *Musa balbisiana* var. *balbisiana* and *Musa ornata*. The boundary of JLB (LSC-IRb) was situated at the coding region between *rpl22* and *rps19* genes. The *NdhA* gene generally spanned 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* was the longest with 1,191 bp, and the *ndhH* and *ndhF* genes deviated from JSB (IRb-SSC) region.

Pi analysis of nucleotide diversity

The nucleotide diversity (π) of the chloroplast genome shows the difference of nucleic acid sequences among different species. Thus, highly variable regions provide potential molecular markers in population genetics (Fig. 7).

Pi analysis shows gene variation. The SNP and indel are the global comparison points. Our results revealed that the Pi values of *matK*, *rps16*, *psaC*, *rpl16*, *ndhF*, *rpl36*, *rpl32*, *ccsA*, *accD*, *rps15*, *ycf1*, and *trnG-UCC* genes were higher, as to be seen in the subsequent barcodes.

Comparative analysis of chloroplasts in Musaceae

This research compared the *Musa basjoo* Siebold chloroplast genome with nine other Musaceae species (Fig. 8). The mauve program was adopted to compare the whole genome of ten Musaceae species (Darling et al. 2004; Doose et al. 2017). The ten plastid lengths varied from 161,347bp (*Musa textilis*) to 172,322bp (*Musa basjoo* Siebold). The *Musa basjoo* Siebold cp genome was similar in gene order to the other Musaceae species analyzed (Fig. 8). As a result, Musaceae plants were 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 (Jansen et al. 2007; Xu et al. 2017). Our study selected the complete plastid genome sequences from ten Musaceae species to study the phylogenetic position of *Musa basjoo* Siebold. The ten selected complete plastome sequences were aligned using MAFFT software (Katoh and Standley 2013). Maximum likelihood (ML) analysis was carried out using the RAxML software (Stamatakis 2014), and most of the nodes have 100% boot support in our ML tree. GTR model and hill clipping algorithm were used to construct the evolutionary tree (Fig. 9).

We obtained a complete genome of *Musa basjoo* Siebold chloroplast genome, which provided information for studying *Musa basjoo* Siebold phylogeny in Musaceae. Furthermore, we used the chloroplast completed genomes of ten Musaceae species for multiple sequence alignment. The results showed that this evolutionary tree was divided into three branches. The *Musa basjoo* Siebold and *Musa itinerans* are the closest relatives within one sister group. The *Musa balbisiana* and *Musa textilis* belong to another sister group. In contrast, *Musa banksii*, *Musa ornata*, *Musa acuminata*, and *Musa beccarii* belong to a third group.

Discussion

In this study, we used Illumina sequencing technology to sequence the plastid sequence of *Musa basjoo* Siebold. Chloroplast genome analysis showed that the genome had 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 *Musa basjoo* Siebold chloroplast genome are similar to other sequenced *Musa* chloroplast genomes. The content of GC in the chloroplast genome of *Musa basjoo* Siebold was 36.45%. In addition, *rps12* was considered to be a trans-spliced gene, which had also been reported in other species. The gene contents and arrangement of the chloroplast genome of *Musa basjoo* Siebold are similar to other chloroplast genome sequences of *Musa*. Some genes in *Musa basjoo* Siebold chloroplast genome begin with *ATC*, *ATA*, and *ATG* codon reported in *Musa* chloroplast genome.

The cp microsatellites (cpSSRs) are often used as molecular markers in evolutionary studies, such as genetic diversity, short repeats in cp genome inherited by a single parent. The cp microsatellite analysis shows 98 SSRs in the chloroplast genome of *Musa 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, 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 *Musa basjoo* Siebold and other sequenced Musaceae. Among the 10 species, the boundaries of IR-SC region were different. According to the positions of *rps19*, *rpl22*, *ndhH*, *ndhF*, *ndhA*, *psbA*, and *trnH*, we identified some types of connections, which were caused by the contraction and expansion of the inverted repeat region.

The dN/dS ratio, Synonymous (dS) and non-synonymous (dN) substitution rate, is used to evaluate the purification selection of protein-coding genes and the sequence difference. Our result showed that most of the gene sequences had little difference (dS < 0.1). The dn / dS analysis revealed that most protein-coding genes were under negative selection, only a few genes were under positive selection (dN/dS > 1), and other plastids had similar findings.

The plastid genome is an excellent resource to infer the relationship between evolution and phylogeny. Many studies have used chloroplast 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 are selected as research objects, and they are confirmed that *Musa basjoo* Siebold and *Musa itinerans* are 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 (Liu et al. 2017). In this study, we obtained and analyzed for the first time the complete chloroplast genome sequence of the Chinese traditional medicinal plant (*Musa basjoo* Siebold) by using Illumina high-throughput sequencing technology. The genome sequencing, assembly, annotation, and comparative analysis revealed that *Musa basjoo* Siebold cp genomes had a typical quadruple structure with a conserved arrangement. Their GC content, arrangement, and codon usage features were 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 *Musa 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, we reveal the complete chloroplast genome of *Musa basjoo* Siebold for the first time. As genus *Musa* and *zingiberales* plants generally grow in tropical areas, they share a close relationship. The previous researches often used *Musa* and *Zingiberales* species to construct the evolutionary tree to illustrate the relationship between the two genera. However, it is the first time that we have used only *Musa* species, 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. The *Musa basjoo* and *Musa itinerans* are the closest relatives within one sister group. The *Musa balbisiana* and *Musa textilis* belong to another sister group. In contrast, *Musa banksii*, *Musa ornata*, *Musa acuminata*, and *Musa beccarii* belong to a third group. This finding may well provide a relatively complete reference for future studies.

Declarations

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

Conflict of interest: The authors have no conflict of interest.

Availability of data and material: The identified long repeat regions (Kurtz et al. 2001) 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

Authors' contributions: FL: Conceptualization, Software, Formal analysis, Writing - Original Draft, Visualization, 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; YZ: Review &editing and data curation.

References

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. Journal of Agricultural Science - Tokyo Nogyo Daigaku (Japan)

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 (15):2723-2732. doi:10.1016/j.biomaterials.2004.07.047

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology : a journal of computational molecular cell biology 19 (5):455-477. doi:10.1089/cmb.2012.0021

Batnini 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. Scientia Horticulturae 245:99-106. doi:https://doi.org/10.1016/j.scienta.2018.09.071

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 (20):6095-6104. doi:10.1093/emboj/16.20.6095

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:i54-62. doi:10.1093/bioinformatics/btg1005

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. doi:10.1186/1471-2164-13-715

Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research 14 (7):1394-1403. doi:10.1101/gr.2289704

Dierckxsens N, Mardulyn P, Smits G (2016) NOVOPlasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Research doi: 10.1093/nar/gkw955. doi:10.1093/nar/gkw955

Doose D, Grand C, Lesire C (2017) MAUVE Runtime: A Component-Based Middleware to Reconfigure Software Architectures in Real-Time. doi:10.1109/IRC.2017.47

Doyle J, Doyle J (1986) A Rapid DNA Isolation Procedure from Small Quantities of Fresh Leaf Tissues. Phytochem Bull 19

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. Nucleic Acids Res 32 (Web Server issue):W273-279. doi:10.1093/nar/gkh458

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. Journal of controlled release : official journal of the Controlled Release Society 95 (2):197-207. doi:10.1016/j.jconrel.2003.11.006

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 and surfaces B, Biointerfaces 102:833-841. doi:10.1016/j.colsurfb.2012.09.010

He P, Ma Q, Dong M, Yang Z, Liu L (2019) The complete chloroplast genome of Leontice incerta and phylogeny of Berberidaceae. Mitochondrial DNA Part B 4:101-102. doi:10.1080/23802359.2018.1536489

Henriquez CL, Abdullah, Ahmed I, Carlsen MM, Zuluaga A, Croat TB, McKain MR (2020) Evolutionary dynamics of chloroplast genomes in subfamily Aroideae (Araceae). Genomics 112 (3):2349-2360. doi:https://doi.org/10.1016/j.ygeno.2020.01.006

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. Molecular ecology resources 9 (2):439-457. doi:10.1111/j.1755-0998.2008.02439.x

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. Scientific Reports 6 (1):34057. doi:10.1038/srep34057

Ivanova Z, Sablok G, Daskalova E, Zahmanova G, Apostolova E, Yahubyan G, Baev V (2017) Chloroplast Genome Analysis of Resurrection Tertiary Relict Haberlea rhodopensis Highlights Genes Important for Desiccation Stress Response. Front Plant Sci 8:204. doi:10.3389/fpls.2017.00204

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Müller KF, Guisinger-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 U S A 104 (49):19369-19374. doi:10.1073/pnas.0709121104

Jiao J, Huang W, Bai Z, Liu F, Ma C, Liang Z (2018) DNA barcoding for the efficient and accurate identification of medicinal polygonati rhizoma in China. PloS one 13 (7):e0201015. doi:10.1371/journal.pone.0201015

Katoh K, Standley D (2013) Katoh K, Standley DM.. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. Mol Biol Evol 30: 772-780. Molecular biology and evolution 30. doi:10.1093/molbev/mst010

Kim DW, Kang JH, Oh DH, Yong CS, Choi H-G (2012) Development of novel flurbiprofen-loaded solid self-microemulsifying drug delivery system using gelatin as solid carrier. Journal of Microencapsulation 29 (4):323-330. doi:10.3109/02652048.2011.651497

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.

doi:10.1371/journal.pone.0000508

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 (22):4633-4642. doi:10.1093/nar/29.22.4633

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. doi:10.1080/23802359.2019.1566794

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 8:968. doi:10.3389/fpls.2017.00968

Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Molecular biology and evolution 13 (5):660-665. doi:10.1093/oxfordjournals.molbev.a025626

Lohse M, Drechsel O, Kahlau S, Bock R (2013) OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic acids research 41. doi:10.1093/nar/gkt289

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 (2):113-117. doi:10.1080/10717540801905124

Munir A, Mehmood A, Azam S (2016) Structural and Function Prediction of Musa acuminata subsp. Malaccensis Protein. International Journal Bioautomotion 20:19-30

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution 32 (1):268-274. doi:10.1093/molbev/msu300

Niu Y-F, Gao C-W, Liu J (2018) The complete chloroplast genome sequence of wild banana, Musa balbisiana variety' Pisang Klutuk Wulung' (Musaceae). Mitochondrial DNA Part B 3:460-461. doi:10.1080/23802359.2018.1462123

Otálvaro F, Görls H, Hölscher 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 (1):61-66. doi:10.1016/s0031-9422(02)00066-3

Pascual-Villalobos MJ, Rodríguez B (2007) Constituents of Musa balbisiana seeds and their activity against Cryptolestes pusillus. Biochemical Systematics and Ecology 35 (1):11-16. doi:https://doi.org/10.1016/j.bse.2006.08.004

Pollux ÉK (2012) Musa (genus). Chrono Press

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 (1):101-113. doi:10.1007/s10709-017-0003-x

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) 22 (8). doi:10.3390/molecules22081330

Stamatakis A (2014) RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics (Oxford, England) 30. doi:10.1093/bioinformatics/btu033

Tamura KT (1998) Cycloartane triterpenes from the fruit peel of Musa sapientum. Phytochemistry (Oxford)

Terakami S, Matsumura Y, Kurita K, Kanamori H, Katayose Y, Yamamoto T, Katayama H (2012) Complete sequence of the chloroplast genome from pear (Pyrus pyrifolia): genome structure and comparative analysis. Tree Genetics & Genomes 8 (4):841-854. doi:10.1007/s11295-012-0469-8

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. Genomics, Proteomics & Bioinformatics 8 (1):77-80. doi:https://doi.org/10.1016/S1672-0229(10)60008-3

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, Salmon 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 (8):810-821. doi:10.1038/s41477-019-0452-6

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics (Oxford, England) 20 (17):3252-3255. doi:10.1093/bioinformatics/bth352

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 Urophysa and Their Contribution to Chloroplast Phylogeny and Adaptive Evolution. Int J Mol Sci 19 (7):1-20. doi:10.3390/ijms19071847

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. doi:10.3389/fpls.2017.00015

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. Theoretical and Applied Genetics 102 (5):683-688. doi:10.1007/s001220051697

Xu F, Wu H, Wang X, Yang Y, Wang Y, Qian H, Zhang Y (2014) RP-HPLC characterization of lupenone and β-sitosterol in rhizoma musae and evaluation of the anti-diabetic activity of lupenone in diabetic Sprague-

Dawley rats. Molecules (Basel, Switzerland) 19 (9):14114-14127. doi:10.3390/molecules190914114

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. Horticulture Research 6 (1):89. doi:10.1038/s41438-019-0171-1

Yang L, Qun LU, Zhiyuan Z, & Jiechun, C. (2013) Advances in studies on Musa basjoo Sieb.Et Zucc. Journal of Guangdong Pharmaceutical University

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 (2):2614-2616. doi:10.1080/23802359.2019.1642167

Tables

Due to technical limitations, table 1, 2, 3, 4, 5 and 6 is only available as a download in the Supplemental Files section.

Figures

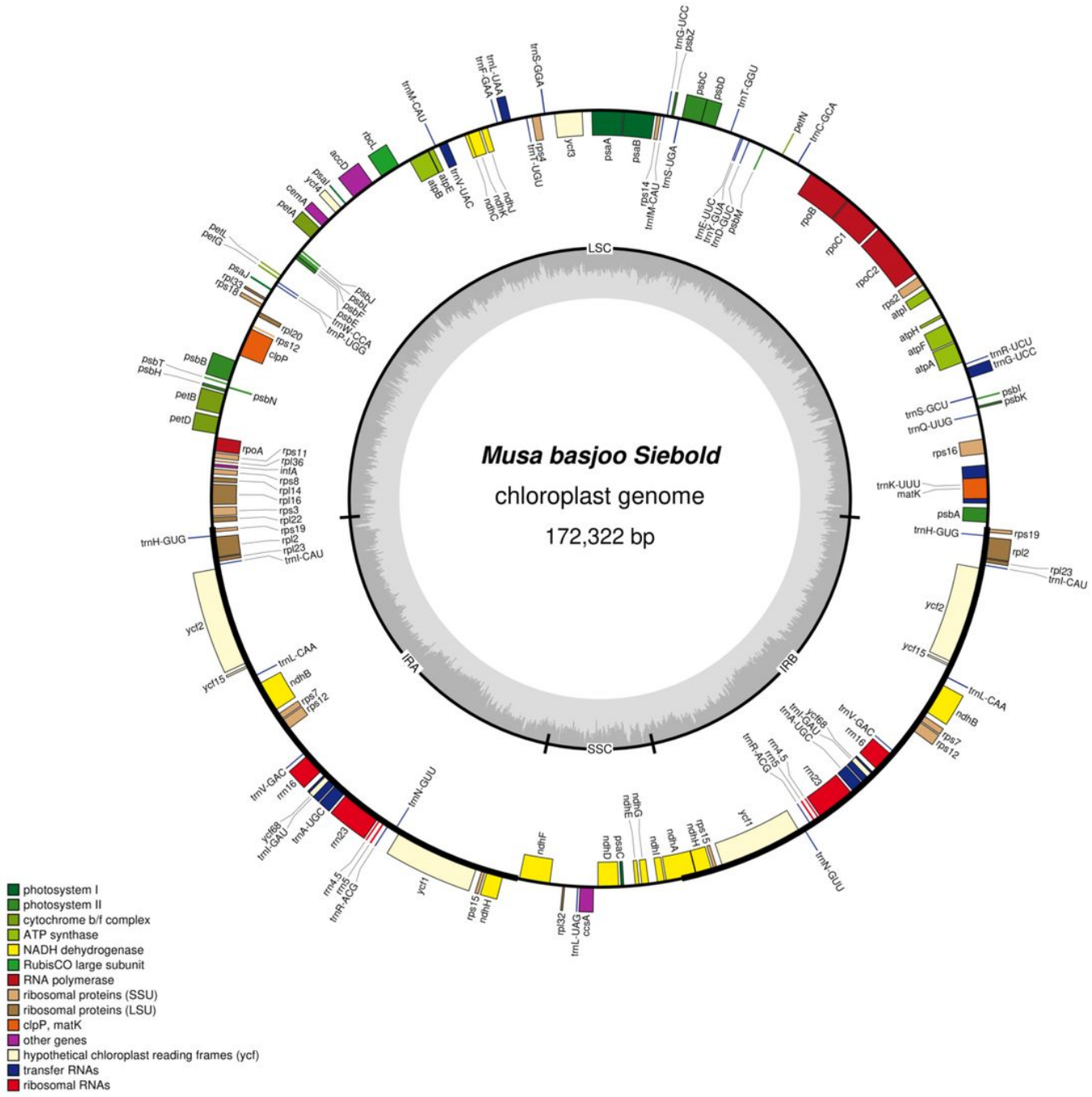

Figure 1

Plastid genome map of Musa basjoo Siebold; Note: The genes inside the circle are transcribed clockwise, and those outside are transcribed counterclockwise. The darker gray in the inner circle shows the GC content, while the lighter gray shows the AT content. The genes of different functions are color-coded.

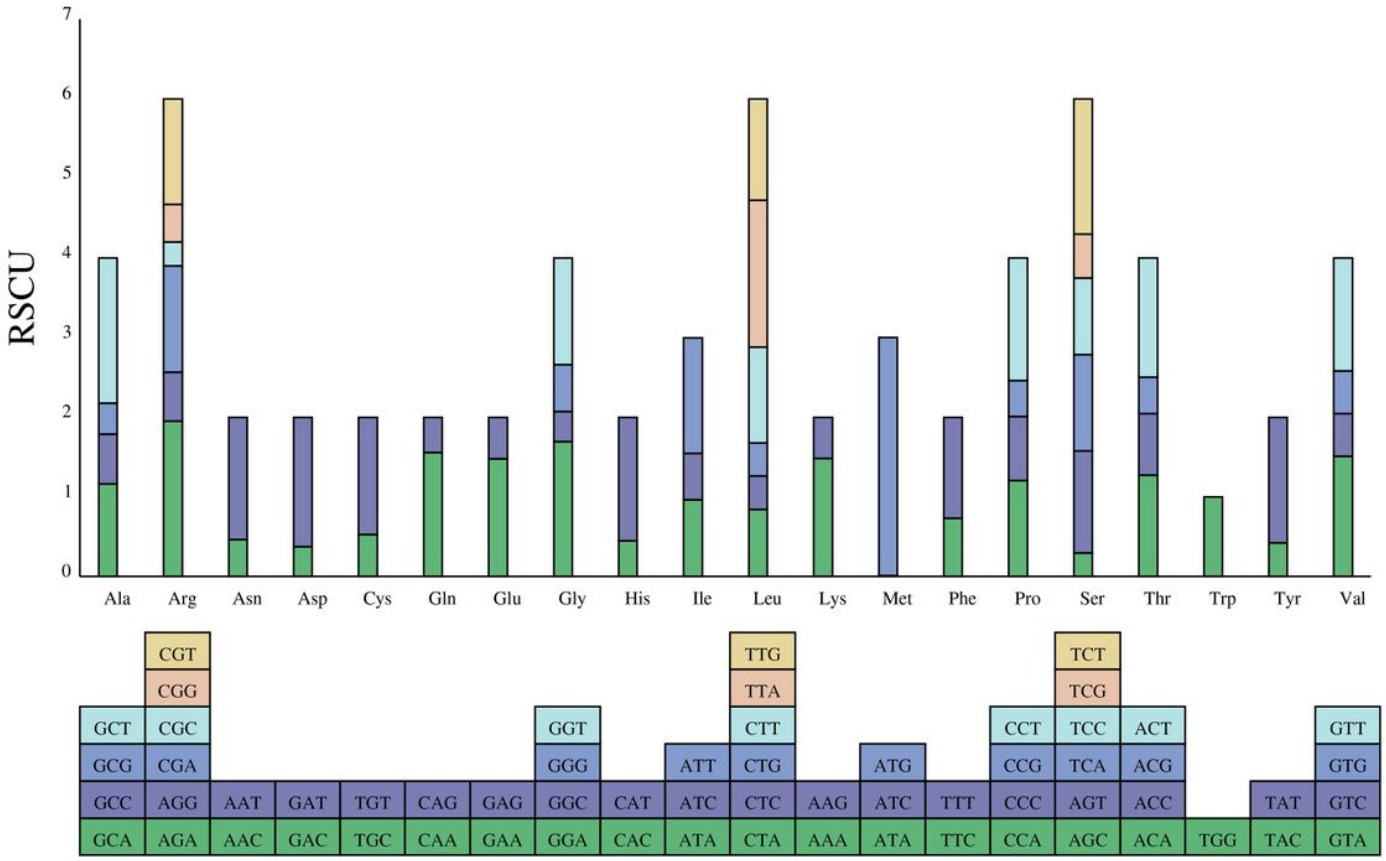

Figure 2

RSCU histogram of Musa basjoo Siebold; Note: The blocks underneath represent all codons encoding each amino acid. The columns on the top represent the sums of RSCU values of all codons.

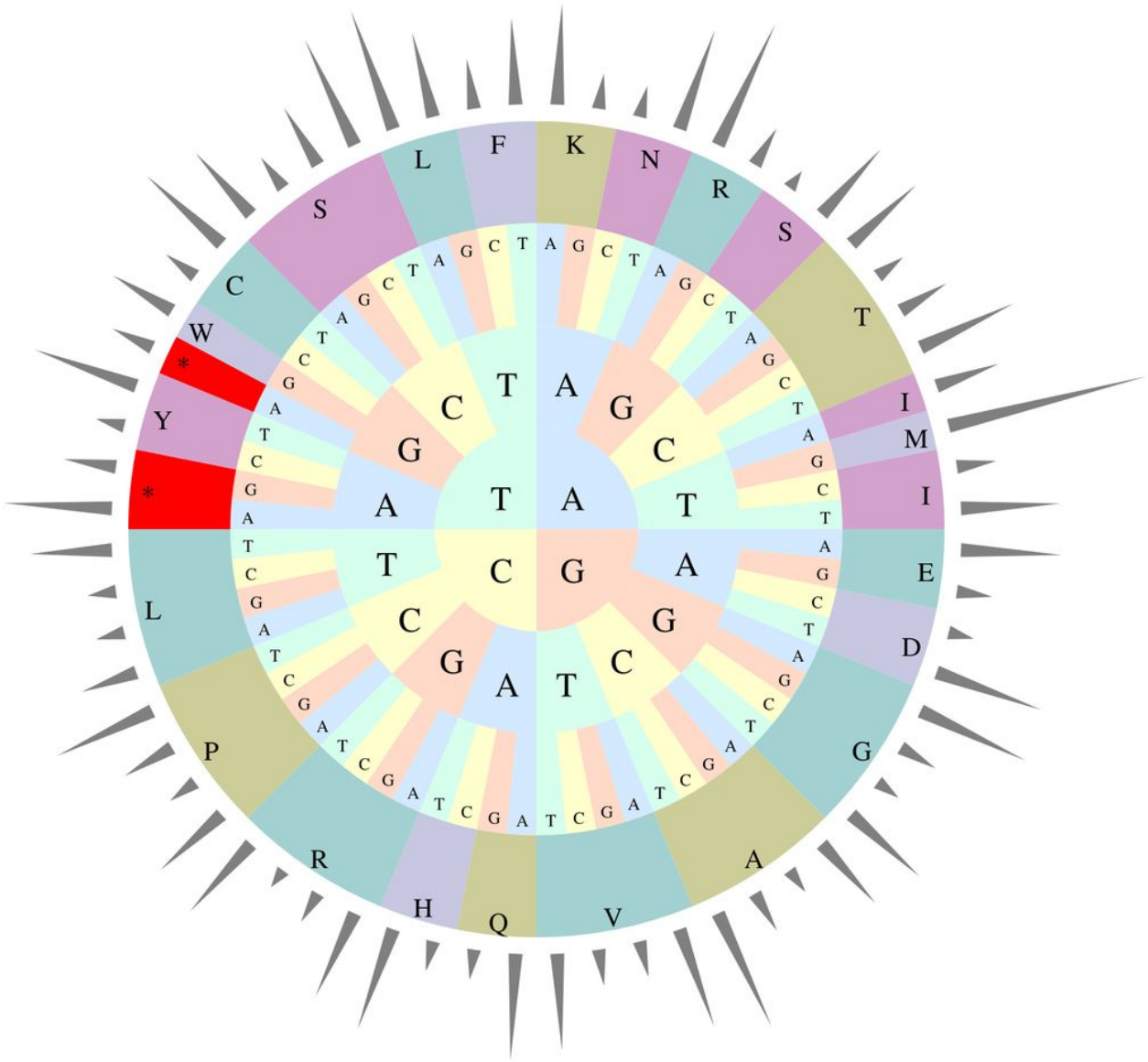

Figure 3

RSCU pie chart of Musa basjoo Siebold; Note: the height of the outermost column is RSCU value, the inner layer is an amino acid, and the innermost three layers are codons.

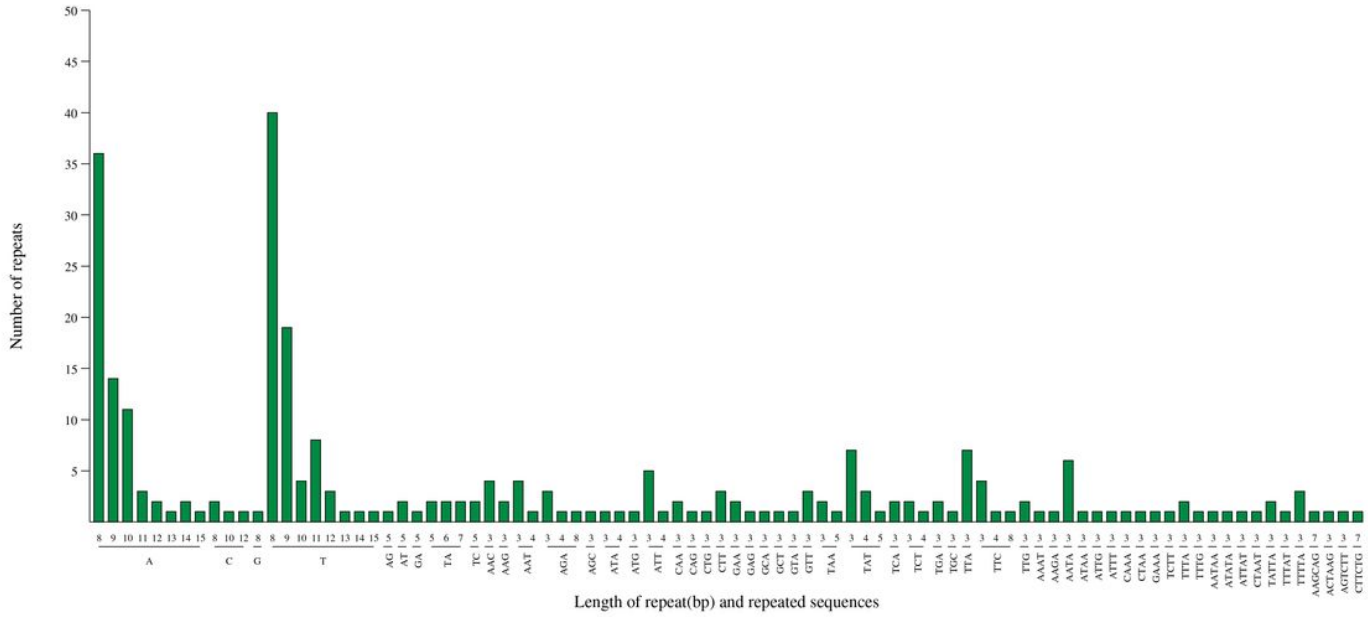

Figure 4

Distribution of simple sequence repeats (SSRs) in the chloroplast genomes of Musa basjoo Siebold; Note: Horizontal coordinates represent SSR duplicate units, and longitudinal coordinates represent the number of SSR types.

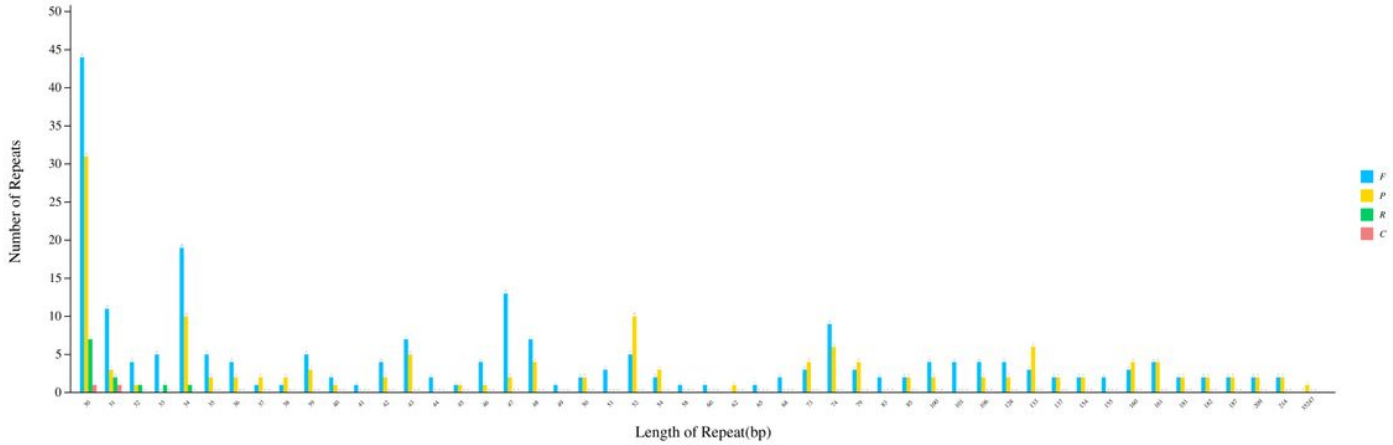

Figure 5

Distribution of long repeat sequences in Musa basjoo Siebold chloroplast genomes; Note: Horizontal coordinate specifies the type of scattered repeat sequence, and longitudinal coordinate marks the number of the scattered repeat sequence. F represents forward repetition, P represents palindrome repetition, R represents reverse repetition, and C represents complementary repetition.

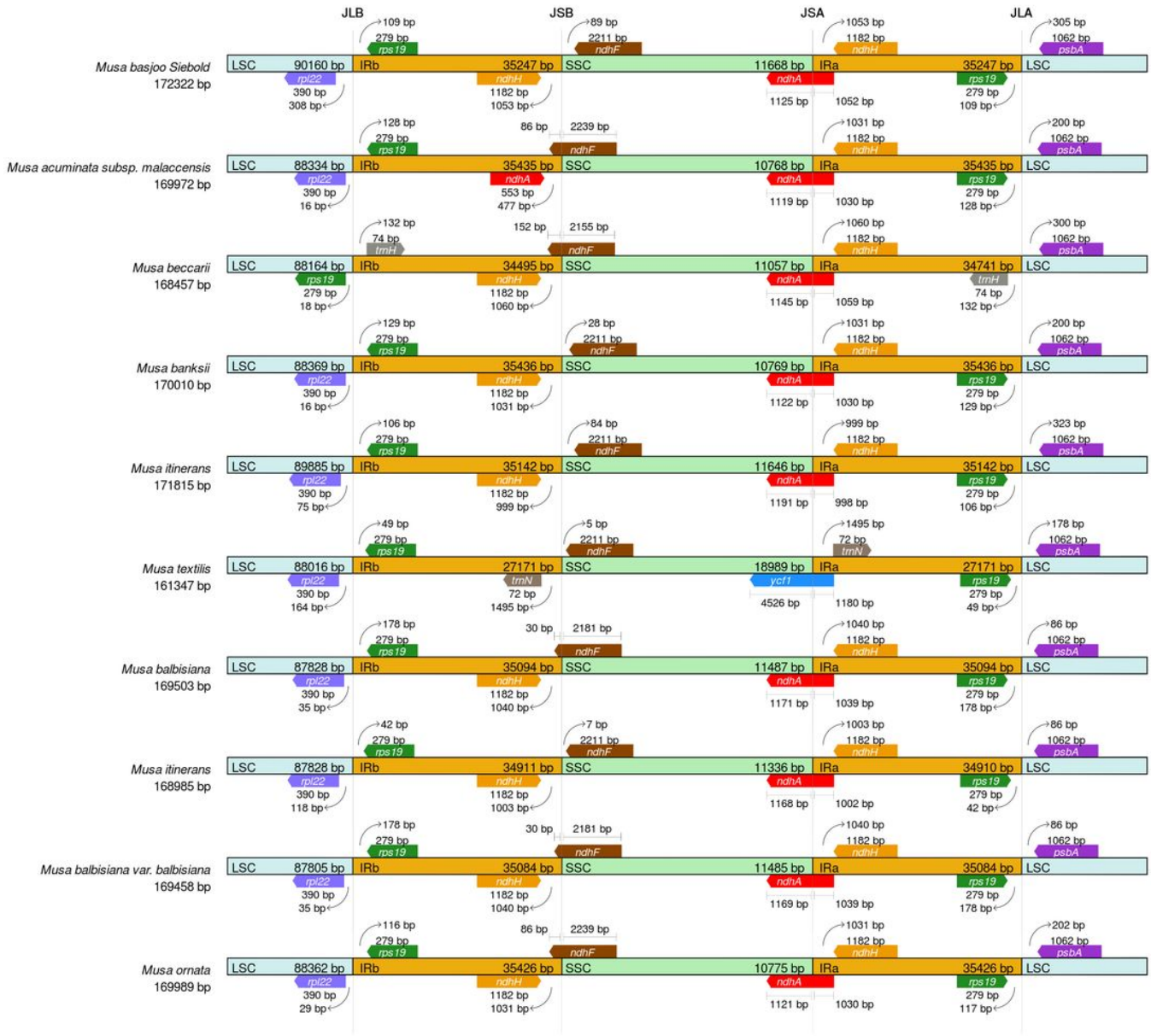

Figure 6

Comparison of the large single-copy (LSC), short single-copy (SSC), and inverted repeat (IR) regions in the plastomes of eight Musaceae species; Note: 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 points, and the figures give the information of the genes near the connection points.

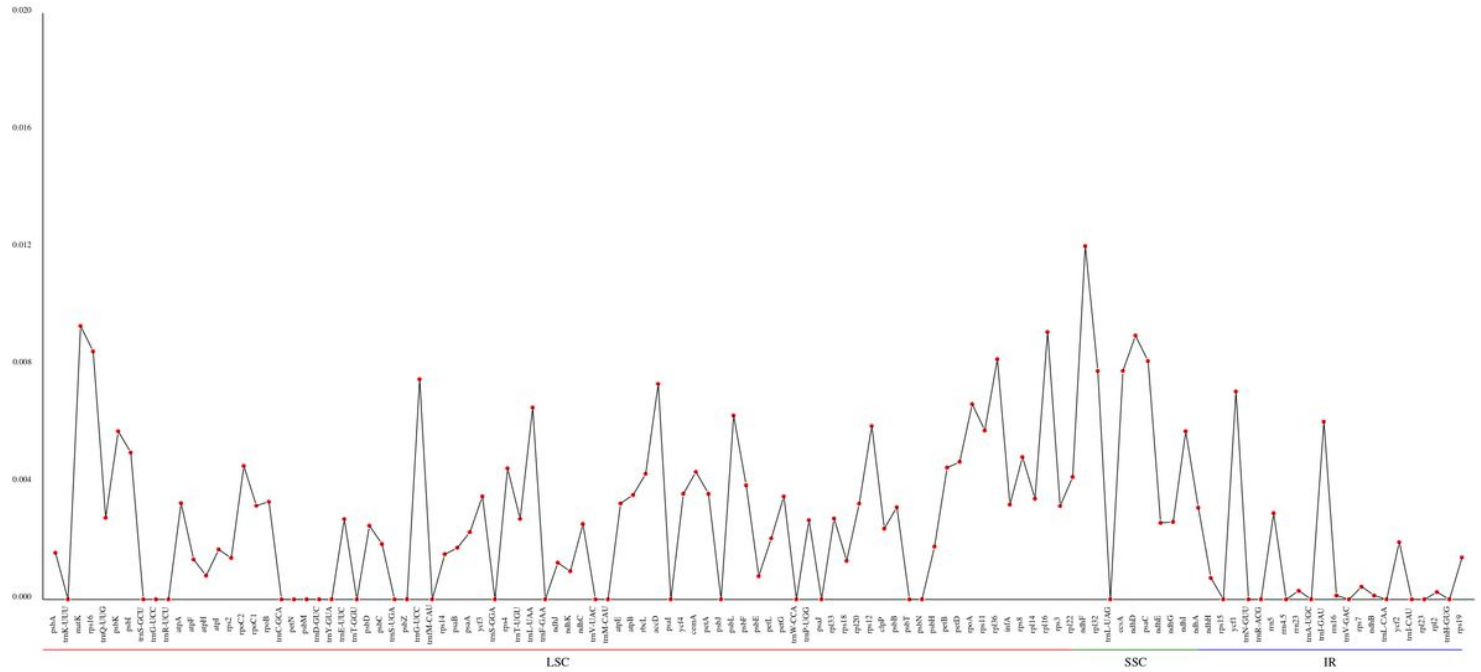

Figure 7

Line graph of the PI values of Musa basjoo Siebold genes; Note: Horizontal coordinate represents gene names and longitudinal coordinates representing PI value.

Figure 8

Mauve alignment of plastid genomes of 8 species of Musaceae. The Musa basjoo Siebold genome is put at the top as the reference genome. Within each of the alignments, local collinear blocks are represented by blocks of the same color connected by lines; Note: 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.

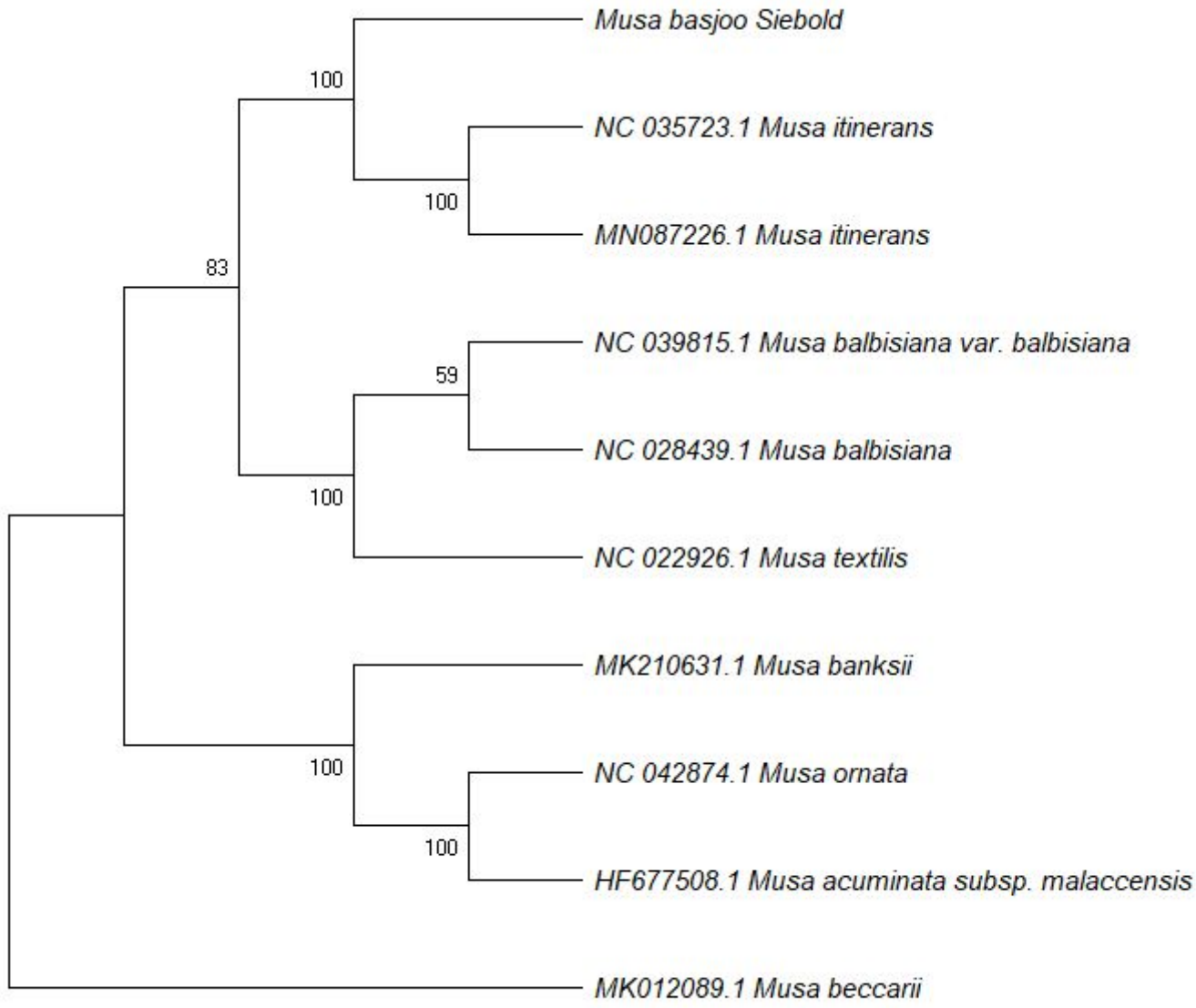

Figure 9

Maximum likelihood (ML) phylogenetic tree reconstruction, including 10 Musaceae species based on concatenated sequences from all chloroplast genomes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- table1.xlsx
- TableS1.xlsx
- Table2.xlsx
- table3.xlsx
- table4.xlsx
- table5.xlsx
- table6.xlsx
- TableS2.xlsx
- TableS3.xlsx