 doi:10.20944/preprints202107.0094.v1

Chloroplast genome of *Phyllanthus emblica* and *Leptopus cordifolius*: Comparative analysis and phylogenetic within family Phyllanthaceae

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

[1]*Department of Biochemistry, Hazara University, Mansehra, KPK, Pakistan*
[2]*Departemet of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan*
[3]*Key Laboratory of Horticulture Plant Biology (Ministry of Education), College of Horticulture and Forestry Sciences, Huazhong Agriculture University, Wuhan 430070, China*
[4]*Alpha Genomics Private Limited, Islamabad, Pakistan*

[5]*Finnish Museum of Natural History, P.O. Box 7, FI-00014 University of Helsinki, Finland*

[6]*Faculty of Biological and Environmental Sciences, PO Box 65 FI-00065 University of Helsinki, Finland*

*Correspondence authors: *Nighat Sultana (nighat.sultana@hu.edu.pk); Péter Poczai (peter.poczai@helsinki.fi)*

Abstract: Family Phyllanthaceae is one of the largest segregates of the eudicot order Malpighiales and its species are herb, shrub, and tree, which are mostly distributed in tropical regions. Certain taxonomic discrepancies exist at genus and family level. Here, we report chloroplast genomes of three Phyllanthaceae species—*Phyllanthus emblica*, *Flueggea virosa*, and *Leptopus cordifolius*— and compare them with six others previously reported Phyllanthaceae chloroplast genomes. The species of Phyllanthaceae displayed quadripartite structure, comprising inverted repeat regions (IRa and IRb) that separate large single copy (LSC) and small single copy (SSC) regions. The length of complete chloroplast genome ranged from 154,707 bp to 161,093 bp; LSC from 83,627 bp to 89,932 bp; IRs from 23,921 bp to 27,128 bp; and SSC from 17,424 bp to 19,441 bp. Chloroplast genomes contained 111 to 112 unique genes, including 77 to 78 protein-coding, 30 transfer RNA (tRNA), and 4 ribosomal RNA (rRNA) that showed similarities in arrangement. The number of protein-coding genes varied due to deletion/pseudogenization of *rps*16 genes in *Baccaurea ramiflora* and *Leptopus cordifolius*. High variability was seen in number of oligonucleotide repeats while analysis of guanine-cytosine (GC) content, codon usage, amino acid frequency, simple sequence repeats analysis, synonymous and non-synonymous substitutions, and transition and transversion substitutions showed similarities in all Phyllanthaceae species. We detected a higher number of transition substitutions in the coding sequences than non-coding sequences. Moreover, the high number of transition substitutions was determined among the distantly related species in comparison to closely related species. Phylogenetic analysis shows the polyphyletic nature of the genus *Phyllanthus* which requires further verification. We also determined suitable polymorphic coding genes, including *rpl*22, *ycf*1, *mat*K, *ndh*F, and *rps*15 which may be helpful for the reconstruction of the high-resolution phylogenetic tree of the family Phyllanthaceae using a large number of species in the future. Overall, the current study provides insight into chloroplast genome evolution in Phyllanthaceae.

Key words: Phyllanthaceae, Phyllanthus, Leptopus, transition and transversion substitutions, chloroplast genome, phylogenetic, polymorphic loci, *rps*16 loss, *atp*F intron loss

 doi:10.20944/preprints202107.0094.v1

1. Introduction

The plant family Phyllanthaceae together with 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]. *Phyllanthus* L. is considered among the largest genera of the family Phyllanthaceae with 900 species [6,7], with tropical to subtropical distribution [8]. Many species of the genus have medicinal uses in Southeast Asian countries, including China, India, and Brazil [6]. *Phyllanthus emblica* L. is cultivated for fruit [9] and has many medicinal properties, including antioxidant, anti-cancer, anti-inflammatory, anti-pyretic diuretic, laxative, stomachic, cardioprotective, and hepatoprotective [6,8,10–13]. The genus *Leptopus* Decne. is comprised of 9 species of herbs and shrubs that grow in open vegetation and in the forest industry [14]. The species of genus *Leptopus* are mostly distributed in Asia, including central, northern and southern China, India, Pakistan and Iran. However, some species also exist in America and Europe [14]. *Leptopus cordifolius* Decne. grows in wild and distributed locations in Pakistan, through north India and Nepal to West Himalaya [14].

The chloroplast is an important organelle that plays a vital role in photosynthesis. The chloroplast genome has a mostly quadripartite structure and consist of large single copy (LSC) region, small single copy (SSC) region, and inverted repeat regions (IRa and IRb) [15–17]. The size of the chloroplast genome of photosynthetic plants ranges from 107 KB to 218 KB and contain 120 to 130 genes [18]. Th chloroplast genome inherits maternally [19] in most angiosperms and paternally in some gymnosperms [20]. Many mutational events are reported in chloroplast genomes, including substitutions, insertion-deletions (InDels), repeats, and inversion [21–23]. The deletion of certain genes from the chloroplast genome or their transfer to nuclear genomes are also reported in several plant lineages, including the species of order Malpighiales to which family Phyllanthaceae belong [16,24,25]. The chloroplast genome is slowly evolving and lacks the meiotic recombination that is seen in the nuclear genome where homologous chromosomes exchange segments [26]. These properties make the chloroplast genome a suitable molecule for studies ranging from population genetics [27,28] to deep divergence of genus and family level [29–32], including plant evolution and phylogeny, providing deep insight into the evolution of the plant [33,34].

 doi:10.20944/preprints202107.0094.v1

In the current study, the chloroplast genome of three species was reported and compared with six other species of the family Phyllanthaceae. The aims of the current study were to: (a) gain insight into chloroplast genome characteristics and structure; (b) elucidate the molecular evolution of the chloroplast genome; (c) gain insight into synonymous and non-synonymous substitutions and transition and transversion substitutions; (d) determine suitable polymorphic loci for phylogenetic inference; and (e) determine the phylogenetic position of the newly reported species within the family Phyllanthaceae.

2. Materials and methods

2.1 Plant Collection and DNA extraction

The healthy leaves of *Phyllanthus emblica* and *Leptopus cordifolius* without any apparent disease symptoms were collected from the Mansehra district, Khyber Pakhtunkhwa, Pakistan. These 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 from Qiagen. After confirmation of the quality and quantity on a 1% agarose gel and using a Thermo Scientific Nanodrop spectrophotometer, DNA was sent for sequencing with Novogene, Hong Kong.

2.2 Sequencing, *de novo* assembly, and annotation of chloroplast genome

The DNA of *Phyllanthus emblica* and *Leptopus cordifolius* 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) 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 the mentioned three species using NOVOPlasty. Use of this method, provides a complete circular chloroplast genome including a Large Single Copy (LSC) region, Inverted repeat (IR) regions, and a Small Single Copy (SSC) region. The accuracy of the assembled genome and the coverage depth was confirmed by mapping short reads to its *de novo* assembled chloroplast genome by using Burrows-Wheeler Aligner (BWA) [35] and visualized in Tablet [36].

De novo assembled chloroplast genome of *Phyllanthus emblica* was annotated using GeSeq [37] whereas the tRNA genes were further confirmed by tRNAscan-SE v.2 [38] with default

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

parameters. The annotations of protein-coding genes were further confirmed against homologous genes by blast search of National Center of Biotechnology information (NCBI). The five column delimited table was formed with tools of GB2sequin [39] for submission to GenBank of NCBI. The assembled chloroplast genome was submitted to GenBank of NCBI.

2.3 Analysis of chloroplast genome features and inverted repeat contraction and expansion

The *de novo* assembled three chloroplast genomes of *Phyllanthus emblica*, *Leptopus cordifolius,* and *Flueggea virosa* were compared with previously reported species (mentioned in Table 1). The length of complete chloroplast genome, LSC, SSC, and IRs were compared among the Phyllanthaceae, including gene content, intron content, and GC content of each part in Geneious R8.1 [40]. We also compared the gene arrangement among species of Phyllanthaceae using Mauve alignment [41] integrated in Geneious R8.1. The contraction and expansion of inverted repeat regions were visualized using IRscope [42].

2.4 Analysis of codon usage, amino acid frequency and prediction of RNA editing sites

The relative synonymous codon usage (RSCU) and amino acid frequency of each species was determined using Geneious R8.1 [40]. The RSCU value of each species were shown with the help of a heatmap using TBtool [43].

2.5 Synonymous, non-synonymous, transition, and transversion substitutions analysis

The synonymous (Ks) and non-synonymous (Ka) substitutions were analyzed in TBtool [43] for 77 protein-coding genes using *Baccaurea ramiflora* Lour. as reference for all other species as this species lies basal to current species following Abdullah et al. [15]. Each gene selection was predicted following Henriquez et al. [44] by considering ration of Ka/Ks <1 purifying selection, Ka/Ks = 1 neutral selection, and Ka/Ks >1 positive selection. The transition and transversion substitutions of complete genome and protein-coding genes were determined by comparing closely related species and far related species. The whole genome was aligned using MAFFT, whereas the protein-coding sequences of each species were concatenated, except for *ycf1* and aligned using MAFFT alignment [45]. For closely related species, the genome of *Phyllanthus emblica* L. (Pakistan) was used as reference for *Breynia fruticosa* (L.) Müll.Arg., *Phyllanthus amarus* Schumach. & Thonn., *Phyllanthus emblica* (China), *Glochidion chodoense* C.S. Lee & H.T. Im and *Sauropus spatulifolius*. The species that were far related

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

were compared using *Baccaurea ramiflora* as reference for *Leptopus cordifolius*, *Flueggea virosa*, and *Phyllanthus emblica* (Pakistan).

2.6 Repeats analyses

MIcroSAtellite (MISA) [46] was used for the determination of simple sequence repeats (SSRs) with the maximum threshold of 10 for mononucleotide SSRs, 5 for dinucleotide SSRs, and 4 for tri-, and 3 for tetra-, penta-, and hexanucleotide SSRs. REPuter [47] was used for the determination of four types of oligonucleotide repeats including forward (F), complementary (C), reverse (R), and palindromic (P). The parameters, such as repeat size ≥ 30 with at least 90% similarities, were adjusted.

2.7 Polymorphism of protein-coding genes

To determine the extent of polymorphism of protein-coding genes for further phylogenetic studies, we extracted all the protein-coding genes of each species. The sequence of each gene was multiple aligned using Geneious R8.1 and analyzed in DnaSP v.6 [48]

2.8 Reconstruction of phylogenetic tree

For inferring of phylogeny, the sequences of 76 protein-coding genes (*ycf*1 not included) of each species were extracted and concatenated in Geneious R8.1 [49]. These concatenated sequences of all the 10 species including *Linum usitatissimum* L. as outgroup from the family Pinaceae were multiple aligned using Multiple Alignment Fast Fourier Transform (MAFFT) [50] extension in Geneious R8.1. The maximum likelihood tree was reconstructed with RAxML [51] using CIPRES gateway [52]. The Interactive Tree of Life [53] 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 7.85 GB data with 22.6 million reads for *Leptopus cordifolius*. The chloroplast genomes assembled from the data exhibited 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 show the average coverage depth of 69X. These three de novo assembled genomes, along with the previously reported genome, provided an opportunity to perform in depth comparative chloroplast genomics in family Phyllanthaceae.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

The gene content and organization of the chloroplast genomes of family Phyllanthaceae are provided in Table 1, Table 2, and Figure 1. The genomes show high similarity in gene content except *rps16* which was missing in *Baccaurea ramiflora and Leptopus cordifolius*. The chloroplast genome of all species displayed quadripartite structure, comprised of IRs (IRa and IRb) that separate LSC and SSC regions. The length of the complete chloroplast genome ranged from 154,707 bp (*Sauropus spatulifolius*) to 161,093 bp (*Baccaurea ramiflora*), IRs from 23,921 bp (*Sauropus spatulifolius*) to 27,128 bp (*Phyllanthus amarus*), LSC from 83,627 bp (*Leptopus cordifolius*) to 89,932 bp (*Phyllanthus emblica* of Pakistan), and SSC from 17,424 bp (*Leptopus cordifolius*) to 19,441 bp (*Breynia fruticosa*).

The chloroplast genomes had 111 to 112 unique genes which showed high similarities in the arrangement within the genomes and found no inversion events related to gene rearrangement as predicted by Mauve alignment (Figure 2). Among these 112 genes, 77 to 78 were protein-coding genes, 30 transfer RNA (tRNA) genes, and 4 ribosomal RNA (rRNA) genes (Table 1). Moreover, 16 to 17 genes were also duplicated in IR regions, including 5 to 6 protein-coding genes, 7 tRNA genes, and 4 rRNA genes. Here, we exclude two pseudogenes of *ycf1*$^{\Psi}$ and *rps19*$^{\Psi}$, which left at the junctions of LSC/IRa and SSC/IRa (Table 1). The *rps12* gene was a trans-splicing gene due to which 5′ part present in LSC in the form of single copy whereas the 3′ part duplicated in IRa and IRb. The average GC content of the complete chloroplast genome was 36.6% to 36.8%. LSC was 34.3% to 34.6%, SSC was 30.1% to 30.8%, and of IRs was 42.3% to 43.3%. Maximum GC content was found in IR regions (42.3–43.3%), followed by LSC region (34.3–34.6%), and then by SSC region (30.1–30.8%). The high GC content of IRs belongs to rRNAs (55.5%) and tRNAs (53%). Summary details of the genomic features are given in Table 2.

We found 17 to 18 intron-containing genes, including 11 to 12 protein-coding genes and 6 tRNA genes. The intron was found in the *atp*F gene of *Baccaurea ramiflora* only while found absent in all other species. Except *clpP* and *ycf3* that contained two introns, all other genes contained one intron (Table 1). Among intron-containing genes, 5 genes (3 protein-coding genes and 2 tRNA genes) were also duplicated in IRs regions.

3.2 Inverted repeat contraction and expansion

Comparative analysis of the junctions of LSC/IR and SSC/IR showed similarities among the species of family Phyllanthaceae except few differences (Figure **3**). At the LSC/IRb junction (JLB), the *rps19* start from the IRb and enter LSC regions in six species. Hence, a pseudo-copy

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

of the *rps19* gene at JLA (IRa/LSC) left in these species (Figure 3). In the remaining three species, the *rps*19 completely exists in the LSC regions at the JLB junction, which was not led to the generation of a pseudo copy at the JLA junction. The *rpl*2 gene exists completely in the IR regions away from the JLB and JLA junctions, except for in *Sauropus spatulifolius* where the *rpl*2 start is in IRb and enters into LSC regions, leaving a pseudo-copy at the junction of JLA. At the SSC/IRb junction (JSB), *ndh*F exists entirely in the SSC region in the six species whereas it integrates into the IRb regions within three species, overlapping with the pseudo-copy of *ycf*1. The *ycf*1 started in IRa and ended in SSC, as shown at the JSA (SSC/IRa) junction. Consequently, a pseudogene of *ycf*1 remains at JSB (IRb/SSC) which ranges in size from 195 bp to 1921 bp. At the junction of JLA (IRa/LSC), the *trn*H gene was found in all species (Figure **3**). So, the IR contraction and expansion was not responsible for complete duplication of a functional copy of a gene, but it leads to the generation of pseudo-copies/copy of *rps*19, *rpl*2 and *ycf*1.

3.3 Codon usage and amino-acid frequency

Codon usage analysis was interpreted in terms of RSCU values. The analysis showed that most amino acids were encoded from those codons for which the 3′ ended with A/T (having RSCU > 1) instead of with C/G (having RSCU < 1) (Figure 4). The amino acid frequency analysis revealed that leucine was the most encoded amino acid while the cysteine was the rarest (Figure 5). Codon-usage analysis and amino-acid frequency show high similarities in all species of the family Phyllanthaceae.

3.4 Analysis of simple sequence repeats and oligonucleotide repeats

We analyzed simple sequence repeats (SSRs) and oligonucleotide repeats. The analysis of MISA revealed 630 SSRs made of A/T motifs in all species (Table S1). We found all six types of SSRs based on the motif types including mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide. However, SRRs were predominantly mononucleotide, followed by dinucleotide (Figure 6a). Most of the SSRs exist in the LSC region, followed by the SSC, and then by the IR region (Figure 6b). Most of the SSRs (84) were predicted in *Phyllanthus emblica* (China) and the lowest SRRs (51) were predicted 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, whereas less reverse and complementary repeats were found (Figure 6c). Most of the repeats ranged in size

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

from 30–34 bp (Figure 6d). The LSC contained a high number of repeats in comparison to SSC and IR, whereas some repeats were also shared between regions of chloroplast genome including LSC/IR, LSC/SSC, and SSC/IR (Figure 6e).

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 acts on the protein-coding genes of the species of family Phyllanthaceae. Two genes—*psb*L and *pet*L—showed a signature of positive selection in *Sauropus spatulifolius* and *Flueggea virosa*, respectively. Moreover, the value of *rpl*23 was 1.04 for *Glochidion chodoense* and *Phyllanthus amarus*.

3.6 Transition and transversion substitutions

The transition and transversion substitutions within the complete chloroplast genome and within the protein-coding regions were analyzed. The comparative analysis of the complete chloroplast genome among the far related species of Phyllanthaceae revealed the existence of 10,950 to 12,603 substitutions, whereas comparison of closely related species revealed the presence of 257 to 2077 substitutions (Figure 7a). We predicted 3237 to 3747 substitutions when comparing protein-coding sequences of far related species and 65 to 685 substitutions in closely related species (Figure 7b). The ratio of transition and transversion substitutions (Ts/Tv) also revealed variations among the far related species and closely related species and the Ts/Tv ratio of far related species varied from 1.34 to 1.40 whereas Ts/Tv ratio of closely related species varied from 0.77 to 1.12 (Figure 7c). The analysis of protein-coding sequences revealed a higher ratio of Ts/Tv as compared to the complete chloroplast genome. The Ts/Tv ratio for far related species was 2.09 to 2.34, and for closely related species was 1.42 to 1.71 (Figure 7c). The higher Ts/Tv in coding regions showed the occurrence of higher transition substitutions within the protein coding regions as compared to transversion substitutions, whereas the recorded lower Ts/Tv ratio in complete genome than coding regions revealed that this may be due to inclusion of intergenic spacer region in which higher transversion substitutions take place compared to transition substitutions.

3.7 Intra species variations in chloroplast genome of *Phyllanthus emblica*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

We determined the intra species variations in the chloroplast genome of *Phyllanthus emblica* that belong to two different countries—Pakistan and China. The analysis predicted 257 SNPs in the whole chloroplast genome, of which 79 belong to coding regions. We also predicted about 108 insertions-deletions (InDels) events with an average length of 5.324 in which 3 Indels belong to coding regions. We have not predicted any InDels or SNPs in genes of transfer RNAs and ribosomal RNAs.

3.8 Polymorphism of protein-coding genes

The polymorphism of all protein-coding genes was accessed which helped us to identify the suitable polymorphic loci for future phylogenetic inference of family Phyllanthaceae (Table S3). We selected fifteen genes of ≥ 200 bp. We also accessed the missing data produced by the high polymorphic regions due to Indels to provide information about the suitability of these loci. These loci include *rpl*22, *ycf*1, *mat*K, *ndh*F, *rps*15 etc. as shown in table 3.

3.9 Phylogenetic analysis

The phylogenetic analysis based on 75 coding sequences resolved the relationship of the limited species of family Phyllanthaceae with high bootstrapping support. The polyphyletic position of the genus *Phyllanthus* as the species of three genera, including *Breynia*, *Sauropus*, and *Glochidion,* was embedded between *Phyllanthus emblica* and *Phyllanthus amarus* (Figure 8). The node showing the polyphyletic nature of the *Phyllanthus* was root by *Flueggea virosa,* then by *Leptopus cordifolius,* and then by *Baccaurea ramiflora*.

4. Discussion

4.1 Chloroplast genome assembly from whole genome sequencing and comparative chloroplast genomics

In this study, we employed whole genome shotgun sequencing to assemble complete chloroplast genomes with a high coverage depth. The conventional method of chloroplast genome sequencing was isolation or enrichment of the chloroplast genome by amplification with long-range PCR after DNA extraction [54,55]. The advancement of Next generation sequencing technologies makes it feasible to assemble chloroplast genome from whole genome shotgun sequences due to the generation of high parallel sequencing data [56]. Several other studies also reported *de novo* assembled chloroplast genomes from whole genome shotgun reads [57,58].

The angiosperm chloroplast genome is conserved in gene content and gene order. Both introns in the protein-coding genes and tRNA genes are conserved. Yet, the loss of some genes or introns is also reported in some species [16,18,33,59]. The chloroplast genome features of the species of family Phyllanthaceae were similar to each other with few differences. The intron of *atp*F gene was deleted in all other 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 two species—*Baccaurea ramiflora* (deletion of exon 1) and *Leptopus* cordifolius (on few fragments exits). The deletion of rps16 was also reported in other species of the order Malphighiales, however, only *Glochidion chodoense* was included in the comparison and *rps*16 was functional and present in the species of family Phyllanthaceae. Hence, this study sheds light on the deletion/pseudogenization of *rps*16 in the family Phyllanthaceae for the first time. The GC content of species of family *Phyllanthaceae* is also similar to other angiosperms and other species of Order Malphighiales [16,44,60]. The *infA* gene was fond completely absent in the chloroplast genome of the species of family Phyllanthaceae. The presence of a pseudogene of *infA* or the complete loss of this gene is also reported in other families of angiosperms, including families Araceae [58,61,62], Malvaceae [15], and Malpighiaceae [16]. The *infA* gene has a vital function as a translation initiation factor. The lack of the complete gene or presence of a pseudogene revealed the existence of a functional copy or transferring of *infA* gene to the nuclear genome [33,63].

Inverted repeat contraction and expansion leads to the generation of both a pseudo copy and a functional copy of a gene, changing duplicated genes to a single copy gene due to transfer of IRs to single copy regions (LSC or SSC), or conversion of a single copy gene to duplicated gene due to transfer from single copy regions (LSC or SSC) to IRs, as reported previously [15,32,57,62]. In the current study, the species of family Phyllanthaceae were also found to be conserved in terms of contraction and expansion of IRs. However, the generation of a pseudo copy of *ycf*1, *rpl*2, and *rps*19 remained similar to other angiosperms [64,65]. The duplication of a complete functional gene such as of *ycf1*, *rps*15 and *rps*19 that is seen in other angiosperms was not detected here [15,57,62].

4.2 Simple sequence repeats and oligonucleotide repeat analyses

Simple sequence repeats (SSRs) and oligonucleotide repeats were also analyzed in the chloroplast genome. We observed the highest number of SSRs in the LSC region as compared to SSC and IR regions. Most of the mononucleotide SSRs were A/T whereas for dinucleotide

 doi:10.20944/preprints202107.0094.v1

SSRs AT/TA SSRs were more abundant. Most of the SSRs were mononucleotide, followed by dinucleotide SSRs, and then by trinucleotide SSRs. These findings are consistent with previous studies [66–71]. However, a higher number of trinucleotides, compared to dinucleotides, has also been reported, [15,55,60,72]. The SSR marker identified in the current study can serve as a resource for population genetics studies of the species of family Phyllanthaceae. We found forward, reverse, complementary, and palindromic repeats in chloroplast genomes using REPuter. These oligonucleotide repeats originate InDels and substitutions [73–75] and can be used as proxy for identification of mutational hotspot regions [61,74,76].

4.3 Synonymous and non-synonymous substitutions and positive selection analysis

The analysis of synonymous and non-synonymous substitutions mostly revealed Ka/Ks < 1 and show that these genes are under high purifying selection pressure. The chloroplast genome performs a special function in the plant during photosynthesis and survives under to high genetic selection pressure. Hence the similar has been reported in other angiosperms, including the other species of order Malpighiales [16,77–79]. However, *psb*L and *pet*L show signature of positive selection in *Sauropus spatulifolius* and *Flueggea virosa*. The *psb*L belong to photosystem II, and *pet*L belong to the cytochrome b/f complex group. The positive selection of these genes shows that these may be helpful for biotic and abiotic stresses faced by these species in their ecological niches.

4.4 Higher transition substitutions in protein-coding genes than non-coding regions and in far related species than closely related species

Higher transition substitutions occur within in protein-coding genes as compared to non-coding regions. Moreover, the transition rate was also high in far related species compared to closely related species. The transfer RNA and ribosomal RNA genes are conserved within species and GC rich, as observed here and reported previously [55,74,80,81]. Therefore, the non-coding part has less GC content than coding sequences as shown in Table 2. The higher Ts/Tv ratio was also observed in the protein-coding sequences of family Araceae [32,58,82], whereas the low Ts/Tv ratio (up to 1) was reported based on the comparison of chloroplast genomes among closely related species, such as Asteraceae [76], Solanaceae [69,70], and Malvaceae [15,54]. The high GC content of coding sequences (up to 37%) compared to non-coding regions (up to 32.5%) might be the reason for higher rate of transition and less transversion substitutions in coding sequences compared to non-coding regions as reported for the chloroplast genome of maize and rice [83]. Another study showed that the ratio of Ts/Tv effected from the function

 doi:10.20944/preprints202107.0094.v1

of regional and flanking base composition and suggested the codon usage bias of chloroplast genomes as a possible reason [84] given that codons ending with A/T at the 3′ end showed a higher abundance and high encoding efficacy as reported in the current study and in the previous reports of angiosperms [15,44,62,74,76,85]. The current study, together with previous reports [32,44,58], also shows that most coding sequence substitutions are synonymous instead of non-synonymous to avoid integration of a new amino acid to protein sequences. Recent studies support that most of the synonymous substitutions are linked to transition substitutions and non-synonymous substitutions to transversion substitutions [62,81] as most are synonymous in the protein-coding sequences. This might also be a reason for the higher ratio of Ts/Tv in protein-coding sequences. Further study is required, however, to determine why the Ts/Tv ratio is higher in far diverse species in comparison to closely related species.

4.5 Phylogenetics analysis and suitable polymorphic loci for further phylogenetic inference

The phylogenetic analysis showed the polyphyletic nature of the genus *Phyllanthus* similar to 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 [86]. Hence, the author suggested the division of the large genus *Phyllanthus* to nine small genera instead of combining the embedded genera to *Phyllanthus*. Our study was based on 75 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. The other species show the similar phylogenetic position as reported previously [4,87]. Certain discrepancies still exist in the phylogeny of the family Phyllanthaceae, which warrant further elaboration [3,4,87]. Here, we identified 15 suitable high polymorphic regions from protein-coding sequences based on the comparative analysis of chloroplast genomes. 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 [18,34,76,88,89].

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 family Phyllanthaceae for the first time. Our study shows the higher transition rate in coding sequences as compared to non-coding sequences and the polyphyletic nature of the genus *Phyllanthus*. We also

 doi:10.20944/preprints202107.0094.v1

determined suitable polymorphic loci, which may be helpful for the broad phylogenetic inference of the family Phyllanthaceae in the future.

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

Conflict of Interest: Authors declare that conflict of interest does not exist.

Author's contribution: Manuscript drafting: A.; U.R.; Data analyses: U.R.; A.; M.M.; A.J.; Data curation: U.R, A., N.S.; A.; Data interpretation: U.R.; A.; P.P.; N.S.; Conceptualization: U.R.; N.S.; P.P.; Review and editing of first draft: N.S.; P.P.; Supervision: N.S., Project administration and resources: N.S.,

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. U. S. A.* **2012**, *109*, 17519–17524.

3. Hoffmann, P.; Kathriarachchi, H.; Wurdack, K. A phylogenetic classification of Phyllanthaceae (Malpighiales; Euphorbiaceae sensu lato). *Kew Bull.* **2006**, 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 nrITS and plastid matK DNA sequence data. *Am. J. Bot.* **2006**, *93*, 637–655.

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.

6. 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. *Evidence-Based Complement. Altern. Med.* **2016**, *2016*.

7. 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.

8. Gaire, B.P.; Subedi, L. Phytochemistry, pharmacology and medicinal properties of Phyllanthus emblica Linn. *Chin. J. Integr. Med.* **2014**, 1–8.

9. Lim, T.K. *Edible Medicinal And Non-Medicinal Plants*; Springer Netherlands: Dordrecht, 2012; ISBN 978-94-007-1763-3.

10. Mandal, A. a Review on Phytochemical, Pharmacological and Potential Therapeutic Uses of Phyllanthus Emblica. *World J. Pharm. Res.* **2017**, *6*, 817–830.

11. 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.

12. 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.

13. Zhao, T.; Sun, Q.; Marques, M.; Witcher, M. Anticancer Properties of *Phyllanthus emblica* (Indian Gooseberry). *Oxid. Med. Cell. Longev.* **2015**, *2015*, 1–7.

14. Vorontsova, M.S.; Hoffmann, P. Revision of the genus leptopus (Phyllanthaceae, Euphorbiaceae sensu lato). *Kew Bull.* **2009**, *64*, 627–644.

15. 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.

16. 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*, 1–12.

17. 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.

18. Daniell, H.; Lin, C.-S.; Yu, M.; Chang, W.-J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* **2016**, *17*, 134.

19. Daniell, H. Transgene containment by maternal inheritance: effective or elusive? *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 6879–6880.

20. 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.

21. 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*, 1–18.

22. 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.

23. 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*.

24. Abdullah; Mehmood, F.; Heidari, P.; Ahmed, I.; Poczai, P. Pseudogenization of trnT-GGU in chloroplast genomes of the plant family Asteraceae. *bioRxiv* **2021**.

25. Alqahtani, A.A.; Jansen, R.K. The evolutionary fate of rpl32 and rps16 losses in the Euphorbia schimperi (Euphorbiaceae) plastome. *Sci. Rep.* **2021**, *11*, 7466.

26. Palmer, J.D. Comparative organization of chloroplast genomes. *Annu. Rev. Genet.* **1985**, *19*, 325–354.

27. Ahmed, I. Evolutionary dynamics in taro, Massey University, Palmerston North, New Zealand, 2014.

28. 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.

29. 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.

30. 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.

31. 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*.

32. Abdullah; Henriquez, C.L.; Mehmood, F.; Hayat, A.; Sammad, A.; Waseem, S.; Tahir, M.; Matthews, P.J.; Croat, T.B.; Poczai, P.; et al. Chloroplast genome evolution in the Dracunculus clade (Aroideae , Araceae). *Genomics* **2021**, *113*, 183–192.

33. 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.* **2007**, *104*, 19369–19374.

34. 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.

35. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760.

36. 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.

37. 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.

38. 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.

39. Lehwark, P.; Greiner, S. GB2sequin - A file converter preparing custom GenBank files for database submission. *Genomics* **2019**, *111*, 759–761.

40. 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**.

41. 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.

42. Amiryousefi, A.; Hyvönen, J.; Poczai, P. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* **2018**, *34*, 3030–3031.

43. 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.

44. 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.

45. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**.

46. Beier, S.; Thiel, T.; Münch, T.; Scholz, U.; Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **2017**, *33*, 2583–2585.

47. 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.

48. 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.

49. 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.

50. 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.

51. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313.

52. Miller, M.A.; Pfeiffer, W.; Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In Proceedings of the 2010 Gateway Computing Environments Workshop, GCE 2010; 2010.

53. Letunic, I.; Bork, P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **2019**, *47*, W256–W259.

54. Cai, J.; Ma, P.F.; Li, H.T.; Li, D.Z. Complete plastid genome sequencing of four tilia species (Malvaceae): A comparative analysis and phylogenetic implications. *PLoS One* **2015**, *10*, 1–13.

55. Amiryousefi, A.; Hyvönen, J.; Poczai, P. The chloroplast genome sequence of bittersweet (Solanum dulcamara): Plastid genome structure evolution in Solanaceae. *PLoS One* **2018**, *13*, 1–23.

56. Nock, C.J.; Waters, D.L.E.; Edwards, M.A.; Bowen, S.G.; Rice, N.; Cordeiro, G.M.; Henry, R.J. Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnol. J.* **2011**, *9*, 328–333.

57. 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.

58. 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.

59. 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*, 1–11.

60. 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.

61. 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.

62. 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.

63. 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 infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. *Plant Cell* **2001**, *13*, 645–58.

64. Abdullah; Waseem, S.; Mirza, B.; Ahmed, I.; Waheed, M.T. Comparative analyses of chloroplast genomes of Theobroma cacao and Theobroma grandiflorum. *Biologia (Bratisl).* **2020**, *75*, 761–771.

65. 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*, 1–25.

66. 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*, 1–15.

67. 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*.

68. 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.

69. 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.

70. 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.

71. 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.

72. 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.

73. McDonald, M.J.; Wang, W.C.; Huang, H. Da; Leu, J.Y. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences.

PLoS Biol. **2011**, *9*.

74. 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.

75. Abdullah; Henriquez, C.L.; Croat, T.B.; Poczai, P.; Ahmed, I. Mutational dynamics of aroid chloroplast genomes II. *Front. Genet.* **2021**, *11*, 610838.

76. 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**, 1–17.

77. 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.

78. Kazakoff, S.H.; Imelfort, M.; Edwards, D.; Koehorst, J.; Biswas, B.; Batley, J.; Scott, P.T.; Gresshoff, P.M. Capturing the Biofuel Wellhead and Powerhouse: The Chloroplast and Mitochondrial Genomes of the Leguminous Feedstock Tree Pongamia pinnata. *PLoS One* **2012**, *7*, e51687.

79. Chen, Z.; Grover, C.E.; Li, P.; Wang, Y.; Nie, H.; Zhao, Y.; Wang, M.; Liu, F.; Zhou, Z.; Wang, X.; et al. Molecular evolution of the plastid genome during diversification of the cotton genus. *Mol. Phylogenet. Evol.* **2017**, *112*, 268–276.

80. Sablok, G.; Amiryousefi, A.; He, X.; Hyvönen, J.; Poczai, P. Sequencing the plastid genome of giant ragweed (Ambrosia trifida, asteraceae) from a herbarium specimen. *Front. Plant Sci.* **2019**, *10*, 218.

81. Abdullah; Shahzadi, I.; Mehmood, F.; Ali, Z.; Malik, M.S.; Waseem, S.; Mirza, B.; Ahmed, I.; Waheed, M.T. Comparative analyses of chloroplast genomes among three *Firmiana* species: Identification of mutational hotspots and phylogenetic relationship with other species of Malvaceae. *Plant Gene* **2019**, *19*, 100199.

82. Wang, W.; Messing, J. High-Throughput sequencing of three Lemnoideae (duckweeds) chloroplast genomes from total DNA. *PLoS One* **2011**, *6*.

83. Morton, B.R. Neighboring base composition and transversion/transition bias in a

comparison of rice and maize chloroplast noncoding regions. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 9717–9721.

84. Morton, B.R.; Bi, I. V.; McMullen, M.D.; Gaut, B.S. Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. *Genetics* **2006**, *172*, 569–577.

85. Avni, R.; Nave, M.; Barad, O.; Baruch, K.; Twardziok, S.O.; Gundlach, H.; Hale, I.; Mascher, M.; Spannagl, M.; Wiebe, K.; et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **2017**, *357*, 93–97.

86. 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.

87. Kathriarachchi, H.; Hoffmann, P.; Samuel, R.; Wurdack, K.J.; Chase, M.W. Molecular phylogenetics of Phyllanthaceae inferred from five genes (plastid atpB, matK, 3′ndhF, rbcL, and nuclear PHYC). *Mol. Phylogenet. Evol.* **2005**, *36*, 112–134.

88. Li, X.; Yang, Y.; Henry, R.J.; Rossetto, M.; Wang, Y.; Chen, S. Plant DNA barcoding: From gene to genome. *Biol. Rev.* **2014**.

89. 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.

 doi:10.20944/preprints202107.0094.v1

Figure Ligands

Figure 1. Circular diagram of *Phyllanthus emblica* chloroplast genome. The genes transcribed clockwise are shown inside the circle, whereas genes transcribed anti-clockwise are shown outside the circle. The circular diagram is shown as representative of all other species as difference exists only related to *rps*16 and inverted repeat contraction and expansion which is described within the manuscript.

Figure 2. Mauve alignment of the species of family Phyllanthaceae show high similarities in gene arrangement and gene content. The small blocks represent genes: Red = rRNA; green = tRNA with intron; Black = tRNA without intron; white = protein-coding genes

Figure 3. Contraction and expansion of inverted repeats at the junction of chloroplast genome. JLB: LSC/IRb; JSB: IRb/SSC; JSA: SSC/IRa; JLA: IRa/LSC.

Figure 4. Heatmap representing the relative synonymous codon usage. The y-axis represents 61 amino acid coding codons, and the x-axis represents species names.

Figure 5. 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 acid.

Figure 6. Comparison of simple sequence repeats and oligonucleotide repeats among the species of 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 repeats; R = reverse; C = complementary; P = palindromic

Figure 7. Comparison of transition and transversion substitutions. (a) the comparison of transition and transversion substitutions within 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 protein-coding sequences.

Figure 8. Maximum likelihood phylogenetic tree reconstructed with RAxML using dataset of 74 protein-coding genes. The 100 bootstrapping support was observed for all nodes. The deletion/pseudogenization of *rps*16 and loss of *atp*F intron are indicated with the symbols.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

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

Category for gene	Group of gene	Name of gene					Number
Photosynthesis-related genes	Photosystem I	*psaA*	*psaB*	*psaC*	*psaI*	*psaJ*	5
	Photosystem II	*psbA*	*psbK*	*psbI*	*psbM*	*psbD*	15
		psbF	*psbC*	*psbH*	*psbJ*	*psbL*	
		psbE	*psbN,*	*psbB*	*psbT*	*psbZ*	
	Cytochrome b/f compelx	*petN*	*petA*	*petL*	*petG*	*petD**	8
		petB*					
	ATP synthase	*atpI*	*atpH*	*atpA*	*atpF*	*atpE*	6
		atpB					
	Cytochrome c-type synthesis	*ccsA*					1
	Assembly/stability of photosystem I	*ycf3***	*ycf4*				2
	NADPH dehydrogenase	*ndhB*^{*,a,}	*ndhH,*	*ndhA**	*ndhI*	*ndhG*	12
		ndhJ	*ndhE*	*ndhF*	*ndhC*	*ndhK*	
		ndhD					
	Rubisco	*rbcL*					1
Transcription and translation related genes RNA genes	Transcription Small subunit of ribosome	*rpoA*	*rpoC2*	*rpoC1**	*rpoB*	*rps16*$^{*,£}$	5
		rps7^{a}	*rps15*	*rps19*§	*rps3*	*rps8*	13
		rps14	*rps11*	*rps12*^{a,*}	*rps18*	*rps4*	
		rps2					
	Large subunit of ribosome	*rpl2*$^{a,*,££}$	*rpl23*^{a,}	*rpl32*	*rpl22,*	*rpl14*	11
		rpl33	*rpl36*	*rpl20*	*rpl16**		
	Ribosomal RNA	*rrn16*^{a},	*rrn4.5*^{a},	*rrn5* ^{a,}	*rrn23*^{a}		8
	Transfer RNA	*trnV-GAC*^{a}	*trnI-CAU*^{a}	*trnA-UGC*^{a,*}	*trnN-GUU*^{a}	*trnP-UGG*	37
		trnW-CCA	*trnV-UAC**	*trnL-UAA**	*trnF-GAA*	*trnR-ACG*^{a}	
		trnT-UGU	*trnG-UCC**	*trnT-GGU*	*trnR-UCU*	*trnE-UUC*	
		trnY-GUA	*trnD-GUC*	*trnC-GCA*	*trnS-GCU*	*trnH-GUG*	
		trnK-UUU*	*trnQ-UUG*	*trnfM-CAU*	*trnG-GCC*	*trnS-UGA*	
		trnS-GGA	*trnL-UAG*	*trnM-CAU*	*trnL-CAA*^{a}	*trnI-GAU*^{*,a}	
Other genes	RNA processing	*matK*					1
	Carbon metabolism	*cemA*					1
	Fatty acid synthesis	*accD*					1
	Proteolysis	*clpP***					1
	Component of TIC complex	*ycf1*					1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

	Hypothetical proteins	*ycf2*[a]					2
Total							131

[*] Gene with one intron, [**] Gene with two introns, [a] Gene with two copies, [£]Gene pseudo in some species, [$] single copy exist of *rps*19 exist in *Glochiodion chodoense*, [££]gene with single functional copy in *Sauropus spatulifolius,*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

Table 2. Genomic features of the chloroplast genomes in the family Phyllanthaceae

Characteristics		*Baccaurea ramiflora*	*Breynia fruticosa*	*Flueggea virosa*	*Glochiodion 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 genes		111 (130*)	112 (131)	112 (131)	112 (130)	112 (130)	112 (130)	112 (131)	112 (131)	112 (130)
Protein-coding genes		77 (83)	78 (84)	78 (84*)	78 (84*)	77 (83*)	78 (84*)	78 (84*)	78 (84*)	78 (84*)
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[b]	19[a]	18[a]	19[a]	19[a]	18[b]
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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

	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] This also include pseudogene or truncated copy of the functional gene; The bracket indicates the total number of genes; *Sequence and assembled in current study, ** assembled from raw reads of NCBI in current study;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

1 Table 3. The polymorphic protein coding sequences

Gene	Total number of mutations	Alignment length	Alignment length without Indels	Nucleotide diversity	Missing data
rpl22	113	513	372	0.10887	27.49
ycf1	1422	5865	5449	0.08388	7.09
matK	342	1541	1521	0.06901	1.30
ndhF	417	2201	2105	0.06127	4.36
rps15	52	291	261	0.06117	10.31
rpl20	66	354	354	0.05841	0
ccsA	170	972	957	0.05717	1.54
rps3	105	687	645	0.05078	6.11
rps8	67	405	405	0.05018	0
rpl16	64	411	408	0.05007	0.73
ndhD	239	1524	1520	0.05007	0.26
accD	224	1524	1458	0.04774	4.33
cemA	66	477	477	0.0456	0
ycf4	77	563	549	0.04417	2.49
rps11	58	417	417	0.04368	0

2

3

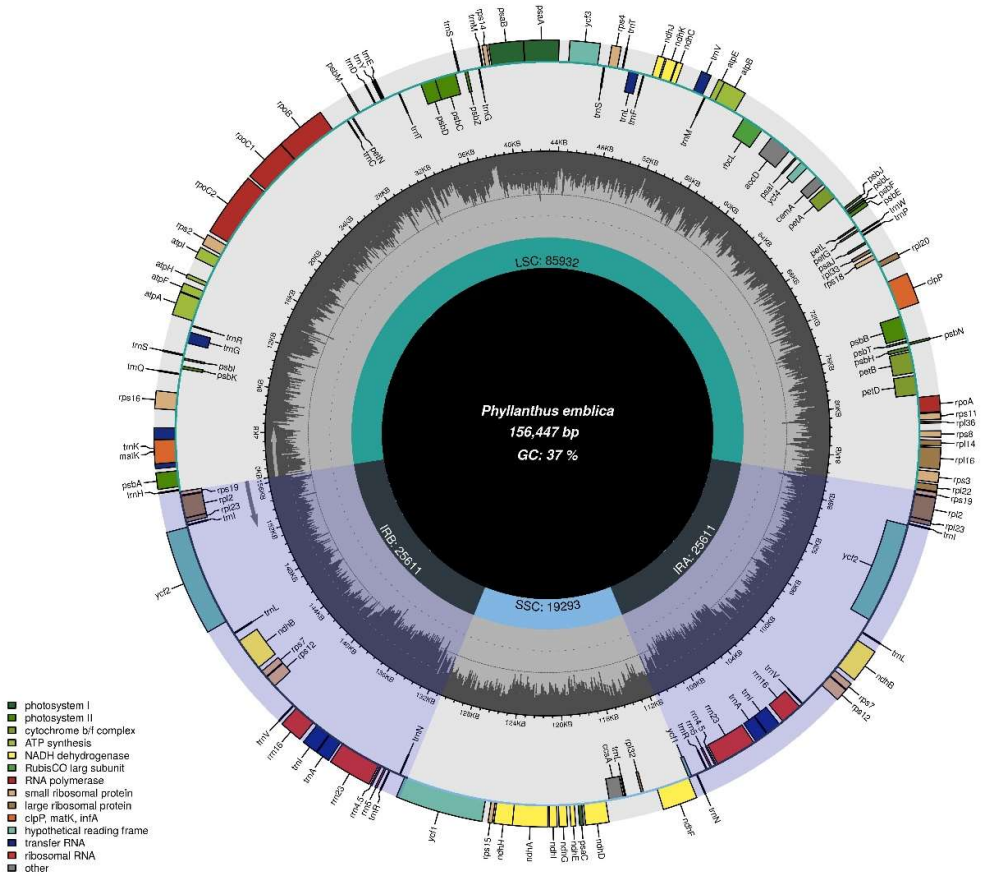

Phyllanthus emblica
156,447 bp
GC: 37 %
LSC: 85932
IRB: 25611
IRA: 25611
SSC: 19293
photosystem I
photosystem II
cytochrome b/f complex
ATP synthesis
NADH dehydrogenase
RubisCO larg subunit
RNA polymerase
small ribosomal protein
large ribosomal protein
clpP, matK, infA
hypothetical reading frame
transfer RNA
ribosomal RNA
other

Figure 1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

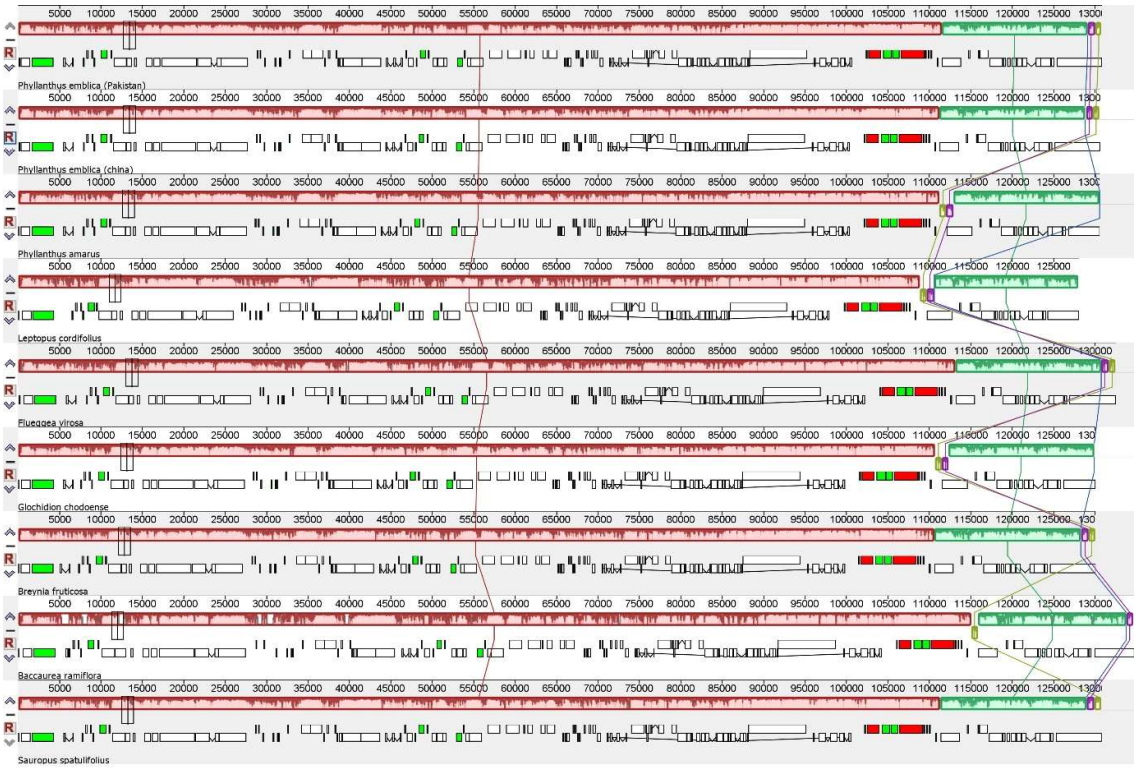


Figure 2

 doi:10.20944/preprints202107.0094.v1

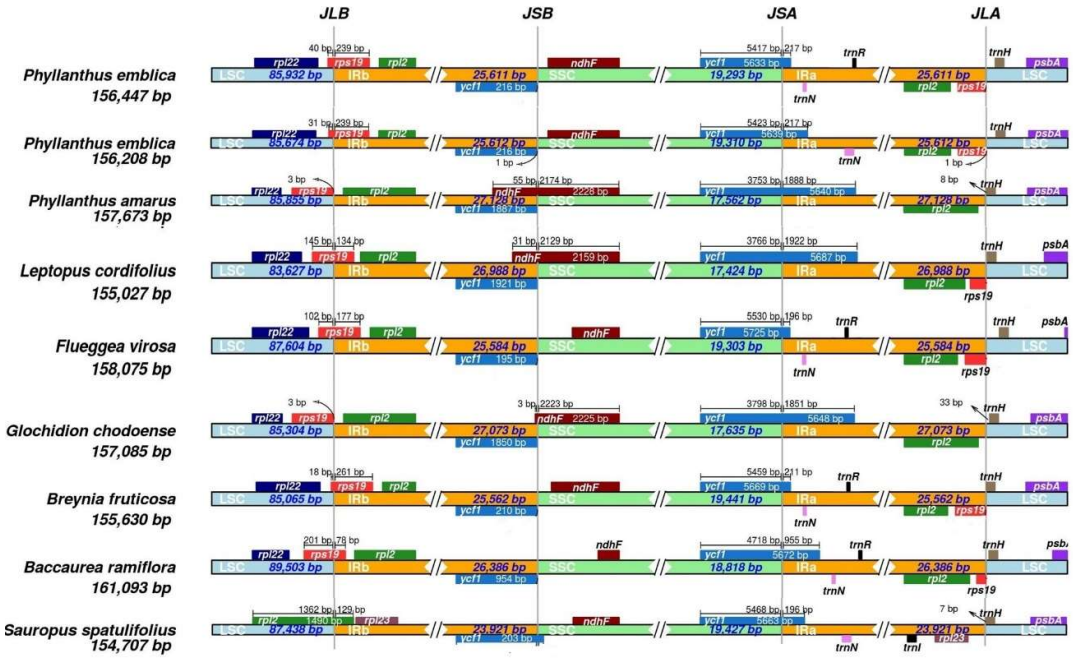


Figure 3

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

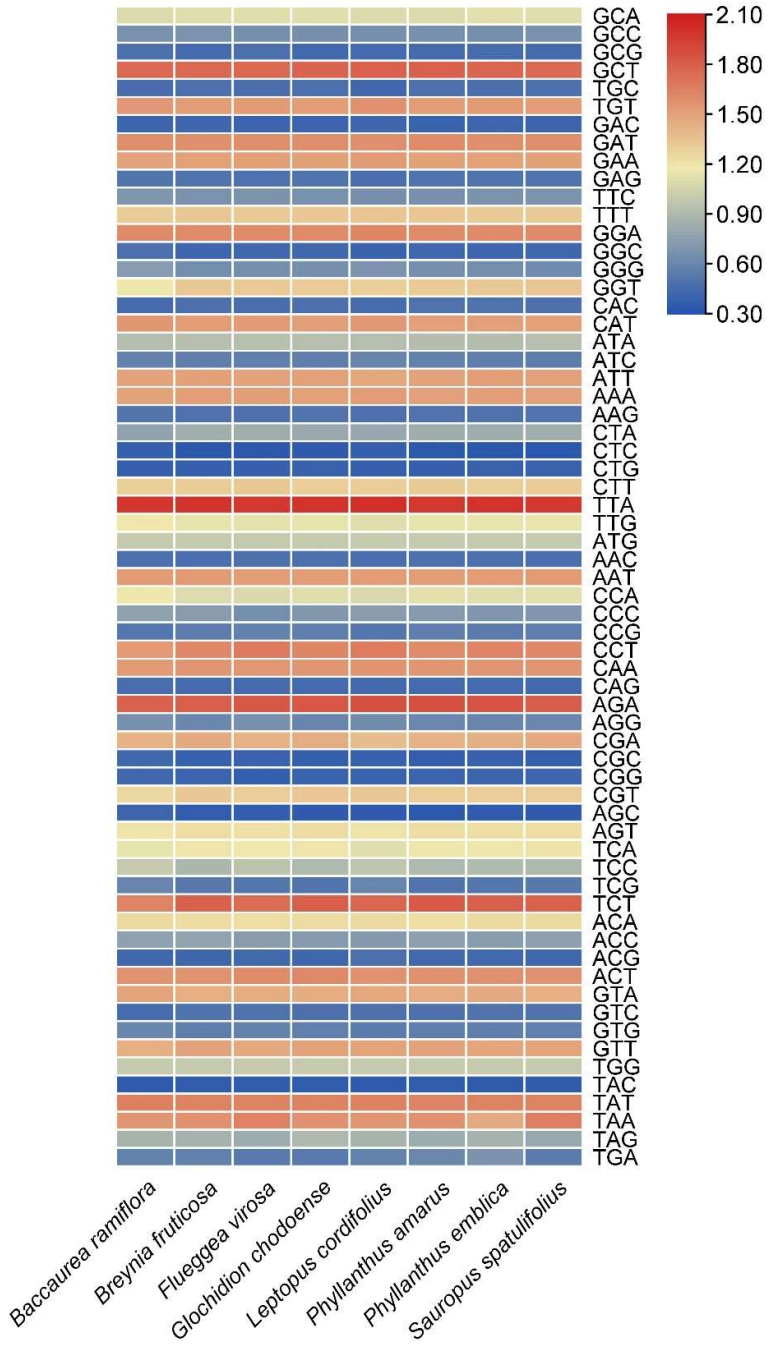


Figure 4

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

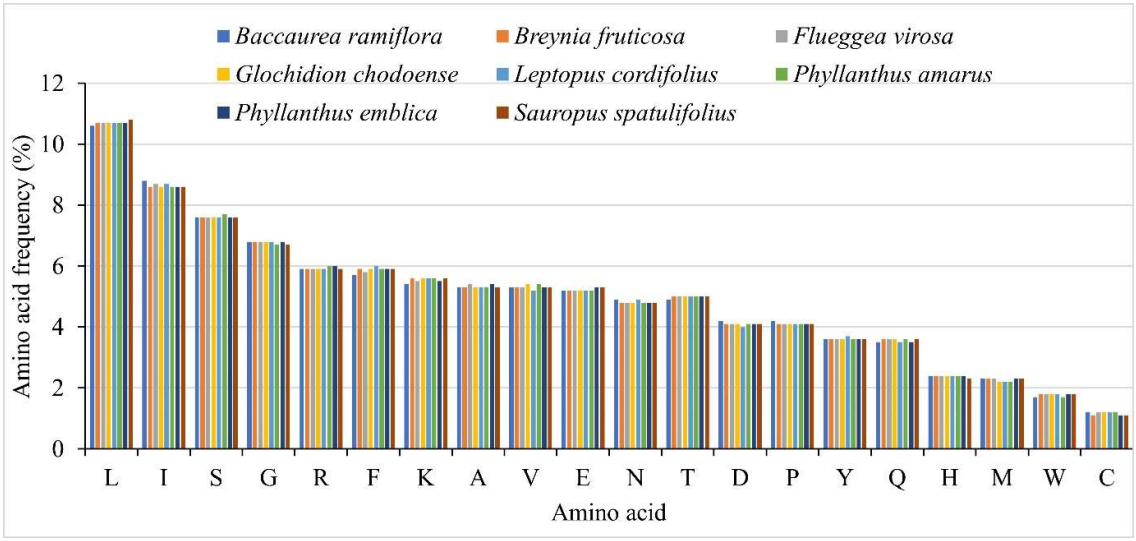


Figure 5

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

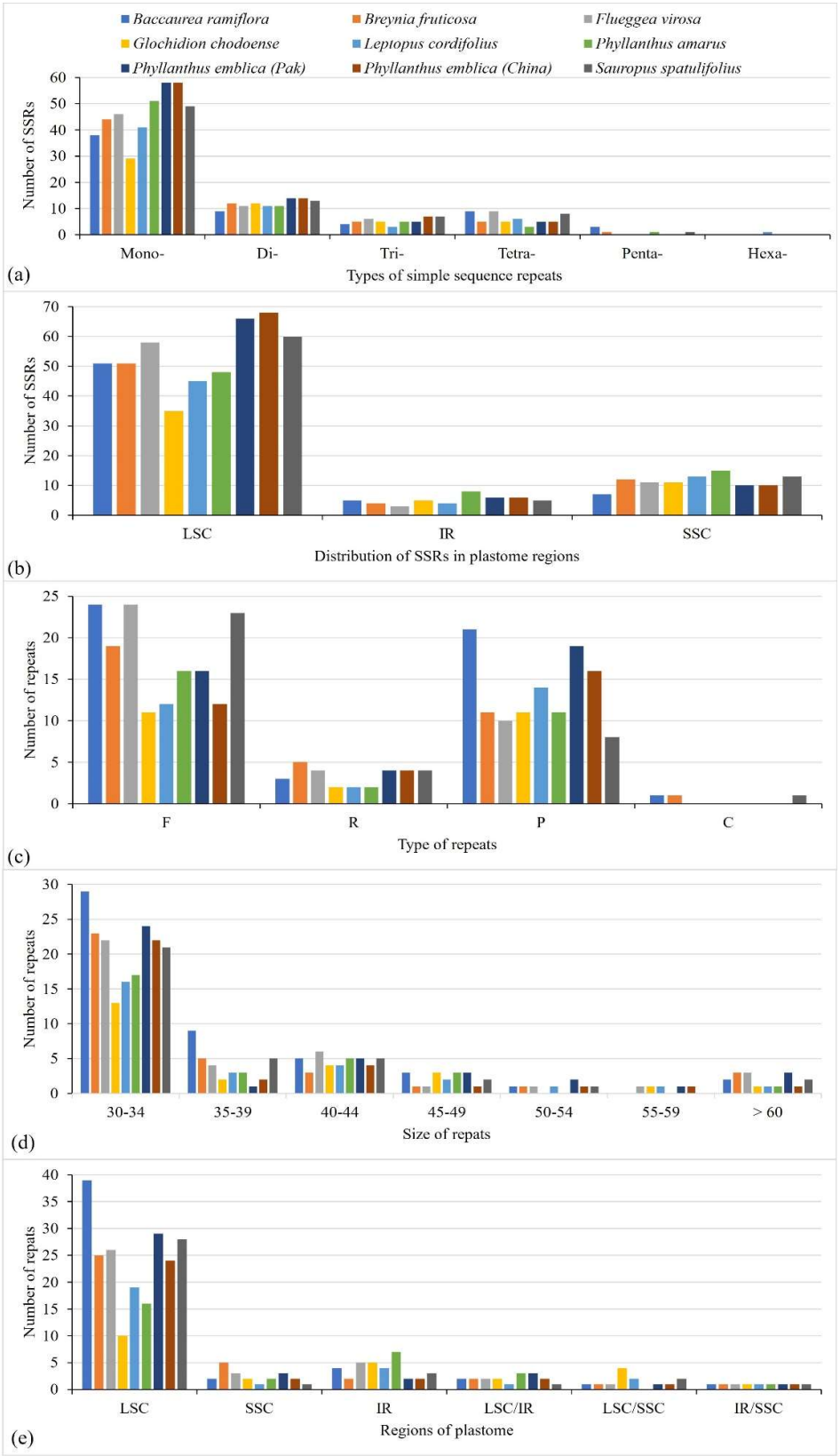


Figure 6

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2021 doi:10.20944/preprints202107.0094.v1

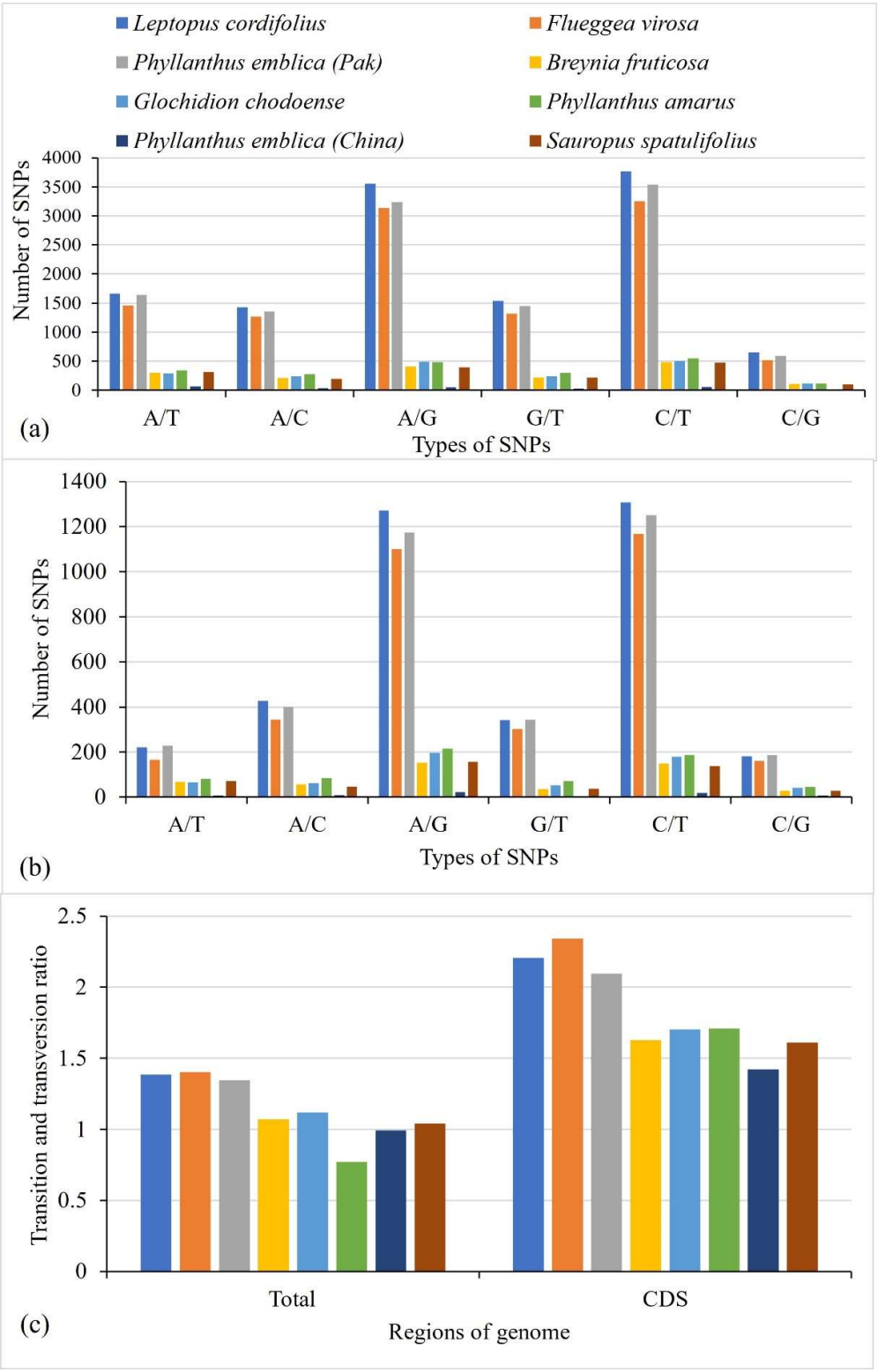
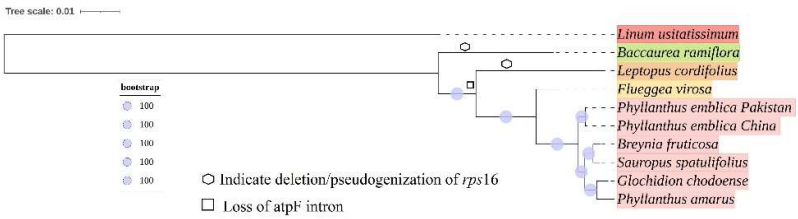


Figure 7

 doi:10.20944/preprints202107.0094.v1



27

28 Figure 8