Comparison among the first representative chloroplast genomes of *Orontium*, *Lasia*, *Zamioculcas*, and *Stylochaeton* of the plant family Araceae: inverted repeat dynamics are not linked to phylogenetic signaling

Abdullah[1], Claudia L. Henriquez[2], Furrukh Mehmood[1,4], Iram Shahzadi[1], Zain Ali[1,5], Mohammad Tahir Waheed[1], Thomas B. Croat[3], Peter Poczai[*,4], Ibrar Ahmed[*,5]

[1]*Department of Biochemistry, Faculty of Biological Sciences, Quaid-i-Azam University, 45320, Islamabad, Pakistan*

[2]*University of California, Los Angeles, Department of Ecology and Evolutionary Biology*

[3]*Missouri Botanical Garden, St. Louis, MO*

[4]*Finnish Museum of Natural History, University of Helsinki, PO Box 7 FI-00014 Helsinki Finland*

[5]*Alpha Genomics Private Limited, Islamabad, 45710, Pakistan*

**corresponding author:*

Péter Poczai (peter.poczai@helsinki.fi)

Ibrar Ahmed (iaqureshi_qau@yahoo.com)

Abstract

The chloroplast genome provides insight into the evolution of plant species. We *de novo* assembled and annotated chloroplast genomes of the first representatives of four genera representing three subfamilies: *Lasia spinosa* (Lasioideae), *Stylochaeton bogneri*, *Zamioculcas zamiifolia* (Zamioculcadoideae), and *Orontium aquaticum* (Orontioideae), and performed comparative genomics using the plastomes. The size of the chloroplast genomes ranged from 163,770–169,982 bp. These genomes comprise 114 unique genes, including 80 protein-coding, 4 rRNA, and 30 tRNA genes. These genomes exhibited high similarities in codon usage, amino acid frequency, RNA editing sites, and microsatellites. The junctions JSB (IRb/SSC) and JSA (SSC/IRa) are highly variable, as is oligonucleotide repeats content among the genomes. The patterns of inverted repeats contraction and expansion were shown to be homoplasious and therefore unsuitable for phylogenetic analyses. Signatures of positive selection were shown for several genes in *S. bogneri*. This study is a valuable addition to the evolutionary history of chloroplast genome structure in Araceae.

Key words:

Araceae, chloroplast genome, substitutions, gene evolution, IR contraction and expansion, phylogenetics

1. Introduction

The chloroplast is an important double membrane bounded organelle that plays a crucial role in photosynthesis and in the synthesis of fatty acids and amino acids [1]. The chloroplast contains its own DNA, and replicates independently from the nuclear genome [1,2]. Chloroplast genomes mostly exhibit a quadripartite structure in which a pair of inverted repeats (IRa and IRb) separate large single copy (LSC) and small single copy (SSC) regions [1–4]. However, in some plant lineages, quadripartite structure is not observed due to the loss of one or two inverted repeats (IRs), for example in Taxodiaceae [5] and Fabaceae [6]. On the other hand, very short IRs are also reported, for example in Pinaceae [7]. Moreover, a mixture of linear and circular chloroplast genomes has also been observed [8].

The structure of chloroplast genomes is conserved regarding gene organization, gene content and intron content [1,9–12]. However, large scale events of gene rearrangement, gene loss/generation of pseudogenes, and intron loss are also reported in various plant lineages [10,13–17]. Inverted repeat contraction and expansion in chloroplast genomes create pseudogenes, cause gene duplication, or convert duplicate into single-copy genes [10,11]. Many other types of mutational events also take place within chloroplast genomes, including insertion-deletion (indels), substitutions, tandem repeat variations and variations in number and type of oligonucleotide repeats [11,18–20]. The single parent inheritance of the chloroplast genome – paternally in some gymnosperms and maternally in most angiosperms – along with adequate levels of polymorphism make it suitable for studies of evolution, domestication, phylogeography, population genetics, and phylogenetics [7,10,11,21–24].

Araceae is an ancient and large monocot plant family, comprising 114 genera and 3750 species [25]. This family is subdivided into eight subfamilies: Gymnostachydoideae, Orotioideae, Lemnoideae, Pothoideae, Monsteroideae, Lasioideae, Zamioculcadoideae, and Aroideae [22,26,27]. Advancements in next generation sequencing have made genomic resources available for species of Lemnoideae [28], Monsteroideae [18,29,30], and Aroideae [10,31,32]; chloroplast genomes of *Symplocarpus renifolius* Schott ex Tzvelev and *S. nipponicus* Makino of Orotioideae have also been reported [33,34]. Previous studies provide insight into some unique evolutionary events of chloroplast genomes in Araceae, including IR contraction and expansion, gene rearrangement, and positive selection [10,18,28]. Moreover, loss/pseudogenation of some important genes are also reported in the genus *Amorphophallus* Blume (Aroideae) [17]. Nevertheless, data for the chloroplast genome

structure and evolution is still lacking for several major aroid clades, including the subfamilies Lasioideae and Zamioculcadoideae. Hence, new genomic resources are required to provide better insight into the evolution of chloroplast genomes in Araceae.

In the current study, we report *de novo* assembled and fully annotated chloroplast genomes of four species from three different subfamilies: *Lasia spinosa* (L.) Thwaites (Lasioideae), *Stylochaeton bogneri* Mayo and *Zamioculcas zamiifolia* (Lodd.) Engl. (Zamioculcadoideae), and *Orontium aquaticum* L. (Orontioideae). We performed comparative chloroplast genomics among these species. This study provides the first insights into the chloroplast genome structure of four genera of Araceae, along with the evolutionary rate of protein-coding genes and types of substitutions occurring within these chloroplast genomes.

2. Materials and Methods

2.1 Sample collection, DNA extraction and sequencing

We collected fresh and healthy leaves of four species (*L. spinosa*, *S. bogneri*, *Z. zamiifolia*, and *O. aquaticum*) from the Araceae Greenhouse at the Missouri Botanical Garden in St. Louis, Missouri. Whole genomic DNA was extracted from the collected leaves using Qiagen DNeasy Minikit (Qiagen, Germantown, Maryland, USA) with some modification following a previous approach [10,18]. DNA quality and quantity were confirmed by 1% agarose gel electrophoresis and Nanodrop (ThermoScientific, Delaware, USA). The libraries were constructed following manufacturer's protocol of Illumina TruSeq kits (Illumina, Inc., San Diego, California) in the Pires laboratory at the University of Missouri, Columbia. The Illumina HiSeq 2000 platform was used to sequence qualified libraries from single end with 100 bp short reads at the University of Missouri DNA Core.

2.2 Genome assembly and annotation

The sequencing of these genomes generated 3.31 GB (*S. bogneri*) to 11.3 GB (*Z. zamiifolia*) of raw data (Table 1). The quality of the generated short read data was compared among species using FastQC and MultiQC [35,36]. The analyses confirmed high quality of the data with high average Phred score ranging from 35.69–37.6. The raw data of the sequenced four species were submitted to the Sequence Read Archive of the National Center for Biotechnology (NCBI) under SRA project number PRJNA613281. The generated sequence data were used to *de novo* assemble chloroplast genomes using Velvet v.1.2.10 [37] by generating contigs with various kmer values of 51, 61, 71, and 81, combined with the *de novo*

assembly option of Geneious R8.1 [38] following previous studies [9,11,39]. The coverage depth analysis was performed by mapping the short reads to their respective *de novo* assembled chloroplast genomes by BWA mem [40]. The assembly of the genomes was then validated by visualizing in Tablet [41]. We observed issues at 4–5 points of repetitive regions, therefore, for further validation we used Fast-Plast v.1.2.2 following exactly the same procedure employed for the assembly of other Araceae species [10,18]. This helped us to corroborate the correct sequence at those points. The coverage depth analyses revealed that the average coverage depths of the genomes ranged from 92.7X–1021X. The *de novo* assembled chloroplast genomes were annotated by GeSeq [42], whereas tRNA genes were further verified by tRNAscan-SE v.2.0.3 [43] and ARAGORN v.1.2.38 [44] by selecting default parameters. The final annotated genomes were submitted to NCBI under specific accession numbers (Table 1). GB2sequin was used to generate five column tab-delimited files from the annotated genomes for NCBI submission [45]. The circular map of these genomes was drawn by OrganellarGenomeDRAW (OGDRAW) [46].

2.3 Characterization, comparative analyses, and phylogenetic inference

We used Geneious R8.1 [38] to compare genomic features and determine amino acid frequency and codon usage. To visualize and compare the junctions of chloroplast genomes, we used IRscope with default parameters [47]. The integrated Mauve alignment [48] of Geneious R8.1 was used to analyze gene arrangement based on Colinear block analyses after removal of IRa from the genomes.

The Predictive RNA editors for Plant (PREP-CP) [49] program was used to determined RNA editing sites in the chloroplast genomes. We also analyzed microsatellites and oligonucleotide repeats using MISA (MIcroSAtellite) and REPuter, respectively. We determined microsatellites with repeat units as: mononucleotide repeats ≥ 10, dinucleotide ≥ 5, trinucleotide ≥ 4, tetranucleotide, pentanucleotide and hexanucleotide ≥ 3. The forward and reverse oligonucleotide repeats were determined with length ≥ 14 bp with 1 editing site, initially. Later, we removed all repeats that contained mismatches from the analyses and left only those repeat pairs that exhibited 100% similarities, following Abdullah et al. 2020 [50].

We determined transition substitutions (Ts), transversion substitutions (Tv) and their ratio (Ts/Tv) in 78 protein-coding genes. For this purpose, we concatenated protein-coding genes of all four species. The sequences of the concatenated protein-coding genes of *L. spinosa*, *S. bogneri*, and *Z. zamiifolia* were pairwise aligned to *O. aquaticum* by Multiple Alignment

using Fast Fourier Transform (MAFFT). The substitution types were determined from each alignment in Geneious R8.1 [38].

We determined the rate of synonymous substitutions (K_s), non-synonymous substitutions (K_a) and their ratio (Ka/Ks) in 75 protein-coding genes. We extracted and aligned protein-coding genes from all four species. The chloroplast genome of *O. aquaticum* was used as a reference, and the rates of evolution of protein-coding genes were recorded. A similar approach was previously applied in other angiosperms [9,11,18,39,51]. The data were interpreted as: $K_a/K_s < 1$, $K_a/K_s = 1$, $K_a/K_s > 1$, representing purifying, neutral and positive selection, respectively.

A phylogenetic analysis was performed among 30 species of Araceae, using *Acorus americanus* (Acoraceae) as an outgroup. MAFFT [52] on XSEDE v.7.402 in CIPRES [53] was used to align complete chloroplast genomes of all species after removal of one IR. The phylogeny was inferred based on this alignment after removal of indels using RAxML-HPC BlackBox v.8.2.12 [54] in CIPRES [53]. The details regarding the species that were used in the phylogenetic analysis are shown in table S1.

3. Results

3.1 Comparative genomics among *de novo* assembled chloroplast genomes

The sizes of the genomes ranged from 163,770 bp (*S. bogneri*) to 169,980 bp in *L. spinose*. The SSC region ranged from 13,967 bp (*O. aquaticum*) to 20,497 bp (*S. bogneri*); LSC ranged from 87,269 bp (*O. aquaticum*) to 91,357 bp (*Z. zamiifolia*); the size of each IR region ranged from 26,702 bp (*S. bogneri*) to 32,053 bp (*L. spinosa*) (Table 2). The chloroplast genomes of the four species were found to be highly conserved in terms of gene organization, gene content and intron content. This highly conserved structure was also confirmed using circular maps of the genomes (Figure 1), as well as from the Colinear Block analyses of Mauve (Figure 2). All species exhibited 114 unique genes, including 80 protein-coding, 30 tRNA, and 4 rRNA genes. We recorded 17 duplicated genes in the IRs of *S. bogneri* and *Z. zamiifolia*, and 18 duplicated genes in *O. aquaticum* and *L. spinosa*. In total, 18 intron-containing genes were observed, including 6 tRNA genes and 12 protein-coding genes. Among the intron-containing genes, 2 tRNA genes and 3 protein-coding genes are located in IRs. The size of introns showed some variation among species, whereas exons showed high similarity (Table S2). The *inf*A gene was found to be a pseudogene in all species. The GC content of the complete chloroplast genomes and of all regions showed high similarities among species, whereas fluctuation in GC content was observed within the different regions

of the same chloroplast genome. The GC content of coding regions, rRNAs, and tRNAs also showed high similarities among species (Table 2).

3.2 Inverted repeats contraction and expansion

The chloroplast genomes showed variations and similarities at the junctions of LSC/IR/SSC. The junctions of JLB (LSC/IRb) and JLA (IRa/LSC) showed similarities across all four species. The junctions of JSB and JSA were highly variable among species. The chloroplast genomes *O. aquaticum* and *L. spinosa* were found to be similar at these two junctions, and IR expansion led to duplication of the complete *ycf*1 gene and the origin of pseudogenes of *rps*15 at JSB. The chloroplast genomes of *S. bogneri* and *Z. zamiifolia* showed less expansion of IRs, which led to the origin of only a pseudogene of *ycf*1 at JSB. The integration of *ndh*F into IRb was only recorded in *L. spinosa*. At JLA in *O. aquaticum*, *trnH-GUG* was found to be completely in the LSC region 12 bp away from the junction, whereas other species showed integration of *trn*H-GUG into the IRa region from 6 bp to 11 bp. The complete details are presented in Figure 3.

3.3 Analyses of codon usage, amino acid frequency and RNA editing

The codon usage analyses revealed high encoding efficacy for those codons that end with A/T as opposed to codons that end with C/G. We recorded a relative synonymous codon usage (RSCU) value ≥ 1 for most codons that end with A/T, whereas RSCU < 1 was recorded for codons that end with C/G (Table S3). The ATG codon is the most common start codon. However, we also observed ACG (in *rpl*2) and GTG (in *rps*19) as start codons. The amino acid frequency analyses revealed high encoding of leucine, whereas the rarest encoding was recorded for cysteine (Figure S1). The RNA editing analyses revealed the presence of 62–74 RNA editing sites in 19–21 genes (Table S4). The RNA editing sites were found in the same genes with a few exceptions: RNA editing sites were detected in *psa*B genes of only *Z. zamiifolia*, whereas the RNA editing site was not found in *rpl*20 of *S. bogneri* and *rps*8 of *L. spinosa*. Most of the RNA editing sites were recorded in *ndh*A, *ndh*B, *ndh*D, *rpo*A, *rpo*B, *rpo*C1, and *rpo*C2 (Table S4). ACG was found as a start codon in gene *rpl2* and the RNA editing analyses confirmed conversion of ACG codon to ATG. Most RNA editing sites were found to be related to conversion of serine to leucine. Moreover, almost all editing sites led to accumulation of hydrophobic amino acids in the polypeptide chain of proteins (Table S4).

3.4 Repeats analyses

The analyses of microsatellites revealed 104–146 repeats in the genomes. Most of the repeats existed in LSC followed by SSC and then IR (Fig. 4a). We recorded an abundance of mononucleotide repeats in all species, especially *Z. zamiifolia*. Dinucleotide repeats were in greater abundance in *O. aquaticum* and *L. spinosa*, whereas *S. bogneri* and *Z. zamiifolia* showed an abundance of mononucleotide repeats and tetranucleotide repeats. *Lasia spinosa*, *Z. zamiifolia* and *O. aquaticum* showed similarities in numbers of tri and tetranucleotide repeats in their respective genomes, but *S. bogneri* showed a lack of trinucleotide repeats relative to tetranucleotide repeats (Fig. 4b). Pentanucleotide and hexanucleotide repeats were in lower abundance than the other types of repeats and were completely lacking in Z. zamiifolia (Fig. 4a). Most repeats of all six microsatellite types were the A/T rather than G/C motif (Table S5). The analyses of oligonucleotide repeats revealed the existence of a higher number of forward and reverse repeats in all four species. We recorded most repeats in the LSC region as compared to the SSC and IR regions. We also found some shared repeats among the three regions of the chloroplast genomes (Figure 4c). The number of repeats ranged from 647 (*O. aquaticum*) to 1471 (*Z. zamiifolia*). We recorded high similarities in the numbers of forward and reverse repeats in *O. aquaticum* and *L. spinosa*, whereas in *S. bogneri* and *Z. zamiifolia* there was a higher abundance of forward repeats (Figure 4d). Most repeats ranged in length of 14–20 bp (Figure 4d), whereas the largest repeats varied from 39 bp (*L. spinosa*) to 75 bp (*S. bogneri*) (Figure 4e). The details about the position and number of repeats are provided in Table S6.

3.5 Analyses of substitutions types

We recorded a greater number of Ts than Tv substitutions. The ratio of Ts/Tv was 2.3, 2.03, and 2.15 in the genomes of *L. spinosa*, *S. bogneri*, and *Z. zamiifolia*, respectively. The majority of Ts were promoted by A/G rather than C/T, whereas the majority of Tv were found to be related to A/C and G/T rather than A/T and C/G (Table 3). For K_s and K_a, we found a higher average of K_s than K_a. Hence, on average, we recorded very low K_a/K_s for all genes, which shows that purifying selection is acting on these genes. The average values recorded for the different groups of genes were: photosystem I group (K_s = 0.1677, K_a = 0.0125, and K_a/K_s = 0.1211), photosystem II (K_s = 0.1208, K_a = 0.0085, and K_a/K_s = 0.0671), cytochrome group (K_s = 0.1757, K_a = 0.0298, and K_a/K_s = 0.2012), ATP synthase group (K_s = 0.1466, K_a = 0.0188, and K_a/K_s = 0.1337), ribosomal small subunit group (K_s = 0.1620, K_a = 0.0633, and K_a/K_s = 0.3491), ribosomal large subunit (K_s = 0.1639, K_a = 0.0677, and K_a/K_s = 0.4612), NADPH dehydrogenase group (K_s = 0.1711, K_a = 0.0781, and K_a/K_s = 0.3935),

and RNA polymerase group (K_s = 0.1813, K_a = 0.3228, K_a/K_s = 0.1800) (Table S7). Some genes showed neutral selection (K_a/K_s = 1) in all species, including *ndh*K, *pet*L, *rpl*16, *ndh*F, *ndh*H, and *rps*15. Interestingly, we found evidence for positive selection in three genes (*ycf*2, *clp*P, *rpl*36) in only *S. bogneri* (Table S7).

3.6 Phylogenetic inference of Araceae

A phylogenetic tree was reconstructed using 31 species based on an alignment of 93,794 nucleotide sites, of which 13,488 were found to be parsimony informative; 8,843 showed distinct site pattern while the remaining sites (67,229 sites) were shared in all species. The resulting phylogeny shows the monophyly of Lasioideae, Zamioculcadoideae and Orontioideae, with Zamioculcadoideae forming a clade with *Stylochaeton*.

Discussion

In the current study, we report *de novo* assembled and fully annotated chloroplast genomes of four species from three subfamilies of Araceae. Comparative chloroplast genomics revealed high similarities in gene content across all species. However, the sizes of these genomes varied due to the variable length of intergenic spacer regions (IGS) and IRs contraction and expansion. The substitution analyses revealed Ts > Tv and $K_s > K_a$. The phylogenetic analysis confirmed the monophyly of Lasioideae, Orontioideae and Zamioculcadoideae.

The chloroplast genomes of the four species showed a highly conserved structure in terms of gene content, intron content and gene organization. Similar gene contents were also reported in other subfamilies of Araceae [10,18,28,34]. Moreover, highly conserved chloroplast genomes are reported in other angiosperms [10,18,55]. However, loss of some important protein-coding genes and tRNA genes has been reported in the genus *Amorphophallus* (Aroideae, Araceae), which might be specific to this genus. The *inf*A gene encodes translation initiation factor I, but we found this gene to be non-functional in all species. This gene is also reported to be non-functional or absent in the chloroplast genomes of other angiosperms including species of Araceae [9–11,18,28,55]. Hence, it is suggested that this gene is either transferred to the nuclear genome as an active functional gene, or a functional copy of this important gene might already exist in the nuclear genome [31,56]. We observed duplication of *ycf*1 genes or origination of pseudogenes of *ycf*1 and *rps*15 due to IR contraction and expansion. The duplication of *ycf*1 and/or *rps*15 is also reported in species of the subfamily Lemnoideae [28] and two species (*Anchomanes hookeri* Schott. and *Zantedeschia aethiopica* Spreng.) of Aroideae [10].

Together with previously published aroid chloroplast genomes, the chloroplast genomes of the current study reveal that IR contraction and expansion might be a species-specific event, opposed to synapomorphies of subfamilies. The duplication of *ycf*1 and origination of the *rps*15 pseudogene in *O. aquaticum* is not observed in species of *Symplocarpus* Salisb. ex Nutt. [34], and both genes are found in the SSC region. The complete duplication of *ycf*1 and *rps*15 is observed in Lemnoideae species [28]; in *S. bogneri* and *Z. zamiifolia*, duplication of partial *ycf*1 is observed, and *rps*15 completely exists in the SSC region. Moreover, complete duplication of *ycf*1 and partial duplication of *rps*15 was observed in *L. spinosa*, similar to *O. aquaticum*. This data suggests that IR contraction and expansion is highly flexible over evolutionary time, and that similar IR boundary architecture across lineages can be the result of homoplasy. In other angiosperms, differential IR contraction and expansion has also been seen in species of the same genus such as *Aquilaria* Lam. [57,58]. This observation is contradictory to a previous study in which resemblance of IR junctions was suggested for phylogenetic inference [59]. However, these authors did not provide conclusive results and suggested further study in wide range of germplasms.

The high RSCU value has also been previously reported for those codons that contain A/T at 3′ end instead of C/G, which might be due to the high AT content of the chloroplast genomes [3,11,18,55]. We found an abundance of leucine and isoleucine, whereas cysteine was found to be the rarest amino acid. These results are in agreement with previous studies of angiosperms, including the family Araceae [3,9,16,18,28]. We found high similarities in RNA editing sites. Most of the conversions by RNA editing led to the addition of hydrophobic amino acids to polypeptide chains of proteins. A similar pattern of conversion has been noted in the chloroplast genomes of other angiosperms [11,16,19]. We found ACG as a start codon for *rpl*2 instead of ATG, and RNA editing analyses confirmed the conversion of ACG to ATG. This is in agreement with a previous study [60].

Oligonucleotide repeats generate mutations in genomes [11,31,50,61]. Hence, these repeats are suggested to be used as a proxy to identify mutational hotspots [11,31,50]. In the current study, we analyzed forward and reverse repeats, as these repeats are considered important for the generation of mutations [11,31,50] and show up to 90% co-occurrence with substitutions as reported in the plant family Malvaceae [50]. Previously, high similarities were reported in repeat numbers and types in some angiosperms [3,4,9,11,19,55]. However, previous studies also show significantly variable number of repeats in other species of Araceae, which does not correlate with genome size or phylogenetic position [10,18]. Here, our result agrees with

previous studies of Araceae. Notably, penta- and hexanucleotide repeats were completely absent in *Z. zamiifolia*, while the number of mononucleotide repeats was almost twice as high as in the other genera. In the current study, we changed the criteria for repeat determination following a recent study of Abdullah et al. 2020 [50], and identified repeats of exact match ≥ 14. However, our result still agrees with a previous study in that repeats in the chloroplast genomes of Araceae are independent of genome size and phylogenetic position.

We observed a greater amount of Ts than Tv substitutions within the protein-coding genes , as expected in DNA sequences [62]. However, fewer Ts than Tv has also been reported previously in chloroplast genomes [9,39,63]. Higher Ts than Tv might occur due to genome composition and codon characteristics [64]. Moreover, higher Ts than Tv was also reported in the complete chloroplast genome of *Dioscorea polystachya* Turcz. [65] and in the coding sequences of the species of Lemnoideae (Araceae). This suggests that species of the plant family Araceae are consistent regarding the existence of higher Ts as compared to Tv.

We observed $K_a/K_s < 1$ due to higher K_s than K_a for most of the protein-coding genes. These results are consistent with previous studies of angiosperm chloroplast genomes, including the family Araceae, as purifying selection pressure mostly acts on the genes of chloroplast genomes [9,11,18,55]. However, a higher K_a/K_s was also reported in some species of Araceae in which most of the genes were under positive selection [34]. We also found three genes under positive selection in the chloroplast genome of *S. bogneri*, including *ycf*2, *clp*P, and *rpl*36, which might be due to the different types of stresses faced by these species in their respective ecological niches. These genes were also found to be under positive selection in various other species [11,16,55,66,67].

This study presents a comparison of the complete chloroplast genomes of four species representing three subfamilies of Araceae including Orontioideae, Lasioideae and Zamioculcadoideae, and the independent taxon *S. bogneri*. IR contraction and expansion appears to be homoplasious among these taxa, precluding the use of the IR architecture in phylogenetic analyses. Unique features of members of the *Stylochaeton* clade [26], represented here by *Z. zamiifolia* (Zamioculcadoideae) and *Stylochaeton* Lepr., include repeat types and amounts in both species , and signatures of positive selection in several genes in *Stylochaeton*.

Authors Contribution

Sample collection, DNA extraction and sequencing: CH and TC; Genomes assembly, coverage depth analyses and annotations: A, CH, and ZA; Data analyses: A, FM and IS; Data interpretation: A and FM; Conceptualization: A, PP, IA, MTW; Data curation: A, CH, FM; Project administration: A, CH; Writing – Original Draft: A; Writing - Review & Editing: CH, IA, PP, MTW; Supervision: PP and IA

Funding

Funding for this study was provided by the GAANN fellowship, the Rettner B. Morris Scholarship, Washington University in St. Louis, J. Chris Pires Lab (NSF DEB 1146603).

Acknowledgment

Authors are thankful for funding and laboratory support to Dr. Barbara Schaal at Washington University in St. Louis and Dr. J. Chris Pires at the University of Columbia, Missouri. Authors are also thankful to Dr. Tatiana Arias for valuable help in the laboratory and data processing. In the aroid greenhouse at the Missouri Botanical Garden, Emily Colletti provide help with living material and authors like to thank for it.

Conflict of interest

No conflict of interest exists.

References

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

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

3. 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.

4. Abdullah; Waseem, S.; Mirza, B.; Ahmed, I.; Waheed, M.T. Comparative analyses of chloroplast genome in *Theobroma cacao* and *Theobroma grandiflorum*. *Biologia (Bratisl).* **2019**.

5. Hirao, T.; Watanabe, A.; Kurita, M.; Kondo, T.; Takata, K. Complete nucleotide sequence of the *Cryptomeria japonica* D. Don. chloroplast genome and comparative chloroplast genomics: Diversified genomic structure of coniferous species. *BMC Plant Biol.* **2008**, *8*.

6. Sabir, J.; Schwarz, E.; Ellison, N.; Zhang, J.; Baeshen, N.A.; Mutwakil, M.; Jansen, R.; Ruhlman, T. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. *Plant Biotechnol. J.* **2014**, *12*, 743–754.

7. Zeb, U.; Dong, W.; Zhang, T.; Wang, R.; Shahzad, K.; Ma, X.; Li, Z. Comparative plastid genomics of *Pinus* species: Insights into sequence variations and phylogenetic relationships. *J. Syst. Evol.* **2019**, jse.12492.

8. Oldenburg, D.J.; Bendich, A.J. The linear plastid chromosomes of maize: terminal sequences, structures, and implications for DNA replication. *Curr. Genet.* **2016**, *62*, 431–442.

9. 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.

10. Henriquez, C.L.; Abdullah; Ahmed, I.; Carlsen, M.M.; Zuluaga, A.; Croat, T.B.;

McKain, M.R. Evolutionary dynamics in chloroplast genomes of subfamily Aroideae (Araceae). *Genomics* **2020**.

11. 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.

12. 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.

13. Schwarz, E.N.; Ruhlman, T.A.; Sabir, J.S.M.; Hajrah, N.H.; Alharbi, N.S.; Al-Malki, A.L.; Bailey, C.D.; Jansen, R.K. Plastid genome sequences of legumes reveal parallel inversions and multiple losses of *rps16* in papilionoids. *J. Syst. Evol.* **2015**, *53*, 458–468.

14. Rabah, S.O.; Shrestha, B.; Hajrah, N.H.; Sabir, M.J.; Alharby, H.F.; Sabir, M.J.; Alhebshi, A.M.; Sabir, J.S.M.; Gilbert, L.E.; Ruhlman, T.A.; et al. *Passiflora* plastome sequencing reveals widespread genomic rearrangements. *J. Syst. Evol.* **2019**, *57*, 1–14.

15. Lopes, A. de S.; Pacheco, T.G.; Santos, K.G. dos; Vieira, L. do N.; Guerra, M.P.; Nodari, R.O.; Souza, E.M. de; Pedrosa, F. de O.; Rogalski, M. The Linum usitatissimum L. plastome reveals atypical structural evolution, new editing sites, and the phylogenetic position of Linaceae within Malpighiales. *Plant Cell Rep.* **2018**, *37*, 307–328.

16. 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**, 100199.

17. Liu, E.; Yang, C.; Liu, J.; Jin, S.; Harijati, N.; Hu, Z.; Diao, Y.; Zhao, L. Comparative analysis of complete chloroplast genome sequences of four major *Amorphophallus* species. *Sci. Rep.* **2019**, *9*, 809.

18. 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.

19. 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.

20. 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.

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

22. 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.

23. 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.

24. Jansen, R.K.; Wojciechowski, M.F.; Sanniyasi, E.; Lee, S.-B.; Daniell, H. Complete plastid genome sequence of the chickpea *(Cicer arietinum)* and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). *Mol. Phylogenet. Evol.* **2008**, *48*, 1204–1217.

25. 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.

26. Cusimano, N.; Bogner, J.; Mayo, S.J.; Boyce, P.C.; Wong, S.Y.; Hesse, M.; Hetterscheid, W.L.A.; Keating, R.C.; French, J.C. Relationships within the Araceae: Comparison of morphological patterns with molecular phylogenies. *Am. J. Bot.* **2011**, *98*, 654–668.

27. Nauheimer, L.; Metzler, D.; Renner, S.S. Global history of the ancient monocot family Araceae inferred with models accounting for past continental positions and previous ranges based on fossils. *New Phytol.* **2012**, *195*, 938–950.

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

29. Liu, X.F.; Zhu, G.F.; Li, D.M.; Wang, X.J. The complete chloroplast genome sequence of *Spathiphyllum cannifolium*. *Mitochondrial DNA Part B Resour.* **2019**, *4*, 1822–

1823.

30. Han, L.; Wang, B.; Wang, Z.Z. The complete chloroplast genome sequence of *Spathiphyllum kochii*. *Mitochondrial DNA* **2016**, *27*, 2973–2974.

31. 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.

32. Han, L.; Chen, C.; Wang, B.; Wang, Z.-Z. The complete chloroplast genome sequence of medicinal plant Pinellia ternata. *Mitochondrial DNA. Part A, DNA mapping, Seq. Anal.* **2016**, *27*, 2921–2.

33. Choi, K.S.; Park, K.T.; Park, S. The Chloroplast Genome of Symplocarpus renifolius : A Comparison of Chloroplast Genome Structure in Araceae. *Genes (Basel).* **2017**, *8*, 324.

34. Kim, S.-H.; Yang, J.; Park, J.; Yamada, T.; Maki, M.; Kim, S.-C. Comparison of Whole Plastome Sequences between Thermogenic Skunk Cabbage *Symplocarpus renifolius* and Nonthermogenic *S. nipponicus* (Orontioideae; Araceae) in East Asia. *Int. J. Mol. Sci.* **2019**, *20*, 4678.

35. Andrews, S. FastQC: A Quality Control tool for High Throughput Sequence Data Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on Sep 15, 2019).

36. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **2016**, *32*, 3047–8.

37. Zerbino, D.R.; Birney, E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **2008**, *18*, 821–829.

38. 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.

39. 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.

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

41. 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.

42. 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.

43. 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.

44. Laslett, D.; Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **2004**, *32*, 11–16.

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

46. Greiner, S.; Lehwark, P.; Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res*. **2019**, *47*, W59–W64.

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

48. 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.

49. Mower, J.P. The PREP suite: Predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acids Res.* **2009**, *37*, W253–W259.

50. 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.* **2020**.

51. 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.

52. Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **2002**, *30*, 3059–66.

53. 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.

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

55. 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). *bioRxiv* **2020**.

56. 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.

57. Wang, Y.; Zhan, D.-F.; Jia, X.; Mei, W.-L.; Dai, H.-F.; Chen, X.-T.; Peng, S.-Q. Complete Chloroplast Genome Sequence of *Aquilaria sinensis* (Lour.) Gilg and Evolution Analysis within the Malvales Order. *Front. Plant Sci.* **2016**, *7*, 1–13.

58. Lee, S.Y.; Ng, W.L.; Mohamed, R.; Terhem, R. The complete chloroplast genome of *Aquilaria malaccensis* Lam. (Thymelaeaceae), an important and threatened agarwood-producing tree species. *Mitochondrial DNA Part B* **2018**, *3*, 1120–1121.

59. Liu, L.; Wang, Y.; He, P.; Li, P.; Lee, J.; Soltis, D.E.; Fu, C. Chloroplast genome analyses and genomic resource development for epilithic sister genera Oresitrophe and Mukdenia (Saxifragaceae), using genome skimming data. *BMC Genomics* **2018**, *19*,

235.

60. Neckermann, K.; Zeltz, P.; Igloi, G.L.; Kössel, H.; Maier, R.M. The role of RNA editing in conservation of start codons in chloroplast genomes. *Gene* **1994**, *146*, 177–182.

61. 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*.

62. Wakeley, J. The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. *Trends Ecol. Evol.* 1996, *11*, 158–162.

63. 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.

64. Morton, B.R.; Oberholzer, V.M.; Clegg, M.T. The influence of specific neighboring bases on substitution bias in noncoding regions of the plant chloroplast genome. *J. Mol. Evol.* **1997**, *45*, 227–31.

65. Cao, J.; Jiang, D.; Zhao, Z.; Yuan, S.; Zhang, Y.; Zhang, T.; Zhong, W.; Yuan, Q.; Huang, L. Development of Chloroplast Genomic Resources in Chinese Yam *(Dioscorea polystachya). Biomed Res. Int.* **2018**, *2018*, 1–11.

66. Piot, A.; Hackel, J.; Christin, P.A.; Besnard, G. One-third of the plastid genes evolved under positive selection in PACMAD grasses. *Planta* **2018**, *247*, 255–266.

67. Yu, X.; Zuo, L.; Lu, D.; Lu, B.; Yang, M.; Wang, J. Comparative analysis of chloroplast genomes of five *Robinia* species: Genome comparative and evolution analysis. *Gene* **2019**, *689*, 141–151.

Figure Legend

Figure 1. Circular maps of chloroplast genomes. Genes present inside the circle are transcribed counter clockwise, whereas genes present outside the circle are transcribed clockwise. Genes are colour coded based on functionality. LSC, IRb, SSC, and IRa of inner circle represent quadripartite structure of genomes.

Figure 2. Colinear block-based analyses of gene arrangement in the chloroplast genomes. The black block: transfer RNA genes, green block: transfer RNA genes with introns, white block: coding-genes, red block: ribosomal RNA genes. Light green and dark green blocks show differential existence of *ycf1* and *rps15* due to inverted repeats contraction and expansion.

Figure 3. Comparison of quadripartite junction sites among chloroplast genomes of four assembled species. Genes present on top of track transcribe on the negative strand, whereas genes present below the track transcribe on the positive strand. The T scale bar shows integration of genes between two adjacent regions. The junctions of genomes are represented as: JLB: IRb/LSC, JSB: IRb/SSC, JSA: SSC/IRa, and JLA: IRa/LSC.

Figure 4. Comparison of repeats among chloroplast genomes of four species. (a) Microsatellites distribution in regions of chloroplast genomes. (b) Numbers of different types of microsatellites. (c) Distribution of oligonucleotide repeats in regions of chloroplast genomes. (d) Types of oligonucleotide repeats. (e) Number of repeats based on the size. LSC: large single copy, SSC: small single copy, IR: inverted repeats, LSC/SSC, LSC/IR, SSC/IR represent those repeats pairs in which one copy exists in one region and another copy in another region. 14-20, 21-26,27-32, >32 showed range of repeat size.

Figure 5. Maximum likelihood tree based on multiple alignment of 30 species of Araceae. We omitted bootstrapping support from those nodes which showed 100 bootstrapping support.

Figure S1. Comparison of amino acid frequency among four species of Araceae.

Table 1. Quality and quantity of whole genome short reads and coverage depth analyses of the *de novo* assembled genomes

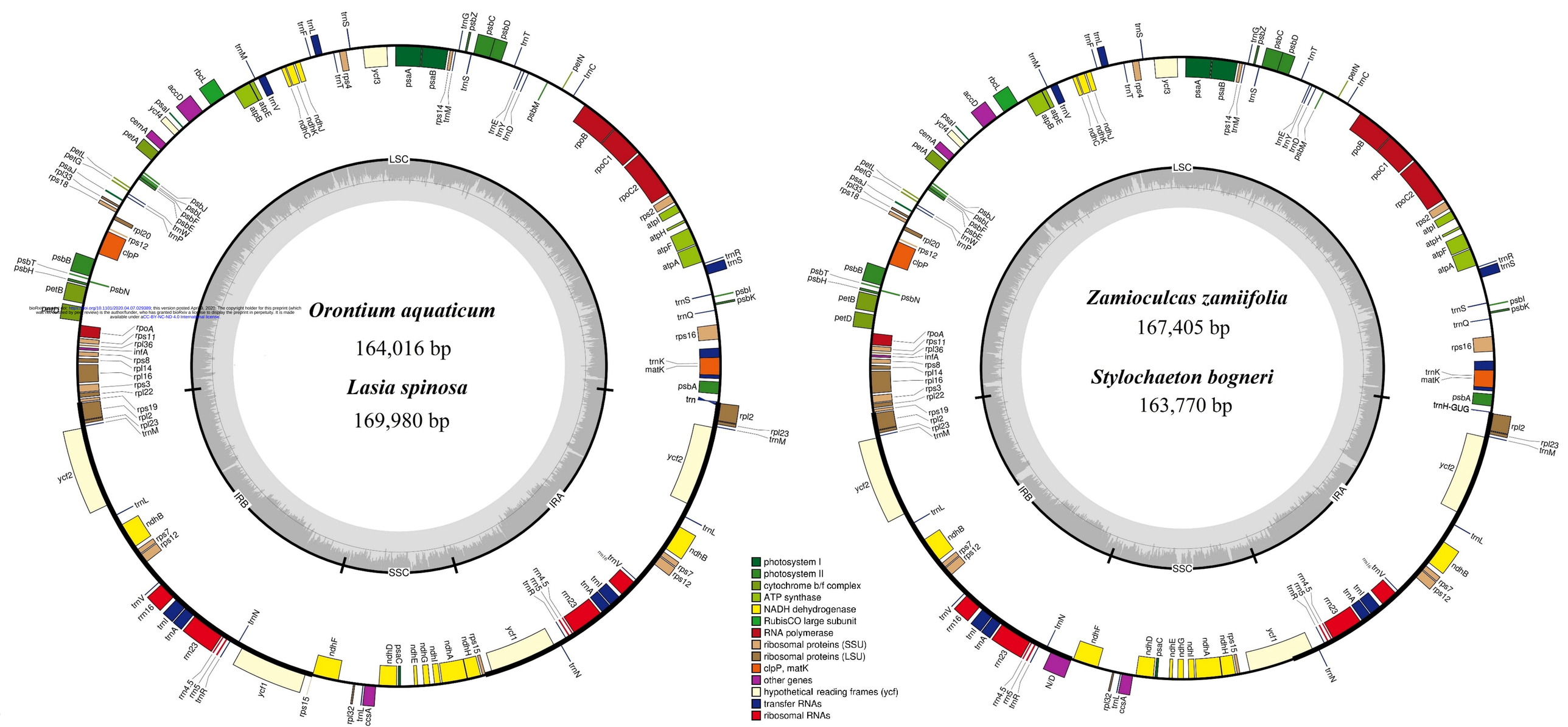
Species	Data in GB	Whole genome reads	Phred Score	Plastome Reads	Average coverage	Maximum coverage	NCBI accession
Orontium aquaticum	5.43	20,860,738	37.2	200,065	124.8	1347	MT226773
Lasia spinosa	8.35	32,000,000	37.6	1,738,094	1021	1929	MT226772
Zamioculcas zamiifolia	11.3	43,288,898	35.69	1,298,593	774.4	1293	MT226775
Stylochaeton bogneri	3.31	12,709,376	37.39	148,896	92.7	749	MT226774

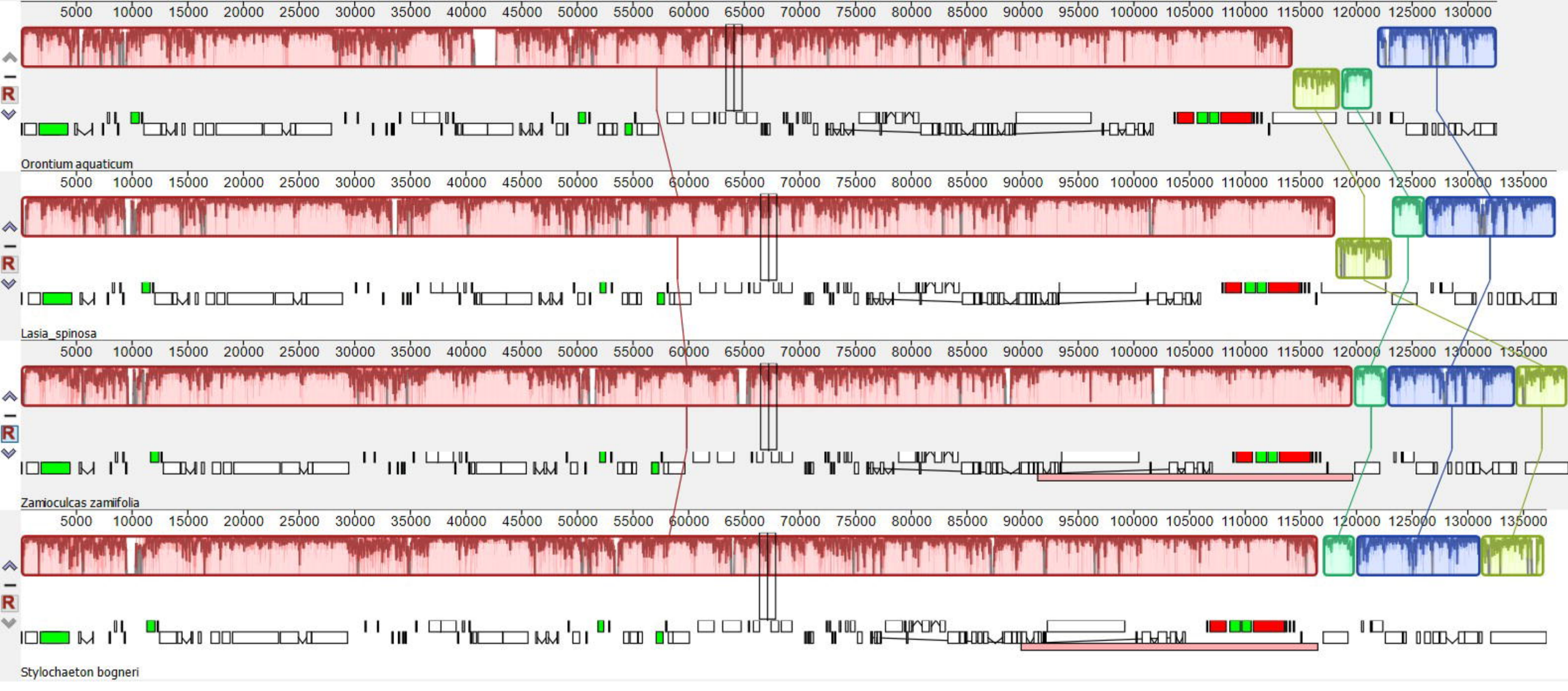
Table 2. Genomic features of *de novo* assembled chloroplast genomes

Characteristics		***Orontium aquaticum***	***Lasia spinosa***	***Zamioculcas zamiifolia***	***Stylochaeton bogneri***
Size (base pair; bp)		164,016	169,980	167,405	163,770
LSC length (bp)		87,269	91,150	91,357	89,869
SSC length (bp)		13,967	18,551	19,326	20,497
IR length (bp)		31,390	32,053	28,361	26,702
Number of genes		131	131	130	130
Protein-coding genes		85	85	84	84
tRNA genes		37	37	37	37
rRNA genes		8	8	8	8
Duplicate genes		18	18	17	17
GC content	Total (%)	37.3	36.1	35.9	35.7
	LSC (%)	35.7	33.9	34.2	34.0
	SSC (%)	31.9	31.0	30.4	29.5
	IR (%)	40.6	41.8	39.7	40.5
	CDS (%)	37.7	37.8	37.5	37.9
	rRNA (%)	55.2	54.6	55.0	55.0
	tRNA (%)	53.1	52.9	53.2	53.1
	All gene %	39.2	39.3	39.0	39.2

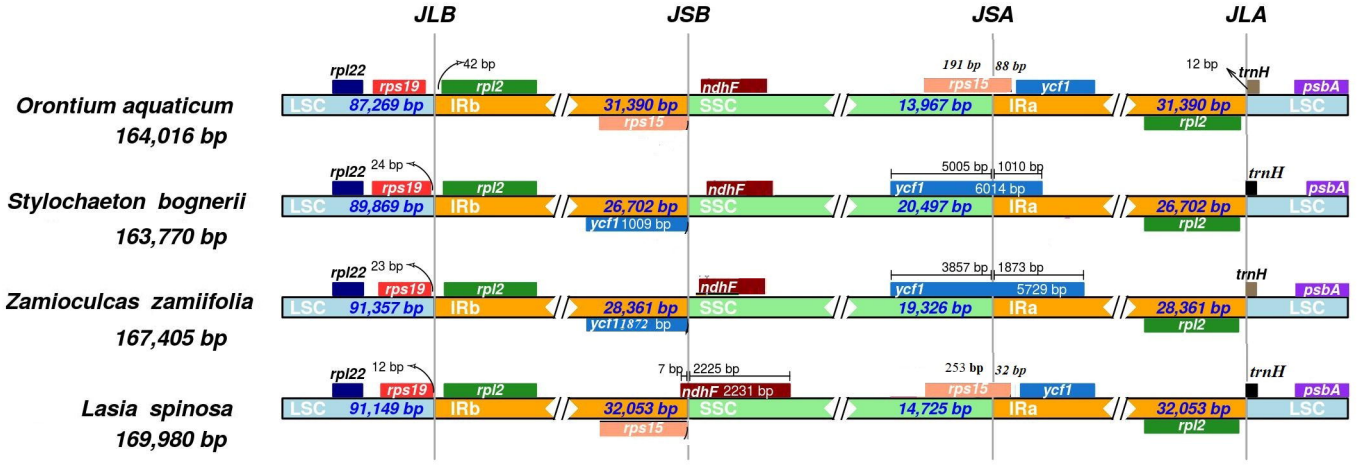
Table 3. Transition and transversion substitutions in protein-coding genes

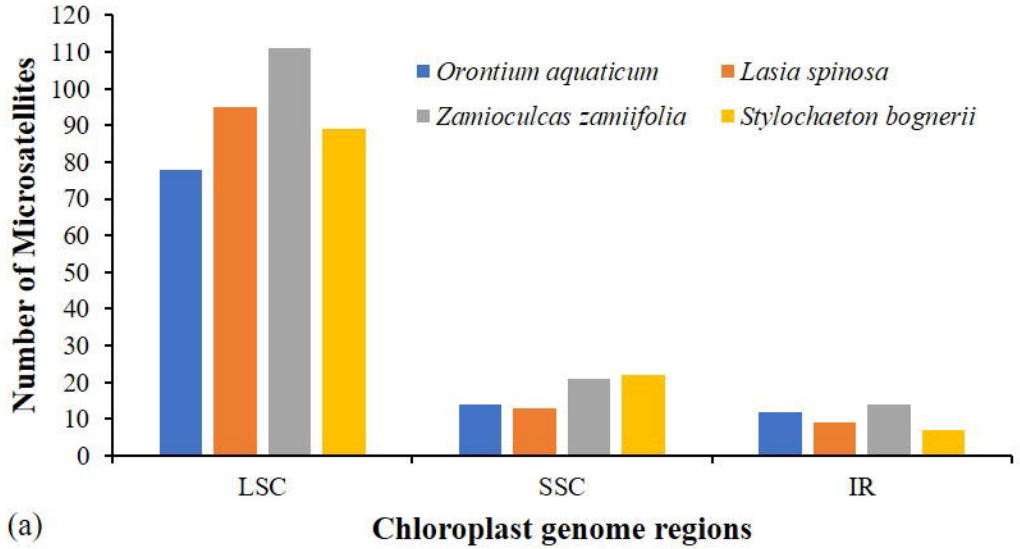
Substitution types	*Lasia spinosa*	*Zamioculcas zamiifolia*	*Stylochaeton bogneri*
A/C	478	466	560
C/T	1375	1267	1457
A/G	1429	1316	1540
A/T	274	254	315
C/G	160	156	193
G/T	305	322	406
Ts/Tv	2.3	2.15	2.03



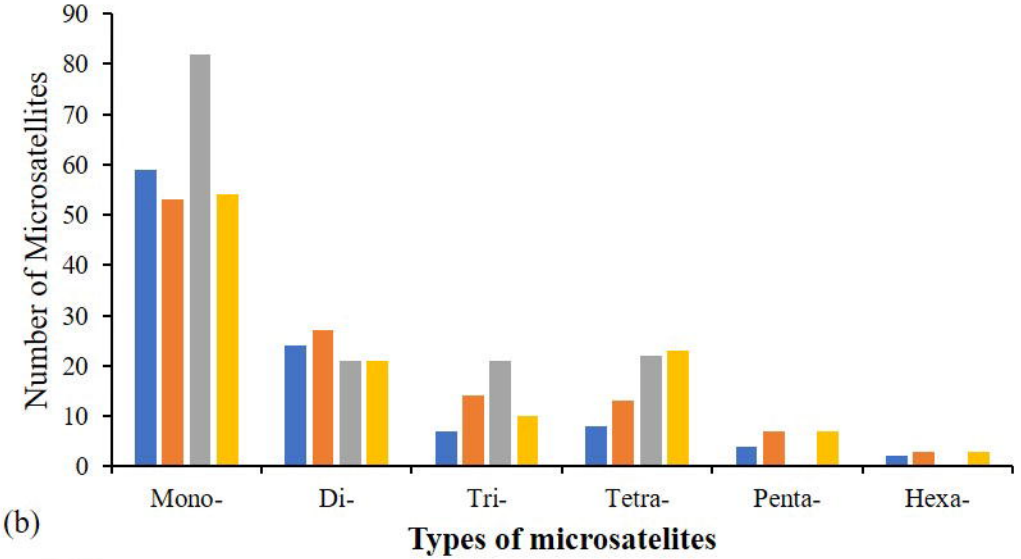

5000 10000 15000 20000 25000 30000 35000 40000 45000 50000 55000 60000 65000 70000 75000 80000 85000 90000 95000 100000 105000 110000 115000 120000 125000 130000
Orontium aquaticum
5000 10000 15000 20000 25000 30000 35000 40000 45000 50000 55000 60000 65000 70000 75000 80000 85000 90000 95000 100000 105000 110000 115000 120000 125000 130000 135000
Lasia_spinosa
5000 10000 15000 20000 25000 30000 35000 40000 45000 50000 55000 60000 65000 70000 75000 80000 85000 90000 95000 100000 105000 110000 115000 120000 125000 130000 135000
Zamioculcas zamiifolia
5000 10000 15000 20000 25000 30000 35000 40000 45000 50000 55000 60000 65000 70000 75000 80000 85000 90000 95000 100000 105000 110000 115000 120000 125000 130000 135000
Stylochaeton bogneri

Inverted Repeats

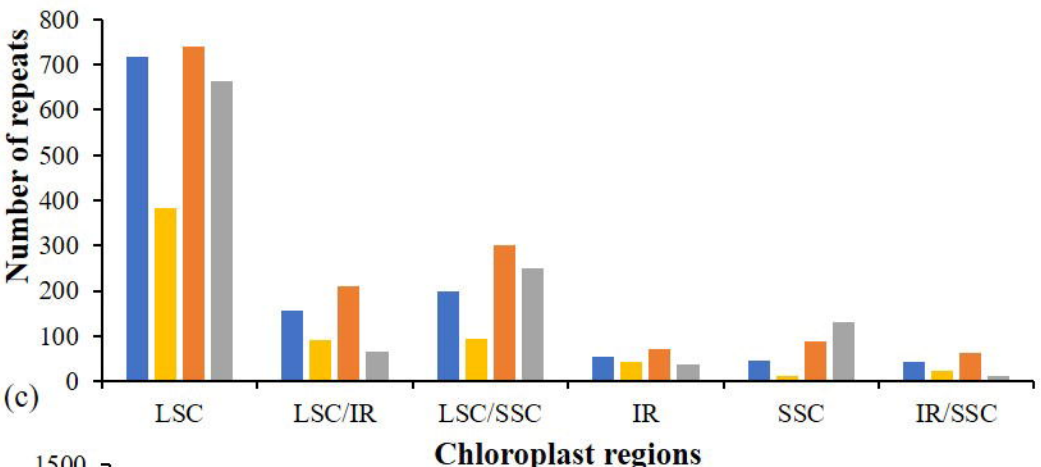



Number of Microsatellites
120
110
100
90
80
70
60
50
40
30
20
10
0
Orontium aquaticum
Lasia spinosa
Zamioculcas zamiifolia
Stylochaeton bognerii
LSC
SSC
IR
Chloroplast genome regions

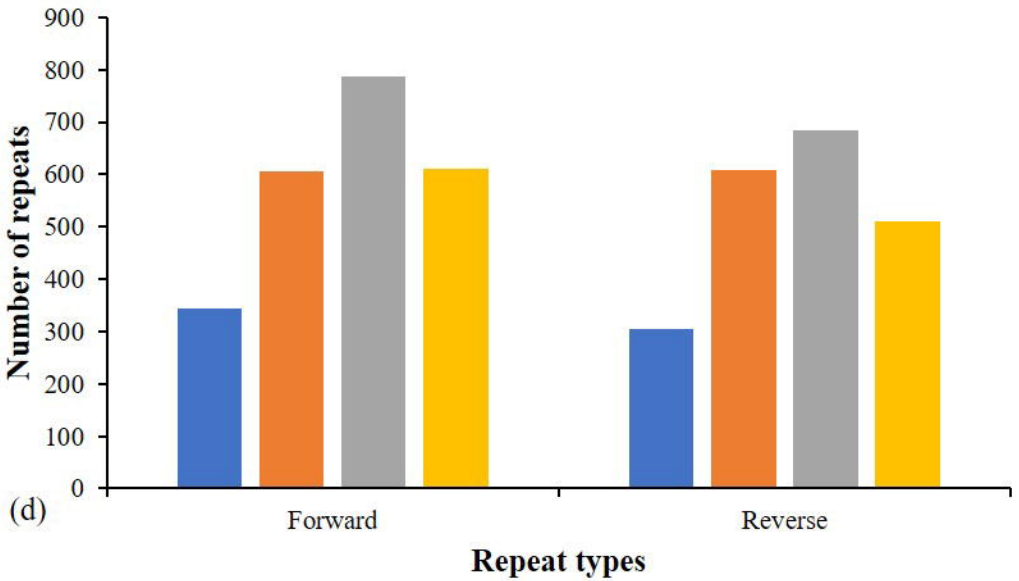
(a)


Number of Microsatellites
90
80
70
60
50
40
30
20
10
0
Mono-
Di-
Tri-
Tetra-
Penta-
Hexa-
Types of microsatelites

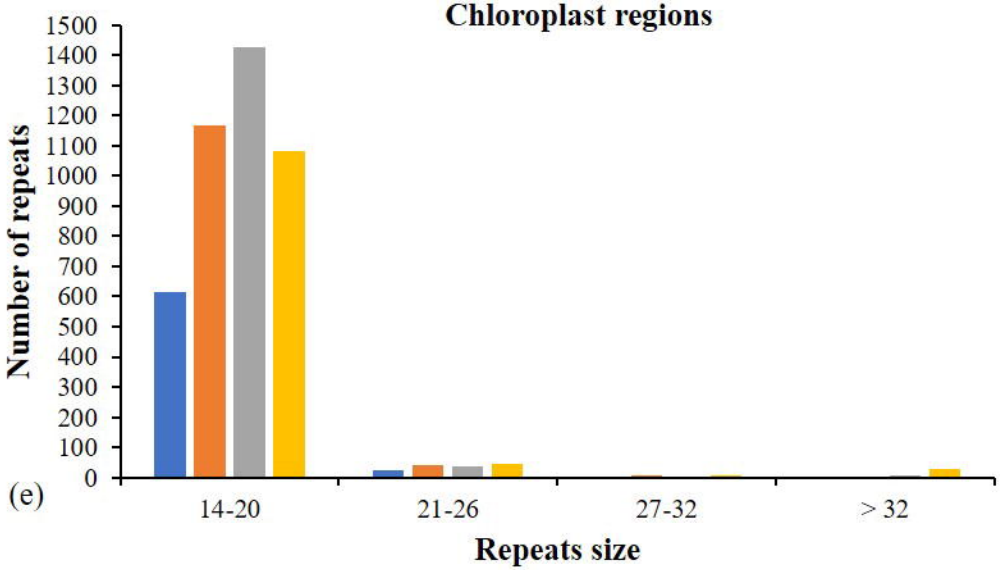
(b)


Number of repeats
800
700
600
500
400
300
200
100
0
LSC
LSC/IR
LSC/SSC
IR
SSC
IR/SSC
Chloroplast regions

(c)


Number of repeats
900
800
700
600
500
400
300
200
100
0
Forward
Reverse
Repeat types

(d)


Number of repeats
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
14-20
21-26
27-32
> 32
Repeats size

(e)