CTU Journal of Innovation and Sustainable Development
ISSN 2615-9422 | e-ISSN 2815-5602

DOI:10.22144/ctujoisd.2023.049

Characterization of single nucleotide polymorphisms in chloroplast genomes of Musaceae Juss.

Mach Bao Ngoc, and Do Hoang Dang Khoa*

NTT Hi-Tech Institute, Nguyen Tat Thanh University, Viet Nam

**Corresponding author (dhdkhoa@ntt.edu.vn)*

Article info.

Received 23 Mar 2023
Revised 13 Apr 2023
Accepted 14 Apr 2023

Keywords

Chloroplast genome, Musaceae, nucleotide diversity, phylogenetic analysis, SNPs

ABSTRACT

Musaceae Juss., also called the bananas and plantains family contains essential food crops with critical economic value and nutritional and medicinal properties. In this study, complete chloroplast genomes of 55 species of Musaceae, including all three genera of Musa, Musella, and Ensete, were used to characterize single nucleotide polymorphisms. Also, nucleotide diversity among surveyed species was observed. The results showed regions of high genetic variability in the chloroplast genome and genes carrying multiple single-nucleotide polymorphisms specific for species and genera, such as ycf1, ycf2, ndhF, matK, accD, infA, and petL. A biased nucleotide conversion toward G, C, and T suggests a trend in the evolution of the Musaceae chloroplast genomes. Phylogenetic analysis revealed a close relationship between Ensete and Musella genera and confirmed the existence of two clades in the genus Musa. This study summarizes nucleotide diversity, focusing on single-nucleotide polymorphisms, which are helpful for further studies on population genetics and developing molecular markers in Musaceae.

1. INTRODUCTION

Musaceae Juss. is a family of monocot plants which comprises over 90 species (Royal Botanic Gardens Kew, 2023) and can be divided into three genera: *Musa* (the largest genus, with 82 accepted species), *Musella* (one species), and *Ensete* (eight species). The family is known as bananas and plantains native to Southeast Asia and Oceania regions.

Bananas are popular, high-nutritional tropical fruits and have been widely consumed because they are non-seasonal crops with affordable prices. These fruits are an abundant source of vitamins (A, B_6, C, and D), minerals, and calories (International Tropical Fruits Network, 2016). Bananas also have excellent medicinal properties. Numerous parts of banana plants are ingredients of traditional medicine to treat disorders and help prevent or relieve the symptoms of chronic diseases (Kumar et al., 2012; Sarma et al., 2019). Because of their valuable characteristics and essential roles in the economy, fruits have been one of the most exported fruit trees worldwide with significant economic value (approximately over 19 million tons exported in 2022 (The Food and Agriculture Organization Corporate Statistical Database (FAOSTAT, 2021). The plantation area of these fruits has reached over 6.7 million hectares by 2021 (FAOSTAT, 2022).

Tracing back history, the Papuans in New Guinea cultivated the first domesticated bananas at least 7000 years ago (Simmonds & Shepherd, 1955; Heslop-Harrison & Schwarzacher, 2007; Perrier et al., 2011). They were naturally parthenocarpic individuals (seedless) from the *Musa banksii* species. The crops were domesticated and cultured by early cultivation methods such as transplantation. Later, the plant proximity brought the cultivated

bananas to the Island of Southeast Asia. They hybridized with other subspecies of *Musa acuminata* and *Musa balbisiana* (possibly independently domesticated). A series of hybridization eventually produced triploid banana cultivars like the modern domesticated bananas today (Sardos et al., 2022).

Domesticated bananas are highly vulnerable to diseases and insect threats since they are sterile (Marín et al., 2003; Jones, 2007; Drenth & Kema, 2021). Their cuttings or whole plants are used to reproduce the next generations vegetatively. Thus, genetic diversity and vulnerability to diseases of domesticated bananas have been a primary concern of global farmers and industries. An enormous plantation area can be threatened or destroyed by a disease.

Wild bananas have a high tolerance against harsh environmental conditions and resistance to many dangerous diseases or insects. For example, Black Sigatoka disease caused by *Mycosphaerella fijiensis M. Morelet* fungus destroyed a significant plantation area of bananas and plantains (Marin et al., 2003; Thangavelu et al., 2020). Fungicides were used to control the disease, which is not just harmful to humans, but also pollutes the environment and increases the cost of production. Meanwhile, resistant traits to this disease were found in wild diploid cultivars, including *M. acuminata subsp. burmanicca*, *M. acuminata subsp. malaccensis*, and *M. acuminata subsp. siamea* (Ahmad et al., 2020). Therefore, breeding research focuses on creating new banana cultivars with disease and insect resistance by hybridizing with wild diploid bananas or gene modification based on knowledge about resistant traits in wild species. Since 1982, an Embrapa Breeding Program has produced commercial triploid or tetraploid banana cultivars resistant to Black Sigatoka disease based on the traits of wild bananas (Amorim et al., 2011).

The whole genome, plastid genome structures, and diversity at the molecular level of the Musaceae family have been studied thoroughly in recent years. The nuclear genome sequence of *Musa acuminata* was published in 2012 (D'Hont et al., 2012). Further research on the Musaceae genome has been conducted to understand better plant phylogeny, evolution, and genetics. Banana Genome Hub, the public database of the banana genome, was constructed, published, and continuously updated (Droc et al., 2013).

Plants' chloroplast genome (also called plastome) has received increased attention from scientists. Chloroplasts are membrane-bound organelles that perform photosynthesis – a process that creates food for plants in the form of sugars and fuels all activities of these organisms. In chloroplasts, various proteins participate in photosynthesis and other crucial biochemical pathways. The organelles have autonomic genomes with sizes varying from 18,000 to 200,000 base pairs (bp) (Dobrogojski et al., 2020), containing photosynthesis and vitality-related genes with high conservation in their structure. The chloroplast genome has been a target of research focused on plant taxonomy and breeding for years (Palmer et al., 1988; Leister, 2003).

Single Nucleotide Polymorphism (SNP) is the most popular type of genetic variation. Its abundance plays a crucial role in the genetic diversity of organisms. SNPs that only appear in a few species are defined as species-specific SNPs, thereby being exploited as promising biological markers for selection and breeding. This study aimed to characterize SNPs in 55 published Musaceae complete chloroplast genome sequences in the NCBI database. The distribution and proportion of SNPs in each gene were recorded and evaluated. In addition, we also analyzed chloroplast genome features, nucleotide diversity, and the phylogenetic relationship of the Musaceae species.

2. MATERIALS AND METHOD

2.1. A sampling of chloroplast genomes in Musaceae

Complete chloroplast genome sequences of 55 species in the Musaceae were obtained from NCBI with the strategy of (1) Sequences were selected based on the following standards: Sequences with "UNVERIFIED" labels were eliminated; (2) Whole chloroplast genome sequences which do not contain unidentified nucleotides (labeled as "N"); (3) Sequences whose accession number starts with

"NC" were prioritized due to the higher completeness and the latest updates; (4) If different research groups published various sequences of one species, one of them was randomly chosen for this study. All data came with gene annotation, and Geneious Prime v2022.2 (Geneious Prime, 2022) was conducted for screening the SNPs in 79 genes of the Musaceae chloroplast genomes. Brief details of whole chloroplast genome sequences used in this study are listed in Table 1 and Table 2.

Table 1. Gene content of chloroplast genomes in Musaceae species

Groups of genes	Names of genes
Photosystem I	*psaA, psaB, psaC, psaI, psaJ*
Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
Cytochrome	*petA, petB*, petD*, petG, petL, petN*
ATP synthases	*atpA, atpB, atpE, atpF*, atpH, atpI*
Large unit of Rubisco	*rbcL*
NADH dehydrogenase	*ndhA*, ndhB*, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
ATP-dependent protease subunit P	*clpP**
Envelope membrane protein	*cemA*
Large units of ribosome	*rpl2*, rpl14, rpl16*, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36*
Small units of ribosome	*rps2, rps3, rps4, rps7, rps8, rps11, rps12*, rps14, rps15, rps16*, rps18, rps19*
RNA polymerase	*rpoA, rpoB, rpoC1*, rpoC2*
Initiation factor	*infA*
Miscellaneous protein	*accD, ccsA, matK*
Hypothetical proteins and conserved reading frames	*ycf1, ycf2, ycf3*, ycf4, ycf15*[a]

**- genes with introns; a-gene only occurs in the plastome of M. acuminata subsp. Malaccensis*

Table 2. List of 55 chloroplast genome sequences in the Musaceae

No.	Accession Number	Species	Size (bp)
1	NC_058947	*Ensete glaucum*	168,248
2	NC_058943	*Ensete livingstonianum*	168,258
3	NC_058930	*Ensete superbum*	168,332
4	NC_058926	*Ensete ventricosum*	168,411
5	LC609627	*Musa acuminata cv. Pisang Lilin*	170,586
6	NC_058951	*Musa acuminata subsp. burmannica*	169,795
7	NC_058945	*Musa acuminata subsp. halabanensis*	169,658
8	HF677508	*Musa acuminata subsp. malaccensis*	169,972
9	NC_058940	*Musa acuminata subsp. microcarpa*	170,081
10	NC_058928	*Musa acuminata subsp. truncata*	170,137
11	MT593357	*Musa acuminata var. chinensis*	170,402
12	NC_058925	*Musa acuminata var. zebrina*	169,873
13	NC_058957	*Musa aurantiaca*	170,058
14	NC_028439	*Musa balbisiana*	169,503
15	NC_039815	*Musa balbisiana var. balbisiana*	169,458
16	NC_058956	*Musa banksii*	169,808
17	NC_058955	*Musa barioensis*	168,559
18	NC_058954	*Musa basjoo*	171,853
19	NC_058953	*Musa beccarii*	168,209
20	NC_058952	*Musa borneensis*	168,703
21	NC_058950	*Musa cheesmanii*	170,714
22	NC_058949	*Musa chunii*	169,309
23	NC_058948	*Musa coccinea*	166,826
24	NC_058946	*Musa gracilis*	166,756
25	NC_056826	*Musa ingens*	168,471
26	NC_035723	*Musa itinerans*	168,985
27	LC609773	*Musa itinerans var. formosana*	171,478
28	NC_056827	*Musa jackeyi*	167,975
29	NC_058944	*Musa johnsii*	167,331

No.	Accession Number	Species	Size (bp)
30	NC_056828	*Musa laterita*	170,565
31	NC_058942	*Musa lokok*	166,902
32	NC_056829	*Musa lolodensis*	168,542
33	NC_058941	*Musa maclayi subsp. maclayi*	167,586
34	NC_056830	*Musa mannii*	170,699
35	NC_056831	*Musa nagensium*	170,304
36	LC609776	*Musa ornata*	170,285
37	NC_058939	*Musa paracoccinea*	167,601
38	NC_058958	*Musa peekelii subsp. angustigemma*	167,660
39	NC_058938	*Musa puspanjaliae*	171,298
40	NC_058937	*Musa rosea*	168,495
41	NC_056832	*Musa rubinea*	172,653
42	NC_058936	*Musa rubra*	169,309
43	NC_058935	*Musa ruiliensis*	167,806
44	NC_058934	*Musa salaccensis*	167,018
45	NC_058933	*Musa sanguinea*	170,501
46	NC_058932	*Musa schizocarpa*	169,821
47	LC610771	*Musa siamensis*	170,349
48	NC_022926	*Musa textilis*	161,347
49	NC_058929	*Musa tonkinensis*	170100
50	NC_056833	*Musa troglodytarum*	168,121
51	NC_058927	*Musa velutina*	169,791
52	LC609625	*Musa x chiliocarpa*	170,959
53	LC609772	*Musa x formobisiana*	171,982
54	NC_056834	*Musa yunnanensis*	169,816
55	LC610747	*Musella lasiocarpa*	170,142

2.2. Single nucleotide polymorphism (SNP) identification in Musaceae chloroplast genomes

Whole chloroplast genome sequences of 55 species were aligned using the MAUVE program embedded in Geneious Prime. Sequences of each gene would be aligned into blocks. Then, the "Extract region" command was used to take out variant sequences of 79 protein-coding genes in those species (Table 1). The gene *ycf15* only exists in *M. acuminata subsp. malaccensis* with a shortened structure because of the presence of termination triplet. Hence, it was eliminated from this comparison. Variant sequences of each gene were aligned by Multiple Sequence comparisons by Log-Expectation (abbreviation: MUSCLE algorithm) (Edgar, 2004).

SNPs in the sequences were detected and highlighted with different colors for each type of nucleotide (A, T, G, or C) by the "Find Variations/SNPs" command. Statistical results were summarized in the "Annotations" tab. Data were filtered to eliminate Indels (insertion or deletion polymorphisms) variations. Variant sequences and their frequency were presented in the data tables. The species-specific SNPs and their occurrence frequency were recorded. Types of substitution and their frequency were identified and calculated to determine the most abundant one and trends in molecular evolution. The dominant substitution type in each gene was indicated, and the predominant type in most of the genes would be labeled as the most abundant one (highest frequency).

2.3. Nucleotide diversity analysis

The chloroplast genome sequence of 55 species in the Musaceae family was aligned by MAUVE alignment version 1.1.3 (Darling, 2010) in Geneious Prime software with default settings. The algorithm was used to identify structural changes in the chloroplast genome, including the rearrangement of genes, inversions, and insertion. Structures of the sequences were rearranged to become locally collinear blocks – variant structures of each gene in different species were vertically aligned. Processed data were imported to DnaSP 6 software (Rozas, 2017) for nucleotide diversity analysis using a sliding window and Pi values calculation. The sliding window size was set to be 2000; the step size was 100. Nucleotide diversity plots of the Musaceae family and two genera *Musa*, *Ensete*, were constructed and evaluated. Genus *Musella* has only

one species. Thus, its nucleotide diversity plot was not created and compared to two other genera and the Musaceae family.

2.4. Characterizing Inverted Repeat (IR) regions in chloroplast genomes of the Musaceae

Raw data downloaded from NCBI were imported to Geneious Prime. Large Single Copy (LSC) region, Small Single Copy region (SSC), and Inverted Repeat (IR) regions were assessed and rearranged to correct the genes' positions, direction, length, and structure in these regions. Geneious Prime determined junctions between the LSC, SSC, and IR regions. Contraction and expansion of IR regions were evaluated based on their length, the positions of borders between regions, and the distribution and structures of genes in these areas. Similarities and differences between species were also compared and recorded.

2.5. Phylogenetic analysis

Whole chloroplast genome sequences of 55 species from the Musaceae family were aligned by the MUSCLE alignment algorithm in Geneious Prime software. Later, the data were imported to jModelTest version 2.1.10 (Guindon & Gascuel, 2003; Darriba et al., 2012) to identify the optimal phylogenetic model ("GTR + I + G"). MrBayes version 3.2.7a (Ronquist et al., 2012) was used to calculate Bayesian Inference (BI value) along with Markov chain Monte Carlo (Brooks et al., 2011) simulation. The default run length was set at 1.000.000 generations. The first 25% of the data obtained will be eliminated. The run would be stopped if, after 20.000 generations, the standard deviation of split frequencies is below 0.01 because, at that point, the evolutionary model is stable. Otherwise, generations will continuously be added until reaching the expected value. The phylogenetic tree was constructed and corrected by FigTree version 1.4.0 software (Rambaut & Drummond, 2012).

3. RESULTS AND DISCUSSION

3.1 RESULTS

3.1.1. Features of SNPs in Musaceae chloroplast genomes

The number of single nucleotide polymorphisms in one gene was proportional to the size of the gene. Significant genes like *ycf1*, *ycf2*, *ndhF*, *matK*, and *rpoC2* had the most SNPs, as opposed to a few numbers of SNPs that occurred in small genes such as *atpH*, *ndhB*, *petL*, and *psaI* (Figure 1). The number of species-specific SNPs also followed this rule. However, the ratio between species-specific and total SNPs in one gene varied independently of the gene's size. Over two-thirds of SNPs in *psbJ*, *psbZ*, *rpl23*, *petN*, *psaJ*, and *aptH* genes were species specific. PsaA, *psbF*, and *rps12* had no species-specific SNP in their structures.

Among the three genera, the *Ensete* and the *Musella* had the most species-specific SNPs, while numerous species of the *Musa* genus had a few numbers, and 17 species did not have any species-specific SNP. *M. ingens*, *M. rubinea*, *M. schizocarpa*, *M. chessmani*, *M. nagensium*, *M. tonkinesis*, and *M. yunnanesis* were exceptions with species-specific SNPs greater than 35 (Table 3). We also recognized potential genes like *infA* and *petL* with genus-specific SNPs (specific for the *Ensete* and *Musella* genera, respectively). In contrast, some genes had abundant species-specific SNPs in most species of the Musaceae family, such as *ycf1*, *ycf2*, *rpoC2*, *ndhF*, *matK*, *ndhD*.

Table 3. Quantity and distribution of species-specific SNPs in the chloroplast genomes of the Musaceae

Species	**Total species-specific SNPs**	***Genes contain species-specific SNPs (number of SNP)***
Ensete glaucum	157	*ycf2 (22), ycf1 (15), ndhF (13), matK (8), ndhD (5), rpoB (4), accD (8), ccsA (5), rpoC1 (2), rpoA (4), atpA (1), psaB (2), atpB (2), psbB (3), rbcL (2), petA (1), ndhA (1), psbC (2), psbA (4), rps3 (4), ndhG (1), clpP (3), cemA (2), rps2 (3), psbD (1), atpF (1), rps8 (1), petB (2), rps11 (1), atpI (2), ndhE (1), rpl16 (1), rpl14 (1), rps16 (1), ycf3 (2), ndhJ (1), psbH (1), rpl33 (1), ndhH (1), infA (1), rpl2 (1), petG (1), psbN (1), psa J(1), rpl23 (2), psbI (1), psbL (1)*
Ensete livingstonianum	104	*ycf2 (1), ycf1 (15), ndhF (12), rpoC2 (10), matK (9), ndhD (5), rpoB (1), accD (4), ccsA (3), rpoC1 (1), rpoA (3), atpA (2), psaB (2), atpB (2), psbB (1), rbcL (3), ndhA (2), psbA (1), ndhK (1), ndhI (1), rps3 (2), ndhG (1), cemA (3), rps2 (1), atpF (2), petB (2), rps11 (2), ycf4 (2), atpI (2), rpl14 (1), rps4 (1), rpl32 (1), rps16 (1), psbH (1), ndhC (1), rpl33 (2), ndhH (1), infA (2), rpl2 (1), ndhB (1), petG (1), atpH (2), rpl23 (2), rpl36 (2), psaI (1)*
Ensete superbum	184	*ycf2 (7), ycf1 (6), ndhF (20), rpoC2 (15), matK (12), ndhD (9), rpoB (14), accD (4), ccsA (4), rpoC1 (6), rpoA (3), atpA (4), psaB (1), rbcL (5), ndhA (3), psbC (3), psbA (3), rpl22 (1), ndhK (4), ndhI (2), rps3 (1), ndhG (2), clpP (4), rps2 (2), psbD (2), rps8 (2), petB (2), rps11 (1), ycf4 (2), atpI (1), ndhE (1), rpl16 (1), rpl14 (1), rpl20 (1), rps4 (1), rpl32 (3), ycf3 (1), ndhJ (3), ndhC (2), atpE (1), ndhH (1), psaC (1), rps18 (1), ndhB (1), petN (4), psbE (1), psbJ (1), psbM (1), psbT (1)*
Ensete ventricosum	106	*ycf2 (7), ycf1 (3), ndhF (7), rpoC2 (9), matK (8), ndhD (4), rpoB (2), accD (6), ccsA (1), rpoC1 (5), atpA (5), psaB (2), psbB (4), rbcL (1), petA (1), ndhA (1), psbC (1), psbA (1), rpl22 (1), ndhK (2), ndhI (1), ndhG (2), psbD (4), atpF (1), rps8 (1), petB (2), rps11 (1), ycf4 (1), atpI (3), ndhE (1), rpl16 (1), rpl14 (2), rps4 (1), ndhJ (1), ndhC (1), petD (1), atpE (1), rpl33 (1), infA (1), rps19 (1), ndhB (2), rps14 (1), atpH (1), psbZ (1), psbJ (1), rps15 (1)*
Musa acuminata cv. Pisang Lilin	2	*ndhF (1), ndhD (1)*
Musa acuminata subsp. Burmannica	1	*ccsA (1)*
Musa acuminata subsp. Halabanensis	10	*ndhF (2), rpoB (2), ccsA (2), rpoA (1), psbB (1), psbA (1), atpF (1)*
Musa acuminata subsp. Malaccensis	2	*rpl14 (1), psaC (1)*
Musa acuminata subsp. Microcarpa	3	*ndhF (2), ccsA (1)*
Musa acuminata subsp. Truncata	0	
Musa acuminata var. chinensis	3	*ycf1 (1), atpA (1), ndhK (1)*
Musa acuminata var. zebrina	0	
Musa aurantiaca	1	*ndhF (1)*
Musa balbisiana	0	

Species	**Total species-specific SNPs**	***Genes contain species-specific SNPs (number of SNP)***
Musa balbisiana var. balbisiana	0	
Musa banksii	10	*ycf1 (1), ndhF (1), rpoC2(1), ndhD (1), accD (1), ccsA (1), rpoA (1), clpP (1), atpF (1), petB (1)*
Musa barioensis	6	*ycf1 (1), rpoB (1), psaB (2), rpl20 (1), ycf3 (1)*
Musa basjoo	5	*rpoC2 (1), psbB (1), rps11 (1), ycf4 (1), atpI (1),*
Musa beccarii	14	*ycf2 (3), ycf1 (2), rpoC2 (1), rpoB (2), accD (3), ccsA (1), rpoC1 (1), atpA (1)*
Musa borneensis	19	*ycf1 (3), ndhF (4), matK (2), rpoB (2), rpoA (1), ndhI (1), rps3 (1), clpP (1), cemA (1), rps11 (1), rpl20 (1), rpl33 (1)*
Musa cheesmanii	70	*ycf2 (4), ycf1 (4), ndhF (9), rpoC2 (6), matK (1), ndhD (5), rpoB (2), accD (1), rpoC1 (2), rpoA (5), psaB (1), psbB (2), petA (3), ndhA (3), ndhI (1), rps3 (2), ndhG (1), cemA (1), psbD (1), rps8 (4), petB (1), rps11 (1), atpI (1), rpl16 (1), rps4 (1), rps16 (1), petD (1), rpl2 (1), rps14 (1), petG (1), psbN (1), petN (1),*
Musa chunii	0	
Musa coccinea	7	*ycf2 (1), ycf1 (1), ndhF (1), rpoB (1), ccsA (3)*
Musa gracilis	3	*ycf1 (1), rpoC2 (1), ndhI (1)*
Musa ingens	111	*ycf2 (7), ycf1 (5), ndhF (11), rpoC2 (5), matK (4), ndhD (10), rpoB (5), accD (3), ccsA (4), rpoC1 (4), rpoA (1), atpA (1), psbB (1), rbcL (2), petA (3), ndhA (1), psbC (4), rpl22 (8), ndhI (2), ndhG (1), clpP (2), cemA (1), rps2 (1), atpF (1), rps 8 (2), petB (1), rps11 (1), ycf4 (2), ndhE (1), rpl16 (1), rpl20 (1), rpl20 (1), rps4 (1), rps16 (1), psbH (3), ndhH (1), rps18 (2), rps14 (1), psbK (1), psbZ (1), rps7 (1),*
Musa itinerans	3	*rpoC2 (1), ndhD (1), rpl14 (1)*
Musa itinerans var. formosana	0	
Musa jackeyi	2	*rpl20 (1), rpl33 (1),*
Musa johnsii	14	*ycf2 (2), ndhF (2), matK (3), rpoB (1), atpB (1), rbcL (1), petA (1), ndhE (1), rpl20 (1), rps14 (1)*
Musa laterita	0	
Musa lokok	6	*ycf2 (1), matK (1), ccsA (1), psbC (1), rpl20 (1), rps18 (1)*
Musa lolodensis	7	*ycf1 (2), matK (1), rpoB (1), accD (1), petA (1), rps4 (1)*
Musa maclayi subsp. Maclayi	0	
Musa mannii	0	
Musa nagensium	56	*ycf2 (1), ycf1 (3), ndhF (5), rpoC2 (5), matK (3), ndhD (1), rpoB (2), accD (2), ccsA (4), rpoC1 (1), atpA (1), psaB (2), psbB (2), rbcL (1), petA (1), ndhA (2), psbC (2), psbA (1), rps3 (1), ndhG (1), cemA (1), rps2 (3), atpF (1)*
Musa ornata	0	
Musa paracoccinea	5	*ycf2 (2), ycf1 (1), rpoC2 (1), rps3 (1)*
Musa peekelii subsp. Angustigemma	1	*ndhF (1)*
Musa puspanjaliae	30	*ycf1 (1), ndhF (3), rpoC2 (3), matK (1), ndhD (1), accD (2), rpoC1 (1), rpoA (4), psaB (1), atpB (2), psbB (1), ndhA (1), rps3 (1), cemA (1), psbD (2), rps11 (1), ycf4 (1), atpI (1), rpl14 (1), rps4 (1),*

Species	**Total species-specific SNPs**	*Genes contain species-specific SNPs (number of SNP)*
Musa rosea	0	
Musa rubinea	39	*ycf2 (1), ycf1 (4), ndhF (2), rpoC2 (2), matK (2), rpoB (4), ccsA (2), rpoC1 (1), rpoA (1), atpA (2), atpB (1), psbB (1), petA (2), ndhA (1), psbC (1), rpl22 (1), ndhK (2), psbD (2), rps11 (1), atpI (1), rps4 (1), ndhJ (1), ndhC (1), atpE (1), psaJ (1),*
Musa rubra	0	
Musa ruiliensis	0	
Musa salaccensis	3	*psbA (1), cemA (1), petD (1)*
Musa sanguinea	0	
Musa schizocarpa	68	*ycf2 (3), ycf1 (1), ndhF (9), rpoC2 (4), matK (4), ndhD (4), rpoB (4), accD (2), ccsA (1), rpoC1 (1), rpoA (1), atpA (1), psbB (1), petA (5), ndhA (2), psbC (3), rpl22 (3), ndhI (1), rps3 (1), ndhG (2), clpP (1), psbD (1), atpF (2), rps8(1), ycf4 (1), ndhE (3), rpl14 (1), rpl20 (10, rpl32 (1), rps16 (1), ndhC (1), psbZ (1)*
Musa siamensis	0	
Musa textilis	0	
Musa tonkinensis	42	*ycf2 (4), ycf1 (8), ndhF (2), rpoC2 (2), matK (5), ndhD (5), rpoB (2), accD (2), ccsA (1), rpoA (1), psaB (1), psbB (1), petA (1), psbC (1), ndhK (1), rps3 (1), rps11 (1), rpl16 (1), atpH (1), rps7 (1)*
Musa troglodytarum	1	*matK (1)*
Musa velutina	19	*ycf2 (2), ycf1 (1), ndhF (1), rpoC2 (1), matK (1), rpoB (2), accD (1), rpoC1 (1), rpoA (1), atpA (2), rpl22 (1), ndhK (1), rps3 (1), ndhG (1), atpF (1), ndhE (1)*
Musa x chiliocarpa	2	*rpl16 (2)*
Musa x formobisiana	0	
Musa yunnanensis	43	*ycf2 (2), ycf1 (2), ndhF (4), rpoC2 (6), matK (6), accD (1), ccsA (1), rpoC1 (2), atpA (3), psbB (1), psbA (1), rpl22 (1), ndhK (2) clpP (1), cemA (1), rps2 (1), rps8 (1), petB (1), ndhE (1), rpl32 (1), ndhB (1), atpH (1), psaJ (1)*
Musella lasiocarpa	135	*ycf2 (20), ycf1 (9), ndhF (8), rpoC2 (9), matK (9), ndhD (9), rpoB (1), accD (4), ccsA (4), rpoC1 (2), rpoA (7), atpA (2), psaB (3), atpB (2), psbB (1), rbcL (2), petA (1), psbA (2), rpl22 (2), ndhK (1), ndhI (2), rps3 (5), ndhG (4), cemA (2), rps2 (1), atpF (4), rps8 (1), ndhE (1), rpl16 (1), rpl14 (1), rpl20 (1), rps4 (1), rps32 (4), rps16 (1), petD (1), psaC (1), rps18 (1), psbK (1), atpH (1), psaJ (2), petL (1)*
TOTAL	1294	

At one specific position of SNP, we may have several types of substitution that differ from species to species. Among the twelve types of nucleotide substitutions, “C to T”, “A to G”, and “T to C” had the highest occurrence frequency.

At one specific position of SNP, we may have several types of substitution that differ from species to species. Among the twelve types of nucleotide substitutions, “C to T”, “A to G”, and “T to C” had the highest occurrence frequency.

In particular, the “C to T” conversion dominated in 28/79 genes, and the “A to G” conversion dominated in 27/79 genes, which is at the same ratio as the “T to C” conversion (Table 4). Trends of nucleotide conversion in the plastid genome showed a trend of gene conversion and provided evidence of evolutionary trends.

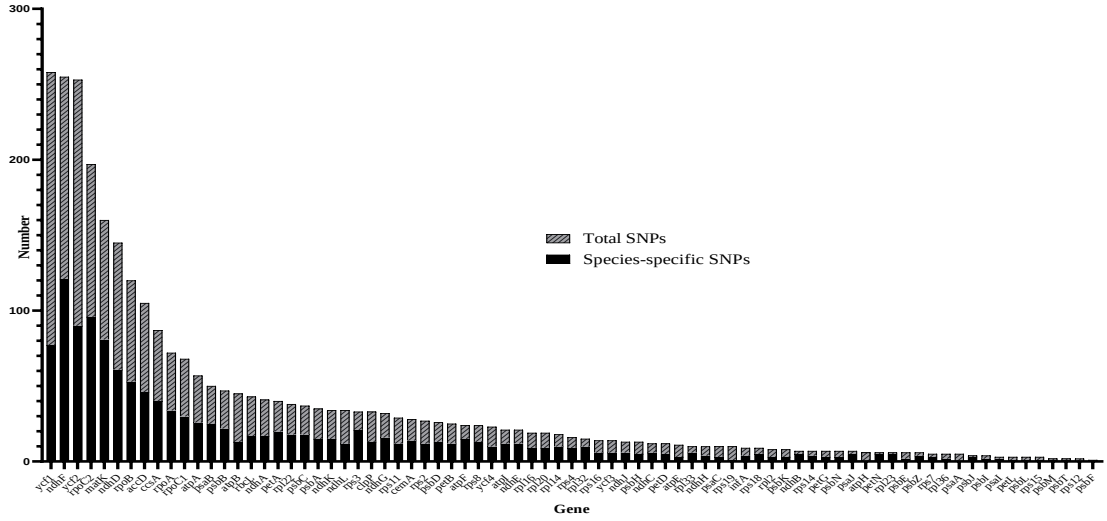

Figure 1. The ratio between the number of species-specific SNPs and total SNP positions in the Musaceae chloroplast genes

Genes are displayed in the order; the genes with a higher number of SNPs are shown first. The ratio of the length of two sections in a stacked column illustrates the ratio between species-specific SNPs and total SNP position in one gene.

Table 4. Summary of nucleotide substitution types creating SNPs and their percentage in each gene

Gene	Total SNPs Positions	Percentage (%)											
		A→G	**A→T**	**A→C**	**G→A**	**G→T**	**G→C**	**C→A**	**C→G**	**C→T**	**T→A**	**T→G**	**T→C**
ycf2	242	18.18	1.19	2.37	14.23	7.91	1.19	7.51	2.77	14.23	2.77	5.93	**21.74**
ycf1	240	14.34	5.81	8.14	9.69	7.36	0.78	5.81	2.71	14.34	6.2	8.14	**16.67**
ndhF	239	**23.14**	2.35	2.35	16.86	7.45	1.96	4.71	0.78	13.73	2.75	5.1	18.82
rpoC2	192	**26.4**	2.03	4.57	13.71	7.61	1.02	6.09	0	10.15	4.57	3.55	20.3
matK	154	**21.3**	3.1	6.9	17.5	8.1	0.6	8.1	1.3	10	3.1	6.9	13.1
ndhD	139	**22.76**	2.07	1.38	16.55	7.59	0.69	4.41	2.07	**22.76**	4.14	2.76	13.1
rpoB	117	17.5	0.8	2.5	10	2.5	1.7	4.2	0	19.2	2.5	4.2	**35**
accD	101	**25.71**	2.86	2.86	15.24	7.62	2.86	5.71	1.9	14.29	2.86	0.95	17.14
ccsA	83	**21.84**	4.6	2.3	20.69	5.75	1.15	11.49	0	14.94	1.15	1.15	14.94
rpoC1	68	**26.47**	2.94	1.47	17.65	5.88	0	4.41	0	14.71	1.47	2.94	22.06
rpoA	66	15.28	1.39	4.17	12.5	16.67	4.17	2.78	0	**18.06**	2.78	4.17	**18.06**
atpA	56	**28.1**	0	1.8	24.6	5.3	3.5	3.5	0	7	5.3	3.5	17.5
psaB	48	32	2	2	16	2	0	0	0	10	0	2	**34**
atpB	45	**33.33**	0	0	22.22	6.67	0	4.44	0	11.11	0	2.22	20
psbB	44	23.4	2.13	4.26	17.02	4.26	0	2.13	0	8.51	2.13	6.38	**29.79**
rbcL	42	**18.6**	2.33	6.98	16.28	4.65	4.65	6.98	0	16.28	0	4.65	**18.6**
ndhA	40	**24.39**	0	2.44	12.2	7.32	0	7.32	2.44	17.07	4.88	2.44	19.51
petA	40	17.5	0	2.5	**22.5**	2.5	0	5	5	12.5	2.5	7.5	**22.5**
psbC	37	10.81	5.41	2.7	16.22	8.11	2.7	5.41	0	18.92	0	5.41	**24.32**
psbA	35	25.53	0	0	**57.45**	0	0	2.13	0	2.13	0	0	12.77
ndhK	34	**29.41**	2.94	2.94	2.94	8.82	0	2.94	0	17.65	5.88	8.82	17.65
ndhL	34	11.76	0	0	20.59	2.94	0	2.94	0	**29.41**	0	2.94	**29.41**
rpl22	34	13.16	15.79	5.26	5.26	2.63	2.63	2.63	7.89	**21.05**	5.26	13.16	5.26
rps3	32	18.18	3.03	3.03	6.06	12.12	3.03	3.03	0	**24.24**	0	6.06	21.21
clpP	31	**33.33**	0	6.06	9.09	3.03	0	3.03	3.03	6.06	6.06	3.03	27.27
ndhG	31	21.9	3.1	3.1	15.6	0	0	3.1	0	21.9	0	3.1	**28.1**
cemA	27	17.86	0	3.57	10.71	0	0	10.71	0	17.86	0	3.57	**35.71**
rps2	27	18.52	0	3.7	**22.22**	3.7	0	7.41	3.7	**22.22**	0	3.7	14.81
psbD	25	**26.92**	3.85	3.85	7.69	0	0	7.69	3.85	23.08	0	3.85	19.23

Gene	Total SNPs Positions	Percentage (%)											
		A→G	**A→T**	**A→C**	**G→A**	**G→T**	**G→C**	**C→A**	**C→G**	**C→T**	**T→A**	**T→G**	**T→C**
atpF	24	8.33	0	4.17	**25**	8.33	0	16.67	0	**25**	4.17	0	8.33
petB	24	20	0	4	16	0	0	0	0	**36**	4	0	20
rps11	24	17.24	3.45	6.9	6.9	0	0	0	0	17.24	10.34	10.34	**27.59**
rps8	24	16.67	0	0	12.5	4.17	0	0	0	12.5	0	12.5	**41.67**
ycf4	23	**21.74**	0	4.35	17.39	4.35	0	17.39	0	17.39	4.35	4.35	8.7
atpI	21	**38.1**	0	0	9.52	4.76	0	9.52	4.76	14.29	0	0	19.05
ndhE	19	19.05	9.52	4.76	**28.57**	4.76	0	4.76	0	4.76	4.76	9.52	9.52
rpl16	18	15.79	5.26	0	10.53	10.53	0	5.26	0	**26.32**	0	5.26	21.05
rpl14	17	0	0	5.56	16.67	5.56	0	5.56	0	22.22	5.56	0	**38.89**
rpl20	17	21.05	0	5.26	10.53	5.26	0	15.79	0	**26.32**	5.26	0	10.53
rps4	16	6.25	0	6.25	25	0	6.25	0	0	**37.5**	0	0	18.75
rpl32	14	**33.33**	6.67	0	13.33	6.67	0	6.67	0	20	0	0	13.33
rps16	14	**42.86**	7.14	7.14	14.29	0	0	0	0	14.29	0	0	14.29
ycf3	14	7.14	0	0	7.14	7.14	0	0	0	28.57	0	0	**50**
ndhJ	13	23.08	0	7.69	7.69	0	7.69	0	0	**30.77**	0	7.69	15.38
psbH	13	**23.08**	0	0	15.38	7.69	0	7.69	0	**23.08**	0	0	**23.08**
ndhC	12	**33.33**	0	8.33	8.33	8.33	0	0	0	16.67	0	0	25
atpE	11	9.09	0	0	18.18	9.09	0	0	0	**27.27**	0	9.09	**27.27**
petD	11	16.67	0	8.33	**25**	0	0	8.33	0	16.67	0	8.33	16.67
ndhH	10	0	0	10	**30**	10	0	10	0	10	0	0	**30**
psaC	10	**40**	0	0	20	0	0	0	0	10	0	10	20
rpl33	10	10	0	10	**30**	10	0	0	0	**30**	0	0	10
infA	9	44.44	0	11.11	11.11	11.11	0	0	0	11.11	0	0	11.11
rps19	9	10	**20**	10	0	0	0	10	0	**20**	0	10	**20**
rpl2	8	0	12.5	0	12.5	0	0	**25**	0	**25**	0	12.5	12.5
rps18	8	**22.22**	0	0	11.11	0	0	**22.22**	11.11	11.11	11.11	0	11.11
ndhB	7	14.29	0	14.29	14.29	0	0	14.29	0	**28.57**	0	0	14.29
petG	7	14.29	0	0	0	0	0	0	0	14.29	0	0	**71.43**
psbK	7	12.5	0	0	12.5	0	0	0	0	12.5	0	25	**37.5**
psbN	7	0	0	0	**57.14**	0	0	0	0	14.29	0	0	28.57
rps14	7	0	0	14.29	14.29	0	0	0	0	**28.57**	14.29	14.29	14.29
atpH	6	16.67	0	0	0	0	0	16.67	0	**66.67**	0	0	0
petN	6	16.67	0	0	**50**	0	0	0	0	0	0	0	33.33
psaJ	6	14.29	0	0	0	0	0	0	0	14.29	**28.57**	14.29	**28.57**
psbE	6	16.67	0	0	33.33	0	0	0	0	16.67	0	0	33.33
rpl23	6	**16.67**	0	0	**16.67**	**16.67**	0	0	**16.67**	**16.67**	0	0	**16.67**
psbZ	5	16.67	0	0	0	16.67	0	16.67	0	**33.33**	0	16.67	0
rpl36	5	0	0	0	0	0	0	0	0	**80**	0	0	20
rps7	5	0	0	0	0	20	0	0	0	**40**	20	0	20
psbI	4	0	0	0	**25**	25	0	**25**	0	**25**	0	0	0
psbJ	4	0	0	0	**50**	0	0	0	0	25	0	0	25
petL	3	**33.33**	0	0	**33.33**	0	0	0	0	**33.33**	0	0	0
psaI	3	0	**33.33**	0	0	0	0	0	0	**33.33**	0	0	**33.33**
psbL	3	0	0	0	0	**66.67**	0	0	0	33.33	0	0	0
rps15	3	**33.33**	0	0	0	**33.33**	0	0	0	0	0	0	**33.33**
psbM	2	**50**	0	0	0	**50**	0	0	0	0	0	0	0
psbT	2	**50**	0	**50**	0	0	0	0	0	0	0	0	0
rps12	2	0	0	0	0	0	0	0	0	0	0	0	**100**
psbF	1	0	0	0	0	0	0	0	0	0	0	**100**	0
Dominant substitution type in (number of genes)		27	2	1	14	4	0	3	1	28	1	1	27

Dominant types with the highest occurrence frequency are presented in bold numbers.

3.1.2. Nucleotide Diversity in chloroplast genomes of Musaceae

A comparison of nucleotide diversity in the *Ensete* and *Musa* genera to that of the Musaceae family was made to detect mutation hotspots in chloroplast genome structure and evolutionary direction. The family had high nucleotide diversity, proven by several hypervariable regions with high Pi values in its plastome. Most species in the Musaceae family belonged to the *Musa* genus. Hence, patterns found in nucleotide diversity of three regions, LSC, SSC, and IR of this genus, represented the whole family.

In the LSC region of the *Ensete*, the *Musa*, and the Musaceae, *matK*, *rps16*, *psbD*, *psbE*, *petL*, and *petA* were notable hypervariable genes. Except for the *Ensete* genus, the region containing *rbcL* and *accD* genes showed markable genetic variability. Despite the similarities in LSC structure and the overall trend of genetic variation in this region among the two genera and the Musaceae family, *rpoA-infA* was a specific hypervariable region of the family but insignificant in the *Musa* and the *Ensete* genera. Likewise, the *ndhC-trnV_UAC* peak only appeared in the LSC region of the *Ensete* genus, as demonstrated in Figure 2.

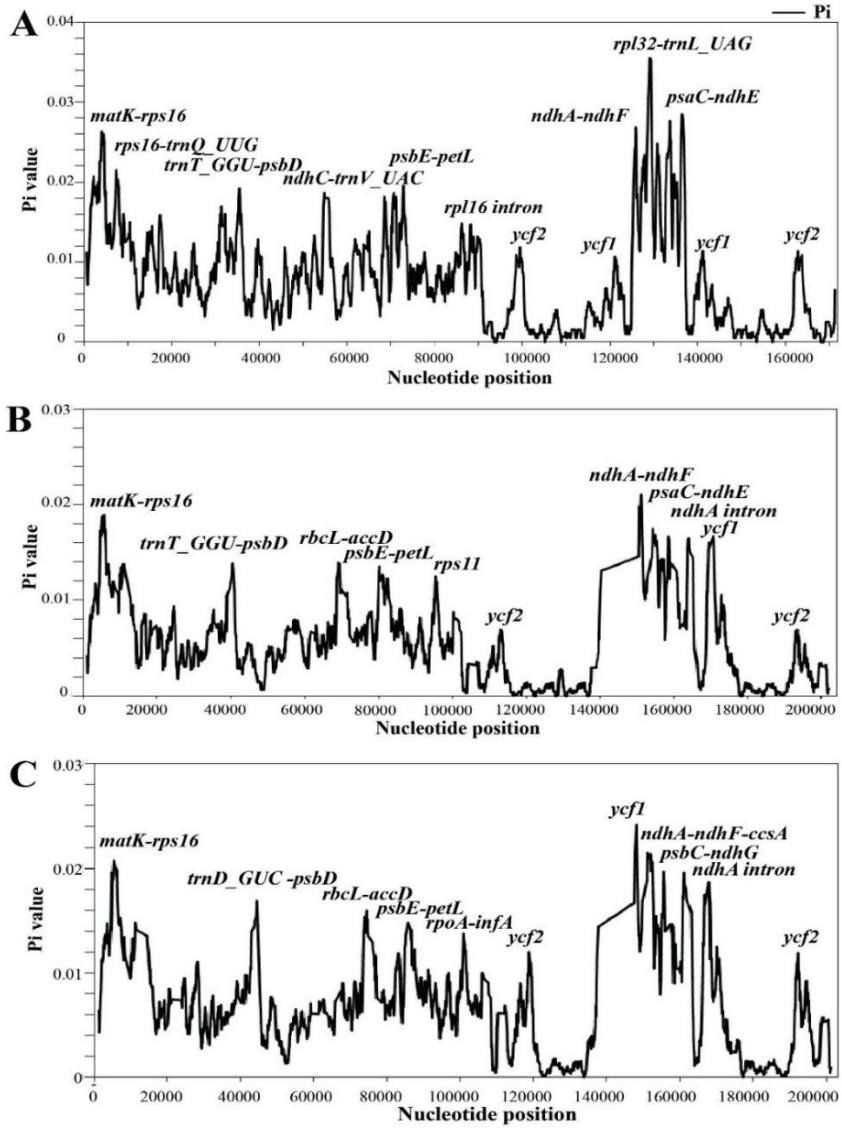

Figure 2. Nucleotide diversity in the chloroplast genomes of the Musaceae

(A) Nucleotide diversity in Ensete genus. (B) Nucleotide diversity in Musa. (C) Nucleotide diversity in Musaceae family. Genes in hypervariable regions with significant Pi values are annotated.

The SSC region of two genera and the Musaceae family was the most variable. Representative genes with high genetic variability were *ndhA* and *ndhF*, the highest peak of the SSC region in the *Musa* genus. The *rpl32-trnL_UAG* region had the top nucleotide diversity in the *Ensete* genus, although it did not have significant variability in the Musaceae family. Similarly, even though *psaC-ndhE* showed excellent variability in the *Ensete* and the *Musa* genera, this pattern did not appear in the Musaceae (Figure 2). The IR regions of the plastid genome in plants were very conservative. The only two exceptional cases, *ycf1* and *ycf2*, were genes with high genetic variability. Because of an uneven expansion of IRa and IRb regions, in Graph (B) and Graph (C) of Figure 2, there was only one peak of *ycf1*, which seemed to be merged with SSC regions. In fact, *ycf1* belonged to IR regions.

3.1.3. Features of Inverted Repeat (IR) regions in Musaceae

In the *Ensete* genus, the region *trnH_GUG-rps19* was situated on the boundary of LSC/IR, while *ndhA-ndhF* was on SSC/IR border, overlap length was 1014-1078 bp. The *Musella* genus's LSC/IR junction site contained the *rps19-rpl12* region, and the *ndhF* gene was found on the SSC/IR boundary. The *Musa* genus had its LSC/IR boundary containing intergenic sequences *rps19-rpl22* or *trnH_GUG-rps19*. In some exceptional cases, including *M. acuminata subsp. halabanensis*, *M. chunii*, their LSC/IR junction sites only contained the *rpl22* gene, and *M. ingens* had its *rpl2* intron on the LSC/IR border (Table 5). Except for *M. textilis*, which had *ycf1* in its SSC/IR junction, most *Musa* species had either *ndhA-ndhF* or only the *ndhF* gene positioned in this zone. Results also showed an expansion of IR regions in the Musaceae since their average size in other angiosperms was approximately 25,000 bp, while those varied from 29,960 to 35,678 bp (Table 5).

Table 5. Comparison of chloroplast genome features in the Musaceae

Genera	Species	Accession number	Length (bp)	GC content (%)	Gene content (Protein coding/ tRNA/ rRNA)	Junction		IR Length (bp)
						LSC/IR	SSC/IR [*ndhA* length]	
Ensete	*Ensete glaucum*	NC_058947	168,248	37.1	79/30/4	IGS (*trnH-GUG/rps19*)	IGS (*ndhA-ndhF*) [1045 bp]	34,636
	Ensete livingstonianum	NC_058943	168,258	37.1	79/30/4	IGS (*trnH-GUG/rps19*)	IGS (*ndhA-ndhF*) [1078 bp]	34,506
	Ensete superbum	NC_058930	168,332	37	79/30/4	IGS (*trnH-GUG/rps19*)	IGS (*ndhA-ndhF*) [1014 bp]	34,547
	Ensete ventricosum	NC_058926	168,411	37.1	79/30/4	IGS (*trnH-GUG/rps19*)	IGS (*ndhA-ndhF*) [1049 bp]	34,358
Musella	*Musella lasiocarpa*	LC610747	170,142	36.6	79/30/4	IGS (*rps19/rpl22*)	*ndhF (26 bp)* [1262 bp]	35,645
Musa	*Musa acuminata cv. Pisang Lilin*	LC609627	170,586	36.8	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1030 bp]	35,469
	Musa acuminata subsp. burmannica	NC_058951	169,795	36.9	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1030 bp]	35,376
	Musa acuminata subsp. halabanensis	NC_058945	169,658	36.9	79/30/4	*rpl22* (2bp)	IGS (*ndhA-ndhF*) [773 bp]	34,991

Genera	Species	Accession number	Length (bp)	GC content (%)	Gene content (Protein coding/ tRNA/ rRNA)	Junction		IR Length (bp)
						LSC/IR	SSC/IR [*ndhA* length]	
	Musa acuminata subsp. malaccensis	HF677508	169,972	36.8	80/30/4	IGS (*rps19/rpl22*)	*ndhF (86 bp*) [1030 bp]	35,433
	Musa acuminata subsp. microcarpa	NC_058940	170,081	36.8	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1030 bp]	35,228
	Musa acuminata subsp. truncata	NC_058928	170,137	36.9	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1030 bp]	35,309
	Musa acuminata var. chinensis	MT593357	170,402	36.8	79/30/4	IGS (*rps19/rpl22*)	*ndhF (86 bp)* [1030 bp]	34,974
	Musa acuminata var. zebrina	NC_058925	169,873	36.9	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1030 bp]	35,338
	Musa aurantiaca	NC_058957	170,058	36.9	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1043 bp]	35,407
	Musa balbisiana	NC_028439	169,503	36.8	79/30/4	IGS (*rps19/rpl22*)	*ndhF (30 bp*) [1039 bp]	35,094
	Musa balbisiana var. balbisiana	NC_039815	169,458	36.8	79/30/4	IGS (*rps19/rpl22*)	*ndhF (30 bp)* [1039 bp]	34,776
	Musa banksii	NC_058956	169,798	36.9	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1030 bp]	35,317
	Musa barioensis	NC_058955	168,559	36.8	79/30/4	IGS (*trnH-GUG/rps19*)	*ndhF (64 bp*) [1059 bp]	34,530
	Musa basjoo	NC_058954	171,853	36.5	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1029 bp]	35,184
	Musa beccarii	NC_058953	168,209	36.8	79/30/4	IGS (*trnH-GUG/rps19*)	*ndhF (64 bp)* [1059 bp]	34,495
	Musa borneensis	NC_058952	168,703	36.8	79/30/4	IGS (*trnH-GUG/rps19*)	*ndhF (64 bp)* [1059 bp]	34,600
	Musa cheesmanii	NC_058950	170,714	36.7	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1005 bp]	35,276
	Musa chunii	NC_058949	169,309	37	79/30/4	*rpl22* (18 bp)	IGS (*ndhA-ndhF*) [1043 bp]	35,328
	Musa coccinea	NC_058948	166,826	37.1	79/30/4	IGS (*rpl2/rps19*)	*ndhF (49 bp*) [1196 bp]	34,129
	Musa gracilis	NC_058946	166,756	37	79/30/4	IGS (*trnH-GUG/rps19*)	IGS (*ndhA-ndhF*) [976 bp]	33,985
	Musa ingens	NC_056826	168,471	36.8	79/30/4	*rpl2* intron	*ndhF (37 bp*) [1025 bp]	33,986
	Musa itinerans	NC_035723	168,958	36.8	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1002 bp]	34,906

Genera	Species	Accession number	Length (bp)	GC content (%)	Gene content (Protein coding/ tRNA/ rRNA)	Junction		IR Length (bp)
						LSC/IR	SSC/IR [*ndhA* length]	
	Musa itinerans var. formosana	LC609773	171,478	36.5	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [998 bp]	34,968
	Musa jackeyi	NC_056827	167,975	36.9	79/30/4	IGS (*trnH-GUG/rps19*)	*ndhF (46 bp*) [1015 bp]	34,357
	Musa johnsii	NC_058944	167,331	37	79/30/4	IGS (*trnH-GUG/rps19*)	*ndhF (55 bp*) [1050 bp]	34,387
	Musa laterita	NC_056828	170,565	36.8	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1030 bp]	35,628
	Musa lokok	NC_058942	166,902	37	79/30/4	IGS (*trnH-GUG/rps19*)	*ndhF (38 bp*) [1022 bp]	34,467
	Musa lolodensis	NC_056829	168,542	36.8	79/30/4	IGS (*trnH-GUG/rps19*)	*ndhF (55 bp)* [1050 bp]	34,576
	Musa maclayi subsp. maclayi	NC_058941	167,586	36.9	79/30/4	IGS (*trnH-GUG/rps19*)	*ndhF (46 bp*) [1015 bp]	34,147
	Musa mannii	NC_056830	170,699	36.8	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1043 bp]	35,678
	Musa nagensium	NC_056831	170,304	36.6	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [994 bp]	35,602
	Musa ornata	LC609776	170,285	36.8	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1043 bp]	35,225
	Musa paracoccinea	NC_058939	167,601	37	79/30/4	IGS (*trnH-GUG/rps19*)	*ndhF (41 bp*) [1188 bp]	34,354
	Musa peekelii subsp. angustigemma	NC_058958	167,660	36.9	79/30/4	IGS (*trnH-GUG/rps19*)	*ndhF (46 bp*) [1015 bp]	34,147
	Musa puspanjaliae	NC_058938	171,298	36.6	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [994 bp]	35,218
	Musa rosea	NC_058937	168,945	37.1	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1030 bp]	35,295
	Musa rubinea	NC_056832	172,653	36.4	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1103 bp]	35,586
	Musa rubra	NC_058936	169,309	37	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1030 bp]	35,204
	Musa ruiliensis	NC_058935	167,796	37.1	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1043 bp]	35,224
	Musa salaccensis	NC_058934	167,018	37	79/30/4	IGS (*trnH-GUG/rps19*)	*ndhF (38 bp)* [1022 bp]	34,322

Genera	Species	Accession number	Length (bp)	GC content (%)	Gene content (Protein coding/ tRNA/ rRNA)	Junction		IR Length (bp)
						LSC/IR	SSC/IR [*ndhA* length]	
	Musa sanguinea	NC_058933	170,501	36.8	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1043 bp]	35,179
	Musa schizacarpa	NC_058932	169,821	36.9	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1030 bp]	35,200
	Musa siamensis	LC610771	170,349	36.8	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1030 bp]	35,404
	Musa textilis	NC_022926	161,347	37.1	79/30/4	IGS (*rps19/rpl22*)	*ycf1* (1179 bp) [1179 bp]	27,171
	Musa tonkinensis	NC_058929	170,100	36.8	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [1019 bp]	35,106
	Musa troglodytarum	NC_056833	168,121	36.8	79/30/4	IGS (*trnH-GUG/rps19*)	*ndhF (46 bp)* [1015 bp]	34,174
	Musa velutina	NC_058927	169,791	36.8	79/30/4	IGS (*trnH-GUG/rps19*)	IGS (*ndhA-ndhF*) [1030 bp]	34,666
	Musa x chiliocarpa	LC609625	170,959	36.6	79/30/4	IGS (*rps19/rpl22*)	*ndhF (30 bp*) [1039 bp]	35,477
	Musa x formobisiana	LC609772	171,982	36.5	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [998 bp]	35,130
	Musa yunnanensis	NC_056834	169,816	36.8	79/30/4	IGS (*rps19/rpl22*)	IGS (*ndhA-ndhF*) [685 bp]	34,012

3.1.4. Phylogenetic relationship reconstruction in Musaceae

The phylogenetic tree showed the relationships between the three genera of the Musaceae family (Figure 3). Posterior probability (Bayesian Inference – BI) values were labeled on the branches if they were less than 1, showing chances of inaccuracy. Unlabeled branches had a BI value equal to 1, meaning 100% accurate.

The *Musella* and *Ensete* genera formed one phylogenetic branch. Therefore, they were genetically closer to each other when compared to their relationship with the *Musa* genus.

Musa, the largest genus of Musaceae, could be divided into two clades based on the genetics of species. The first clade (A) comprised 15 species, in which the branches of *M. jackeyi*, *M. troglodytarum*, *M. maclayi subsp. maclayi* and *M. peekelii subsp. angustigemma* were uncertain (posterior probability = 0.87). Other branches with an even higher chance of being inaccurate were those of *M. johnsii*, *M. maclayi subsp. maclayi* and *M. peekelii subsp. angustigemma (*posterior probability = 0.57). Clade (B) had 35 species, and most of the branches had high accuracy, except for subclade (C) (posterior probability = 0.99), *M. x chiliocarpa* (0.57), *M. balbisiana*, *M. balbisiana var. balbisiana*, and *M. textilis* (0.65).

3.2. DISCUSSION

3.2.1. Variability of SNPs in the Musaceae

Most SNPs in the Musaceae plastome occur in significant genes such as *ycf2*, *ycf1*, *ndhF*, *rpoC2*, and *matK*; those genes are located in different regions of the chloroplast genome. This phenomenon insists on the presence and the unbalanced distribution of mutation hotspots in the plastid genome.

A large number of SNPs was found in crucial genes like *ycf1*, *ycf2*, *ndhF*, *rpoC2*, *matK*, etc., indicating

the evolutionary dynamics. In other plants, these genes were also proven to have high variability (Duan et al., 2020; Moghaddam et al., 2022; Zhang et al., 2022). The genes that have undergone the most changes in evolution are believed to have significant roles, and their structure tends to vary continuously.

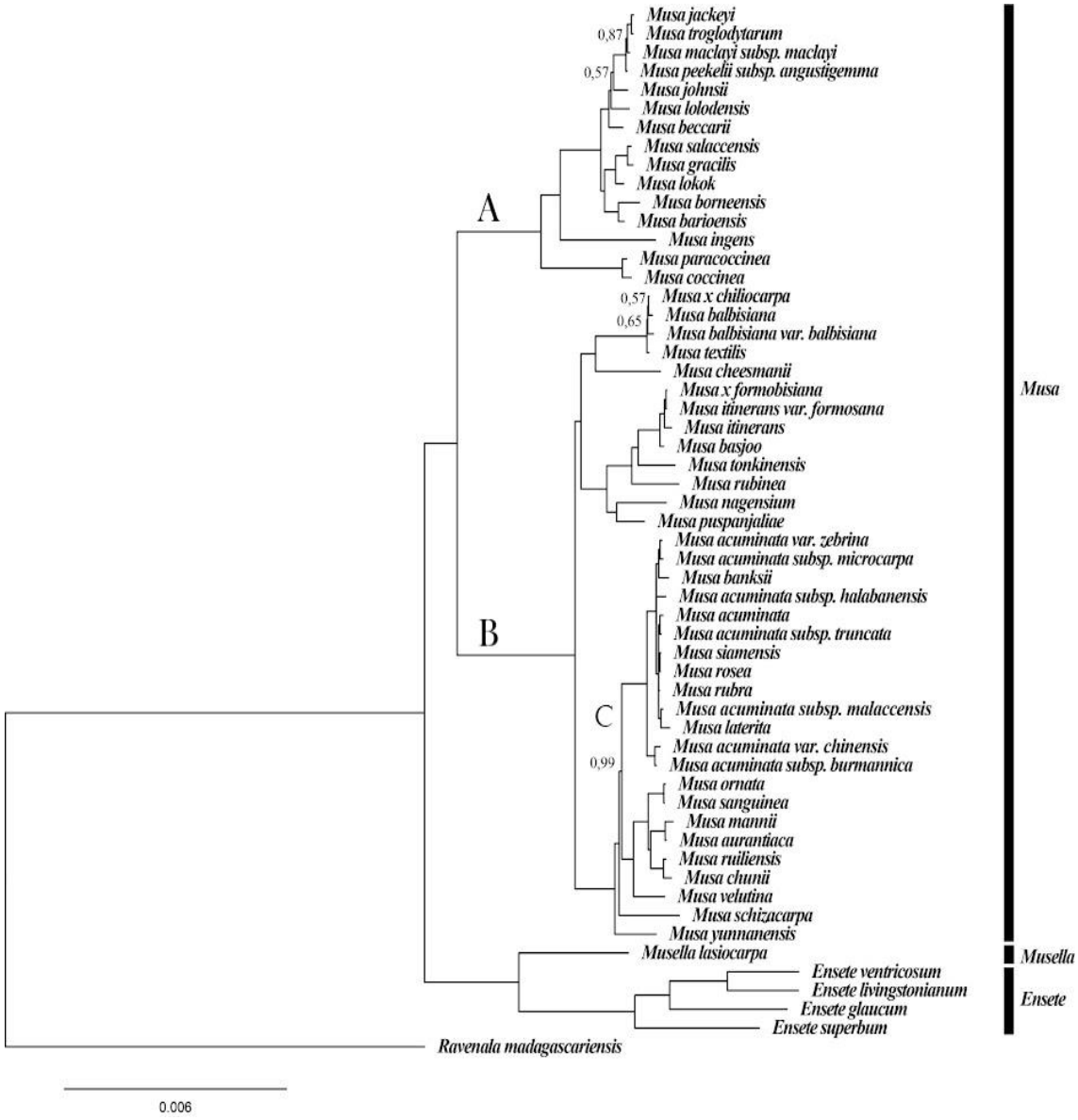

Figure 3. Phylogenetic relationship among species in three genera of the Musaceae

The number presented on the branches shows posterior probability. Unlabeled branches have their posterior probability equal to 1. The length of each branch shows genetic divergence. A and B show two clades of Musa species.

In recent years, SNPs discovery has been conducted for major trait mapping, genetic diversity, and population assessment in the Musaceae family (Till et al., 2010; Mmeka et al., 2013; Igwe et al., 2021; Rijzaani et al., 2022). Our results provide insight into SNPs and their distribution in the Musaceae plastid genome. The findings are beneficial for cross-genera comparative analysis of the phenotype-genotype correlation and gene modification research in the future.

Chloroplast genomes are known to be highly rich in AT content (Sayers et al., 2022). In some species, like liverwort *Marchantia polymorpha*, this content can reach more than 70% (Ohyama et al., 1986, 1996). It is possibly because of either bias gene conversion or mutational bias toward AT (Birdsell, 2002; Marais, 2003; Guo et al., 2021). The Musaceae has approximately 63% - 64% AT content in their plastid genomes. Nucleobase substitution trends consist of C to T, T to C (almost at the same ratio), and A-G conversion are recognized (Table 4). It could be considered bias conversion toward GC and T.

Bias gene conversion toward GC is often found in the nuclear genomes (Ibarra-Laclette et al., 2011; Kostka et al., 2012; Pessia et al., 2012). High GC content correlates with increased recombination and gene conversion activity (Lesecque et al., 2013; Arbeithuber et al., 2015; Mugal et al., 2015; Muyle

& Marais, 2016; Liu et al., 2018). The reasons and mechanism behind the trend of gene conversion in the Musaceae plastome might reflect the evolutionary direction of these species and require further studies to clarify.

3.2.2. Genetic Divergence in Musaceae Taxa

Plastid genomes of the Musaceae family are distinctly diverse, and most of the hypervariable regions are in LSC and SSC. IR regions are conservative in the number and structure of genes, except for the case of the hypervariable gene *ycf1*. Notable hypervariable areas of the plastome are *ndhA-ndhF*, *matK-rps16*, *psbD*, *psbE-petL*, and *petA* are promising candidates for biomarkers development along with the genes containing SNPs specific for a genus like *infA* (specific for the *Ensete* genus) and *petL* (specific for the *Musella* genus).

Genetic divergence in the Musaceae family might be helpful in the improvement and enhancement of desired traits in domesticated bananas thanks to the valuable genetic information they provide and the possibility of gene modification based on the beneficial characteristic of closely related species. The efficiency of closely related species discrimination is believed to be improved by combining multiple markers.

3.2.3. Differentiation of Inverted Repeat (IR) regions in Musaceae

Except for the case of *M. textilis*, only a slight difference in the sizes of IR regions was recognized in Musaceae species. However, since the average size of these regions in the angiosperms is approximately 25,000 bp, undoubtedly, expansion happened in IR regions of the Musaceae family.

Noticeable differences were noticed in the junction sites of LSC/IR and SSC/IR (Table 5). Genes transfer between regions of the plastome is usually caused by the rearrangement of junction sites, which are the consequences of the changes in the IR region's structure. Hypothetically, the extreme expansion of the IR region in the *Musa* genus was because of the additional complete gene integration, including *rps15*, *ndhD*, and *ycf1*, and a part of *ndhA* genes (approximately 1030 bp long) (Martin et al., 2013). These events have been proven to happen at Musaceae's SSC/IRa boundaries and are considered the largest IR extension observed in monocots (Martin et al., 2013).

Our results also confirm structural event occurrence in the chloroplast of the Musaceae and provide new insights into the evolutionary direction. Thus, IR regions are believed to have phylogenomic and evolutionary dynamics.

3.2.4. Phylogenetic relationship among Musaceae members

Phylogenetic trees can be constructed based on different genetic information. Phylogenetic trees of several species from the Musaceae family have been created based on Internal Transcribed Spacer (ITS) in their nuclear ribosome and a plastid gene *trnL-F* (Liu et al., 2010) or major chloroplast genes such as *matK* and *ndhF* (Probojati et al., 2021). A set of multiple nuclear gene sequences or the Musaceae's whole nuclear genome has also been conducted for phylogenetic tree construction (Christelová et al., 2011; Šimoníková et al., 2022). Recently, a phylogenetic tree of 45 species in three genera of Musaceae was generated based on their whole plastome and analyzed in detail (Fu et al., 2022).

Whole genome sequences are a better data resource for generating phylogenetic trees with high accuracy, especially in large families with many

species like Musaceae, since the accuracy of the tree is inversely proportional to the species diversity in the family. Moreover, the overall high conservation of plastome also helps ensure the precision of phylogeny analysis. Therefore, we conducted whole plastome sequences to build the phylogenetic tree.

Our findings from Musaceae phylogeny demonstrate the relationship in the Musaceae with minimal uncertainty in some branches. The *Ensete* genus is closely related to the monospecies genus *Musella* (sister taxa), and the *Musa* genus consists of two groups (Figure 3) that follow Hakkinen's sections (Häkkinen, 2013): (A) is *Callimusa* and (B) is *Musa*. Further studies are recommended to clarify Musaceae's phylogenesis better and identify the species divergence time.

4. CONCLUSION

The results show genetic diversity in the plastome of the Musaceae family and manifest the availability of promising biomarkers for closely related species discrimination. We also provide insights into the key features of the Musaceae's chloroplast genome (structure, gene content, structural variation, genetic diversity)

Evolutionary trends of the plastome have been evidenced by the expansion of IR regions, the addition of genes into these regions, and the bias nucleotide substitution towards G, C, and T. Further studies are recommended to explain the reasons, the mechanism behind these phenomena and to clarify the correlation between SNPs and traits. It is believed that the strong support for phylogeny analysis and breeding research will come from the surge in genome sequencing technology development, as more whole genomes of Musaceae chloroplast and nucleus are being sequenced.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this study.

ACKNOWLEDGMENTS

This research was funded by Nguyen Tat Thanh University (Ho Chi Minh City, Vietnam), under a project coded 2022.01.127/HĐ-KHCN to Mach Bao Ngoc.

REFERENCE

Ahmad, F., Martawi, N. M., Poerba, Y. S., De Jong, H., Schouten, H., & Kema, G. H. J.. (2020). Genetic mapping of Fusarium wilt resistance in a wild banana Musa acuminata ssp. malaccensis accession. *Theoretical and Applied Genetics, 133*(12), 3409–3418. https://doi.org/10.1007/s00122-020-03677-y

Amorim, E. P., dos Santos-Serejo, J. A., Amorim, V. B. O., Ferreira, C. F., & Silva, S. O. (2011, October). Banana breeding at Embrapa cassava and fruits. *In VII International Symposium on Banana: ISHS-ProMusa Symposium on Bananas and Plantains: Towards Sustainable Global Production 986* (pp. 171-176).

Arbeithuber, B., Betancourt, A. J., Ebner, T., & Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences of the United States of America, 112*(7), 2109–2114. https://doi.org/10.1073/pnas.1416622112

Birdsell, J. A. (2002). Integrating Genomics, Bioinformatics, and Classical Genetics to Study the Effects of Recombination on Genome Evolution. *Molecular Biology and Evolution, 19*(7), 1181–1197. https://doi.org/10.1093/oxfordjournals.molbev.a004176

Brooks, S., Gelman, A., Jones, G., & Meng, X. L. (Eds.). (2011). *Handbook of Markov chain monte Carlo. CRC press.*

Christelová, P., Valárik, M., Hřibová, E., De Langhe, E., & Doležel, J.. (2011). A multi gene sequence-based phylogeny of the Musaceae (banana) family. *BMC Evolutionary Biology, 11*(1), 103. https://doi.org/10.1186/1471-2148-11-103

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M., Da Silva, C., Jabbari, K., Cardi, C., Poulain, J., Souquet, M., Labadie, K., Jourda, C., Lengellé, J., Rodier-Goud, M., … Wincker, P.. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature, 488*(7410), 213–217. https://doi.org/10.1038/nature11241

Darling, A. E., Mau, B., & Perna, N. T.. (2010). progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLOS ONE, 5*(6), e11147. https://doi.org/10.1371/journal.pone.0011147

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D.. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods, 9*(8), 772–772. https://doi.org/10.1038/nmeth.2109

Dobrogojski, J., Adamiec, M., & Luciński, R.. (2020). The chloroplast genome: a review. *Acta Physiologiae Plantarum, 42*(6). https://doi.org/10.1007/s11738-020-03089-x

Drenth, A., & Kema, G. (2021). The Vulnerability of Bananas to Globally Emerging Disease Threats. *Phytopathology, 111*(12), 2146–2161. https://doi.org/10.1094/PHYTO-07-20-0311-RVW

Droc, G., Lariviere, D., Guignon, V., Yahiaoui, N., This, D., Garsmeur, O., Dereeper, A., Hamelin, C., Argout, X., Dufayard, J.-F., Lengelle, J., Baurens, F.-C., Cenci, A., Pitollat, B., D'Hont, A., Ruiz, M., Rouard, M., Bocs, S (2013). *The Banana Genome Hub Database*. DOI: http://10.1093/database/bat035

Duan, H., Guo, J., Xuan, L., Wang, Z., Li, M., Yin, Y., & Yang, Y.. (2020). Comparative chloroplast genomics of the genus Taxodium. *BMC Genomics, 21*(1). https://doi.org/10.1186/s12864-020-6532-1

Edgar, R. C.. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Food and Agriculture Organization of the United Nations (FAO) (n.d.). FAOSTAT (2021). *Statistics | Food and Agriculture Organization of the United Nations*. https://www.fao.org/faostat/en/#home

Food and Agriculture Organization of the United Nations (FAO) (n.d.). FAOSTAT (2022). *Statistics | Food and Agriculture Organization of the United Nations*. https://www.fao.org/faostat/en/#home

Fu, N., Ji, M., Rouard, M., Yan, H.-F., & Ge, X.-J.. (2022). Comparative plastome analysis of Musaceae and new insights into phylogenetic relationships. *BMC Genomics, 23*(1). https://doi.org/10.1186/s12864-022-08454-3

Geneious Prime 2022.2. (2022, June 14). Geneious | *Bioinformatics Software for Sequence Data Analysis*. https://www.geneious.com/

Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology, 52*(5), 696–704. https://doi.org/10.1080/10635150390235520

Guo, Y., Yang, J., Li, H., & Zhao, H. (2021). Chloroplast Genomes of Two Species of Cypripedium: Expanded Genome Size and Proliferation of AT-Biased Repeat Sequences. *Frontiers in Plant Science, 12*. https://doi.org/10.3389/fpls.2021.609729

Häkkinen, M. (2013). Reappraisal of sectional taxonomy in Musa (Musaceae). *Taxon, 62*(4), 809–813. https://doi.org/10.12705/624.3

Heslop-Harrison, J. S., & Schwarzacher, T. (2007). Review: Domestication, Genomics and the Future for Banana. *Annals of Botany, 100*(5), 1073-1084. https://doi.org/10.1093/aob/mcm191

Ibarra-Laclette, E., Albert, V. A., Herrera-Estrella, A., & Herrera-Estrella, L.. (2011). Is GC bias in the nuclear genome of the carnivorous plant Utricularia driven by ROS-based mutation and biased gene conversion?. *Plant Signaling & Behavior, 6*(11), 1631–1634. https://doi.org/10.4161/psb.6.11.17657

Igwe, D. O., Ihearahu, O. C., Osano, A. A., Acquaah, G., & Ude, G. N.. (2021). Genetic Diversity and Population Assessment of Musa L. (Musaceae) Employing CDDP Markers. *Plant Molecular Biology Reporter, 39*(4), 801–820. https://doi.org/10.1007/s11105-021-01290-x

International Tropical Fruits Network. (2016, March 16). Banana - Nutritional Value. Retrieved March 21, 2023, from https://www.itfnet.org/v1/2016/03/banana-nutritional-value/

Jones, D. R. (2007, September). Disease and pest constraints to banana production. *In III International Symposium on Banana: ISHS-ProMusa Symposium on Recent Advances in Banana Crop Protection for Sustainable 828* (pp. 21-36).

Kostka, D., Hubisz, M. J., Siepel, A., & Pollard, K. S. (2012). The Role of GC-Biased Gene Conversion in Shaping the Fastest Evolving Regions of the Human Genome. *Molecular Biology and Evolution, 29*(3), 1047–1057. https://doi.org/10.1093/molbev/msr279

Kumar, K. P. S., Bhowmik, D., Duraivel, S., & Umadevi, M. (2012). Traditional and Medicinal Uses of Banana. *Journal of Pharmacognosy and Phytochemistry, 1*(3).

Leister, D. (2003). Chloroplast research in the genomic age. *Trends in Genetics, 19*(1), 47-56. https://doi.org/10.1016/S0168-9525(02)00003-3

Lesecque, Y., Mouchiroud, D., & Duret, L. (2013). GC-Biased Gene Conversion in Yeast Is Specifically Associated with Crossovers: Molecular Mechanisms and Evolutionary Significance. *Molecular Biology and Evolution, 30*(6), 1409–1419. https://doi.org/10.1093/molbev/mst056

Liu, H., Huang, J., Sun, X., Li, J., Hu, Y., Yu, L., Liti, G., Tian, D., Hurst, L. D., & Yang, S. (2018). Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nature Ecology and Evolution, 2*(1), 164–173. https://doi.org/10.1038/s41559-017-0372-7

Liu, Z., Kress, W. J., & Li, Z. (2010). Phylogenetic Analyses of the Banana Family (Musaceae) Based on Nuclear Ribosomal (ITS) and Chloroplast (trnL-F) Evidence. *Taxon, 59*(1), 20-28. https://doi.org/10.2307/27757047

M. Mmeka, E. C., Adesoye, A. I., Vroh, B. I., & Ubaoji, K. I. (2013). Single nucleotide polymorphism (SNP) markers discovery within Musa spp (plantain landraces, AAB genome) for use in beta carotene

(Provitamin A) trait mapping. *American Journal of Biology and Life Sciences, 1*(1), 11-19.

Marais, G. A. B. (2003). Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics, 19*(6), 330–338. https://doi.org/10.1016/s0168-9525(03)00116-1

Marin, D. H., Romero, R. A., Guzman, M., & Sutton, T. B. (2003). Black Sigatoka: an increasing threat to banana cultivation. *Plant disease, 87*(3), 208-222.

Martin, G., Baurens, F.-C., Cardi, C., Aury, J.-M., & D'Hont, A.. (2013). The Complete Chloroplast Genome of Banana (Musa acuminata, Zingiberales): Insight into Plastid Monocotyledon Evolution. *PLOS ONE, 8*(6), e67350. https://doi.org/10.1371/journal.pone.0067350

Moghaddam, M., Ohta, A., Shimizu, M., Terauchi, R., & Kazempour-Osaloo, S.. (2022). The complete chloroplast genome of Onobrychis gaubae (Fabaceae-Papilionoideae): comparative analysis with related IR-lacking clade species. *BMC Plant Biology, 22*(1). https://doi.org/10.1186/s12870-022-03465-4

Mugal, C. F., Weber, C. C., & Ellegren, H. (2015). GC-biased gene conversion links the recombination landscape and demography to genomic base composition. *BioEssays, 37*(12), 1317–1326. https://doi.org/10.1002/bies.201500058

Muyle, A., & Marais, G. (2016). Mating Systems in Plants, Genome Evolution and. *Encyclopedia of Evolutionary Biology*, 480-492. https://doi.org/10.1016/B978-0-12-800049-6.00320-6

Ohyama, K. (1996). Chloroplast and Mitochondrial Genomes from a Liverwort, Marchantia polymorpha—Gene Organization and Molecular Evolution—. *Bioscience, Biotechnology, and Biochemistry, 60*(1), 16–24. https://doi.org/10.1271/bbb.60.16

Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, S., Inokuchi, H., & Ozeki, H. (1986). Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. *Nature, 322*(6079), 572–574. https://doi.org/10.1038/322572a0

Palmer, J. D., Jansen, R. K., Michaels, H. J., Chase, M. W., & Manhart, J. R. (1988). Chloroplast DNA Variation and Plant Phylogeny. *Annals of the Missouri Botanical Garden, 75*(4), 1180-1206. https://doi.org/10.2307/2399279

Perrier, X., Langhe, E. D., Donohue, M., Lentfer, C., Vrydaghs, L., Bakry, F., Carreel, F., Hippolyte, I., Horry, P., Jenny, C., Lebot, V., Risterucci, M., Tomekpe, K., Doutrelepont, H., Ball, T., Manwaring, J., & Denham, T. (2011). Multidisciplinary perspectives on banana (Musa spp.) domestication. *Proceedings of the National Academy of Sciences of the United States of America, 108*(28), 11311-11318. https://doi.org/10.1073/pnas.1102001108

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., & Marais, G. a. B. (2012). Evidence for Widespread GC-biased Gene Conversion in Eukaryotes. *Genome Biology and Evolution, 4*(7), 675–682. https://doi.org/10.1093/gbe/evs052

Probojati, R. T., Listyorini, D., Sulisetijono, S., & Wahyudi, D. (2021). Phylogeny and estimated genetic divergence times of banana cultivars (Musa spp.) from Java Island by maturase K (matK) genes. *Bulletin of the National Research Centre, 45*(1), 1-13. https://doi.org/10.1186/s42269-021-00492-3

Rambaut, A., & Drummond, A. J. (2012). FigTree version 1.4. 0.

Rijzaani, H., Bayer, P. E., Rouard, M., Doležel, J., Batley, J., & Edwards, D.. (2022). The pangenome of banana highlights differences between genera and genomes. *The Plant Genome, 15*(1). https://doi.org/10.1002/tpg2.20100

Ronquist, F., Teslenko, M., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology, 61*(3), 539-542. https://doi.org/10.1093/sysbio/sys029

Royal Botanic Gardens Kew (n.d.). Musaceae Juss. Plant of the World Online. https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:30000849-2

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular biology and evolution, 34*(12), 3299-3302.

Sardos, J., Breton, C., Perrier, X., Carpentier, S., Paofa, J., Rouard, M., & Roux, N. (2022). Hybridization, missing wild ancestors and the domestication of cultivated diploid bananas. *Frontiers in Plant Science, 13*. https://doi.org/10.3389/fpls.2022.969220

Sarma, U., Govila, V. K., & Yadav, A. (2019). The traditional and therapeutic use of banana and its plant-based delicacies in ethnic Assamese cuisine and religious rituals from Northeast India. *Journal of Ethnic Foods, 7*(21). https://doi.org/10.1186/s42779-020-00053-5

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., … Sherry, S. T. (2022*). Database resources of the national center for biotechnology information. Nucleic acids research, 50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Simmonds, N. W., & Shepherd, K. (1955). The taxonomy and origins of the cultivated bananas. *Botanical Journal of the Linnean Society, 55*(359), 302-312.

Šimoníková, D., Čížková, J., Zoulová, V., Christelová, P., & Hřibová, E.. (2022). Advances in the Molecular Cytogenetics of Bananas, Family Musaceae. *Plants, 11*(4), 482. https://doi.org/10.3390/plants11040482

Thangavelu, R., Loganathan, M., Arthee, R., Prabakaran, M., & Uma, S. (2020). Fusarium wilt: a threat to banana cultivation and its management. *CABI Reviews, (2020)*, 1-24.

Till, B. J., Jankowicz-Cieslak, J., Sági, L., Huynh, O. A., Utsushi, H., Swennen, R., Terauchi, R., & Mba, C.. (2010). Discovery of nucleotide polymorphisms in the Musa gene pool by Ecotilling. *Theoretical and Applied Genetics, 121*(7), 1381–1389. https://doi.org/10.1007/s00122-010-1395-5

Zhang, Z., Zhang, D.-S., Zou, L., & Yao, C.-Y.. (2022). Comparison of chloroplast genomes and phylogenomics in the Ficus sarmentosa complex (Moraceae). *PLOS ONE, 17*(12), e0279849. https://doi.org/10.1371/journal.pone.0