Chloroplast Genome of Rambutan and Comparative Analyses in *Sapindaceae*

Fei Dong
College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, Fujian, China
Zhicong Lin
College of Agriculture, Fujian Agriculture and Forestry University, Fuzhou, 350002, Fujian, China
Jing Lin
Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Key Laboratory of Genetics, Fujian Agriculture and Forestry University, Fuzhou, 350002
Ray Ming
Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
Wenping Zhang (wenpingzhang@fafu.edu.cn)
Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Key Laboratory of Genetics, Fujian Agriculture and Forestry University, Fuzhou, 350002

Research Article

Keywords: Rambutan, Nephelium lappaceum, Chloroplast genome, Sapindaceae, RNA editing, Phylogeny

Posted Date: December 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-128918/v1

Abstract

Background: Rambutan (*Nephelium lappaceum* L.) is an important fruit tree belongs to the family *Sapindaceae* and widely cultivated in Southeast Asia. The chloroplast of plants, as a photosynthetic organelle plays an important role in the photosynthesis and secondary metabolic activities. The chloroplast genome sequencing has become an integral part in understanding the genomic machinery and the phylogenetic histories of rambutan organelles.

Results: We sequenced its chloroplast genome and assembled 161,321 bp circular DNA. It is characterized by a typical quadripartite structure composed of a large (86,068 bp) and small (18,153 bp) single-copy region interspersed by two identical inverted repeats (IRs) (28,550 bp). We identified 132 genes including 78 protein-coding, 29 tRNA and 4 rRNA genes, with 21 genes duplicated in the IRs. Sixty-three simple sequence repeats (SSRs) and 98 repetitive sequences were detected. Twenty-nine codons showed biased usage and 49 potential RNA editing sites were predicted across 18 protein-coding genes in the rambutan chloroplast genome. In addition, coding gene sequence divergence analysis of *N. lappaceum* suggested that *ccsA, clpP, rpoA, rps12, psbJ* and *rps19* were under positive selection, which might reflect specific adaptations of *N. lappaceum* to its particular living environment. Comparative chloroplast genome analyses from five species in *Sapindaceae* revealed that a higher similarity was conserved in the IR regions than in the LSC and SSC regions. The phylogenetic analysis showed that *N. lappaceum* chloroplast genome has the closest relationship with that of *Pometia tomentosa.*

Conclusions: The understanding of the chloroplast genomics of rambutan and comparative analysis of *Sapindaceae* species would provide insight into future research on the breeding of rambutan and *Sapindaceae* evolutionary studies.

Background

Rambutan (*Nephelium lappaceum* L.) is an important tropical fruit in the family *Sapindaceae* and originated in Indonesia and Malay Peninsula [1]. It is widely cultivated in Southeast Asia and the coastal areas of South China. Malaysians refer to it as “rambutan”, because of the fruit surface covered with thick and elongated spines. The fruits of rambutan are popular in the general population due to its rich nutrients, delicate and characteristic flavor and delicious taste. Rambutan peel extract is rich in phenolic content and exhibited antibacterial activity against many pathogenic bacteria, suggesting its antioxidant and/or antimicrobial properties[2]. Rambutan has the potential to be used as natural antioxidants and anti-aging agent in pharmaceutical and food industries to replace synthetic ones[3, 4].

The *Sapindaceae* family contains over 150 genera and 2000 species with several economically important crops widely distributed in tropical and subtropical regions[5]. However, genomic research on *Sapindaceae* family, especially in the *N. lappaceum* has been relatively scarce. This lack of genetic information making it difficult to meet the need for improving the quality and agronomic characteristics of rambutan through breeding and gene editing.

Chloroplast (cp) are photosynthetic organelles that provide energy to green plants, it plays an important role in the photosynthesis and secondary metabolic activities[6, 7]. The chloroplast genomes are maternally inherited in most plants and are highly conserved with its composition and sequence. The typical chloroplast genomes of angiosperms are circular DNA molecule which has a characteristic quadripartite structure with a large single-copy (LSC) region, a small single-copy (SSC) region, and two inverse repeats (IRs) regions[6]. The length of the genome is between 120 and 170 kb and usually encode 110 to 130 genes, and about 40 genes are specialized participating in photosynthesis, transcription and translation[8, 9]. The first chloroplast genome from tobacco (*Nicotiana tabacum*) was sequenced in 1986[10], With the rapid development of next-generation sequencing technologies, the cost of whole genome sequencing is dropping rapidly[11]. Complete chloroplast genome sequences now could be easily acquired with relatively low cost. It has been an explosion in the number of available chloroplast genome sequences. Over 4,300 complete chloroplast genome sequences have been submitted in the National Center for Biotechnology Information (NCBI) organelle genome database.

Within the *Sapindaceae* family, the complete chloroplast genomes of eight plant species have been sequenced and were available from the NCBI database. Nevertheless, no chloroplast genome in the genus *Nephelium* Linn has been reported. In this study, we report the first complete chloroplast genome of *N. lappaceum*, exploring its general features, SSRs and long repeats,

codon usage and analysis of IR contraction and expansion. In addition, nine chloroplast genome sequences were used for analysis of molecular evolution in the *Sapindaceae* family. We constructed a phylogenetic tree to understand the phylogenetic relationship of *Sapindaceae* plants. The chloroplast genome sequence and the comprehensive chloroplast genomic analysis of *N. lappaceum* would provide a theoretical basis for molecular identification and further understanding of the evolutionary history of *Sapindaceae* family.

Results

Chloroplast Genome Features of N. lappaceum

The structure of *N. lappaceum* chloroplast genome was analogous to most chloroplast genomes of plants with a typical quadripartite structure. We assemble a closed circular chloroplast genome with 161,321 bp in *N. lappaceum*. The chloroplast genome contains a pair of inverted repeat regions (IRs) of 28,550 bp, a large single-copy region (LSC) of 86,068 bp and a small single-copy region (SSC) of 18,153 bp (Fig. 1). The size of *N. lappaceum* chloroplast genome was slightly larger than that in *S. mukorossi* (160,481 bp), *P. tomentosa* (160,818 bp), *D. Longan* (160,833 bp) and shorter than that in *L. chinensis* chloroplast genome (162,524 bp) of *Sapindoideae* (Table 1). The number of chloroplast genes in *N. lappaceum* was 132, the same with those in *D. Longan* and *L. chinensis* (Table 1). In addition, there was no significant difference in GC content among the five analytical genomes in *Sapindoideae*.

Table 1
Comparison of the general features of the five *Sapindoideae* chloroplast genomes.

Genome feature	*Dimocarpus longan*	*Litchi chinensis*	*Pometia tomentosa*	*Sapindus mukorossi*	*Nephelium lappaceum*
GenBank	MG214255	KY635881	MN106254	KM454982	MT936934
Size (bp)	160833	162524	160818	160481	161321
LSC (bp)	85707	85750	85666	85650	86068
SSC (bp)	18270	16568	18360	18873	18153
IR (bp)	28428	30103	28396	27979	28550
Total genes	132	132	133	135	132
Protein genes	87	87	88	88	87
tRNA genes	37	37	37	39	37
rRNA genes	8	8	8	8	8
GC (%)	37.79%	37.80%	37.87%	37.66%	37.77%

The overall nucleotide composition of rambutan is: 30.79% A, 31.44% T, 19.27% C, and 18.50% G, with a total GC content of 37.77%. In total, 132 genes were annotated on this chloroplast genome, including 78 protein-coding genes, 29 transfer RNA genes (tRNA) and 4 ribosomal RNA genes (rRNA). Among them, a total of 21 genes were found duplicated in the IR regions, including nine protein-coding genes (*rps3, rps7, rps12, rps19, rpl2, rpl22, rpl23, ndhB* and *ycf2*), eight tRNA genes (*trnA-UGC, trnI-CAU, trnI-GAU, trnL-CAA, trnM-CAU, trnN-GUU, trnR-ACG* and *trnV-GAC*) and four rRNA genes (*rrn4.5 s, rrn5s, rrn16s* and *rrn23s*) (Additional file 1: Table S2). The genes structure analysis showed that 21 genes contains introns, and 19 of them (11 protein-coding genes and 8 tRNA genes) have one intron, while two genes (*ycf3* and *clpP*) have two introns (Additional file 1: Table S3).

Characterization of SSRs and repeat sequences

A total of 63 SSRs were detected from rambutan chloroplast genome, of which 45 were mononucleotide, 3 dinucleotide, 8 trinucleotide, 5 tetranucleotide and two pentanucleotide (Additional file 1: Table S4). Moreover, we compared the distribution pattern and number of SSRs with eight other chloroplast genomes in *Sapindaceae* family (Additional file 1: Table S5). The

number of mononucleotide repeats is more than the sum of other types (Fig. 2A), and the number and types of chloroplast SSRs vary in different species. *S. mukorossi* (91 SSRs) possess the highest number of SSRs while *E. cavaleriei* (62 SSRs) possesses the lowest. Furthermore, the chloroplast genome of *D. longan*, *L. chinensis*, *P. tomentosa*, *D. viscosa*, *K. paniculate* and *X. sorbifolium* contained 79, 75, 74, 77, 87 and 83 SSRs, respectively (Fig. 2B). In this study, a total of 98 larger repeats (> 10 bp) were identified in *N. lappaceum* chloroplast genome composed of 42 forward, 11 reverse, 41 palindromic and 4 complement repeats (Additional file 1: Table S6) using REPuter[12]. Among them, the largest repeat was a palindromic repeat with a size of 48 bp.

Codon usage analysis and RNA editing sites prediction

We used 53 protein coding sequences from rambutan chloroplast genome for calculate codon usage frequency and relative synonymous codon usage (RSCU) frequency (Additional file 1: Table S7). All protein coding sequences contain 21,434 codons. In detail, leucine and cysteine are the highest and lowest number of amino acids, they have 2,232 codons (approximately 10.41% of the total) and 236 codons (approximately 1.10% of the total), respectively. While Met (ATG) and Trp (TGG) are encoded by only one codon showed no biased usage (RSCU = 1). 30 codons with RSCU values more than 1, indicating they showed biased usage (Fig. 3). Among them, excluding the leucine (UUG) codon was G-ending, the remaining 29 biased usage codons of *N. lappaceum* were all A/T-ending in the third codon. In addition, there were 49 potential RNA editing sites were found across 18 protein-coding genes in *N. lappaceum* chloroplast genome and the *ndhB* gene contained the most RNA editing sites (9) (Additional file 1: Table S8). We also observed that RNA editing sites were all C to U conversion, and took place at the first (30.6%) or second (69.4%) positions of the codons, indicating that editing in the third codon position disappeared quicker than that in the second or first codon position. Furthermore, serine codons were more frequently edited than codons of other amino acids and the conversion from serine to leucine occurred most frequently.

Comparative genomes analysis

The comparative analysis based on mVISTA was performed between the chloroplast genomes of rambutan with other four *Sapindoideae* species with the annotated *D. longan* chloroplast genome as a reference. The five *Sapindoideae* subfamily chloroplast genomes length between the confines of 160,481 to 162,524 bp. The chloroplast genome of *L. chinensis* has the largest size, whereas *S. mukorossi* has the smallest size. Interestingly, the SSC region (16,568 bp) of *L. chinensis* is the shortest, whereas the SSC region (18,873 bp) of *S. mukorossi* chloroplast genome is the longest (Fig. 4). The IR (A/B) regions exhibited less divergence than the SSC and LSC regions. In addition, the coding regions were more highly conserved than the non-coding regions. Among the five chloroplast genomes, four rRNA genes (*rrn16S, rrn23S, rrn5S, rrn4.5S*) were the most conserved, while 7 genes (*matK, rpoC2, psbB, rpoA, ndhF, ndhD* and *ycf1*) showed the most diversity in the coding regions. The highly divergent regions were found in the intergenic spacers and introns, including *trnH-GUG-psbA, trnR-UCU-atpA, petN-psbM, psbZ-trnG-GCC, ndhC-trnV-UAC, psbE-petL, rpl16-rps3* and *rpl32-trnL-UAG*.

Expansion and contraction of IR regions

We compared the IR regions and the junction sites of the LSC and SSC regions of nine *Sapindaceae* family chloroplast genomes (including *N. lappaceum*) (Fig. 5). The IR regions vary in different chloroplast genomes, ranging from 26,923 bp in *E. cavaleriei* to 30,103 bp in *L. chinensis*. In our study, the *ycf1* gene was located at the SSC/IRA junction in all of the nine chloroplast genomes and the fragment located at the IRa region ranged from 962 bp to 3,183 bp. Moreover, most junctions between LSC and IRa in this study was located downstream of the *trnH-GUG*, except the *S. mukorossi*. In addition, the LSC/IRb junction of three species *D. viscosa*, *E. cavaleriei* and *K. paniculate* was located within the coding region of *rpl22* and created a location of 110, 40 or 63 bp at the LSC/IRb border. The remaining chloroplast genomes share a similar pattern, the LSC/IRb junction was located in intergenic regions of *rpl16* and *rps3*, and the IRb/SSC junction between IRb and SSC region (JSB) of five species (*S. mukorossi, X. sorbifoliun, D. viscosa*, *E. cavaleriei* and *K. paniculate*) was located between the gene of ycf1 and *ndhF*. However, other four chloroplast genomes only have *ndhF* located or near the JSB.

Synonymous (Ks) and non-synonymous (Ka) substitution rate analysis

To explore molecular evolution of orthologous genes shared by nine *Sapindaceae* species, particularly genes undergoing purifying or positive selection, we calculated the Ka/Ks ratio of 622 orthologous pairs with 78 protein coding genes (Additional

file 1: Table S9). Overall, the average Ka/Ks ratio of the nine chloroplast genomes was 0.20. Total 612 orthologous pairs had a Ka/Ks ratio less than 1 in the nine comparison groups, out of which 546 orthologs had a Ka/Ks ratio less than 0.5 (Fig. 6), suggesting that most genes are undergoing strong purifying selection pressures. Moreover, 66 orthologs of 31 genes with a Ka/Ks ratio between 0.5 and 1, 10 orthologous pairs of 6 genes (*ccsA, rpoA, rps12, psbJ, clpPc* and *rps19*) with a Ka/Ks ratio greater than 1 were detected in this study, suggesting that these genes might have experienced positive selection in the procedure of evolution. Among them, the Ka/Ks ratio of the *ycf1* gene was greater than 0.5 in eight comparison groups, the *rpoA* and *ycf2* gene with Ka/Ks ratio greater than 0.5 was also observed in the comparison of seven and six groups, respectively. Besides, *clpP, matK* and *rps15* gene with Ka/Ks ratio > 0.5 were founded in four out of the eight comparison groups.

Phylogenetic analysis

We performed multiple sequence alignments using the whole chloroplast genome sequences of nine *Sapindaceae* species and two *Anacardiaceae* species as outgroups (Fig. 7). All nodes in the ML trees have 100% bootstrap support values, and these 11 chloroplast genome sequences are clustered into three groups. In detail, the five species (*D. longan, L. chinensis, P. tomentosa, N. lappaceum* and *S. mukorossi*) from *Sapindoideae* clustered into one group, four species (*K. paniculata, D. viscosa, E. cavaleriei* and *X. sorbifolium*) from *Dodonaeoide*ae are in one group, and the two species (*A. occidentale* and *M. indica*) in *Anacardiaceae* are cluster into one group. In the *Sapindoideae* group, the *N. lappaceum* chloroplast genome sequence showed the closest relationship with *P. tomentosa*, followed by *D. longan* and *L. chinesis*, as far as *S. mukorossi*. The three groups of this phylogenetic tree of the 11 chloroplast genome sequences are consistent with traditional taxonomy, suggesting that the chloroplast genome could effectively resolve the phylogenetic positions and relationships of species.

Disscussion

We assembled *N. lappaceum* complete chloroplast genome sequence and deposited it to GenBank under accession number: MT936934, *N. lappaceum* chloroplast genome was consistent with the characteristics of most angiosperm species in structure and gene content. Although there are some differences in the sizes of the overall genome, LSC, SSC and IR regions, the numbers of genes and GC content are similar among the five *Sapindoideae* chloroplast genomes, which to some extent reflects the high conservation of angiosperm chloroplast genomes[6]. Intron plays an important role in RNA stability, regulation of gene expression and alternative splicing which have been reported in many other species[13, 14]. There were two genes (*ycf3* and *clpP*) included two introns in the *N. lappaceum* chloroplast genome. It has been reported that *ycf3* gene is essential for the accumulation of the photosystem I (PSI) complex and acts a chaperone that interacts with the PSI subunits at a post-translational level[15, 16]. Besides, *clpP* gene functions as the proteolytic subunit of the ATP-dependent Clp protease in plant chloroplasts and is essential for the development and/or function of plastids with active gene expression in previous studies[17, 18]. Thus, study of *ycf3* and *clpP* gene will contribute to further investigation of chloroplast in *N. lappaceum*.

Simple sequence repeats (SSRs), also known as microsatellites, are tandem repeats (1 ~ 6 bp units repeated multiple times) distributed across the entire genome which have been widely applied as molecular markers for determining genetic variations across species and evolutionary studies because of its unique uniparental in inheritance[19–21]. we identified 63 SSRs in *N. lappaceum* chloroplast genome and most of SSRs were distributed in IGS regions. Mononucleotide SSRs were identified most frequently (68% on average) among the nine analyzed chloroplast genomes of *Sapindaceae* species, and vast majority of mononucleotide repeats consist of short polyA or polyT repeats sharing a similar pattern in most angiosperm chloroplast genomes[22, 23]. Moreover, repetitive sequences are helpful in phylogenetic study and play a vital role in genome rearrangement[24]. These results can provide chloroplast molecular markers that can be used to quickly identify species, confirm hybrid progeny when breeding.

Codon usage biases are found in all eukaryotic and prokaryotic genomes and have been proposed to regulate different aspects of translation process[25]. High RSCU values of the codons are probably attributed to amino acid functions or peptide structures that avoid transcriptional errors in chloroplast genomes[26, 27]. There are 30 codons showed biased usage and most of them were A/T-ending in the third codon. This phenomenon also exists in previous studies[22]. Codon usage bias of chloroplast genome reflects a selective pressure to increase translation efficiency[28], and research on codon preferences can help us to better understand gene expression and molecular evolution mechanisms of *N. lappaceum.*

We observed *ndhB* gene contained the most RNA editing sites within the 49 potential RNA editing sites, and 16 editing sites were U_A type, indicating there was a U_A bias for the distribution of RNA editing sites that was in accordance with previous reports in other species[23, 29]. RNA editing is a post-transcriptional regulation pattern involved in the insertion, deletion, or modification of nucleotides that widely exists in land plants[30]. The first chloroplast RNA editing event of land plant was discovered in the mRNA transcript of *rpl2* gene in maize chloroplast genome in 1991[31]. The most frequent editing events in plants are C-to-U changes, however, U-to-C editing has also been observed[32, 33]. Additionally, RNA editing usually occurs in the first or second base of codons, resulting in the conversion of hydrophilic amino acid to hydrophobic[34].

Comparative analysis of chloroplast genomes is an essential step in genomics which can provide insight into complex evolutionary relationships. The mVISTA analysis showed that five *Sapindoideae* chloroplast genomes were conserved, with a high degree of synteny and gene order conservation, and the coding region was more conserved than the non-coding region, which is consistent with reports on other angiosperms[35], suggesting an evolutionary conservation of these genomes at the genome-scale level. In addition, *ycf1* gene showed the greatest degree of differentiation. Previous study reported that *ycf1* is helpful to provide phylogenetic information at the species level and more variable than *matK* in *Orchidaceae*[36]. Furthermore, *ycf1* performed better to identify DNA barcodes of high resolution at species level than any of the *matK*, *rbcL* and *trnH-psbA*[35]. These variable genic regions found in our study can be regarded as molecular markers for DNA barcoding and phylogenetic studies in *Sapindaceae*.

Although most land plants have relatively conserved cp genomes, the end of the inverted repeats (IRa and IRb) regions differs among various plant species. The expansion or contraction of the IR regions represent important evolutionary events often results in size variation of different chloroplast genomes and is helpful to studying the chloroplast genome evolution history[37, 38]. In this study, our results suggested that the boundary of IR/LSC and IR/SSC might be conversed among chloroplast genomes of closely related family species but some differences also occurs between relatively distantly related family species, such as gene overlap length, duplicate of the *ycf1* and *rps3* genes, even the distance of *trnH-GUG* from the border near the LSC/IRB junctions, indicating that the expansion and contraction of the IR region led to length and structure changes of chloroplast genomes.

The ratio between nonsynonymous (Ka) and synonymous (Ks) nucleotide substitution has been widely used as important markers in genome or gene evolution studies[8]. Ka/Ks = 1 signifies neutral evolution, Ka/Ks > 1 indicates that the gene is affected by positive selection, whereas Ka/Ks < 1 indicates that the gene is affected by purifying selection[39]. Additionally, a Ka/Ks ratio of 0.5 was considered as a useful cut-off value to identify genes under positive selection in previous studies[40]. In our study, the *ccsA, rpoA, rps12, psbJ, clpP* and *rps19* gene with Ka/Ks > 1. It is noteworthy that the *ycf1* gene also exhibited high Ka/Ks ratios with Ka/Ks > 0.5 in eight comparison groups. This result is in keeping with the previous observations that the *ycf1* gene was more variable than the *matK* and *rbcL* genes in most plant, and could be using as an effective biological tool for plant phylogeny study[35]. The positive selection of genes in *N. lappaceum* possibly provided help for adaptations to its particular living environment.

Numerous studies have shown that chloroplast genome sequences have been successfully used in taxonomic and phylogenetic studies[41], and contribute to describe the evolutionary relationships between species[42]. In this study, the topology of the trees consists of two main branches: *Dodonaeoideae* and evolutionary younger *Sapindoideae.* And generic relationships of the two subfamilies are basically congruent with the taxonomy of these families. The availability of the completed *N. lappaceum* chloroplast genome provided us with sequence information that can be used to confirm the phylogenetic position of *N. lappaceum* and understand the phylogenetic relationships among *Sapindaceae*. However, we had used only a small number of species in *Sapindaceae*, further research on other chloroplast genome as well as nuclear genome sequences of *Sapindaceae* should be sequenced to provide more sufficient evidence to accurately illustrate the evolution of family *Sapindaceae.*

Conclusions

We assembled the first complete chloroplast genome of rambutan using Illumina sequencing technology and compared its structure with other *Sapindaceae* species. The chloroplast genome of *N. lappaceum* exhibits similar quadripartite structure, gene

order, G + C content when compared with other *Sapindaceae* chloroplast genomes. A total of 63 SSRs and 98 repeat sequences were identified in *N. lappaceum* chloroplast genome. The research on codon usage of *N. lappaceum* shows that some amino acids have obvious codon usage bias and the codon preferences may help us understand the evolution mechanisms of *N. lappaceum*. With PREP prediction, we detected 49 RNA editing loci in 18 protein-coding genes in *N. lappaceum.* Moreover, the expansion and contraction of the IR regions, leading to the variations in nine *Sapindaceae* chloroplast genome size. There are 6 genes (*ccsA, rpoA, rps12, psbJ, clpP* and *rps19*) were detected with a Ka/Ks ratio > 1, suggesting that these genes experienced positive selection in the evolution. Additionally, phylogenetic analysis using 9 complete chloroplast genome sequences in *Sapindaceae* strongly supports the close relationship of *N. lappaceum* and *P. tomentosa* among sequenced chloroplast genomes in *Sapindaceae*.

Methods

Plant material, DNA extraction, and sequencing

Young, healthy leaves of the major cultivar of rambutan, Baoyan7, were collected from Baoting (N18°23', E109°21') in Hainan Province, China. The leaves were frozen in liquid nitrogen, and maintained at − 80 °C. The total genomic DNA was extracted by 2X cetyltrimethylammonium bromide (CTAB) method [43]. And a library with insert sizes of 300–500 bp was constructed and then sequenced on an Illumina HiSeq2500 platform (Illumina, San Diego, CA, USA) double terminal sequencing method (150 pair-ends).

Chloroplast genome assembly and annotation

First, FastQC software was performed to evaluate the quality of Illumina paired-end raw reads[44], and low-quality reads were filtered. The remaining clean reads were used for assembly via NOVOPlasty[45] using *Dimocarpus longan* chloroplast genome(GenBank: MG214255)[46] as the reference genome to generate the first version of rambutan genome. Next, all clean reads were mapped onto the first version genome and the mapped reads were assembled using SPAdes3.14.1[47] and assembled contigs were corrected using the pair-end short reads from HiSeq2500 by Pilon version 1.23 (https://github.com/broadinstitute/pilon)[48] to generate the second version of rambutan chloroplast genome. These two versions were compared and mutually corrected to get the final complete rambutan chloroplast genome.

The chloroplast genome was annotated by online program GeSeq (https://chlorobox.mpimp-golm.mpg.de/geseq.html)[49] and CPGAVAS2[50]. Genome features like start/stop codons and intron/exon borders were manually corrected through the comparison of other reported *Sapindaceae* family chloroplast genomes. In addition, tRNA genes were identified by tRNAscan-SE 2.0 (http://lowelab.ucsc.edu/tRNAscan-SE/)[51]. A circular map of the revised annotated rambutan chloroplast genome was illustrated by using Organellar Genome DRAW (OGDRAW) (https://chlorobox.mpimpgolm.mpg.de/OGDraw.html)[52].

Chloroplast genome analysis

The simple sequence repeats (SSR) in nine chloroplast genome sequences of *Sapindaceae* (including *N. lappaceum*) (Additional file 1: Table S1) were identified by online tool MISA (https://webblast.ipk-gatersleben.de/misa/)[53] and the threshold settings were as follows: ten was applied to mononucleotide repeats, five to dinucleotide repeats and four to trinucleotide repeats, three for tetra-, penta-, and hexanucleotide repeats[54]. Repetitive sequences including forward, reverse, palindrome, and complement sequences were analyzed by REPuter program[12], and the parameter was set with a minimum length of repeat region set to 10 bp and the minimum sequence identity set to 90%.

The expansion and contraction of the inverted repeat (IR) regions at junction sites from eight *Sapindaceae* family chloroplast genome sequences, including *Dimocarpus longan* (MG214255), *Litchi chinensis* (KY635881), *Pometia tomentosa* (MN106254), *Sapindus mukorossi* (KM454982), *Dodonaea viscosa*(MF155892), *Eurycorymbus cavaleriei* (MG813997), *Koelreuteria paniculate* (KY859413) and *Xanthoceras sorbifolium* (KY779850), were examined and plotted using IRscope online program (https://irscope.shinyapps.io/irapp/)[55]. Codon usage of the *N. lappaceum* chloroplast genome was analyzed via GALAXY platform (https://galaxy.pasteur.fr) [56] with CodonW online tool. The length of protein-coding gene less than 300 nucleotide

and the repetitive genes sequences were removed to reduce deviation of the results[57]. Finally, 53 CDS in *N. lappaceum* were selected for further codon usage analysis. Besides, putative RNA editing sites were predicted using PREP-Cp web server (http://prep.unl.edu/cgi-bin/cp-input.pl)[58] with a cutoff value of 0.8.

Genome Comparison

We downloaded four whole chloroplast genome sequences of *Sapindoideae* subfamily from the National Center for Biotechnology Information (NCBI) Organelle Genome and Nucleotide Resources database, including *D. longan*[46], *L. chinensis*, *P. tomentosa*[59] and *S. mukorossi*[60]. The mVISTA online program (Shuffle-LAGAN mode)[61, 62] was used to compare chloroplast genome sequence of rambutan with the other species from *Sapindoideae* subfamily, in which the annotation of *D. longan* as the reference.

Positive selection analysis of protein sequence

We analyzed synonymous (Ks) and non-synonymous (Ka) substitution rates to investigate the molecular evolutionary process of *Sapindaceae* family, The protein-coding genes of *N. lappaceum* were separately compared with eight closely related species in *Sapindaceae* family: *D. longan*, *L. chinensis*, *P. tomentosa*, *S. mukorossi, D. viscosa*[63], *E. cavaleriei*[64], *K. paniculate*[65] and *X. sorbifolium*[66] using ParaAT 2.0[67], then the Ka/Ks value was calculated by KaKs_calculator 2.0[68] with NG method[69].

Phylogenetic analysis

In order to deeply detect the evolutionary relationship of *Sapindaceae* family, we aligned 9 complete chloroplast genomes (including *N. lappaceum*) with MAFFT version 7[70]. The best fitting nucleotide substitution model (TVM + I + G) was chosen by jModelTest v2.1.7[71]. Phylogenetic analysis was then inferred by ML (maximum-likelihood) method based on the TVM + I + G substitution model in PAUP* 4.0[72] with 1000 bootstrap replicates. *Anacardium occidentale* (KY635877) and *Mangifera indica* (KY635882)[73] in *Anacardiaceae* family were set as the outgroup.

Abbreviations

Cp
Chloroplast; SSRs:Simple sequence repeats; IRs:Inverted repeats; LSC:Large single-copy; SSC:Small single-copy; ML:Maximum-likelihood

Declarations

Acknowledgements

This work was supported by startup fund from Fujian Agriculture and Forestry University.

Author contributions

F.D. performed most of the data analysis, and wrote the manuscript. W.Z. collected experiment materials and data.
Z. L. and J. L. helped in genome assembly strategy design. R.M., W. Z. and F. D. designed the project and revised the manuscript.

Availability of Data and Materials

The chloroplast genome assembly and annotation were deposited in GenBank under the accession number of MT936934. All other data and material generated in this manuscript are available from the corresponding author upon reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Funding

Not applicable.

References

1. Lim TK: **Edible Medicinal and Non Medicinal Plants** 2015: Springer; 2012.
2. Palanisamy UD, Cheng HM, Masilamani T, Subramaniam T, Ling LT, Radhakrishnan AK: **Rind of the rambutan, Nephelium lappaceum, a potential source of natural antioxidants**. *Food Chem* 2008, **109**(1):54–63.
3. Zhuang Y, Ma Q, Guo Y, Sun L: **Protective effects of rambutan (Nephelium lappaceum) peel phenolics on H2O2-induced oxidative damages in HepG2 cells and d-galactose-induced aging mice**. *Food Chemical Toxicology* 2017, **108**(Pt B):554–562.
4. c NNMPab, B TTL, A JVC, A KR: **Evaluation of antimicrobial activity of rambutan (Nephelium lappaceum L.) peel extracts**. *Int J Food Microbiol* 2020, **321**.
5. Harrington MG, Edwards KJ, Johnson SA, Chase MW, Gadek PA: **Phylogenetic Inference in Sapindaceae sensu lato Using Plastid matK and rbcL DNA Sequences**. *Syst Bot* 2005, **30**(2):366–382.
6. Wicke S, Schneeweiss GM, Depamphilis CW, Kai FM, Quandt D: **The evolution of the plastid chromosome in land plants: gene content, gene order, gene function**. *Plant MolBiol* 2011, **76**(3–5):273–297.
7. Bobik K, Burch-Smith TM: **Chloroplast signaling within, between and beyond cells**. *Front Plant Sci* 2015, **6**:781.
8. Wolfe KH, Li W, Sharp PM: **Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs**. *Proc Natl Acad Sci U S A* 1987, **84**(24):9054–9058.
9. Palmer JD: **Comparative Organization of Chloroplast Genomes**. *Annu Rev Genet* 1985, **19**(1):325–354.
10. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Sugiura M: **The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression**. *Plant Mol Biol Rep* 1986, **5**(9):2043–2049.
11. Li C, Lin F, An D, Wang W, Huang R: **Genome Sequencing and Assembly by Long Reads in Plants**. *Genes* 2017, **9**(1):6.
12. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R: **REPuter: the manifold applications of repeat analysis on a genomic scale**. *Nucleic Acids Res* 2001, **29**(22):4633–4642.
13. Nguyen Dinh S, Sai TZT, Nawaz G, Lee K, Kang H: **Abiotic stresses affect differently the intron splicing and expression of chloroplast genes in coffee plants (Coffea arabica) and rice (Oryza sativa)**. *J Plant Physiol* 2016, **201**:85–94.
14. Mirzaei S, Mansouri M, Mohammadi-Nejad G, Sablok G: **Comparative assessment of chloroplast transcriptional responses highlights conserved and unique patterns across Triticeae members under salt stress**. *Photosynth Res* 2017, **136**(3):357–369.
15. Naver H, Boudreau E, Rochaix JD: **Functional studies of YCF3: Its role in assembly of photosystem I and Interactions with some of its subunits**. *Plant Cell* 2002, **13**(12):2731–2745.
16. Boudreau E, Takahashi Y, Lemieux C, Turmel M, Rochaix JD: **The chloroplast ycf3 and ycf4 open reading frames of Chlamydomonas reinhardtii are required for the accumulation of the photosystem I complex**. *Embo J* 1997, **16**(20):6095–6104.
17. Clarke AK, Schelin J, Porankiewicz J: **Inactivation of the clpP1 gene for the proteolytic subunit of the ATP-dependent Clp protease in the cyanobacterium Synechococcus limits growth and light acclimation**. *Plant MolBiol* 1998, **37**(5):791–801.
18. Bruce CA, Cunningham KA, Stern DB: **The plastid clpP gene may not be essential for plant cell viability**. *Plant Cell Physiol* 2003(1):93–95.

19. Varshney RK, Sigmund R, Borner A, Korzun V, Stein N, Sorrells ME, Langridge P, Graner A: **Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice**. *Plant Sci* 2005, **168**(1):195–202.
20. Yang A, Zhang J, Tian H, Yao X: **Characterization of 39 novel EST-SSR markers for Liriodendron tulipifera and cross-species amplification in L. chinense (Magnoliaceae)**. *Am J Bot* 2012, **99**(11):e460-464.
21. Li B, Lin F, Huang P, Guo W, Zheng Y: **Development of nuclear SSR and chloroplast genome markers in diverse Liriodendron chinense germplasm based on low-coverage whole genome sequencing**. *Biol Res* 2020, **53**(1):21.
22. Gao B, Yuan L, Tang T, Hou J, Pan K, Wei N: **The complete chloroplast genome sequence of Alpinia oxyphylla Miq. and comparison analysis within the Zingiberaceae family**. *PLoS One* 2019, **14**(6).
23. Yan C, Du J, Gao L, Li Y, Hou X: **The complete chloroplast genome sequence of watercress (Nasturtium officinale R. Br.): Genome organization, adaptive evolution and phylogenetic relationships in Cardamineae**. *Gene* 2019, **699**:24–36.
24. Cavalier-Smith T: **Chloroplast Evolution: Secondary Symbiogenesis and Multiple Losses**. *Curr Biol* 2002, **12**(2):R62-R64.
25. Hershberg R, Petrov DA: **Selection on codon bias**. *Annu Rev Genet* 2008, **42**(1):287–299.
26. Raman G, Park S, Lee EM, Park SJ: **Evidence of mitochondrial DNA in the chloroplast genome of Convallaria keiskei and its subsequent evolution in the Asparagales**. *Sci Rep* 2019, **9**(1):5028.
27. Purabi M, Rofinayasmin BO, Katharina M, Ramakrishnan N, Jennifer AH: **Codon usage and codon pair patterns in non-grass monocot genomes**. *Ann Bot* 2017(6):1–17.
28. Morton BR, Wright SI: **Selective Constraints on Codon Usage of Nuclear Genes from Arabidopsis thaliana**. *Mol Biol Evol* 2006, **24**(1):122–129.
29. Wang W, Yu H, Wang J, Lei W, Gao J, Qiu X, Wang J: **The Complete Chloroplast Genome Sequences of the Medicinal Plant Forsythia suspensa (Oleaceae)**. *Int J Mol Sci* 2017, **18**(11):2288.
30. Smith HC, Gott JM, Hanson MR: **A guide to RNA editing**. *RNA-Publ RNA Soc* 1997, **3**(10):1105–1123.
31. Hoch B: **Editing of a chloroplast mRNA by creation of an initiation codon**. *Nature* 1991, **353**(6340):178–180.
32. Maier RM, Zeltz P, Kossel H, Bonnard G, Gualberto JM, Grienenberger JM: **RNA editing in plant mitochondria and chloroplasts**. *Plant MolBiol* 1996, **32**(1):343–365.
33. Schmitzlinneweber C, Barkan A: **RNA splicing and RNA editing in chloroplasts**. *Topics in Current Genetics* 2007, **19**:213–248.
34. Shikanai T: **RNA editing in plant organelles: machinery, physiological function and evolution**. *Cellular Molecular Life Sciences* 2006, **63**(6):698–708.
35. Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S: **ycf1, the most promising plastid DNA barcode of land plants**. *Sci Rep* 2015, **5**:8348.
36. Neubig KM, Whitten WM, Carlsward BS, Blanco MA, Endara L, Williams NH, Moore M: **Phylogenetic utility of ycf1 in orchids: a plastid gene more variable than matK**. *Plant Syst Evol* 2009, **277**(1–2):75–84.
37. Dugas DV, Hernandez D, Koenen EJM, Schwarz E, Straub S, Hughes CE, Jansen RK, Nageswara-Rao M, Staats M, Trujillo JT: **Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in clpP**. *Sci Rep* 2015, **5**:16958.
38. Yu X, Tan W, Zhang H, Gao H, Tian X: **Complete Chloroplast Genomes of Ampelopsis humulifolia and Ampelopsis japonica: Molecular Structure, Comparative Analysis, and Phylogenetic Analysis**. *Plants-Basel* 2019, **8**(10):410.
39. Yang Z, Bielawski JP: **Statistical methods for detecting molecular adaptation**. *Trends Ecol Evol* 2000, **15**(12):496–503.
40. Swanson WJ, Wong A, Wolfner MF, Aquadro CF: **Evolutionary Expressed Sequence Tag Analysis of Drosophila Female Reproductive Tracts Identifies Genes Subjected to Positive Selection**. *Genetics* 2004, **168**(3):1457–1465.
41. Gitzendanner MA, Soltis PS, Wong GK, Ruhfel BR, Soltis DE: **Plastid phylogenomic analysis of green plants: A billion years of evolutionary history**. *American Journal of Botany* 2018, **105**(3):291–301.
42. Du YP, Bi Y, Yang FP, Zhang MF, Zhang XH: **Complete chloroplast genome sequences of Lilium: Insights into evolutionary dynamics and phylogenetic analyses**. *Sci Rep* 2017, **7**(1).

43. Porebski S, Bailey LG, Baum BR: **Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components**. *Plant Mol Biol Rep* 1997, **15**(1):8–15.
44. Andrews S: **FastQC A Quality Control tool for High Throughput Sequence Data**. *Babraham Institute* 2015.
45. Nicolas D, Patrick M, Guillaume S: **NOVOPlasty: de novo assembly of organelle genomes from whole genome data**. *Nucleic Acids Res* 2017(4):4.
46. Wang K, Li L, Zhao M, Li S, Sun H, Lv Y, Wang Y: **Characterization of the complete chloroplast genome of longan (Dimocarpus longan Lour.) using illumina paired-end sequencing**. *Mitochondrial DNA Part B* 2017, **2**(2):904–906.
47. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing**. *J Comput Biol* 2012, **19**(5):455–477.
48. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman JR, Young S: **Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement**. *PLoS One* 2014, **9**(11):e112963.
49. Michael T, Pascal L, Tommaso P, Ulbricht-Jones ES, Axel F, Ralph B, Stephan G: **GeSeq – versatile and accurate annotation of organelle genomes**. *Nucleic Acids Res* 2017(W1):W1.
50. Shi L, Chen H, Jiang M, Wang L, Wu X, Huang L, Liu C: **CPGAVAS2, an integrated plastome sequence annotator and analyzer**. *Nucleic Acids Res* 2019, **47**(W1):W65-W73.
51. Chan PP, Lowe TMJMoMB: **tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences**. *Methods Mol Biol* 2019, **1962**:1–14.
52. Lohse M, Drechsel O, Bock R: **OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes**. *Curr Genet* 2007, **52**(5):267–274.
53. Beier S, Thiel T, Munch T, Scholz U, Mascher M: **MISA-web: a web server for microsatellite prediction**. *Bioinformatics* 2017, **33**(16):2583–2585.
54. Li Q, Wan JM: **[SSRHunter: development of a local searching software for SSR sites]**. *Yi Chuan* 2005, **27**(5):808–810.
55. Amiryousefi A, Hyvonen J, Poczai P: **IRscope: an online program to visualize the junction sites of chloroplast genomes**. *Bioinformatics* 2018, **34**(17):3030–3031.
56. Afgan E, Baker D, Den Beek MV, Blankenberg D, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Eberhard C: **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update**. *Nucleic Acids Res* 2016, **44**(W1):W3-W10.
57. Wright F: **The effective number of codons used in a gene**. *Gene* 1990, **87**(1):23–29.
58. Mower JP: **The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments**. *Nucleic Acids Res* 2009, **37**:253–259.
59. Wang Y, Yuan X, Zhang J: **The complete chloroplast genome sequence of Pometia tomentosa**. *Mitochondrial DNA Part B* 2019, **4**(2):3950–3951.
60. Yang B, Li M, Ma J, Fu Z, Tian J: **The complete chloroplast genome sequence of Sapindus mukorossi**. *Mitochondrial DNA Part A* 2016, **27**(3):1825–1826.
61. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak IJNAR: **VISTA: computational tools for comparative genomics**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W273-279.
62. Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S: **Glocal alignment: finding rearrangements during alignment**. *Bioinformatics* 2003, **19**:54–62.
63. Saina JK, Gichira AW, Li ZZ, Hu GW, Wang QF, Liao K: **The complete chloroplast genome sequence of Dodonaea viscosa: comparative and phylogenetic analyses**. *Genetica* 2017, **146**(1):101–113.
64. Du X, Xin G, Ren X, Liu H, Hao N, Jia G, Liu W: **The complete chloroplast genome of Eurycorymbus cavaleriei (Sapindaceae), a Tertiary relic species endemic to China**. *Conserv Genet Resour* 2018.
65. Kim SC, Baek SH, Hong KN, Lee JW: **Characterization of the complete chloroplast genome of Koelreuteria paniculata (Sapindaceae)**. *Conserv Genet Resour* 2018, **10**(4):69–72.

66. Chen SY, Zhang XZ: **Characterization of the complete chloroplast genome of Xanthoceras sorbifolium, an endangered oil tree**. *Conserv Genet Resour* 2017, **9**(4):595–598.
67. Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, Dai L: **ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments**. *Biochem Biophys Res Commun* 2012, **419**(4):779–781.
68. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J: **KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies**. *Genomics,Proteomics & Bioinformatics* 2010, **8**(1):77–80.
69. Nei M, Gojobori T: **Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions**. *Mol Biol Evol* 1986, **3**(5):418–426.
70. Katoh K, Rozewicki J, Yamada KD: **MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization**. *Brief Bioinform* 2017.
71. Darriba D, Taboada GL, Doallo R, Posada D: **jModelTest 2: more models, new heuristics and parallel computing**. *Nat Methods* 2012, **9**(8):772–772.
72. Cummings MP: **PAUP* (Phylogenetic Analysis Using Parsimony (and Other Methods))**. *Dictionary of Bioinformatics Computational Biology* 2004.
73. Azim MK, Khan IA, Zhang Y: **Characterization of mango (Mangifera indicaL.) transcriptome and chloroplast genome**. *Plant MolBiol* 2014, **85**(1–2):193–208.

Figures

Figure 1

Gene map of N. lappaceum chloroplast genome. Genes drawn outside and inside of the circle are transcribed clockwise and counterclockwise, respectively. Genes belonging to different functional groups are color coded. The darker gray in the inner circle corresponds to GC content. SSC region, LSC region, and inverted repeats (IRA and IRB) are indicated.

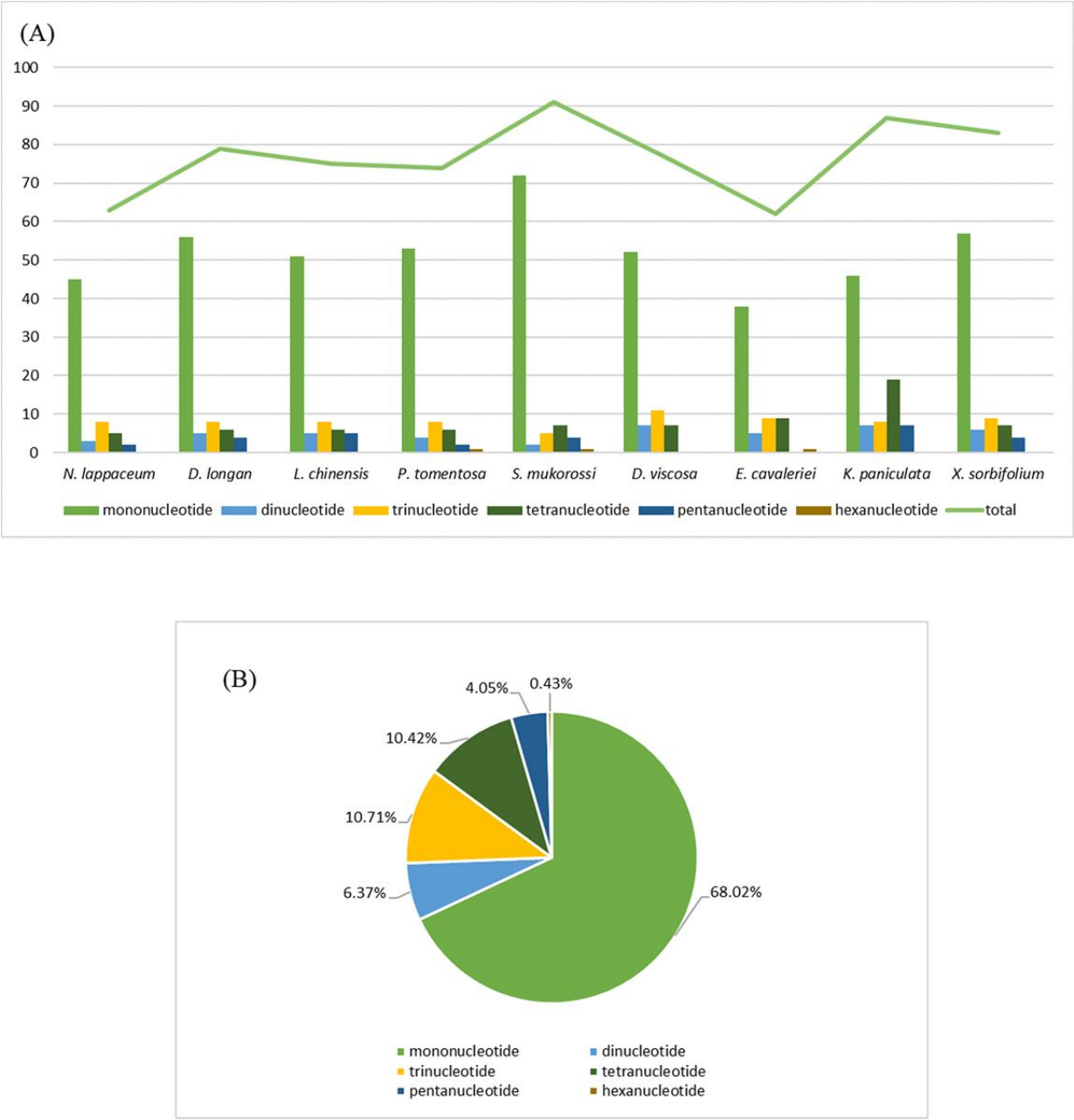

Figure 2

Analysis of simple sequence repeats (SSRs) in nine Sapindaceae (including N. lappaceum) chloroplast genomes. (A) Number of different SSRs types detected in nine Sapindaceae (including N. lappaceum) chloroplast genomes. (B) Presence of different SSRs types in all SSRs of nine Sapindaceae (including N. lappaceum) chloroplast genomes.

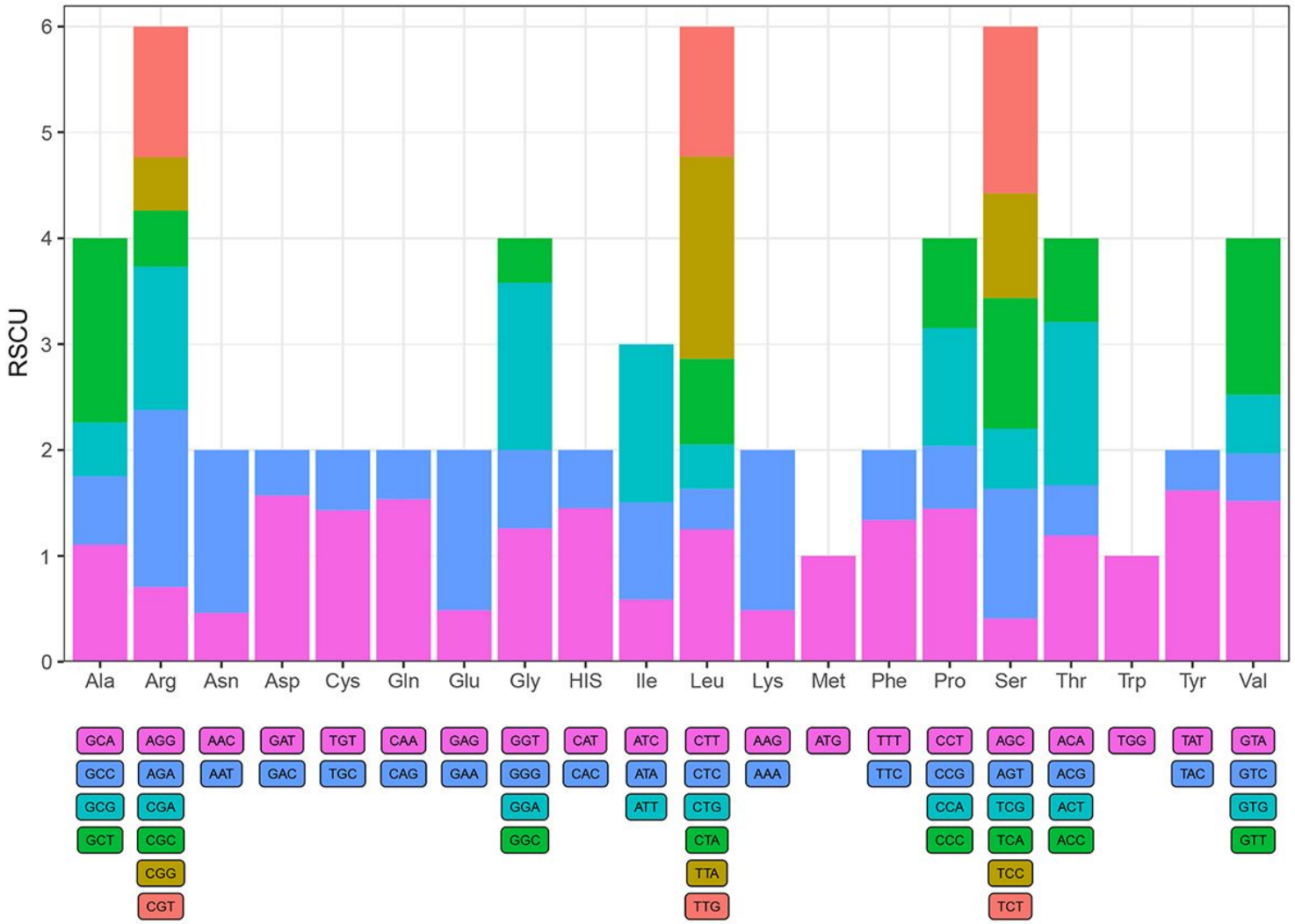

Figure 3

Codon content of 20 amino acids and stop codons in all protein-coding genes of N. lappaceum chloroplast genome. The colour of the histogram corresponds to the colour of codons.

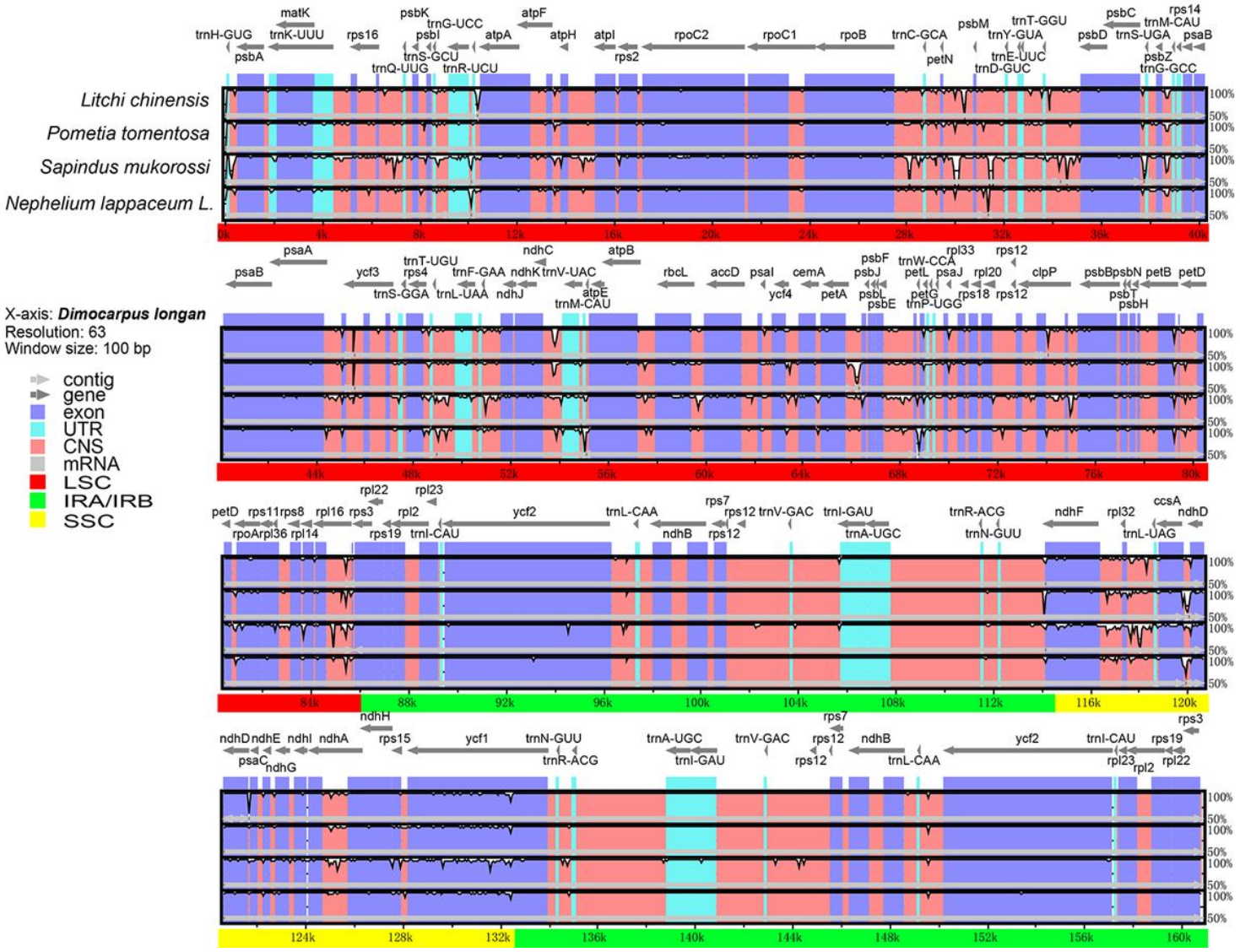

Figure 4

Comparison of five Sapindoideae subfamily chloroplast genomes (including N. lappaceum), with D. longan as a reference. Gray arrows and thick black lines above the alignment indicate the direction of the gene. Purple bars represent exons, blue bars represent untranslated regions (UTRs), pink bars represent conserved non-coding sequences (CNS), and gray bars represent mRNA. The y-axis indicates the percent identity between 50% and 100%.

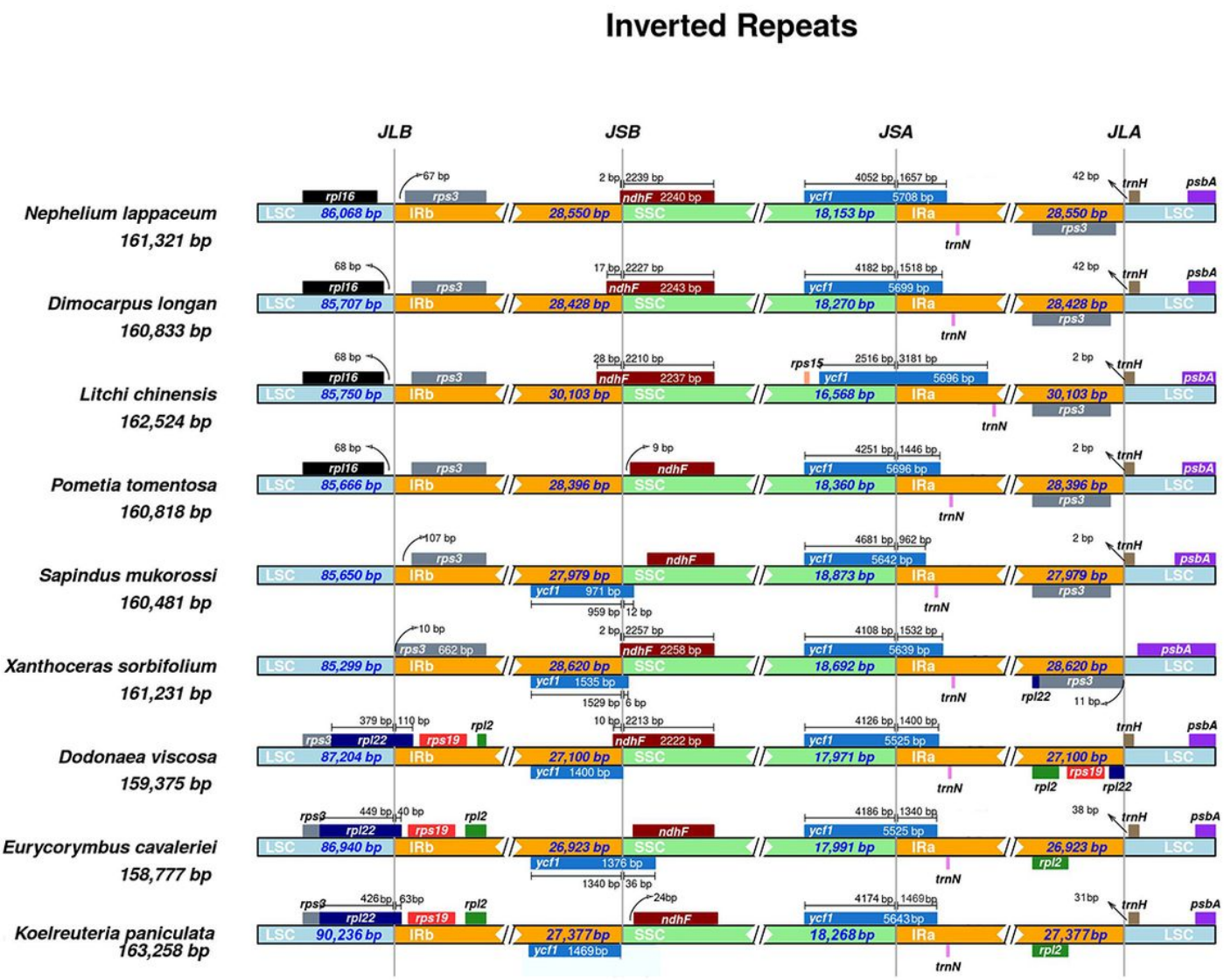

Figure 5

Comparison of the borders of the LSC, SSC and IR regions among nine Sapindaceae chloroplast genomes. For each species, genes transcribed in positive strand are depicted on the top of their corresponding track from right to left direction, while the genes on the negative strand are depicted below from left to right. The numbers at arrows refer to the distance of the start or end position of a given gene from the corresponding junction site. The T bar above or below the genes indicate the extent of their parts with their corresponding values in base pair. The plotted genes and distances in the vicinity of the junction sites are the scaled projection of the genome. JLB (IRb /LSC), JSB (IRb/SSC), JSA (SSC/IRa) and JLA (IRa/LSC) denote the junction sites between each corresponding two regions of the genome.

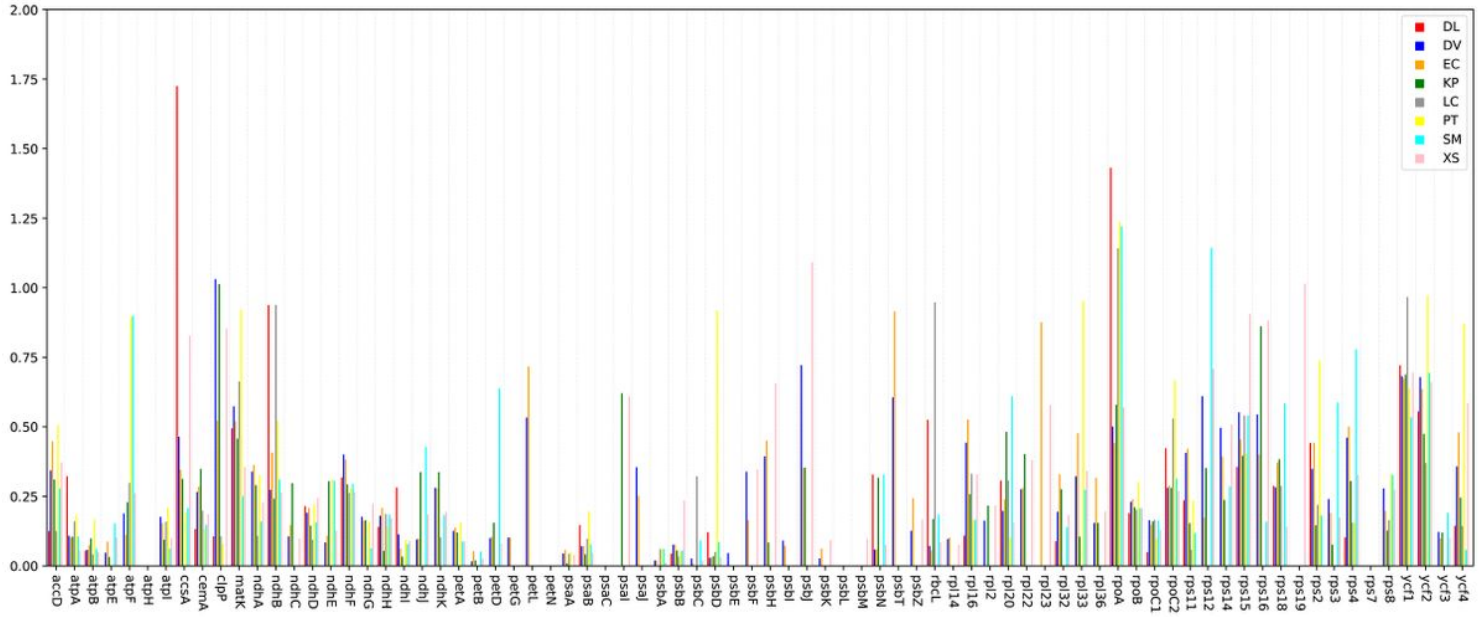

Figure 6

The Ka/Ks ratios of 78 protein-coding genes of the N. lappaceum chloroplast genome versus eight closely related species of Sapindaceae.

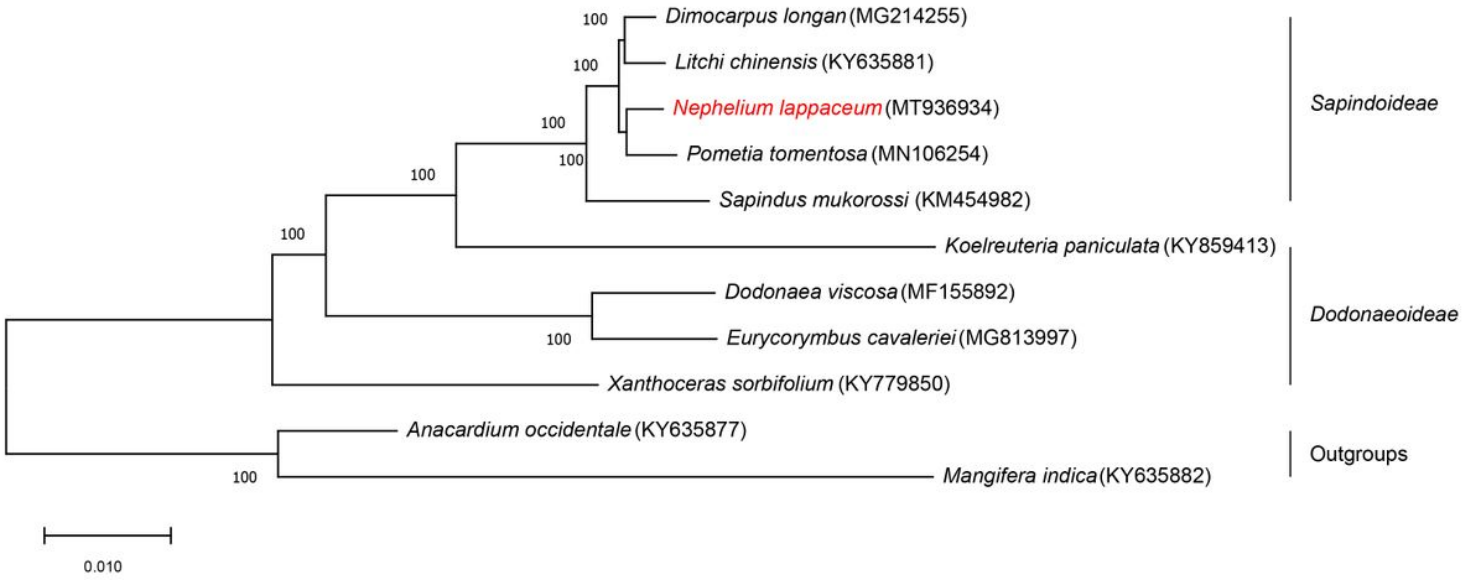

Figure 7

The Maximum likelihood (ML) phylogenetic tree of the Sapindaceae family based on chloroplast genome sequences. The numbers in each node were tested by bootstrap analysis with 1000 replicates. Anacardium occidentale and Mangifera indica were set as the outgroups. The position of N. lappaceum is indicated in red text.

Supplementary Files

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

- TableS1TableS8.xlsx
- TableS9.xlsx