The evolution of extremely diverged plastomes in Selaginellaceae (lycophyte) is driven by repeat patterns and the underlying DNA maintenance machinery

Qiao-Ping Xiang1, Jun-Yong Tang1,2, Ji-Gao Yu1,2, David Roy Smith3, Yan-Mei Zhu1, Ya-Rong Wang1, Jong-Soo Kang1, Jie Yang1,2 and Xian-Chun Zhang1,*

1State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, The Chinese Academy of Sciences, Beijing 100093, China,
2University of Chinese Academy of Sciences, Beijing 100049, China, and
3Department of Biology, University of Western Ontario, London, N6A 5B7, Ontario, Canada

SUMMARY

Two factors are proposed to account for the unusual features of organellar genomes: the disruptions of organelle-targeted DNA replication, repair, and recombination (DNA-RRR) systems in the nuclear genome and repetitive elements in organellar genomes. Little is known about how these factors affect organellar genome evolution. The deep-branching vascular plant family Selaginellaceae is known to have a deficient DNA-RRR system and convergently evolved organellar genomes. However, we found that the plastid genome (plastome) of Selaginella sinensis has extremely accelerated substitution rates, a low GC content, pervasive repeat elements, a dynamic network structure, and it lacks direct or inverted repeats. Unexpectedly, its organelle DNA-RRR system is short of a plastid-targeted Recombinase A1 (RecA1) and a mitochondrion-targeted RecA3, in line with other explored Selaginella species. The plastome contains a large collection of short- and medium-sized repeats. Given the absence of RecA1 surveillance, we propose that these repeats trigger illegitimate recombination, accelerated mutation rates, and structural instability. The correlations between repeat quantity and architectural complexity in the Selaginella plastomes support these conclusions. We, therefore, hypothesize that the interplay of the deficient DNA-RRR system and the high repeat content has led to the extraordinary divergence of the S. sinensis plastome. Our study not only sheds new light on the mechanism of plastome divergence by emphasizing the power of cytonuclear integration, but it also reconciles the longstanding contradiction on the effects of DNA-RRR system disruption on genome structure evolution.

Keywords: Selaginella sinensis, network plastome, perfect repeats, dynamic mitogenome, GC content, RNA editing.

INTRODUCTION

With the modern advancements in sequencing technologies, many unusual organellar genomes have been uncovered recently from diverse plant lineages, including Selaginella (Bendich, 2014; Cortona et al., 2017; Guisinger et al., 2011; Hecht et al., 2011; Kang et al., 2020; Kozik et al., 2019; Mower et al., 2019; Odahara et al., 2021; Oldenburg & Bendich, 2004; Sloan et al., 2012; Weng et al., 2014; Su et al., 2019; Zhang, Zhang, et al., 2019). The mechanisms underlying the evolution of these abnormal organellar genomes are becoming a fascinating question. So far, two hypotheses have been proposed: one is that disruptions to the nuclear-encoded, organelle-targeted DNA replication, recombination, and repair (DNA-RRR) systems are responsible for the unusual organellar genome evolution; the other is that the number of repetitive elements is positively associated with aberrant organellar genomes. DNA-RRR genes have been regarded as key players in organellar genome evolution (Kang et al., 2020; Weng et al., 2014). This opinion is stimulated mainly by the experiment results using mutants of model organisms (e.g., Arabidopsis and Physcomitrella) and bacteria addressing functions of RecA homologous and other genes of the DNA-RRR system (Cerutti et al., 1992; Cox, 2013; Odahara et al., 2009, 2015; Rowan et al., 2010; Shedge et al., 2007). Kang et al. (2020) reported a deficient
DNA-RRR system in wild vascular plants and revealed the convergent evolution of plastome and mitochondrion genome (mitogenome) in Selaginellaceae caused solely by intact, dual-targeted recombinase gene. In studies on bacterial endosymbionts, it was reported that a deficient DNA-RRR system sometimes co-occurred with gene-order conservation (Garcia-Gonzalez et al., 2013; Tamas et al., 2002), or a highly dynamic architecture (Sloan & Moran, 2013). Therefore, how disruptions to the DNA-RRR system impacts genomic structure is still poorly understood, particularly in non-model organisms.

The correlations between repeat patterns and substitution rates, genome rearrangements and structure stability in some angiosperms and bacterial endosymbionts have been well explored (Blazier et al., 2016; Guisinger et al., 2008; Ruhlman et al., 2017; Sloan et al., 2012; Weng et al., 2014). In studying the plastomes of the seed plant genus Erodium, Blazier et al. (2016) concluded that the overall repeat content negatively correlated with genome stability, regardless of the genes observed in the DNA-RRR system. However, the above hypotheses have not been rigorously tested when taking into consideration the nuclear-encoded RRR system and repeat patterns in organellar genomes.

The lycophyte family Selaginellaceae (spikemoss) is the earliest-diverged lineage of vascular plants and contains a single genus, which is remarkable not only in high species diversity but also in great habitat amplitude (Jermy, 1990; Weststrand & Korall, 2016; Zhang et al., 2013). Because of its key systematic position in vascular plants and the small genome size, the Selaginellaceae has become a model of evolutionary studies in plants (Banks et al., 2011; VanBuren et al., 2018; Wang et al., 2020; Xu et al., 2018). Members of Selaginella are renowned for having unusual plastome features, such as genome-wide GC biases (>50%) (Kang et al., 2020; Smith, 2009; Zhang et al., 2020), contrasting the near-universal AT bias of organellar DNA in most eukaryotes (Smith, 2012). C-to-U RNA editing is pervasive in Selaginella plastomes, but U-to-C RNA editing is missing (Smith, 2020; Tsuji et al., 2007). The presence of the directed repeats (DR; including ribosome operon), which is different from the large inverted repeats (IR), the landmark of most land plant plastomes, was inferred as the ancestral state of the genus (Mower et al., 2019; Zhang, Zhang, et al., 2019). Plastomes and mitogenomes generally exhibit distinct features in plants. However, in the Selaginellaceae, they typically share similar genetic features (Kang et al., 2020; Smith & Keeling, 2015). The available Selaginella mitogenomes cannot be assembled into any predominant forms (Hecht et al., 2011; Kang et al., 2020). In addition to unusual genome organization, the mitogenomes have extremely high GC contents (63–68%), genome-wide C-to-U RNA editing, elevated substitution rates, and no detectable tRNA genes, paralleling some of the trends observed in the plastomes of Selaginella.

Kang et al. (2020) revealed that the DNA-RRR system of Selaginella organellar genomes encoded by the nuclear genome was deficient. For example, the plastid-targeted RecA1 and mitochondrion-targeted RecA3 are absent, and only dual-targeted RecA2 exists in the Selaginellaceae. The function of the RecA gene family was reported to play a key part in the accurate pairing of various types of homologous recombinases, via suppressing the illegitimate recombination between short repeats (Cox et al., 2000; Heinzhorst et al., 2004; Odahara et al., 2009, 2015; Rowan et al., 2010). The dual-targeted RecA2 in the Selaginellaceae was proposed to be responsible for the convergent evolution of the two compartments (Kang et al., 2020).

Previous studies have suggested that in the Selaginella sinensis group, plastome architecture might be particularly interesting as it consistently showed very long branches in phylogenies based on plastid DNA (ptDNA) fragments (Korall & Kenrick, 2002, 2004). The S. sinensis group comprises about 10 species with a disjunct distribution in the old world from Asia (China and Yemen) and Africa to Australia. Species of the S. sinensis group share some synapomorphy characters such as creeping stems, ventral rhizophores, similar dorsal and ventral vegetative leaves, isomorphic sporophylls, and often with only one megasporophyll at the base of strobilus. It belongs to subgenus Stachygyynandrum (Jermy, 1990; Weststrand & Korall, 2016), but the phylogenetic position is unclear and varies depending on the methods of analysis and different datasets (Korall & Kenrick, 2002, 2004). The extreme substitution rates in the S. sinensis group and the rate heterogeneity of ptDNA in the family might be responsible for these problems (Barrett et al., 2016; Wei et al., 2017; Zhang et al., 2020). Therefore, S. sinensis provides an opportunity to understand the evolutionary mechanisms responsible for genome divergence, given the unusual sequence characters and a potential genus-wide-deficient DNA-RRR system.

Here, using a combination of Illumina short-read and PacBio long-read sequencing data, we characterize the plastome and mitogenome of S. sinensis, investigate its DNA-RRR genes based on whole-genome data, and carry out systematic comparisons within the phylogenetic framework of the family. We reveal the unprecedented S. sinensis plastome, unremarkable mitogenome, and the deficient DNA-RRR system. We hypothesize that the interplay between the pervasive repeats in the plastome and the absence of RecA1 surveillance is responsible for the extremely high mutation rate, decreased GC content, and dynamic network structure of the plastome. Our study reconciles the longstanding opposing opinions on the evolutionary mechanism of organellar genome divergence and emphasizes the importance of integrating the genome information of different counterparts.
RESULTS

Structure and gene content of the organellar genomes in *Selaginella sinensis*

Plastome structure and gene content. Fifteen putative contigs were obtained from the reference-based plastome assembly of *S. sinensis* (Figure 1a; Table S1). Genes were identified and annotated on 12 of these contigs, but no recognizable gene was found on the other three contigs (Figure 1b). Close inspection of these contigs indicated that they did not derive from a single, circular-mapping chromosome, but instead came from a complex network of overlapping, reticulate plastome molecules. For example, in the hotspot Region II, there are many connection modes, including circular contigs (i–j–k–f–h and i–g–k–f–h) and linear contigs (h–i–j–k–l and h–i–g–k–l) (Figure 1a). This contrasts with the plastome conformations from other described *Selaginella* species, which are intact chromosomes with canonical quadripartite structures (Mower et al., 2019; Smith, 2009; Zhang, Xiang, et al., 2019), or multipartite chromosomes (Kang et al., 2020).

To decipher the underlying pattern of the putative plastome structure of *S. sinensis*, we identified three recombination activity regions (Regions I, II, and III) (Figure 1a). These regions were hotspots for rearrangements and could be used to bridge together distinct contigs (Figure 1a). For instance, four possible rearrangement patterns were associated with Region I and Region III, while there were 64 possible rearrangements associated with Region II. Given the architecture and position of the three conserved regions, there were at least 1024 putative isomers, and each of them could cover all contigs. By coverage statistical analysis, we assembled two small subgenomes C1 (37 604 bp) and C2 (43 521 bp), and a large genome-sized molecule (C1 + C2, 81 010 bp) (Figure S1) *in silico*. We further tested their existence using PacBio long-reads.
The PacBio long-reads sequencing of *S. sinensis* supported a reticulated plastome architecture deduced from the Illumina data, and provided precise connection patterns and gross frequency among the plastome contigs (Table S2). In total, the PacBio long-reads revealed 40 possible connections (surrounding Regions I, II, and III), and 70% of the connections were supported by more than 1000 reads. For example, there were more than 20,000 PacBio long-reads supporting contigs m–l, k–l–n and g–k–l. Although both Region I (1762 bp) and Region III (2822 bp) were relatively longer repeats, the rearrangements mediated by homologous recombination through these two repeats (longer than 1000 bp) were unbalanced. Two-fold more PacBio long-reads mapped to contigs b–c–e and a–c–e in comparison with contigs a–c–d and b–c–d (Figure 1a; Table S2). The middle-sized repeats (i: 260 bp, k: 462 bp, l: 169 bp, and h: 370 bp in Region II), also mediated unbalanced recombination (Table S2). The result of asymmetric recombination via smaller (100–1000 bp) repeats perfectly matched the previous hypothesis that the smaller repeats might engage in asymmetric recombination under some conditions, notably in repair mutants and when DNA damage was increased (Christensen, 2018; Klein et al., 1994; Marechal & Brisson, 2010; Palmer & Herbert, 1988; Shedge et al., 2007). Indeed, the *Selaginella* species were natural DNA repair mutants including *S. sinensis*. However, it is unclear how the relationships between the asymmetric recombination mediated by longer and middle-sized repeats (longer than 100 bp) and the deficient DNA-RRR system.

The relationship between the PacBio long-read coverage and the different connection patterns varied from contig to contig. We further evaluated the two subgenomes C1 and C2, as well as the putative assembled genome with all contigs C1 + C2. We found that there were five and 104 PacBio long-reads supporting C1 and C2, respectively; no PacBio long-reads supporting the master genomic form, C1 + C2 (Table S3). Perhaps C1 + C2 was an assembly artifact or it only existed in some special development stages of the plant (or plastids). Consequently, we argued that the *S. sinensis* plastome has an active tendency towards rearrangements and, therefore, a highly dynamic conformation.

The *S. sinensis* plastome has a reduced gene content (Figure 1b; Figure S2; Table S4). In total, 48 putative plastid-encoded genes were detected, including 44 protein-coding genes and four rRNA genes. This is the smallest plastid gene set yet identified in lycophyte. Not a single plastDNA-located tRNA was identified, paralleling data from the *Selaginella* mtDNA analyses (Hecht et al., 2011; Kang et al., 2020). NADH dehydrogenase complexes (*ndh* genes) are absent, and only a small portion of the ribosomal subunit genes (*rpl* genes) were identified from the plastDNA assembly in *S. sinensis*, which is similar to reports from dry-tolerant spikemosses (Xu et al., 2018; Zhang, Zhang, et al., 2019). What is more, we were not able to recognize all the genes for the plastid-encoded RNA polymerase (*rpo* genes) (Figure S2), which might be explained by the extremely high substitution rates in the *S. sinensis* plastome (Blazier et al., 2016).

**Mitogenome structure and gene content.** Forty-four putative contigs were obtained from the reference-based mitogenome assembly of *S. sinensis* (Table S5). Both ends of each contig have one or multiple configurations, which results in a complicated mitogenome structure, consistent with previously reported *Selaginella* mitogenomes (Hecht et al., 2011; Kang et al., 2020). The assembled mitogenome is a complicated network and there are numerous possible tailing paths among the 44 contigs (Figure 2a). The mitogenomes lacking a circular molecule had been well addressed in plants (Backert & Börner, 2000; Bendich, 1996; Kang et al., 2020; Kozik et al., 2019; Li & Cullis, 2021; Manchekar et al., 2009; Mower et al., 2012; Oldenburg & Bendich, 1996, 1998, 2001). For the convenience of analysis and description, these contigs were numbered 1–44 according to their lengths (Figure 2a; Table S5). The filtered mitochondrial PacBio reads were used to verify the connection nodes and pathways. For annotating mitochondrial genes, eight scaffolds consisting of 44 contigs were used for gene annotation, and connection nodes of each scaffold were confirmed by PacBio reads (Figure 2b). Each scaffold is formed by the connection of multiple contigs, and some contigs appear more than once in these scaffolds. The coverage of the eight scaffolds ranged from 143× to 185×, and the GC content ranged from 68% to 69.4% (Table S6). We completely annotated 17 protein-coding genes, nine of which (*nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, and *nad9*) were Complex I (NADH Dehydrogenase Subunits) related, one (*cob*) involved in Complex III (Cytochrome bc1 Complex Subunits), three (*cox1*, *cox2*, and *cox3*) participated in Complex IV (Cytochrome c Oxidase Subunits), and four genes (*atp1*, *atp6*, *atp8*, and *atp9*) related to Complex V (ATP Synthase Subunits). The annotated protein-coding genes of the *S. sinensis* mitogenome are almost identical with those of *Selaginella moellendorffii* and *Selaginella nipponica*, except for the absence of a plant-typical core set of twin-arginine translocase (*tatC*) (Figure S3).

**Repeat elements in plastomes of *Selaginella sinensis* and other lycophytes.**

Here, plastome repeats were divided into three main types: long repeats (classical IR/DR region including ribosome operon, generally greater than 5000 bp), medium-sized repeats (from 100 to 5000 bp), and short repeats (generally less than 100 bp). We compared the abundance of repeats among the plastomes of *S. sinensis* and 12 other species...
with differing levels of structural complexity, representing different lineages of lycophytes with or without RecA1. The *S. sinensis* plastome lacks DR or IR structures, with pervasive medium and short repeats, and the longest repeat is 2822 bp (contig v) (Figures 1 and 3; Figure S4; Table S7). The plastomes of the other 12 taxa have either IR or DR structures (Figure 3).

There are seven medium repeats in the abnormal plastome of *S. sinensis*, which is the greatest among the *Selaginella* species, and outgroups (*Isoetes engelmannii* and *Huperzia serrata*). *Selaginella bisulcata* with coexisting IR and DR contains four medium repeats. The multipartite plastomes of *S. nipponica* and *S. pallidissima* each contain two medium repeats. The remaining IR or DR plastomes (*Selaginella doederleinii*, *S. moellendorfii*, *Selaginella tamariscina*, *Selaginella lyalii*, *Selaginella remotifolia*, *Selaginella vardei*, *Selaginella lepidophylla*, *I. engelmannii*, and *H. serrata*) contain only one or no medium repeat (Figure 3) (Kang et al., 2020; Zhang, Zhang, et al., 2019). *Selaginella sinensis* has the greatest number of short repeats (93) among the *Selaginella* species (from 6 to 30), but the difference is not so remarkable compared with the...
outgroups *I. engelmannii* (76) and *H. serrata* (39). In addition, repeats in *S. sinensis* are distributed throughout the plastome (both coding and non-coding regions), but repeats in other plastomes are mostly distributed in non-coding regions (Figure S4; Table S7).

**Nucleotide substitution rates of the organellar genomes among Selaginella species**

Nucleotide substitution rate analyses suggested that the *S. sinensis* plastome experienced a dramatically elevated evolutionary rate based on the 31 protein-coding genes shared with other *Selaginella* ptDNA (see “Results” section) (Figure S5). Both the non-synonymous (dN) and synonymous (dS) substitution rates (0.37 and 5.38, respectively) in the *S. sinensis* plastid genes are extremely accelerated compared with those in other *Selaginella* species (0.16 and 1.27, respectively), and the difference of the substitution rates between *S. sinensis* and other *Selaginella* species is significant according to the Bonferroni post-hoc tests (P < 0.001) (Figure S5). The dN of *S. sinensis* is at least twice that of other *Selaginella* species, and four times higher than those of other land plants were (Figure S5a). The dS of *S. sinensis* is even more extreme, to nearly five times that of other *Selaginella* species, and nine times that of other land plants (Figure S5b). Considering the extreme substitution rate of *S. sinensis*, the saturation should be questioned here. To eliminate the influence of branch accumulation and mutation saturation on the nucleotide substitution, we used *Huperzia* and *Isoetes* as the references respectively. For the plastomes, gene pairs taken from 31 shared genes in each *Selaginella* species and two outgroup species were constructed to calculate dN and dS. The dS median value of *S. sinensis* is 2.02, the greatest one in *Selaginella*, 52% faster than that of *S. remotifolia* (dS median value 1.33, the second greatest one), and 136% faster than that of *S. moellendorffii* (dS median value 0.9, the least one here) with *Huperzia* as the reference. The results are similar using *Isoetes* as the reference (Table S8).

Different from the plastome, substitution rates of the mitogenome (based on 16 shared protein-coding genes) of *S. sinensis* are similar to those of other *Selaginella* species (Figure S6). Both dN and dS of the *S. sinensis* mitogenome (dN 0.62 and dS 1.53) are slightly higher than those of the other two explored *Selaginella* species, but their difference is not significant (dN: 0.474, dS: 1.286 in *S. moellendorffii*,

---

**Figure 3.** Comparison of plastome structure, repeat number, GC content, and synonymous and non-synonymous substitution rate, among the representative species in a phylogenetic framework of Selaginellaceae. Phylogenetic framework was modified from Weststrand and Korall (2016). Middle was the different complex categories of structure in the representative Selaginella species and their repeat richness. L, long repeat (IR/DR, >5000 bp) number; M, medium repeat number (100–5000 bp); S, short repeat (<100 bp) number. Network plastome of *Selaginella sinensis* was achieved in this study, the multipartite plastome of *Selaginella nipponica* was from Kang et al. (2020), the DR plastome of *Selaginella vardei*, the IR plastome of *S. lepidophylla*, and the DR/IR coexisting plastome of *Selaginella bisulcata* were from Zhang, Xiang, and Zhang (2019). Right panel represented the GC content, and the synonymous and non-synonymous substitution rate of the plastome, respectively.
and \(dN\): 0.402, \(dS\): 1.141 in \(S.\) nipponica). The genus \(Selaginella\) as a whole, exhibits dramatically elevated levels of \(dN\) and \(dS\) as compared with other land plants (Figure S6). The \(dS\) median value of \(S.\) sinensis is 0.44, which is similar with \(S.\) moellendorffii (0.43) and \(S.\) nipponica (0.42) using \(Huperzia\) as the reference (Table S9).

Correlation of repeat abundance with structural complexity and substitution rates respectively among the \(Selaginella\) plastomes

It is known that repeat-mediated recombination plays an important role in organellar genome stability, including genome structures and nucleotide substitution rates (Christensen, 2013, 2018; Marechal & Brisson, 2010). We assessed the correlation of repeat abundance with structural complexity and substitution rates among \(Selaginella\) plastomes and took into consideration the medium repeats and short repeats due to the lack of long repeats (IR/DR) in the \(S.\) sinensis plastome. A significant positive correlation between the number of medium repeats and structural complexity (using gene order shuffling as proxy) in \(Selaginella\) was expressed by correlation regression analyses \((P = 0.04, R^2 = 0.879)\) (Figure 4a). The number of short repeats showed an even more significant positive correlation with structural complexity in \(Selaginella\) \((P = 0.02, R^2 = 0.9315)\) (Figure 4b). The results of regression analyses, applied to test the quantitative relationship between repeat abundance and substitution rates among plastomes of different \(Selaginella\) species, indicated that the richness of short repeats and substitution rates were significantly correlated \((P = 1.82e-13, R^2 = 0.9267)\) (Figure 4c), but the quantity of medium repeats showed a weak positive correlation with substitution rates \((P = 1.74e-13, R^2 = 0.74)\) (Figure 4d).

GC content and RNA editing frequency in organellar genomes of \(Selaginella\) sinensis

One of the unusual features of the \(S.\) sinensis plastome is that it has become a relatively “normal” GC content: approximately 36% (Figure S7; Table S10). In contrast with most land plant plastomes, which are AT-rich, plastomes of most \(Selaginella\) species stand out as being GC-rich ptDNAs: over 50% (Figure 3; Figure S7) (Kang et al., 2020; Mower et al., 2019; Smith, 2009; Zhang, Zhang, et al., 2019; Zhang et al., 2020). The \(S.\) sinensis plastome represents the first exception to the GC-biased feature of

Figure 4. Regression analysis between rearrangements (inversion distance as proxy), repeats, and mutations in the \(Selaginella\) plastomes.
(a) Regression analysis of rearrangements against numbers of medium repeats.
(b) Regression analysis of rearrangements against numbers of short repeats.
(c) Regression analysis of mutations against numbers of short repeats.
(d) Regression analysis of mutations against numbers of medium repeats.

© 2022 Society for Experimental Biology and John Wiley & Sons Ltd.,
The Plant Journal, (2022), 111, 768–784
Selaginella plastomes. GC content of ptDNA varies slightly across different regions (coding or non-coding region) and contigs (Table S11), and subgenomes C1 and C2 have GC contents of 34.8 and 36.8% respectively (Table S3). Therefore, the low GC content is genome-wide, not local in the S. sinensis plastome.

To detect any potential mutational biases in the plastome of S. sinensis, we analyzed the shared protein-coding genes with six other Selaginella species (see “Results” section), using I. engelmannii as the reference. The transition from A:T to G:C in S. sinensis plastome dramatically decreased compared with other Selaginella species. Conversely, both transition and transversion to A:T in S. sinensis plastome are markedly increased (Figure S8).

In addition to having higher GC-biased nucleotide composition, Selaginella plastomes are renowned for undergoing extensive C-to-U RNA editing. For example, more than 3400 C-to-U RNA editing sites were uncovered in the Selaginella uncinata ptDNA, and those of Selaginella kraussiana and S. lepidophylla have >1300 and >700 editing sites, respectively (Smith, 2020). However, the S. sinensis ptDNA undergoes only a modest amount of editing (approximately 148 sites) (Table S12). These editing sites all occur in 31 of the 48 putative plastid genes. There is no detectable RNA editing event in other 13 protein-coding genes and four rrn genes. The editing sites mainly locate in the first and second codon positions with the potential to cause amino acid changes. The genes atpB and petA contain the largest number of editing sites (12 sites each gene). The S. sinensis plastome experienced about one-tenth editing hits observed in the S. uncinata ptDNA (Figure S9). We further compared the GC content of each gene before and after editing and found that the per-genome GC content after editing in S. uncinata is still much higher than in S. sinensis (Figure S9). This result suggested that the retroprocessing might not be responsible for the genome-level GC content difference in Selaginella.

Different from the plastome, the S. sinensis mitogenome keeps the character of GC bias, with GC content 68.7% (Figure S10), which is consistent with those of S. moellendorffii (68.1%) and S. nipponica (68.2%). The GC content of the Selaginella mitogenomes is much higher than the quillworts (49% of I. engelmannii) (Grewe et al., 2009) and clubmosses (44.2% of Huperzia squarrosa) (Liu et al., 2012) in lycophytes. In accordance with the abundant RNA editing sites reported in S. uncinata and S. moellendorffii (Smith, 2020; Tsuji et al., 2007), we identified 1936 C-to-U RNA editing sites in the 17 protein coding regions of the S. sinensis mitogenome (Table S13). Of these, 441 editing sites (22%) are silent, whereas the remaining 1475 sites introduce codon changes. The number of RNA editing sites among 17 genes varies greatly. There are 1061 editing sites in Complex I-related genes, which account for most of the RNA editing events in mitochondria DNA, particularly for nad4 and nad5, both containing more than 200 editing events respectively. There are more than 400 RNA editing sites in Complex IV- and Complex V-related genes respectively. In line with the GC content trend, the number of RNA editing sites in Selaginella is much higher than that in I. engelmannii (Grewe et al., 2009) and H. squarrosa (Liu et al., 2012), but similar among Selaginella species.

Nuclear-encoded, organelle-targeted DNA-RRR and PPR protein families

RecA (Recombinase A), OSB (Organellar single-stranded DNA-binding proteins), and Why (Whirly) families represent the protein families involved in DNA repair systems, which mainly suppress AT biased and error-prone mutation mediated by short repeats (Guisinger et al., 2008; Marechal & Brisson, 2010; Odahara et al., 2009; Shedge et al., 2007; Weng et al., 2014; Zhang et al., 2016). To explore the potential factors associated with the structural diversity of the Selaginellaceae organellar genomes, we investigated nuclear-encoded DNA-RRR protein families based on the whole-genome data of Arabidopsis thaliana, Amborella trichopoda, Marchantia polymorpha, S. moellendorffii, S. sinensis, S. lepidophylla, and S. tamariscina (Table 1; Figure S11). The Selaginella nuclear genomes retain only RecA2 and OSB3 of RecA and OSB protein families, and their products target both the plastid and the mitochondrion. The single-target members of these gene families, specific to plastid or mitochondrion, e.g., RecA1, RecA3, OSB1, and OSB2, are all absent. The Why protein family is puzzling. Why2 protein has a dual-targeted function in Selaginella according to its TAEGETP value, but it has not been reported as having a dual-targeted function. However, S. sinensis has an additional member (Why1/3 could not be distinguished in the phylogenetic tree of Figure S11c, mainly based on the prediction results of TARGETP), which might only target to plastid (Table 1). This issue needs further investigation.

Thus far, Selaginella species have the largest number of nuclear-encoded PPR proteins in land plants (Banks et al., 2011). Based on the whole nuclear genome data, we found 803 PPR homologs in S. sinensis, but 1279 counterparts in S. moellendorffii using two searching strategies (see “Results” section) (Table S14).

DISCUSSION

Repeat elements playing on the stage set by the deficient DNA-RRR system shape the extraordinary plastome divergence in Selaginella

The S. sinensis plastome has a dynamic, reticulated architecture, unlike the circular mapping, quadripartite or multipartite structures of other characterized Selaginella ones (Kang et al., 2020; Mower et al., 2019; Smith, 2009; Zhang, Zhang, et al., 2019). Its mitogenome shows a similarly...
Figures S9 and 10). These observations lend support for plastomes, despite usually assem-
dering as circular molecules, could exist as a complex assemblage of linear, circular, and branched molecules as recognized in plant mitogenomes (Ben-

Two hypotheses have been proposed to interpret the organellar genome divergence: one is that the disruptions of nuclear-encoded and organelle-targeted DNA-RRR system may be responsible for the unusual organellar genome structures; the other is that the number of repeats is related with the aberrant organellar genomes. Our observations in Selaginella (Selaginellaceae, Lycophyte) revealed the limitation of both hypotheses. First, species of Selaginellaceae, including S. sinensis, share a common, deficient DNA-RRR system characterized by the universal absence of RecA1 and RecA3 genes (plastid-targeted and mitochondrion-targeted respectively), and the remaining of intact dual-targeted RecA2 gene (Table 1) (Kang et al., 2020). However, the S. sinensis plastome is extraordinary in its highly dynamic network structure, extremely high substitution rate, lacking DR, pervasive medium and short repeats, low GC content, and low RNA editing frequency. Obviously, the divergence of Selaginella plastomes (Figure 3), could not be interpreted by the genome-wide deficient DNA-RRR system only (Table 1). Furthermore, our observations challenge the opinion that overall repeat content is negatively correlated with genome stability regardless of the genes in the DNA-RRR system (Blazier et al., 2016). The plastome of I. engelmannii (Isoetaceae, sister family with Selaginellaceae in lycophyte), also contains many short repeats (76), but the plastome structure and nucleotide substitution rate is relatively conservative (Figure 3). However, RecA1 is present in the DNA-RRR system of I. engelmannii, but absent in S. sinensis (Table 1).

Table 1 Genome screening results of DNA-RRR system related genes in seven land plants

<table>
<thead>
<tr>
<th>Protein</th>
<th>Subcellular target</th>
<th>Arabidopsis thaliana</th>
<th>Amborella trichopoda</th>
<th>Selaginella moellendorfii</th>
<th>Selaginella sinensis</th>
<th>Selaginella lepidophylla</th>
<th>Selaginella tamariscina</th>
<th>Marchantia polymorpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>RecA2</td>
<td>Plastid/ mitochondrion</td>
<td>2(^a), 0.005(^b), 0.932(^b)</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, 0.125, 0.549</td>
<td></td>
</tr>
<tr>
<td>RecA3</td>
<td>Mitochondrion</td>
<td>1.019, 0.006</td>
<td>0.006</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>1.093, 0.000</td>
</tr>
<tr>
<td>OSB1</td>
<td>Mitochondrion</td>
<td>1.096, 0.000</td>
<td>0.000</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
</tr>
<tr>
<td>OSB2</td>
<td>Plastid</td>
<td>2.052, 0.382</td>
<td>0.002</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
</tr>
<tr>
<td>OSB3</td>
<td>Plastid/ mitochondrion</td>
<td>2.0937, 0.045</td>
<td>4.0934, 0.000</td>
<td>3.0487, 0.511</td>
<td>1.0371, 0.615</td>
<td>1.0145, 0.853</td>
<td>2.0693, 0.113</td>
<td>0, --, --</td>
</tr>
<tr>
<td>Why1</td>
<td>Plastid</td>
<td>1.073, 0.857</td>
<td>2.083, 0.477</td>
<td>0, --, --</td>
<td>1.092, 0.571</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>1.304, 0.636</td>
</tr>
<tr>
<td>Why2</td>
<td>Plastid</td>
<td>1.0013, 0.821</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
<td>0, --, --</td>
</tr>
</tbody>
</table>

DNA-RRR, DNA replication, repair and recombination.
\(^a\)Homolog number.
\(^b\)Average frequency of mitochondrial targeting peptide.
\(^c\)Average frequency of plastid targeting peptide.
The Plant Journal

Repeats and DNA-RRR drive genome evolution

777

ribosome operon, greater than 5000 bp), medium repeats (100–5000 bp), and short repeats (less than 100 bp). We compared the abundance of repeats among the plastomes with different structural complexity, representing different lycophyte lineages with or without RecA1 (Figure 3; Table 1) (Kang et al., 2020). There is no DR in the S. sinensis plastome (Figure 3), which is the landmark of the Selaginella plastome. There are much more medium (seven) and short repeats (93) in S. sinensis than in other species (Figure 3; Figure S4; Table S7). The DR plastomes of S. doederleini, S. moellendorffii, S. tamariscina, S. iyalii, S. remotifolia, and S. vardei contain only one or even no medium repeats, 6–20 short repeats (Figure 3), which is consistent with our previous report (Zhang, Zhang, et al., 2019). A repeat is regarded as the by-product of the DNA-RRR processes (Perry & Wolfe, 2002). Owing to the shared deficient DNA-RRR system, the repeat generation is supposed to be similar in different species of Selaginella. What causes such significant difference on repeat contents in Selaginella plastomes? In our opinion, DR is long enough to mediate gene conversion and to correct the genome sequences by eliminating the generated repeats, which results in fewer repeats in DR plastomes of most Selaginella species (Mower et al., 2019; Zhang, Zhang, et al., 2019). The mechanism of illegitimate repeat purge mediated by DR does not exist in the S. sinensis plastome, which accumulate the illegitimate repeats regularly produced (Figure 3).

In the DNA-RRR system, RecA1 has been known to suppress the illegitimate recombination via the short repeats in plastomes of moss and seed plant, which causes AT biased and error-prone mutation, also structure instability (Cerutti et al., 1992; Cox, 2013; Odahara et al., 2009, 2015; Rowan et al., 2010; Shedge et al., 2007). The function of RecA1 should be conserved in higher plants including lycophytes. Because of the absence of RecA1, the illegitimate recombination via abundant short repeats is out of control and becomes extremely active in the S. sinensis plastome, which causes the increased AT content (accordingly decreased GC content), extremely high mutation rate, and complicated structure as observed. This interpretation is further supported by the significant correlation between quantity of short repeats with structure complexity and substitution rates in the plastomes of different Selaginella species (Figure 4b,c). The contribution of asymmetric recombination to the dynamic network structure, is supported by the significant correlation between quantity of medium repeats (100–5000 bp) and structure complex of the Selaginella plastomes (Figure 4a).

Because of the AT-biased mutation caused by the pervasive short repeats in the S. sinensis plastome, the GC content is decreased to 36.2% in comparison with those (above 50%) in other Selaginella species. The positive correlations among the GC content, the frequency of C-to-U RNA editing events in organellar DNA, and the amount of PPR families (Dong et al., 2019; Fujii & Small, 2011; Hecht et al., 2011; Malek et al., 1996; Rüdinger et al., 2008, 2012; Salone et al., 2007; Schallenberg-Rudinger et al., 2014; Tsuji et al., 2007; Wolf & Karol, 2012). We will discuss this point below.

Based on the above results, here we propose the “Deficient DNA-RRR system + Repeat pattern” hypothesis to explain the divergent evolution of organellar genomes in Selaginella, particularly the extremely diverged S. sinensis plastome (Figure 5). We argue that the impact of RecA1-lacking DNA-RRR system on the plastome divergent evolution depends on the repeat patterns, which includes the presence or absence of DR/IR, and quantity of medium repeats and short repeats. The deficient DNA-RRR system could result in highly diverged genomic complexity when there are rich medium and short repeats in the non-DR plastome, such as in S. sinensis. Alternatively, it could result in relatively conserved structures when there are few or no medium and short repeats, such as the DR plastomes of other Selaginella species, e.g., S. vardei (Zhang, Xiang, et al., 2019). This hypothesis combines the function of missing genes in a deficient DNA-RRR system with the specific plastome features correlated with the missing genes. It is powerful in explaining the plastome divergence in Selaginella, which could also help to explain the genomic complexity of the bacterial endosymbiont with a deficient DNA-RRR system (Garcia-Gonzalez et al., 2013; Sloan & Moran, 2013; Tamas et al., 2002).

Evolutionary scenario “If you don’t use it, you lose it”: decreasing RNA editing frequency co-occurring with the PPR family contraction

The GC content in organellar genomes of Selaginella is on the top level of land plants, ranging from 50.2% to 56.5% in plastomes (Mower et al., 2019; Zhang, Zhang, et al., 2019), and 63.5% to 68.1% in mitogenomes (Hecht et al., 2011; Kang et al., 2020). The 68.6% GC content of S. sinensis mitogenome is consistent with the genus level. However, the S. sinensis plastome is distinctive from other available Selaginella plastomes in that it does not have a GC bias, with GC content around 36.2% in line with most land plants (Figure S7; Table S10) (Smith, 2012). The sequence analysis reveals moderate A:T to G:C transitions in the S. sinensis plastome compared with other GC-rich Selaginella species (with a significant excess of A:T to G:C transitions), which would spontaneously cause GC content to equilibrate close to the level usually seen across land plants (Figure S8).

According to the phylogeny based on the combined nuclear DNA and ptDNA sequences, the S. sinensis group was resolved as a member of the Stachyandrum clade (Weststrand & Korall, 2016). Therefore, the plastome features of S. sinensis including GC content were derived...
from the common ancestor of *Selaginella*, which possessed DR structure and skewed GC content (>50%) (Zhang, Zhang, et al., 2019). Therefore, the similar nucleotide composition between the plastomes of *S. sinensis* and other land plants is independently evolved.

Generally, the GC content of organellar DNA is positively correlated with the frequency of C-to-U RNA editing events (Hecht et al., 2011; Malek et al., 1996; Tsuji et al., 2007; Wolf & Karol, 2012), which is supported by the observation in the *Selaginella* organellar genomes (Figures S9 and S10). The *S. uncinata* plastome has the highest GC content (54.8%) and most RNA editing events (3415) yet reported from any organism (Oldenkott et al., 2014). The *S. sinensis* plastome has the dramatically decreased GC content (36.2%) and lowest recorded editing frequency (approximately 148 sites) in *Selaginella*. Although the GC content of the *S. sinensis* plastome is the lowest in the genus, it falls well within the range of most land plants (Figure S7; Table S10) (Smith, 2009). As expected, the RNA editing frequency of the *S. sinensis* plastome is largely equivalent to those in other land plants. Theoretically, the mutational burden hypothesis is usually adopted to interpret the relationship between high mutation rate and reduced RNA editing frequency, assuming that the maintenance of proper editosome recognition sites imposes a mutational burden on an allele (Lynch et al., 2006). Although the extremely high mutation rate and the dramatically decreased RNA editing frequency in the *S. sinensis* plastome fit the hypothesis well, it could not interpret the high mutation rates and numerous RNA editing sites of most *Selaginella* organellar genomes. Retroprocessing is an RNA-mediated gene conversion model for the loss of RNA editing sites (Sloan et al., 2010). However, a comparison of the GC content of each gene before and after editing indicates that almost all the GC contents after editing in *S. uncinata* are still much higher than those of *S. sinensis* (Figure S9), which implies that the retroprocessing hypothesis might be not applicable here. Accordingly, we propose that the decreased GC-biased mutation in the *S. sinensis* plastome returns the normal GC content genome-wide, restores most editing targets on the DNA level, correspondingly reduces editing events on the RNA level. This interpretation is straightforward, which need not invoke more assumptions.

The nuclear-encoded PPR proteins, featured by tandem arrays of a weakly conserved 35 amino-acid motif (Small & Peeters, 2000), are widely accepted to be the recognition factors for RNA editing events (Cheng et al., 2016; Gutmann et al., 2020). Most angiosperms encode 400–600 PPRs (Fujii & Small, 2011), but there are great variations in liverworts ranging from 160 to 2700 (Dong et al., 2019). The positive correlation between the number of RNA editing sites and the number of PPR proteins has been widely reported from empirical and in silico data (Fujii & Small, 2011; Rüdinger et al., 2008; Rüdinger et al., 2012; Salone et al., 2007; Schallenberg-Rüdinger et al., 2014). It
has been demonstrated that reversion of an editing site could lead to the loss of the relevant editing factor (Hayes & Mulligan, 2011). Selaginella moellendorffii records the RNA editing events of ptDNA and mitochondrion DNA, also the greatest PPR protein family of land plants (Banks et al., 2011; Cheng et al., 2016; Hecht et al., 2011; Smith, 2009). If the positive correlation holds in Selaginella, the PPR protein family is expected to shrink in S. sinensis compared with S. moellendorffii. Our genome-wide investigation reveals that there are 803 PPR homologs in S. sinensis, but 1279 PPR homologs in S. moellendorffii (Table S14). This dramatic contraction of the PPR family annotates the evolutionary scenario well “If you don’t use it, you lose it.”

CONCLUSION

Selaginella sinensis has a dynamic network architecture in both its plastome and mitogome. The plastome is characterized by an extremely accelerated mutation rate, low GC content, and reduced RNA editing frequency, which are departures from organellar genomes of other Selaginella species. The imperfect organelle DNA-RRR system, with only intact dual-targeted recombinase genes present, shapes the convergent evolution between plastome and mitogome in the same species. The lacking of plastid-targeted recombinase genes, fortuitously meeting with the pervasive short repeats results in the lineage-specific plastome divergence in S. sinensis. We propose that the impact of the disrupted DNA-RRR system on the organellar genome evolution depends on the function of the disrupted (including missing) genes and the specific organellar genome features. Previous hypotheses either focus on the DNA-RRR system encoded by nuclear DNA or on repeats in the organellar genome respectively, and could not account for the complicated observations. Our study not only reconciles the longstanding contradictory opinions on the effects of disruption to the DNA-RRR system on genome evolution but also emphasizes the importance of integrating nuclear and organellar genome data to understand better the evolutionary mechanism of organellar genomes.

EXPERIMENTAL PROCEDURES

Sample collection, and DNA and RNA sequencing

Living plants of S. sinensis were collected from the field in Beijing, China. The voucher specimens of the collection (voucher numbers YM Zhu 203) were deposited in the Herbarium of Institute of Botany, CAS (PE). Total genomic DNA of S. sinensis was isolated from fresh plant tissues with the CTAB method as described in Li et al. (2013). Library construction was performed with the NEB-Next DNA Library Prep Kit (New England Biolabs, Ipswich, MA, USA). The sequencing was performed on an Illumina HiSeq 2000 platform (Illumina, San Diego, CA, USA) with approximately 11.42 Gb paired-end reads (the genome coverage was approximately 87×) generated from 270-bp insert libraries. A 10-kb SMRT cell library was constructed for PacBio Sequel sequencing (Pacific Biosciences, Menlo Park, CA, USA). Approximately 25.9 Gb sub-reads with an average >11 kb length were generated from two cells. All PacBio long-reads were self-corrected by Canu (v.2.2) (meronyThreads = 12, mbcuration = 2) and the genomeSize = 130 m minReadLength = 2000 minOverlapLength = 500 corOutCoverage = 120 corMinCoverage = 2 useGrid = false (Koren et al., 2017).

Total RNA was isolated using a RNAprep Pure plant kit (Tiangen Biotech Co., Ltd, Beijing, China) from fresh plant tissues. RNA quality and quantity were assessed with Qubit and Nanodrop spectrophotometer. Two cDNA libraries for Illumina sequencing were generated using TruSeq Stranded RNA sample prep kit (Illumina). One library enriched in poly(A) mRNA due to oligo(d-T) retro-transcription and one total RNA library after the depletion of rRNAs using Ribo-Zero™ rRNA Removal Kits (Plant) (Epicenter, Madison, WI, USA). The two libraries were sequenced on Illumina HiSeq X ten platform (2.57 Gb reads each library). The above-mentioned works of nucleotide extraction, library construction, and sequencing were performed at Beijing Biomarker Biotechnology Co., Ltd (Beijing, China).

Organellar genomes assembly

The plastome and mitogome assembly of S. sinensis were conducted as described in Kang et al. (2020). In brief, Illumina paired-end reads were initially mapped to plastome sequences of S. uncinata (NC_041575) and S. moellendorffii (MG272484), and mitogome sequences of H. squarrosa (NC_017755), I. engelmannii (FJ010859, FJ536259, FJ390841, FJ176330, FJ628360) and S. moellendorffii (JF338143–JF338147), using the Geneious platform (v.11.1.4, Map to Reference function, Sensitivity: Medium Sensitivity/Fast, Fine Tuning: Iterate 10 times) (Kearse et al., 2012). The mapped plastome or mitogome reads were then assembled into contigs using Velvet function in Geneious, respectively. In addition, GetOrganelle pipeline (https://github.com/Kinggerm/GetOrganelle) was also used to assemble the illumina reads into contigs (Jin et al., 2020). Then, the assembly contigs and connection pathways of plastome and mitogome were visualized by Bandage v.0.8.1 (Wick et al., 2015). Based on the depth of each set of contigs, the plastome of S. sinensis was assembled into one master genome and two subgenomes with unequal sizes in Bandage, and the final assembly was re-confirmed by mapping in Geneious. The filtered organellar PacBio reads were used to verify the connection nodes and tiling fragment pathways generated by Illumina data. In addition, Trinity (v.2.0.8) was used for de novo assembly of RNA-sequencing data of S. sinensis (Grabherr et al., 2011).

Quantification of repeat-mediated plastome rearrangements

The PacBio long-reads data were used to verify the assembly of the S. sinensis plastome, particularly the node regions where recombination could be mediated. The corrected PacBio long-reads were extracted in Geneious using the Illumina reads assembled plastome sequence as a reference, with medium-low sensitivity in 5–10 iterate. We checked all the long reads related with the three rearrangement hotspots (Region I, II, and III in Figure 1c), by assuming that homologous recombination occurred depending on whether they mapped consistently to the 24 conformations. To quantify the abundance of each conformation, the number of consistent reads spanning each region was counted.
Repeats finding

Repeat sequences of *S. sinensis* were identified using Geneious iFind. Repeat function with following parameter: Minimum repeat length: 16, Maximum mismatches: 0%. Exclude repeats up to 16 bp longer than contained repeat. Exclude contained repeats when longer repeat has frequency at least: 2, Maximum repeats [approximate] to find: 100.

Gene annotation of the organellar genomes

The local BLAST (v.2.2.30+; E-value < 10^{-5}) was used to perform the initial annotation of the *S. sinensis* plastome (Altschul et al., 1990). The putative positions and structures of plastid genes determined by making comparison with the plastomes of other *Selaginella* species, and mitochondrial genes compared with corresponding ones in *S. moellendorffii*, H. squarrosa, and *l. engelmannii*. However, only limited plastid genes were annotated successfully based on the initial annotation. Thus, three more methods were performed to obtain further results. First, PSI-BLAST and tBLASTX (E-value < 10^{-5}) were used against the NCBI NR database to annotate all potential protein-coding genes within all intergenic sequences, which were translated into all six potential reading frames. Second, possible genes were identified from the annotated open reading frames (ORFs) based on the stability of gene clusters and the conservatism of their locations. Meanwhile, some ORFs can be annotated in Geneious with following parameter: Minimum size: 100, resulting in the identification of some other genes according to the relative location and the length of ORFs. Lastly, tRNAscan-SE (Lowe & Eddy, 1997) (http://lowelab.ucsc.edu/tRNAscan-SE/) and Aragorn (Laslett & Canbäck, 2004) (http://mbio-serv2.mbioekol.lu.se/ArAGOrN/) were used to search and annotate tRNA genes. To investigate the missing genes further, the genes that present in closely related *Selaginella* species were used as a query to search against the transcriptomes (mRNA and total RNA) by BLASTP and BLASTN (E-value < 10^{-5}). The organellar genome maps (circular and linear) were generated in OGDRAW software (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html) (Lohse et al., 2007).

Nuclear-encoded DNA-RRR and PPR protein homologs search

For identifying DNA-RRR protein family, the pipeline was described in Kang et al. (2020) for RecA, OSB, and Why genes identification. In brief, we examined seven representative land plant species (Table 1), and protein databases (except *S. sinensis*, our data unpublished) of six species were downloaded from public database the Joint Genome Institute (JGI), https://phytozone.jgi.doe.gov/pz/portal.html, for *Arabidopsis thaliana*, *A. trichopoda*, *S. moellendorffii*, *S. tamariscina*, *Marchantia polymorpha*, and National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA388671, for *S. lepidophyllum*). DNA-RRR proteins families were detected by initial BLASTP (E-value < 10^{-5}) searches with *A. thaliana* DNA-RRR proteins families (download from NCBI) as seed sequences. The searched homologs were used to construct phylogenetic trees using IQ-TREE (detailed in “Experimental Procedures” section in this paper). Then, subcellular localization of each homolog was detected by TARGETP v.1.1 (Emanuelsson et al., 2007). Combining the two results from phylogeny and subcellular localization, we eventually identified the homologs of *Why*, *RecA*, and *OSB* families.

For identifying the PPR protein family, we integrated the search results of HHMmer (v.3.3.2) and motif analysis. First, the 474 PPR proteins of *A. thaliana* (Cheng et al., 2016), were used as seed sequences to search nuclear genome of *S. sinensis* by HHMmer package (E-value < 10^{-5}). Furthermore, we identified the DWY-type PPR by analyzing motif on the web WEBLOGO (http://weblogo.berkeley.edu/).

Phylogenetic analysis and substitution rate estimation

To estimate nucleotide substitution rates, 31 shared plastid protein-coding genes (atpA, atpE, atpF, atpH, atpI, petA, petB, petD, petG, petL, psaA, psaB, psaC, psaJ, psbA, psbC, psbD, psbE, psbF, psbH, psbI, psbK, psbL, psbT, rbcL, rps3, rps8, rps11, and rps19) of *S. sinensis* and other 20 land plants (including two bryophytes, 12 lycophytes, one gymnosperm, and six angiosperms) were used for phylogenetic analysis. For details, the protein-coding genes were aligned based on codon alignment mode and constructed using MAFFT with following parameters, Alignment Mode: Codon; Code Table: The Bacterial, Archaeal and Plant Plastid Code; Strategy: Auto (Katoh & Standley, 2013). The poorly aligned regions and the positions with gaps were removed using Gblocks (v.0.91b). Based on the concatenated data matrix of 31 plastid gene sequences, maximum likelihood (ML) analyses were conducted using IQ-TREE (v.1.6.8) (Nguyen et al., 2015) with 1000 bootstrap replicates and the GTR+G model selected by ModelFinder (Kalyaanamoorthy et al., 2017). We used the CODEML module in the PAML (v.4.9) (Yang, 2002) to calculate dN and dS based on the concatenated data matrix and the ML tree topology with branch model = 1 and run mode = 0. Because of the great sequence divergence of *S. sinensis* and the expected long branch attraction during the phylogenetic analyzing, its position was adjusted manually according to the *Selaginella* phylogeny based on the result of the combined nuclear DNA and ptDNA sequences (Weststrand & Koralli, 2016). For mitochondria, 16 mitochondria shared protein-coding genes (atp1, atp6, atp8, cob, cox1, cox2, cox3, nad1, nad2, nad3, nad4, nad4L, nad5, nad6 and nad9) were used to analyze the substitution rate using the same method.

In addition, to estimate the point mutation types and mutation bias of the plastome sequence, we compared 37 shared protein-coding genes (atpA, atpB, atpE, atpF, atpH, atpI, ccsA, clpP, petA, petB, petD, petG, petL, psaA, psaB, psaC, psaJ, psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbK, psbL, psbM, psbT, psbZ, rbcL, rpl22, rps3, rps8, rps11, and rps19) of *S. sinensis*, *S. moellendorffii*, *S. vardei*, *S. tamariscina*, *S. sanguinolenta*, and *S. hainanensis* with *l. engelmannii* as the reference. The mutation types were divided into transitions (A:T to G:C and C:G to A:T) and transversions (A:T to C:G and G:C to T:A).

GC content and RNA editing identification

The GC content of plastome and mitogenome in different species was calculated in Geneious. We calculated the GC content of overall region, shared coding regions (CDS), intergenic regions, canonical R region in other seven lycophytes and three different regions in *S. sinensis*. Forty-three shared protein-coding genes (except for *rrn5*, *rrn5*, *rrn6*, *rrn23*, and *psaM* of 48 plastid genes in *S. sinensis*) of each species were concatenated and were used to calculate the GC content.

To obtain RNA editing information in the plastome and mitogenome of *S. sinensis*, the transcriptome data were mapped to the CDS sequence of each gene in Geneious. The editing types and sites were checked and counted manually. In addition, PRE-PACT 3.0 (Lenz et al., 2018) (http://www.prepect.de/prepect-main.php) was used to estimate and identify the type, locus, and number of the RNA editing sites of each gene. A comparative analysis was made between *S. sinensis* and *S. uncinata* that the detailed...
information of RNA editing events had been published (Oldenkott et al., 2014).

**Plastome rearrangement estimation and regression analyses**

Plastome rearrangement was evaluated based on shared gene order. Multiple plastome alignment of the 13 species (described in Figure 3) was performed using the progressive Mauve algorithm (Darling et al., 2010). The orders of 31 shared plastome gene were used to generate a FASTA format file with a custom Perl script for genome rearrangement estimation. The rearrangement measure, reversal distance, was estimated by comparing gene order between 11 Selaginella species and *H. serrata* respectively, using the online web server Common Interval Rearrangement Explorer (CREx) (Bernt et al., 2007). We used the analyzing kit provided by Excel to calculate the correlation between the number of repeats (medium repeats and short repeats) and recombination, between the number of repeats and the number of nucleotide substitutions. The *F*-test was applied for significance test.

**ACKNOWLEDGEMENTS**

We thank Dr. Shan-Shan Dong, Dr. Yang Liu, and Dr. Hong-Rui Zhang for suggestions on data analysis, Dr. Ran Wei for comments on the preliminary manuscript. We thank the editor and the two anonymous reviewers for their professional and constructive comments. This work was supported by the National Natural Science Foundation of China (31770237, 32170233), and the Beijing Municipal Natural Science Foundation (8202019).

**AUTHOR CONTRIBUTIONS**

Q-PX conceived and supported the study, designed the experiment, wrote the manuscript with input from X-CZ, J-YT, J-GY, Y-MZ, and DRS, and revised the manuscript. J-YT, J-GY, Y-MZ, Y-RW, J-SK, and YJ collected and analyzed the data. J-YT revised the manuscript. J-YT and J-GY, Y-MZ, and DRS, and revised the manuscript. J-GY, Y-MZ, Y-RW, J-SK, and DRS, and revised the manuscript. J-GY, Y-MZ, and DRS, and revised the manuscript. J-GY, Y-MZ, and DRS, and revised the manuscript.

**CONFLICT OF INTERESTS**

The authors declare that they have no competing interests.

**DATA AVAILABILITY STATEMENT**

The genomes’ sequences have been deposited at the GenBank under the accession numbers ON479702–ON479711. The detailed annotations of the plastome and mitogenome are deposited in the online public database FigShare (https://doi.org/10.6084/m9.figshare.19743325).

**SUPPORTING INFORMATION**

Additional Supporting Information may be found in the online version of this article.

**Figure S1.** One possible transforming scheme of the *Selaginella sinensis* plastome. The circular subgenome C1 + C2 (81 010 bp; average coverage 780.1×) contains all contigs, and circular subgenomes C1 (37 604 bp; average coverage 698.3×) and C2 (43 521 bp; average coverage 815.9×) contain partial contigs. Subgenomes C1 and C2 could be further assembled into larger circular-mapping molecules C1 + C2, based on the existence of the conserved regions.

**Figure S2.** Gene content comparison of plastomes among *Selaginella sinensis* and other lycophytes.

**Figure S3.** Gene content comparison of mitogenomes among *Selaginella sinensis* and other lycophytes.

**Figure S4.** The positions of genes and repeats in surveyed plastomes. Each plastome was represented by a black line, the upper scale indicated the position of plastid genes, the green blocks indicated the annotated genes, the gene name was displayed in the blocks, and the arrow indicated the gene direction.

**Figure S5.** Substitution rate divergence of 31 protein-coding genes of plastomes.

**Figure S6.** Substitution rate divergence of 16 protein-coding genes of mitogenomes.

**Figure S7.** The GC content of the sequenced plastomes in land plants.

**Figure S8.** Comparison of the nucleotide mutation bias among plastomes of *Selaginella sinensis* and other Selaginella species.

**Figure S9.** Comparison of RNA editing events and GC contents before and after editing between *Selaginella sinensis* and *Selaginella uncinata*.

**Figure S10.** Comparison of RNA editing events and GC contents before and after editing between *Selaginella sinensis* and *Selaginella moellendorfii*.

**Figure S11.** The ML phylograms of seven representative species based on the amino acid sequences of functional domains of three DNA-RRR genes using IQ-TREE.

**Table S1.** The Illumina data assembly information of the *Selaginella sinensis* plastome.

**Table S2.** Verification of possible conformations of the *Selaginella sinensis* plastome by PacBio data at Region I, II, and III.

**Table S3.** Basic information of the master genome and subgenomes of the *Selaginella sinensis* plastome.

**Table S4.** The information of plastid genes in *Selaginella*.

**Table S5.** The assembly information of *Selaginella sinensis* mitogenome contigs.

**Table S6.** The information of 8 scaffolds of the *Selaginella sinensis* mitogenome.

**Table S7.** The information of repeats of *Selaginella* plastomes.

**Table S8.** Median substitution rates of 31 shared plastid protein-coding genes between *Selaginella* species and outgroup.

**Table S9.** Median substitution rates of 16 shared mitochondria protein-coding genes between *Selaginella* species and outgroup.

**Table S10.** The GC content of plastomes used to plot in Figure S7.

**Table S11.** The GC content of different parts of plastomes in *Selaginella sinensis* and other lycophytes.

**Table S12.** The information of plastid genes with RNA editing events.

**Table S13.** Substitution rate divergence of 16 protein-coding genes of mitogenomes.

**Table S14.** Substitution rate divergence of 31 protein-coding genes of plastomes.

**REFERENCES**


Repeats and DNA-RRR drive genome evolution


