Comparative Chloroplast Genomics in Phyllanthaceae Species

Rehman, Umar

Multidisciplinary Digital Publishing Institute
2021-08-26

Rehman, U.; Sultana, N.; Abdullah; Jamal, A.; Muzaffar, M.; Poczai, P. Comparative Chloroplast Genomics in Phyllanthaceae Species. Diversity 2021, 13, 403.

http://hdl.handle.net/10138/349080

diversity

MDPI

Article

Comparative Chloroplast Genomics in Phyllanthaceae Species

Umar Rehman [1], Nighat Sultana [1,*], Abdullah [2], Abbas Jamal [3], Maryam Muzaffar [2,4] and Peter Poczai [5,6,*]

[1] Department of Biochemistry, Hazara University, Mansehra P.O. Box 21300, Pakistan; umarrehman07@hotmail.com
[2] Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan; abd.ullah@bs.qau.edu.pk (A.); mmaryam@bs.qau.edu.pk (M.M.)
[3] Key Laboratory of Horticulture Plant Biology (Ministry of Education), College of Horticulture and Forestry Sciences, Huazhong Agriculture University, Wuhan 430070, China; abbasjamal1@yahoo.com
[4] Alpha Genomics Private Limited, Islamabad 45710, Pakistan
[5] Finnish Museum of Natural History, University of Helsinki, P.O. Box 7, FI-00014 Helsinki, Finland
[6] Faculty of Biological and Environmental Sciences, University of Helsinki, P.O. Box 65, FI-00065 Helsinki, Finland
* Correspondence: nighat.sultana@hu.edu.pk (N.S.); peter.poczai@helsinki.fi (P.P.)

Abstract: Family Phyllanthaceae belongs to the eudicot order Malpighiales, and its species are herbs, shrubs, and trees that are mostly distributed in tropical regions. Here, we elucidate the molecular evolution of the chloroplast genome in Phyllanthaceae and identify the polymorphic loci for phylogenetic inference. We de novo assembled the chloroplast genomes of three Phyllanthaceae species, i.e., *Phyllanthus emblica*, *Flueggea virosa*, and *Leptopus cordifolius*, and compared them with six other previously reported genomes. All species comprised two inverted repeat regions (size range 23,921–27,128 bp) that separated large single-copy (83,627–89,932 bp) and small single-copy (17,424–19,441 bp) regions. Chloroplast genomes contained 111–112 unique genes, including 77–78 protein-coding, 30 tRNAs, and 4 rRNAs. The deletion/pseudogenization of *rps*16 genes was found in only two species. High variability was seen in the number of oligonucleotide repeats, while guanine-cytosine contents, codon usage, amino acid frequency, simple sequence repeats, synonymous and non-synonymous substitutions, and transition and transversion substitutions were similar. The transition substitutions were higher in coding sequences than in non-coding sequences. Phylogenetic analysis revealed the polyphyletic nature of the genus *Phyllanthus*. The polymorphic protein-coding genes, including *rpl*22, *ycf*1, *mat*K, *ndh*F, and *rps*15, were also determined, which may be helpful for reconstructing the high-resolution phylogenetic tree of the family Phyllanthaceae. Overall, the study provides insight into the chloroplast genome evolution in Phyllanthaceae.

Keywords: chloroplast genome; Phyllanthaceae; *Leptopus*; *Phyllanthus*; polymorphic loci; polyphyletic relationships

check for updates

Citation: Rehman, U.; Sultana, N.; Abdullah; Jamal, A.; Muzaffar, M.; Poczai, P. Comparative Chloroplast Genomics in Phyllanthaceae Species. *Diversity* **2021**, *13*, 403. https://doi.org/10.3390/d13090403

Academic Editors: Michael Wink and Mario A. Pagnotta

Received: 7 July 2021
Accepted: 24 August 2021
Published: 26 August 2021

1. Introduction

The family Phyllanthaceae and its closely related sister family Picorodendraceae comprise a separate clade of phyllantoid taxa within the order Malpighiales [1,2]. The species of the family Phyllanthaceae are predominantly tropical in distribution and include herbs, shrubs, and trees [3,4]. This family consists of two subfamilies (Phyllanthoideae and Antidesmatoideae), 10 tribes, 57 genera, and 2050 species [5]. Certain taxonomic discrepancies exist within Phyllanthaceae, i.e., *Phyllanthus* is polyphyletic since species of genera *Breynia*, *Glochidion*, *Sauropus*, and *Synostemon* were found embedded in its phylogeny [3,6,7]. Moreover, taxonomic discrepancies exist at intra-genus level since several subgenera and sections of *Phyllanthus* were paraphyletic [6,8]. Hence, researchers have suggested either merging all embedded genera into *Phyllanthus* L. to create a giant genus or dividing Phyllanthaceae into several monophyletic genera for grouping morphological similar species [3,6,8]. *Phyllanthus* L. is considered to be the largest genera of the family

Phyllanthaceae having 900 species [8,9], distributed both tropically and sub-tropically [10]. *Phyllanthus emblica* L. is commonly used as medicine in Asia [11] for its antioxidant, anti-cancer, anti-inflammatory, anti-pyretic, diuretic, laxative, stomachic, cardioprotective, and hepatoprotective properties [9,10,12–15]. The other genus, *Leptopus* Decne., is made up of nine species consisting of herbs and shrubs, which are distributed in Asia, America, and Europe and grow in open fields and in forests [16]. One of the species of this genus, *Leptopus cordifolius*, grows in the hilly areas of Pakistan and is also distributed from North India and Nepal to West Himalaya [16].

The chloroplast is an important organelle that plays a vital role in photosynthesis, also possessing a chloroplast genome. The chloroplast genome inherits maternally [17] in most angiosperms and paternally in some gymnosperms [18]. The chloroplast genome has a quadripartite structure and consists of a large single-copy (LSC) region, a small single-copy (SSC) region, and inverted repeat regions (IRa and IRb) [19–21]. The chloroplast genome normally ranges from 107 to 218 kb and contains 120 to 130 genes [22]. Many mutational events are reported in the chloroplast genome, including substitutions, insertion-deletions (InDels), repeats, and inversions [23–25]. The deletion of certain genes from the chloroplast genome or their transfer to nuclear genomes has also been reported in several plant lineages, including the species of the order Malpighiales to which the family Phyllanthaceae belongs [20,26,27]. The chloroplast genome evolves slowly and lacks the meiotic recombination of the nuclear genome, where homologous chromosomes exchange segments [28]. These properties make the chloroplast genome a suitable molecule for studies ranging from population genetics [29,30] to deep divergence at the genus and family levels [31–34], including plant evolution and phylogeny, providing deep insight into the evolution of plants [35,36].

In this study, the chloroplast genome of three species was de novo assembled and compared with previously reported chloroplast genomes of six other species of the family Phyllanthaceae. This provided us with deep insight into features of chloroplast genomes of selected species and similarities and differences among the species, elucidating the molecular evolution of the chloroplast genome in the family Phyllanthaceae. Moreover, suitable polymorphic loci were identified, which may be helpful for future phylogenetic inference to resolve taxonomic discrepancies of the family Phyllanthaceae.

2. Materials and Methods

2.1. Plant Collection and DNA Extraction

The leaves of *Phyllanthus emblica* and *Leptopus cordifolius* without apparent disease symptoms were collected during September 2018 from the district Mansehra, Khyber Pakhtunkhwa, Pakistan. Coordinates of collection points were also recorded with the Global Positioning System (GPS) for *Phyllanthus emblica* (34°31′50″ N, 72°88′10″ E) and *Leptopus cordifolius* (34°40′84″ N, 73°31′62″ E). The leaves were washed with distilled water and ethanol before drying with silica. Whole genomic DNA was extracted from the silica dried leaves using the DNeasy Plant Mini Kit manufactured by Qiagen. After confirmation of the quality and quantity on 1% agarose gel and by Thermo Scientific Nanodrop spectrophotometer, DNA was sent for sequencing to Novogene, Hong Kong.

2.2. Sequencing, De Novo Assembly, and Annotation of Chloroplast Genome

The DNA of *Phyllanthus emblica* L. and *Leptopus cordifolius* Decne. was sequenced with HiSeq 2500 using paired-end run with 150 bp short read size and 350 bp insert size. We also retrieved the 1.2 Gb raw reads of *Flueggea virosa* (Roxb. Ex Willd.) Royle (SRR7121487) [37] from the Sequence Read Archive (SRA) of the National Center for Biotechnology, which were sequenced by BGI-seq with 100 bp short reads. The whole genome shotgun was used for the de novo assembly of chloroplast genomes of selected three species using NOVOPlasty. The analysis provided the complete chloroplast genomes of all three species, including LSC regions, IR regions, and SSC region. The accuracy of assembled genomes and their coverage depth were confirmed by mapping short reads to their de novo assembled chloroplast genome using Burrows–Wheeler Aligner (BWA) [38] and visualized in Tablet [39].

De novo assembled chloroplast genomes were annotated using GeSeq [40], whereas the tRNA genes were confirmed by tRNAscan-SE v.2 [41] and ARAGORN v.1.2.3.8 [42] with default parameters. The annotations of protein-coding genes were further confirmed against homologous genes by blast search of NCBI. The five column delimited table of annotations was generated using GB2sequin [43] to submit de novo assembled genomes to GenBank of NCBI along with genomic features.

2.3. Analysis of Chloroplast Genome Features and Inverted Repeat Contraction and Expansion

The de novo assembled chloroplast genomes of *Phyllanthus emblica* (Pakistan), *Leptopus cordifolius*, and *Flueggea virosa* were compared with six previously reported chloroplast genomes of *Baccaurea ramiflora* Lour., *Phyllanthus amarus* Schumach. & Thonn., *Phyllanthus emblica* L. (China), *Glochidion chodoense* C.S. Lee & H.T.Im., *Breynia fruticosa* (L.) Müll.Arg, and *Sauropus spatulifolius* Beille. The lengths of complete chloroplast genome, LSC, SSC, and IRs were compared among the nine species of Phyllanthaceae. Moreover, gene contents, intron contents, and GC contents of all species were compared using Geneious R8.1 [44]. The gene arrangement among the selected species of Phyllanthaceae was determined using Mauve alignment [45] integrated in Geneious R8.1. The contraction and expansion of inverted repeat regions were visualized using IRscope [46].

2.4. Analysis of Codon Usage, Amino Acid Frequency, and Repeats

The relative synonymous codon usage (RSCU) and amino acid frequency of each species were determined using Geneious R8.1 [44]. The RSCU value of each species was shown with the help of a heatmap using TBtool [47]. MIcroSAtellite (MISA) [48] was used for the determination of simple sequence repeats (SSRs), with a maximum threshold of 10 for mononucleotide SSRs, 5 for dinucleotide SSRs, 4 for trinucleotide SSRS, and 3 for tetra-, penta-, and hexanucleotide SSRs. REPuter [49] was used for the determination of four types of oligonucleotide repeats, including forward (F), complementary (C), reverse (R), and palindromic (P). Parameters such as repeats size $\geq$ 30 with at least 90% similarities were adjusted.

2.5. Synonymous, Non-Synonymous, Transition, and Transversion Substitution Analyses

The synonymous (K_s) and non-synonymous (K_a) substitutions were analyzed in TBtool [47] for 77 protein-coding genes using *Baccaurea ramiflora* Lour. as a reference for all other species, as this species lies basal to the current species following Abdullah et al. [19]. Each gene selection was predicted by considering the ratios of Ka/Ks < 1 purifying selection, Ka/Ks = 1 neutral selection, and Ka/Ks > 1 positive selection [50]. The positive selection on protein coding genes or their codons was further analyzed using the Mixed Effects Model of Evolution (MEME) [51] and Fast Unconstrained Bayesian Approximation (FUBAR) [52] as implemented in the DATAMONKEY web server [53], according to previous parameters [54]. The transition and transversion substitutions of complete genome and protein-coding genes were determined by comparison with closely as well as with distantly related species. The whole genome was aligned using MAFFT, whereas the protein-coding sequences of each species were concatenated, except for *ycf*1, and also aligned using MAFFT alignment [55]. For closely related species, the genome of *P. emblica* (Pakistan) was used as a reference for *Breynia fruticosa*, *Phyllanthus amarus*, *Phyllanthus emblica* (China), *Glochidion chodoense*, and *Sauropus spatulifolius*. Similarly, for distantly related species, *Baccaurea ramiflora* was used as a reference for the selected species of *Leptopus cordifolius*, *Flueggea virosa*, and *Phyllanthus emblica* (Pakistan).

2.6. Polymorphism of Protein-Coding Genes

To determine the extent of polymorphism of protein-coding genes for further phylogenetic studies, all protein-coding genes of each species were extracted. The sequence of each gene was multiple aligned using Geneious R8.1 and analyzed in DnaSP v.6 [56].

2.7. Reconstruction of Phylogenetic Tree

For inferring phylogeny, the sequences of 76 protein-coding genes, except *ycf*1, of each species were extracted and concatenated in Geneious R8.1 [44]. The *ycf*1 was excluded from the analysis as this gene exists at the junction of IR/SSC, with one part of *ycf*1 located in the IR region and the other part in the SSC region. The part situated in the IR region showed a different rate of evolution than the part in the SSC region due to rate heterotachy [57–59]. These concatenated sequences of all 10 species, including *Linum usitatissimum* L. as outgroup from the family Linaceae, were multiple aligned using MAFFT [60] extension in Geneious R8.1. The maximum likelihood tree was reconstructed with best fit model GTR + F + G4 and determined by jModelTest 2 [61] using IQ-tree [62] with the same parameters as described previously [54]. The phylogenetic tree was also reconstructed with the identified polymorphic loci following the same approach. However, the best fit model TVM + F + G4 was predicted by jModelTest 2 and applied. The Interactive Tree of Life [63] was used to improve the presentation of the tree.

3. Results

3.1. Chloroplast Genome Features and Organization

HiSeq 2500 produced about 10.52 GB data with 31.4 million reads for *Phyllanthus emblica* and 7.58 GB data with 22.6 million reads for *Leptopus cordifolius*. Chloroplast genomes assembled from the data exhibited a high average coverage depth of 855X for *Phyllanthus emblica* and 375X for *Leptopus cordifolius*. The *Flueggea virosa* which was assembled from the SRA data of NCBI showed an average coverage depth of 138X. These three de novo assembled genomes along with the previously reported genome provided an opportunity to perform in-depth comparative chloroplast genomics of members of the family Phyllanthaceae.

The gene contents and organization of the chloroplast genomes of the family Phyllanthaceae are provided in Tables 1 and 2 and Figure 1. The genomes showed high similarity in gene contents, except *rps*16, which was missing in *Baccaurea ramiflora* and *Leptopus cordifolius*. The chloroplast genomes of all selected species displayed quadripartite structure, comprising IRs (IRa and IRb) that separate the LSC and SSC regions. The length of the complete chloroplast genome ranged from 154,707 (*Sauropus spatulifolius*) to 161,093 bp (*Baccaurea ramiflora*), IRs from 23,921 (*Sauropus spatulifolius*) to 27,128 bp (*Phyllanthus amarus*), LSC from 83,627 (*Leptopus cordifolius*) to 89,932 bp (*Phyllanthus emblica* from Pakistan), and SSC from 17,424 (*Leptopus cordifolius*) to 19,441 bp (*Breynia fruticosa*).

Chloroplast genomes possessed 111–112 unique genes that showed high similarities in the arrangement within the genomes and no inversion events related to gene rearrangement were found, as predicted by Mauve alignment (Figure 2). Among these 111–112 genes, 77 –78 were protein-coding genes, 30 transfer RNA (tRNA) genes, and 4 ribosomal RNA (rRNA) genes (Table 1). Moreover, 16–17 genes were also duplicated in IR regions, including 5–6 protein-coding genes, 7 tRNA genes, and 4 rRNA genes. Here, two pseudogenes of *ycf*1$^{\Psi}$ and *rps*19$^{\Psi}$ were excluded, which exist at the junctions of LSC/IRa and SSC/IRa (Table 1). The *rps12* gene was a trans-splicing gene; the 5′ end was present in LSC in the form of a single copy, whereas the 3′ end was duplicated in IRa and IRb. The average GC content of the complete chloroplast genome was 36.6–36.8%. LSC was 34.3–34.6%, SSC 30.1–30.8%, and IR 42.3–43.3%. Maximum GC contents were found in IR regions (42.3–43.3%), followed by the LSC region (34.3–34.6%) and the SSC region (30.1–30.8%). The highest GC content of IRs belongs to rRNAs (55.5%) and tRNAs (53%); the genomic features are given in a detailed summary in Table 2.

Table 1. Gene content and functional classification of the chloroplast genomes of the family Phyllanthaceae.

Category for Gene	Group of Gene	Name of Gene					Number
Photosynthesis-related genes	Photosystem I	*psa*A	*psa*B	*psa*C	*psa*I	*psa*J	5
	Photosystem II	*psb*A	*psb*K	*psb*I	*psb*M	*psb*D	15
		*psb*F	*psb*C	*psb*H	*psb*J	*psb*L	
		*psb*E	*psb*N	*psb*B	*psb*T	*psb*Z	
	Cytochrome b/f complex	*pet*N	*pet*A	*pet*L	*pet*G	*pet*D *	6
		*pet*B *					
	ATP synthase	*atp*I	*atp*H	*atp*A	*atp*F	*atp*E	6
		*atp*B					
	Cytochrome c-type synthesis	*ccs*A					1
	Assembly/stability of photosystem I	*ycf*3 **	*ycf*4				2
	NADPH dehydrogenase	*ndh*B [*,a]	*ndh*H	*ndh*A *	*ndh*I	*ndh*G	12
		*ndh*J	*ndh*E	*ndh*F	*ndh*C	*ndh*K	
		*ndh*D					
	Rubisco	*rbc*L					1
Transcription and translation related genes RNA genes	Transcription Small subunit of ribosome	*rpo*A	*rpo*C2	*rpo*C1 *	*rpo*B	*rps*16 [*,£Ψ]	5
		*rps*7 [a]	*rps*15	*rps*19 [$]	*rps*3	*rps*8	14
		*rps*14	*rps*11	*rps*12 [a,*]	*rps*18	*rps*4	
		*rps*2	*rps*19 [Ψ]				
	Large subunit of ribosome	*rpl*2 [a,*,$]	*rpl*23 [a]	*rpl*32	*rpl*22	*rpl*14	11
		*rpl*33	*rpl*36	*rpl*20	*rpl*16 *		
	Ribosomal RNA	*rrn*16 [a]	*rrn*4.5 [a]	*rrn*5 [a]	*rrn*23 [a]		8
	Transfer RNA	*trn*V-GAC [a]	*trnI*-CAU [a]	*trnA*-UGC [a,*]	*trn*N-GUU [a]	*trn*P-UGG	37
		*trn*W-CCA	*trn*V-UAC *	*trn*L-UAA *	*trn*F-GAA	*trn*R-ACG [a]	
		*trn*T-UGU	*trn*G-UCC *	*trn*T-GGU	*trn*R-UCU	*trn*E-UUC	
		*trn*Y-GUA	*trn*D-GUC	*trn*C-GCA	*trn*S-GCU	*trn*H-GUG	
		*trn*K-UUU *	*trn*Q-UUG	*trnf*M-CAU	*trn*G-GCC	*trn*S-UGA	
		*trn*S-GGA	*trn*L-UAG	*trn*M-CAU	*trn*L-CAA [a]	*trn*I-GAU [*,a]	
Other genes	RNA processing	*mat*K					1
	Carbon metabolism	*cem*A					1
	Fatty acid synthesis	*acc*D					1
	Proteolysis	*clp*P **					1
	Component of TIC complex	*ycf*1	*ycf*1 [Ψ]				2
	Hypothetical proteins	*ycf*2 [a]					2
Total							131

* Gene with one intron, ** gene with two introns, [a] gene with two copies, [£Ψ] gene pseudo in *Baccaurea ramiflora* and *Leptopus cordifolius*, [$] single copy existed of *rps*19 in *Glochidion chodoense* and of *rpl*2 in *Sauropus spatulifolius*, [Ψ] pseudo-copy also existed along with a functional copy.

Table 2. Comparison of chloroplast genomes of all nine species of the family Phyllanthaceae.

Characteristic		***Baccaurea ramiflora***	***Breynia fruticosa***	***Flueggea virosa***	***Glochidion chodoense***	***Leptopus cordifolius***	***Phyllanthus amarus***	***Phyllanthus emblica* (Pak)**	***Phyllanthus emblica* (China)**	***Sauropus spatulifolius***
Size (base pair; bp)		161,093	155,630	158,075	157,085	155,027	157,673	156,477	156,208	154,707
LSC length (bp)		89,503	85,065	87,604	85,304	83,627	85,855	89,932	85,674	87,438
SSC length (bp)		18,818	19,441	19,303	17,635	17,424	17,564	19,293	19,310	19,427
IR length (bp)		26,386	25,562	25,584	27,073	26,988	27,128	25,611	25,612	23,921
Number of unique genes		111 (128)	112 (129)	112 (129)	112 (129)	111 (128)	112 (129)	112 (129)	112 (129)	112 (128)
Protein-coding genes		77 (83)	78 (84)	78 (84)	78 (83)	77 (83)	78 (84)	78 (84)	78 (84)	78 (83)
tRNA genes		30 (37)	30 (37)	30 (37)	30 (37)	30 (37)	30 (37)	30 (37)	30 (37)	30 (37)
rRNA genes		4 (8)	4 (8)	4 (8)	4 (8)	4 (8)	4 (8)	4 (8)	4 (8)	4 (8)
Duplicate genes		19 [a]	19 [a]	19 [a]	18 [a]	19 [a]	18 [a]	19 [a]	19 [a]	18 [a]
GC content	Total (%)	36.7	36.7	36.6	36.7	36.8	36.6	36.7	36.8	36.6
	LSC (%)	34.4	34.5	34.3	34.4	34.6	34.4	34.4	34.5	34.4
	SSC (%)	30.8	30.2	30.4	30.2	30.1	30.2	30.2	30.2	30.1
	IR (%)	42.7	43	43	42.3	42.3	36.6	43.1	43.1	43.2
	CDS (%)	37.8	37.4	37.5	37.4	37.3	37.4	37.4	37.4	37.3
	rRNA (%)	55.5	55.4	55.4	55.4	55.4	55.3	55.5	55.5	55.4
	tRNA (%)	53.4	53.4	53.3	53.2	53.2	53.1	53	53	53.3
	Non-coding regions (%)	32.5	32.4	32.4	32.6	32.9	32.4	32.6	32.6	32.4
Accession number		MT900598	MT863745	BK059210 **	MK056235	MZ424188 *	MN736962	MN122078 *	MN711725	MT089915

[a] Duplicating genes in IRs including pseudogenes, information in parentheses indicates the total number of functional genes, * represents sequenced and assembled in the current study, ** represents assembled from raw reads of NCBI.

There were 17–18 intron-containing genes of which 11–12 were protein-coding genes and 6 were tRNA genes. Moreover, an intron was found in *atp*F gene in only *Baccaurea ramiflora*, which was missing in all other species examined. All protein-coding genes contained one intron, except *clp*P and *ycf*3, which had two introns as shown in Table 1. Of these intron-containing genes, three were protein-coding genes and two tRNA genes were also duplicated in IR regions.

3.2. Phylogenetic Analysis

Phylogenetic analysis of the selected species of the family Phyllanthaceae was performed based on 76 protein-coding genes and suitable polymorphic loci. Both methods provided similar results and resolved the phylogenetic relationship within the family Phyllanthaceae with high bootstrapping support of 100. The genus *Phyllanthus* was found to be polyphyletic, with *P. amarus* closely related to *G. chodoense* instead of *P. emblica* (Figure 3). The nodes "*P. amarus* and *G. chodoense*" and "*B. fruticosa* and *S. spatulifolius*" were sisters to one another and were rooted by *P. emblica*. This whole clade was rooted by *F. virosa*, followed by *L. cordifolius* and then by *B. ramiflora*, which lies basal to the selected species in the tree.

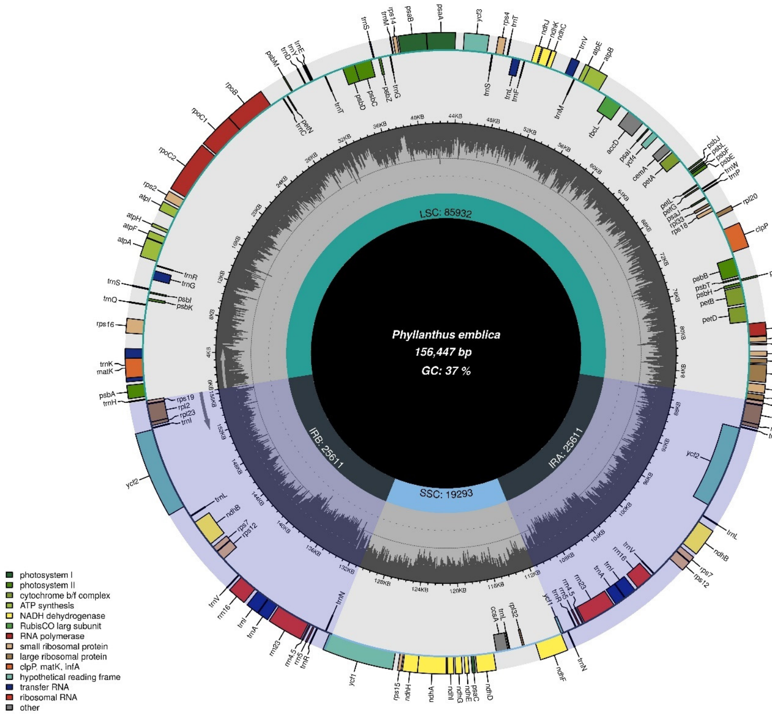

Figure 1. Circular representation of *Phyllanthus emblica* chloroplast genome (a representative of all other species). The genes transcribed clockwise are shown inside the circle, whereas genes transcribed anti-clockwise are shown outside the circle.

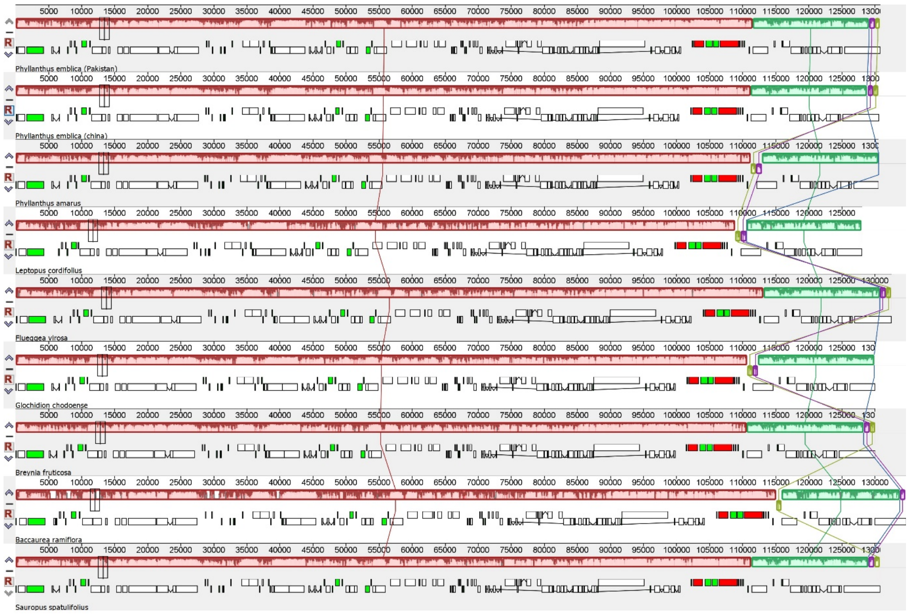

Figure 2. Mauve alignment of nine species of the family Phyllanthaceae shows high similarities in gene arrangement and gene contents. The small blocks represent genes: red = rRNA; green = tRNA with intron; black = tRNA without intron; white = protein-coding genes. The large reddish color represents LSC and IR regions, while the greenish part represents SSC regions. The small purple and green blocks represent part of the *ycf*1^{Ψ} gene, e.g., in *Leptopus cordifolius* the size of *ycf*1^{Ψ} was 1032 bp and both blocks are present at the start of the large greenish block, while in *Flueggea virosa* the size of *ycf*1^{Ψ} was 220 bp and both blocks are present at the end of the large greenish block.

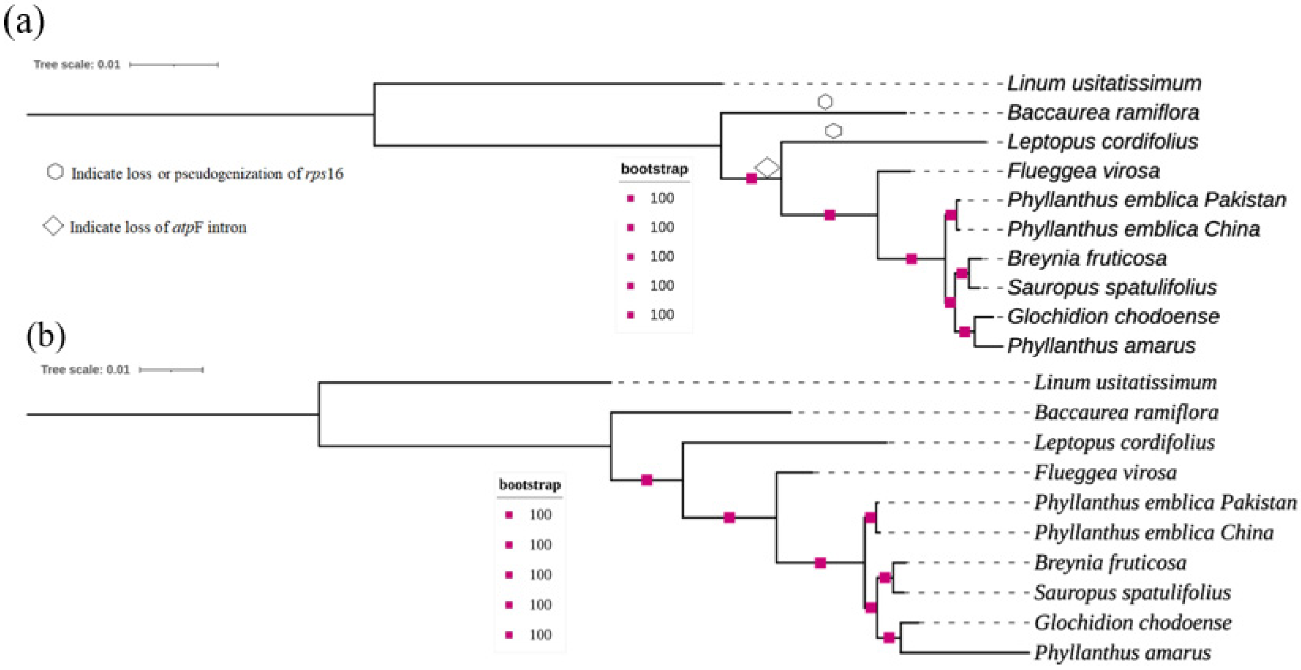

Figure 3. Maximum likelihood phylogenetic tree reconstructed with a dataset of 76 protein-coding genes (**a**) and identified polymorphic loci (**b**). The 100 bootstrapping support was observed for all nodes. The deletion/pseudogenization of *rps*16 and loss of *atp*F intron are indicated by the symbols only in the tree constructed with coding sequences to avoid duplication of the same data.

3.3. Inverted Repeat Contraction and Expansion

Comparative analysis of the junctions of LSC/IR and SSC/IR showed similarities among species of the family Phyllanthaceae, except for a few differences that mostly occurred at the junction of LSC/IRb (JLB) (Figure 4). At JLB, the *rps*19 started from the IRb and entered the LSC regions in six species. Hence, truncated *rps*19 pseudogene remained at JLA (IRa/LSC) in these species (Figure 4). In the remaining three species, the *rps*19 completely existed in the LSC regions at the JLB junction and did not lead to the generation of a pseudogene at the JLA junction. The *rpl*2 gene was completely found in the IR regions away from the JLB and JLA junctions, except in *Sauropus spatulifolius,* where the *rpl*2 started from IRb and entered the LSC regions, retaining a pseudogene at the junction of JLA instead of a functional copy due to contraction of IRs. Moreover, at JLA the *rpl*23 was present in IRb and IRa instead of *rpl*2. At the SSC/IRb junction (JSB), *ndh*F was situated entirely in the SSC region in the six species of *Baccaurea ramiflora*, *Flueggea virosa*, *Phyllanthus emblica* (China), *Phyllanthus emblica* (Pak), *Breynia fruticosa*, and *Sauropus spatulifolius,* while it was integrated into the IRb regions within the remaining three species of *Phyllanthus amarus*, *Glochidion chodoense,* and *Leptopus cordifolius,* where it overlapped with the pseudogene of *ycf*1. The *ycf*1 started in IRa and ended in SSC, as seen at the JSA (SSC/IRa) junction. Consequently, pseudogene of *ycf*1 remained at JSB (IRb/SSC), which ranges in size from 195 to 1921 bp. At the junction of JLA (IRa/LSC), the *trn*H gene was found in all species (Figure 4). Thus, IR contraction and expansion were not responsible for complete duplication of a functional copy of a gene, but they led to the generation of truncated *rps*19, *rpl*2, and *ycf*1 pseudogenes.

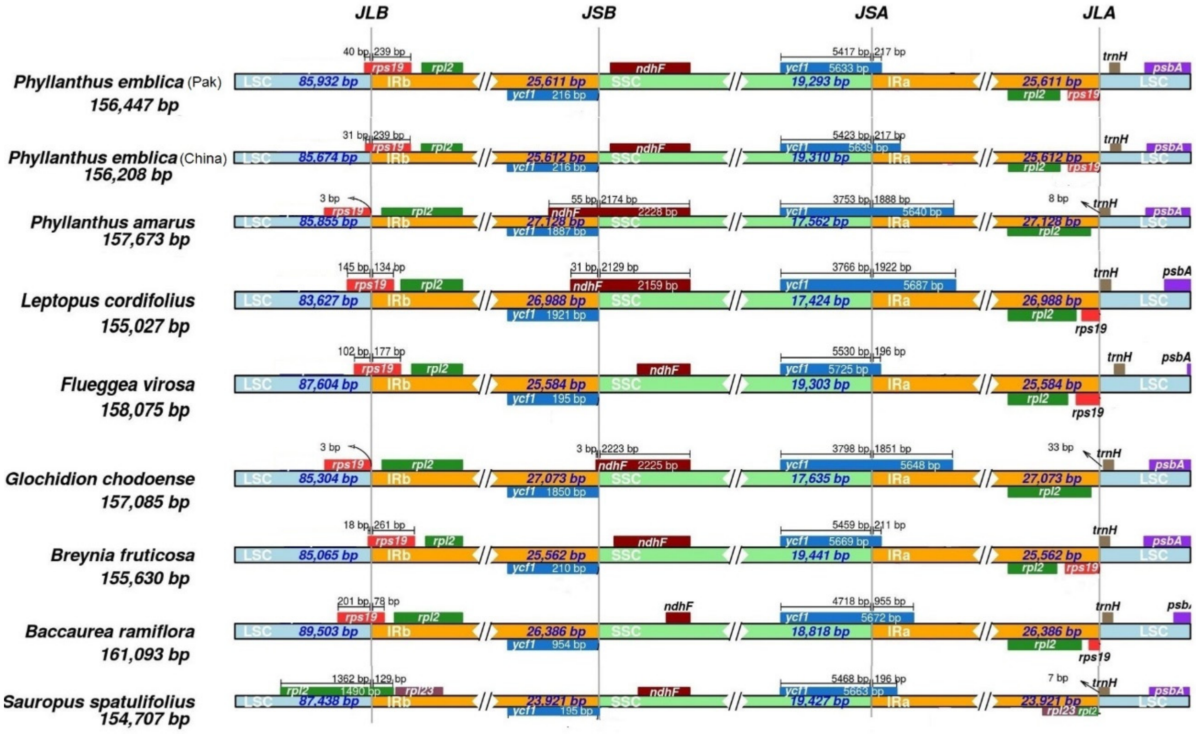

Figure 4. Contraction and expansion of inverted repeats at the junction of chloroplast genome. JLB: LSC/IRb; JSB: IRb/SSC; JSA: SSC/IRa; JLA: IRa/LSC. Arrows illustrate the distance of genes from the junction site as shown for *rps*19 at JLB and for *trn*H at JLA. The scale bar above some genes illustrates the number of base pairs that each gene occupies in specific regions of the chloroplast, e.g., the scale bar above *ycf*1 represents the part of the gene located in the IR region and the SSC region.

3.4. Codon Usage and Amino Acid Frequency, and Repeats

Codon usage analysis was interpreted in terms of RSCU values. The analysis showed that most of the amino acids were encoded from codons ended with A/T at 3′ (having RSCU > 1) instead of C/G (having RSCU < 1) (Figure S1). The amino acid frequency analysis revealed that leucine was the most encoded amino acid while cysteine was the rarest (Figure S2). Codon usage analysis and amino acid frequency showed high similarities in all species of the family Phyllanthaceae. The simple sequence repeats (SSRs) and oligonucleotide repeats were analyzed in the Cp genome of selected species. The analysis of MISA revealed 630 SSRs mostly consist of A/T motifs in all species (Table S1). All six types of SSRs were based on motif types, including mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide. However, SSRs were predominantly mononucleotide, followed by dinucleotide (Figure S3a). The highest number of SSRs was found in the LSC region, followed by the SSC region and the IR (Figure S3b). The highest number of SSRs (84) was observed in *Phyllanthus emblica* (China), and the lowest number of SSRs (51) was seen in *Glochidion chodoense*. The REPuter program detected 311 oligonucleotide repeats with sizes of 30–96 bp. The number of repeats varied from 24 (*Glochidion chodoense*) to 49 (*Baccaurea ramiflora*). Most of the oligonucleotide repeats were forward and palindromic, but reverse and complementary repeats were also found (Figure S3c). Most of the repeats ranged in size from 30 to 34 bp (Figure S3d). The number of repeats was higher in LSC than in SSC and IR, whereas some repeats were also shared between LSC/IR, LSC/SSC, and SSC/IR regions of the chloroplast genome (Figure S3e).

3.5. Synonymous and Non-Synonymous Substitutions

The synonymous (Ks) and non-synonymous (Ka) substitutions and their ratio (Ka/Ks) revealed high similarities among the species of Phyllanthaceae (Table S2). The lowest average values of Ka = 0.0298, Ks = 0.1866, and Ka/Ks = 0.1597 were recorded, which showed that high purifying selection pressure acted on the protein-coding genes of species of the family Phyllanthaceae. Two genes—*psb*L and *pet*L—showed a signature of positive selection in *Sauropus spatulifolius* and *Flueggea virosa*, respectively, based on the values of Ka/Ks. However, the results of FUBAR and MEME did not support signature of positive selection.

3.6. Transition and Transversion Substitutions

The transition and transversion substitutions within the complete chloroplast genome and in the protein-coding regions were analyzed. The comparative analysis of the complete chloroplast genomes among the distantly related species of Phyllanthaceae revealed the existence of 10,950–12,603 substitutions (Figure 5a), whereas comparison of closely related species revealed the presence of 257–2077 substitutions (Figure 6a). Furthermore, 3237–3747 substitutions were observed in protein-coding sequences of distantly related species (Figure 5b) and 65–685 substitutions in closely related species (Figure 6b). The ratio of transition and transversion substitutions (Ts/Tv) also revealed variations in both distantly related and closely related species. The Ts/Tv ratio for distantly related species varied from 1.34 to 1.40 (Figure 5c), whereas for closely related species it varied from 0.77 to 1.0 (Figure 6c). The analysis revealed that the Ts/Tv ratio was higher in protein-coding sequences than in the complete chloroplast genome. The Ts/Tv ratio for distantly related species was 2.09–2.34 (Figure 5c) and for closely related species 1.42–1.71 (Figure 6c). The higher Ts/Tv ratio in protein-coding regions showed the occurrence of higher transition substitutions than transversion substitutions. The lower Ts/Tv ratio in the complete genome than in coding regions was due to inclusion of an intergenic spacer region in which higher transversion substitutions take place than transition substitutions.

3.7. Intraspecies Variations in Chloroplast Genomes of Phyllanthus Emblica

The intraspecies variations in the chloroplast genomes of *Phyllanthus emblica* belonging to two different countries—Pakistan and China—with totally different climatic changes were determined. Slight differences emerged in genome size, LSC, and SSC, as shown in Table 2. The numbers of protein-coding genes, unique genes, and IR regions were the same in both species. The comparative analysis of both *P. emblica* chloroplast genomes revealed 257 SNPs in the whole chloroplast genome, 79 of which belong to coding regions. In total, 108 insertion-deletion (InDel) events with an average length of 5.32 in which 3 InDels belong to coding regions were observed. No InDels or SNPs in genes of transfer RNAs and ribosomal RNAs were seen.

3.8. Polymorphism of Protein-Coding Genes

The polymorphism of all protein-coding genes was assessed to facilitate identification of suitable polymorphic loci for future phylogenetic inference of the family Phyllanthaceae (Table S3). The 15 genes of ≥200 bp were selected. The missing data produced by the highly polymorphic regions due to InDels to provide information about the suitability of these loci were also assessed and are presented in Table 3. These loci included in the analysis were *rpl*22, *ycf*1, *mat*K, *ndh*F, and *rps*15, as shown in Table 3.

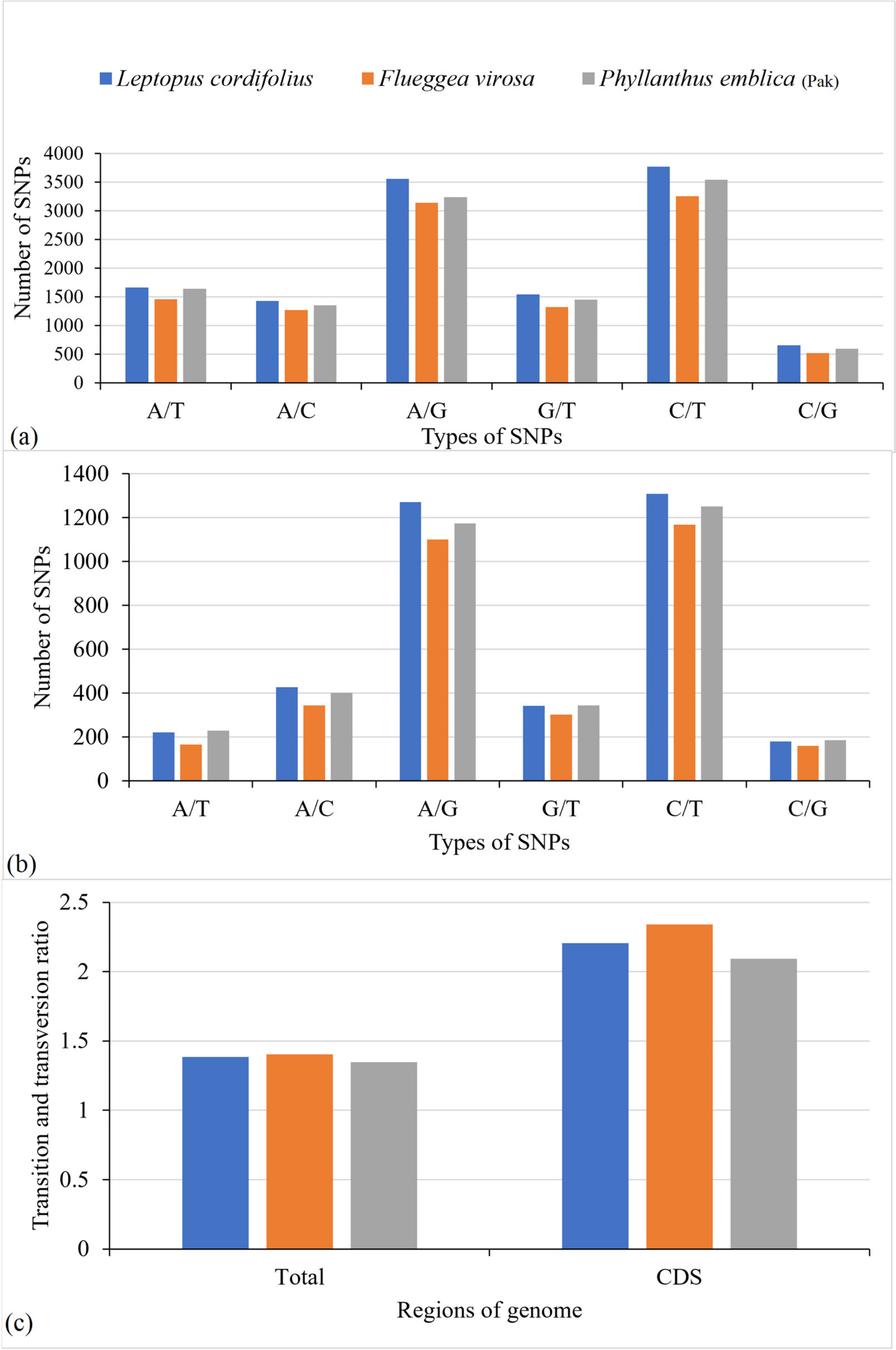

Figure 5. Comparison of transition and transversion substitutions in distantly related species. (**a**) The comparison of transition and transversion substitutions within the complete chloroplast genome. (**b**) Comparison of transition and transversion substitutions within protein-coding sequences. (**c**) Comparison of the ratio of transition and transversion substitutions in the complete chloroplast genome and in protein-coding sequences. *Baccaurea ramiflora* was used as a reference for all species.

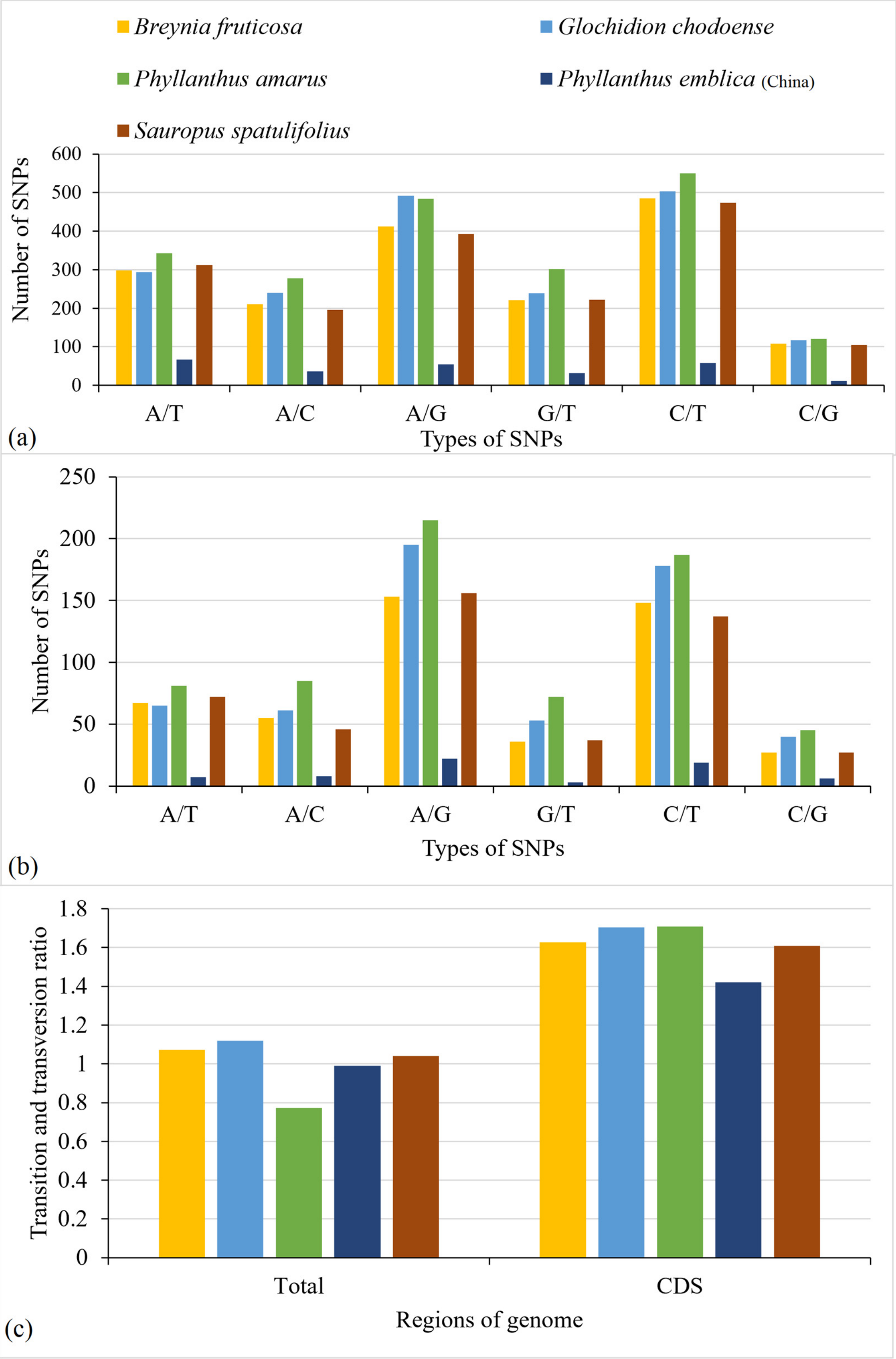

Figure 6. Comparison of transition and transversion substitutions in closely related species. (**a**) The comparison of transition and transversion substitutions within the complete chloroplast genome. (**b**) Comparison of transition and transversion substitutions within protein-coding sequences. (**c**) Comparison of the ratio of transition and transversion substitutions in the complete chloroplast genome and in protein-coding sequences. *Phyllanthus emblica* (Pakistan) was used as a reference for all species.

Table 3. Polymorphic protein-coding sequences.

Gene	Total Number of Mutations	Alignment Length	Alignment Length without InDels	Nucleotide Diversity	Missing Data
*rpl*22	113	513	372	0.10887	27.49
*ycf*1	1422	5865	5449	0.08388	7.09
*mat*K	342	1541	1521	0.06901	1.30
*ndh*F	417	2201	2105	0.06127	4.36
*rps*15	52	291	261	0.06117	10.31
*rpl*20	66	354	354	0.05841	0
*ccs*A	170	972	957	0.05717	1.54
*rps*3	105	687	645	0.05078	6.11
*rps*8	67	405	405	0.05018	0
*rpl*16	64	411	408	0.05007	0.73
*ndh*D	239	1524	1520	0.05007	0.26
*acc*D	224	1524	1458	0.04774	4.33
*cem*A	66	477	477	0.0456	0
*ycf*4	77	563	549	0.04417	2.49
*rps*11	58	417	417	0.04368	0

4. Discussion

4.1. Chloroplast Genome Assembly from Whole Genome Sequencing and Comparative Chloroplast Genomics

Although the angiosperm's chloroplast genome is believed to be conserved in both gene content and gene order, with conserved intronic regions in protein-coding genes and tRNA genes. However, the loss of some genes or introns were also reported [20,22,35,64]. Several studies have reported de novo assembled chloroplast genomes from whole genome shotgun reads [65,66]. In the present study, whole genome shotgun sequencing was employed to assemble the complete chloroplast genomes of *Phyllanthus emblica* (Pak), *Leptopus cordifolius*, and *Flueggea virosa* with high coverage depth.

We observed that the chloroplast genome features of the species belonging to the family Phyllanthaceae were highly similar to each other. The intron of *atp*F gene was deleted in all of the species, except *Baccaurea ramiflora*, which was present at a basal point in the current species. The *rps*16 gene was found to be non-functional in the two species of *Baccaurea ramiflora* (deletion of exon 1) and *Leptopus cordifolius* (gene loss with only few base pairs remaining). The deletion of the *rps*16 gene has previously been reported [27] in the species of the order Malpighiales and a functional copy was found for some species in nuclear genome but was not reported in the family Phyllanthaceae due to inclusion of *Glochidion chodoense* in the comparison containing a functional *rps*16 gene. Here, comparison of several species helped us to determine the deletion/pseudogenization of *rps*16 for the first time in the family Phyllanthaceae. The GC content of species of the family *Phyllanthaceae* is similar to other angiosperms and to other species of the order Malpighiales [20,67,68]. Another gene, infA, was completely absent in the chloroplast genome of the species of the family Phyllanthaceae. The deletion and pseudogenization of the infA gene were also reported in other families of angiosperms, including Araceae [57,66,69], Malvaceae [19], and Malpighiaceae [20]. Since the product of the *inf*A gene has a vital function as a translation initiation factor, it can be inferred that the functional copy of *inf*A has probably been transferred to the nuclear genome [35,70].

Previously, it was documented that IR's contraction and expansion can generate a pseudo-copy as well as a functional gene copy or may change duplicated genes to a single-copy gene due to transfer of IRs to single-copy regions (LSC or SSC) or may convert a single-copy gene to a duplicated gene when transferred from LSC or SSC to IRs [19,34,57,65]. In our study, the analyzed species were found to be conserved in terms of contraction and expansion of IRs. Likewise, the generation of a pseudo-copy of *ycf*1, *rpl*2, and *rps*19 remained similar to other angiosperms [71,72]. However, the duplication of complete functional genes, such as *ycf*1, *rps*15, and *rps*19, observed in other angiosperms was not detected here [19,57,65].

4.2. Simple Sequence Repeats and Oligonucleotide Repeat Analyses

The highest number of simple sequence repeats (SSRs) was observed in the LSC region as compared with lower numbers in SSC and IR regions. Mononucleotide SSRs (A/T) and dinucleotide SSRs (AT/TA) were more abundant than trinucleotide SSRs, consistent with previous studies [54,73–77]. However, a few studies have reported higher numbers of trinucleotides than of dinucleotides [19,68,78,79]. The SSR marker identified in the current study can serve as a resource for population genetics studies of species of the family Phyllanthaceae. REPuter detected different oligonucleotide repeats, including forward, reverse, complementary, and palindromic repeats in the chloroplast genomes. These repeats originate InDels and substitutions [80–82] and can be used as proxies for identification of mutational hotspot regions [69,81,83]. High variations were detected in the number of oligonucleotide repeats, but further analysis involving more species is needed to gain insight into the evolutionary events leading to loss/gain of repeats. Previously, correlations between the identified number of repeats and genome size and phylogenetic position of species have not been established [65].

4.3. Phylogenetics Analysis and Suitable Polymorphic Loci for Further Phylogenetic Inference

The phylogenetic analysis showed that the polyphyletic nature of the genus *Phyllanthus* was similar to the recently obtained results using nuclear (ITS, *PHY*C) and chloroplast (*mat*K, *acc*D-*psa*I, *trn*S-*trn*G) markers along with morphological data [6]. Hence, the author suggested the division of the genus *Phyllanthus* into nine small genera instead of combining the embedded genera to *Phyllanthus*. Our study was based on 76 protein-coding sequences, which supports the polyphyletic nature of the genus *Phyllanthus* with a limited number of species. Further sequencing will thus be required to gain broad insight into the phylogenetics of *Phyllanthus* and the tribe Phyllantheae. A similar phylogenetic position has been reported in other species earlier [4,7]. Certain discrepancies still exist in the phylogeny of the family Phyllanthaceae, which warrant further elaboration [3,4,7]. Here, we have identified 15 suitable highly polymorphic regions from protein-coding sequences based on the comparative analysis of chloroplast genomes. The phylogenetic analysis based on these loci provided similar results to the dataset of 76 genes, which suggests a high efficacy of these loci for phylogenetic inference. These will serve as suitable markers for quality phylogenetic inference of the family Phyllanthaceae, as lineage-specific molecular markers are more authentic, robust, and cost-effective [22,36,83–85].

In conclusion, our study provides insight into the molecular evolution of the chloroplast genome and sheds light on the deletion/pseudogenization of *rps*16 in the family Phyllanthaceae for the first time. However, sequencing and analysis of a broad number of species may elucidate the evolution and biological factors that have led to deletion of this gene. The phylogenetic analysis of our limited species showed the polyphyletic nature of the genus *Phyllanthus* but sequencing multiple species of this genus may be helpful to reconstruct a high-resolution phylogenetic tree and obtain insight into its systematics. The suitable polymorphic markers determined in this study may also be validated and applied in the future to a large number of species for broad phylogenetic inference of the family Phyllanthaceae.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/d13090403/s1, Figure S1. Heatmap representing the relative synonymous codon usage. The y-axis represents 61 amino acid coding codons, and the x-axis represents species names. *Phyllanthus emblica* (Pak) was included in the comparison. Figure S2. Comparison of amino acid frequency within the species of Phyllanthaceae. The x-axis represents the amino acid with standard symbol, and the y-axis represents the frequency of amino acids. *Phyllanthus emblica* (Pak) was included in the comparison. Figure S3. Comparison of simple sequence repeats and oligonucleotide repeats among species of the family Phyllanthaceae. (a) Comparison of six types of SSRs. (b) Distributions of SSRs in the LSC, SSC, and IR regions. (c) Comparison of four types of oligonucleotide repeats. (d) Oligonucleotide repeat comparison based on size. (e) Distribution of oligonucleotide repeats in three main regions of the plastome. LSC = large single copy; SSC = small single copy; IR = inverted repeat; F = forward; R = reverse; C = complementary; P = palindromic. Table S1. Simple sequence repeats analysis among species of the family Phyllanthaceae. Table S2. Comparison of synonymous and non-synonymous substitutions among species of the family Phyllanthaceae. Table S3. Polymorphic analysis of protein-coding genes.

Author Contributions: Manuscript drafting, A. and U.R.; data analyses, U.R., A., M.M. and A.J.; data curation, U.R. and A.; data interpretation, U.R., A., P.P. and N.S.; conceptualization, U.R., N.S. and P.P.; review and editing of first draft, N.S. and P.P.; supervision, N.S.; project administration and resources, N.S. All authors have read and agreed to the published version of the manuscript.

Funding: Open access funding provided by University of Helsinki.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The newly assembled chloroplast genomes have been submitted to the National Center for Biotechnology under accession numbers BK059210 (*Flueggea virosa*), MZ424188 (*Leptopus cordifolius*), and MN122078 (*Phyllanthus emblica*). All of the analyzed data are available in the main manuscript or as Supplementary Material.

Conflicts of Interest: The authors have no conflict of interest to declare.

References

1. Chase, M.W.; Christenhusz, M.J.M.; Fay, M.F.; Byng, J.W.; Judd, W.S.; Soltis, D.E.; Mabberley, D.J.; Sennikov, A.N.; Soltis, P.S.; Stevens, P.F.; et al. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **2016**, *181*, 1–20.
2. Xi, Z.; Ruhfel, B.R.; Schaefer, H.; Amorim, A.M.; Sugumaran, M.; Wurdack, K.J.; Endress, P.K.; Matthews, M.L.; Stevens, P.F.; Mathews, S.; et al. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 17519–17524. [CrossRef] [PubMed]
3. Hoffmann, P.; Kathriarachchi, H.; Wurdack, K. A phylogenetic classification of Phyllanthaceae (Malpighiales; Euphorbiaceae sensu lato). *Kew Bull.* **2006**, *61*, 37–53.
4. Kathriarachchi, H.; Samuel, R.; Hoffmann, P.; Mlinarec, J.; Wurdack, K.J.; Ralimanana, H.; Stuessy, T.F.; Chase, M.W. Phylogenetics of tribe Phyllantheae (Phyllanthaceae; Euphorbiaceae sensu lato) based on *nr*ITS and plastid *mat*K DNA sequence data. *Am. J. Bot.* **2006**, *93*, 637–655. [CrossRef] [PubMed]
5. Christenhusz, M.J.M.; Byng, J.W. The number of known plants species in the world and its annual increase. *Phytotaxa* **2016**, *261*, 201–217. [CrossRef]
6. Bouman, R.W.; Keßler, P.J.A.; Telford, I.R.H.; Bruhl, J.J.; Strijk, J.S.; Saunders, R.M.K.; van Welzen, P.C. Molecular phylogenetics of *Phyllanthus sensu lato* (Phyllanthaceae): Towards coherent monophyletic taxa. *Taxon* **2021**, *70*, 72–98. [CrossRef]
7. Kathriarachchi, H.; Hoffmann, P.; Samuel, R.; Wurdack, K.J.; Chase, M.W. Molecular phylogenetics of Phyllanthaceae inferred from five genes (plastid *atp*B, *mat*K, 3′*ndh*F, *rbc*L, and nuclear *PHY*C). *Mol. Phylogenet. Evol.* **2005**, *36*, 112–134. [CrossRef]
8. Bouman, R.W.; Keßler, P.J.A.; Telford, I.R.H.; Bruhl, J.J.; Van Welzen, P.C. Subgeneric delimitation of the plant genus *Phyllanthus* (Phyllanthaceae). *Blumea J. Plant. Taxon. Plant. Geogr.* **2018**, *63*, 167–198. [CrossRef]
9. Mao, X.; Wu, L.; Guo, H.; Chen, W.; Cui, Y.; Qi, Q.; Li, S.; Liang, W.; Yang, G.; Shao, Y.; et al. The genus *Phyllanthus*: An ethnopharmacological, phytochemical, and pharmacological review. *Evid. Based Complement. Altern. Med.* **2016**, *2016*, 7584952. [CrossRef]
10. Gaire, B.P.; Subedi, L. Phytochemistry, pharmacology and medicinal properties of *Phyllanthus emblica* Linn. *Chin. J. Integr. Med.* **2014**, 1–8. [CrossRef]
11. Lim, T.K. *Edible Medicinal and Non-Medicinal Plants*; Springer: Dordrecht, The Netherlands, 2012; ISBN 978-94-007-1763-3.

12. Mandal, A. A Review on Phytochemical, Pharmacological and Potential Therapeutic Uses of *Phyllanthus Emblica*. *World J. Pharm. Res.* **2017**, *6*, 817–830. [CrossRef]
13. Lu, C.-C.; Yang, S.-H.; Hsia, S.-M.; Wu, C.-H.; Yen, G.-C. Inhibitory effects of *Phyllanthus emblica* L. on hepatic steatosis and liver fibrosis in vitro. *J. Funct. Foods* **2016**, *20*, 20–30. [CrossRef]
14. Luo, W.; Zhao, M.; Yang, B.; Ren, J.; Shen, G.; Rao, G. Antioxidant and antiproliferative capacities of phenolics purified from *Phyllanthus emblica* L. fruit. *Food Chem.* **2011**, *126*, 277–282. [CrossRef]
15. Zhao, T.; Sun, Q.; Marques, M.; Witcher, M. Anticancer Properties of *Phyllanthus emblica* (Indian Gooseberry). *Oxidative Med. Cell. Longev.* **2015**, *2015*, 950890. [CrossRef] [PubMed]
16. Vorontsova, M.S.; Hoffmann, P. Revision of the genus *leptopus* (Phyllanthaceae, Euphorbiaceae sensu lato). *Kew Bull.* **2009**, *64*, 627–644. [CrossRef]
17. Daniell, H. Transgene containment by maternal inheritance: Effective or elusive? *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 6879–6880. [CrossRef]
18. Neale, D.B.; Sederoff, R.R. Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in *Loblolly pine*. *Theor. Appl. Genet.* **1989**, *77*, 212–216. [CrossRef] [PubMed]
19. Abdullah; Mehmood, F.; Shahzadi, I.; Waseem, S.; Mirza, B.; Ahmed, I.; Waheed, M.T. Chloroplast genome of *Hibiscus rosa-sinensis* (Malvaceae): Comparative analyses and identification of mutational hotspots. *Genomics* **2020**, *112*, 581–591. [CrossRef] [PubMed]
20. Menezes, A.P.A.; Resende-Moreira, L.C.; Buzatti, R.S.O.; Nazareno, A.G.; Carlsen, M.; Lobo, F.P.; Kalapothakis, E.; Lovato, M.B. Chloroplast genomes of *Byrsonima* species (Malpighiaceae): Comparative analysis and screening of high divergence sequences. *Sci. Rep.* **2018**, *8*, 2210. [CrossRef] [PubMed]
21. Li, D.-M.; Zhao, C.-Y.; Liu, X.-F. Complete Chloroplast Genome Sequences of *Kaempferia Galanga* and *Kaempferia Elegans*: Molecular Structures and Comparative Analysis. *Molecules* **2019**, *24*, 474. [CrossRef] [PubMed]
22. Daniell, H.; Lin, C.-S.; Yu, M.; Chang, W.-J. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Biol.* **2016**, *17*, 134. [CrossRef] [PubMed]
23. Li, B.; Cantino, P.D.; Olmstead, R.G.; Bramley, G.L.C.; Xiang, C.L.; Ma, Z.H.; Tan, Y.H.; Zhang, D.X. A large-scale chloroplast phylogeny of the Lamiaceae sheds new light on its subfamilial classification. *Sci. Rep.* **2016**, *6*, 34343. [CrossRef] [PubMed]
24. Xu, J.-H.; Liu, Q.; Hu, W.; Wang, T.; Xue, Q.; Messing, J. Dynamics of chloroplast genomes in green plants. *Genomics* **2015**, *106*, 221–231. [CrossRef] [PubMed]
25. Saina, J.K.; Li, Z.Z.; Gichira, A.W.; Liao, Y.Y. The complete chloroplast genome sequence of tree of heaven (*Ailanthus altissima* (mill.) (sapindales: Simaroubaceae), an important pantropical tree. *Int. J. Mol. Sci.* **2018**, *19*, 929. [CrossRef]
26. Abdullah; Mehmood, F.; Heidari, P.; Ahmed, I.; Poczai, P. Pseudogenization of *trn*T-*GGU* in chloroplast genomes of the plant family Asteraceae. *bioRxiv* **2021**. [CrossRef]
27. Alqahtani, A.A.; Jansen, R.K. The evolutionary fate of *rpl*32 and *rps*16 losses in the *Euphorbia schimperi* (Euphorbiaceae) plastome. *Sci. Rep.* **2021**, *11*, 7466. [CrossRef] [PubMed]
28. Palmer, J.D. Comparative organization of chloroplast genomes. *Annu. Rev. Genet.* **1985**, *19*, 325–354. [CrossRef] [PubMed]
29. Ahmed, I. *Evolutionary Dynamics in Taro*; Massey University: Palmerston North, New Zealand, 2014.
30. Li, L.-F.; Wang, H.-Y.; Zhang, C.; Wang, X.-F.; Shi, F.-X.; Chen, W.-N.; Ge, X.-J. Origins and Domestication of Cultivated Banana Inferred from Chloroplast and Nuclear Genes. *PLoS ONE* **2013**, *8*, e80502. [CrossRef]
31. Henriquez, C.L.; Arias, T.; Pires, J.C.; Croat, T.B.; Schaal, B.A. Phylogenomics of the plant family Araceae. *Mol. Phylogenet. Evol.* **2014**, *75*, 91–102. [CrossRef]
32. Zhai, W.; Duan, X.; Zhang, R.; Guo, C.; Li, L.; Xu, G.; Shan, H.; Kong, H.; Ren, Y. Chloroplast genomic data provide new and robust insights into the phylogeny and evolution of the Ranunculaceae. *Mol. Phylogenet. Evol.* **2019**, *135*, 12–21. [CrossRef]
33. Li, Y.; Zhang, Z.; Yang, J.; Lv, G. Complete chloroplast genome of seven *Fritillaria* species, variable DNA markers identification and phylogenetic relationships within the genus. *PLoS ONE* **2018**, *13*, e0194613. [CrossRef] [PubMed]
34. Abdullah; Henriquez, C.L.; Mehmood, F.; Hayat, A.; Sammad, A.; Waseem, S.; Waheed, M.T.; Matthews, P.J.; Croat, T.B.; Poczai, P.; et al. Chloroplast genome evolution in the Dracunculus clade (Aroideae, Araceae). *Genomics* **2021**, *113*, 183–192. [CrossRef] [PubMed]
35. Jansen, R.K.; Cai, Z.; Raubeson, L.A.; Daniell, H.; dePamphilis, C.W.; Leebens-Mack, J.; Muller, K.F.; Guisinger-Bellian, M.; Haberle, R.C.; Hansen, A.K.; et al. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 19369–19374. [CrossRef]
36. Ahmed, I.; Matthews, P.J.; Biggs, P.J.; Naeem, M.; Mclenachan, P.A.; Lockhart, P.J. Identification of chloroplast genome loci suitable for high-resolution phylogeographic studies of *Colocasia esculenta* (L.) *Schott* (Araceae) and closely related taxa. *Mol. Ecol. Resour.* **2013**, *13*, 929–937. [CrossRef]
37. Liu, H.; Wei, J.; Yang, T.; Mu, W.; Song, B.; Yang, T.; Fu, Y.; Wang, X.; Hu, G.; Li, W.; et al. Molecular digitization of a botanical garden: High-depth whole-genome sequencing of 689 vascular plant species from the Ruili Botanical Garden. *Gigascience* **2019**, *8*, 1–9. [CrossRef]
38. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef]
39. Milne, I.; Bayer, M.; Cardle, L.; Shaw, P.; Stephen, G.; Wright, F.; Marshall, D. Tablet-next generation sequence assembly visualization. *Bioinformatics* **2009**, *26*, 401–402. [CrossRef] [PubMed]

40. Tillich, M.; Lehwark, P.; Pellizzer, T.; Ulbricht-Jones, E.S.; Fischer, A.; Bock, R.; Greiner, S. GeSeq—Versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **2017**, *45*, W6–W11. [CrossRef]
41. Lowe, T.M.; Chan, P.P. tRNAscan-SE On-line: Integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **2016**, *44*, W54–W57. [CrossRef] [PubMed]
42. Laslett, D.; Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **2004**, *32*, 11–16. [CrossRef]
43. Lehwark, P.; Greiner, S. GB2sequin—A file converter preparing custom GenBank files for database submission. *Genomics* **2019**, *111*, 759–761. [CrossRef]
44. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C.; et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **2012**, *28*, 1647–1649. [CrossRef] [PubMed]
45. Darling, A.C.E.; Mau, B.; Blattner, F.R.; Perna, N.T. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Res.* **2004**, *14*, 1394–1403. [CrossRef] [PubMed]
46. Amiryousefi, A.; Hyvönen, J.; Poczai, P. IRscope: An online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* **2018**, *34*, 3030–3031. [CrossRef]
47. Chen, C.; Chen, H.; Zhang, Y.; Thomas, H.R.; Frank, M.H.; He, Y.; Xia, R. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant.* **2020**, *13*, 1194–1202. [CrossRef] [PubMed]
48. Beier, S.; Thiel, T.; Münch, T.; Scholz, U.; Mascher, M. MISA-web: A web server for microsatellite prediction. *Bioinformatics* **2017**, *33*, 2583–2585. [CrossRef] [PubMed]
49. Kurtz, S.; Choudhuri, J.V.; Ohlebusch, E.; Schleiermacher, C.; Stoye, J.; Giegerich, R. REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **2001**, *29*, 4633–4642. [CrossRef]
50. Lawrie, D.S.; Messer, P.W.; Hershberg, R.; Petrov, D.A. Strong purifying selection at synonymous sites in *D. melanogaster. PLoS Genet.* **2013**, *9*, e1003527. [CrossRef]
51. Murrell, B.; Wertheim, J.O.; Moola, S.; Weighill, T.; Scheffler, K.; Kosakovsky Pond, S.L. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **2012**, *8*, e1002764. [CrossRef]
52. Murrell, B.; Moola, S.; Mabona, A.; Weighill, T.; Sheward, D.; Kosakovsky Pond, S.L.; Scheffler, K. FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection. *Mol. Biol. Evol.* **2013**, *30*, 1196–1205. [CrossRef]
53. Delport, W.; Poon, A.F.Y.; Frost, S.D.W.; Pond, S.L. Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **2010**, *26*, 2455–2457. [CrossRef]
54. Mehmood, F.; Abdullah; Ubaid, Z.; Shahzadi, I.; Ahmed, I.; Waheed, M.T.; Poczai, P.; Mirza, B. Plastid genomics of *Nicotiana* (Solanaceae): Insights into molecular evolution, positive selection and the origin of the maternal genome of Aztec tobacco (*Nicotiana rustica*). *PeerJ* **2020**, *8*, e9552. [CrossRef]
55. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]
56. Rozas, J.; Ferrer-Mata, A.; Sánchez-DelBarrio, J.C.; Guirao-Rico, S.; Librado, P.; Ramos-Onsins, S.E.; Sánchez-Gracia, A. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* **2017**, *34*, 3299–3302. [CrossRef] [PubMed]
57. Abdullah; Henriquez, C.L.; Mehmood, F.; Carlsen, M.M.; Islam, M.; Waheed, M.T.; Poczai, P.; Croat, T.B.; Ahmed, I. Complete chloroplast genomes of *Anthurium huixtlense* and *Pothos scandens* (Pothoideae, Araceae): Unique inverted repeat expansion and contraction affect rate of evolution. *J. Mol. Evol.* **2020**, *88*, 562–674. [CrossRef]
58. Lockhart, P.; Novis, P.; Milligan, B.G.; Riden, J.; Rambaut, A.; Larkum, T. Heterotachy and Tree Building: A Case Study with Plastids and *Eubacteria*. *Mol. Biol. Evol.* **2006**, *23*, 40–45. [CrossRef]
59. Zhu, A.; Guo, W.; Gupta, S.; Fan, W.; Mower, J.P. Evolutionary dynamics of the plastid inverted repeat: The effects of expansion, contraction, and loss on substitution rates. *New Phytol.* **2016**, *209*, 1747–1756. [CrossRef] [PubMed]
60. Katoh, K.; Kuma, K.I.; Toh, H.; Miyata, T. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **2005**, *33*, 511–518. [CrossRef] [PubMed]
61. Darriba, D.; Taboada, G.L.; Doallo, R.; Posada, D. jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **2012**, *9*, 772. [CrossRef]
62. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef]
63. Letunic, I.; Bork, P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **2019**, *47*, W256–W259. [CrossRef]
64. Choi, K.S.; Kwak, M.; Lee, B.; Park, S.J. Complete chloroplast genome of *tetragonia tetragonioides*: Molecular phylogenetic relationships and evolution in caryophyllales. *PLoS ONE* **2018**, *13*, e0199626. [CrossRef]
65. Henriquez, C.L.; Abdullah; Ahmed, I.; Carlsen, M.M.; Zuluaga, A.; Croat, T.B.; Mckain, M.R. Evolutionary dynamics of chloroplast genomes in subfamily Aroideae (Araceae). *Genomics* **2020**, *112*, 2349–2360. [CrossRef] [PubMed]
66. Abdullah; Henriquez, C.L.; Mehmood, F.; Shahzadi, I.; Ali, Z.; Waheed, M.T.; Croat, T.B.; Poczai, P.; Ahmed, I. Comparison of chloroplast genomes among Species of Unisexual and Bisexual clades of the monocot family Araceae. *Plants* **2020**, *9*, 737. [CrossRef] [PubMed]

67. Henriquez, C.L.; Abdullah; Ahmed, I.; Carlsen, M.M.; Zuluaga, A.; Croat, T.B.; Mckain, M.R. Molecular evolution of chloroplast genomes in Monsteroideae (Araceae). *Planta* **2020**, *251*, 72. [CrossRef]
68. Shahzadi, I.; Abdullah; Mehmood, F.; Ali, Z.; Ahmed, I.; Mirza, B. Chloroplast genome sequences of *Artemisia maritima* and *Artemisia absinthium*: Comparative analyses, mutational hotspots in genus *Artemisia* and phylogeny in family Asteraceae. *Genomics* **2020**, *112*, 1454–1463. [CrossRef] [PubMed]
69. Ahmed, I.; Biggs, P.J.; Matthews, P.J.; Collins, L.J.; Hendy, M.D.; Lockhart, P.J. Mutational dynamics of aroid chloroplast genomes. *Genome Biol. Evol.* **2012**, *4*, 1316–1323. [CrossRef] [PubMed]
70. Millen, R.S.; Olmstead, R.G.; Adams, K.L.; Palmer, J.D.; Lao, N.T.; Heggie, L.; Kavanagh, T.A.; Hibberd, J.M.; Gray, J.C.; Morden, C.W.; et al. Many parallel losses of *inf* A from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. *Plant. Cell* **2001**, *13*, 645–658. [CrossRef]
71. Abdullah; Waseem, S.; Mirza, B.; Ahmed, I.; Waheed, M.T. Comparative analyses of chloroplast genomes of *Theobroma cacao* and *Theobroma grandiflorum*. *Biologia* **2020**, *75*, 761–771. [CrossRef]
72. Poczai, P.; Hyvönen, J. The complete chloroplast genome sequence of the CAM epiphyte Spanish moss (*Tillandsia usneoides*, Bromeliaceae) and its comparative analysis. *PLoS ONE* **2017**, *12*, e0187199. [CrossRef]
73. Lin, M.; Qi, X.; Chen, J.; Sun, L.; Zhong, Y.; Fang, J.; Hu, C. The complete chloroplast genome sequence of *Actinidia arguta* using the PacBio RS II platform. *PLoS ONE* **2018**, *13*, e0197393. [CrossRef]
74. Hu, Y.; Woeste, K.E.; Zhao, P. Completion of the Chloroplast Genomes of Five Chinese Juglans and Their Contribution to Chloroplast Phylogeny. *Front. Plant. Sci.* **2017**, *7*, 1955. [CrossRef]
75. Wang, C.-L.; Ding, M.-Q.; Zou, C.-Y.; Zhu, X.-M.; Tang, Y.; Zhou, M.-L.; Shao, J.-R. Comparative analysis of four *Buckwheat* species based on morphology and complete chloroplast genome sequences. *Sci. Rep.* **2017**, *7*, 6514. [CrossRef] [PubMed]
76. Mehmood, F.; Abdullah; Ubaid, Z.; Bao, Y.; Poczai, P. Comparative Plastomics of *Ashwagandha* (*Withania*, Solanaceae) and Identification of Mutational Hotspots for Barcoding Medicinal Plants. *Plants* **2020**, *9*, 752. [CrossRef] [PubMed]
77. Mehmood, F.; Abdullah; Shahzadi, I.; Ahmed, I.; Waheed, M.T.; Mirza, B. Characterization of *Withania somnifera* chloroplast genome and its comparison with other selected species of Solanaceae. *Genomics* **2020**, *112*, 1522–1530. [CrossRef]
78. Amiryousefi, A.; Hyvönen, J.; Poczai, P. The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): Plastid genome structure evolution in Solanaceae. *PLoS ONE* **2018**, *13*, e0196069. [CrossRef] [PubMed]
79. Iram, S.; Hayat, M.Q.; Tahir, M.; Gul, A.; Abdullah; Ahmed, I. Chloroplast genome sequence of *Artemisia scoparia*: Comparative analyses and screening of mutational hotspots. *Plants* **2019**, *8*, 476. [CrossRef]
80. McDonald, M.J.; Wang, W.C.; Da Huang, H.; Leu, J.Y. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biol.* **2011**, *9*, e1000622. [CrossRef]
81. Abdullah; Mehmood, F.; Shahzadi, I.; Ali, Z.; Islam, M.; Naeem, M.; Mirza, B.; Lockhart, P.; Ahmed, I.; Waheed, M.T. Correlations among oligonucleotide repeats, nucleotide substitutions and insertion-deletion mutations in chloroplast genomes of plant family Malvaceae. *J. Syst. Evol.* **2021**, *59*, 388–402. [CrossRef]
82. Abdullah; Henriquez, C.L.; Croat, T.B.; Poczai, P.; Ahmed, I. Mutational dynamics of aroid chloroplast genomes II. *Front. Genet.* **2021**, *11*, 610838. [CrossRef]
83. Abdullah; Mehmood, F.; Rahim, A.; Heidari, P.; Ahmed, I.; Poczai, P. Comparative plastome analysis of *Blumea*, with implications for genome evolution and phylogeny of Asteroideae. *Ecol. Evol.* **2021**, *11*, 7810–7826. [CrossRef] [PubMed]
84. Li, X.; Yang, Y.; Henry, R.J.; Rossetto, M.; Wang, Y.; Chen, S. Plant DNA barcoding: From gene to genome. *Biol. Rev.* **2014**, *90*, 157–166. [CrossRef] [PubMed]
85. Nguyen, V.B.; Park, H.-S.; Lee, S.-C.; Lee, J.; Park, J.Y.; Yang, T.-J. Authentication markers for five major *Panax* species developed via comparative analysis of complete chloroplast genome sequences. *J. Agric. Food Chem.* **2017**, *65*, 6298–6306. [CrossRef] [PubMed]