Complete chloroplast genome sequence of *Buchanania latifolia* (Anacardiaceae): genome structure, and phylogenetic relationships

Chunmin Mao
Southwest Forestry University
Rui Rao
Southwest Forestry University
Quwen Lei
Southwest Forestry University
WanTing Chen
Liangliang Yue
yueliangliang@swfu.edu.cn
Southwest Forestry University

Research Article

Keywords: Buchanania latifolia Roxb., Anacardiaceae, Chloroplast genome, SSR loci, Phylogeny

Posted Date: June 11th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4552236/v1

Additional Declarations: The authors declare no competing interests.

Abstract

The chloroplast (cp) genomes are valuable resource with multiple applications, encompassing species identification, phylogenetic reconstruction, and evolutionary investigations. In this study, the complete chloroplast genome sequence of *Buchanania latifolia*was de novo sequenced, assembled and annotated. The chloroplast genome of *B. latifolia* exhibits a typical quadripartite structure, with a total length of 160,088 bp, containing 88 protein-coding sequences (CDS), 37 tRNA genes, and 8 rRNA genes, with an overall GC content of 37.7%. A total of 99 SSR loci and 63 repeat sequences were identified, which can be utilized for marker development, phylogenetic and population studies of *B. latifolia*. Codon usage analysis revealed a preference for Leu codons ending with A/U. Additionally, the study investigated IR boundaries, DNA polymorphism, positive selection suites, and phylogenetic position. Comparative analysis with five other species from the Anacardiaceae family confirmed the nearly identical and highly conserved chloroplast genome features of *B. latifolia*, which can be valuable for understanding the plastid evolution and evolutionary relationships within Anacardiaceae. Phylogenetic analysis reveals that *B. latifolia*is positioned at the base of Anacardiaceae, sister to *Choerospondias axillaris*, *Lannea coromandelica*, and *Sclerocarya birrea*. These findings could provide important genetic information for further research into breeding of Anacardiaceae, phylogeny, and evolution of *B. latifolia*.

Introduction

Buchanania latifolia Roxb. also known as Chironji, is a tree species of Anacardiaceae (Weber, 2018), native to the eastern Himalayas, Southwest China, and part of Southeast Asia. It's valued for its edible fruits and oil. As a pioneer species, it is commonly found in hot-dry valleys and forests in Yunnan, China. Initially named *Coniogeton arborescens*, it was reclassified under the genus *Buchanania* by Blume in 1850 (Parveen and Ilyas, 2021). The genus *Buchanania*, with *B. latifolia* as a notable species, is considered to occupy a basal position in the second ancestral clade of the Anacardiaceae family. This genus comprises approximately 25 species predominantly found in the tropical regions of Asia and Oceania. However, prior to this study, no plastid genome data of the genus *Buchanania* were available for phylogenetic reconstruction, which could significantly enhance our understanding of phylogenetic relationship, divergence time and biogeographic patterns of these basal clades within the Anacardiaceae family.

The next-generation sequencing (NGS) technology is generating vast amounts DNA data including both nuclear and organelle genomes. The chloroplast genome (cp genome), with its self-replication mechanism, relatively independent evolution, appropriate natural evolutionary rate, and unique maternal inheritance is typically highly conserved in terms of genome size, structure, gene content, and gene types in most terrestrial plants (Park and Lee, 2016, Li et al., 2019a). These conserved and divergent gene features provide crucial information for identifying, classifying, and reconstructing phylogenetic relationships between species and families (Jansen et al., 2007). In this study, the chloroplast genome of *B. latifolia* was sequenced, assembled, annotated and uploaded to NCBI database. The key features of the chloroplast genome were analyzed, including simple sequence repeat (SSR) loci and optimal codon usage boundaries.A plastid phylogenomic analysis was conducted to examine the phylogenetic relationships among the basal Anacardiaceae taxa. The publication of this plastome will widen our edge of knowledge for the applications both of DNA barcoding and for the biogeographic investigations of this tropical distributed woody clade of Angiosperm.

Materials and Methods

Plant material, sequencing, and genome assembly

Fresh leaves of the *Buchanania latifolia* were collected from Xinping County, Yunnan Province, China. (102°7'41.75"E, 24°1'54.84'N). The voucher specimen (yue2021006) was deposited in Wetland College of Southwest Forestry University, Kunming, China (email: mcm18213835817@126.com). Total DNA was extracted from fresh leaves using the E. Z. N. A ® HP Plant DNA Kit (Omega Bio-tek, GA, United States) following the manufacturer's protocol. Shotgun libraries (350 bp) were constructed using a TruSeq DNA Sample Prep Kit (Illumina, United States). Next-generation sequencing (NGS) was conducted on an Illumina nova-seq 6000 sequencing platform (Illumina, CA, United States), from which approximately 5 Gb of raw NGS data was obtained. Raw reads were filtered using the NGS QC Toolkit (Patel and Jain, 2012) to remove low quality reads and to trim off the sequence adaptors. The filtered reads were then fed into the GetOrganelle (Jin et al., 2020) pipeline for *de novo* genome assembly. The Chloroplast genome of *B. latifolia* was annotated using Plastid Genome Annotator (PGA) (Qu et al., 2019), using the platome of *Mangifera indica* (GenBank accession number: NC_035239) as a reference. As for the annotated results, the start & stop codons were checked manually, and the intron/exon boundaries of protein-coding genes were checked by using the Geneious Prime (Kearse et al., 2012). The chloroplast genome maps were drawn using CPGview (http://www.1kmpg.cn/cpgview/). Finally, the annotated chloroplast genome sequence was submitted to GenBank with an accession number of OM000214.

Genome structure analysis

The Geneious software was used to export the types and numbers of annotated genes in the cp genome into a CSV file (Kearse et al., 2012). Geneious was also used to identify the LSC, SSC, and IR regions in each cp genome sequence and to calculate the AT and GC content.

Codon preference analysis

The CDS coding sequences of the *B. latifolia* cp genome were extracted using Geneious (Kearse et al., 2012). The number of codons and the relative usage of synonymous codons were analyzed using CodonW 1.4.4 (Peden and John, 2000). Relative Synonymous Codon Usage (RSCU) refers to the relative probability of using a synonymous codon when encoding a specific amino acid. An RSCU > 1 indicates that the codon is used frequently, an RSCU = 1 indicates no preference, and an RSCU < 1 indicates low usage.

Repeat sequence analysis

Simple sequence repeats (SSRs) were identified and counted using MISA-web (Beier et al., 2017), in which the parameters for SSR identification were set at minimal repeat numbers of 10, 5, 4, 3, 3, and 3 for mono-, di-, tri-, tetra-, penta- and hexa- nucleotide repeats, respectively. For large repeats, REPuter (Kurtz et al., 2001) was used to identify forward, palindrome, reverse, and complementary long repeats. The hamming distance was set as 3, while the minimal repeat size was 30 bp.

Comparative analysis of chloroplast genome

IRscope was used to compare IR/SSC and IR/LSC boundaries and junctions among the six Anacardiaceae species (https://irscope.shinyapps.io/irapp/) (Amiryousefi et al., 2018). The mVISTA (https://genome.lbl.gov/vista/mvista/submit.shtml) (Frazer et al., 2004) was used to compare the complete chloroplast genome of the six Anacardiaceae species. Before nucleotide divergence analysis, the cp genome were aligned using MAFFT v7 (Katoh and Standley, 2013). The nucleotide variability of the six cp genomes were calculated using DnaSP v5.10 (Librado and Rozas, 2009), in which highly mutational hotspot regions were identified through a sliding window analysis. The length of the window frame was set at 4400 bp and a 200 bp step size was selected.

Nucleotide substitution rate (Ka/Ks)

A total of 79 protein-coding genes shared by the *B. latifolia* cp genome were extracted using Geneious. After aligning them with MAFFT, the online program ALTER (https://www.sing-group.org/ALTER/) was used to convert the file from '.fas' format to '.aln' format. Subsequently, the '.aln' files were converted to '.axt' format using AXTConvertor from the KaKs-Calculator in a Linux system (Zhang et al., 2006). Finally, the ratio of non-synonymous (Ka) to synonymous (Ks) substitutions was calculated using the KaKs-Calculator, selecting the NG model and setting the genetic code type to 11(bacterial and plant plastid codon). When Ka = 0 and Ks = 0, the value of Ka/Ks was represented by NA.

Phylogenetic analysis

We conducted a phylogenetic analysis of 37 chloroplast genomes in the family Anacardiaceae, using three species from the family Burseraceae, *Commiphora gileadensis, Commiphora wightii*, and *Bursera fagaroides*, as outgroups(The accession numbers can be found in Supporting Table 1). The Phylogenetic trees of complete chloroplast genomes were constructed using the maximum likelihood (ML) method. ML analysis was conducted using IQ-TREE (Nguyen et al., 2015), during which a substitution model of GTR + F + I were automatically calculated and applied with a bootstrap replicates of 1,000 times.

Results

Genome structure of B. latifolia chloroplast genome

The chloroplast (cp) genome of *B. latifolia* exhibits a quadripartite structure, characterized by a large single-copy (LSC) region, a small single-copy (SSC) region, and a pair of inverted repeats (IRa and IRb), measuring 87,689 bp, 17,941 bp, and 27,217 bp, respectively (refer to Fig. 1; Table 1). The overall nucleotide composition reveals a predominance of A (30.8%) and T (31.5%), with C (19.2%) and G (18.5%) constituting the remaining nucleotides, resulting in a total AT content of 62.3% and GC content of 37.7%. A comprehensive annotation identified a total of 133 genes within the cp genome of *B. latifolia*, including 88 protein-coding genes (CDS), 37 transfer RNA genes (tRNAs), and 8 ribosomal RNA genes (rRNAs). Notably, the AT content is higher than the GC content (62.3% vs. 37.7%). Moreover, the GC content varies across different regions, with the LSC, SSC, and IR regions exhibiting values of 35.8%, 32%, and 42.6%, respectively, with the IR regions demonstrating higher GC content compared to the LSC and SSC regions. Functional classification of the genes in the cp genome of *B. latifolia* categorizes them into four classes: photosynthesis-related genes (44), genes involved in self-replication (74), other annotated genes (6), and genes with unknown functions (7) (see Table 2).

Table 1
Characteristics of the cp genome of *B. latifolia.*

Category	Item	Describe
Chloroplast genome structure	Cp gene/bp	160088
	LSC/bp	87713
	SSC/bp	17941
	IRA/IRB/bp	27217
Gene composition	Cp gene	133
	CDS	88
	tRNA	37
	rRNA	8
GC Content (%)	Cp gene	37.7
	LSC	35.8
	SSC	32
	IRA/IRB	42.6

A total of 19 duplicated genes including one NADH dehydrogenase subunit gene (ndhB), five self-replication genes (*rpl2*, *rpl23*, *rps12*, *rps19*, *rps7*), four rRNA genes (*rrn16*, *rrn23*, *rrn4.5*, *rrn5*), seven tRNA genes (*trnA-UGC*, trnI-CAU, *trnI-GAU*, *trnL-CAA*, *trnN-GUU*, *trnR-ACG*, *trnV-GAC*), and two unknown function protein genes (*ycf15*, *ycf2*) were illustrated. Most genes are non-coding genes, and a few genes contain one or two introns.Among them, ndhA, ndhB, petB, petD, atpF, rpl16, rpl2, rps12, rps16, rpoC1, *trnA-UGC*, *trnI-GAU*, *trnK-UUU*, *trnL-UAA*, *trnT-CGU*, and *trnV-UAC* contain one intron, while *rps12*, *clpP* and *ycf3* contain three introns. It is worth noting that further research using advanced molecular techniques to determine the functions of the 7 genes with unknown functions were necessary (Table 2).

Table 2
Genes present in chloroplast genome of *B. latifolia*

Category	Gene group	Gene name
Photosynthesis	Subunits of photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*
	Subunits of photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*
	Subunits of NADH dehydrogenase	*ndhA**, *ndhB*(2)*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
	Subunits of cytochrome b/f complex	*petA*, *petB**, *petD**, *petG*, *petL*, *petN*
	Subunits of ATP synthase	*atpA*, *atpB*, *atpE*, *atpF**, *atpH*, *atpI*
	Large subunit of rubisco	*rbcL*
	Subunits photochlorophyllide reductase	-
Self-replication	Proteins of large ribosomal subunit	*rpl14*, *rpl16**, *rpl2*(2)*, *rpl20*, *rpl22*, *rpl23(2)*, *rpl32*, *rpl33*, *rpl36*
	Proteins of small ribosomal subunit	*rps11*, *rps12**(2)*, *rps14*, *rps15*, *rps16**, *rps18*, *rps19(2)*, *rps2*, *rps3*, *rps4*, *rps7(2)*, *rps8*
	Subunits of RNA polymerase	*rpoA*, *rpoB*, *rpoC1**, *rpoC2*
	Ribosomal RNAs	*rrn16(2)*, *rrn23(2)*, *rrn4.5(2)*, *rrn5(2)*
	Transfer RNAs	*trnA-UGC*(2)*, *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, *trnG-GCC*, *trnH-GUG*, *trnI-CAU(2)*, *trnI-GAU*(2)*, *trnK-UUU**, *trnL-CAA(2)*, *trnL-UAA**, *trnL-UAG*, *trnM-CAU*, *trnN-GUU(2)*, *trnP-UGG*, *trnQ-UUG*, *trnR-ACG(2)*, *trnR-UCU*, *trnS-GCU*, *trnS-GGA*, *trnS-UGA*, *trnT-CGU**, *trnT-GGU*, *trnT-UGU*, *trnV-GAC(2)*, *trnV-UAC**, *trnW-CCA*, *trnY-GUA*, *trnfM-CAU*
Other genes	Maturase	*matK*
	Protease	*clpP***
	Envelope membrane protein	*cemA*
	Acetyl-CoA carboxylase	*accD*
	c-type cytochrome synthesis gene	*ccsA*
	Translation initiation factor	*infA*
	other	-
Genes of unknown function	Conserved hypothetical chloroplast ORF	*ycf1*, *ycf15(2)*, *ycf2(2)*, *ycf3***, *ycf4*

Notes: Gene*: Gene with one introns; Gene**: Gene with two introns; #Gene: Pseudo gene; Gene (2): Number of copies of multi-copy genes.

Analysis of codon preference

There are significant differences in codon usage patterns among different species. Each amino acid is encoded by at least one codon and up to six codons. This unequal usage of synonymous codons is known as codon bias. Natural selection and base mutations are considered the main factors influencing codon bias. Based on 88 coding sequences (CDS), the codon usage frequency and Relative Synonymous Codon Usage (RSCU) of the cp genome were calculated. These CDS consist of 22,855 codons each, encoding 20 amino acids in the chloroplast genome. Among these, six codons encode arginine (Arg), leucine (Leu), and serine (Ser), while only one codon encodes methionine (Met) and tryptophan (Trp). Among them, leucine (Leu: 10.53%) is the most frequently utilized amino acid, while cysteine (Cys: 1.12%) is the least utilized amino acid in the cp genome. According to RSCU analysis, except for methionine (Met) and tryptophan (Trp), almost all amino acids are encoded by 2–6 synonymous codons. The relative synonymous codon usage for methionine (Met) and tryptophan (Trp) is 1. There were 30 high-frequency codons with RSCU > 1, among which 13 and 16 high-frequency codons ended with A and U, accounting for 93.55% of the total. Only one high-frequency codons ended with G, and no high-frequency codons ending with C were found. A preference of A or U to end codons was indicated for cp genome of the *B. latifolia* chloroplast.

Repeat sequences analysis

In this study, 99 SSRs were identified in the cp genome of *B. latifolia*, including 77 mononucleotides (Mono-), 8 dinucleotides (Di-), 5 trinucleotides (Tri-), 7 tetranucleotides (Tetra-), and 2 hexanucleotides (Hexan-) (Fig. 3A). 16 SSRs were located in the IR region, 22 SSRs in the SSC region, and the LSC region contained the highest number of SSRs, with 61 (Fig. 3B). The repeat units were primarily composed of A or T, with *B. latifolia* cp genome being A/T types rather than G/C types. Furthermore, dinucleotides are primarily composed of AT/AT, with only 1 or 2 occurrences of other nucleotide types. (Fig. 3C).

Within this cp genome, a total of 63 repeat sequences were identified. Among these, forward repeat sequences constituted 39.68% (25 repeats) of the total repeats, while reverse repeat sequences accounted for 3.17% (2 repeats). The remaining 57.14% (36 repeats) were attributed to palindromic repeat sequences. Notably, no complementary repeat sequences were detected (refer to Fig. 4A). Furthermore, the analysis revealed that repetitive sequences ≤ 40 bp in length were the most prevalent, with 37 occurrences. Subsequently, there were 10 occurrences in the 41–50 bp range, and 16 occurrences in the > 50 bp range (Fig. 4B).

IR expansion and contraction

The chloroplast genome of *B. latifolia* was compared with those of five reported chloroplast genomes from the Anacardiaceae family (Fig. 5). The main chloroplast marker genes, *rpl22*, *ndhF*, *ycf1*, *and trnH*, were found at the boundaries of LSC/IRA, IRA/SSC, SSC/IRB, and IRB/LSC, respectively. Among these genes, rpl22 primarily located in the LSC region or crossed the LSC/IRA boundary. In *B. latifolia*, *Dobinea delavayi*, *Lannea coromandelica*, and *Pegia nitida*, the *rpl22* gene was partly located within the IRb region, the IRb part of the *rpl22* genes ranged from 39 to 69 bp, while in *Searsia paniculata* and *Semecarpus reticulatus*, the entire *rpl22* gene was within the IRb region. The *ndhF* gene was mainly located in the LSC region or across the IRb/SSC boundary, showing positions ranging from 2208 to 2259 within the SSC region. Notably, among all the compared six Anacardiaceae species, the ycf1 gene crossed the IRa/SSC boundary, with positions ranging from 952 to 1479 bp within the IRa region and from 3996 to 4568 bp within the SSC region. The *ndhF* gene was located within the SSC region, with distances to the IRb/SSC junction of 5, 56, 61, 64, 143, and 154 bp. The *trnN* gene was entirely located within the IRa region and contracted by 971 to 1479 bp. The *trnH* gene is located within the LSC region, positioned 74–75 bp away from the IRa/LSC boundary.

Structural comparison and divergence hotspot identification analysis

Using the annotation of *B. latifolia* as reference, the chloroplast genome sequences of the five Anacardiaceae species were compared by mVISTA (Fig. 6). The sequence divergences remarkably differed among regions. The alignment result indicates that the chloroplast genome is extremely conserved with only few variations detected. The data revealed that the non-coding region was more divergent than coding counterparts. IR regions of all cp genomes were less diverged than the LSC and SSC regions. The rRNA genes were highly conserved comparing to other genes. The exons of the ycf1 gene were regions representing the highest polymorphism.

The diversity of nucleic acids can reveal variations in nucleic acid sequences among different species. Regions with high variability can serve as potential molecular markers for population genetics. The nucleotide diversity (Pi) value for the six cp genomes ranged from 0 to 0.14438, with an average of 0.02078. At a cut-off point set at Pi ≥ 0.1, four hypervariable regions, including *rps16 (exonl)-trnQ* (0.12183), *atpF (exon1)-atpH* (0.11283), *ndhF-rpl32*(0.1098), *rpl32-trnL* (0.10633), and *ycf1* (0.14433), were identified. Among the four hypervariable regions, two were in the LSC region, the other two were in the SSC region. None was detected in the IR region (Fig. 7).

Adaptive evolution analyses

Taking *B. latifolia* as a reference, alterations in 5 Anacardiaceae cp genomes were examined to uncover patterns of selection among protein coding genes (Table 3). Overall, the Ka/Ks ratios of 79 protein-coding genes shared by the cp genome of *B. latifolia* were compared. The results indicate that only the *psbT* (1.61) gene in *L. coromandelica* has a Ka/Ks ratio > 1. Additionally, in *S. paniculata*, the rpl22 (1.09), *rpl32* (1.23), *rps16* (1.64) and *ycf2* (2.04) genes have Ka/Ks ratios greater than 1, while all other genes have Ka/Ks ratios < 1.

Table 3
The 79 protein-coding genes in cp genome of the B. latifolia and 5 Anacardiaceae species were used for ka/ks analysis

Gene	*D. delavayi*	*L. coromandelica*	*P. nitida*	*S. paniculata*	*S. reticulatus*	*Gene*	*D. delavayi*	*L. coromandelica*	*P. nitida*	*S. paniculata*	*S. reticulatus*
accD	0.40	0.40	0.42	0.33	0.45	psbI	0.00	0.00	0.00	0.00	0.00
atpA	0.07	0.04	0.03	0.03	0.04	psbJ	NA	NA	NA	NA	NA
atpB	0.06	0.07	0.05	0.02	0.10	psbK	0.36	0.68	0.82	0.55	0.18
atpE	0.20	0.00	0.00	0.00	0.00	psbL	NA	0.00	0.00	NA	0.00
atpF	0.18	0.21	0.26	0.18	0.35	psbM	0.20	0.20	0.20	0.15	NA
atpH	0.00	0.00	0.00	0.00	0.00	psbN	0.00	0.00	0.00	0.00	0.00
atpI	0.04	0.06	0.06	0.06	0.09	psbT	0.52	1.61	0.98	0.73	0.64
ccsA	0.32	0.35	0.27	0.37	0.29	psbZ	0.00	0.17	0.17	0.00	NA
cemA	0.22	0.11	0.13	0.65	0.40	rbcL	0.30	0.16	0.17	0.18	0.21
clpP	0.03	0.03	0.04	0.19	0.10	rpl14	0.17	0.25	0.18	0.06	0.08
matK	0.52	0.58	0.48	0.60	0.76	rpl16	0.05	0.05	0.04	0.05	0.10
ndhA	0.23	0.14	0.15	0.23	0.04	rpl2	0.16	0.13	0.11	0.11	0.22
ndhB	0.21	0.16	0.31	0.21	0.16	rpl20	0.63	0.40	0.73	0.22	0.31
ndhC	0.08	0.15	0.22	0.17	0.15	rpl22	0.53	0.50	0.54	1.09	0.83
ndhD	0.16	0.21	0.14	0.17	0.13	rpl23	NA	NA	NA	0.29	NA
ndhE	0.18	0.40	0.30	0.12	0.15	rpl32	0.31	0.56	0.70	1.23	0.78
ndhF	0.22	0.32	0.29	0.26	0.27	rpl33	0.36	0.37	0.28	0.09	0.00
ndhG	0.31	0.24	0.31	0.40	0.19	rpl36	0.13	0.04	0.05	0.20	0.17
ndhH	0.04	0.08	0.06	0.07	0.03	rpoA	0.20	0.14	0.14	0.15	0.25
ndhI	0.24	0.52	0.53	0.27	0.64	rpoB	0.10	0.07	0.11	0.17	0.11
ndhJ	0.11	0.21	0.12	0.56	0.28	rpoC1	0.11	0.18	0.12	0.20	0.05
ndhK	0.11	0.11	0.14	0.35	0.19	rpoC2	0.37	0.35	0.41	0.32	0.31
petA	0.11	0.13	0.17	0.19	0.05	rps11	0.06	0.09	0.10	0.06	0.05
petB	0.03	0.03	0.03	0.03	0.21	rps12	0.18	0.12	0.35	0.00	0.00
petD	0.00	0.00	0.00	0.08	0.00	rps14	0.14	0.20	0.19	0.19	0.07
petG	0.00	NA	NA	NA	0.00	rps15	0.22	0.47	0.18	0.20	0.86
petL	NA	NA	NA	0.18	0.74	rps16	0.21	0.97	0.65	1.64	0.64
petN	NA	0.00	0.32	NA	NA	rps18	0.10	0.30	0.15	0.15	0.30
psaA	0.06	0.06	0.05	0.08	0.06	rps19	0.00	0.05	0.11	0.00	0.00
psaB	0.02	0.02	0.03	0.01	0.00	rps2	0.09	0.09	0.08	0.16	0.12
psaC	0.00	0.00	0.00	0.00	0.00	rps3	0.25	0.16	0.15	0.14	0.14
psaI	0.15	0.61	0.61	NA	NA	rps4	0.11	0.17	0.15	0.35	0.24
psaJ	0.33	0.33	NA	0.33	0.33	rps7	0.32	NA	NA	NA	NA
psbA	0.02	0.02	0.02	0.02	0.06	rps8	0.16	0.08	0.07	0.07	0.05
psbB	0.03	0.05	0.01	0.08	0.07	ycf1	0.73	0.87	0.99	0.98	0.77
psbC	0.02	0.03	0.01	0.04	0.08	ycf15	0.00	0.00	0.00	0.00	NA
psbD	0.00	0.00	0.00	0.00	0.00	ycf2	0.81	0.88	0.83	2.04	0.97
psbE	0.00	0.00	0.00	0.00	0.00	ycf3	0.19	0.29	0.17	0.29	0.29
psbF	0.00	0.00	0.00	0.00	0.00	ycf4	0.21	0.16	0.14	0.42	0.23
psbH	0.04	0.00	0.00	0.00	0.17						

Phylogenetic relationships of Anacardiaceae species

B. latifolia is positioned within the basal clade of the Anacardiaceae family, sister to the clade containing *Choerospondias axillaris*, *Lannea coromandelica*, and *Sclerocarya birrea* (Fig. 8). *Buchanania* is grouped on the same branch with *Choerospondias*, *Lannea*, and *Sclerocarya* of the Spondioideae. *Anacardium* and *Mangifera* are clustered with *Semecarpus* into the clade of Semecarpeae.

Discussion

Homology: Feature of Chloroplast Genomes

In this study, the chloroplast genome features and phylogenetic relationships of *Buchanania latifolia* were comprehensively analyzed. The chloroplast genome of *B. latifolia* showed a high degree of homology compared to other reported chloroplast genomes of Anacardiaceae species, including *Dracontomelon delavayi*, *Lannea coromandelica*, *Pistacia nitida*, *Spondias paniculata*, and *Swintonia reticulata*. The genome lengths of these species ranged from 159,485 to 162,509 bp, with a maximum difference of 3,024 bp. The overall GC content is comparable to that of other species in the Anacardiaceae family, such as *Rhus chinensis* (37.79%), *Pistacia weinmannifolia* (37.84%), *Toxicodendron vernicifluum* (37.96%), and *Cotinus species* (37.9%-38.1%)

(Wang et al., 2020, Zheng et al., 2018, Liu et al., 2023). The number and types of genes were also very similar, reflecting the highly conserved characteristics of chloroplast genomes. The homology among closely related woody species is typically considered reasonable due to their long generation times and the relatively low number of substitutions occurring over a given period. Substitutions from parent plants can only be passed to offspring during the process of germination.

The study of codon preference aids in understanding the evolutionary processes of plant species and optimizing the expression of exogenous genes in chloroplasts, enabling the prediction of gene function and expression levels (Li et al., 2019b). Consistent with previous findings, *Buchanania latifolia* demonstrates a preference for A or U bases in its chloroplast genome codons, a common trait in plant chloroplast genomes (Zhou et al., 2008). In B. latifolia, 30 high-frequency codons were identified, with 29 of them ending with an A or U base, likely influenced by natural selection and mutations (Necşulea and Lobry, 2007). Previous studies have also noted that in the chloroplast genomes of Anacardiaceae, high-frequency codons tend to utilize A or U bases as the third codon base (Liu et al., 2023, Xin et al., 2023, Wang et al., 2020).

The most prevalent SSRs in the chloroplast genomes of *B. latifolia* were mononucleotide repeats. Similar to other plants, chloroplast SSRs predominantly consist of short poly-A or poly-T repeats, with mononucleotide repeats being the most common forms (Tao et al., 2023, Vu et al., 2020, Djedid et al., 2021, Provan et al., 2001, Yang et al., 2020). Moreover, the majority of SSRs are located in the LSC and SSC regions, consistent with previous findings on chloroplast genomes (Alshegaihi, 2024, Liu et al., 2023, Wang et al., 2020). Palindrome sequences accounted for 39.68%, forward repeated sequences for 57.14%, and reverse sequences for 3.17%. No complementary sequences were found. Forward and palindromic repeats were the most common repeat types, with most dispersed repeats being less than 40 bp, as reported in previous studies (Liang et al., 2020, Kirov et al., 2020, Tian et al., 2021, Yuan et al., 2005).

The IR region of the chloroplast genome is thought to be the most conservative section. The expansion and contraction of the IR region are pivotal factors influencing the length variation observed in plant chloroplast genomes, typically categorized into two types (Yi et al., 2013, Zhang et al., 2013). Such expansions and contractions in the IR region across most species manifest as minor deviations in the IR/SC boundary within a few fixed genes, which may lead to pseudogenization of certain genes. The examination of the chloroplast genome of *B. latifolia* and the Anacardiaceae family in this study aligns with prior reports in angiosperms (Liu et al., 2023, Wang et al., 2020, Xin et al., 2023). Typically, the LSC/IRb boundary is situated on or near *rps19*, *rpl2*, or *rpl22*, while IRb/SSC is generally positioned on *ycf1* or between *ycf1* and *ndhF*. The SSC/IRa boundary typically lies on *ycf1*, and the IRa/LSC boundary usually falls on or near *rps19*, *rpl2*, *rpl12*, and *trnH* (Alshegaihi, 2024).

Intergenic spacers are more divergent than introns and protein-coding sequences (Meng et al., 2018). However, pseudogenes suffer the same fate as intergenic spacers due to a lack of functional importance, leading to less conservative strains. Pseudogenization was common in the evolution of chloroplast genomes, such as *accD*, *ccsA*, *ycf1*, *rps19* and *psbB* pseudogenes (Krawczyk et al., 2018, Li et al., 2021). The most divergent genes among the six Anacardiaceae species were two pseudogenes *accD* and *ycf1*, an intron of the *rps16* genes, and the intergenic spacer *ndhF-rpl32*. These genes or regions can be utilized for developing molecular markers for species identification and population studies (Magdy et al., 2019).,

The synonymous (Ks) and non-synonymous (Ka) nucleotide substitution pattern is a well-recognized marker for assessing genome evolution; and the Ka/Ks ratio reflects selection pressure on genes (Yang and Nielsen, 2000, Guo et al., 2017). Ka/Ks < 1, Ka/Ks = 1, and Ka/Ks > 1 indicate genes that underwent purifying, neutral, and positive selections, respectively (Yang and Nielsen, 2000). Generally, synonymous mutations occur more frequently than nonsynonymous mutations within genes, causing the Ka/Ks values to be below 1 (Makałowski and Boguski, 1998). Our results using *B. latifolia* and five species from the Anacardiaceae family indicate that only the *psbT* gene in *L. coromandelica* and the *rpl22*, *rpl32*, *rps16* and *ycf2* genes in *S. paniculata* have Ka/Ks values > 1, suggesting strong positive selection acting on these two genes. Ka/Ks values for all other detected genes are below 1, indicating widespread purifying selection on these chloroplast genomes.

Application on Phylogenetic Reconstruction

The informative sites provided by single genes or gene combinations are often limited. Differences in the evolutionary trajectories of various gene sequences can result in low resolution of the constructed phylogenetic tree, posing challenges in elucidating the evolutionary relationships within the Anacardiaceae family. To optimize phylogenetic outcomes, we conducted a phylogenomic analysis of Anacardiaceae based on complete chloroplast genomes using the maximum likelihood method. Overall, branch support at the generic level was generally high, with only relatively lower support observed for the branch of the genus Rhus and its sister groups (BS = 68), while support for all other branches was robust (BS = 100). This suggests that our analysis accurately reflects the phylogenetic relationships within Anacardiaceae. *B. latifolia* is sister to *Choerospondias*, *Lannea*, and *Sclerocarya*, consistent with previous phylogenetic trees constructed using sequences such as nuclear ribosomal external transcribed spacer (ETS), the chloroplast *trnL* intron and *trnL-F* intergenic spacer (*trnL-F* region), and the chloroplast rps16 intron (Weeks et al., 2014). However, *Choerospondias*, *Lannea*, and *Sclerocarya* all belong to the tribe/subfamily Spondiadeae/Spondioideae. Phylogenetically, they are distant from *Anacardium* and *Mangifera*, which are in the same tribe/subfamily Anacardieae/Anacardioideae as *Buchanania*. Although *Buchanania* is typically considered a member of the tribe/subfamily Anacardieae/Anacardioideae, it is here regarded as a sister group to those traditionally belonging to the tribe/subfamily Spondiadeae/Spondioideae, which aligns with Weeks' view (Weeks et al., 2014).

Conclusions

In this study, the complete chloroplast genome sequence of *B. latifolia* was assembled and annotated de novo. The chloroplast genome of *B. latifolia* exhibits a typical quadripartite structure, with a length of 160,088 bp, containing 88 protein-coding sequences (CDS), 37 tRNA genes, and 8 rRNA genes, with a GC content of 37.7%. A total of 99 SSR loci and 63 repeat sequences were identified, which can be utilized for marker development, phylogenetic and population studies of *B. latifolia*. Codon usage analysis revealed a preference for Leu codons ending with A/U. Additionally, the study investigated IR boundaries, DNA polymorphism, positive selection suites, and phylogenetic position. Comparative analysis with five other species from the Anacardiaceae family confirmed the nearly identical and highly conserved chloroplast genome features of *B. latifolia*, which can be valuable for understanding the plastid evolution and evolutionary relationships within Anacardiaceae. Phylogenetic analysis reveals that *B. latifolia* is positioned at the base of Anacardiaceae, sister to *Choerospondias*, *Lannea*, and *Sclerocarya*. These fndings could provide important genetic information for further research into breeding of Anacardiaceae, phylogeny, and evolution of *B. latifolia*. This study reveals the structural characteristics of the chloroplast genome, SSR loci, and phylogenetic analysis, providing valuable genetic information for understanding the origin and evolution of Anacardiaceae in the future.

Declarations

Data Availability

The genome sequence data that support the fndings of this study are openly available in GenBank of NCBI at https://www.ncbi.nlm.nih.gov/ under the accession OM000214.1. The associated BioProject, SRA, and Bio-Sample numbers are PRJNA783515, SRR17036892, and SAMN23436132 respectively.

Funding

This research was funded by the National Natural Science Foundation of China (grant No. 42161015). The Key Project of Basic Research of Yunnan Province, China (202202AS07337) to Gao Chen. This work is supported by Yunnan Science and Technology Talent and platform Program (202305AM340008).

Acknowledgements

Thanks to Zhefei Zeng and Kaiwen Jiang for their support of field sampling.

Authors and Affiliations

(1) Yunnan Key Laboratory of Plateau Wetland Conservation, Restoration and Ecological Services; National Plateau Wetlands Research Center; Dianchi Lake Ecosystem Observation and Research Station of Yunnan Province; Southwest Forestry University, Kunming, PR, 650224, China

Chunmin Mao, Rui Rao, WanTing Chen, Liangliang Yue

(2) Technology Centre of Kunming Customs

Quwen Lei

Contributions

LLY designed the study. CMM performed statistical and bioinformatic analyzes and wrote the first draft. RR proofreads the image text and provides suggestions for revision. QWL sampling, data analysis. WTC Molecular Materials Management. CMM and RR contribute equally in this study. All-authors read and approved the final manuscript.

Corresponding author

Correspondence to Liangliang Yue(yueliangliang@swfu.edu.cn)

Ethics declarations

Ethical approval

Not applicable.

Informed consent.

Not applicable.

Conflict of interest

The authors declare no competing interests.

Additional information

Communicated by Ewa Liangliang Yue.

References

1. Alshegaihi R M 2024. The complete chloroplast genome of the halophyte flowering plant Suaeda monoica from Jeddah, Saudi Arabia. Molecular Biology Reports [J], 51. https://doi.org/10.1007/s11033-023-09069-x
2. Amiryousefi A, Hyvönen J, Poczai P 2018. IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics [J], 34: 3030-3031. https://doi.org/10.1093/bioinformatics/bty220
3. Beier S, Thiel T, Münch T, et al. 2017. MISA-web: a web server for microsatellite prediction. Bioinformatics [J], 33: 2583-2585. https://doi.org/10.1093/bioinformatics/btx198
4. Djedid I K, Terzaghi M, Brundu G, et al. 2021. Genetic diversity and differentiation of eleven medicago species from campania region revealed by nuclear and chloroplast microsatellites markers. Genes [J], 13: 97. https://doi.org/10.3390/genes13010097
5. Frazer K A, Pachter L, Poliakov A, et al. 2004. VISTA: computational tools for comparative genomics. Nucleic acids research [J], 32: W273-W279. https://doi.org/10.1093/nar/gkh458
6. Guo Y, Liu J, Zhang J, et al. 2017. Selective modes determine evolutionary rates, gene compactness and expression patterns in Brassica. The Plant Journal [J], 91: 34-44. https://doi.org/10.1111/tpj.13541
7. Jansen R K, Cai Z, Raubeson L A, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proceedings of the National Academy of Sciences [J], 104: 19369-19374. https://doi.org/10.1073/pnas.0709121104
8. Jin J-J, Yu W-B, Yang J-B, et al. 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome biology [J], 21: 1-31. https://doi.org/10.1186/s13059-020-02154-5
9. Katoh K, Standley D M 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution [J], 30: 772-780. https://doi.org/10.1093/molbev/mst010
10. Kearse M, Moir R, Wilson A, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics [J], 28: 1647-1649. https://doi.org/10.1093/bioinformatics/bts199
11. Kirov I, Odintsov S, Omarov M, et al. 2020. Functional Allium fistulosum centromeres comprise arrays of a long satellite repeat, insertions of retrotransposons and chloroplast DNA. Frontiers in Plant Science [J], 11: 562001. https://doi.org/10.3389/fpls.2020.562001
12. Krawczyk K, Wiland-Szymańska J, Buczkowska-Chmielewska K, et al. 2018. The complete chloroplast genome of a rare orchid species Liparis loeselii (L.). Conservation genetics resources [J], 10: 305-308. https://doi.org/10.1007/s12686-017-0809-y
13. Kurtz S, Choudhuri J V, Ohlebusch E, et al. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic acids research [J], 29: 4633-4642. https://doi.org/10.1093/nar/29.22.4633
14. Li W, Zhang C, Guo X, et al. 2019a. Complete chloroplast genome of Camellia japonica genome structures, comparative and phylogenetic analysis. PLoS One [J], 14: e0216645. https://doi.org/10.1371/journal.pone.0216645
15. Li X, Yang J-B, Wang H, et al. 2021. Plastid NDH pseudogenization and gene loss in a recently derived lineage from the largest hemiparasitic plant genus Pedicularis (Orobanchaceae). Plant and Cell Physiology [J], 62: 971-984. https://doi.org/10.1093/pcp/pcab074
16. Li Y, Sylvester S P, Li M, et al. 2019b. The complete plastid genome of Magnolia zenii and genetic comparison to Magnoliaceae species. Molecules [J], 24: 261. https://doi.org/10.3390/molecules24020261
17. Liang H, Zhang Y, Deng J, et al. 2020. The complete chloroplast genome sequences of 14 Curcuma species: insights into genome evolution and phylogenetic relationships within Zingiberales. Frontiers in Genetics [J], 11: 802. https://doi.org/10.3389/fgene.2020.00802
18. Librado P, Rozas J 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics [J], 25: 1451-1452. https://doi.org/10.1093/bioinformatics/btp187

19. Liu Q, Yang N, Dong W, et al. 2023. Molecular evolution and phylogenomic analysis of complete chloroplast genomes of Cotinus (Anacardiaceae). Ecology and Evolution [J], 13: e10134. https://doi.org/10.1002/ece3.10134
20. Magdy M, Ou L, Yu H, et al. 2019. Pan-plastome approach empowers the assessment of genetic variation in cultivated Capsicum species. Horticulture Research [J], 6. https://doi.org/10.1038/s41438-019-0191-x
21. Makałowski W, Boguski M S 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proceedings of the National Academy of Sciences [J], 95: 9407-9412. https://doi.org/10.1073/pnas.95.16.9407
22. Meng X-X, Xian Y-F, Xiang L, et al. 2018. Complete chloroplast genomes from Sanguisorba: identity and variation among four species. Molecules [J], 23: 2137. https://doi.org/10.3390/molecules23092137
23. Necşulea A, Lobry J R 2007. A new method for assessing the effect of replication on DNA base composition asymmetry. Molecular biology and evolution [J], 24: 2169-2179. https://doi.org/10.1093/molbev/msm148
24. Nguyen L-T, Schmidt H A, Von Haeseler A, et al. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution [J], 32: 268-274. https://doi.org/10.1093/molbev/msu300
25. Park J H, Lee J 2016. The complete plastid genome of Scopolia parviflora (Dunn.) Nakai (Solanaceae). Korean Journal of Plant Taxonomy [J], 46: 60-64. https://doi.org/10.11110/kjpt.2016.46.1.60
26. Parveen T, Ilyas O 2021. Tree diversity and natural regeneration in Tropical Dry Deciduous Forest of Panna Tiger Reserve, India. https://doi.org/10.21203/rs.3.rs-629351/v1
27. Peden, John F 2000. Analysis of codon usage. University of Nottingham [J], 90: 73–74.
28. Provan J, Powell W, Hollingsworth P M 2001. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends in ecology & evolution [J], 16: 142-147. https://doi.org/10.1016/S0169-5347(00)02097-8
29. Qu X-J, Moore M J, Li D-Z, et al. 2019. PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. Plant Methods [J], 15: 1-12. https://doi.org/10.1186/s13007-019-0435-7
30. Tao L, Duan H, Tao K, et al. 2023. Complete chloroplast genome structural characterization of two Phalaenopsis (Orchidaceae) species and comparative analysis with their alliance. BMC Genomics [J], 24: 359. https://doi.org/10.1186/s12864-023-09448-5
31. Tian C, Li X, Wu Z, et al. 2021. Characterization and comparative analysis of complete chloroplast genomes of three species from the genus Astragalus (Leguminosae). Frontiers in Genetics [J], 12: 705482. https://doi.org/10.3389/fgene.2021.705482
32. Vu H-T, Tran N, Nguyen T-D, et al. 2020. Complete chloroplast genome of Paphiopedilum delenatii and phylogenetic relationships among Orchidaceae. Plants [J], 9: 61. https://doi.org/10.3390/plants9010061
33. Wang L, He N, Li Y, et al. 2020. Complete chloroplast genome sequence of Chinese lacquer tree (Toxicodendron vernicifluum, Anacardiaceae) and its phylogenetic significance. BioMed Research International [J], 2020. https://doi.org/10.1155/2020/9014873
34. Weber A 2018. A garden as a niche: botany and imperial politics in the early nineteenth century Dutch empire. Studium [J], 11: 178-190. https://doi.org/10.18352/studium.10177
35. Weeks A, Zapata F, Pell S K, et al. 2014. To move or to evolve: contrasting patterns of intercontinental connectivity and climatic niche evolution in "Terebinthaceae" (Anacardiaceae and Burseraceae). Front Genet [J], 5: 409. https://doi.org/10.3389/fgene.2014.00409
36. Xin Y, Yu W-B, Eiadthong W, et al. 2023. Comparative analyses of 18 complete chloroplast genomes from eleven Mangifera species (Anacardiaceae): sequence characteristics and phylogenomics. Horticulturae [J], 9: 86. https://doi.org/10.3390/horticulturae9010086
37. Yang J-P, Zhu Z-L, Fan Y-J, et al. 2020. Comparative plastomic analysis of three Bulbophyllum medicinal plants and its significance in species identification. Acta Pharmaceutica Sinica [J]: 2736-2745.
38. Yang Z, Nielsen R 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular biology and evolution [J], 17: 32-43. https://doi.org/10.1093/oxfordjournals.molbev.a026236
39. Yi X, Gao L, Wang B, et al. 2013. The complete chloroplast genome sequence of Cephalotaxus oliveri (Cephalotaxaceae): evolutionary comparison of Cephalotaxus chloroplast DNAs and insights into the loss of inverted repeat copies in gymnosperms. Genome biology and evolution [J], 5: 688-698. https://doi.org/10.1093/gbe/evt042
40. Yuan Y-M, Wohlhauser S, Möller M, et al. 2005. Phylogeny and biogeography of Exacum (Gentianaceae): a disjunctive distribution in the Indian Ocean Basin resulted from long distance dispersal and extensive radiation. Systematic biology [J], 54: 21-34. https://doi.org/10.1080/10635150590905867
41. Zhang H, Li C, Miao H, et al. 2013. Insights from the complete chloroplast genome into the evolution of Sesamum indicum L. PLoS One [J], 8: e80508. https://doi.org/10.1371/journal.pone.0080508
42. Zhang Z, Li J, Zhao X Q, et al. 2006. KaKscalculating Ka and Ks through model selection and model averaging. 4: 0-263. https://doi.org/10.1016/S1672-0229(07)60007-2
43. Zheng W, Li K, Wang W, et al. 2018. The complete chloroplast genome of the threatened Pistacia weinmannifolia, an economically and horticulturally important evergreen plant. Conservation genetics resources [J], 10: 535-538. https://doi.org/10.1007/s12686-017-0871-5
44. Zhou M, Long W, Li X 2008. Patterns of synonymous codon usage bias in chloroplast genomes of seed plants. Forestry Studies in China [J], 10: 235-242. https://doi.org/10.1007/s11632-008-0047-1

Figures

Buchanania latifolia

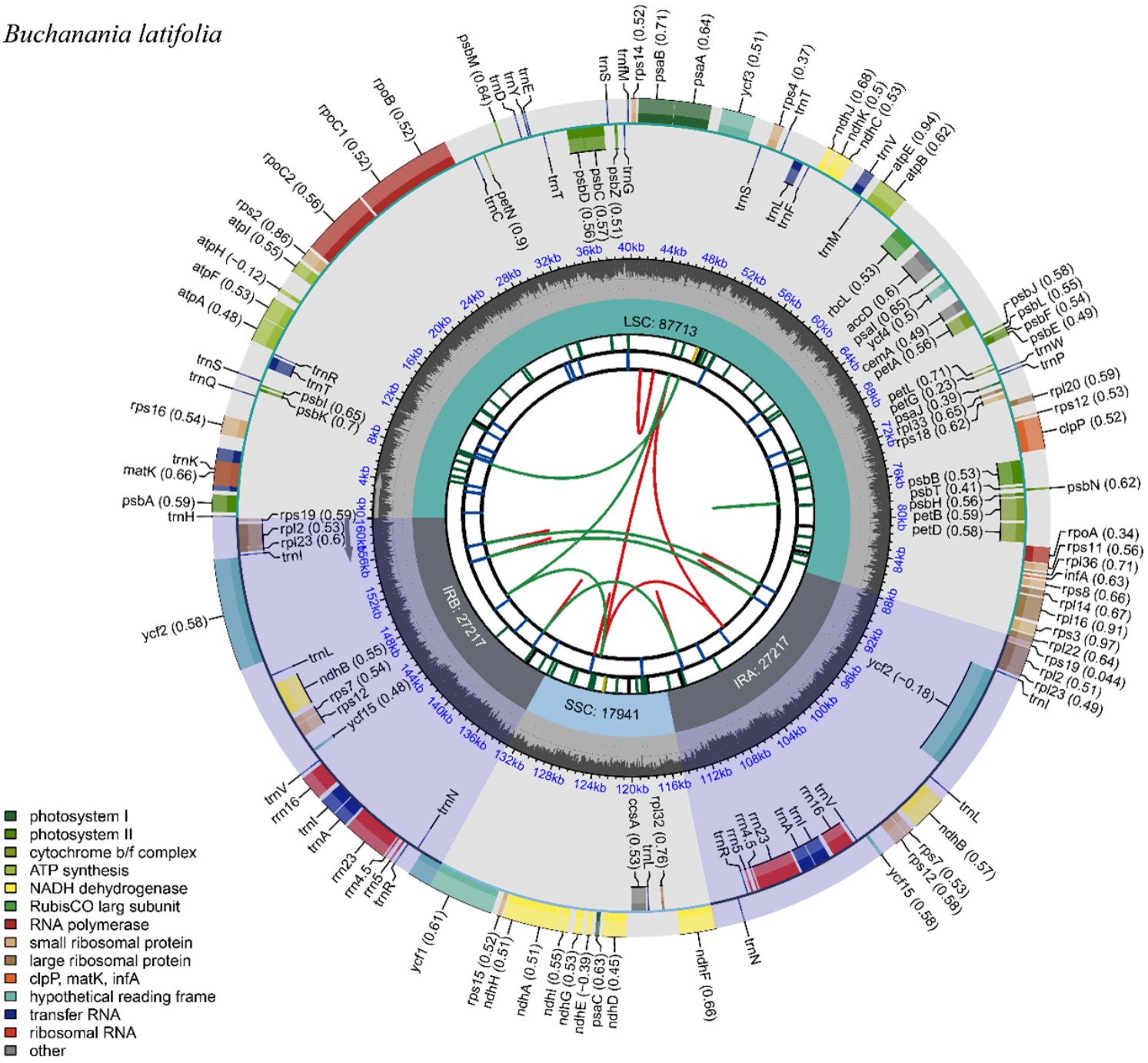

Figure 1

Gene map for cp genome of the *B. latifolia.*

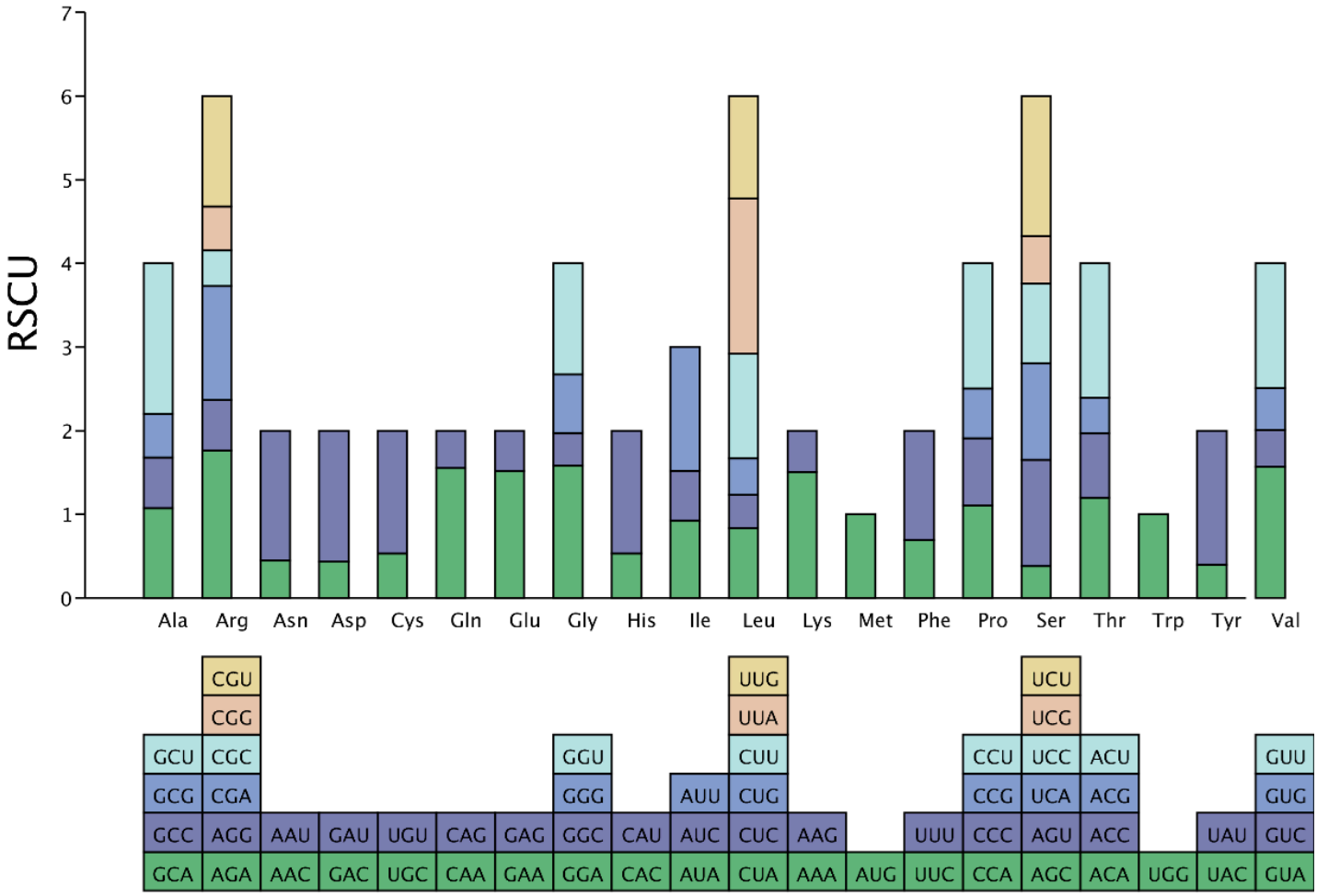

Figure 2

Counting of relative synonymous codon usage (RSCU) of amino acids in the cp genome of *B. latifolia*. The colors of the histogram correspond to the colors of the codons.

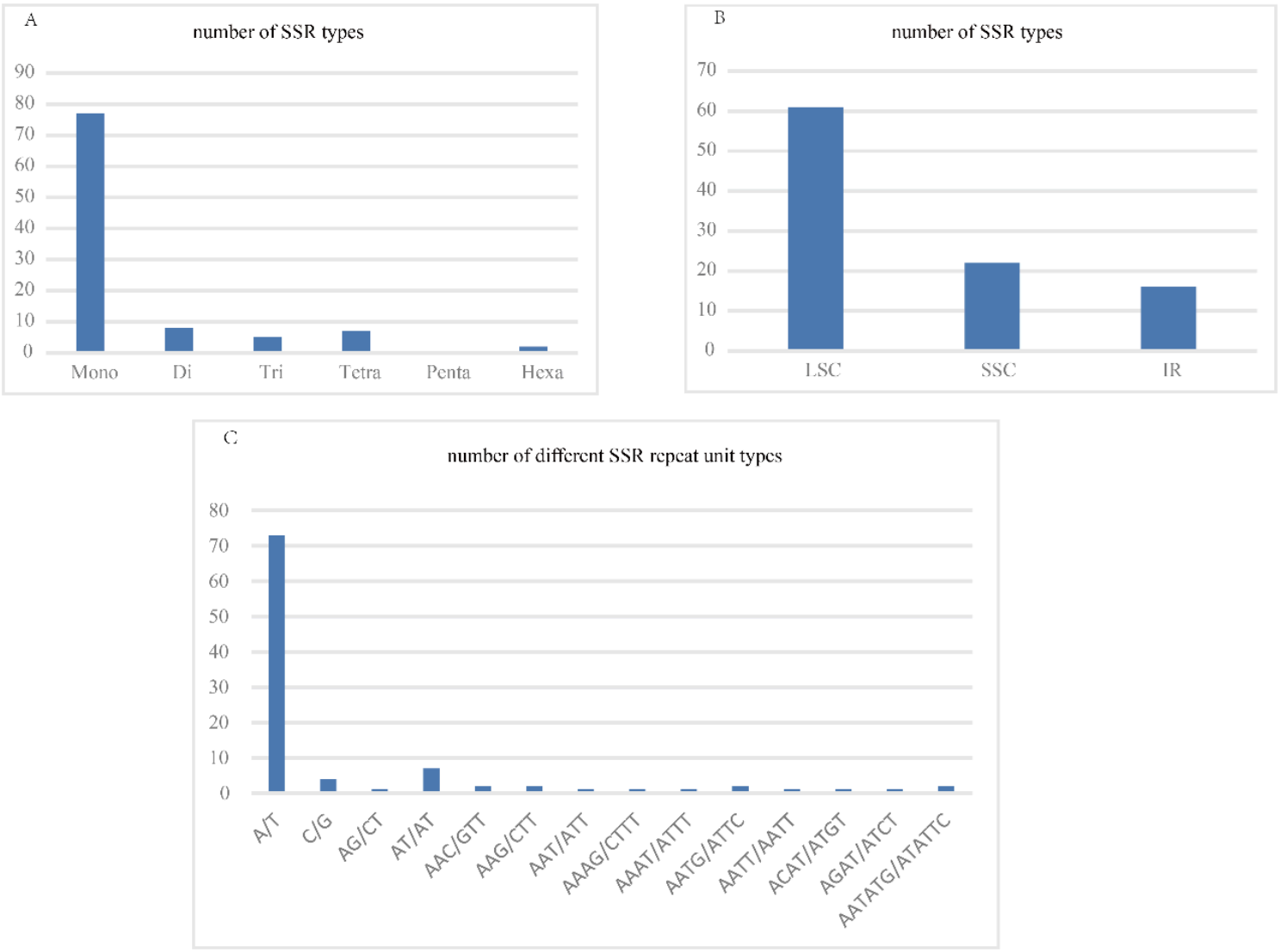

Figure 3

Analysis of chloroplast SSRs. A: Number of SSR types. B: Number of SSRs in different copy regions. C: Number of different SSR repeat unit types.

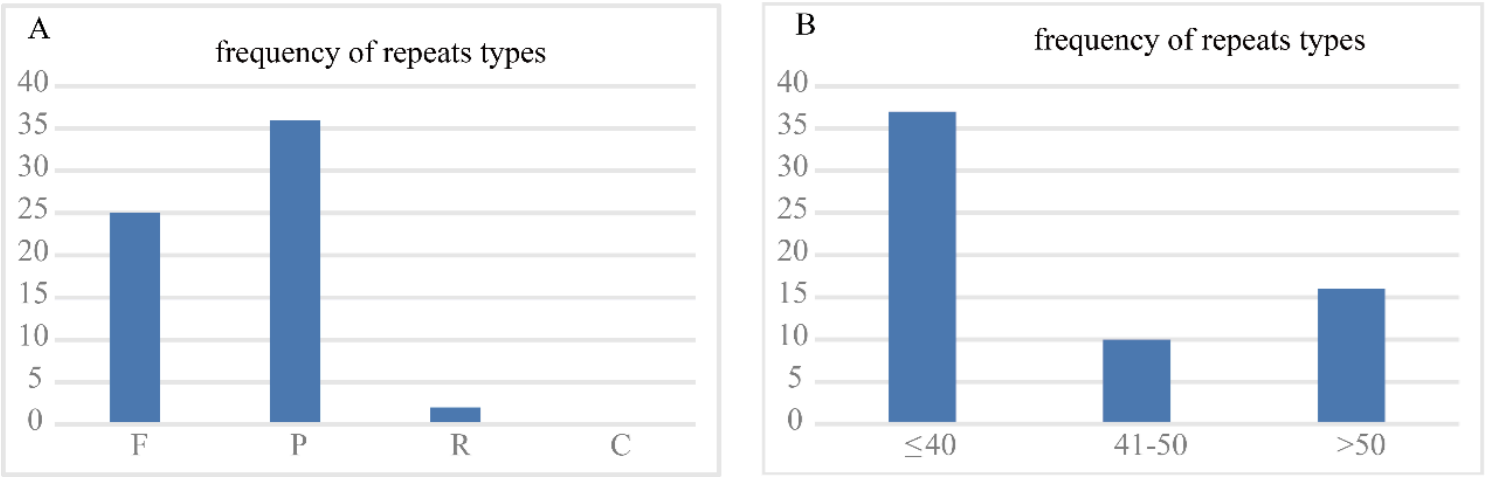

Figure 4

Analysis of chloroplast genome repeat sequences. A: Frequency of four types of repeat sequences. B: Number of tandem copies. Abbreviations: C, complement; F, forward; P, palindrome; R, reverse.

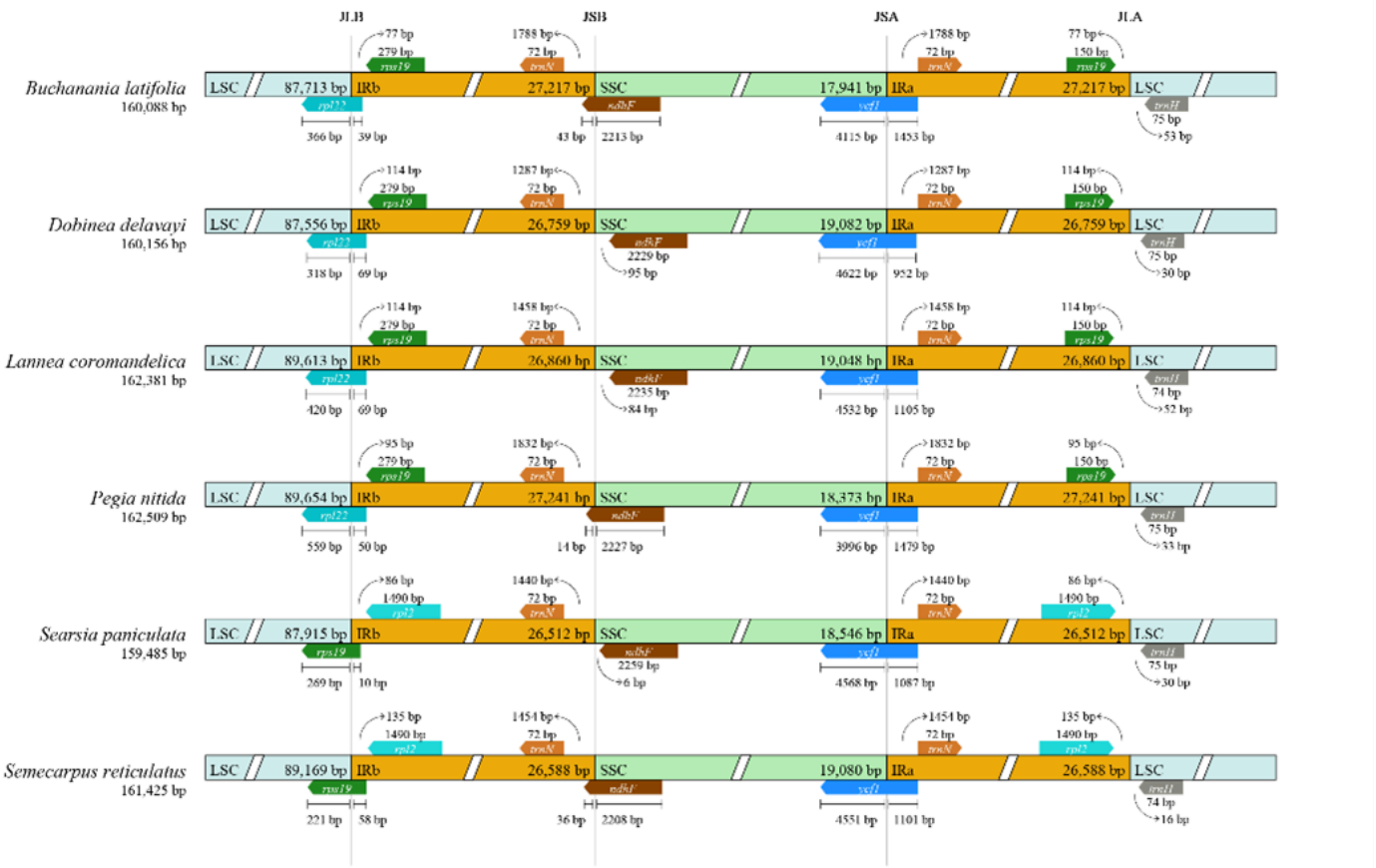

Figure 5

The borders of the large single-copy (LSC), small single-copy (SSC), and inverted repeat (IR) regions in cp genomes among 6 Anacardiaceae species or varieties. Genes are denoted in colored boxes. The distribution of base lengths (bp) at boundaries are indicated.

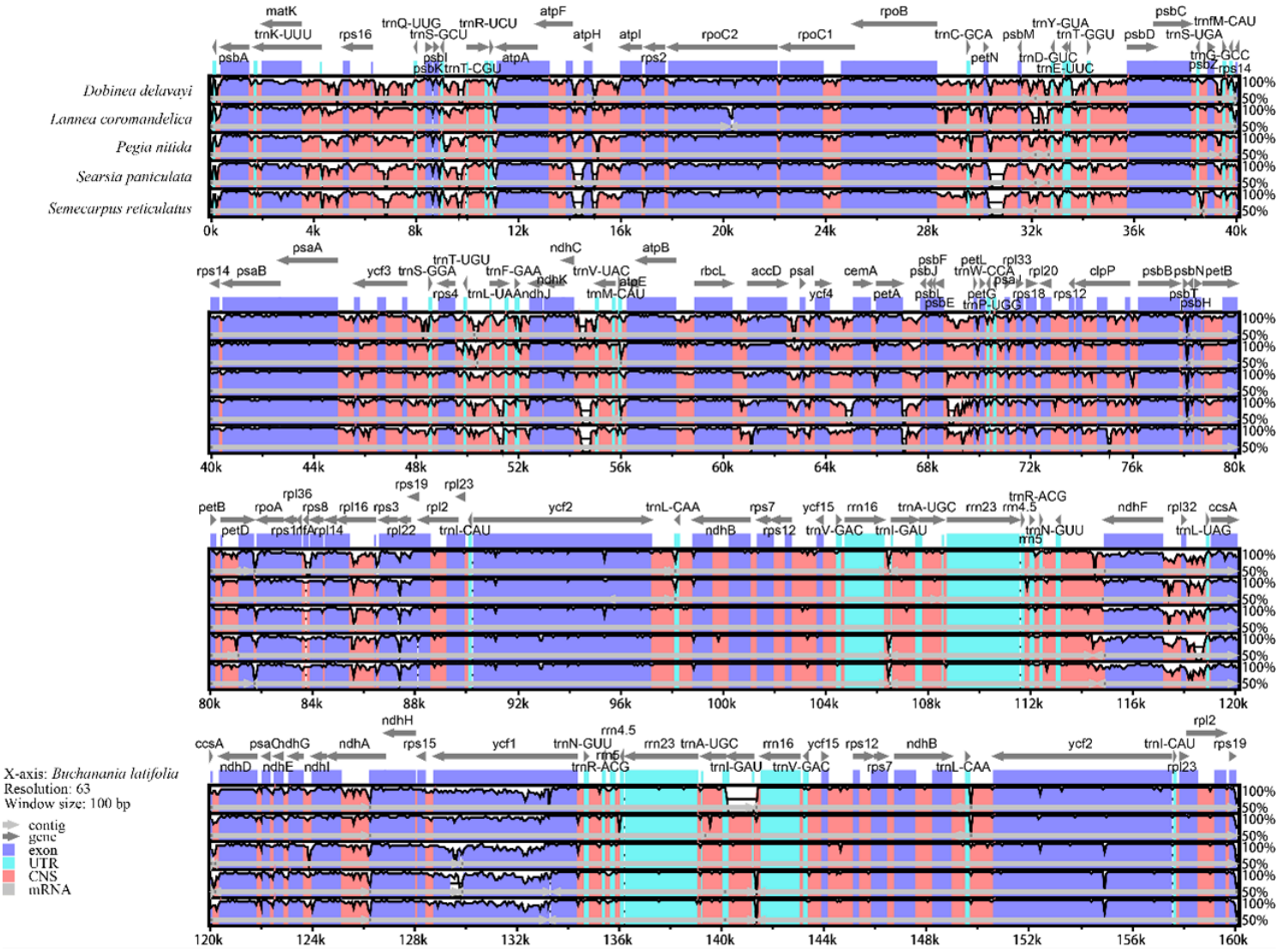

Figure 6

Sequence alignment of eight Phalaenopsis chloroplast genomes using mVISTA. The vertical scale indicates the percentage of identity, ranging from 50 to 100%. The horizontal axis indicated the coordinates within the chloroplast genome. Genome regions were color coded as exon, intron, and conserved non-coding sequences (CNS) and mRNA.

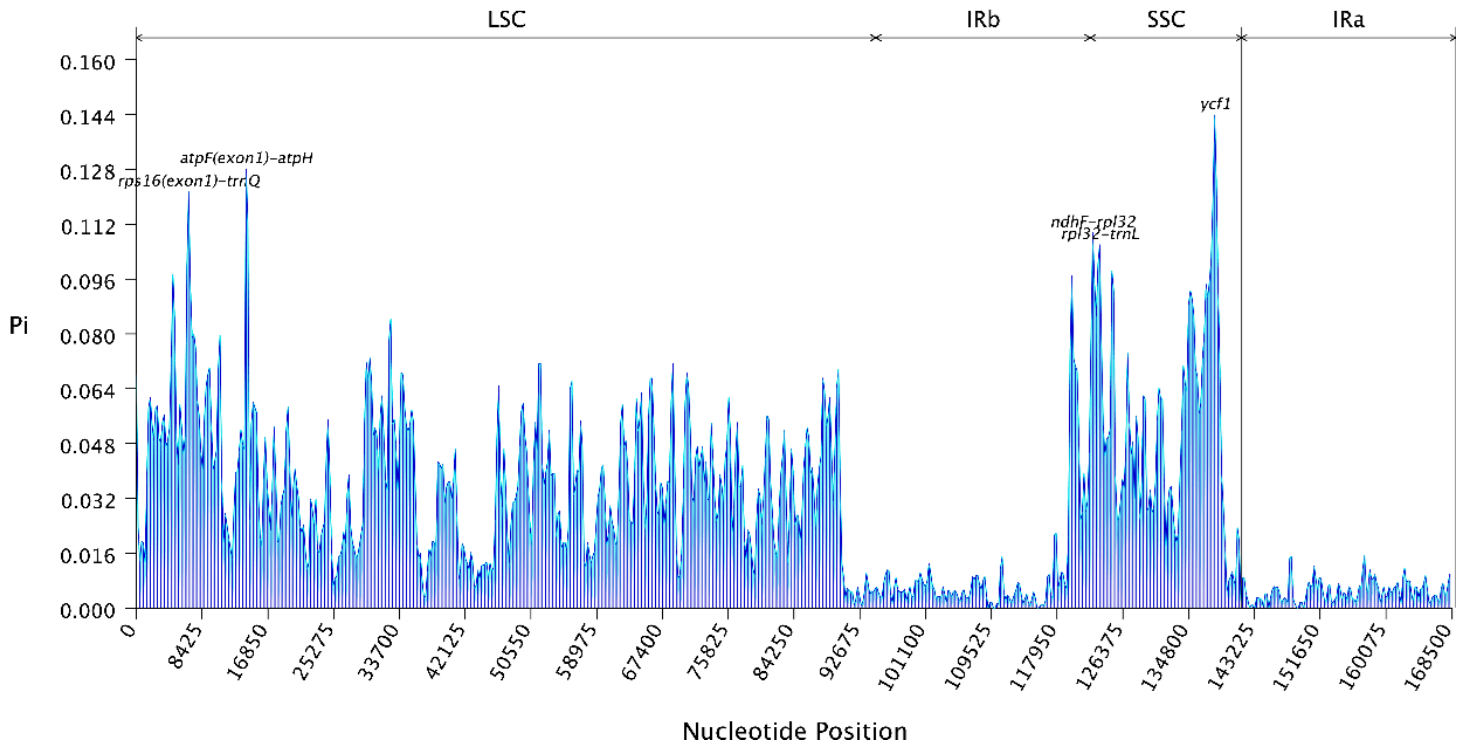

Figure 7

Nucleotide Variety across the cp genomes of the six Anacardiaceae species. X-axis, position of the midpoint of a window; Y-axis, nucleotide diversity of each window.

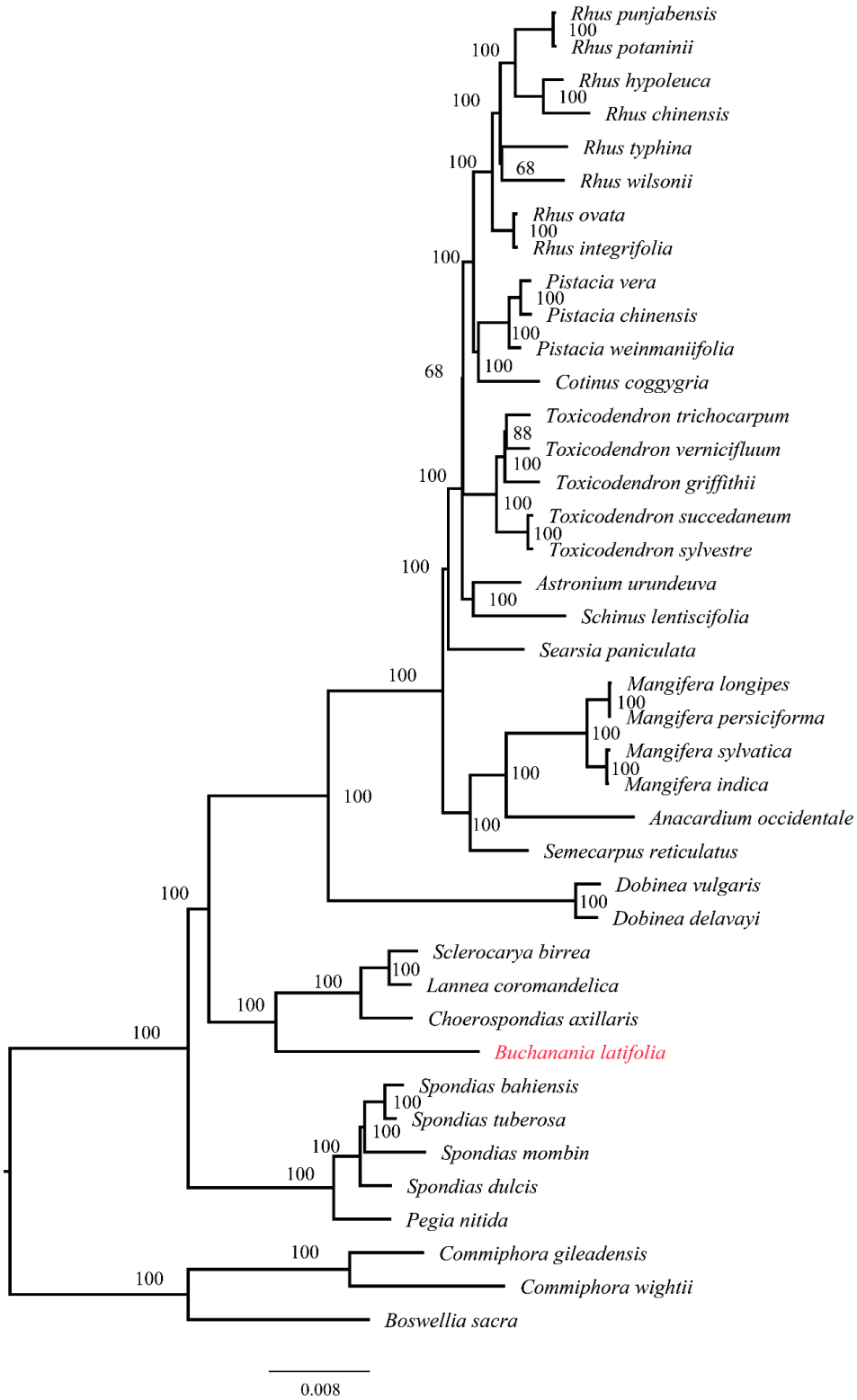

Figure 8

Phylogenetic tree reconstruction of Anacardiaceae species based on maximum-likelihood (ML) analysis using the program IQ-Tree v.6.1 (Nguyen et al., 2015) with 1,000 bootstrap replications in 37 complete chloroplast genome sequences. Using three species from the Burseraceae family as outgroups.