Complete Chloroplast Genome Sequence of *Musa balbisiana* Corroborates Structural Heterogeneity of Inverted Repeats in Wild Progenitors of Cultivated Bananas and Plantains

Santoshkumar M. Shetty, Maria Ulfia Md Shah, Kavyashree Makale, Yusmin Mohd-Yusuf, Norzulaani Khalid, and Rofina Yasmin Othman*

**Abstract**

Complete genome sequencing of cytoplasmically inherited chloroplast DNA provides novel insights into the origins of clonally propagated crops such as banana and plantain (*Musa* spp.). This study describes the structural organization of the chloroplast genome of *M. balbisiana* Colla and its phylogenetic relationship with other wild progenitors of the domesticated banana cultivars. The *M. balbisiana* chloroplast genome was sequenced using Illumina HiSeq 2000 platform, followed by a combination of de novo short-read assembly and reference-guided mapping of contigs to generate complete plastome sequence. The *M. balbisiana* chloroplast genome is 169,503 bp in length, exhibits a typical quadripartite structural organization with a large single-copy (LSC; 87,828 bp) region and a small single-copy (SSC; 11,547 bp) region interspersed between inverted repeat (IRa/b; 35,064 bp) regions. Overall, its gene content, size, and gene order were identical to that of *M. acuminata* Colla with extensive expansion of the inverted repeat–small single-copy (IR-SSC) junctions. Comparative analyses revealed the conserved IRa-SSC expansion in three wild *Musa* species and members of the order Zingiberales. In contrast, IRb-SSC expansion was conspicuously absent in the sister taxon *M. textilis* Nee and related species of Zingiberales. Interestingly, phylogenomic assessment based on whole-plastome and protein-coding gene sets have provided robust support for the association of *M. balbisiana* and *M. textilis* as a sister group, despite the variation in IRb-SSC expansion. Although the current study substantiates the infrageneric IRb-SSC fluctuations in Musaceae, extensive taxon sampling is necessary to confirm whether the accessions of section *Musa* have undergone independent IRb-SSC expansion relative to section Callimusa.

**Core Ideas**

- First structural description of complete nucleotide sequence of *M. balbisiana* chloroplast genome
- Evolutionarily conserved IRa-SSC expansion events in the genus *Musa* and Zingiberales species
- Existence of infrageneric fluctuations in the IRb-SSC junction expansion among wild *Musa* species
- *M. balbisiana* and *M. textilis* share close evolutionary relationship albeit variable IRb-SSC expansion
- A significant addition to a growing body of *Musa* genomics resources

**Chloroplasts** are a quintessential feature of land plants and algae, harboring oxygencic photosynthetic machinery and an array of metabolic pathways. Believed to have originated from endosymbiotic associations between cyanobacteria and eukaryotic cells, chloroplasts have evolved complex mechanisms of intracellular gene regulation.


S.M. Shetty and M.U.M. Shah contributed equally to this work and are co-senior authors. Received 22 Sept. 2015. Accepted 8 Apr. 2016. *Corresponding author (yasmin@um.edu.my).

**Abbreviations:** cpDNA, chloroplast DNA; cpSSR, chloroplast SSR; IR, inverted repeat; LSC, large single-copy; NUPT, nuclear plastid DNA-like sequence; ORF, open reading frame; PCR, polymerase chain reaction; rRNA, ribosomal RNA; SH-aLRT, Shimodaira–Hasegawa approximate likelihood ratio; SSC, small single-copy; SSR, simple-sequence repeat; tRNA, transfer RNA.
transfer, protein translocation, and metabolite exchange (McFadden, 2001). In general, chloroplasts are maternally inherited and have their own genome, sizes ranging between 120 and 180 kb in flowering plants with a typical quadripartite structure (Provan et al., 2001). The first report on the complete genome sequence of chloroplasts was of the liverwort, *Marchantia polymorpha* L. (Ohyma et al., 1986) followed by tobacco (*Nicotiana tabacum* L.) (Shinozaki et al., 1986). The number of chloroplast genomes and transcriptomes sequenced has been dramatically increased with the advent of next-generation sequencing technologies (Stull et al., 2013; Small et al., 2013). Chloroplast genomes are targeted for gene transfer because of several unique advantages such as gene containment by maternal inheritance, high protein levels, site-specific integration of transgenes, and the lack of posttranscriptional gene silencing (Daniell et al., 2002; Bock, 2014). Both protein-coding genes and noncoding regions of chloroplast origin have been widely used to identify the parental origin of cultivars and in the taxonomic studies of plants (Shaw et al., 2007; Li et al., 2010, 2013). Chloroplast DNA microsatellite markers, either alone or in conjunction with nuclear microsatellites, have been extensively used to investigate the population genetic structure and phylogeography of plants (Ebert and Peakall, 2009). A targeted approach using the species-specific primers has proven to be more productive than the universal primers in identifying polymorphic loci in chloroplast genomes (Wheeler et al., 2014).

Banana and plantain are monocotyledonous plants, belonging to the genus *Musa* in the family Musaceae of the order Zingiberales. Banana ranks the fourth most important food crop, after rice (*Oryza sativa* L.), wheat (*Triticum aestivum* L.), and maize (*Zea mays* L.), with worldwide production of more than 106 Tg y⁻¹ (FAO-STAT, 2015). Based on the molecular analyses, wild banana species have been reclassified in to two sections, *Musa* (by merging *Eumusa* with *Rhodochlamys*) and *Callimusa* (x = 9, 10) with the inclusion of *Australimusa* and *Ingentimusa* (Häkkinen, 2013). Fertile diploid wild banana species have undergone hybridization to yield diploid (2n = 2x = 22), triploid (2n = 3x = 33), and rarely tetraploid (2n = 4x = 44) hybrids, which are widely distributed across the tropical and subtropical regions of Asia–Pacific and Africa (Perrier et al., 2011). Most of the dessert and cooking banana are intraspecific and interspecific hybrids of *M. acuminate* (A genome, 2n = 22) and *M. balbisiana* (B genome, 2n = 22), with main cultivated banana grouped into AA, AB, AAA, AAB, and ABB genomic groups (Simmonds and Shepherd, 1955; Heslop-Harrison and Schwarzrader, 2007). The *Musa Schizocarpa* N.W. Simmonds (S genome) (Cheesman, 1947) and *M. textilis* (T genome) (Lalusin and Villavicencio, 2015) are also considered as wild progenitors of cultivated banana and are endemic to Papua New Guinea and Philippines, respectively.

Primitive clones of *M. acuminate* and *M. balbisiana* are believed to have evolved in parallel in wetter Southeast Asian countries and drier parts of Asia (northeastern India–Myanmar and southern China) respectively (Perrier et al., 2011). While there were no distinct subspecies defined in *M. balbisiana*, numerous homologous and heterogenomic clones are abundant in the Indian subcontinent, which is a major center of diversity (Daniells et al., 2001, Uma et al., 2006). The *M. balbisiana* plants are tall and robust, have erect leaves, and the fruits are inedible and nonfleshy (Price, 1995; Thomas et al., 1998). Heterogenomic banana hybrids derived from *M. balbisiana* were found to be better adapted to withstand drought and extremes of temperatures than homologous cultivars originated from *M. acuminata* alone (De Langhe, 2002). Prior studies have concluded that in Musaceae chloroplast genomes are maternally inherited, whereas their mitochondrial genomes are inherited paternally (Fauré et al., 1994; Carreel et al., 2002). Sequencing of maternally inherited chloroplast genomes provides novel insights into the origin of clonally propagated crops such as banana cultivars. Recently, the sequencing and annotation of nuclear genomes of the wild accessions of *M. acuminate* subsp. *malaccensis* (Ridl.) N.W. Simmonds and *M. balbisiana* ‘Pisang Klutuk Wulung’ (PKW) have been completed (D’Hont et al., 2012; Davey et al., 2013).

Concomitantly, reports of the complete chloroplast genome sequences of *M. acuminate* subsp. *malaccensis* and *M. textilis* have revealed a typical quadripartite structure and variation in chloroplast genome size (Martin et al., 2013; Barrett et al., 2014). With the availability of nuclear genome sequences of the two wild banana species, there is an excellent opportunity to comparatively analyze the dynamics of chloroplast DNA transfer to nuclear genomes. Both simple-sequence repeats (SSRs) and minisatellites derived from chloroplast genome have been used to design polymorphic polymerase chain reaction (PCR) markers to differentiate among different *Musa* cultivars (Martin et al., 2013).

In monocots, IR expansion events were found to be correlated with the divergence pattern of monocot phylogeny (Wang et al., 2008). There is also speculation that the *M. acuminate* and related members of the order Zingiberales may have undergone independent expansions at their IRa and SSC junction (IRa-SSC) relative to Poales taxa (Martin et al., 2013). A preliminary comparative analysis of the available banana chloroplast genome sequences has revealed that variation in their size resulting from the unequal expansions at the IRb and SSC junction (IRb-SSC). In the absence of comparative studies, the question of whether these IR-SSC expansion events are congruent with the phylogenetic relationships among *Musa* species remains unanswered. Recently, using the plastome-wide protein-coding gene sets, Barrett et al. (2014) have elucidated the phylogenetic relationships among nine taxa representative of eight families of the order Zingiberales. Previous studies have used combination of nuclear and chloroplast genes to resolve the phylogenetic relationship among *Musa* cultivars (Li et al., 2010; Liu et al., 2010; Christelová et al., 2013).
2011). However, phylogenomic studies involving whole-chloroplast genomes of *Musa* cultivars are currently very limited. In the present study, we have used the whole-plastome as well as concatenated protein-coding gene sets to ascertain the phylogenetic relationship among three wild banana species and to other members of the order Zingiberales.

**Materials and Methods**

**DNA Extraction, Sequencing, and Genome Assembly**

Fresh leaf samples of *M. balbisiana* were collected from the germplasm collection of Rimba Ilmu Botanical Gardens of University of Malaya, Kuala Lumpur, Malaysia (3°7’53” N, 101°39’26” E). Chloroplast DNA was isolated using a modified high-salt buffer protocol by Shi et al. (2012) with the following modifications. Approximately 50 g of young leaves were collected, washed, and homogenized with Buffer A (1.25 M NaCl, 0.25 M ascorbic acid, 50 mM Tris·Cl [pH 8.0], 7 mM EDTA, 4% PVP-40 [w/v], 1% BSA, and 0.1% [v/v] β-mercaptoethanol [β-ME], pH 3.6). The homogenate was filtered through Miracloth (Calbiochem) and the filtrate was spun twice (300 g for 15 min at 4°C), to separate cell wall debris and nuclei pellet from chloroplast organelle. Chloroplasts were pelleted (3500 g for 15 min at 4°C) and washed twice with Buffer B (1.25 M NaCl, 50 mM Tris·Cl [pH 8.0], 25 mM EDTA, 4% PVP-40 [w/v], 1% BSA, and 0.1% [v/v] β-ME, pH 8.0) and suspended with Buffer C (100 mM NaCl, 100 mM Tris·Cl [pH 8.0], and 50 mM EDTA). Chloroplast organelles were lysed using a lysis buffer (20% sodium sarcosinate [w/v], 50 mM Tris·Cl [pH 8.0], and 25 mM EDTA) and treated with Proteinase K for protein removal. Chloroplast DNA (cpDNA) obtained was purified by phenol and chloroform extraction before pelleted and resuspended in 60 μL nuclease-free water. The purity of cpDNA samples was measured by using Nanodrop 2000 ultraviolet–visible spectrophotometer (Thermo Fischer Scientific); sample with A260/280 ratio >1.8 was processed for chloroplast genome sequencing.

Whole-chloroplast genome sequencing was outsourced to Beijing Genomics Institute (BGI-Shenzhen, Shenzhen, China), to generate 100-bp paired-end reads from a library of 500-bp insert size on an Illumina HiSeq 2000 platform (Illumina). Raw sequencing data was filtered using the following criteria: (i) reads with ambiguous bases (represented by letter N) >2% of bases or poly-A structure constituted were removed, (ii) reads that have 40 bases with quality score ≤7 for the library were filtered, (iii) reads with adaptor contamination were removed, and (iv) small insert-size reads in which Read 1 and Read 2 overlapped ≥10 bp and PCR duplications were removed. De novo assembly was performed using CLC Genomics Workbench 8.0 (http://www.clcbio.com) with parameter settings as the K-mer size of 22, bubble size of 50, mismatch cost of 2, insertion cost of 3, deletion cost of 3, length fraction of 0.5, and similarity fraction of 0.8. Contigs of <200 bp in length were excluded prior to scaffolding. Gaps in the scaffolds were subsequently filled by remapping the short reads iteratively using Geneious software version 7.5 (http://www.geneious.com; Kearse et al., 2012). The scaffolds with >1000× read mapping coverage were subjected to NCBI BLAST to determine their homology to chloroplast genomes. De novo scaffolds with positive relation to chloroplasts were ordered on to the reference plastome of *M. acuminata* subsp. *malaccensis* (HF677508). Evidently, scaffolds with highest read coverage represent inverted repeat regions of chloroplast genome and will be represented twice in complete genome assembly. Orientations of the ordered contigs were further assessed by iterative mapping at their junctions. Paired-end reads were remapped to consensus assembly with multiple iterations to fill gaps in final consensus sequence. Primers were designed across the junctions of contigs representing LSC, SSC, and IRa/b regions (Supplemental Table S1). The PCR reactions were performed using MJ Mini thermal cycler (Bio-rad Laboratories) in 25 μL reaction mixture containing 50 ng template DNA, 10 μM forward and reverse primers, and 2× PCRBI Taqmix Red (PCR Biosystems) containing Taq DNA polymerase (1 Unit), 6 mM MgCl₂, and 2 mM dNTPs. Thermocycler profile for amplification was 1 min of initial denaturation at 95°C, 40 cycles of 15 s at 95°C, 45 s at optimal primer annealing temperature, and 1 min for extension at 72°C, followed by final extension for 10 min at 72°C. Sanger dyeoxy sequencing of PCR amplicons was performed, and orientation of the contigs was confirmed by aligning the amplicon sequences across the representative contigs.

**Chloroplast Genome Annotation and Comparative Analyses**

Chloroplast genome annotation was performed on Dual Organellar GenoMe Annotator, DOGMA (http://dogma. ccbu.utexas.edu) to predict genes encoding proteins, ribosomal RNA (rRNA), and transfer RNA (tRNA) (Wyman et al., 2004). The initial annotations were further verified by comparing with homologous gene annotations of related chloroplast genes by nucleotide BLAST search. The tRNA gene annotations were further verified by tRNAscan-SE Search server (http://lowelab.ucsc.edu/ tRNAscan-SE/). A circular map of *M. balbisiana* chloroplast genome was subsequently drawn using web-based program, OGDRAW V1.2 (http://ogdraw.mpimp-golm.mpg. de/) (Lohse et al., 2013). Relative synonymous codon usage of *M. balbisiana* genes was examined using CodonW program (http://mobyle.pasteur.fr/cgi-bin/portal.py#jobs::CodonW). The mVISTA program (http://genome. lbl.gov/vista/mvista/submit.shtml) was used to compare the complete chloroplast genomes of *M. balbisiana* with *M. acuminata* and *M. textilis* (KF601567). Default parameters were used to align the chloroplast genomes in LAGAN program and sequence conservation profile was visualized using mVISTA plot (Brudno et al., 2003; Frazer et al., 2004). The IR-SSC expansion–contraction events in the
genus *Musa* and related species of the order Zingiberales were compared with that of basal angiosperm *Amborella trichopoda* Baill.

**Examination of Repeat Structure**

The detection of microsatellite SSRs was performed using MISA perl script (available at http://pgrc.ipk-gatersleben.de/misa/) with the following parameter settings: minimum threshold of 10 repeat units for mononucleotide, five repeat units for dinucleotides, four repeat units for trinucleotides, and three repeat units for tetra-, penta- and hexanucleotide repeats. Tandem repeats were detected by analyzing the plastome sequence in Tandem Repeat Finder web server (https://tandem.bu.edu/trf/trf.html), with match, mismatch and insertion–deletion parameters of 2, 7, 7 respectively, minimum alignment score of 80, and maximum period size of 500. Redundant and nested repeats were removed, whereas repeats located in the IR regions were counted only once. Plastid to nuclear transfer of organellar DNA fragments was investigated by nucleotide BLAST of *M. balbisiana* chloroplast genome (with one IR removed) against *M. balbisiana* nuclear genome database (http://banana-genome.cirad.fr/blast). Further, BLAST hits ≥100 bp in length and identity over 90% were used to exclude the short and fragmented sequences in downstream analyses (Yoshida et al., 2014).

**Phylogenomic Analyses**

Maximum likelihood phylogenetic analyses were conducted using complete chloroplast genome sequences of *M. balbisiana*, selected monocotyledons, basal angiosperms, and gymnosperms plastomes (Supplemental Table S2). In total, 28 selected monocot complete plastome sequences were aligned using MAFFT program version 5.0 (Katoh and Standley, 2013) at Trex webserver (http://www.trex.uqam.ca/). Poorly aligned regions in the chloroplast plastome nucleotide alignment were eliminated using the web-based version of the Gblocks program with options for less stringent selection of conserved blocks (Talavera and Castresana, 2007). The maximum likelihood phylogenetic inference of concatenated Gblocks alignment of chloroplast genomes was performed using RAxML version 8.2.4 (Stamatakis, 2014) at CIPRES Science Gateway (Miller et al., 2010). Analyses were performed using the general time reversible nucleotide substitution model under the GAMMA model of rate heterogeneity. The tree topology was assessed by rapid bootstrap analyses using 1000 pseudo-replicates to obtain branch support values (Stamatakis et al, 2008). Similarly, 38 shared protein-coding genes derived from 45 selected plant species were concatenated prior to alignment in MAFFT alignment. Phylogenetic trees were built using maximum likelihood tree estimation program PHYML (Guindon and Gascuel, 2003) with general time reversible nucleotide substitution model, and tree topology searching operation was performed using best of NNI and SPR option. Branch support values were evaluated using nonparametric Shimodaira–Hasegawa approximate likelihood ratios (SH-aLRT) statistic (Anisimova et al., 2011). Newick trees were edited in FigTree graphical viewer (version 1.4.2) to obtain publication quality phylogenetic trees (http://tree.bio.ed.ac.uk/software/figtree/).

**Results and Discussion**

**Genome Sequencing, Assembly, and Junction Validation**

In the present study, we report the complete chloroplast genome sequence of *M. balbisiana*. Illumina high-throughput sequencing of chloroplast genome of *M. balbisiana* generated 1995.79 Mb of data at an estimated sequencing depth of 13,305×. However, stringent quality control measures to remove the low-quality and ambiguous bases reduced the number of reads to 1836.28 Mb with a sequencing depth of 12,241×. The sequencing read data of *M. balbisiana* chloroplast genome can be accessed through NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra) with Bioproject ID PRJNA 286171. De novo assembly generated 92,426 contigs and 138 scaffolds, the majority of which contained gaps of variable lengths. Subsequently, five contigs with highest read mapping coverage and significant homology to chloroplast genomes sequences were selected for ordering on *M. acuminata* reference chloroplast genome for final assembly. It has been observed that the K-mer size of 22 produced optimal results by producing contigs corresponding near-perfect single-copy and IR segments. A single scaffold with consensus length of 87,992 bp, with total read count of 1,983,912 and average coverage of 2024 corresponding to the LSC, was identified. Two scaffolds with a consensus length of 2772 and 33,000 bp, with an average coverage of 3651× and 3561×, respectively, constituted the IR regions. The scaffolds corresponding to SSC showed total read count of 252,634 and an average coverage of 1878×. Subsequently, contigs were ordered based on *M. acuminata* reference plastome, and junctions between the single-copy and IR regions were verified by Sanger capillary sequencing of PCR amplicons. Further, the low-coverage regions and gaps in the draft assembly were resolved by PCR-based gap filling to obtain a complete sequence of *M. balbisiana* chloroplast genome.

**Structural Features of *Musa balbisiana* Chloroplast Genome**

Akin to other plastid genomes, *M. balbisiana* chloroplast genome shows a typical quadripartite structure with 169,503 bp in size, consisting of LSC (87,828 bp) and SSC (11,547 bp) regions separated by IR (35,064 bp) regions (Fig. 1). The complete chloroplast genome with gene annotation was deposited in NCBI GenBank with accession number KT595228. Remarkably, *M. balbisiana* SSC was relatively larger in size than that of *M. acuminata*, whereas LSC and IRa/b were smaller in size. Of the 113 genes annotated, 79 genes encode proteins, 30 genes encode tRNA, and four genes encode rRNA (Table 1). The genes repeated in IRs comprised of...
11 protein-coding genes, eight tRNA, and four rRNA genes. The presence of introns was detected in 17 genes, of which two genes (clpP and ycf3) had two introns each, whereas the rest possessed only a single intron each. The percentage of AT and GC nucleotides for the whole-chloroplast genome was 62.3 and 37.8%, respectively; LSC, SSC, and IR regions showed GC content of 35.14, 30.91, and 39.74%, respectively. Two copies of ycf15 and ycf68 genes with internal stop codons were annotated as pseudo genes in *M. balbisiana*. Two copies of *ndhA* genes were found across IR-SSC junctions with partially duplicated copy at IRb-SSC junction. Trans-splicing of rps12 gene was evident from 5’ exon located in LSC and two 3’ exons located in IR regions. Pseudogenisation of *ndhA* has been attributed to the incomplete duplication of the 5’ end of *ndhA* in IRa and SSC junction (Martin et al., 2013). The number and order of genic features of *M. balbisiana* chloroplast genome was found to be similar to those of *M. acuminata* chloroplast genome. Apparently, the size of the *M. balbisiana* chloroplast genome is larger than that of *M. textilis* and slightly smaller than *M. acuminata*. The high degree of conservation among these three wild *Musa* species is evident from the genome-wide comparison of the three chloroplast genome sequences as shown in the mVISTA plot (Fig. 2). Inverted repeat regions showed a higher degree of sequence conservation than the LSC and SSC regions. Despite the IRb-SSC fluctuations, *M. balbisiana* and *M. textilis* shared a high degree of sequence similarity, both in noncoding and protein-coding genes. In contrast, *M. acuminata* showed numerous insertions and deletions across the whole-chloroplast genome particularly in the intergenic regions and in ycf genes. The phylogenetic implications of these
plastome-wide sequence variations among the three Musa species are discussed later in this paper.

A total of 29,460 codons represent the coding capacity of 79 protein-coding genes in M. balbisiana (Table 2). Among these codons, 2977 (10.1%) encode for leucine and 344 (1.2%) for cysteine, which are the most and the least prevalent amino acids, respectively. Among these codons, 2977 (10.1%) encode for leucine and 344 (1.2%) for cysteine, which are the most and the least prevalent amino acids, respectively. The alternation of initiator codon ACG in the genes least prevalent amino acids, respectively. The alternation of initiator codon ACG in the genes

Table 1. List of annotated genes in Musa balbisiana chloroplast genome.

<table>
<thead>
<tr>
<th>Category</th>
<th>Group</th>
<th>Genes names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photosynthesis</td>
<td>Photosystem I</td>
<td>psaA, B, C, I, J</td>
</tr>
<tr>
<td></td>
<td>Cytochrome b6/f</td>
<td>petA, Bt, D, G, I, N</td>
</tr>
<tr>
<td></td>
<td>ATP synthase</td>
<td>atpA, B, F, H, J</td>
</tr>
<tr>
<td></td>
<td>Rubisco</td>
<td>ibc</td>
</tr>
<tr>
<td></td>
<td>NADH oxidoreductase</td>
<td>ndhA§, B§§, C, D, E, F, G, H§, I, I, K</td>
</tr>
<tr>
<td></td>
<td>Large subunit ribosomal proteins</td>
<td>rpl2§, 14, 16A, 20, 22, 23, 32, 33, 36</td>
</tr>
<tr>
<td></td>
<td>Ribosomal RNAs</td>
<td>tmr23§, 16§, 55§, 4.5§</td>
</tr>
<tr>
<td></td>
<td>Other proteins</td>
<td>accD, cca1, ccmA, clpP1, cemA, cemK, ndhA</td>
</tr>
<tr>
<td></td>
<td>Unknown function</td>
<td>ycf1§, ycf2§, ycf3§, ycf4§, ycf15§, ycf68§</td>
</tr>
</tbody>
</table>

† Gene containing two introns.
‡ Gene containing a single intron.
§ Two gene copies in the inverted repeats.
¶ Pseudogene.

The expansion of M. balbisiana IRa-SSC junction was analogous to the one observed in M. acuminate and considered to be the largest IR expansion observed in the monocotyledons (Martin et al., 2013). Interestingly, its sister taxon, M. textilis, and other representative species of the order Zingiberales, Heliconia collinsiana Griggs, Curcuma roscoaeana Wall., Ravenala madagascariensis Sonn., and Zingiber spectabile Griff., showed only IRA-SSC expansion relative to A. trichopoda but lacked the IRb-SSC expansion (Fig. 3). Martin et al. (2013) have previously suggested that the IRA-SSC expansion event may be of common occurrence in other Musaceae and related members of the Zingiberales. However, the absence of ndhA, ndhH, rps15 genes and pseudogenisation of the ycf1 gene at the IRb-SSC junction in M. textilis suggests that IRb-SSC expansion has occurred independently of IRA-SSC expansion in members of Musaceae and the order Zingiberales. We speculate that the constituent members of the order Zingiberales might have undergone independent IRA-SSC and IRb-SSC expansion events. In another study, Wang et al. (2008) has observed a positive correlation between the IRA/LSC expansion and divergence of the major monocotyledon lineages. It remains to be confirmed, however, the fluxes in the IRb-SSC expansion correlates with the diversification of the Musaceae lineages.

Identification and Characterization of Repeats

In total, 59 chloroplast SSRs (cpSSRs) were identified in the M. balbisiana chloroplast genome consisting of 29 mononucleotides, seven dinucleotides, five trinucleotides, 11 tetranucleotides, one pentanucleotide and hexanucleotide, and eight complex repeats (Supplemental Table S4). All but one of 20 homopolymer repeats detected were overwhelmingly represented by poly As and poly Ts. The majority of the predicted SSRs were located in the

Comparative Analyses Reveal Structural Heterogeneity of Inverted Repeat Regions in Musa Species

With the inclusion of ndhA, ndhH, rps15, and a full copy of ycf1 genes, the IRs of M. balbisiana show extensive expansion in to the SSC relative to A. trichopoda (Fig. 3).
Fig. 2. Visualization of the whole-chloroplast genome identity plots of *Musa balbisiana*, *M. acuminate*, and *M. textilis* as analyzed by mVISTA program. Peaks and valleys represent percentage sequence conservation.
untranslated, introns, and intergenic regions; while a few SSRs were detected in \textit{trnK-UUU}, \textit{ycf1}, and \textit{ycf2} gene-coding regions. Of the 23 tandem repeats identified, only nine repeats were located in genic regions (Supplemental Table S5). The unit length of the most repeated minisatellite was 24, which repeated nine times in tandem. The average repeat length of minisatellites was found to be 65.47 bp, whereas the minimum repeat length and maximum repeat length was 14 and 90 bp, respectively. The largest open reading frames (ORFs) of the chloroplast genome \textit{ycf1} and \textit{ycf2} were also the genes harboring the most number of sequence repeats. The A/T rich cpSSRs were also reported in \textit{Saccharum} and other related species of Poaceae (Melotto-Passarin et al., 2011). The overall composition of the cpSSRs in \textit{M. balbisiana} was within the range of those reported for related monocots and similar to those present within \textit{M. acuminata} subsp. \textit{malaccensis} chloroplast genome (Huotari and Korpe-lainen, 2012; Martin et al., 2013). The cpSSRs identified in this study could prove useful in designing PCR markers to ascertain the parental origins of allopolyploid cultivars of edible banana and plantain.

### Nuclear Chromosomal Localization of \textit{Musa balbisiana} Chloroplast DNA

Taking advantage of the availability of nuclear and chloroplast genome sequences of \textit{M. balbisiana}, the abundance of nuclear plastid DNA-like sequences (NUPTs) were analyzed. A total of 418 NUPTs (>100 bp in length) were detected in \textit{M. balbisiana} pseudo chromosomes (Fig. 4). The cumulative length of the \textit{M. balbisiana} NUPTs was 126,688 kb with largest NUPT length of 3694 bp. In total, 16 NUPTs >1 kb in length were detected, which were considerably higher than those observed in \textit{M. acuminata} (6 NUPTs). As expected, more than 50% of the NUPTs were originated from the LSC region; however, the highest number of NUPTs per gene was found in the largest ORFs \textit{ycf1} and \textit{ycf2} genes of IR region.

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>Codon</th>
<th>Number</th>
<th>RSCU†</th>
<th>tRNA</th>
<th>Amino acids</th>
<th>Codon</th>
<th>Number</th>
<th>RSCU†</th>
<th>tRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phe</td>
<td>UUU</td>
<td>1057‡</td>
<td>1.28</td>
<td>–</td>
<td>Asp</td>
<td>GAU</td>
<td>1009</td>
<td>1.61</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>UUC</td>
<td>596</td>
<td>0.72</td>
<td>tmf-GAA</td>
<td>GAC</td>
<td>241</td>
<td>0.39</td>
<td>tmf-GUC</td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>UUA</td>
<td>907</td>
<td>1.83</td>
<td>tmf-UAA</td>
<td>GAA</td>
<td>1199</td>
<td>1.46</td>
<td>tmf-UUC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UUG</td>
<td>627</td>
<td>1.26</td>
<td>tmf-CA</td>
<td>GAG</td>
<td>438</td>
<td>0.54</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUU</td>
<td>586</td>
<td>1.18</td>
<td>–</td>
<td>Pro</td>
<td>CCU</td>
<td>477</td>
<td>1.51</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>CUC</td>
<td>222</td>
<td>0.45</td>
<td>–</td>
<td>CCC</td>
<td>249</td>
<td>0.79</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUA</td>
<td>434</td>
<td>0.87</td>
<td>tmf-UAG</td>
<td>CCA</td>
<td>381</td>
<td>1.21</td>
<td>tmf-UGG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUG</td>
<td>201</td>
<td>0.41</td>
<td>–</td>
<td>CCG</td>
<td>155</td>
<td>0.49</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Ile</td>
<td>AUU</td>
<td>1218</td>
<td>1.44</td>
<td>–</td>
<td>Thr</td>
<td>ACU</td>
<td>583</td>
<td>1.52</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>AUC</td>
<td>508</td>
<td>0.60</td>
<td>tmf-GAU</td>
<td>ACC</td>
<td>300</td>
<td>0.78</td>
<td>tmf-GGU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUA</td>
<td>812</td>
<td>0.96</td>
<td>–</td>
<td>ACA</td>
<td>476</td>
<td>1.24</td>
<td>tmf-UGU</td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>AUG</td>
<td>691</td>
<td>1.00</td>
<td>tmf-CAU</td>
<td>ACG</td>
<td>173</td>
<td>0.45</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>GGU</td>
<td>570</td>
<td>1.42</td>
<td>–</td>
<td>Ala</td>
<td>GCU</td>
<td>671</td>
<td>1.82</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>GUC</td>
<td>215</td>
<td>0.54</td>
<td>tmf-GAC</td>
<td>GCC</td>
<td>233</td>
<td>0.63</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GUA</td>
<td>600</td>
<td>1.5</td>
<td>tmf-UC</td>
<td>GCA</td>
<td>419</td>
<td>1.14</td>
<td>tmf-UGC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GUG</td>
<td>216</td>
<td>0.54</td>
<td>–</td>
<td>GCG</td>
<td>148</td>
<td>0.40</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>GGU</td>
<td>616</td>
<td>1.30</td>
<td>–</td>
<td>Ser</td>
<td>AGU</td>
<td>499</td>
<td>1.26</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>GGC</td>
<td>188</td>
<td>0.40</td>
<td>tmf-GCC</td>
<td>AGC</td>
<td>120</td>
<td>0.30</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GGA</td>
<td>810</td>
<td>1.71</td>
<td>tmf-UCC</td>
<td>UCU</td>
<td>669</td>
<td>1.69</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GGG</td>
<td>285</td>
<td>0.60</td>
<td>tmf-GGG</td>
<td>UCC</td>
<td>382</td>
<td>0.97</td>
<td>tmf-S-GGA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UAA</td>
<td>16</td>
<td>1.12</td>
<td>–</td>
<td>UCA</td>
<td>474</td>
<td>1.20</td>
<td>tmf-S-GUA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UGA</td>
<td>17</td>
<td>1.19</td>
<td>–</td>
<td>UCG</td>
<td>225</td>
<td>0.57</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UAG</td>
<td>10</td>
<td>0.70</td>
<td>–</td>
<td>Arg</td>
<td>GGU</td>
<td>393</td>
<td>1.31</td>
<td>tmf-RAG</td>
</tr>
<tr>
<td></td>
<td>UGG</td>
<td>501</td>
<td>1.00</td>
<td>tmf-CCA</td>
<td>GGC</td>
<td>93</td>
<td>0.31</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UGU</td>
<td>252</td>
<td>1.47</td>
<td>tmf-GCA</td>
<td>GGA</td>
<td>408</td>
<td>1.36</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UGC</td>
<td>92</td>
<td>0.53</td>
<td>–</td>
<td>GGG</td>
<td>149</td>
<td>0.50</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UAU</td>
<td>864</td>
<td>1.56</td>
<td>–</td>
<td>AGA</td>
<td>578</td>
<td>1.93</td>
<td>tmf-S-GCU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UAC</td>
<td>241</td>
<td>0.44</td>
<td>tmf-GUA</td>
<td>AGG</td>
<td>176</td>
<td>0.59</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAU</td>
<td>536</td>
<td>1.54</td>
<td>–</td>
<td>Asn</td>
<td>AAU</td>
<td>1090</td>
<td>1.52</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>CAC</td>
<td>159</td>
<td>0.46</td>
<td>tmf-GUG</td>
<td>AAC</td>
<td>340</td>
<td>0.48</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAA</td>
<td>826</td>
<td>1.54</td>
<td>tmf-QUG</td>
<td>AAA</td>
<td>1189</td>
<td>1.49</td>
<td>tmf-K-UUU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAG</td>
<td>246</td>
<td>0.46</td>
<td>–</td>
<td>AAG</td>
<td>410</td>
<td>0.51</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>

† RSCU, relative synonymous codon usage.
‡ Numbers shown in italic indicate the most frequently used codon.

Table 2. Codon usage and codon–anticodon recognition pattern for transfer RNAs (tRNAs) in \textit{Musa balbisiana} Colla chloroplast genome.
Since the draft version of the M. balbisiana nuclear genome was used in the analyses, the number of NUPTs detected might be refined on the availability of the final genome annotation. In comparison, at least 563 NUPTs with cumulative length of 134,491 bp were reported in M. acuminata, which correspond to 0.41% of the nuclear genome (Martin et al., 2013).

In general, NUPTs have been found to be localized uniformly throughout the whole chromosomes in plants species with large genomes, whereas in small genomes, NUPTs tend to preferentially located in the regions surrounding the centromeres (Michalovová et al., 2013). The NUPT content in polyplastid species with multiple plastids per cell was found to be 80 times more than those with monoplastid species (Smith et al., 2011). The number of NUPTs was also found to be higher in species with larger nuclear genome sizes (Yoshida et al., 2014). Thus, polyploid Musa cultivars may have a higher percentage of NUPT content than those of their diploid progenitors.

**Phylogenomic Analyses**

A maximum likelihood phylogenetic reconstruction was performed with two data partitions, a set of monocotyledon whole-chloroplast genomes, and a set of concatenated protein-coding genes derived from monocotyledons, basal angiosperms, and gymnosperms. The RAxML inferred maximum likelihood phylogenetic tree of whole-chloroplast genomes showed an early diverging Acorus (Acorales) species from the rest of the monocotyledons followed by Colocasia and Lemna species of Alismatales (Fig. 5). The latter was positioned as a sister group to Lilium (Liliales) and Dioscorea (Dioscoreales) followed by Asparagales, which was placed as an early diverging sister taxa to the Commelinid clade. The latter was resolved into three monophyletic orders, namely Arecales, Poales, Zingiberales, and an unplaced family of Dasypogonaceae with bootstrap support value of 99%. The placement of Arecales as the sister taxon to Zingiberales is in congruence with the previous study describing monocotyledons phylogeny based on 93 nuclear single-copy genes (D’Hont et al., 2012). The inclusion of noncoding regions (introns and intergenic regions) of chloroplast genomes in phylogenetic reconstruction may have resulted in well-supported Arecales and Zingiberales sister group topology. Previous studies have proven the utility of noncoding regions of chloroplast genomes in increasing the resolution at low taxonomic levels (Shaw et al., 2007). On the other hand, a recent study involving a plastome-wide gene set coding for 75 proteins derived from 132 monocotyledon taxa has positioned Dasypogonaceae as a sister clade to Arecales and strongly supported the association between Zingiberales and Poales as sister groups (Barrett et al., 2016).
whole-plastome data has positioned Heliconia (Heliconiaceae) as sister taxon to Musaceae and Ravenala (Sterilitziaceae) with low bootstrap branch support value of 52, whereas Zingiberaceae (Curcuma and Zingiber) formed a well-supported separate clade. In contrast, phylogenetic relationships inferred from 81 plastome-wide protein-coding genes of Zingiberales species has placed Heliconia as an early diverging sister taxon and found weak support for deep relationship among constituent families of the order Zingiberales (Barrett et al., 2014).

In the Musaceae clade, M. balbisiana and M. textilis were grouped together in a single clade with robust branch support value (100%), while M. acuminata was positioned as their sister taxon. Apparently, the high levels of plastome-wide sequence conservation between M. balbisiana and M. textilis has undermined the influence of IRb-SSC fluctuations on the phylogenetic assessment of the three Musa species. This topology is incongruent to the tree topology derived from dense sampling of 33 wild banana species, wherein M. balbisiana has been placed in a single clade along with M. acuminata, whereas M. textilis was positioned in a separate clade with other members of section Callimusa (Li et al., 2010). Though current topology is based on large character sampling, limited taxon sampling has precluded the positioning of M. balbisiana as a primary sister taxon of M. textilis.

Mapping of IR-SSC expansion events using parsimony principle was performed to unveil the most likely scenarios of the IR-SSC fluctuation events in Musa species and in the representative species of the order Zingiberales. The IRA-SSC expansion might have occurred following the divergence of Arecales and Zingiberales, making it a common feature among wild Musa species and related members of the order Zingiberales. Martin et al. (2013) have suggested that M. acuminata and related species might have undergone independent IRA-SSC expansion relative to those belonging to the order Poales. On the other hand, the occurrence of IRb-SSC expansion events in both M. balbisiana and M. acuminata, and its absence in sister taxon M. textilis strongly supports the subgeneric variation in IRb-SSC expansion in the Musaceae family. In the absence of thorough taxon sampling, the parsimonious scenario of IRb-SSC expansion in Musa species was found to be incongruent with the phylogenetic relationships inferred from whole-chloroplast genome data. While the IRA-SSC fluctuations has been observed in all monocotyledons, the expansion of IRb-SSC is described only in Panicoideae, where IR is extended into 29 bp of ndhF gene (Guisinger et al., 2010). Adequate taxon sampling from the sections Musa and Callimusa, as well as from the genus Ensete, is necessary to substantiate the phenomenon of infrageneric variation in IRb-SSC expansion.

For a more comprehensive phylogenetic reconstruction, 38 protein-coding plastid genes derived from 45 selected plant species belonging to angiosperms and gymnosperms were subjected to maximum likelihood tree inference. The phylogenetic tree was rooted with gymnosperms (Pinus thunbergii Parl. and Cycas revoluta Thunb.) and a basal angiosperm A. trichopoda as outgroups and has resolved three distinct clades: monocotyledons, basal angiosperms, and gymnosperms with significant branch support values (Fig. 6). The positioning

---

Fig. 4. Chloroplast DNA insertions in Musa balbisiana nuclear genome. The x-axis represents 11 pseudochromosomes of M. balbisiana, and y-axis represents the number of plastid DNA insertions >100 bp in length (cumulative high-scoring sequence pair length).
The close phylogenetic relationship between *M. balbisiana* and *M. textilis* reported in this study is in agreement with the previous phylogenetic inference based on chloroplast *trnL-F* and nuclear ribosomal ITS genes of 38 *Musa* accessions (Liu et al., 2010). Phylogenetic assessment of *Musa* species based on multiple nuclear genes has also found *M. balbisiana* to be remarkably distinct from *M. acuminata* and other taxa belonging to the section *Musa* (Christelová et al., 2011). Our observation also supports earlier reports that found *M. balbisiana* to form a distinct clade or lineage relative to *M. acuminata* (Li et al., 2010). In another study involving the nuclear ribosomal ITS and chloroplast (*atpB-rbcL*, *rps16* and *trnL-F*) genes of 42 *Musa* accessions, *M. balbisiana* was placed alongside *M. acuminata* in a single clade albeit as a distinct lineage (Li et al., 2010). As a result of comprehensive
taxon sampling of Musa species, phylogenetic relationships described by the Li et al. (2010) were markedly different from the current study. Thus, it is imperative that the whole-plastome data sets must also be accompanied by an extensive taxon sampling to unequivocally resolve the phylogenetic relationships among Musa taxa.

Apparently, whole-plastome sequences of taxonomically different and geographically isolated species will be required to unravel the evolutionary dynamics of IR expansion–contraction and its phylogenomic implications in Musaceae. As observed in the previous studies, whole-plastome-wide data partitions used in the current analyses have failed to resolve and provide robust support for deep phylogenetic relationships among constituent families of Zingiberales and among commelinid orders (Barrett et al., 2013; Barrett et al., 2014). The lack of comprehensive taxon sampling across the Musa genus and variable number of phylogenetically informative characters in whole-plastome and protein-coding gene sets might have resulted in the contradicting phylogenetic inferences in the present study.

Conclusions

The first structural description of complete sequence of M. balbisiana plastome has provided the evidence to substantiate the occurrence of IR-SSC fluctuations in the wild Musa species. The current study has found that the IRA-SSC expansion is conserved among the three wild Musa species investigated. Furthermore, comparative analyses have confirmed that the other Zingiberales taxa have also undergone similar IRA-SSC expansion, further reaffirming their sister relationship status with Musaceae. In contrast, the expansion of IRb-SSC junction appears to be an independent event in the section Musa. Albeit lacking the IRb-SSC junction expansion, M. textilis showed high degree of sequence conservation with M. balbisiana chloroplast genome, which was reflected in the phylogenomic analyses involving both the whole-plastome and protein-coding gene sets. We strongly believe that the complete chloroplast genome sequence of M. balbisiana is a significant addition to an increasing body of genetic resources that will aid in the improvement of banana cultivars. Moreover, cpSSRs identified in
this study could be used complementarily with nuclear SSRs to deduce the parental origin of heterogenomic cultivars of edible banana and plantain derived from the wild M. balbisiana. The incongruence between the phylogenetic inferences made in the current and previous studies should incite the future investigations to evaluate the population genetic structure of wild banana species in their multiple centers of diversity. It remains to be seen whether the magnitude of IRa-SSC expansion is conserved across the Musaceae family. Additional plastome sequencing of Musa and Ensete taxa could provide novel clues to understand the dynamics and evolutionary implications of IR-SSC fluctuations in Musa cultivars. Such studies will also reveal whether the fluxes in the IRb-SSC expansion correlate with the diversification of the Musaceae lineages. Future investigations involving extensive taxon sampling should be conducted to confirm whether the members of section Musa (M. balbisiana and M. acuminata) have undergone independent IRb-SSC expansion, since such expansion is conspicuously absent in the section Callimusa (M. textilis) and in the related members of the order Zingiberales.

Acknowledgments
This work was supported by High Impact Research grant (UM.C/625/1/HIR/MOHE/SC/14), University of Malaya Research Grant (RG018-09BIO) and Postgraduate Research Grant (PS308-2010A). We would like to thank the anonymous reviewers for their useful suggestions and comments.

References


SHETTY ET AL.: MUSA BALBISIANA CHLORoplast GENOME 13 of 14


