# Integrated analysis of four new sequenced fern Chloroplast Genomes: Genome structure and Comparative Analysis

**Ruifeng Fan**
Heilongjiang University of Chinese Medicine https://orcid.org/0000-0002-6113-733X
**Wei Ma**
Heilongjiang University of Chinese Medicine
**Shilei Liu**
Heilongjiang University of Chinese Medicine
**Qingyang Huang** ( ruifeng1021@126.com )

**Research article**

**Keywords:** fern, chloroplast genome, IR border, phylogenetic, RNA editing

**Posted Date:** December 4th, 2019

**DOI:** https://doi.org/10.21203/rs.2.18299/v1

# Abstract

Background: Dryopteris goeringiana (Kunze) Koidz, Arthyrium brevifrons Nakai ex Kitagawa, Dryopteris crassirhizoma Nakai, and Polystichum tripteron (Kunze) Presl are fern species with their chloroplast genomes uncovered, except D. crassirhizoma . In this study, high throughput sequences were performed to get better understand of their chloroplast genomes and comparison analysis with other fern species.

Methods: Fresh fern leaves were used for sequencing. Simple Sequence Repeat (SSR) analysis, high-variety regions comparison, IR border comparison, nucleotide diversity, RNA editing, and phylogenetic analysis were performed.

Results: They have different genome structure characteristics in terms of genome and its area size, gene types and location. The quantity and type of SSRs in D. crassirhizoma is very similar to D. goeringiana , with (ATAA)2, (ATCT)1 and (TTTA)1 only present in D. crassirhizoma . SSRs contain more AT bases than GC. There were divergent genes existed in fern species. Ten genes have a Pi value >0.20. C-to-U RNA editing was most common. Phylogenetic analyses showed that species in the same genus clustered together.

Conclusion: The genomic structure and genetic resources of D. goeringiana , A. brevifrons , D. crassirhizoma , and P. tripteron contribute to further studies on phylogenetics, population genetics, and conservation biology of ferns.


# Introduction

With the development of next generation sequencing (NGS) technology, the details of the most subtle nuclear gene components in eukaryotic cells become more clearer, and the cytoplasmic organelle genome is also facilitated in a simpler and more time-saving way [1]. Especially chloroplast, the organelle that involved in biochemical processes, including amino acid, sugar, lipid, vitamin, starch and pigment syntheses, sulfate reduction and nitrogen metabolism, and for the most important, it could convert light energy into chemical energy through photosynthesis in plants [2–4]. Chloroplast genomes are relatively conserved in terms of gene order, gene content and substitution rate when compared with nuclear DNA [5–9]. Most chloroplast obtains a typically circular structure composed including one large single-copy region (LSC), one short single-copy (SSC) and two inverted repeats (IRs), with its length ranging from 120 to 170 kilobases (kb). In recent years, chloroplast genome has become a valuable and ideal resource for species identification, population genetics, plant phylogenetics and genetic engineering, giving for their similar structures, highly conserved sequences and stable maternal heredity way [10]. However, gain and loss of gene, duplications of gene content and rearrangements in gene order appear to be phylogenetically informative, and reflect species differentiation events [2, 6, 7].

An increasing number of chloroplast genomes have been reported in recent years especially when the NGS becomes more cheaper and faster. The chloroplast genomes of many plants have been sequenced, including bryophytes [11, 12], lycophytes [13, 14], monilophytes [1, 15–18] and seed plants [6, 19],

especially in seed plants that have attracted many studies. As one of the largest groups of vascular plant, ferns contain approximately 2129 species in China [20]. Ferns have great potential functions for their medicinal characteristics, which could be used as treatment for several illnesses [21]. Genus Dryopteris (Dryopteridaceae) comprised with 225 - 300 species, and mostly living in temperate forests and montane areas, was considered as ideal material for addressing questions about diversification, hybridization and polyploidy [22]. Polystichum (Dryopteridaceae) is one of the largest genera of ferns commonly occurred in temperate and subtropical regions, in lowlands and montane to alpine areas [23], which contains of 500 species, among which 208 species known to occur in China [24]. Athyrium Roth (Athyriaceae), the lady-fern genus, contains about 220 described species [25]. Moreover, *A. brevifrons* is often used as a wild vegetable in northeastern China, with high nutritional value.

According to the former studies, we have known that there were only 60 ferns for which complete chloroplast genomes have been reported [16, 18]. There are still many common ferns needed to be sequenced and reported, including *D. goeringiana, A. brevifrons,* and *P. tripteron.* Hence, chloroplast genome sequencing and comparative analysis about these three ferns species as well as *D. crassirhizoma* were performed. Moreover, the comparison analysis of simple sequence repeats (SSRs), nucleotide diversity, RNA editing events and phylogenetic analysis were conducted. We aimed to further uncover the chloroplast genomes in *D. goeringiana, A. brevifrons, D. crassirhizoma,* and *P. tripteron,* and to obtain clear comparison when compared with other reported ferns species.

## Results

# Chloroplast DNA sequencing and genome features

Species of *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P. tripteron* produced 30,640,836 (30955× coverage), 34,833,094 (34754× coverage), 26,981,820 (27353× coverage) and 29,722,892 (30215× coverage) paired-end (150 bp) raw reads, respectively (Table 1). The chloroplast genome sizes of *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P. tripteron* were 149,468 bp, 151,341 bp, 148,947 bp and 148539 bp, respectively (Table 1, Figure 1). The chloroplast genomes displayed a typical quadripartite structure, including LSC (82,504 bp, 82,459 bp, 82,384 bp, and 82,799 bp), SSC (21,600 bp, 21,708 bp, 21,623 bp and 21,660 bp) and IRa/IRb (22,682 bp, 23,588bp, 22,471 bp and 22,040 bp) in *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P. tripteron,* respectively. The genome GC content among four species ranged from 42.40 to 43.76 (Table 2).

**Table 1.** Sequencing data and summary of complete chloroplast genomes for *D. crassirhizoma*, *A. brevifrons*, *D. goeringiana* and *P. tripteron*.

| Items | Species | | | |
|---|---|---|---|---|
| | *D. crassirhizoma* | *A. brevifrons* | *D. goeringiana* | *P. tripteron* |
| Sequencing | | | | |
| Total Reads | 30640836 | 34833094 | 26981820 | 29722892 |
| Total Bases (bp) | 4626766236 | 5259797194 | 4074254820 | 4488156692 |
| Q30 (%) | 89.06 | 89.07 | 90.97 | 91.52 |
| GC content (%) | 41.32 | 42.57 | 42.42 | 42.26 |
| LSC | | | | |
| Length (bp) | 82,504 | 82,459 | 82,384 | 82,799 |
| Percentage (%) | 55.20 | 54.49 | 55.31 | 55.74 |
| SSC | | | | |
| Length (bp) | 21,600 | 21,708 | 21,623 | 21,660 |
| Percentage (%) | 14.45 | 14.34 | 14.52 | 14.58 |
| IR | | | | |
| Length (bp) | 22,682 | 23,588 | 22,471 | 22,040 |
| Percentage (%) | 15.18 | 15.59 | 15.09 | 14.84 |
| Total | | | | |
| Total length (bp) | 149468 | 151341 | 148947 | 148539 |
| GC content (%) | 43.19 | 43.76 | 43.12 | 42.40 |

From up to down: LSC: Large single copy; SSC: small single copy; IR: Inverted repeat.

**Table 2.** Comparison of coding and non-coding regions size among *D. crassirhizoma*, *A. brevifrons*, *D. goeringiana* and *P. tripteron*.

| Items | Species | | | |
|---|---|---|---|---|
| | *D. crassirhizoma* | *A. brevifrons* | *D. goeringiana* | *P. tripteron* |
| Protein coding genes | | | | |
| Gene numbers | 89 | 89 | 89 | 89 |
| Gene total length (bp) | 81927 | 81984 | 81921 | 81774 |
| Gene average length (bp) | 920 | 921 | 920 | 918 |
| Gene length / Genomes (%) | 54.8 | 54.2 | 55 | 55.1 |
| tRNA genes | | | | |
| Gene numbers | 32 | 33 | 33 | 31 |
| Gene total length (bp) | 2363 | 2388 | 2282 | 2259 |
| Gene average length (bp) | 73 | 72 | 69 | 72 |
| Gene length / Genomes (%) | 1.58 | 1.58 | 1.53 | 1.52 |
| rRNA genes | | | | |
| Gene numbers | 8 | 8 | 8 | 8 |
| Gene total length (bp) | 9060 | 9062 | 9060 | 9062 |
| Gene average length (bp) | 1132.5 | 1132.75 | 1132.5 | 1132.75 |
| Gene length / Genomes (%) | 6.06 | 5.99 | 6.09 | 6.11 |

From up to down: tRNA, transfer RNA; rRNA, ribosomal RNA.

Four fern species were composed by 89 protein-coding genes, eight ribosomal RNA (rRNA) and different amount (38, 38, 36, and 35 in *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P. tripteron,* respectively) of transfer RNA (tRNA) genes. After removing the duplications, there were 84 protein-coding genes, four rRNA and different amount (32, 33, 33, and 31 in *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P.*

*tripteron*, respectively) of tRNA genes remained. Five protein-coding genes are duplicated, including *ndhB*, *psbA*, *rps12*, *rps7*, and *ycf2* in IR regions (Additional file 1). The duplicated tRNA are mainly *trnL-UAA*, *trnA-UGC*, *trnI-GAU*, *trnG-UCC*, with some differences in different species (Additional file 2).

The LSC region was composed of 67 protein-coding and 20 tRNA genes in *D. crassirhizoma;* 66 protein-coding and 21 tRNA genes in *A. brevifrons*, *D. goeringiana* and *P. tripteron.* SSC region has 15 protein-coding and 2 tRNA genes in *D. crassirhizoma*, *D. goeringiana*, and *P. tripteron;* 15 protein-coding and 3 tRNA genes in *A. brevifrons.* In the IR regions, they obtained 10 protein-coding and 8 rRNA genes. However, the tRNA genes differed in them, with 10 in *D. crassirhizoma*, 8 in *A. brevifrons*, 10 in *D. goeringiana*, and 8 in *P. tripteron* (Table 3). trnN-GUU in SSC region was only presence in *A. brevifrons*, and was absent in *D. crassirhizoma*, *D. goeringiana*, and *P. tripteron.* The result suggests a possible absence of trnI-GAU in the chloroplast DNA IRb region of *P. tripteron (*Table 3).

**Table 3.** List of genes distributed in different regions of *D. crassirhizoma*, *A. brevifrons*, *D. goeringiana* and *P. tripteron* chloroplast genomes.

| Items | | Species | | | |
|---|---|---|---|---|---|
| | | *D. crassirhizoma* | *A. brevifrons* | *D. goeringiana* | *P. tripteron* |
| LSC | | | | | |
| Protein coding genes | shared | *rpl23, rpl2, rps19, rpl22, rps3, rpl16, rpl14, rps8, infA, rpl36, rps11, rpoA, petD, petB, psbH, psbN, psbT, psbB, clpP, rps12, rpl20, rps18, rpl33, psaJ, petG, petL, psbE, psbF, psbL, psbJ, petA, cemA, ycf4, psaI, accD, rbcL, atpB, atpE, ndhC, ndhK, ndhJ, rps4, ycf3, psaA, psaB, rps14, psbD, psbC, psbZ, petN, psbM, rpoB, rpoC1, rpoC2, rps2, atpI, atpH, atpF, atpA, ycf12, psbI, psbK, chlB, rps16, matK, ndhB* | | | |
| tRNA genes | shared | *trnI-CAU, trnP-UGG, trnW-CCA, trnM-CAU, trnV-UAC, trnF-GAA, trnL-UAA, trnS-GGA, trnfM-CAU, trnT-GGU, trnS-UGA, trnG-GCC, trnC-GCA, trnE-UUC, trnY-GUA, trnD-GUC, trnR-UCU, trnS-GCU, trnQ-UUG* | | | |
| | varied | *trnG-UCC* | *trnG-UCC-1, trnG-UCC-2* | *trnG-UCC-1, trnG-UCC-2* | *trnG-UCC-1, trnG-UCC-2* |
| SSC | | | | | |
| Protein coding genes | shared | *ndhF, rpl21, rpl32, ccsA, ndhD, psaC, ndhE, ndhG, ndhI, ndhA, ndhH, rps15, ycf1, chlN, chlL* | | | |
| tRNA genes | Same | *trnP-GGG, trnL-UAG* | | | |
| | varied | - | *trnN-GUU* | - | - |
| IR regions | | | | | |
| Protein coding genes | shared | *ycf2×2, psbA×2, rps7×2, rps12×2, ndhB×2* | | | |
| tRNA genes | shared | *trnH-GUG×2, trnA-UGC×2, trnR-ACG×2* | | | |
| | varied | *trnN-GUU×2* | *trnI-GAU×2* | *trnN-GUU×2* | *trnN-GUU×2* |
| | | *trnI-GAU×2* | | *trnI-GAU×2* | |
| rRNA genes | shared | *rrn16×2, rrn23×2, rrn4.5×2, rrn5×2* | | | |

We found 14 intron-containing genes in all these four chloroplast genomes, 11 of which contained one intron, the other three genes (*clpP*, *rps12* and *ycf3)* contained two introns in *D. crassirhizoma*, *A. brevifrons*, *D. goeringiana*. *matK* has two introns in *P. tripteron*, but have one in *D. crassirhizoma*, *A. brevifrons* and *D.*

*goeringiana* (Table 4). Especially *rps12,* which contains one exon in LSC region and the other 2 reside in IR regions, was considered as a trans spliced gene separated by two introns.

**Table 4.** List of genes contain introns of seven fern chloroplast genomes.

| Genes contain introns | Species | | | | | | |
|---|---|---|---|---|---|---|---|
| | *D. crassirhizoma* | *A. brevifrons* | *D. goeringiana* | *P. tripteron* | *P. amoena* | *D. fragrans* | *A. gigantea* |
| *atpF* | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *clpP* | 2 | 2 | 2 | 2 | 3 | 1 | 2 |
| *matK* | 1 | 1 | 1 | 2 | | | |
| *ndhA* | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *ndhB* | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *NdhE* | | | | | | 1 | |
| *NdhF* | | | | | | 1 | |
| *NdhG* | | | | | | 1 | |
| *petA* | 1 | 1 | 1 | 1 | | | |
| *petB* | 1 | 1 | 1 | 1 | 1 | | 1 |
| *petD* | 1 | 1 | 1 | 1 | 1 | | 1 |
| *rpl16* | 1 | 1 | 1 | 1 | 1 | | 1 |
| *rpl2* | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *rpl20* | | | | | | 1 | |
| *rpoC1* | 1 | 1 | 1 | 1 | 1 | | 1 |
| *rpoB* | | | | | | 1 | |
| *rps12* | 2 | 2 | 2 | 2 | 3 | | 2 |
| *rps16* | 1 | 1 | 1 | 1 | 1 | | 1 |
| *ycf3* | 2 | 2 | 2 | 2 | 3 | | 2 |
| *trnG-UCC* | | | | | 1 | | 1 |
| *trnV-UAC* | | | | | 1 | | 1 |
| *trnA-UGC* | | | | | 1 | | 1 |
| *trnI-GAU* | | | | | 1 | | 1 |
| *trnL-UAA* | | | | | 1 | | |
| *trnT-UGU* | | | | | 1 | | |
| *trnL-CAA* | | | | | | | 1 |
| *psaA* | | | | | | 1 | |
| *cemA* | | | | | | 1 | |

From up to down: *P. amoena*, *Polypodiodes amoena*; *D. fragrans*, *Dropteris fragrans*. *A. gigantea*, *Alsophila gigantea*.

# SSR analysis

In this study, perfect SSRs in the chloroplast genomes were detected, and 75, 108, 69 and 108 microsatellites were found in *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P. tripteron,* respectively (Figure 2a). The majority of the SSRs in these chloroplast genomes consist of mononucleotide repeat motifs, varying from 55 in *D. goeringiana* to 90 in *P. tripteron* (Figure 2b). Dinucleotide SSRs are the second most common, ranging from 8 in *D. goeringiana* and *P. tripteron* to 12 in *A. brevifrons (*Figure 2c). Furthermore, two trinucleotide SSRs are present in *D. crassirhizoma, A. brevifrons* and *D. goeringiana,* with three in *P. tripteron* (Figure 2d). Additionally, 8 pentanucleotide SSRs were found in *D. crassirhizoma,* 9 in *A. brevifrons,* 4 in *D. goeringiana,* and 6 in *P. tripteron* (Figure 2e). Hexanucleotide repeats were only found in *A. brevifrons* and *P. tripteron,* with 2 in *A. brevifrons* and 1 in *P. tripteron* (Figure 2f). The kind of SSRs in *D. crassirhizoma* is very similar with that in *D. goeringiana,* but was vaster, with (ATAA)2, (ATCT)1 and (TTTA)1 were only present at *D. crassirhizoma,* compared with *D. goeringiana* (Figure 2e). Only 6 SSRs that were appeared in these four ferns simultaneously, they were A, C, G, T, AT and AGAT. Moreover, SSRs were mostly composed with AT bases than GC.

# Structural comparative assessment of chloroplast genomes

Chloroplast genomes of other four ferns (*Cyrtomium devexiscapulae, Dryopteris decipiens, Lepisorus clathratus* and *Polypodium glycyrrhiza)* were selected for the comparison analysis with *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P. tripteron.* Among these chloroplast genomes, *L. clathratus* was the largest (156,998 bp), whereas *P. glycyrrhiza* had the smallest chloroplast genomes size (129,221 bp). Within the genus of Dryopteris, the genomes of chloroplast ranged from 148,974 to 150,987, had a 2,013 bp difference. However, the length of chloroplast genomes differed greatly in genus Polypodium, with 151,341 bp in *A. brevifrons,* and 129,221 bp in *P. glycyrrhiza,* had a 22,120 bp difference. Additionally, *rpoC2, rpoB, psbC, pasA, rbcL, ycf2, ycf1* and *ndhB* were identified to be divergent among these chloroplast genomes (Figure 3).

# LSC, SSC and IR border regions analysis

The study of IR border positions and the adjacent genes among chloroplast genomes were performed to obtain the detailed comparison in *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P. tripteron, L. clathratus, D. decipiens* and *C. devexiscapulae* (Figure 4). In this study, IR region size in *L. clathratus* (27, 112 bp) was higher than others, with the least difference of about 3,524 bp. Other than that, IR region in other six ferns were similar with each other, ranging from 22,040 bp to 23,588 bp, where some IR expansion and contraction were still observed.

The *trnI-CAU* was located in the LSC border, 40 bp in *D. goeringiana* and *D. decipiens,* and 53 bp in *D. crassirhizoma* away from the LSC/IRb border. The *trnI-CAU* was 66 bp, 69bp, 111 bp and 112 bp away from the LSC/IRb border in *A. brevifrons, L. clathratus, C. devexiscapulae* and *P. tripteron,* respectively. The IRb/SSC junction was location in the *ndhF* region in all these species chloroplast genomes, except *D. goeringiana,* with 18bp to 52bp extended into IRb region. The SSC/IRa junction was location in the *chlL*

region in all these species, also except *D. goeringiana,* with 47bp to 67bp extended into IRa region. *D. goeringiana* presented opposite gene order in the junctions of SSC/IRa and IRb/SSC with six species. *ndhB* was located in the IRa/LSC border, with 376 bp away from LSC region in genus Dryopteris, as well as 299 bp, 369bp and 304 bp in *P. tripteron, A. brevifrons,* and *C. devexiscapulae,* respectively. However, *ndhB* extended into the LSC region in *L. clathratus.* The result is shown in Figure 4.

# Nucleotide diversity analysis

Chloroplast genomes sequences contain nucleotides that are highly variable, which is helpful for the screening of suitable loci. These suitable loci are valuable for resolving closely related species or genera, as well as DNA barcoding. We found the *Pi* value were ranged from 0.0000 to 0.2778 (rpl16) through the comparative analysis. The genes (*trnM-CAU, trnE-UUC, psb*Z, *trnN-GUU, trnI-CAU, rpl21, psbM, rpl32, trnV-UAC,* and *rpl16)* with *Pi* value > 0.20 were selected as highly variable loci. The result is shown in Figure 5.

# RNA editing

A total of 268 RNA editing events were found in four sequenced chloroplast genomes in this study, including 85 in *D. crassirhizoma,* 55 in *A. brevifrons,* 50 in *D. goeringiana,* and 78 in *P. tripteron.* Among all the 268 RNA editing events, the C-to-U and mutations were most common, reaching 120 (44.8%), followed by U-to-C 103 (38.4%), A-to-G 36 (13.4%) and G-to-A 9 (3.4%) of all the RNA editing events.

# Phylogenetic analysis

We investigated the phylogenetic relatedness among the chloroplast genomes of 43 fern species, among which 35 nodes with support values greater than 90%, containing 27 nodes with support values of 100%. The fern species in the same genus were clustered together to a certain degree. *D. crassirhizoma, D. goeringiana* were closer to *D. decipiens. P. tripteron* was identified as a sister genus to *C. devexiscapulae.* It was worth noting that *A. brevifrons* formed a single clade with *Athyrium sinense,* instead of *P. glycyrrhiza (*Figure 6).

## Discussion

This study reported four complete chloroplast genomes of *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P. tripteron* for the first time, ranging from 148,539 bp to 151,341 bp in length. The whole chloroplast genomes size of more than 20 ferns were compared in the study of Gao et al [26] and RuizRuano et al [1], which revealed that the length of ferns ranging between 131,760 bp (*Equisetum hyemale)* and 181,684 bp (*C. devexiscapulae)* based on the current researches, and the four ferns detected in our study is within this range. The genomes of these four ferns chloroplast exhibit a typical quadripartite structure, including LSC, SSC and two IRs, as reported in other spore plants [27, 28]. Additionally, the gene number and order were largely similar with other ferns, but there were still some differences among species, which should let the

length variation in the IR and SSC regions take the responsibility [29]. We compared the chloroplast genomes of *D. crassirhizoma, D. goeringiana* and *D. fragrans* [26], and found a genomes loss of 4033 bp in *D. fragrans,* which caused a longest SSC and shorter IRs. Therefore, the *D. fragrans* was found has dispersed gene distribution and longer sequence length, because of more intergenic sequences. As demonstrated by Gao et al [26] and Lu et al [16], they revealed that fern chloroplast genome extends the intergenic sequences but reduces the overlapping genes, thus, sequence utilization is more specific and genes are more independent.

Four fern genomes encoded 132–135 genes, composed with 89 protein-coding, eight ribosomal RNAs and 35–38 transfer RNA genes in this study. The genes *clpP, ycf3,* and *rps12* were found containing two introns, was also reported in many other ferns, as *Alsophila podophylla* [30] and *Alsophia gigantea* [31]. *matK* only in *P. tripteron* was found had two introns in this study. It was worth noting that, a plastid research of *D. crassirhizoma* detected by Xu et al [32] had the different result with this study. They demonstrated that *ycf2, rpoB, clpP* and *ndhF* encoded four, three, two and two introns, respectively. As reported in other angiosperm cp genomes [29, 33], rps12 is a coding unequally divided gene with its 5' terminal exon located in the LSC region, while two copies of the 3' terminal exon and intron are in IRs.

SSRs with its length ranging from 1–6 or more base pairs, play an important role in genetic molecular markers for population genetics [34, 35] and are also widely applied for plant genotyping [36, 37]. In this study, the number of SSRs ranging from 69 to 108, with lowest in *D. goeringiana* and highest in *P. tripteron.* The quantity and type of SSRs showed a high similarity in *D. crassirhizoma* and *D. goeringiana,* which might for the same genus they were belonging. Although *A. brevifrons* and *P. tripteron* obtained the same quantity of SSRs, but the type of SSRs differed largely.

It had been reported that chloroplast genomes SSRs are commonly composed of short polyadenine polyadenine (Poly A) or polythymine (poly T) repeats, but rarely tandem guanine (G) or cytosine (C) repeats, which is consistent with this study [38]. However, this conclusion is not working in *D. fragrans,* which have high GC content than AT in SSRs. Taken the living environment into consideration, the author speculated that high GC content of repeat structures may allow *D. fragrans* coping with heat and large temperature differences [26]. The kind and amount of SSRs in *D. fragrans, D. crassirhizoma* and *D. goeringiana* exist great differences though they all belonging to the Dryopteridaceae family. This implies that great differences exist within species of the same family, which might for the different conditions they survive.

We also revealed that the sequence of chloroplast genome was similar with each other in this study and other four fern species (Figure 3). However, relatively small variation was observed within them in several comparable genomic regions. Furthermore, LSC and SSC regions were less similar than IR regions as previously reported [39, 40]. Similar result was also reported in chloroplast genome of higher plant, which demonstrated that the lower sequence divergence in IR regions was possibly owing to copy correction between IR sequences by gene conversion when compared with SSC and LSC regions [41]. Moreover, some previous reports revealed that LSC region contains most of the divergent genes, and display a trend

towards more rapid evolution [29, 32]. The divergent regions included *rpoC2, rpoB, psbC, pasA, rbcL, ycf2, ycf1* and *ndhB* in this study, as reported previously [32].

In spite of high conservation in IR regions, the expansion and contraction are still take responsible for variations in chloroplast genome size and rearrangement [39, 42], therefore, play a vital role in evolution [15, 43]. *chlL* and *ndhF* extended into IR regions with different length. All these ferns have relatively similar boundary characteristics with the expansion and contraction of IR regions except *D. goeringiana,* which presented opposite in the junctions of SSC/IRa and IRb/SSC with other six fern species in our study and other study [32].

The nucleotide diversity analysis also demonstrated that genes located in IR regions are less variable than in the SC regions. Additionally, genes with *Pi* value > 0.20 were mainly located at SC regions. However, none of the genes with intron (*atpF, clpP, matK, ndhA, ndhB, petA, petB, petD, rpl16, rpl2, rpoC1, rps12, rps16,* and *ycf3)* have a *Pi* value > 0.20, except *rpl16,* in other word, we can speculated that variability of genes without intron was generally higher than genes with intron regions. The result was consistent with a study of plastome in ferns [18].

It was reported that fern and hornwort chloroplast genomes have a relatively high levels of RNA editing compared with seed plants [44, 45], whereas only 30–40 RNA editing sites typically occurred. Most editing events in these four chloroplast genomes were C-to-U events in this study, which is consistent with *D. fragrans* [26]. It has been reported that the excess of C-to-U RNA editing developed in early stages of vascular plant evolution [46].

Chloroplast genome data are valuable for resolving species definitions since organelle-based “barcodes” can be established for certain species and then applied to unmask interspecies phylogenetic relationship [47]. The phylogenetic relationships of *A. brevifrons, D. goeringiana* and *P. tripteron* were rarely been studied for there was none studied had put focus on the chloroplast of them. In our study, we combined 43 species of ferns participated in the phylogenetic tree analysis, which contained almost all the ferns that been studied before. We found that *D. crassirhizoma, D. goeringiana* were closer to *D. decipiens. P. tripteron* was identified as a sister genus to *C. devexiscapulae.* Interestingly, *D. decipiens* and *C. devexiscapulae* were found clustered into one branch in the study of Wei et al [18], and have close relationship with *D. crassirhizoma* in the study of Xu et al [32]. According to the result, we could guess that *D. crassirhizoma, D. goeringiana, D. decipiens, P. tripteron,* and *C. devexiscapulae* have closer relationship with each other (Figure 6).

# Conclusions

In this study, we conducted and compared chloroplast genome skimming of *D. goeringiana, A. brevifrons, D. crassirhizoma* and *P. tripteron.* Through the study, we found that they all obtain the typical quadripartite structure with slightly differences in genes and gene orders. SSR features differed largely among different species, so it might be able to be used in developing molecular markers for genetic diversity and molecular identification. The genomes in this study also show some variations and IR region expansion and

contraction when compared with the other fern species. In addition, nucleotide diversity provides a way to study the genetic variation of the four species. Phylogenetic analysis gives a suggestion of the relationship between ferns species. In conclusion, the genomic characteristics, comparison and genetic resources presented in this study contribute to further studies on phylogenetics, population genetics, and conservation biology of ferns.

## Materials And Methods

# Materials

Wild fern species of *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P. tripteron* were collected from Maoer Mountain, Maoershan Town, Shangzhi City, Heilongjiang Province (N 45°17’51.45”, E127°36’00.03”), China. The four species were identified by Ruifeng Fan from the Heilongjiang University of Chinese Medicine. Voucher specimens were deposited in the Northeast Agricultural University Herbarium with the collector number of 2018–21 (*D. crassirhizoma)*,, 2018–22 (*A. brevifrons)*,, 2018–32 (*D. goeringiana)*,, and 2018–33 (*P. tripteron)*.. None of these species were endangered or protected. The ferns used in this study is not personal saved and this research was permitted by Institute of Natural Resources and Ecology.

# Chloroplast DNA extraction, sequencing

Fresh wild plant leaves of them were selected and immersed into liquid nitrogen immediately and then stored at −80°C before DNA extraction. Approximately 5g of fresh leaves was collected for chloroplast DNA isolation using an improved extraction method [48]. After the DNA extraction, the quality and quantity of samples were evaluated by NanoDrop® 2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA) and Qubit® 3.0 fluorometer (Invitrogen, Carlsbad, CA, USA), respectively. Samples with total amount > 1 μg and $OD_{260/280}$ = 1.8 ~ 2.0 were used for library preparation.

Total of 1 μg chloroplast DNA was used for the library construction according to the protocol of Illumina TruSeq™ Nano DNA Sample Prep Kit (Illumina Inc, USA). Then, libraries were put on an Illumina HiSeq 4000 platform (Biozeron Co., Ltd., China) for sequencing [49].

# Genome assembly and annotation

Prior to assembly, reads with low quality were removed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). After that, the chloroplast genomes were assembled in three steps[50]. First, the clean reads were assembled into contigs using SOAPdenovo 2.04 [51]. Second, clean reads were mapped to the contigs for assembly and optimization using SOAPGapCloser 1.12 [52]. Third, the redundant sequence is removed to get the final assembly result.

Genewise (https://www.ebi.ac.uk/Tools/psa/genewise/), AUGUSTUS (http://bioinf.uni-greifswald.de/augustus/) and EVidenceModeler 1.1.1 were used for the gene comparison, prediction and

combination, respectively. Protein-coding genes, tRNA genes, and rRNA genes were predicted using DOGMA tool [53]. Chloroplast genomes were then blasted against a series of databases, including Clusters of Orthologous Groups [54], Swiss-Prot [55], Gene Ontology [56], and Kyoto Encyclopedia of Genes and Genomes [57, 58]. Finally, the circular chloroplast genome maps of *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P. tripteron* were drawn using OrganellarGenomeDRAW 1.2 [59].

# Simple Sequence Repeat (SSR) Analysis

MIcroSAtellite identification tool (MISA, http://pgrc.ipk-gatersleben.de/misa/) was applied to detect the SSR sequence of *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P. tripteron.* The minimum number of repeats were set to eight, five, four, three, three and three for mononucleotide (mono-), dinucleotides (di-), trinucleotides (tri-), tetranucleotides (tetr-), pentanucleotide (penta-) and hexanucleotides, respectively. The distance between two SSRs is no longer than 100 bp. Finally, the primer sequences of SSRs were designed using Primer3 (http://www.simgene.com/Primer3).

# Comparison analysis

To further compare the genomes of *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P. tripteron,* more fern species (*L. clathratus, D. decipiens, C. devexiscapulae, and P. glycyrrhiza)* were used for the genome comparison, where mVISTA (http://genome.lbl.gov/vista/mvista/about.shtml) was used in Shuffle-LAGAN mode. IR border comparation was conducted in these eight fern species except P. glycyrrhiza.

We measured parsimony-informative characters per site (Pi values) by MAFFT 7.123b (http://mafft.cbrc.jp/alignment/software/) and Variscan 2.0 (http://www.ub.es/softevol/variscan) to detect the most variable chloroplast genes. The Pi value was calculated with the step size of 200 bp and window 300 bp. Genes with *Pi* value >0.20 were marked.

# Phylogenetic analysis

A total of 43 fern species were used for the phylogenetic analysis. ClustalW2 was used to align the chloroplast DNA sequences under default parameters [60], and the alignment was checked manually. The Maximum-likelihood (ML) methods were performed for the genome-wide phylogenetic analyses using PhyML 3.0 [61], respectively. Nucleotide substitution model selection was estimated by jModelTest 2.1.10 [62] and smart model selection by PhyML 3.0. ML analysis with 1,000 bootstrap replicates was performed using GTR + G model to calculate the bootstrap values of the topology. Finally, the results were processed through iTOL 3.4.3 [63].

## Declarations

## Acknowledgements

None.

## Funding

This research was funded by Science and technology research projects of Education Department of Heilongjiang Province (12541747).

## Availability of data and materials

The whole chloroplast genomes of *D. crassirhizoma, A. brevifrons, D. goeringiana* and *P. tripteron* were deposited toGenBank (https://www.ncbi.nlm.nih.gov/genbank/) and could be obtained with the accession number of MN712463, MN712464, MN712465 and MN712466, respectively.

## Author information

### Affiliations

[1] School of Pharmacy, Heilongjiang University of Chinese Medicine, Harbin 150040, Heilongjiang, China

[2] Experimental Teaching & Practical Training Center, Heilongjiang University of Chinese Medicine, Harbin 150040, Heilongjiang, China

[3] Department of ecology, Institute of Natural Resources and Ecology, Heilongjiang Academy of Science, Harbin 150040, Heilongjiang, China

Ruifeng Fan[1], Wei Ma[1], Shilei Liu[2], and Qingyang Huang[3]

### Contributions

Conceptualization, R.F. and Q.H.; methodology, W.M.; software, S.L.; resources, Q.H.; writing-original draft preparation, R.F.; writing-review and editing, Q.H., S.L., and W.M.; visualization, W.M.;

### Corresponding author

Correspondence to Qingyang Huang

## Ethics declarations

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare no conflict of interest.

## Abbreviations

| Simple name | Full name |
|---|---|
| *D. goeringiana* | *Dryopteris goeringiana* (Kunze) Koidz |
| *A. brevifrons* | *Arthyrium brevifrons* Nakai ex Kitagawa |
| *D. crassirhizoma* | *Dryopteris crassirhizoma* Nakai |
| *P. tripteron* | *Polystichum tripteron* (Kunze) Presl |
| *C. devexiscapulae* | *Cyrtomium devexiscapulae* |
| *D. decipiens* | *Dryopteris decipiens* |
| *L. clathratus* | *Lepisorus clathratus* |
| *P. glycyrrhiza* | *Polypodium glycyrrhiza* |
| *P. amoena* | *Polypodiodes amoena* |
| *D. fragrans* | *Dropteris fragrans* |
| *A. gigantea* | *Alsophila gigantea* |
| SSR | Simple Sequence Repeat |
| NGS | next generation sequencing |
| SSC | short single-copy |
| LSC | large single-copy |
| IR | inverted repeats |
| tRNA | transfer RNA |
| rRNA | ribosomal RNA |
| MISA | MIcroSAtellite |
| mono- | mononucleotide |
| di- | dinucleotides |
| tri- | trinucleotides |
| tetr- | tetranucleotides |
| penta- | pentanucleotide |
| ML | Maximum-likelihood |

## References

1.Ruiz-Ruano FJ, Navarro-Domínguez B, Camacho JPM, Garrido-Ramos MA: *Full plastome sequence of the fern Vandenboschia speciosa (Hymenophyllales): structural singularities and evolutionary insights. Journal of Plant Research* 2019, 132(1):3–17.

2.Bausher MG, Singh ND, Lee SB, Jansen RK, Daniell H: *The complete chloroplast genome sequence of Citrus sinensis(L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. BMC Plant Biology* 2006, *6(*1):21–20.

3.Jarvis P, Soll J: *Toc, Tic, and chloroplast protein import. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 2001, *1541(*1–2):64–79.

4.Leister D: *Chloroplast research in the genomic age [Review]. Trends in Genetics* 2003, *19(*1):47–56.

5.Ruhlman TA, Jansen RK: *The plastid genomes of flowering plants. In: Maliga P(ed) Chloroplast biotechnology: methods and protocols. Methods in molecular biology* 2014, *vol 1132. Springer Science+Business Media(*New York):pp 3–38.

6.Xu JH, Liu Q, Hu W, Wang T, Xue Q, Messing J: *Dynamics of Chloroplast Genomes in Green Plants. Genomics* 2015, *106(*4):221–231.

7.Green BR: *Chloroplast genomes of photosynthetic eukaryotes. Plant Journal* 2011, *66(*1):34–44.

8.Helena K: *The evolutionary processes of mitochondrial and chloroplast genomes differ from those of nuclear genomes. Science of Nature* 2004, *91(*11):505–518.

9.Wolfe KH, Li WH, Sharp PM: *Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A* 1987, *84(*24):9054–9058.

10.Nock CJ, Baten A, King GJ: *Complete chloroplast genome of Macadamia integrifolia confirms the position of the Gondwanan early-diverging eudicot family Proteaceae. Bmc Genomics* 2014, *15(*S9):S13.

11.Park M, H P, H L, BH L, J. L: *The Complete Plastome Sequence of an Antarctic Bryophyte Sanionia uncinata (Hedw.) Loeske. International journal of molecular sciences* 2018, *19(*3):709.

12.Wolf P, Karol K: *Plastomes of bryophytes, lycophytes and ferns. In: Bock R, Knoop V (eds) Advances in photosynthesis and respiration. Genomics of Chloroplasts and Mitochondria* 2012, *vol 35(*Springer, Dordrecht):pp 89–102.

13.Guo Z, Zhang H, Shrestha N, Zhang X: *Complete chloroplast genome of a valuable medicinal plant,Huperzia serrata(Lycopodiaceae), and comparison with its congener1. Applications in Plant Sciences* 2016, *4(*11):1600071.

14.Tsuji S, Ueda K, Nishiyama T, Hasebe M, Yoshikawa S, Konagaya A, Nishiuchi T, Yamaguchi K: *The chloroplast genome from a lycophyte (microphyllophyte), Selaginella uncinata, has a unique inversion, transpositions and many gene losses. Journal of Plant Research* 2007, *120(*2):281–290.

15.Logacheva MD, Krinitsina AA, Belenikin MS, Khafizov K, Konorov EA, Kuptsov SV, Speranskaya AS: *Comparative analysis of inverted repeats of polypod fern (Polypodiales) plastomes reveals two hypervariable regions. BMC Plant Biology* 2017, *17(*2):255.

16.Lu Je, Ning Z, Du Xu, Wen J, Li. Dh: *Chloroplast phylogenomics resolves key relationships in ferns. Journal of Systematics Evolution* 2015, *53(*5):448–457.

17.Wolf PG, Der JP, Duffy AM, Davidson JB, Grusz AL, Pryer KM: *The evolution of chloroplast genes and genomes in ferns. Plant Molecular Biology* 2011, *76(*3–5):251–261.

18.Wei R, Yan YH, Harris AJ, Kang JS, Shen H, Xiang QP, Zhang XC: *Plastid Phylogenomics Resolve Deep Relationships among Eupolypod II Ferns with Rapid Radiation and Rate Heterogeneity. Genome Biology Evolution* 2017, *9(*6):1646.

19.Sun Y, Moore MJ, Zhang S, Soltis PS, Soltis DE, Zhao T, Meng A, Li X, Li J, Wang H: *Phylogenomic and structural analyses of 18 complete plastomes across nearly all families of early-diverging eudicots, including an angiosperm-wide analysis of IR gene content evolution ☐. Molecular Phylogenetics Evolution* 2016, *96(*1):93–101.

20.Z. W, PH. R, D. H: *Introduction. In Flora of China, Flora of China editorial board. Beijing: Science Press, St. Louis: Missouri Botanical Garden Press;.* 2013:p. 21.

21.Chen Y-H, Chang F-R, Lin Y-J, Wang L, Chen J-F, Wu Y-C, Wu M-J: *Identification of phenolic antioxidants from Sword Brake fern (Pteris ensiformis Burm.). Food chemistry* 2007, *105(*1):48–56.

22.Sessa EB, Zimmer EA, Givnish TJ: *Phylogeny, divergence times, and historical biogeography of New World Dryopteris (Dryopteridaceae). American Journal of Botany* 2012, *99(*4):730–750.

23.Zhang L: *Taxonomic and nomenclatural notes on the fern genus Polystichum (Dryopteridaceae) in China. Phytotaxa* 2012, *60(*1):57–60.

24.Zhang L, Barrington D: *Polystichum Roth. In: Wu, Z. -Y., Raven, P. H. & Hong, D. -Y. (eds.). Flora of China. Vol. 2–3. Science Press: Beijing & Missouri Botanical Garden Press, St Louis.*

25.Wei R, Zhang X-C: *Athyrium sessilipinnum: A new lady fern (Athyriaceae) from southern China. Brittonia* 2016, *68(*4):440–447.

26.Gao R, Wang W, Huang Q, Fan R, Wang X, Feng P, Zhao G, Bian S, Ren H, Chang Y: *Complete chloroplast genome sequence of Dryopteris fragrans (L.) Schott and the repeat structures against the thermal environment. Scientific Reports* 2018, *8(*1):16635.

27.Liu F, Pang S: *Chloroplast genome of Sargassum horneri (Sargassaceae, Phaeophyceae): comparative chloroplast genomics of brown algae. Journal of applied phycology* 2016, *28(*2):1419–1426.

28.Park M, Park H, Lee H, Lee B-h, Lee J: *The complete plastome sequence of an Antarctic bryophyte Sanionia uncinata (Hedw.) Loeske. International journal of molecular sciences* 2018, *19(*3):709.

29.Asaf S, Khan AL, Khan MA, Waqas M, Kang SM, Yun BW, Lee IJJSR: *Chloroplast genomes of Arabidopsis halleri ssp. gemmifera and Arabidopsis lyrata ssp. petraea: Structures and comparative analysis. Scientific Reports* 2017, *7(*1):7556.

30.Liu S, Ping J, Zhen W, Wang T, Su Y: *Complete chloroplast genome of the tree fern Alsophila podophylla (Cyatheaceae). Mitochondrial DNA Part B* 2017, *3(*1):48–49.

31.Ting Wang, Yongfeng Hong, Zhen Wang, Su. Y: *Characterization of the complete chloroplast genome of Alsophila gigantea (Cyatheaceae), an ornamental and CITES giant tree fern. mitochondrial DNA Part B* 2018, *4(*1):967–968.

32.Xu L, Xing Y, Wang B, Liu C, Wang W, Kang T: *Plastid genome and composition analysis of two medical ferns: Dryopteris crassirhizoma Nakai and Osmunda japonica Thunb. Chinese Medicine* 2019, *14(*1):9.

33.Abdul Latif Khan, Sajjad Asaf, In-Jung Lee, Ahmed Al-Harrasi, Al-Rawahi. A: *First chloroplast genomics study of Phoenix dactylifera (var. Naghal and Khanezi): A comparative analysis. Plos One* 2018, *13(7*):e0200104.

34.Doorduin L, Gravendeel B, Lammers Y, Ariyurek Y, Chinawoeng T, Vrieling K: *The Complete Chloroplast Genome of 17 Individuals of Pest Species Jacobaea vulgaris: SNPs, Microsatellites and Barcoding Markers for Population and Phylogenetic Studies. DNA Research* 2011, *18(*2):93–105.

35.He S, Wang Y, Volis S, Li D, Yi T: *Genetic Diversity and Population Structure: Implications for Conservation of Wild Soybean (Glycine soja Sieb. et Zucc) Based on Nuclear and Chloroplast Microsatellite Variation. International journal of molecular sciences* 2012, *13(*10):12608–12628.

36.Jianhua X, Shuo W, Shi-Liang Z: *Polymorphic chloroplast microsatellite loci in Nelumbo (Nelumbonaceae). American Journal of Botany* 2012, *99(*6):240–244.

37.Ai-Hong Y, Jin-Ju Z, Xiao-Hong Y, Hong-Wen H: *Chloroplast microsatellite markers in Liriodendron tulipifera (Magnoliaceae) and cross-species amplification in L. chinense. American Journal of Botany* 2011, *98(*5):123–126.

38.Dai-Yong Kuang, Hong Wu, Ya-Ling Wang, Lian-Ming Gao, Shou-zhou Zhang, Lu. L: *Complete chloroplast genome sequence of Magnolia kwangsiensis (Magnoliaceae): implication for DNA barcoding and population genetics. Genome* 2011, *54(*8):663–673.

39.Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK: *Comparative chloroplast genomics: analyses including new sequences from the angiosperms Nuphar advena and Ranunculus macranthus. BMC Genomics* 2007, *8(*1):174.

40.Asaf S, Khan AL, Khan AR, Waqas M, Kang S-M, Khan MA, Lee S-M, Lee I-J: *Complete chloroplast genome of Nicotiana otophora and its comparison with related species. Frontiers in Plant Science* 2016, *7:*843.

41.Khakhlova O, Bock R: *Elimination of deleterious mutations in plastid genomes by gene conversion. The plant journal* 2006, *46(*1):85–94.

42.Yang M, Zhang X, Liu G, Yin Y, Chen K, Yun Q, Zhao D, Al-Mssallem IS, Yu J: *The complete chloroplast genome sequence of date palm (Phoenix dactylifera L.). Plos One* 2010, *5(*9):e12762.

43.Daniell H, Lin CS, Ming Y, Chang WJ: *Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biology* 2016, *17(*1):134.

44.Masanori K, Yuhei Y, Takeshi F, Tohoru M, Koichi Y: *RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucleic Acids Research* 2003, *31(*9):2417–2423.

45.Wolf PG, Rowe CA, Hasebe M: *High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern Adiantum capillus-veneris. Gene* 2004, *339(*none):0–97.

46.Koichiro T, Yoshihiro M, Yukiko Y, Yasunari O: *Evolutionary dynamics of wheat mitochondrial gene structure with special remarks on the origin and effects of RNA editing in cereals. Genes Genetic Systems* 2008, *83(*4):301–320.

47.Jun-Bo Yang, Shi-Xiong Yang, Hong-Tao Li, Jing Yang, Li. D-Z: *Comparative Chloroplast Genomes of Camellia Species. Plos One* 2013, *8(*8):e73053.

48.McPherson H, Van der Merwe M, Delaney SK, Edwards MA, Henry RJ, McIntosh E, Rymer PD, Milner ML, Siow J, Rossetto M: *Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. BMC ecology* 2013, *13(*1):8.

49.Borgström E, Lundin S, Lundeberg J: *Large scale library generation for high throughput sequencing. Plos One* 2011, *6(*4):e19119.

50.Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T: *Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Research* 2008, *36(*19):e122-e122.

51.Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y: *SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience* 2012, *1(*1):18.

52.Zhao Q-Y, Wang Y, Kong Y-M, Luo D, Li X, Hao P: *Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study.* In: *BMC bioinformatics: 2011.* BioMed Central: S2.

53.Wyman SK, Jansen RK, Boore JL: *Automatic annotation of organellar genomes with DOGMA. Bioinformatics* 2004, *20(*17):3252–3255.

54.Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN: *The COG database: an updated version includes eukaryotes. BMC bioinformatics* 2003, *4(*1):41.

55.Magrane M: *UniProt Knowledgebase: a hub of integrated protein data. Database* 2011, *2011.*

56.Ashburner M,., Ball CA, Blake JA, Botstein D,., Butler H,., Cherry JM, Davis AP, Dolinski K,., Dwight SS, Eppig JT: *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics* 2000, *25(*1):25–29.

57.Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: *The KEGG resource for deciphering the genome. Nucleic Acids Research* 2004, *32(*suppl_1):D277-D280.

58.Minoru K, Susumu G, Masahiro H, Aoki-Kinoshita KF, Masumi I, Shuichi K, Toshiaki K, Michihiro A, Mika H: *From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Research* 2006, *34(*supp_1):354–357.

59.Lohse M, Drechsel O, Bock R: *OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Current Genetics* 2007, *52(*5–6):267–274.

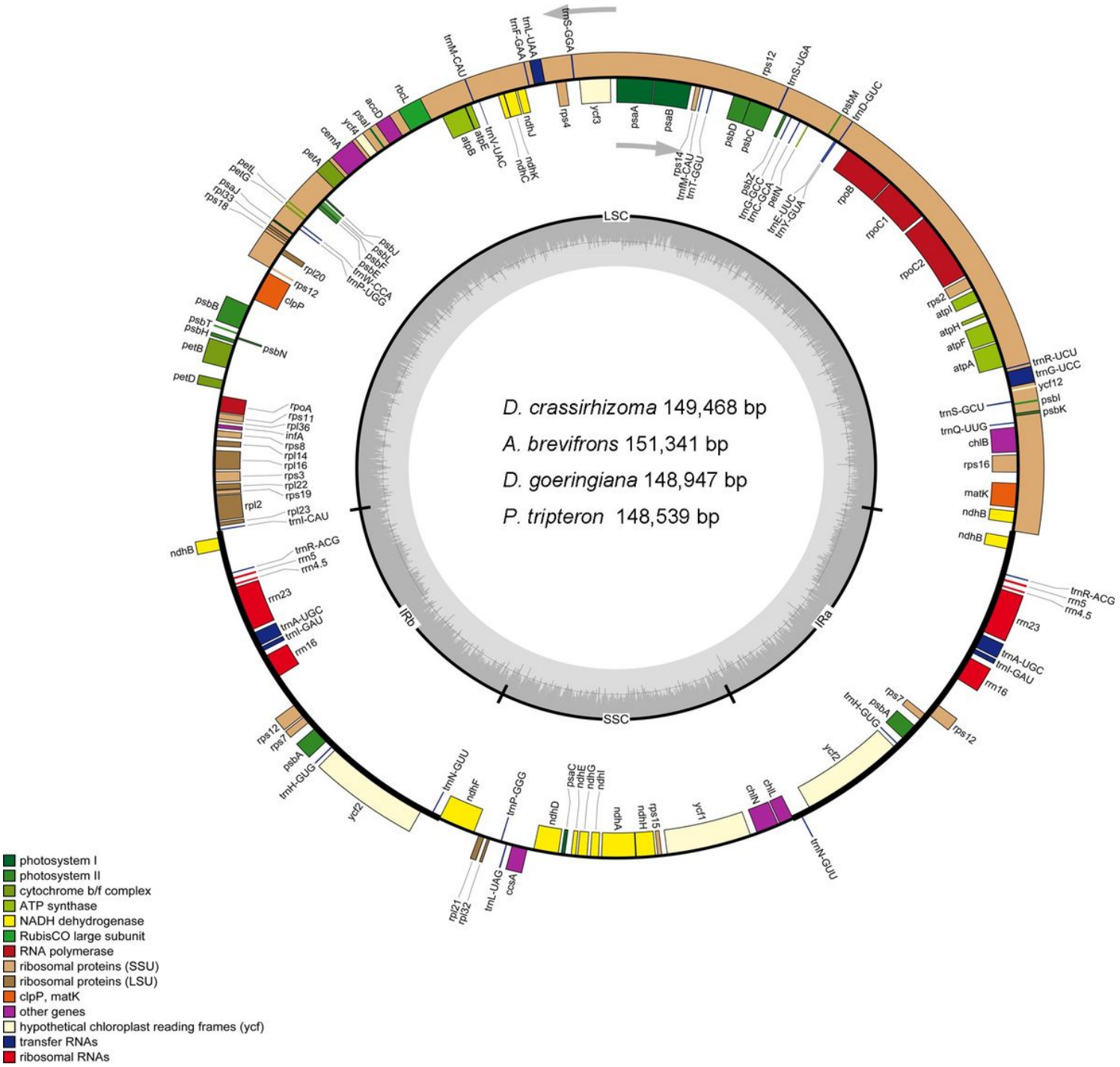
60.Larkin MA, Blackshields G, Brown N, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R: *Clustal W and Clustal X version 2.0. Bioinformatics* 2007, *23(*21):2947–2948.

61.Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O: *New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology* 2010, *59(*3):307–321.

62.Darriba D, Taboada GL, Doallo R, Posada D: *jModelTest 2: more models, new heuristics and parallel computing. Nature Methods* 2012, *9(*8):772.

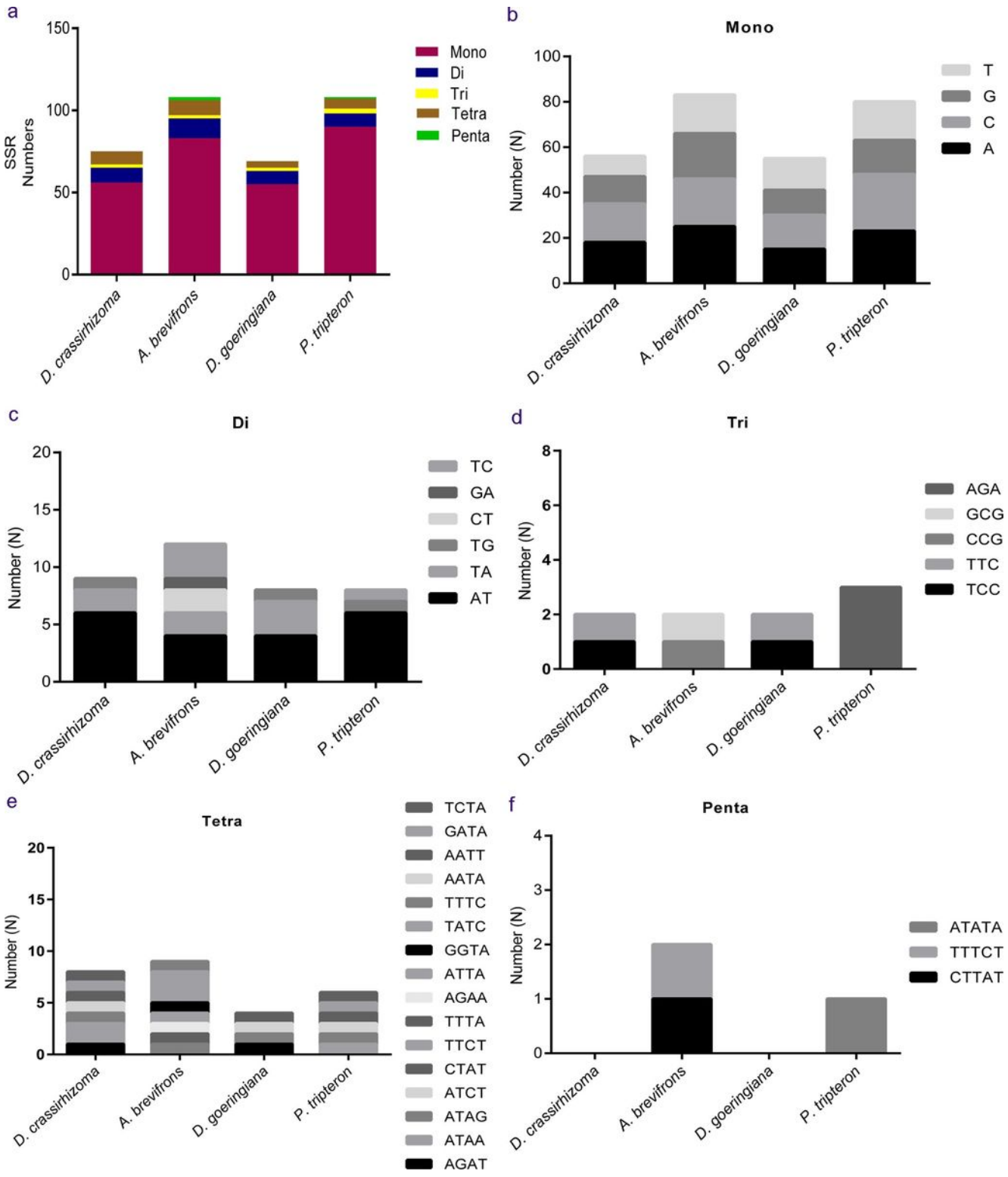
63.Letunic I, Bork P: *Interactive Tree of Life (ITOL) v3: An online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Research* 2016, *44 (web server issue):*gkw290.

# Figures



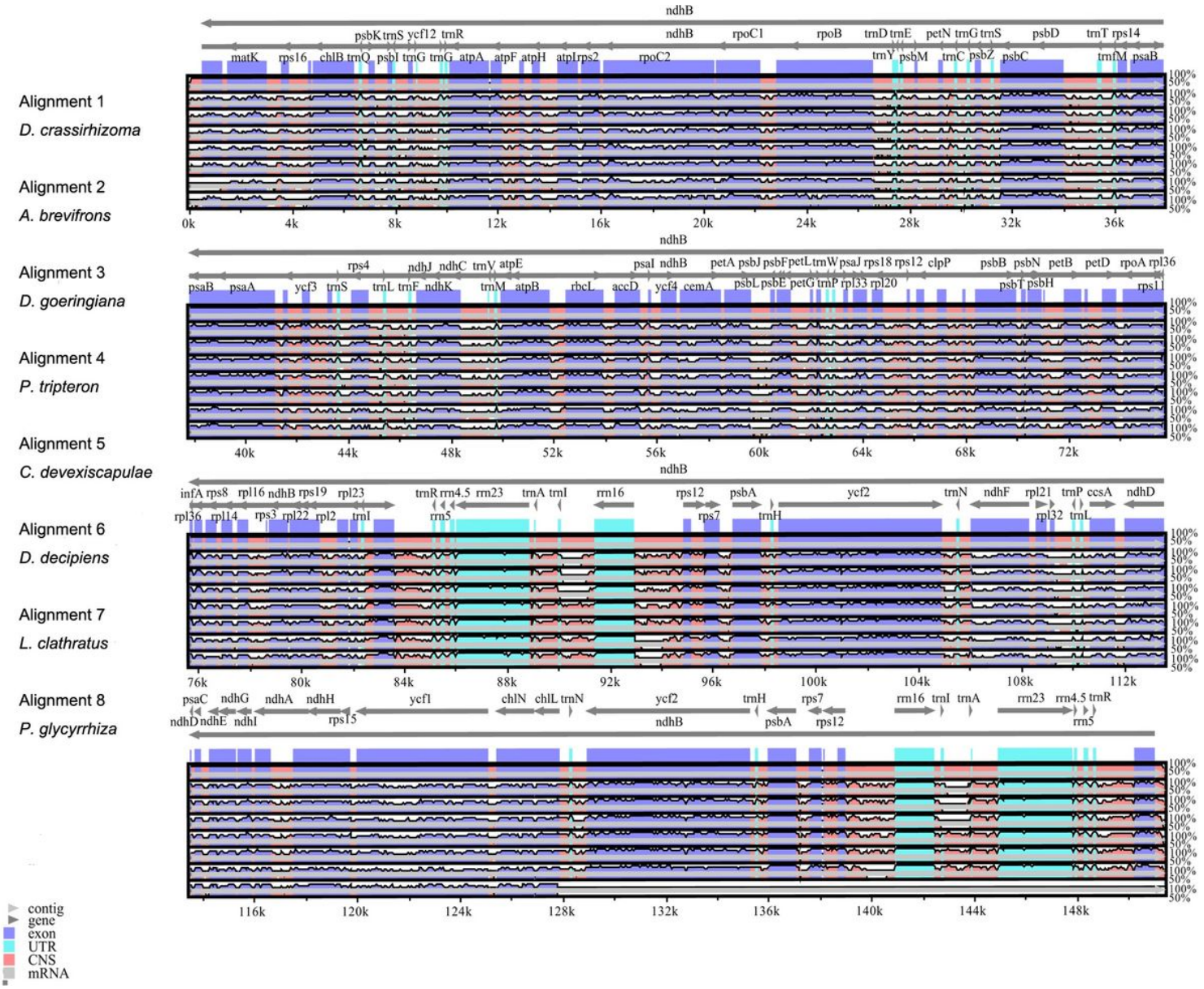

## Figure 1

Chloroplast genome map of D. goeringiana, A. brevifrons, D. crassirhizoma and P. tripteron genomes. Genes drawn inside the circle are transcribed clockwise, and those outside the circle are transcribed counter-clockwise. The light gray inner circle corresponds to the AT content, the dark gray to the GC content. Genes belonging to different functional groups are shown in different colors.



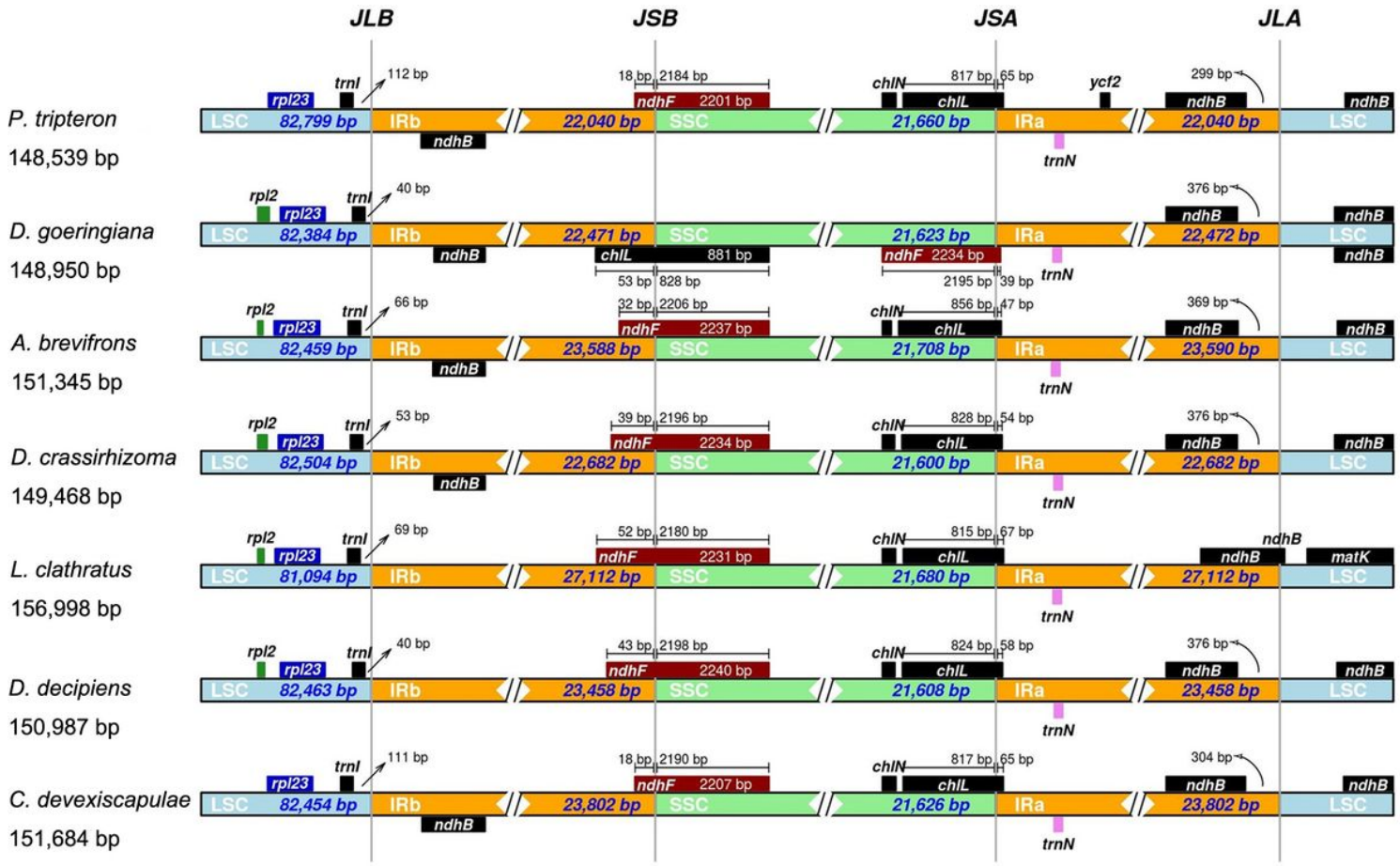

## Figure 2

Analysis of simple sequence repeats in the four fern chloroplast genomes. (a) Number of different SSR types detected in the four genomes; (b) Number and types of identified mononucleotide (Mono) SSRs in the four genomes; (c) Number and types of dinucleotides (Di) SSRs in the four genomes; (d) Number and types of trinucleotides (Tri) SSRs in the four genomes; (e) Number and types of tetranucleotides (Tetr) SSRs in the four genomes; (f) Number and types of pentanucleotide (Penta) SSRs in the four genomes.



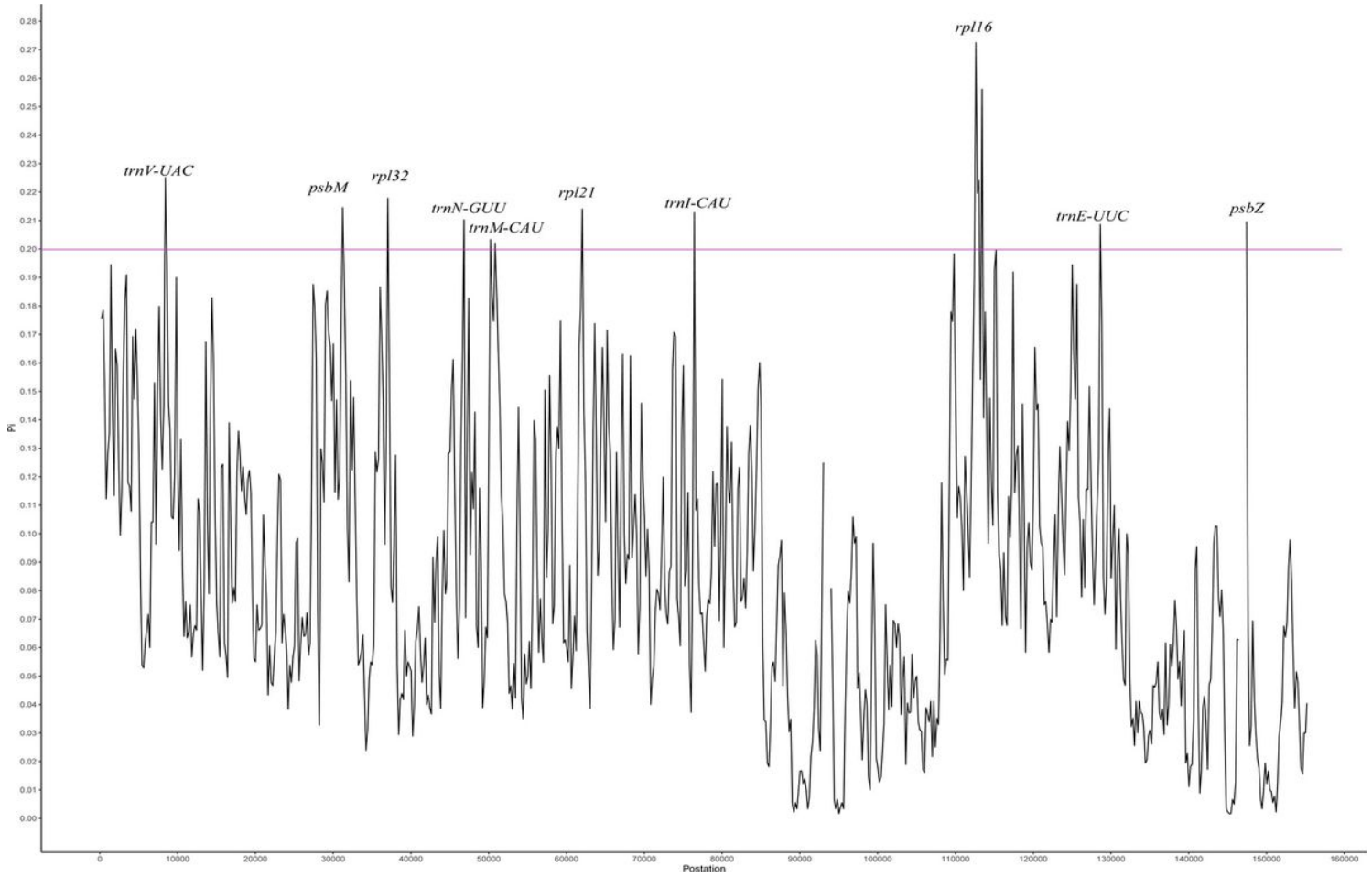

## Figure 3

Sequence alignment of fern species. Grey arrows above the alignment indicate the orientation of genes. Purples bars represent exons, blue ones represent introns and ncRNAs, pink ones represent non-coding sequences. X-axis represents the genome coordinate positions; Y-axis represents the percent identify within 50-100%. Dashed rectangles indicate highly divergent regions. Use D. goeringiana as the reference.




## Figure 4

Comparison of the borders of LSC, SSC and IR regions among the seven chloroplast genomes. The rectangular strips of each row represent a genome. The different colors represent different partitions. The black vertical line represents the boundary; the genes on both sides are indicated by small squares of different colors, the gene name is indicated on it, the gene in the forward chain is above; and the number represent the distant from gene and the boundary. Use D. goeringiana as the reference. LSC, large-single copy. SSC, small-single copy. IR, inverted repeat regions.




## Figure 5

Comparative analysis of the nucleotide variability by Pi values within four fern species. Pi value > 0.20 was wrote down. X-axis: position of the genome; Y-axis: nucleotide diversity (Pi) value.

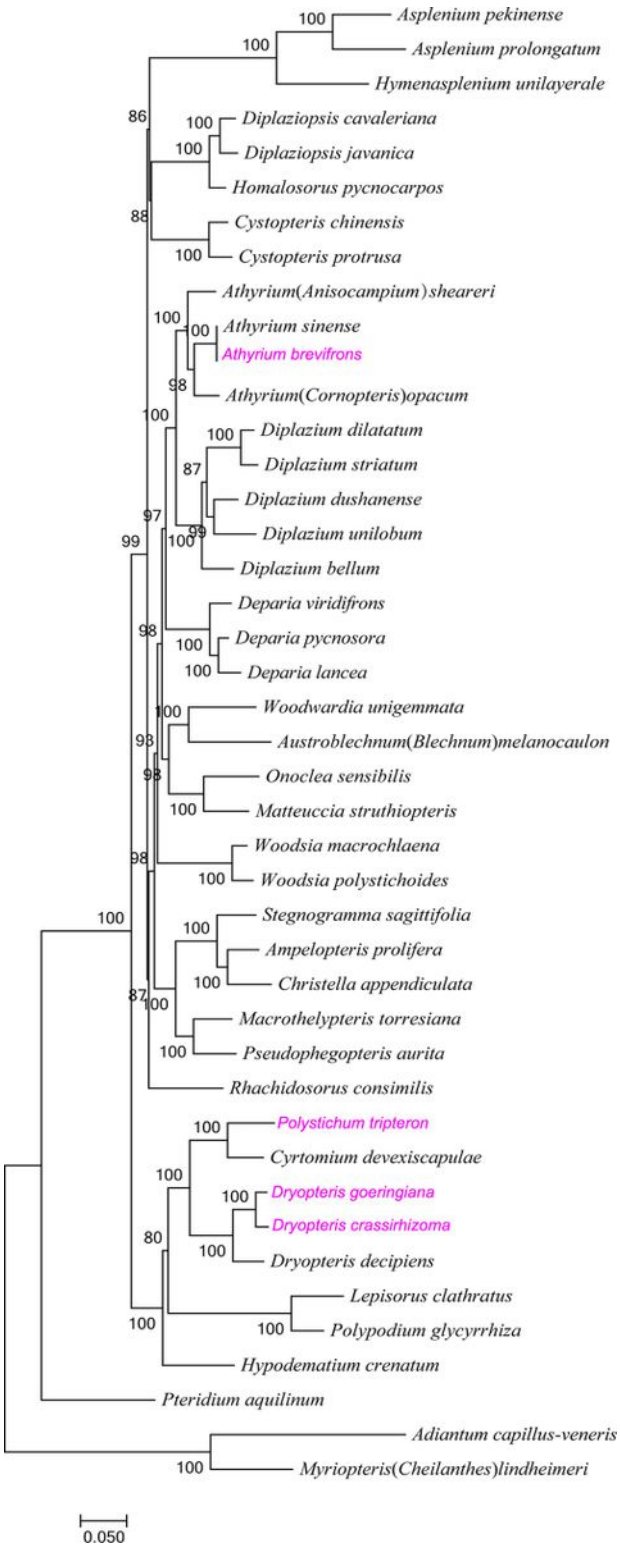



**Figure 6**

Molecular phylogenetic tree on 43 fern species. The tree was constructed using Maximum-Likelihood (ML) algorithm and the general time reversible (GTR) + G model. The species studied in this study was colored with rose red.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.xlsx
- Additionalfile2.xlsx