

SCIENTIFIC REPORTS



OPEN

Complete Chloroplast Genome Sequence of *Decaisnea insignis*: Genome Organization, Genomic Resources and Comparative Analysis

Bin Li^{1,2,3}, Furong Lin^{1,2,3}, Ping Huang^{1,2,3}, Wenying Guo^{1,2,3} & Yongqi Zheng^{1,2,3}

Decaisnea insignis is a wild resource plant and is used as an ornamental, medicinal, and fruit plant. High-throughput sequencing of chloroplast genomes has provided insight into the overall evolutionary dynamics of chloroplast genomes and has enhanced our understanding of the evolutionary relationships within plant families. In the present study, we sequenced the complete chloroplast genome of *D. insignis* and used the data to assess its genomic resources. The *D. insignis* chloroplast genome is 158,683 bp in length and includes a pair of inverted repeats of 26,167 bp that are separated by small and large single copy regions of 19,162 bp and 87,187 bp, respectively. We identified 83 simple sequence repeats and 18 pairs of large repeats. Most simple-sequence repeats were located in the noncoding sections of the large single-copy/small single-copy region and exhibited a high A/T content. The *D. insignis* chloroplast genome bias was skewed towards A/T on the basis of codon usage. A phylogenetic tree based on 82 protein-coding genes of 33 angiosperms showed that *D. insignis* was clustered with *Akebia* in Lardizabalaceae. Overall, the results of this study will contribute to better understanding the evolution, molecular biology and genetic improvement of *D. insignis*.

Lardizabalaceae, a small family with approximately 50 species in 9 genera, is a core component of Ranunculales and belongs to the basal eudicots^{1,2}. *Decaisnea insignis* (Griffith) Hook. f. & Thomson, which is widely distributed from central to south-western China and the Himalayan foothills, is the only species in the genus *Decaisnea*; it is nicknamed “dead man’s fingers” as it possesses racemes of strikingly deep purplish-blue elongated fruits³. This plant is economically important, as it is readily cultivated as an ornamental plant and its fruits are deemed a delicacy. It has also been used in traditional Chinese medicine as an antirheumatic and antitussive drug for a long time⁴. *D. insignis* is a type of wild resource plant and has a wide range of uses; thus, *D. insignis* is worthy of development and utilization. To support the development and utilization of this species, markers that are variable at the population level need to be developed from genomic resources. However, despite its importance, few studies have described the DNA sequences of *D. insignis*. Therefore, the molecular techniques is required to analyse the genetic diversity and phylogenetic relationship of this plant.

Chloroplast genomes have a typical quadripartite structure consisting of a large single copy region (LSC), a small single copy region (SSC) and a pair of inverted repeats (IRs) in most plants. The chloroplast genome is a highly conserved circular DNA ranging from 115 to 165 kb with a stable genome, gene content and gene order^{5–7}. The substitution rates in plant chloroplast genomes are much lower than those in nuclear genomes^{8,9}. The angiosperms’ chloroplast genome has a uniparental inheritance and stable structure, providing sufficient genetic markers for genome-wide evolutionary studies at different taxonomic levels^{10–12}. Although the chloroplast genome shows evolutionary conservation in plants, an accelerated rate of evolution has been widely observed in particular genes or some lineages^{5,13}. For example, *rbcl*, *matK*, and *ycf1* have been used as DNA barcodes for barcoding

¹State Key Laboratory of Tree Genetics and Breeding, Chinese Academy of Forestry, Beijing, China. ²Research Institute of Forestry, Chinese Academy of Forestry, Beijing, China. ³Key Laboratory of Tree Breeding and Cultivation of State Forestry Administration, Chinese Academy of Forestry, Beijing, China. Correspondence and requests for materials should be addressed to Y.Z. (email: zyq8565@126.com)

plants^{14, 15}. As a result of these characteristics, chloroplast genomes are considered to be good models for testing lineage-specific molecular evolution. With the development of high-throughput sequencing technologies, new approaches for chloroplast genome sequencing have been gradually proposed due to their high-throughput, time-saving and low-cost¹⁶.

In the present study, we reconstructed the whole chloroplast genome of *D. insignis* by using next-generation sequencing and applying a combination of *de novo* and reference-guided assembly. The objectives of this study were to establish and characterize the organization of the complete chloroplast genome of *D. insignis* and conduct comparative genomic studies to gain in-depth insights into the overall evolutionary dynamics of chloroplast genomes. Our data will also provide genomic resources for this species to determine its phylogenetic relationship with related species as well as a genetic diversity evaluation and plant molecular identification.

Results and Discussion

Chloroplast genome assembly. A total of 2.48×10^7 reads with an average read length of 150 bp were obtained. The *de novo* assembled contigs were analysed locally by BIANSTN using the *Akebia quinata* genome as a reference; seven contigs were retained. The gaps between the *de novo* contigs were checked by amplification. The total reads were re-mapped to the chloroplast genome, and correction of the sequences was confirmed. The coverage of the chloroplast genome was $2166\times$, and the sequence of the chloroplast genome was deposited in GenBank (accession number: KY200671).

Organization and gene content. The chloroplast genome of *D. insignis* was 158,683 bp in length (Fig. 1). The genome presented a typical quadripartite structure with two inverted repeats (each 26,167 bp in length) separated by one small and one large single-copy region (19,162 and 87,187 bp in length, respectively). These values were similar to those of *Akebia* (Table 1)¹⁷. The GC content of the chloroplast DNA was 38.5%. The GC content of the LSC (36.9%) and SSC regions (33.3%) was lower than that in IR regions (43.1%).

The genome consisted of 113 different coding genes, of which 79 were protein-coding genes, 30 were distinct tRNA genes, and 4 were rRNA genes (Fig. 1, Supplementary Table S1). Of these, five protein-coding, four rRNA, and seven tRNA genes were duplicated in the IR regions. The LSC region comprised 62 protein-coding and 22 tRNA genes, whereas the SSC region comprised 12 protein-coding genes and one tRNA gene. Twelve genes contained introns, *clpP* and *ycf3* comprised two introns, and the rest of the genes had one intron. In *rps12*, a trans-splicing event was observed, with the 5' end located in the LSC region and the duplicated 3' end in the IR region, as previously reported^{18, 19}.

Among the 79 protein-coding genes, 75 genes had the standard AUG as the initiator codon, but *psbC* and *rps19* used GUG, while *rpl2* and *ndhD* used ACG. RNA editing events of the AUG initiation site to GUG had been reported for *psbC*²⁰ and *rps19*²¹. Previous studies on non-canonical translational mechanisms suggested that the translational efficiency of the GUG codon was relatively higher compared with the canonical AUG as the initiation codon²². In Brassicaceae, *psbC* and *rps19* also used AUG as the initiation codon²¹. ACG and GUG were used as the start codons for *rpl2* and *rps19*, respectively, as reported in *Oryza minuta*²³.

Codon usage. We calculated the codon usage frequency and relative synonymous codon usage frequency (RSCU) in the *D. insignis* chloroplast genome. Codon usage plays an important part in shaping chloroplast genome evolution. Mutational bias has been reported to have an essential role in shaping this evolutionary phenomenon^{24, 25}. The total protein coding genes comprised 78,375 bp that encoded 26,325 codons. Of these codons, 2,697 (10.3%) encoded leucine, whereas only 311 (1.2%) encoded cysteine (Fig. 2 and Supplementary Table S2), which were the most and the least frequently used amino acids in the *D. insignis* chloroplast genome, respectively. The AT content was 53.94%, 61.25%, and 68.70% at the 1st, 2nd, and 3rd codon positions, respectively. The preference for a high AT content at the 3rd codon position was similar to the A and T concentrations reported in other plants²⁶. A general excess of A- and U-ending codons was noted. Except for TGA, CTA, and ATA, all preferred synonymous codons (RSCU > 1) ended with an A or U (Supplementary Table S2). Usage of the start codon AUG and tryptophan UGG had no bias (RSCU = 1).

Repeat sequence. SSRs have been described as a major tool that can be used to unravel genome polymorphisms across species and perform population genetics of species on the basis of repeat length polymorphisms in plant molecular studies^{27–29}. Because the chloroplast genome sequences are highly conserved, SSR primers for chloroplast genomes are transferable across species and genera. Considering the role of chloroplast genome SSRs as important phylogenetic markers, we applied a length threshold of greater than 10 repeats for mono-, 5 repeats for di-, 4 repeats for tri-, and 3 repeats for tetra-, penta-, and hexanucleotide repeat patterns. Eighty-three perfect microsatellites were analysed in *D. insignis* (Fig. 3). Mononucleotide SSRs were the richest with a proportion of 72.29%, and the mononucleotide A and T repeat units occupied the highest portion of 96.67%. Our findings agreed with the observation that chloroplast SSRs were generally composed of polyadenine (poly A) and polythymine (poly T) and rarely contained tandem guanine (G) and cytosine (C) repeats^{30, 31}. Furthermore, there were 9 di-, 7 tri-, 6 tetra-, and one hexanucleotide repeats in the *D. insignis* chloroplast genome. Most SSRs were present in the noncoding regions of this chloroplast genome, and only one coding gene, *ycf1*, contained 4 SSRs. Forty-six spacer regions and eight intron regions harboured SSRs; the *trnK-rps16* and *petA-psbJ* spacers had the highest number of indels (four), followed by the *atpF* intron and *trnT-trnL* (three). Most of these SSRs (95.18%) were present in the single copy region (Fig. 3 and Supplementary Table S3). Interestingly, the number of identified SSRs in the *D. insignis* chloroplast genome was low compared with the previously characterized SSRs. Additionally, we did not find a larger abundance of di- and tri-nucleotide repeats. Slipped strand mispairing (SSM) and intramolecular recombination had been suggested as the likely mechanism that led to most SSRs³². Based on the identified SSRs, we designed 79 primer pairs (except the four repeat SSRs in IR region), which could be used for future in-depth studies of phylogeography and the population structure pattern of this species (Supplementary Table S3).

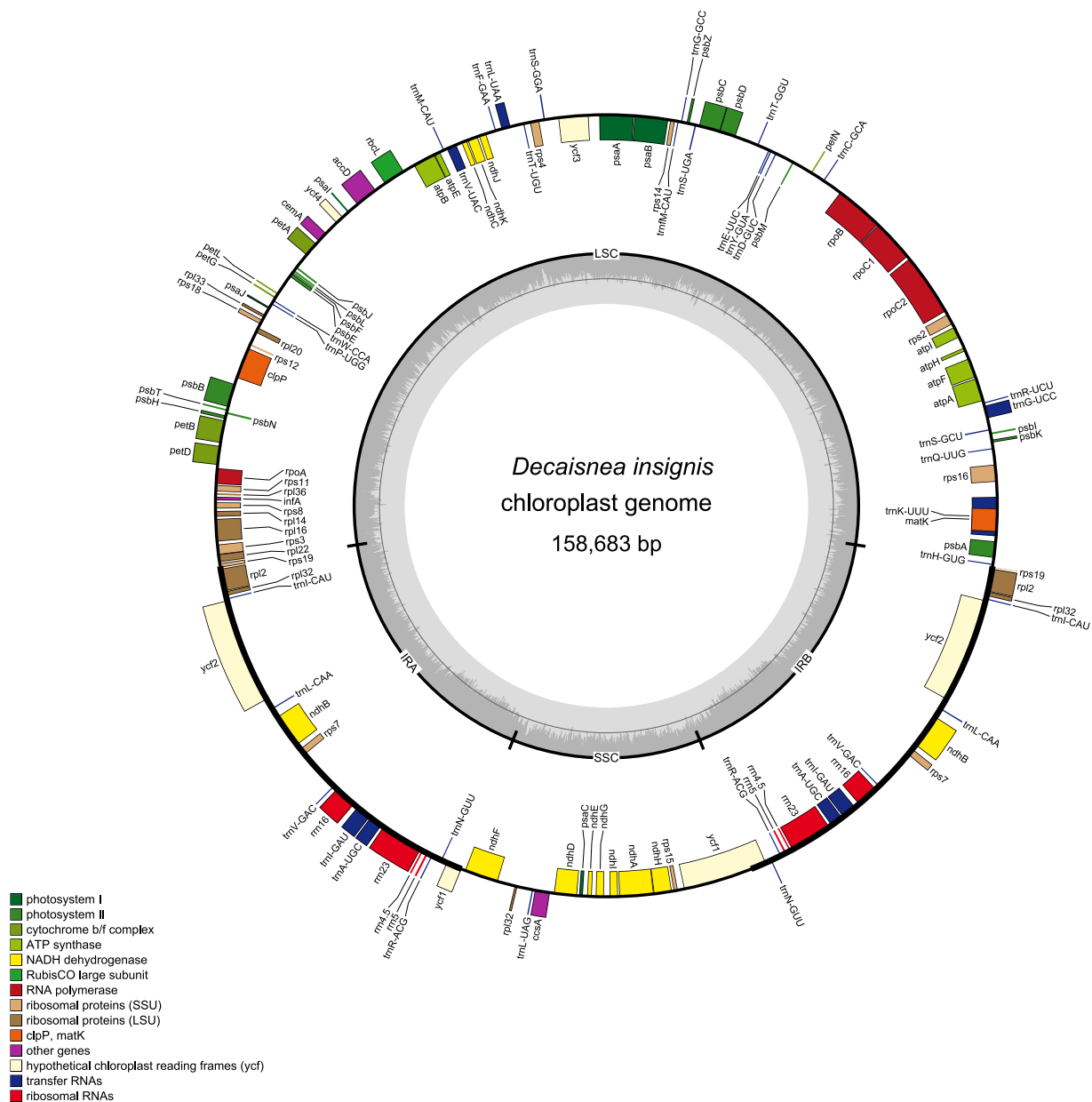


Figure 1. Chloroplast genome map of *D. insignis*. The genes drawn outside of the circle are transcribed clockwise, while those inside are counterclockwise. Small single copy (SSC), large single copy (LSC), and inverted repeats (IRA, IRB) are indicated.

Species	<i>A. trifoliata</i>	<i>A. quinata</i>	<i>D. insignis</i>
Genome size (bp)	158,339	157,817	158,683
LSC (bp)	87,057	86,543	87,187
IR (bp)	26,129	26,143	26,167
SSC (bp)	19,024	18,988	19,162
Total number of genes	113	113	113
Protein coding genes	79	79	79
rRNA	4	4	4
tRNA	30	30	30
GC%	38.7	38.7	38.5

Table 1. Characteristics of the chloroplast genomes in three species.

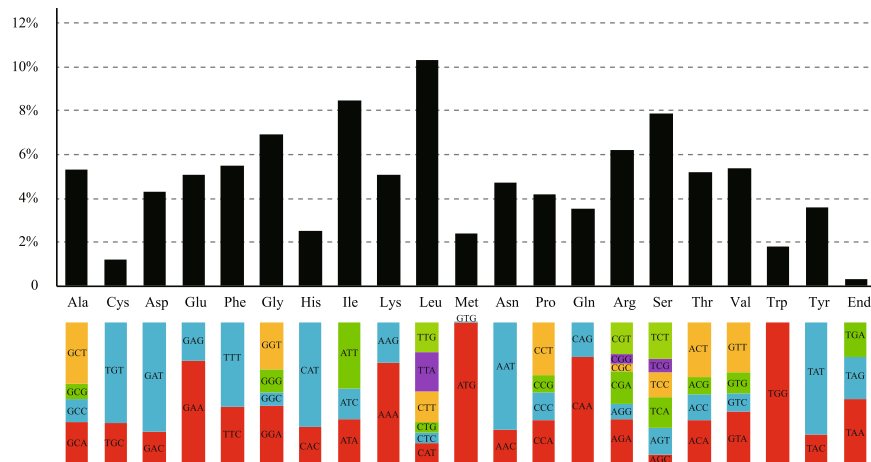


Figure 2. Codon content of 20 amino acid and stop codon of 82 coding genes of *D. insignis* chloroplast genome. Color of the histogram is the proportion of codon usage for amino acid and stop codon.

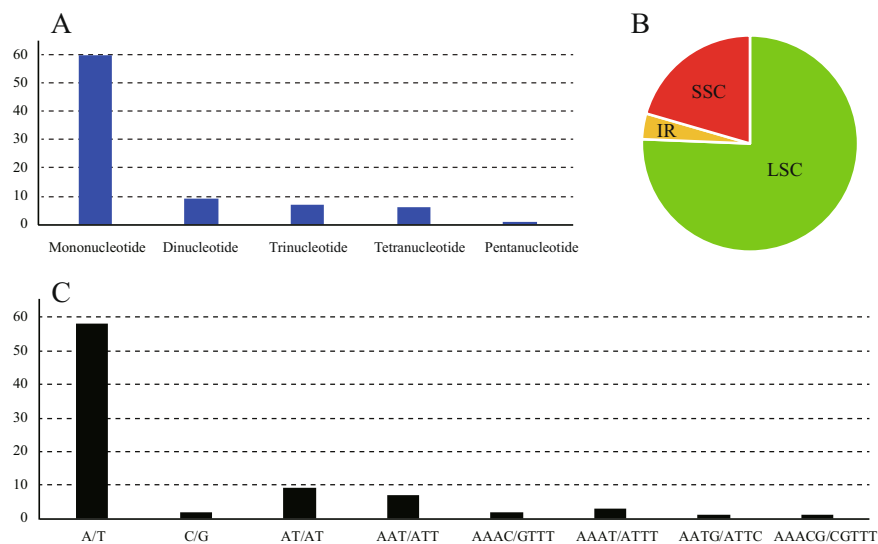


Figure 3. The distribution, type and presence of simple sequence repeats (SSRs) in the chloroplast genome of *D. insignis*. (A) Number of different SSRs types. (B) Proportion of SSRs in the LSC, SSC, and IR regions. (C) Number of identified SSR motifs in different repeat class types.

In addition to the SSRs, we explored the role of long repeats, as identified by REPuter³³, with a minimal repeat size of 30 bp and a Hamming distance of 90. We identified 18 repeats, including 7 forward and 11 palindromic repeats (Table 2). The length of the repeats ranged from 30 to 73 bp. Approximately 40% of these repeats fell exclusively within genes, mainly in *ycf2* (Table 2). Similar to the SSRs, *D. insignis* contained a lower number of repeat elements compared with other plants^{34, 35}. The presence of these repeats indicated that the locus was a crucial mutation hotspot in the genome because the repeat sequences led to sequence variations and genome rearrangements due to the slipped-strand mispairing and improper recombination³⁶. Additionally, these repeats played an important role in developing genetic markers for phylogenetic studies.

Comparative analysis of the *D. insignis* chloroplast genome and two *Akebia* species. A comparative analysis based on mVISTA was performed between the chloroplast genomes of *D. insignis* and two *Akebia* species (*A. quinata* and *A. indica*) to investigate the levels of sequence divergence (Fig. 4). The organization of the chloroplast genome between the *D. insignis* and *Akebia* genomes revealed a high degree of synteny and gene order conservation, suggesting an evolutionary conservation of these genomes at the genome-scale level. As expected, the IR region was more conserved than the LSC and SSC regions among the three genomes. Meanwhile, as seen in other flowering plants, the coding region was more highly conserved than the non-coding regions. The most dissimilar coding regions of the three chloroplast genomes were *matK*, *ndhF*, and *ycf1*, which were located in the LSC, SSC, and SSC regions, respectively. The *matK* and *ycf1* coding regions had been observed to be divergent in chloroplast genomes and could serve as markers for DNA barcoding and phylogenetic analysis^{12, 15}. The most divergent regions were localized in the intergenic spacers and introns, including the *trnH-psbA*, *rps16-trnQ*,

No.	Size (bp)	Start position1	Start position2	Type	Location Region	E-value
1	33	8813	37730	F	<i>psbI-trnS/trnS-psbZ</i>	1.41E-05
2	30	8819	47631	P	<i>psbI-trnS/trnS</i>	5.53E-07
3	30	16853	16853	P	<i>rps2-rpoC2</i>	2.40E-05
4	30	37736	47631	P	<i>trnS-psbZ/trnS</i>	6.73E-04
5	30	39077	39077	P	<i>trnM-rps14</i>	2.40E-05
6	73	40902	43126	F	<i>psaB/psaA</i>	1.33E-28
7	58	40917	43141	F	<i>psaB/psaA</i>	1.27E-21
8	39	45906	101569	F	<i>ycf3</i> sec intron/ <i>rps7-trnV</i>	2.34E-14
9	39	45906	144262	P	<i>ycf3</i> sec intron/ <i>trnV-rps7</i>	2.34E-14
10	30	66548	66595	P	<i>petA-psbI/petA-psbI</i>	2.40E-05
11	44	77100	77100	P	<i>psbN-psbH</i>	2.29E-17
12	32	85121	85121	P	<i>rpl16</i> intron	3.84E-10
13	42	92063	92084	F	<i>ycf2</i>	4.61E-14
14	42	92063	153744	P	<i>ycf2</i>	4.61E-14
15	42	92084	153765	P	<i>ycf2</i>	4.61E-14
16	40	117507	117507	P	<i>rpl32-trnL</i>	5.86E-15
17	42	153744	153765	F	<i>ycf2</i>	4.61E-14
18	31	153755	153776	F	<i>ycf2</i>	1.54E-09

Table 2. Distribution and localization of tandem repeats in *D. insignis* chloroplast genome. F, forward repeat; P, palindrome repeat.

ycf3-trnS, *petA-psbI*, *ndhF-rpl32*, *rps32-trnL* as well as *rps16* introns. *TrnH-psbA* loci were highly variable in most plant groups, and inversions or mononucleotide repeats occurred within these loci, which might result in incorrect alignments or sequencing difficulties^{37,38}. *Rps16-trnQ*, *ycf3-trnS*, *petA-psbI*, *rps16* introns, *ndhF-rpl32* and *rps32-trnL* had been used in previous phylogenetic studies^{39,40}. The most variable one of the identified loci was *ycf1* encoding a protein of approximately 1,800 amino acids¹⁵. These non-coding regions could be used to assess phylogenetic relationships within the Lardizabalaceae species.

IR expansion/contraction also represents a highly variable pattern, which can be used to study the phylogenetic classification of plants. Moreover, the IR boundary expansion/contraction is regarded to be an evolutionary event and has been shown to be the reason for size variation in chloroplast genomes. Detailed comparisons of the IR-SSC and IR-LSC boundaries among the three Lardizabalaceae chloroplast genomes were presented in Fig. 5. The LSC/IRb border was located within the coding region of *rps19* and created a pseudogene of 39 or 87 bp at the IRb/LSC border. The IRa/SSC border extended into *ycf1*, resulting in a pseudogene in the three compared chloroplast genomes. The length of the *ycf1* pseudogene was 1,039 bp in the two *Akebia* species and 1,043 bp in *D. insignis*. Furthermore, *ndhF* deviated from the IRb/SSC in *A. trifoliata*, *A. quinata* and *D. insignis* by 166, 123, and 237 bp, respectively. The *trnH-GUG* gene was located in the LSC, which ranged from 27 to 79 bp from the IRa/LSC border. Overall, the IR boundary regions of the *D. insignis* chloroplast genome were slightly different from those of the other genomes in Lardizabalaceae.

Synonymous and nonsynonymous substitution rate. The synonymous and nonsynonymous nucleotide substitution patterns are very important markers in gene evolution studies. Estimation of these mutations plays a pivotal role in understanding the dynamics of molecular evolution⁴¹. In most genes, nonsynonymous nucleotide substitutions occur less frequently than synonymous substitutions due to the action of purifying selection. Accordingly, the ratio $\omega = dN/dS$ has become a standard measure of selective pressure with $\omega = 1$, >1 , <1 signifying neutral evolution, positive selection, and negative or purifying selection, respectively⁴². Using *D. insignis* as the outgroup, the nonsynonymous substitution (dN), synonymous substitution (dS) and dN/dS of gene groups and some genes in *A. trifoliata* and *A. quinata* were computed and compared (Fig. 6). As expected, a rate of heterogeneity existed among genes and gene groups. After sorting the genes into functional categories, significant differences were revealed among the groups. Analysis of gene groups indicated that the photosynthetic apparatus genes (*psa*, *psb*, *pet*) and *atp* had the lowest dN values relative to the other gene groups. Moreover, the photosynthetic apparatus genes and *atp* had the lowest dN/dS. The *ycf1* and *matK* genes (dN/dS > 0.5) had the highest ratios of dN and dS, indicating that these genes were selected for sequence diversity. For dS values, there were no notable differences among the genes, except in the *matK* gene. It was noteworthy that the *matK*, *ycf1*, *ccsA* and ribosomal protein genes (*rpl*, *rpo*, *rps*) evolved faster than other genes. The gene function and locus-specific variation, selection pressure and gene expression level had been shown to influence the rates of sequence evolution in chloroplast genomes^{43,44}.

Phylogenetic inference. Chloroplast genome sequences have been widely used to reconstruct plant phylogenies, and the rapid improvements in sequencing technologies have led to the routine sequencing of complete chloroplast genomes^{45,46}. *D. insignis* belongs to the Lardizabalaceae family in the early diverging eudicots. To identify its phylogenetic position, 82 common genes were extracted from the chloroplast genomes of 33 species from all families of early diverging eudicots and others⁴⁷. *Ceratophyllum demersum* was set as the outgroup.

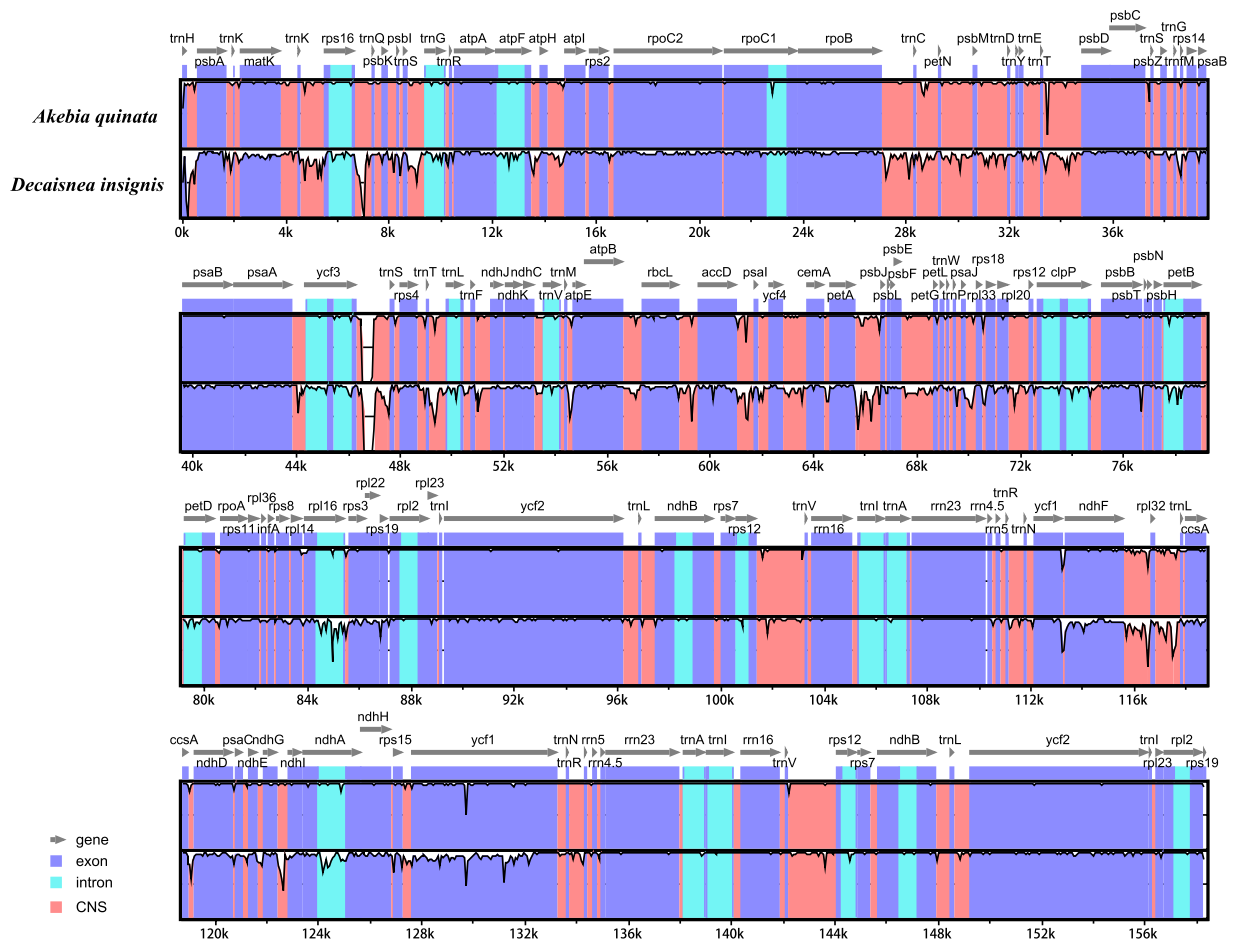


Figure 4. Visualization alignment of chloroplast genome sequences of *D. insignis*, *A. quinata* and *A. indica*. VISTA based similarity graphical information portraying sequence identity of *D. insignis* with reference *A. indica* chloroplast genomes. Grey arrows above the alignment indicate the orientation of genes. Purple bars represent exons, blue ones represent introns, and pink ones represent non-coding sequences (CNS). A cut-off of 50% identity was used for the plots. The Y-scale axis represents the percent identity within 50–100%.

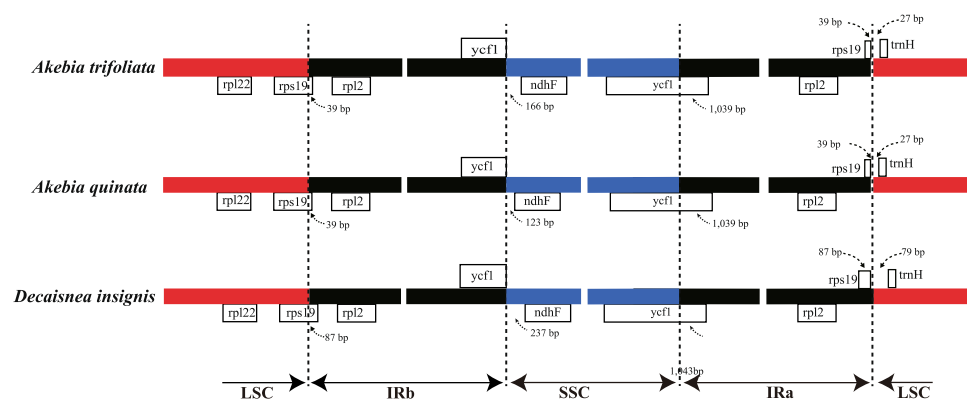


Figure 5. Comparison of chloroplast genome borders of LSC, SSC, and IRs among three Lardizabalaceae species.

After concatenating the alignment, all positions containing gaps and missing data were eliminated, and the sequence alignment comprised 68,668 characters. In the maximum likelihood (ML) and Bayesian inference (BI) tree, most of the nodes had a 100% bootstrap value and 1.0 Bayesian posterior probability. Both the ML and BI trees had similar phylogenetic topologies, which strongly supported the position of *D. insignis* as a sister of the closely related *Akebia* species in the family Lardizabalaceae (Fig. 7). The early diverging eudicot lineages, including the five major lineages (Ranunculales, Sabiales, Proteales, Buxales, and Trochodendrales), formed

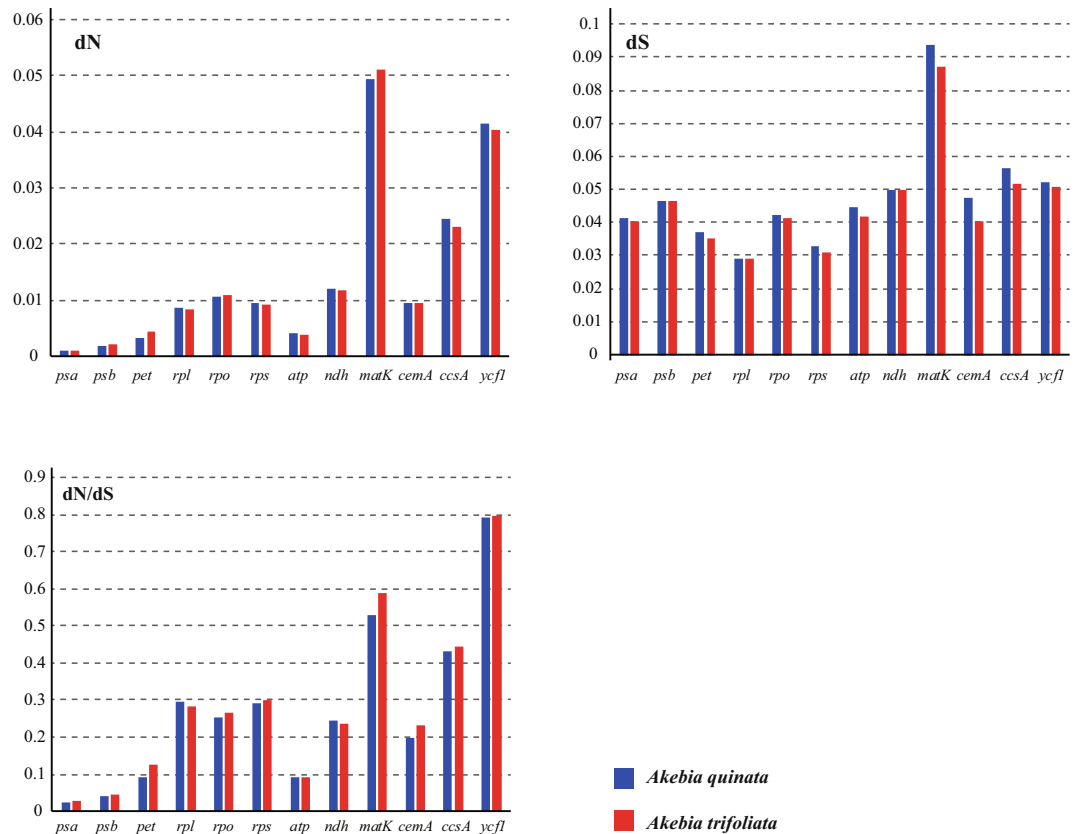


Figure 6. Nonsynonymous substitution (dN), synonymous substitution (dS), and dN/dS values for individual genes or gene groups.

monophyly clades with 100% bootstrap support in the ML analyses and 1.0 Bayesian posterior probability in BI. Lardizabalaceae was a member of Ranunculales and was present at the basal position of the order. The phylogenetic positions of this group were in agreement with recent studies⁴⁸. Although taxon sampling was inadequate and we could not perform a deeper phylogenetic analysis of Lardizabalaceae, our data would provide an example for future genome-scale phylogenetic studies in Lardizabalaceae.

Conclusions

Using Illumina high-throughput sequencing technology, we obtained the complete sequence of the *D. insignis* chloroplast genome. The genomic organization and gene order of *D. insignis* were in agreement with the previously reported chloroplast genomes in Lardizabalaceae. The ML and BI phylogenetic trees strongly supported the position of Lardizabalaceae as a member of the order Ranunculales. The data obtained in this study will be beneficial for further investigations on *D. insignis*. Moreover, the availability of the chloroplast genomes provides a powerful genetic resource for the molecular phylogeny and biological study of this wild resource plant.

Methods

Taxon sampling, DNA extraction and sequencing. Fresh leaves of *D. insignis* were collected from a tree at the Research Institute of Forestry, Chinese Academy of Forestry. Fresh leaves were immediately dried with silica gel before DNA extraction. Total genomic DNA was extracted and purified following the method of Li *et al.*⁴⁹. DNA was randomly fragmented into 400–600 bp using an ultrasonicator. An Illumina paired-end cpDNA library was constructed using the NEBNext[®] Ultra[™] DNA Library Prep Kit following the manufacturer's instructions. Paired-end sequencing (2 × 150 bp) was carried out on an Illumina HiSeq 4000 platform.

Genome assembly and genome annotation. The paired-end reads were qualitatively assessed and assembled with SPAdes 3.6.1⁵⁰. Chloroplast genome sequence contigs were selected from the initial assembly by performing a BLAST search using the *Akebia quinata* chloroplast genome sequence as a reference (GenBank accession KX611091)¹⁷. The selected contigs were assembled with Sequencher 5.4.5. Ambiguous nucleotides, or gaps in the chloroplast genome sequences were filled by PCR amplification and Sanger sequencing. The four junctions between the inverted repeats (IRs) and small single copy (SSC)/large single copy (LSC) regions were checked by amplification with specific primers followed by Sanger sequencing⁷¹. Chloroplast genome annotation was performed with Plann⁵² based on chloroplast genome sequence of *Akebia quinata*. A chloroplast genome map was drawn using Genome Vx software⁵³.

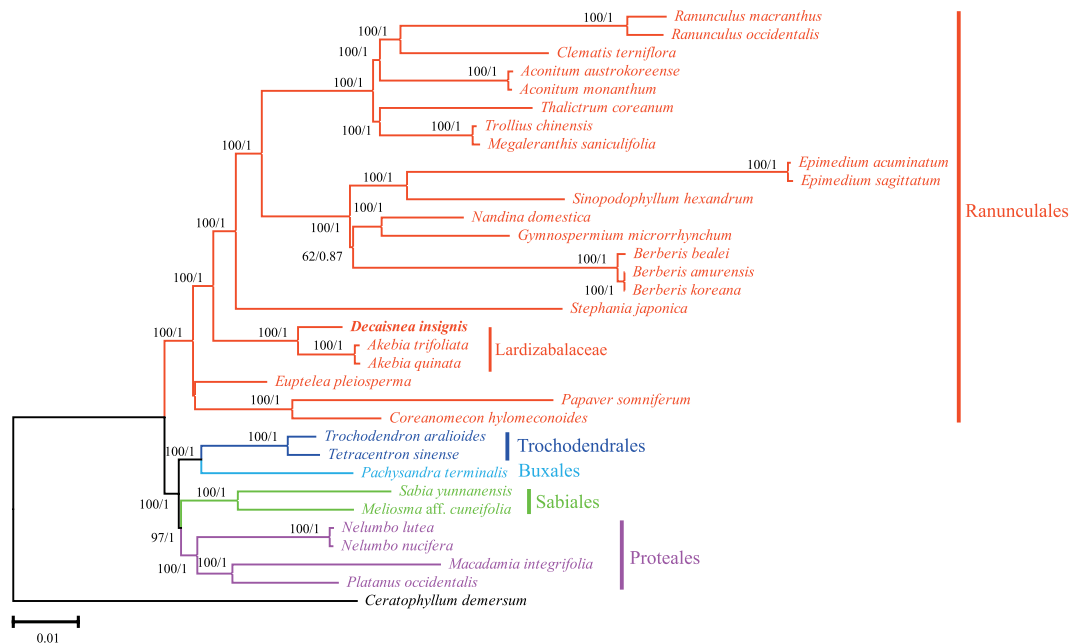


Figure 7. Phylogenetic tree reconstruction of 33 taxa using maximum likelihood and Bayesian inference based on concatenated sequences of 82 genes. ML topology was shown with ML bootstrap support value/Bayesian posterior probability given at each node.

Codon usage. Codon usage was determined for all protein-coding genes. To examine deviations in synonymous codon usage by avoiding the influence of the amino acid composition, the relative synonymous codon usage (RSCU) was determined using MEGA 6 software⁵⁴.

Analysis of single sequence repeats and tandem repeats. Perl script MISA (MicroSatellite; <http://pgrc.ipk-gatersleben.de/misa>) was used to detect single sequence repeats (SSR) within the chloroplast genome, with the parameters set at >10 for mononucleotide, >5 for dinucleotide, >4 for trinucleotide, and >3 for tetra-nucleotide, pentanucleotide, and hexanucleotide SSRs. REPuter was used to visualize the repeat sequences in *D. insignis* by forward vs. reverse complement (palindromic) alignment³³. The minimal repeat size was set at 30 bp, and the identity of repeats was $\geq 90\%$.

Comparative genome analysis. The complete cp genome of *D. insignis* was compared with two *Akebia* species, *A. quinata* and *A. indica*, using the mVISTA program in a Shuffle-LAGAN mode⁵⁵. *A. indica* was set as a reference. The chloroplast genome borders of LSC, SSC, and IRs were compared according to their annotations.

Synonymous and nonsynonymous substitution rate analysis. The relative rates of sequence divergence were analysed using the PAML v4.4 package⁵⁶. *D. insignis* was used as an outgroup. The program yn00 was employed to estimate dN, dS, and dN/dS under a F3 \times 4 substitution matrix using the Nei-Gojobori method. Genes with the same functions were grouped following previous studies^{5, 57–59}. Analyses were carried out on datasets corresponding to the same functions, i.e., for *atp*, *pet*, *ndh*, *psa*, *psb*, *rpl*, *rpo*, and *rps*, and datasets corresponding to singular genes, i.e., for *cemA*, *matK*, *ccsA*, and *ycf1*.

Phylogenetic analysis. The chloroplast genome is uniparentally inherited and does not undergo recombination; thus, its constituent genes should track the same evolutionary history^{45, 46, 60}. Therefore, in this study, we concatenated the 82 chloroplast genes for phylogenetic analysis without concern for strongly conflicting phylogenetic signals. A molecular phylogenetic tree was constructed using 33 angiosperms. The 30 completed chloroplast genome sequences representing the lineages of angiosperms, especially early diverging eudicots, were downloaded from the NCBI Organelle Genome Resource database. GenBank information for all of the chloroplast genomes used for the present phylogenetic analyses can be found in Supplementary Table S4. The sequences were aligned using MAFFT v7⁶¹, and the alignment was manually adjusted.

The best-fitting model of sequence evolution was identified with jModeltest⁶² based on the Akaike Information Criterion (AIC). Maximum likelihood (ML) analysis was performed using the RAxML v 8.0.5 software package⁶³ with 1,000 non-parametric bootstrap replicates.

Bayesian inference (BI) was implemented with MrBayes 3.2.2⁶⁴. Two independent Markov chain Monte Carlo (MCMC) chains were run, each with three heated and one cold chain for 10 million generations. The trees were sampled every 1,000 generations, with the first 25% discarded as burn-in. The remaining trees were used to build a 50% majority-rule consensus tree. Analysis was run to completion, and the average standard deviation of the split frequencies was <0.01.

References

1. Reveal, J. L. & Chase, M. W. APG III: Bibliographical Information and Synonymy of Magnoliidae. *Phytotaxa* **19**, 71–134 (2011).
2. The Angiosperm Phylogeny, G. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **181**, 1–20, doi:10.1111/boj.12385 (2016).
3. Wang, H. F., Friedman, C. R., Zhu, Z. X. & Qin, H. N. Early reproductive developmental anatomy in Decaisnea (Lardizabalaceae) and its systematic implications. *Ann. Bot.* **104**, 1243–1253, doi:10.1093/aob/mcp232 (2009).
4. Zhou, Y.-F. & Liu, W.-Z. Laticiferous canal formation in fruits of Decaisnea fargesii: a programmed cell death process? *Protoplasma* **248**, 683–694 (2011).
5. Dong, W., Xu, C., Cheng, T. & Zhou, S. Complete chloroplast genome of *Sedum sarmentosum* and chloroplast genome evolution in Saxifragales. *PLOS ONE* **8**, e77965, doi:10.1371/journal.pone.0077965 (2013).
6. Asaf, S. *et al.* Complete chloroplast genome of Nicotiana otophora and its comparison with related species. *Frontiers in Plant Science* **7**, doi:10.3389/fpls.2016.00843 (2016).
7. Wambugu, P. W., Brozyska, M., Furtado, A., Waters, D. L. & Henry, R. J. Relationships of wild and domesticated rices (*Oryza* AA genome species) based upon whole chloroplast genome sequences. *Sci. Rep.* **5**, 13957, doi:10.1038/srep13957 (2015).
8. Duchene, D. & Bromham, L. Rates of molecular evolution and diversification in plants: chloroplast substitution rates correlate with species-richness in the Proteaceae. *BMC Evol. Biol.* **13**, 65, doi:10.1186/1471-2148-13-65 (2013).
9. Smith, D. R. Mutation Rates in Plastid Genomes: They Are Lower than You Might Think. *Genome Biol. Evol.* **7**, 1227–1234, doi:10.1093/gbe/evv069 (2015).
10. Wu, F. H. *et al.* Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. *BMC Plant Biol.* **10**, 68, doi:10.1186/1471-2229-10-68 (2010).
11. Zhang, Y., Iaffaldano, B. J., Zhuang, X., Cardina, J. & Cornish, K. Chloroplast genome resources and molecular markers differentiate rubber dandelion species from weedy relatives. *BMC Plant Biol.* **17**, 34, doi:10.1186/s12870-016-0967-1 (2017).
12. Dong, W., Liu, J., Yu, J., Wang, L. & Zhou, S. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLOS ONE* **7**, e35071, doi:10.1371/journal.pone.0035071 (2012).
13. Gaut, B., Yang, L., Takuno, S. & Eguiarte, L. E. The Patterns and Causes of Variation in Plant Nucleotide Substitution Rates. *Annual Review of Ecology, Evolution, and Systematics*, **42**, 245–266, doi:10.1146/annurev-ecolsys-102710-145119 (2011).
14. Hollingsworth, P. M., Graham, S. W. & Little, D. P. Choosing and using a plant DNA barcode. *PLOS ONE* **6**, e19254, doi:10.1371/journal.pone.0019254 (2011).
15. Dong, W. *et al.* ycf1, the most promising plastid DNA barcode of land plants. *Sci. Rep.* **5**, 8348, doi:10.1038/srep08348 (2015).
16. Lima, M. S., Woods, L. C., Cartwright, M. W. & Smith, D. R. The (in)complete organelle genome: exploring the use and non-use of available technologies for characterizing mitochondrial and plastid chromosomes. *Mol. Ecol. Resour.*, doi:10.1111/1755-0998.12585 (2016).
17. Li, B. *et al.* Development of chloroplast genomic resources for *Akebia quinata* (Lardizabalaceae). *Conservation Genetics Resources* **8**, 447–449, doi:10.1007/s12686-016-0593-0 (2016).
18. Liu, T.-J. *et al.* Complete plastid genome sequence of *Primula sinensis* (Primulaceae): structure comparison, sequence variation and evidence for accD transfer to nucleus. *PeerJ* **4**, e2101, doi:10.7717/peerj.2101 (2016).
19. Yao, X. *et al.* Chloroplast genome structure in *Ilex* (Aquifoliaceae). *Sci. Rep.* **6**, 28559, doi:10.1038/srep28559 (2016).
20. Kuroda, H. *et al.* Translation of psbC mRNAs starts from the downstream GUG, not the upstream AUG, and requires the extended Shine-Dalgarno sequence in tobacco chloroplasts. *Plant Cell Physiol* **48**, 1374–1378, doi:10.1093/pcp/pcm097 (2007).
21. Hu, S. *et al.* Plastome organization and evolution of chloroplast genes in Cardamine species adapted to contrasting habitats. *BMC Genomics* **16**, 306, doi:10.1186/s12864-015-1498-0 (2015).
22. Rohde, W., Gramstat, A., Schmitz, J., Tacke, E. & Pruber, D. Plant viruses as model systems for the study of non-canonical translation mechanisms in higher plants. *J. Gen. Virol.* **75**(Pt 9), 2141–2149, doi:10.1099/0022-1317-75-9-2141 (1994).
23. Asaf, S. *et al.* The Complete Chloroplast Genome of Wild Rice (*Oryza minuta*) and Its Comparison to Related Species. *Frontiers in Plant Science* **8**, doi:10.3389/fpls.2017.00304 (2017).
24. Liu, Q. P. & Xue, Q. Z. Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. *Journal of Genetics* **84**, 55–62, doi:10.1007/Bf02715890 (2005).
25. Ivanova, Z. *et al.* Chloroplast Genome Analysis of Resurrection Tertiary Relict *Haberlea rhodopensis* Highlights Genes Important for Desiccation Stress Response. *Frontiers in Plant Science* **8**, doi:10.3389/fpls.2017.00204 (2017).
26. Wang, Y. *et al.* Complete Chloroplast Genome Sequence of *Aquilaria sinensis* (Lour.) Gilg and the Evolution Analysis within the Malvalesorder. *Frontiers in Plant Science* **7**, doi:10.3389/fpls.2016.00280 (2016).
27. Zhou, S. *et al.* How many species of bracken (*Pteridium*) are there? Assessing the Chinese brackens using molecular evidence. *Taxon* **63**, 509–521, doi:10.12705/633.9 (2014).
28. Qi, W. *et al.* High-throughput development of simple sequence repeat markers for genetic diversity research in *Crambe abyssinica*. *BMC Plant Biol.* **16**, 139, doi:10.1186/s12870-016-0828-y (2016).
29. Yu, J. *et al.* PMDBase: a database for studying microsatellite DNA and marker development in plants. *Nucleic Acids Res.* **45**, D1046–D1053, doi:10.1093/nar/gkw906 (2017).
30. Wang, L., Wuyun, T.-n., Du, H., Wang, D. & Cao, D. Complete chloroplast genome sequences of *Eucommia ulmoides*: genome structure and evolution. *Tree Genetics & Genomes* **12**, 1–15, doi:10.1007/s11295-016-0970-6 (2016).
31. Sablok, G. *et al.* ChloroMitoSSRDB: Open Source Repository of Perfect and Imperfect Repeats in Organelle Genomes for Evolutionary Genomics. *DNA Res.*, doi:10.1093/dnares/dss038 (2013).
32. Ochoterena, H. Homology in coding and non-coding DNA sequences: a parsimony perspective. *Plant Syst. Evol.* **282**, 151–168, doi:10.1007/s00606-008-0095-y (2009).
33. Kurtz, S. *et al.* REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **29**, 4633–4642 (2001).
34. Ni, L., Zhao, Z., Dorje, G. & Ma, M. The Complete Chloroplast Genome of Ye-Xing-Ba (*Scrophularia dentata*; Scrophulariaceae), an Alpine Tibetan Herb. *PLOS ONE* **11**, e0158488, doi:10.1371/journal.pone.0158488 (2016).
35. Yang, Y. *et al.* Comparative Analysis of the Complete Chloroplast Genomes of Five *Quercus* Species. *Front Plant Sci* **7**, 959, doi:10.3389/fpls.2016.00959 (2016).
36. Borsch, T. & Quandt, D. Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. *Plant Syst. Evol.* **282**, 169–199, doi:10.1007/s00606-009-0210-8 (2009).
37. Liu, C. *et al.* PTIGS-Idit, a system for species identification by DNA sequences of the psbA-trnH intergenic spacer region. *BMC Bioinformatics* **12**(Suppl 13), S4, doi:10.1186/1471-2105-12-S13-S4 (2011).
38. Pang, X. *et al.* Utility of the trnH-psbA intergenic spacer region and its combinations as plant DNA barcodes: a meta-analysis. *PLOS ONE* **7**, e48833, doi:10.1371/journal.pone.0048833 (2012).
39. Shaw, J. *et al.* The tortoise and the hare II: Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *Am. J. Bot.* **92**, 142–166 (2005).
40. Shaw, J., Lickey, E. B., Schilling, E. E. & Small, R. L. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. *Am. J. Bot.* **94**, 275–288 (2007).
41. Drouin, G., Daoud, H. & Xia, J. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Mol. Phylogenet. Evol.* **49**, 827–831, doi:10.1016/j.ympev.2008.09.009 (2008).

42. Yang, Z. H. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).
43. Matsuoka, Y., Yamazaki, Y., Ogihara, Y. & Tsunewaki, K. Whole chloroplast genome comparison of rice, maize, and wheat: implications for chloroplast gene diversification and phylogeny of cereals. *Mol. Biol. Evol.* **19**, 2084–2091 (2002).
44. Gaut, B. S., Muse, S. V. & Clegg, M. T. Relative rates of nucleotide substitution in the chloroplast genome. *Mol. Phylogenet. Evol.* **2**, 89–96, doi:10.1006/mpev.1993.1009 (1993).
45. Sun, L. *et al.* Chloroplast phylogenomic inference of green algae relationships. *Sci. Rep.* **6**, 20528, doi:10.1038/srep20528 (2016).
46. Goremkyin, V. V., Nikiforova, S. V., Cavalieri, D., Pindo, M. & Lockhart, P. The Root of Flowering Plants and Total Evidence. *Syst. Biol.* **64**, 879–891, doi:10.1093/sysbio/syv028 (2015).
47. Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G. & Soltis, D. E. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc. Nat. Acad. Sci. USA* **107**, 4623–4628, doi:10.1073/pnas.0907801107 (2010).
48. Sun, Y. *et al.* 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. *Mol. Phylogenet. Evol.* **96**, 93–101, doi:10.1016/j.ympev.2015.12.006 (2016).
49. Li, J., Wang, S., Jing, Y., Wang, L. & Zhou, S. A modified CTAB protocol for plant DNA extraction. *Chin. Bull. Bot.* **48**, 72–78 (2013).
50. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477, doi:10.1089/cmb.2012.0021 (2012).
51. Dong, W., Xu, C., Cheng, T., Lin, K. & Zhou, S. Sequencing angiosperm plastid genomes made easy: A complete set of universal primers and a case study on the phylogeny of Saxifragales. *Genome Biol. Evol.* **5**, 989–997, doi:10.1093/gbe/evt063 (2013).
52. Huang, D. I. & Cronk, Q. C. B. Plann: A Command-Line Application for Annotating Plastome Sequences. *Applications in Plant Sciences* **3**, 1500026, doi:10.3732/apps.1500026 (2015).
53. Conant, G. C. & Wolfe, K. H. GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics* **24**, 861–862, doi:10.1093/bioinformatics/btm598 (2008).
54. Tamura, K., Stecher, G., Peterson, D., Filipiński, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729, doi:10.1093/molbev/mst197 (2013).
55. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32**, W273–W279, doi:10.1093/Nar/Gkh458 (2004).
56. Yang, Z. H. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591, doi:10.1093/molbev/msm088 (2007).
57. Chang, C. C. *et al.* The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): Comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Mol. Biol. Evol.* **23**, 279–291 (2006).
58. Guisinger, M. M., Kuehl, J. N. V., Boore, J. L. & Jansen, R. K. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. *Proc. Nat. Acad. Sci. USA* **105**, 18424–18429 (2008).
59. Wu, M. *et al.* The Complete Chloroplast Genome of *Guadua angustifolia* and Comparative Analyses of Neotropical–Paleotropical Bamboos. *PLOS ONE* **10**, e0143792, doi:10.1371/journal.pone.0143792 (2015).
60. Jansen, R. K. *et al.* Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Nat. Acad. Sci. USA* **104**, 19369–19374, doi:10.1073/pnas.0709121104 (2007).
61. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780, doi:10.1093/molbev/mst010 (2013).
62. Santorum, J. M., Darriba, D., Taboada, G. L. & Posada, D. jmodeltest.org: selection of nucleotide substitution models on the cloud. *Bioinformatics* **30**, 1310–1311, doi:10.1093/bioinformatics/btu032 (2014).
63. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313, doi:10.1093/bioinformatics/btu033 (2014).
64. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542, doi:10.1093/sysbio/sys029 (2012).

Acknowledgements

This work was supported by the National Forest Germplasm Resources Shared Service Platform 2016.

Author Contributions

B.L. and Y.Z. designed the experiment; B.L., F.L., P.H., and W.G. collected samples and performed the experiment; B.L. analyzed the data and wrote the manuscript; All of the authors have read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-10409-8

Competing Interests: The authors declare that they have no competing interests.

Accession Code: *D. insignis* chloroplast genome are available in GenBank database (accession number: KY200671).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017