International Journal of Agricultural Technology 2022Vol. 18(5):2109-2122
Available online http://www.ijat-aatsea.com
ISSN 2630-0192 (Online)

The complete chloroplast genome of *Ensete glaucum* (Roxb.) Cheesman

Nguyen, H. D.[1], Le, V. M.[2], Nguyen, P. A. T.[3], Dinh, H. N.[1], Phan, N. H.[1], Vu, M. T.[1] and Do, H. D. K.[1*]

[1]NTT Hi-Tech Institute, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam; [2]Research Center of Ginseng and Medicinal Materials, National Institute of Medicinal Materials, Ho Chi Minh City, Vietnam; [3] Department of Molecular Biotechnology, Biotechnology Research and Development Institute, Can Tho University, Can Tho City, Vietnam.

Nguyen, H. D., Le, V. M., Nguyen, P. A. T., Dinh, H. N., Phan, N. H., Vu, M. T. and Do, H. D. K. (2022). The complete chloroplast genome of *Ensete glaucum* (Roxb.) Cheesman. International Journal of Agricultural Technology 18(5):2109-2122.

Abstract *Ensete glaucum* (Roxb.) Cheesman, also called snow banana, originated in Asia and has ornamental and medicinal value. Results found its complete chloroplast genome, which is 168,483 bp in length and composed of a large single-copy region (LSC; 88,233 bp), a small single-copy region (SSC; 11,138 bp), and two inverted repeat regions (IR; 34,636 bp). The completely sequenced genome includes 135 coding regions of 87 protein-coding genes, 40 tRNAs, and 8 rRNAs. An analysis of repeat composition identified 31 simple sequence repeats and 44 long repeats, mostly in non-coding regions. Notably, the *ycf1* and *ycf2* genes contain various repeats within the coding sequences. A maximum likelihood phylogenetic analysis revealed a close relationship between *E. glaucum* and *Musella lasiocarpa* rather than *Musa* species. Within *E. glaucum*, the Vietnam sample had a chloroplast genome more similar to a sample from Taiwan than the Indian variety.

Keywords: Banana evolution, Musaceae, Phylogenetic relationship, Snow banana

Introduction

Most land plants possess three genomes, the nuclear, mitochondria and chloroplast (cpDNA) genomes (Dobrogojski *et al.*, 2020). The cpDNA is typically quadripartite and ranges from 16 to over 200 kb and contains approximately 130 genes that perform photosynthesis and related metabolism in the chloroplast (Daniell *et al*., 2016). The 1000 Plant Transcriptome and 10,000 Plant Genome Projects were recently initiated (Carpenter *et al*. 2019; Chen *et al*. 2011; Cheng *et al*. 2018). Data have been used to investigate various aspects of plants, such as phylogeny (Gitzendanner *et al*., 2018; One Thousand Plant Transcriptomes Initiative, 2019) and metabolite pathways (Chakraborty, 2018; He *et al*. 2016; Liu *et al*., 2020). Specifically, a one-

[*]**Corresponding Author:** Do, H. D. K.; **Email:** dhdkhoa@ntt.edu.vn

billion-year history of green plants was proposed from the genetic sequencing of representative plants, to help understand how plants dominate on Earth (Gitzendanner *et al*., 2018). The genomic data revealed the synthesis pathways of chemical compounds in medicinal plants and unknown metabolomics (Chakraborty, 2018; He *et al*. 2016). Another application of genomic data was to develop molecular markers for trait mapping, plant breeding and identification (Garrido-Cardenas *et al.,* 2018; Grover and Sharma, 2016; Henry, 2012; Semagn *et al.,* 2006; Vu *et al*., 2021). Such studies have revealed the essential role of genomic data for exploring the evolution of land plants and other living organisms on Earth.

Musaceae Juss. 1789 is a monocotyledonous family that includes three genera: *Musa* L. 1753 (82 species), *Musella* (Franch.) C. Y. Wu ex H. W. Li 1978 (one species), and *Ensete* Bruce ex Horan.1862 (seven species) (POWO 2021). Various genomic studies based on the internal transcribed spacer (ITS) and chloroplast gene sequences (*rps16*, *atpB-rbcL*, and *trnL-F*) have revealed the phylogenetic relationships among these three genera of Musaceae, with *Musella* closer to *Ensete* than to *Musa* (Li *et al*., 2010; Christelová *et al*., 2011; Wu *et al*., 2021; Feng *et al*., 2022). A divergence time analysis indicated that Musaceae arose approximately 69 Mya at the Cretaceous–Tertiary boundary (Christelová *et al*., 2011). The chloroplast genomes of *Musa*, *Musella*, and *Ensete* species have also been reported (Liu *et al*., 2018; Yemataw *et al*., 2018; Zhang *et al*., 2018; Feng *et al*., 2020; Song *et al*., 2022). The entire banana genome was sequenced and provided useful information for the hybridization of banana cultivars (Martin *et al*., 2016; Martin *et al*., 2020).

Within Musaceae, *Ensete glaucum* (Roxb.) Cheesman 1948, also called the snow banana or elephant foot banana by the Vietnamese, has a native range from central Nepal to Papuasia (POWO, 2021). This species is being used in various traditional remedies and as an ornamental plant (Inta *et al.,* 2013; Joga *et al*., 2021; Ochiai 2012). In addition, the microcrystalline cellulose in *E. glaucum* is a potential biomaterial for drug delivery (Pachuau *et al*., 2019). Such works revealed potential medical applications of *E. glaucum*. In this study, the complete chloroplast genome of *E. glaucum* was sequenced using Oxford Nanopore Technology (ONT) to enlarge the genomic data for further studies of *Ensete* in particular and Musaceae. The *E. glaucum* cpDNA was characterized by size, gene content, and repeats. The genomic data extracted from cpDNA sequences of *Ensete*, *Musa* and *Musella* were also used to reconstruct the phylogenetic relationships among species of Musaceae.

Materials and methods

Fresh *E. glaucum* leaves were collected in Ninh Thuan Province, Vietnam (108.767135E, 11.9903180N) and stored in liquid nitrogen. A specimen was deposited at the Research Center of Ginseng and Medicinal Materials under voucher number TTS-MK 231. Total genomic DNA was isolated using the modified CTAB method (Doyle and Doyle 1987). The qualified DNA extract was used to prepare a sequencing library with the SQK-LSK109 ligation sequencing kit (ONT), following the manufacturer's instructions and then sequenced using a single FLO-MIN106 (R9.4) flow cell on a MinION Mk1B device (ONT) for 30 h. The MinKNOW interface was used to monitor the sequencing process and raw reads were base called using Guppy 5 (ONT). Then the processed reads were assembled using Canu and Geneious Prime 2021.1 to complete the chloroplast genome sequence (Koren *et al*. 2017). The obtained chloroplast genome was annotated using Geneious Prime 2021.1 with the reference chloroplast genomes of *Musa ingens* (NCBI accession number MW864253), *Ensete glaucum* (NCBI accession number LC610748), and *Musella lasiocarpa* (NCBI accession number KY807173). The newly sequenced *E. glaucum* chloroplast genome was deposited to NCBI under accession number MZ856381. A map of the chloroplast genome was illustrated using OGDRAW (Greiner *et al.*, 2019).

To identify repeats in the *E. glaucum* chloroplast genome, Phobos embedded in Geneious Prime was used to find simple sequence repeats (SSRs) and REPuter was used to locate long repeats (Kurtz *et al*., 2001; Mayer, 2006-2010). For the SSRs, the features were set to a minimum length of 10 bp for mononucleotides, 12 bp for dinucleotides, 15 bp for trinucleotides, 16 bp for tetranucleotides, 20 bp for pentanucleotides, and 24 bp for hexanucleotides. For long repeats, a minimum length of 20 bp was selected to identify forward (direct), reverse, complementary, and palindromic repeats.

For phylogenetic analysis, 79 protein-coding regions in 15 chloroplast genomes of 12 Musaceae species (*Musa laterita* [NCBI accession number MW864255], *M. acuminata* subsp. *malaccensis* [HF677508], *M. yunnanensis* [MW864261], *M. itinerans* [NC_035723], *M. rubinea* [MW864259], *M. nagensium* [MW864258], *M. balbisiana* [NC_028439], *M. troglodytarum* [MW864260], *M. beccarii* [MK012089], *M. ingens* [MW864253], *Musella lasiocarpa* [KY807173 and LC610747] and *Ensete glaucum* [MZ286962, MZ856381 and LC610748]) were used together with that of *Ravenala madagascariensis* Sonn. 1782 (Strelitziaceae Hutch. 1934) as the outgroup. The sequences were aligned using MUSCLE in Geneious Prime (Edgar, 2004). jModeltest 2.0 was used to estimate the best model for the data matrix, which

was identified as the TVM+G model (Darriba *et al*., 2012). The IQ-TREE package was used to construct a phylogenetic tree using the maximum likelihood method with 1000 bootstrap replicates (Minh *et al*., 2020). The phylogenetic tree was illustrated manually using Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

Results

Features of the chloroplast genome

The ONT device produced 1,277,260 raw reads, which ranged from 41 to 27,789 bp. After polishing the raw data, 9,729 reads remained, with lengths from 310 to 19,378 bp. Among the corrected reads, 3,514 (range 338 to 19,024 bp) were assembled to complete the *E. glaucum* chloroplast genome with 91× coverage. The results revealed a typical quadripartite *E. glaucum* chloroplast genome (168,483 bp) that includes large (LSC; 88,233 bp) and small (SSC; 11,138 bp) single-copy regions separated by two inverted repeat regions (IR; 34,636 bp) (Figure 1). This genome sequence had 37% GC content and 135 coding regions, including 87 protein-coding genes, 40 tRNAs, and 8 rRNAs (Table 1). Of the coding regions, 19 parts were duplicated in the IR regions (*rpl2*, *rpl23*, *rps12*, *rps15*, *ndhB*, *ycf1*, *ycf2*, *trnH-GUG*, *trnI-CAU*, *trnL-CAA*, *trnV-GAC*, *trnI-Gau*, *trnR-ACG*, *trnN-GUU*, *rrn4.5*, *rrn5*, *rrn16*, and *rrn23*). The chloroplast genome newly sequenced in this study showed high similarity to the two *E. glaucum* genomes available on NCBI with accession numbers MZ283962 (99.1%) and LC610748 (99.9%). The junction between the LSC and IR regions of *E. glaucum* was located in the *trnH-GUG/rps19* intergenic space (IGS), whereas these junctions in *Musa* and *Musella* extended to the *rps19-rpl22* IGS. However, the LSC/IR junction at *trnH-GUG/rps19* IGS was also found in some *Musa* species, such as *M. beccarii* and *M. troglodytarum*.

The SSR analysis revealed that mononucleotide repeats accounted for 87% of the SSRs in the *E. glaucum* cpDNA (Figure 2A); dinucleotides and trinucleotides comprised 10% and 3%, respectively. There were no tetra-, penta-, or hexanucleotides. Most of the SSRs were located in non-coding regions (77%; Figure 2B). Of the SSRs, 23% were in the *ycf1*, *rps14*, and *rpoC2* coding sequences (Table 2). The *E. glaucum* SSRs ranged from 10 to 17 bp and were made of A and T nucleotides (Table 2).

Similar to the SSRs, the long repeats were mainly in non-coding regions (59%), and forward was the major type (Figure 3A and 3B). The coding regions containing long repeats include *ycf1*, *ycf2*, *psaA*, *psaB*, *accD*, *trnG-UCC*, *trnG-GCC*, *trnS-GCU*, *trnS-UGA*, *trnS-GGA*, *trnfM-CAU*, *trnP-UGG*, *trnF-GAA*,

trnV-UAC, *trnV-UAC*, and *trnA-UGC* (Table 2). The long repeats ranged from 20 to 48 bp (Table 2).

Table 1. Gene content of *Ensete glaucum* chloroplast genome

Groups of genes	**Names of genes**
Ribosomal RNAs	*rrn4.5(2x), rrn5(2x), rrn16(2x), rrn23(2x)*
Transfer RNAs	*trnA-UGC*(2x), trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG_UCC*, trnG-GCC, trnH-GUG(2x), trnI-CAU(2x),trnI-GAU*(2x), trnK-UUU*, trnL-UAA*, trnL-UAG, trnL-CAA(2x), trnfM-CAU, trnM-CAU, trnN-GUU(2x), trnP-UGG, trnQ-UUG, trnR-UCU, trnR-ACG(2x),trnS-GCU, trnS-UGA, trnS-GGA, trnT-GGU, trnT-UGU, trnV-UAC*, trnV-GAC(2x), trnW-CCA, trnY-GUA*
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*(2x), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
ATP-dependent protease subunit P	*clpP**
Envelope membrane protein	*cemA*
Large units of ribosome	*rpl2*(2x), rpl14, rpl16*, rpl20, rpl22, rpl23(2x), rpl32, rpl33, rpl36*
Small units of ribosome	*rps2, rps3, rps4, rps7(2x), rps8, rps11, rps12*(2x), 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(2x), ycf3*, ycf4, ycf15(2x)*

*- genes with introns; 2x-duplicated genes; $^{\Psi}$-pseudogenes.

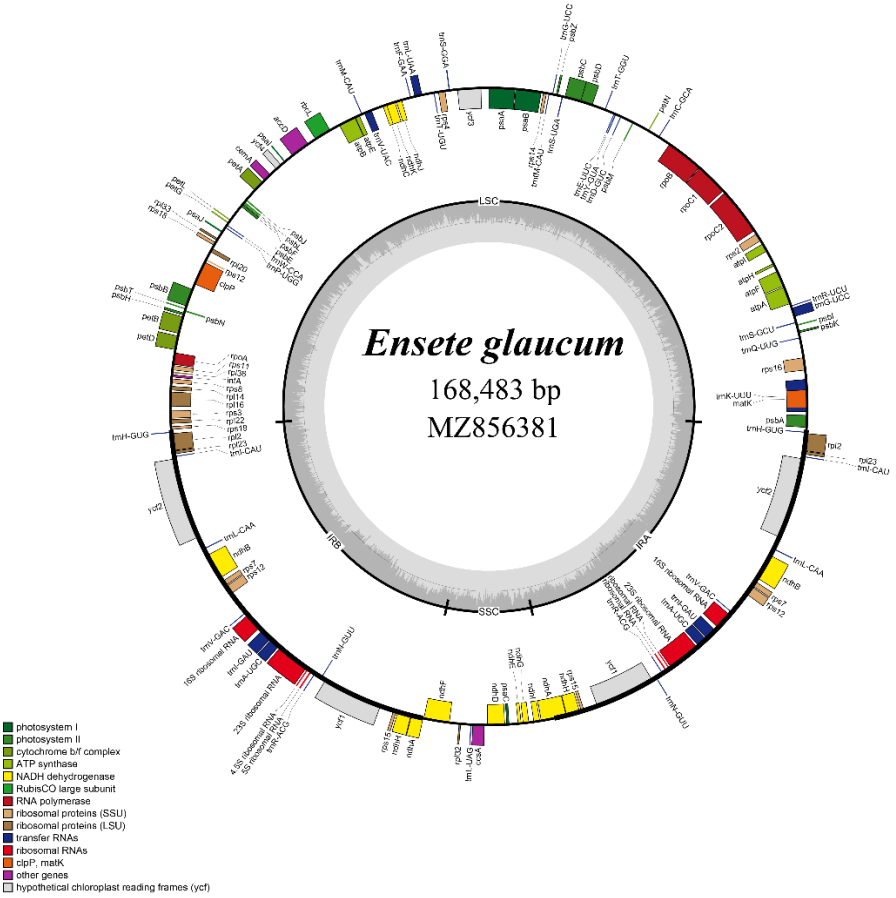

Figure 1. Map of the *Ensete glaucum* chloroplast genome: The genes inside (outside) the circle are transcribed clockwise (counterclockwise); The grey inside the circle indicates the GC content; LSC, large single copy; SSC, small single copy; IRA-IRB, inverted repeat regions

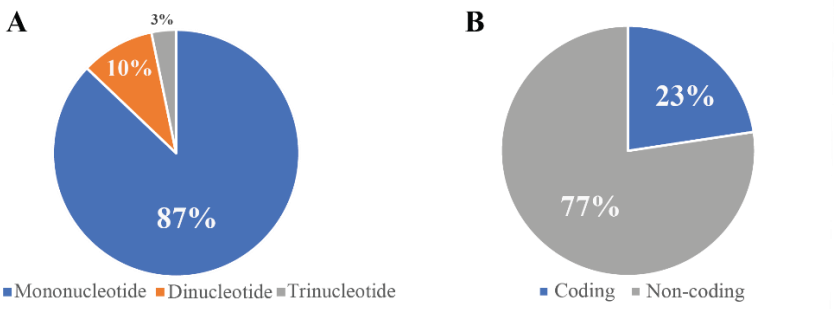

Figure 2. Percentage of SSRs in the *Ensete glaucum* chloroplast genome. A) Types and B) Locations of the SSRs

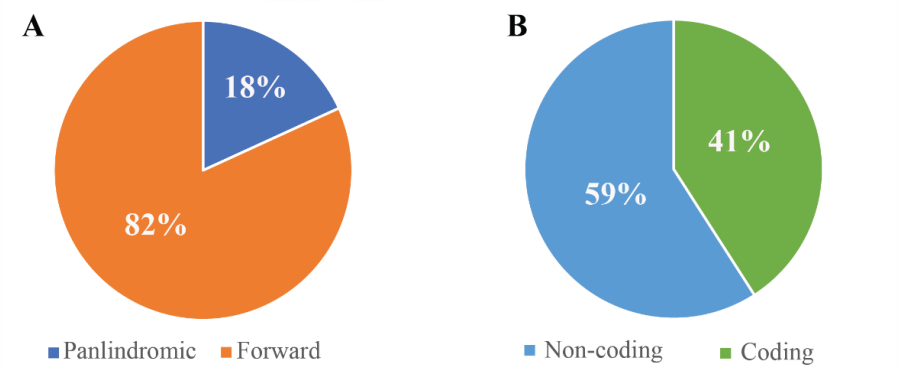

Figure 3. Percentage of long repeats in the *Ensete glaucum* chloroplast genome. A) Types and B) Locations of the repeats

Table 2. Features of repeats in chloroplast genome of *Ensete glaucum*

Type	Length	Location	Sequence
Dinucleotide	19	*rps12-clpP*	TATATATATATATATATAT
Dinucleotide	17	*trnF-GAA-ndhJ*	TATATATATATATATAT
Dinucleotide	12	*matK-trnK-UUU*	ATATATATATAT
Mononucleotide	16	*ycf1**	AAAAAAAAAAAAAAAA
Mononucleotide	15	*rps19-trnH-GUG*	TTTTTTTTTTTTTTT
Mononucleotide	14	*ndhA intron*	AAAAAAAAAAAAAA
Mononucleotide	13	*trnE-UUC-trnT-GGU*	TTTTTTTTTTTTT
Mononucleotide	13	*ycf1**	TTTTTTTTTTTTT
Mononucleotide	12	*atpB-rbcL*	TTTTTTTTTTTT
Mononucleotide	12	*ycf1**	AAAAAAAAAAAA
Mononucleotide	11	*rps16-trnQ-UUG*	AAAAAAAAAAA
Mononucleotide	11	*rps2-rpoC2*	TTTTTTTTTTT
Mononucleotide	11	*rps14**	TTTTTTTTTTT
Mononucleotide	11	*ycf3 intron*	AAAAAAAAAAA
Mononucleotide	11	*trnF-GAA-ndhJ*	AAAAAAAAAAA
Mononucleotide	11	*accD-psaI*	TTTTTTTTTTT
Mononucleotide	10	*matK-trnK-UUU*	AAAAAAAAAA
Mononucleotide	10	*rps16 intron*	AAAAAAAAAA

Table 2. (Con.)

Type	Length	Location	Sequence
Mononucleotide	10	*atpH-atpI*	TTTTTTTTTT
Mononucleotide	10	*rpoC2**	TTTTTTTTTT
Mononucleotide	10	*rpoC1 intron*	TTTTTTTTTT
Mononucleotide	10	*petN-psbM*	TTTTTTTTTT
Mononucleotide	10	*ycf3-trnS-GGA*	AAAAAAAAAA
Mononucleotide	10	*trnF-GAA-ndhJ*	AAAAAAAAAA
Mononucleotide	10	*trnF-GAA-ndhJ*	AAAAAAAAAA
Mononucleotide	10	*ycf4-cemA*	AAAAAAAAAA
Mononucleotide	10	*rpl22-rps19*	TTTTTTTTTT
Mononucleotide	10	*ycf1**	AAAAAAAAAA
Mononucleotide	10	*rpl32-trnL-UAG*	TTTTTTTTTT
Mononucleotide	10	*ycf1**	TTTTTTTTTT
Trinucleotide	15	*atpH-atpI*	TAATAATAATAATAA
Forward	39	*ycf2**	TGATAGTGACGATATCGATATTGATGATAGTGACGATAT
Forward	38	*rps12-trnV-GAC*	ACTATGAAATTAATATTTCTATAACTATGAAATTAATA
Forward	36	*ycf1**	GCGATGTAGAAAGTGAGGAAGAAAGCGATGTAGAAA
Forward	34	*psaB-psaA**	GTTCTATACATATGACCAGCGATCAGGAAAAGAA
Forward	29	*trnF-GAA-ndhJ*	AAATAATAATAAATAAGTTAAAAAAAAAA
Forward	28	*ycf2**	GCTAACTATGACGAATGCGCTAACTATG
Forward	26	*trnF-GAA-ndhJ*	TAAAAAAAAAATAATAATAAATAAGT
Forward	26	*ycf1**	TCAAGGCATAAGAAAATTAATAAGTC
Forward	26	*ycf1**	ATATTGAGAACAATAGTAATATTGAG
Forward	25	*trnT-UGU-trnL-UAA*	AATTATTTCTTAAAACTAACTATTT
Forward	25	*ycf4-cemA*	CCGAAAAGGACTCTTATTCTATGTC
Forward	25	*trnN-GUU-ycf1*	ATATAATAATCAAGAAATTGCAATA
Forward	24	*trnQ-UUG-psbK*	AACACATGTAGATTAGATATAGAA
Forward	24	*ycf2**	CCGAAATCTGATTCAAATCCAATA
Forward	24	*ycf1**	TGAAAATAATATTAGTGAAAATAA
Forward	23	*rps16-trnQ-UUG*	ATAATCTAGTTTATCTATTTTCA
Forward	22	*trnK-UUU intron*	AAGAATAGTTAGGATTCATTAA

Table 2. (Con.)

Type	Length	Location	Sequence
Forward	22	*trnG-UCC-trnG-GCC**	GATGCGGGTTCGATTCCCGCTA
Forward	22	*psaB-psaA**	GCAATATCGGTCAGCCATAAAC
Forward	22	*ndhC-trnV-UAC*	ATCCTATAATTAAATACTATGA
Forward	22	*rps19-trnH-GUG*	AGAAAATCCTTTAGCTAGAAAA
Forward	21	*trnK-UUU intron*	ACTTACATGAGCATTTCAGAA
Forward	21	*trnS-GCU-trnS-UGA-trnS-GGA**	AGAGAGGGATTCGAACCCTCG
Forward	21	*petN-psbM*	TTTCATTTCATTTTTTCATTT
Forward	21	*trnfM-CAU-trnP-UGG**	GACAGGATTTGAACCCGTGAC
Forward	36	*accD-psaI*	TTACTTATAAATAAATAATAT
Forward	21	*rps8-rpl14*	TTGATTATAAAATTATTTATT
Forward	21	*rpl16-rps3*	TTTATATAGTTATTAAGTTTA
Forward	21	*ycf1**	AGTTCTAGTTCTAAAACGAAT
Forward	20	*trnH-GUG-psbA*	AACAATATTGTATCAAACAA
Forward	20	*rps16 intron*	AACCTAAGACAAATTAGATT
Forward	20	*atpI-rpoC2*	TTTTGTTTTTTCTTTTTTTT
Forward	20	*petA-psbJ*	TCTTTTCTTGTTTCTTCGTA
Forward	48	*accD**	CAATGATTTCCATGAAGAAGTAGAGTCTGATTTCTATGAAGAAGTAGA
Forward	48	*ycf2**	CTTTTTGTCCAAGTCACTTCCCTTTTTGTCCAAGTCACTTCCCTTTTT
Forward	47	*ycf1**	AGTCAAAAGAAAAATGAAAATAATATCAGTCAAAAGAAAAATGAAAA
Palindromic	46	*rpl32-trnL-UAG*	TCTACTTTTCACAATAGAAAAATATTTTTCTATTGTGAAAAGTAGA
Palindromic	44	*psaC-ndhE*	ATATACATTAATAATGTATTATATAATACATTATTAATGTATAT
Palindromic	31	*petA-psbJ*	AAGAGTAAGAAAAGAACTCAACGGGACCTTA
Palindromic	26	*petD-rpoA*	ATTAATGTATCTAAGAATAGTGACTT
Palindromic	20	*trnH-GUG-psbA*	AAACAAAGTAGCAATACCCC
Palindromic	20	*trnF-GAA-trnV-UAC**	ATAGCTCAGTTGGTAGAGCA
Palindromic	20	*trnVUAC-trnA-UGC**	GCTCTACCAACTGAGCTATA
Palindromic	23	*petN-psbM*	ATAGTATGGTAGAAAGAATTATATATAATTCTTTCTACCATACTAT

Asterisks indicate the repeats in the coding regions

Phylogenetic relationships

The phylogenetic analysis showed the monophyly of *Musa*, *Musella*, and *Ensete* with high support (Figure 4). In Musaceae, *Ensete* was closer to *Musella* than to *Musa*. Within *Ensete*, the sample from Vietnam was more closely

related to a Taiwan sample (accession no. LC610748) than an Indian sample (accession no. MZ286962).

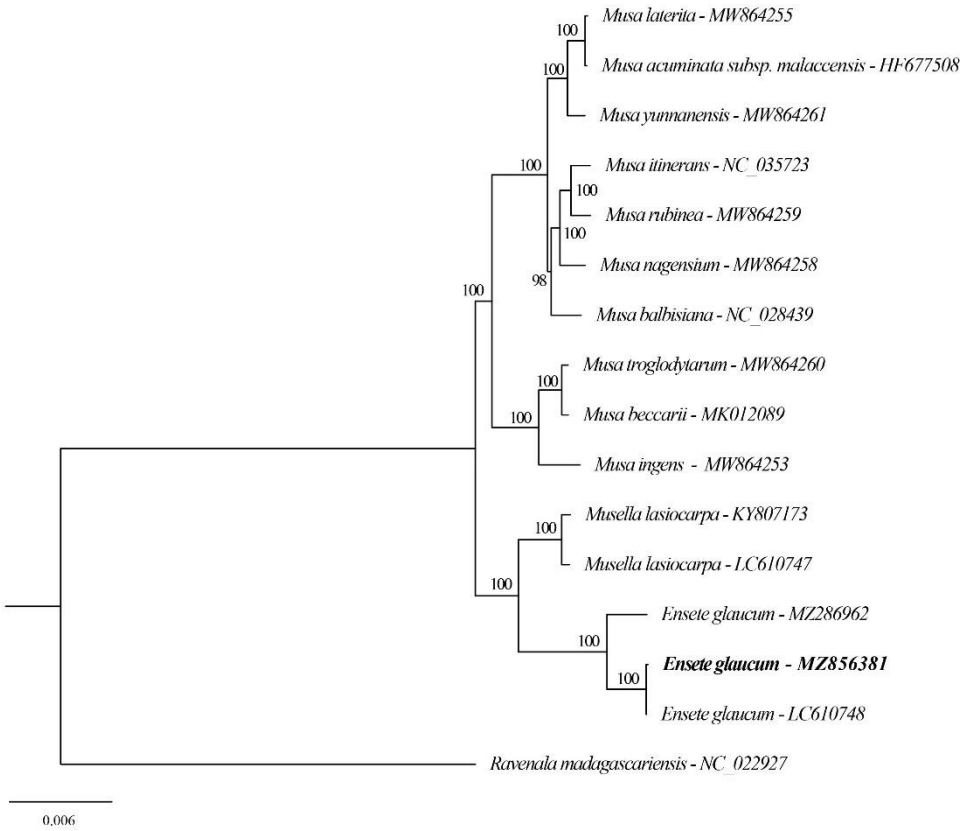

Figure 4. Phylogenetic tree of Musaceae species inferred from 79 protein-coding regions of chloroplast genomes. Numbers at the nodes are the bootstrap values. The chloroplast genome sequenced in this study is in bold italics

Discussion

The newly sequenced *E. glaucum* chloroplast genome has a quadripartite structure, as reported in the cpDNAs of other land plants and other Musaceae (Daniell *et al*., 2016; Liu *et al*., 2018; Yemataw *et al*., 2018; Zhang *et al*., 2018; Feng *et al*., 2020; Song *et al*., 2022). The gene content and order were similar among Musaceae species, with approximately 135 coding regions, including 87 protein-coding genes, 40 tRNAs, and 8 rRNAs. However, IR region expansion increased the gene number. There were three categories of junctions between IR and LSC regions in *Musa* species, in which the IR expanded to include *rpl2*, *rps19*, and partial *rps19/rpl22* IGS (Wu *et al*., 2021). Previously, the junctions

of IR and LSC regions of monocots were classified into five types based on the expansion of IR to *rpl2/trnH_GUG* IGS (type I), *trnH_GUG/rps19* IGS (type II), *rps19/rpl22* (type III), *rpl22/rps3* IGS (type IV), and from *rps3* (type V) (Do *et al*., 2020). In Musaceae, the *Musa* species have type I to III IR/LSC borders, whereas *E. glaucum* and *M. lasiocarpa* both have type II. The variable junctions in *Musa* species revealed the dynamic evolution of cpDNA among *Musa* species. Although *Ensete* includes seven species, only the *E. glaucum* cpDNA was characterized here. Therefore, further study should examine all members of *Ensete* to investigate the evolutionary history of the chloroplast genome within *Ensete*.

In the chloroplast genome, repeat sequences play an essential role in structural evolution. Extensive cpDNA rearrangements have been found in different species, such as Pinaceae and Campanulaceae (Cosner *et al*., 1997; Haberle *et al*., 2008; Wu *et al*., 2011). The SSRs can also be used to study genetic populations and molecular markers of plants (Powell *et al*., 1995; Godwin *et al*., 1997). In Musaceae, the cpDNA repeats have been analyzed in *Musa* species (Martin *et al*., 2013; Novák *et al*., 2014; Song *et al*., 2022). This study is the first to examine the repeat makeup in *Ensete glaucum* cpDNA. Although no cpDNA structural rearrangement has been reported in Musaceae, repeats provide useful information for further studies of the genetic evolution of Musaceae.

In this study, the phylogenetic analysis showed a close relationship between *Musella* and *Ensete* with high support values, as noted previously (Christelová *et al*., 2011; Janssens *et al*., 2016; Wu *et al*., 2021). Of the seven *Ensete* species, previous studies have used only one *E. glaucum* sample, while the current study included three *E. glaucum* cpDNAs to reconstruct the phylogeny of Musaceae. It showed that the sample from Vietnam (MZ856381) was more closely related to that from Taiwan (LC610748) than that from India (MZ286962). It is suggested that variability among *E. glaucum* chloroplast genomes in its native distribution from Central Nepal to Papuasia and provided valuable information for further genomic studies of *Ensete* species in particular and Musaceae in general.

This study reports the detailed features of the chloroplast genome of *E. glaucum*, including its genome size, gene content and order, SSRs, and long repeats. This information is essential for further studies of the evolutionary history of *Ensete* chloroplast genomes, and those of other Musaceae.

Acknowledgements

This work was funded by NTTU Foundation for Science & Technology, Nguyen Tat Thanh University under grant number 2021.01.161/HD-KHCN; and by Vingroup Joint Stock

Company through the Domestic Master/ PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), Vingroup Big Data Institute (VINBIGDATA) to Nguyen Hoang Danh under grant number VINIF.2021.ThS.58; and Department of Science and Technology of Ninh Thuan Province under grant number 11/2020/HĐ-SKHCN.

References

Carpenter, E. J., Matasci, N., Ayyampalayam, S., Wu, S., Sun, J., Yu, J., Vieira, F. R. J., Bowler, C., Dorrell, R. G., Gitzendanner, M. A., Li, L., Du, W., Ullrich, K. K., Wickett, N. J., Barkmann, T. J., Barker, M. S., Leebens-Mack, J. H. and Wong, G. K-S. (2019). Access to RNA-Sequencing Data from 1,173 Plant Species: The 1000 Plant Transcriptomes Initiative (1KP). GigaScience, 8:1-7.

Cosner, M., Jansen, R., Palmer, J. and Downie, S. R. (1997). The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. Curr Genet, 31:419-429.

Chakraborty, P. (2018). Herbal Genomics as Tools for Dissecting New Metabolic Pathways of Unexplored Medicinal Plants and Drug Discovery. Biochimie Open, 6:9-16.

Chen, S., Xiang, L, Guo, X. and Li, Q. (2011). An Introduction to the Medicinal Plant Genome Project. Frontiers of Medicine, 5:178-84.

Cheng, Shifeng *et al*. Cheng, S., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P-M., Li, F-W., Melkonian, B., Mavrodiev, E.V., Sun, W., Fu, Y., Yang, H., Soltis, D. E., Graham, S. W., Soltis, P. S., Liu, X., Xu, X. and Wong, G. K-S. (2018). 10KP: A Phylodiverse Genome Sequencing Plan. GigaScience, 7:1-9.

Christelová, P., Valárik, M., Hřibová, E., Langhe E. D. and Dolezel, J. (2011). A multi gene sequence-based phylogeny of the Musaceae (banana) family. BMC Ecology and Evolution, 11:103.

Daniell, H., Lin, C. S., Yu, M. and Chang W. J. (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biology, 17:134.

Darriba, D., Taboada, G. L., Doallo, R. and Posada, D. (2012). JModelTest 2: More Models, New Heuristics and Parallel Computing. Nature Methods, 9:772-772.

Do, H. D. K., Kim, C., Chase, M. W. and Kim, J. H. (2020). Implications of plastome evolution in the true lilies (monocot order Liliales). Molecular Phylogenetics and Evolution, 148:106818.

Dobrogojski, J., Adamiec, M. and Luciński, R. (2020). The chloroplast genome: a review. Acta Physiologiae Plantarum, 42:98.

Doyle, J. J. and Doyle, J. L. (1987). A Rapid DNA Isolation Procedure for Small Quantities of Fresh Leaf Tissue. Phytochemical bulletin, 19:11-15.

Edgar, R. C. (2004). MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity. BMC Bioinformatics, 5:113.

Feng, H., Chen, Y., Li, C., Xu, X., Luo, H. and He, C. (2022). Organelle DNA sequence data provide new insights into the maternal and paternal lineages of *Musa* species germplasms. Genetic Resources and Crop Evolution, 69:737-754.

Feng, H., Chen, Y., Xu, X., Luo, H., Wu, Y. and He, C. (2020) The complete chloroplast genome of *Musa beccarii*. Mitochondrial DNA Part B. 5:2384-2385.

Garrido-Cardenas, J.A., Mesa-Valle, C. and Manzano-Agugliaro, F. (2018). Trends in Plant Research Using Molecular Markers. Planta, 247:543-57.

Godwin, I. D., Aitken, E. A. B. and Smith, L. W. (1997), Application of inter simple sequence

repeat (ISSR) markers to plant genetics. ELECTROPHORESIS, 18:1524-1528.
Greiner, S., Lehwark, P. and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Research, 47:W59-W64
Grover, A. and Sharma, P. C. (2016). Development and Use of Molecular Markers: Past and Present. Critical Reviews in Biotechnology, 36:290-302.
Gitzendanner, M. A., Soltis, P. S., Wong, G. K., Ruhfel, B. R. and Soltis, D. E. (2018). Plastid Phylogenomic Analysis of Green Plants: A Billion Years of Evolutionary History. American Journal of Botany, 105:291-301.
Haberle, R. C., Fourcade, H. M., Boore, J. L. and Jansen, R. K. (2008). Extensive Rearrangements in the Chloroplast Genome of *Trachelium caeruleum* Are Associated with Repeats and tRNA Genes. Journal of Molecular Evolution, 66:350-361.
He, Y., Xiao, H., Deng, C., Xiong, L., Nie, H. and Peng, C. (2016). Survey of the Genome of *Pogostemon* Cablin Provides Insights into Its Evolutionary History and Sesquiterpenoid Biosynthesis. Scientific Reports, 6:26405.
Henry, R. J. (2012). Evolution of DNA Marker Technology in Plants. In: Robert J. Henry ed. Molecular Markers in Plants, Oxford, Blackwell Publishing Ltd., pp. 1-19.
Inta, A., Trisonthi, P. and Trisonthi, C. (2013). Analysis of Traditional Knowledge in Medicinal Plants Used by Yuan in Thailand. Journal of Ethnopharmacology, 149:344-51.
Janssens, S. B., Vandelook, F., De Langhe, E., Verstraete, B., Smets, E., Vandenhouwe, I. and Swennen, R. (2016). Evolutionary dynamics and biogeography of Musaceae reveal a correlation between the diversification of the banana family and the geological and climatic history of Southeast Asia. New Phytol, 210:1453-1465.
Joga, R. J., Sangma, E., Karmakar, B., Lyngdoh, V. and Aochen, C. (2021). Phytochemical Investigations on the Therapeutic Properties of *Ensete Glaucum* (Roxb.) Cheesman. Indian Journal of Traditional Knowledge, 20:68-73.
Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H. and Phillippy, A. A. (2017). Canu: Scalable and Accurate Long-Read Assembly via Adaptive k -Mer Weighting and Repeat Separation. Genome Research, 27:722-36.
Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J. and Giegerich, R. (2001). REPuter: The Manifold Applications of Repeat Analysis on a Genomic Scale. Nucleic Acids Research, 29:4633-4642.
Li, L. F., Häkkinen, M., Yuan, Y. M., Hao, G. and Ge, X. J. (2010). Molecular phylogeny and systematics of the banana family (Musaceae) inferred from multiple nuclear and chloroplast DNA fragments, with a special reference to the genus Musa. Molecular Phylogenetics and Evolution, 57:1-10.
Liu, J., Gao, C. W. and Niu, Y. F. (2018). The complete chloroplast genome sequence of flowering banana, *Musa ornata*. Mitochondrial DNA Part B Resources, 3:960–961.
Liu, X., Zhu, X., Wang, H., Liu, T., Cheng, J. and Jiang, H. (2020). Discovery and Modification of Cytochrome P450 for Plant Natural Products Biosynthesis. Synthetic and Systems Biotechnology, 5:187-99.
Martin, G., Baurens, F. C., Cardi, C., Aury, J. M. and D'Hont, A. (2013). The Complete Chloroplast Genome of Banana (*Musa acuminata*, Zingiberales): Insight into Plastid Monocotyledon Evolution. PLOS ONE, 8:e67350.
Martin, G., Baurens, F. C., Droc, G., Rouard, M., Cenci, A., Kilian, A., Hastie, A., Doležel, J., Aury, J-M.,Alberti, A., Carreel, F. and Angélique D'Hont, A. (2016). Improvement of the banana "*Musa acuminata*" reference sequence using NGS data and semi-automated bioinformatics methods. BMC Genomics, 17:243.
Martin, G., Cardi, C., Sarah, G., Ricci, S., Jenny, C., Fondi, E., Perrier, X., Glaszmann, J.-C.,

D'Hont, A. and Yahiaoui, N. (2020). Genome ancestry mosaics reveal multiple and cryptic contributors to cultivated banana. Plant Journal, 102:1008-1025.

Mayer, C. (2006-2010). Phobos 3.3.12. Retrieved from http://www.rub.de/ecoevo/cm/cm_phobos.htm.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A. and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution, 37:1530-34.

Novák, P., Hřibová, E., Neumann, P., Koblížková, A., Doležel, J. and Macas, J. (2014) Genome-Wide Analysis of Repeat Diversity across the Family Musaceae. PLOS ONE, 9:e98918.

Ochiai, Y. (2012). From Forests to Homegardens: A Case Study of *Ensete Glaucum* in Myanmar and Laos. Tropics, 21:59-66.

One Thousand Plant Transcriptomes Initiative. (2019). One Thousand Plant Transcriptomes and the Phylogenomics of Green Plants. Nature, 574:679-685.

Pachuau, L., Dutta, R. S., Hauzel, L., Devi, T. B. and Deka, D. (2019). Evaluation of Novel Microcrystalline Cellulose from Ensete Glaucum (Roxb.) Cheesman Biomass as Sustainable Drug Delivery Biomaterial. Carbohydrate Polymers, 206:336-643.

Powell, W., Morgante, M., McDevitt, R., Vendramin, G. G. and Rafalski, J. A. (1995). Polymorphic simple sequence repeat regions in chloroplast genomes: Applications to the population genetics of pines. The Proceedings of the National Academy of Sciences, 92:7759-63.

POWO. (2021). "Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew." Retrieved from http://www.plantsoftheworldonline.org.

Semagn, K., Bjørnstad, A. and Ndjiondjop, M. N. (2006). An Overview of Molecular Marker Methods for Plants. African Journal of Biotechnology, 5:2540-68.

Song, W., Ji, C., Chen, Z., Cai, H., Wu, X., Shi, C. and Wang, S. (2022). Comparative Analysis the Complete Chloroplast Genomes of Nine *Musa* Species: Genomic Features, Comparative Analysis, and Phylogenetic Implications. Frontiers in Plant Science, 13:832884

Vu, T. N., Pham, L. B. H., Nguyen, N. L., Luu, H. L., Huynh, T. T. H., Nguyen, H. H., Ha, H. H. and Le, T. T. H. (2021). Molecular Markers for Analysis of Plant Genetic Diversity. Vietnam Journal of Biotechnology, 18:589-608.

Wu, C. S., Lin, C. P., Hsu, C. Y., Wang, R. J. and Chaw, S. M. (2011). Comparative Chloroplast Genomes of Pinaceae: Insights into the Mechanism of Diversified Genomic Organizations. Genome Biology and Evolution, 3:309-319.

Wu, C. S., Sudianto, E., Chiu, H. L., Chao, C. P. and Chaw, S. M. (2021). Reassessing Banana Phylogeny and Organelle Inheritance Modes Using Genome Skimming Data. Frontiers in Plant Science, 12:713216.

Yemataw, Z., Muzemil, S., Ambachew, D., Tripathi, L., Tesfaye, K., Chala, A., Farbos, A., O'Neill, P., Moore, K., Grant, M. and Studholme, D. J. (2018). Genome sequence data from 17 accessions of *Ensete ventricosum*, a staple food crop for millions in Ethiopia. Data in Brief, 18:285-293.

Zhang, L., Guo, X., Wang, Z., Wang, M. and Hu, Q. (2018). Characterization of the complete chloroplast genome of *Musella lasiocarpa*. Mitochondrial DNA Part B: Resources, 3:728-729.

(Received: 13 April 2022, accepted: 30 July 2022)