Received: 29 June 2020 | Revised: 18 January 2021 | Accepted: 21 January 2021

DOI: 10.1002/ece3.7312

ORIGINAL RESEARCH

Ecology and Evolution Open Access WILEY

Highly variable chloroplast genome from two endangered Papaveraceae lithophytes *Corydalis tomentella* and *Corydalis saxicola*

Fengming Ren[1,2] | **Liqiang Wang**[3] | **Ying Li**[1,4] | **Wei Zhuo**[2] | **Zhichao Xu**[1,4] | **Haojie Guo**[5] | **Yan Liu**[2] | **Ranran Gao**[1] | **Jingyuan Song**[1,4]

[1]Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Key Lab of Chinese Medicine Resources Conservation, State Administration of Traditional Chinese Medicine of the People's Republic of China, Beijing, China

[2]Medicinal Biological Technology Research Center, Chongqing Institute of Medicinal Plant Cultivation, Bio-Resource Research and Utilization Joint Key Laboratory Sichuan and Chongqing, Chongqing, China

[3]College of Pharmacy, Heze University, Heze, China

[4]Engineering Research Center of Chinese Medicine Resource, Ministry of Education, Beijing, China

[5]Wuhu Institute of Technology, Wuhu, China

Correspondence
Jingyuan Song, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Key Lab of Chinese Medicine Resources Conservation, State Administration of Traditional Chinese Medicine of the People's Republic of China, Beijing 100193, China.
Email: jysong@implad.ac.cn

Funding information
2019 Basic Scientific Research Project of Chongqing, Grant/Award Number: 19KF10-2012; Science and Technology Project of Traditional Chinese Medicine in Chongqing, Grant/Award Number: ZY201702143; National Natural Science Foundation of China, Grant/Award Number: 81874339; National Science and Technology Major Project for "Significant New Drugs Development", Grant/Award Number: 2019ZX09201005-006-003

Abstract

The increasingly wide application of chloroplast (cp) genome super-barcode in taxonomy and the recent breakthrough in cp genetic engineering make the development of new cp gene resources urgent and significant. *Corydalis* is recognized as the most genotypes complicated and taxonomically challenging plant taxa in Papaveraceae. However, there currently are few reports about cp genomes of the genus *Corydalis*. In this study, we sequenced four complete cp genomes of two endangered lithophytes *Corydalis saxicola* and *Corydalis tomentella* in *Corydalis*, conducted a comparison of these cp genomes among each other as well as with others of Papaveraceae. The cp genomes have a large genome size of 189,029–190,247 bp, possessing a quadripartite structure and with two highly expanded inverted repeat (IR) regions (length: 41,955–42,350 bp). Comparison between the cp genomes of *C. tomentella*, *C. saxicola*, and Papaveraceae species, five NADH dehydrogenase-like genes (*ndh*F, *ndh*D, *ndh*L, *ndh*G, and *ndh*E) with *psa*C, *rpl*32, *ccs*A, and *trn*L-UAG normally located in the SSC region have migrated to IRs, resulting in IR expansion and gene duplication. An up to 9 kb inversion involving five genes (*rpl*23, *ycf*2, *ycf*15, *trn*I-CAU, and *trn*L-CAA) was found within IR regions. The *acc*D gene was found to be absent and the *ycf1* gene has shifted from the IR/SSC border to the SSC region as a single copy. Phylogenetic analysis based on the sequences of common CDS showed that the genus *Corydalis* is quite distantly related to the other genera of Papaveraceae, it provided a new clue for recent advocacy to establish a separate Fumariaceae family. Our results revealed one special cp genome structure in Papaveraceae, provided a useful resources for classification of the genus *Corydalis*, and will be valuable for understanding Papaveraceae evolutionary relationships.

KEYWORDS

chloroplast genome, *Corydalis saxicola*, *Corydalis tomentella*, Papaveraceae, taxonomic study

1 | INTRODUCTION

Chloroplasts (cp), generally considered to have originated from ancient cyanobacteria, are the main site of photosynthesis and energy conversion in plant cells, containing the major enzyme systems for photosynthesis and a highly conserved genome (Ahlert et al., 2003; Moore et al., 2010). With the development of high-throughput sequencing technology, cp genomics has made rapid progress (Li et al., 2015). The National Center for Biotechnology Information (NCBI) database included 377 complete cp genome sequences in 2010 and had more than 10,381 sequences in 2020 (https://www.ncbi.nlm.nih.gov/genome/browse/), a nearly 30-fold increase over 10 years. Currently, cp genomics research is an intense area of botanical and genomic study.

Correct understanding of the relationship between different biological groups is the main focus of phylogenetic biology, the basis of taxonomy and naming, and a foundation for research in other branches of biology (Chen et al., 2016). Compared with traditional molecular markers, the cp genomes provide specific advantages for establishing plant phylogenetic relationships and taxonomic research (Guo et al., 2017). The length of cp genomes is usually 115–165 kb, a modest size that is easily sequenced. The longer sequence provided more sufficient information for phylogenetic analysis. Relatively conserved gene sequences allow produce co-linearity among plant groups, and the evolution rates of coding regions and noncoding regions are significantly different to be suited for phylogenetic analysis of different ranks (Clegg et al., 1994). Taxonomists have used cp genomes to study plant phylogenetics and advocated for use of cp genomes as a super DNA barcode for species identification (Guo et al., 2017).

In recent years, a large number of cp genome have been sequenced, providing abundant data that can be used for plant phylogeny research to more accurately reveal the true evolutionary relationships between species and effectively solve difficult phylogenetic relationship problems in the study of complex plant taxa (Guo et al., 2019; Jansen et al., 2006; Zhang et al., 2017). Cp genomes have been successfully used as a "super barcode" to identify many taxonomically difficult species (Cui et al., 2019; Ying et al., 2019). With the reduced cost of sequencing and the development of bioinformatics technology, cp genome will be extensively used in future studies of plant taxonomy.

Corydalis DC., the largest genus of Papaveraceae, is recognized as one of the most taxonomically challenging plant taxa (Magnus et al., 1996). It has extremely complex morphological variation because of typical reticulate evolution and intense differentiation during evolution (Wu et al., 1996). Taxonomic study of the genus on the basis of morphological characteristics has been very difficult (Lu et al., 2018). Cp genomes have been proven effective for phylogenetic research of many taxonomically complex taxa. However, there currently are few reports about cp genomes of the genus *Corydalis*, but see two plants, *Corydalis trisecta* and *Corydalis conspersa* (Kanwal et al., 2019; Wu et al., 2020). Therefore, it is necessary to sequence the cp genomes of *Corydalis* plants in order to provide more accurate basis for the classification and identification of this genus.

In this study, high-throughput sequencing and comparative genomics were used to study the cp genomes of two *Corydalis* plants: *Corydalis saxicola* and *Corydalis tomentella*. They belong to Sect. *Thalictrifoliae* Fedde of the genus *Corydalis*, which grows in dry cracks of limestone (Figure 1) and is known as lithophytes. There are little available soil and water on the limestone, so they have been subjected to extreme environmental conditions, such as high temperature, drought, and high PH (Ren et al., 2019). Then, we asked whether the cp genome structures of these two lithophytes had special variation under the extremely harsh lithophytic environment, and whether these variation would affect their classification and identification. We sequenced four complete cp genome sequences from these two plants, described their genomic characteristics, conducted comparisons between these genomes and other Papaveraceae cp genomes, and analyzed the phylogenetic relationships on the basis of common protein CDS. Our study aim was to assess structural variation, and provide valuable resources for classification of the genus *Corydalis*.

2 | MATERIALS AND METHODS

2.1 | Materials, DNA extraction and sequencing

Plant materials were provided by the Chongqing Institute of Medicinal Plant Cultivation (CQIMPC) and identified by researcher Zhengyu Liu as *C. tomentella* Franch. and *C. saxicola* Bunting. The

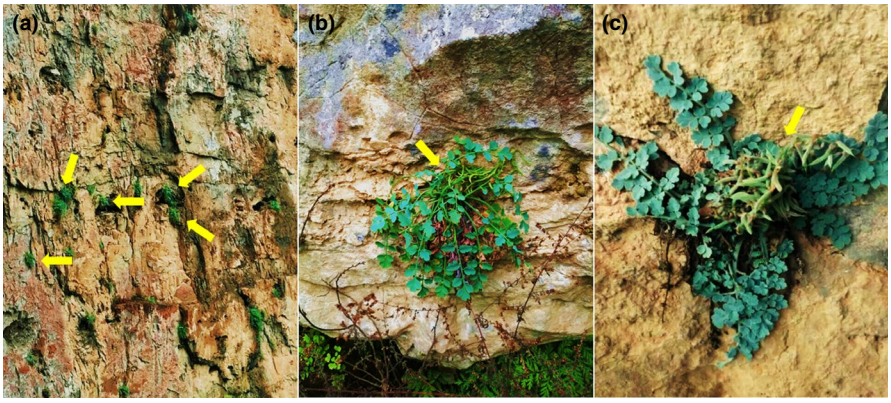

FIGURE 1 The habitat of *C. saxicola* and *C. tomentella*. (a) The distant view of steep cliff growing *C. saxicola*; (b) the close shot of *C. saxicola*; (c) the close shot of *C. tomentella*. The yellow arrows indicated the *Corydalis* plants

voucher specimens of the two species were deposited in CQIMPC, and the specimen accession numbers were *NC-CQIMPC201651*, *NC-CQIMPC201652 NC-CQIMPC201661*, and *NC-CQIMPC201662*, respectively. We collected young leaves from selected plants that were vigorous, healthy, and disease-free. These leaves were wiped with 70% alcohol and repeatedly washed with sterile water before genomic DNA extraction. Total DNA was extracted using a Tiangen plant genomic DNA extraction kit (Tiangen Biotech Co.), and the DNA quality and concentration were detected using 1% agarose electrophoresis and a Nanodrop 2000. The DNA was sheared to yield approximately 500 bp long fragments for paired-end library construction. The library was sequenced on Illumina HiSeq 4000 Platform (Illumina) according to the standard protocol of manufacturer's manual. Approximately 3–5 Gb raw paired-end reads (2 × 150 bp) were obtained for each specimen.

2.2 | Genome assembly and annotation

The cp genome were assembled on a Linux system. First, raw sequencing data were filtered using Trimmomatic (Version 0.36) to get the high-quality clean data (Bolger et al., 2014). In the second step, we used the thirteen chloroplast genome sequences of Papaveraceae species which were downloaded from GenBank to establish a Basic Local Alignment Search Tool (BLASTn) database. Then the clean data were mapped to the BLAST database, and the mapped reads which were considered as reads from chloroplast genome were extracted. Next step, the extracted reads were assembled to contigs using SOAPdenovo2 (Luo et al., 2012). At last, SSPACE was used to construct the scaffold of the chloroplast genome (Boetzer et al., 2011), and GapCloser was used to fill gaps (Luo et al., 2012). The completed genomes were annotated using CPGAVAS2 (Shi et al., 2019), and the results were modified for starter and terminator revisions by Apollo software (Lee et al., 2009). CPGAVAS2 software was used to convert revised GFF3 format annotation results into a sqn format for NCBI submission. Sequin software was used to check and correct unsatisfactory comments in the sqn file, and the corrected results were submitted to the NCBI database. Physical maps of the cp genomes were drawn by GenomeDRAW (Marc et al., 2013) using a GB format file exported from the sqn file by sequin software.

2.3 | Genome structure analyses and genome comparison

GC content was analyzed using MEGA6.06 software (Tamura et al., 2013). The distribution of codon usage was investigated using CodonW software with the RSCU ratio (Sharp & Li, 1987; Zhou et al., 2017). MISA software (http://pgrc.ipk-gatersleben.de/misa/) was used to detect simple sequence repeats (SSRs) (MISA-Microsatellite Identification Tool, 2017). Parameters were set as follows: no less than 8 single-base repeat units; no less than 4 units with 2, 3 bases in one unit; and no less than 3 units with 4, 5, 6 bases in one unit (Huang et al., 2020). Tandem Repeats Finder v4.0.4 software (Benson, 1999) was used to detect tandem repeat sequences, and the default parameter was set to 2-7-7 -80-10-50-500-f-d-m (Li et al., 2014). REPuter software (http://bibiserv.techfak.uni-bielefeld.de/reputer) was used to detect scattered repeating sequences (>30 bp) using the parameter: hamming distance = 3 (Stefan et al., 2001). VISTA software was used to compare multiple cp genomes (Frazer et al., 2004).

2.4 | Phylogenetic analysis

A total of 13 cp whole genome sequences were used in cluster analysis. Eleven genomes were from Papaveraceae (*C. tomentella* MT093187 MT077878, *C. saxicola* MT077878 MT077879, *Papaver somniferum* NC029434, *Papaver orientale* NC037832, *Papaver rhoeas* MF943221, *Meconopsis racemosa* MH394401 NC039625, *Macleaya microcarpa* NC039623, and *Coreanomecon hymenoides* NC031446), and *Coptis chinensis* (NC001879) and *Nicotiana tabacum* (NC036485) genomes were included as outgroups. Of the Papaveraceae genomes, four genomes were newly sequenced in this study, and nine genomes were downloaded from the NCBI database. Common protein coding sequences were extracted from the cp genome sequences (Li et al., 2014), and multiple global alignments of the protein coding sequences was performed using the Clustalw module in MEGA6.06 software. Maximum-Likelihood (ML) phylogenetic tree was constructed by MEGA6.06 software (Tamura et al., 2013). The program operating parameters were set as follows: a Tamura–Nei nucleotide substitution model with 1,000 bootstrap repetitions, accompanied by Gamma distributed with invariant site (G + I) rates, and partial deletion of gaps/missing data. The model with the highest bootstrap values at each node was determined to be the most appropriate model.

3 | RESULTS

3.1 | Chloroplast genomes features

Approximately, 5.12, 5.23, 2.68, and 2.77 Gb raw paired-end reads (2 × 150 bp) were obtained from the Illumina HiSeq 4000 Platform for MHJ-1, MHJ-2, YHL-1, and YHL-2, respectively. The raw sequencing data were filtered using Trimmomatic, 4.54, 4.61, 2.25, and 2.30 Gb of clean data were used to assemble the complete chloroplast genome. The complete *C. tomentella* genomes were 190,198–190,247 bp long and exhibited a typical angiosperm circular cp structure, containing four regions: large single-copy region (LSC: 96,530–96,701 bp), small single-copy region (SSC: 9,636–9,664 bp), and a pair of inverted repeats (IR: 41,955–42,002 bp) (Figure 2). The GC content of the genome and each genomic region was also typical of angiosperm cp structure. Specific lengths and contents are shown in Figure 2 and Table 1. The lengths of the two complete *C. saxicola* genomes were 189,029 and 189,155 bp, which were slightly smaller

than those of *C. tomentella*. The cp genome structure, size of each region, and GC content were similar between the two species (Table 1).

CPGAVAS2 was used to annotate the cp genomes of *C. tomentella* and *C. saxicola*. Removing duplicate genes, a total of 119 annotated genes (Figure 2, Table 2 and Table S1), including 78 protein-coding genes, 37 tRNA genes, and four rRNA genes, were identified from the *C. tomentella*. There were 28 genes in the IR region, of which 15 were involved in gene expression. Introns greatly affect regulated selective splicing in the genome. There were 19 genes that contain introns in the *C. tomentella* cp genome. Most intron genes contained only one intron, while the *ycf*3 gene contained two introns. There were 12 introns with a length of more than 700 bp, and the longest gene was *trn*K-UUU with a length of 2,478 bp. The gene features of *C. saxicola* cp genome were similar to those of *C. tomentella*. The

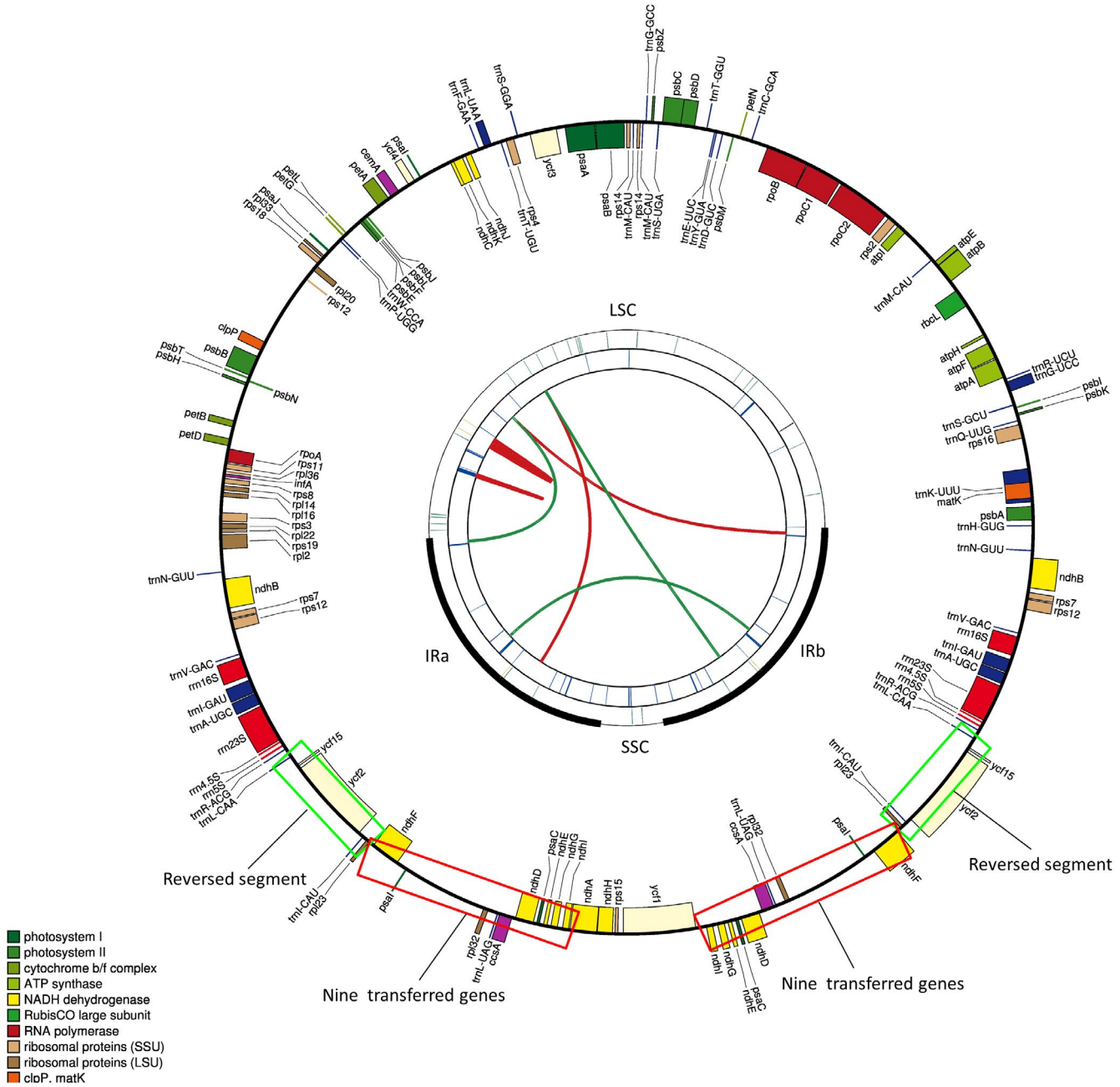

FIGURE 2 Schematic representation of the chloroplast genomes of *C. tomentella*. The map contains four rings. From the center going outward, the first circle shows forward and reverse repeats connected with red and green arcs, respectively. The next circle shows tandem repeats marked with short bars. The third circle shows microsatellite sequences identified by MISA. The fourth circle is drawn using drawgenemap and shows the gene structure of the plastome. The genes are colored on the basis of their functional categories. Genes inside and outside of the circle are transcribed in clockwise and counterclockwise directions, respectively. IR, inverted repeat; LSC, large single copy; SSC, small single copy. The red rectangles indicated the nine gens (*ndh*F, *ndh*D, *ndh*L, *ndh*G, *ndh*E, *psa*C, *ccs*A, *rpl32*, and *trn*L-UAG) normally located in the SSC region have migrated to IRs; the green rectangles indicated the reversed segment involving five genes (*rpl*23, *ycf*2, *ycf*15, *trn*I-CAU, and *trn*L-CAA)

TABLE 1 Summary of chloroplast genome features of *C. tomentella* and *C. saxicola*

Species	Voucher No.	Genbank No.	Length (bp)				GC content (%)			
			Total	IR	LSC	SSC	Total	IR	LSC	SSC
Corydalis tomentella	MHJ1	MT093187	190,247	41,955	96,701	9,636	40.3	42.2	39.2	35.4
	MHJ2	MT077878	190,198	42,002	96,530	9,664	40.2	42.2	39.0	35.4
Corydalis saxicola	YHL1	MT077877	189,155	42,350	94,744	9,711	40.2	42.2	39.1	35.1
	YHL2	MT077879	189,029	42,164	94,993	9,708	40.3	42.2	39.1	35.1

TABLE 2 List of genes in the two *Corydalis* chloroplast genomes

Group of genes	Gene names	Number of genes
Photosystem I	*psa*A, *psa*B, *psa*C(×2), *psa*I(×2), *psa*J	5 (2)
Photosystem II	*psb*A, *psb*B, *psb*C, *psb*D, *psb*E, *psb*F, *psb*I, *psb*J, *psb*K, *psb*L, *psb*M, *psb*N, *psb*T, *psb*Z	14
Cytochrome b/f complex	*pet*A, *pet*B*, *pet*D*, *pet*G, *pet*L, *pet*N	6
ATP synthase	*atp*A, *atp*B, *atp*E, *atp*F*, *atp*H, *atp*I	6
NADH-dehydrogenase	*ndh*A*, *ndh*B*(×2), *ndh*C, *ndh*D(×2), *ndh*E(×2), *ndh*F(×2), *ndh*G(×2), *ndh*H, *ndh*I(×2), *ndh*J, *ndh*K,	11 (6)
RubisCO large subunit	*rbc*L	1
DNA dependent RNA polymerase	*rpo*A, *rpo*B, *rpo*C1*, *rpo*C2	4
Small subunit of ribosome	*rps*2, *rps*3, *rps*4, *rps*7(×2), *rps*8, *rps*11, *rps*12*(×2), *rps*14, *rps*15, *rps*16*, *rps*18, *rps*19	12 (2)
Large subunit of ribosome	*rpl*2*(×2), *rpl*14, *rpl*16*, *rpl*20, *rpl*22, *rpl*23(×2), *rpl*32(×2), *rpl*33, *rpl*36	9 (3)
Proteins of unknown function	*ycf*1, *ycf2*(×2), *ycf*3**, *ycf*4, *ycf*15(×2)	5 (2)
Other genes	*ccs*A(×2), *cem*A, *inf*A, *mat*K, *clp*P**	5 (1)
Transfer RNAs	37 tRNAs(*C. tomentella*); 38 tRNAs(*C. saxicola*)	37/38
Ribosomal RNAs	*rrn*16S(×2), *rrn*23S(×2), *rrn*4.5S(×2), *rrn*5S(×2)	4 (4)

Note: One or two asterisks followed genes indicate the number of contained introns, respectively. (×2) indicates the number of the repeat unit is 2. The numbers in parenthesis at the line of "Number" indicate the total number of repeated genes.

C. saxicola cp genome contained 120 genes, including 78 protein-coding genes, 38 tRNA genes, and four rRNA genes. Nineteen genes contained introns. The longest intron gene in the *C. saxicola* cp genome was *trn*K-UUU, and its length was also 2,478 bp (Figure 2, Table 2 and Table S1).

3.2 | Variation in genome structural

VISTA software was used to make multiple comparisons of the *C. tomentella* and *C. saxicola* cp genome sequences, and results show that intraspecific variation was small but there were still some interspecific differences (Figure 3). The coding and noncoding regions of *C. saxicola* samples were conserved, while the coding regions of *C. tomentella* samples were conserved but there were differences in several consecutive intergenic regions of *rps*12-*clp*P, *clp*P-*psb*B, and *pet*B-*psb*H. Comparing *C. tomentella* and *C. saxicola*, the most highly divergent regions mainly were observed in coding regions and intergenic regions, including *rpl*20, *rrn*23s, *trn*H-GUG, *trn*N-GUU, *rps*12-*clp*P, *clp*P-*psb*B, *pet*B-*psb*H, and *ycf*1-*ndh*L. On the basis of morphological features and cluster analysis of DNA barcodes, it was found that the two species are closely related (Ren et al., 2019). The cp genome differences between the two species have potential for use as molecular markers for species authentication.

Comparisons with the *N. tabacum* outgroup and Papaveraceae family plants *P. rhoeas*, *P. orientale*, *P. somniferum*, and *C. hylomeconoides* showed that *C. tomentella* and *C. saxicola* cp genomes have distinct cp genome structures. The differences included genome size, number of genes, and a disruption of gene collinearity (Figure 4). First, the *C. tomentella* and *C. saxicola* cp genome sizes (189.1–190.2 kb) were larger than those of *N. tabacum* (155.9 kb) and *P. somniferum* (152.9 kb). Second, the length of intergenic regions in *C. tomentella* and *C. saxicola* cp genomes were longer than those in *N. tabacum* and *P. somniferum*, as seen, for example, in the lengths of intergenic regions for *psa*I/*rpl32* (7 kb) in the IR region and *rps12*/*clp*P (5 kb) in the LSC region. Third, *C. tomentella* and *C. saxicola* cp genome structures were significantly different from those of the other six species, including large-scale gene replication, movement, reversal, and changes in the number and arrangement of genes. Fourth, *C. tomentella* and *C. saxicola* IR regions *were* highly dilated (41.9–42.5 kb). The *ndh*F, *ndh*D, *ndh*L, *ndh*G, *ndh*E, *psa*C, *ccs*A, *trn*L-UAG, and *rpl32* genes, usually located in the SSC region, migrated to the IR regions to become double-copy genes (Figure 1). A few *rpl*19 and *rpl*2 genes migrated from the IR region to the LSC region. In

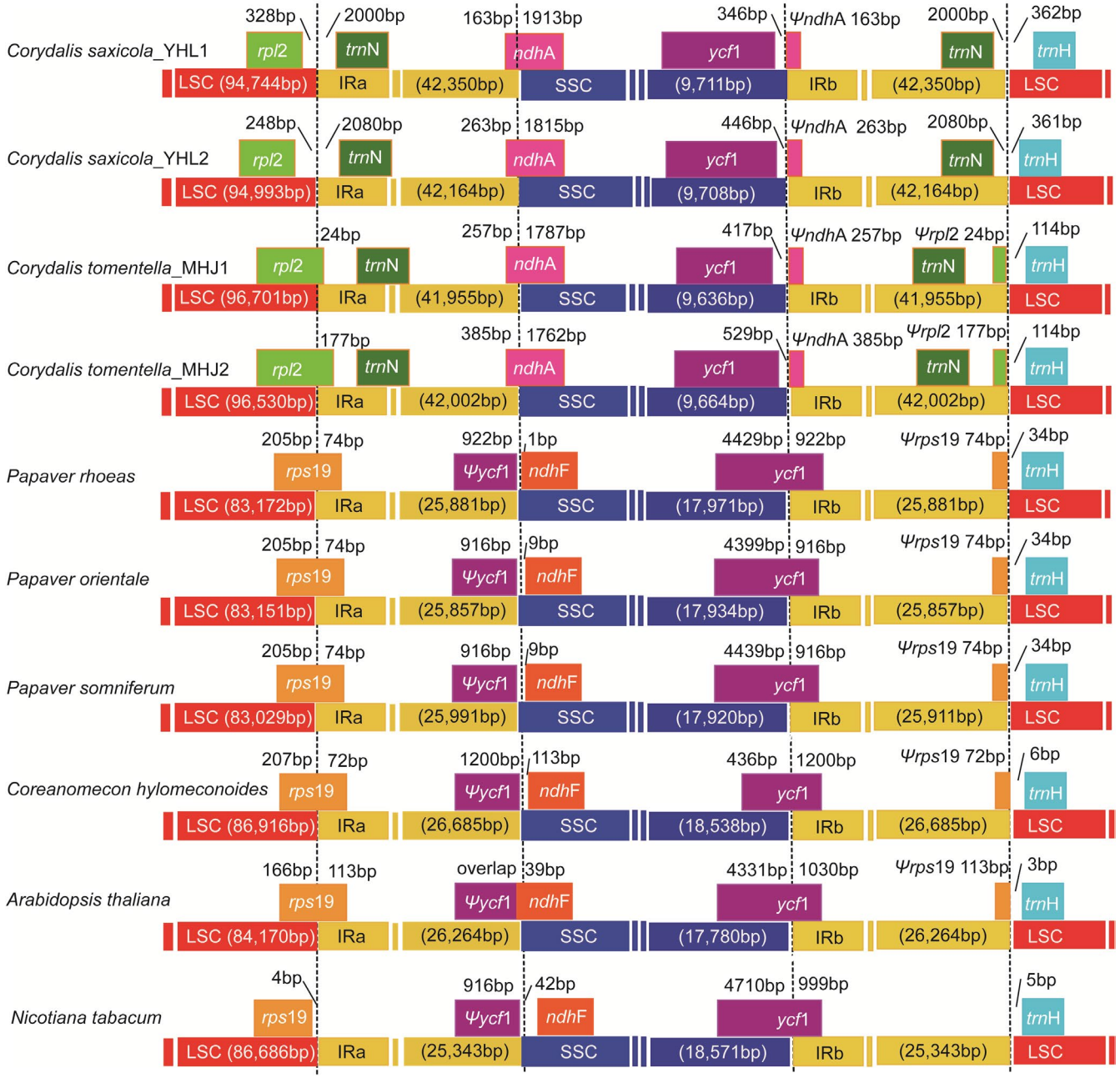

FIGURE 3 Sequence identity plot comparison of the *C. tomentella* and *C. saxicola* cp genomes. Gray arrows and thick black lines above the alignment indicate genes with their orientation and the position of the inverted repeats (IRs), respectively. A cutoff of 70% identity was used for the plots, and the Y-scale represents the percent identity ranging from 50% to 100%

particular, in *C. tomentella* and *C. saxicola*, there is a large fragment (containing *rpl*23, *trn*L-CAU, *ycf*2, *ycf*15, and *trn*L-CAA) that moved within the IR region. Gene migration increased the length of the IR region and decreased the length of the SSC region. Fifth, the LSC region was highly conserved, but the *acc*D gene was lost and the position of the *rbc*L gene changed substantially. In short, both the coding and noncoding regions of *C. tomentella* and *C. saxicola* cp genomes differ greatly from those of other Papaveraceae and tobacco.

Inverted repeat regions are the most conserved regions in the plant plastome, contraction, and expansion at their borders are regarded as the major causes of size variation (Chumley et al., 2006; Xin et al., 2019). We selected four phylogenetically close species (*P. rhoeas*, *P. orientale*, *P. somniferum*, and *C. hylomeconoides*) and two model species (*N. tabacum* and *A. thaliana*) as references for cp genome structure comparisons. Figure 5 displays the detailed information about the boundaries between IR/SSC and IR/LSC in the eight species.

Except for *C. tomentella* and *C. saxicola*, the IRb/SSC boundaries were generally positioned in the coding region of the *ycf*1 gene, resulting in duplication of the 3′ end of this gene. This duplication also produced a variably sized pseudogene *ycf*1 at the IRa/SSC border. The length of the *ycf*1 pseudogene varied from 916 to 1,200 bp. However, the *ycf*1 genes in *C. tomentella* and *C. saxicola* cp genomes

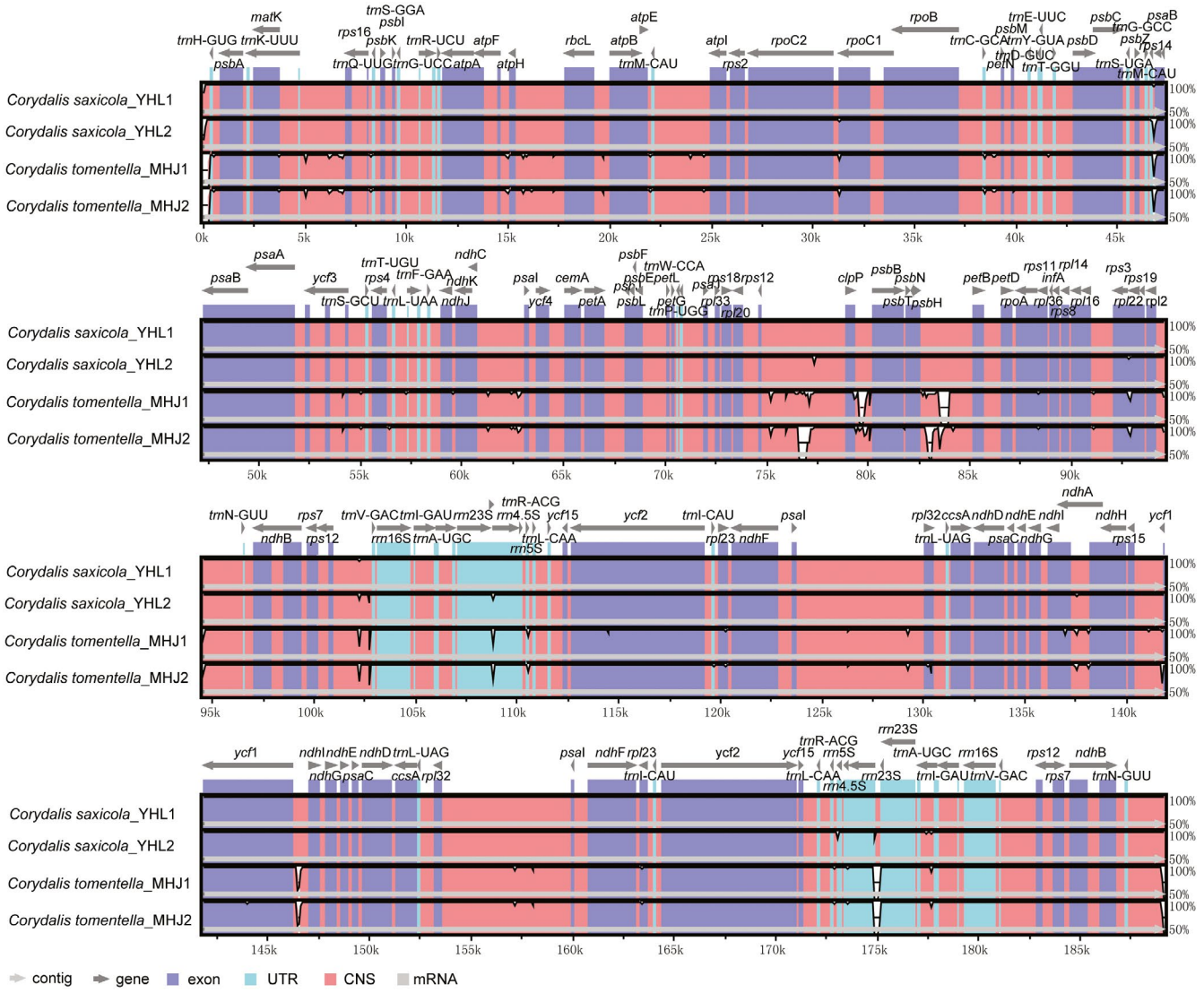

FIGURE 4 Sequence identity plot comparison of the cp genomes of *C. tomentella*, *C. saxicola*, *P. somniferum*, *P. rhoeas*, and *C. hymenoides*. Gray arrows and thick black lines above the alignment indicate genes with their orientation and the position of the inverted repeats (IRs), respectively. A cutoff of 70% identity was used for the plots, and the Y-scale represents the percent identity ranging from 50% to 100%

have been transferred to the SSC region to become a single copy gene. Except for *C. tomentella*, *C. saxicola*, and *N. tabacum*, the LSC/IRb borders of other species were located within the *rps*19 coding region. Correspondingly, a 3′-truncated *rps*19 pseudogene with a length of 74–113 bp was located at the IRb/LSC border. In the *C. tomentella* cp genome, the LSC/IRb border was located in the *rpl*2 coding region. Additionally, in *C. tomentella* and *C. saxicola* cp genomes, the IRa/SSC boundaries were positioned in the *ndh*A coding region, and *trn*N was situated in the IRa and IRb regions, away from the LSC/IRa and IRb/LSC borders. The *trn*H gene was present in LSC regions, away from the IRb/LSC border.

3.3 | Codon usage bias, SSRs, and repeat sequences

Coding sequence codon usage patterns for the *C. tomentella* and *C. saxicola* cp genomes were calculated on the basis of relative synonymous codon usage (RSCU) values. We defined codons with RSCU values greater than 1.00 to be used more frequently, and vice versa. All protein-coding genes in the *C. tomentella* and *C. saxicola* cp genomes were encoded by 52,244 codons and 51,125 codons, respectively (Table S2). The most prevalent amino acid was Leucine in the cp genomes of *C. tomentella* (5,656; 10.83%) and *C. saxicola* (5,528; 10.81%). Conversely, the least frequently utilized amino acid was Cysteine in the cp genomes of these two species (591–634; 1.16%–1.18%). The third position nucleotides in each codon of all the coding genes had a high AT content, at 65.83% and 65.91% for *C. tomentella* and *C. saxicola*, respectively.

Simple sequence repeats are short tandem repeats of 1–6 bp DNA sequences that are widely distributed throughout the cp genome (Lee et al., 2019). In this study, CPGAVAS2 software was used to analyze the sequences and the classification statistics of SSRs with a length greater than or equal to 8 bp. Here, we analyzed the distribution and the type of SSRs contained in *C. tomentella* and *C. saxicola* cp

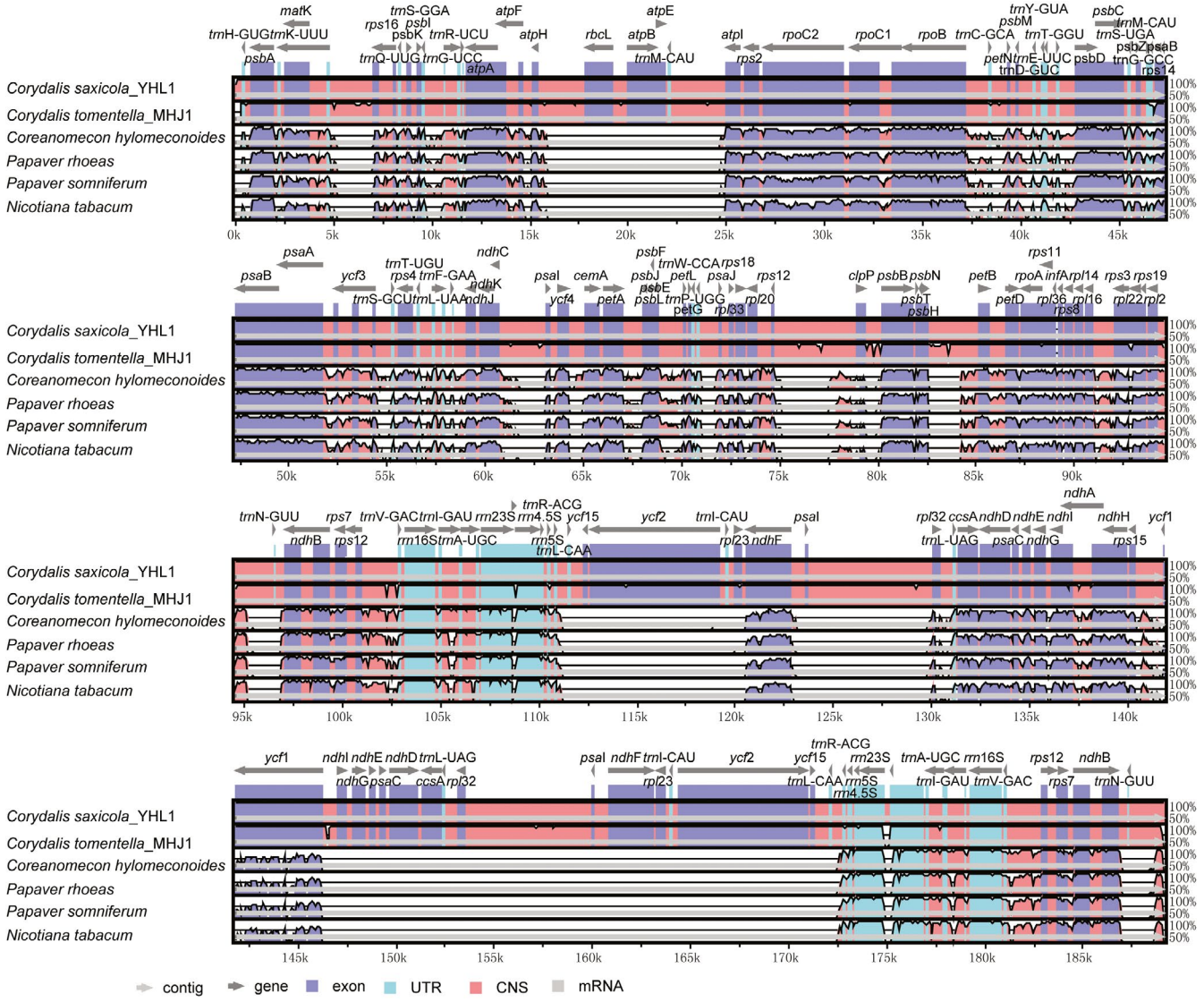

FIGURE 5 Comparison of the borders of LSC, SSC, and IR regions among the eight chloroplast genomes. Number above the gene features indicates the distance between the ends of genes and the border sites. Ψ: pseudogenes

TABLE 3 Interspersed repeat sequences and tandem repeat sequences of *C. saxicola* and *C. tomentella*

Species	Voucher No.	SSR		Interspersed repeat sequences			
		Total	Mono SSR	Total	T	F	P
Corydalis tomentella	MHJ1	172	100	111	61	39	11
	MHJ2	174	102	112	62	39	11
Corydalis saxicola	YHL1	171	96	132	82	23	27
	YHL2	170	96	133	83	26	24

Abbreviations: F, Forward repeats; P, palindromic repeats; T, tandem repeats.

genomes. A total of 172 SSRs were identified in the whole *C. tomentella* cp genome (take MHJ1 as an example), including 100 mono-, 34 di-, and one compound nucleotide SSRs. Among all SSR types, A and T were the most commonly used bases and 116 SSRs in the *C. tomentella* cp genome had A, T, or AT repeat units (Table 3 and Table S3). For *C. saxicola*, 170 SSRs (take YHL2 as an example) were categorized as 96 mono-, 36 di-, six tri- and six compound nucleotide SSRs, including 115 SSRs with A, T, or AT repeat units (Table 3 and Table S3).

In addition to SSRs, forward repeats (F) and palindromic repeats (P) are also called interspersed repeat sequences (length ≥ 30 bp). In the *C. tomentella* cp genome, there were 112 interspersed repeat sequences, comprised of 64 tandem repeats, 39 forward repeats, and

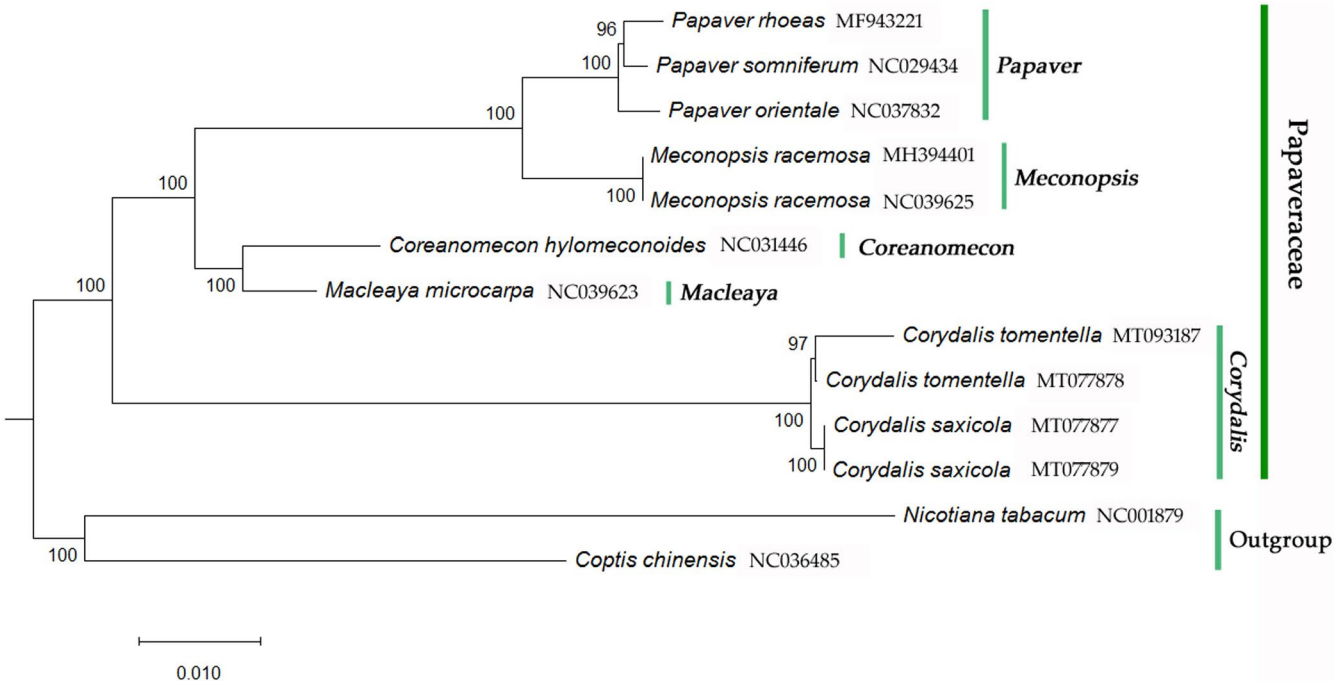

FIGURE 6 ML tree of *C. saxicola* and *C. tomentella* and its relative species based on common protein coding sequences

11 palindromic repeats (Table 3). A total of 132 long repeats were present in *C. saxicola* cp genome, comprised of 82 tandem repeats, 23 forward repeats, and 27 palindromic repeats (Table 3). Comparing the cp genomes of the two species, the *C. saxicola* genome had a greater total number of repeats than the *C. tomentella* cp genome, and the cp genome repeat content in both species was significantly higher than that of most species.

3.4 | Phylogenetic analysis

With *C. chinensis* and *N. tabacum* as outgroups, 70 common protein coding sequences from 13 cp genome sequences were extracted from *C. saxicola*, *C. tomentella*, and six Papaveraceae species to build a Maximum Likelihood (ML) phylogenetic tree (Figure 6). The ML tree has high bootstrap values at each node, indicating a highly credible tree. In this ML tree, the Papaveraceae family is monophyletic, and all samples from Papaveraceae are clustered in a clade. In Papaveraceae, the samples from the genus *Papaver* (*P. somniferum*, *P. orientale*, and *P. rhoeas*) are clustered in a clade; the samples from *Corydalis* (*C. saxicola* and *C. tomentella*) are clustered in a clade; the samples from *Meconopsis* (*M. racemosa*) are clustered in a clade; and *C. hymenoides* and *M. microcarpa* are clustered in a clade. Except for *Coreanomecon* and *Macleaya*, which had only one sample, species in the same genus are clustered into one branch, consistent with previous classification of Papaveraceae genera. At the species level, the *C. saxicola* and *C. tomentella* samples are clustered into separate branches, indicating that the cp genome clustering analysis could effectively distinguish them, while these two closely related species were not monophyletic in the phylogenetic analysis based on short sequence DNA barcodes (Ren et al., 2019). At the same time, *C. saxicola* and *C. tomentella* are clustered in a clade in the ML phylogenetic tree that is distant from other Papaveraceae genera. It shows that *C. saxicola* and *C. tomentella*, both from Sect. *Thalictrifoliae* in *Corydalis*, have a close genetic relationship.

4 | DISCUSSION

4.1 | High variability of genome size and the expansion of IRs

Corydalis saxicola and *C. tomentella* cp genomes are the large cp genomes due to the expansion of IR regions. Most angiosperms cp genomes are highly conserved, typically 115–165 kb in size and possessing a quadripartite structure with two IR regions (IRa and IRb) separating the LSC region and the SSC region (Xin et al., 2019). The sizes of *C. saxicola* and *C. tomentella* cp genomes are larger than those of most flowering plants, such as *N. tabacum* (Sajjad et al., 2016; Shinozaki et al., 1986; Yukawa et al., 2006), 30–40 kb larger than those reported genomes in Papaveriaceae, such as *P. somniferum* (Sun et al., 2016) and *C. hymenoides* (Kim & Kim, 2016). Distinctions between different cp genomes mainly result from the variability of the length and direction of IR regions (Duan et al., 2020). In terms of length, IR regions of the genus *Taxodium* (*T. distichum*, *T. mucronatum* and *T. ascendens*) contracted to about 282 bp (Saski et al., 2005), while IR regions were entirely absent in *Pisum sativum* and *Cryptomeria japonica* (Hirao et al., 2008; Ki & Hae, 2005). In contrast, the length of *Pelargonium hortorum* IR regions expanded to 76 kb (Duan et al., 2020). Numerous studies have shown that IR region lengths

are the main factor influencing cp genome size (Yan et al., 2017). In our study, IR region lengths for the two newly sequenced species were 41,955 to 42,350 bp, which significantly increased their cp genome sizes over that of other Papaveraceae species. Genes normally located in the SSC region, such as *ndh*C, *ndh*D, *ndh*E, *ndh*F, *ndh*G, *ndh*L, *rpl*32, and *trn*L-UAG, have moved to IR regions, contributing to the expanded size of *C. saxicola* and *C. tomentella* IRs.

4.2 | Gene inversions, duplications, and deletions

Inversions usually serve as useful phylogenetic markers (Cosner et al., 2004; Kim et al., 2005). An up to 9 kb inversion containing five genes (*rpl*23, *ycf*2, *ycf*15, *trn*L-CAU, and *trn*L-CAA) was found in the IR regions of *C. tomentella* and *C. saxicola* cp genomes. Relatively large inversions have been found in the cp genomes of some other flowering plants. The 22.8 kb inversion is present in all Asteraceae, except *Barnadesioideae* (Jansen & Palmer, 1987; Martin et al., 2014), the 36 and 78 kb inversions have been detected in core genistoid legumes and Fabaceae subtribe *Phaseolinae*, respectively (Bruneau & Palmer, 1990; Jansen, 2011). These distinctive inversions serve as phylogenetic markers. The inversion in *C. tomentella* and *C. saxicola* is quite distinct from other sequenced Papaveraceae species. To determine if it can be used as a phylogenetic marker of genus *Corydalis*, more species will need to be sequenced. In some plants, the large inversions have been found to be associated with short inverted repeats in cp genome (Joachim et al., 2017; Yi et al., 2013). In Geraniaceae, Campanulaceae and some Fabaceae species, a mass of short inverted repeats have been found to be present at their inversion endpoints (Cosner et al., 2004; Yan et al., 2017). However, we didn't detect large numbers of short inverted repeats emerged in inversion endpoints in *C. tomentella* and *C. saxicola*.

Several NDH (NADH dehydrogenase-like) genes (*ndh*D, *ndh*E, *ndh*F, *ndh*G, *ndh*L) are duplicated in the *C. tomentella* and *C. saxicola* cp genomes, which could provide an explanation for their robust adaptability to harsh environments. Large-scale duplication of cp genes tends to occur only in highly rearranged genomes and can be explained by repeated expansion and contraction of IRs (Mercedes & Bartolomé, 2010; Ruhlman et al., 2015). In this study, genes that are normally located in the SSC region (*ndh*D, *ndh*E, *ndh*F, *ndh*G, *ndh*L, *psa*C, *rpl*32, *ccs*A, and *trn*L-UAG) have migrated to IRs resulting in IR expansion and gene duplication. We found that most of these duplicated genes belong to the NDH complex. Because plastid NDH genes are dispensable under optimal growth conditions, they have been lost in a number of autotrophic and heterotrophic lineages, although they are widely retained across land plants (Ruhlman et al., 2015; Yan et al., 2017). For example, plastid NDH genes have been partially lost or pseudogenized in parasitic plants, such as several orchids and *Petrosavia* (Petrosaviaceae), and autotrophs plants, such as *Najas* (Hydrocharitaceae) and *Erodium* (Geraniaceae) (Mercedes & Bartolomé, 2010), even they have been completely lost in *Selaginella tamariscina* (Xu et al., 2018). Conversely, it is rare for NDH genes to undergo large-scale duplication and augmentation, and the effects of the increased genes resulting from gene duplication on plant growth and development have rarely been discussed in previous research. The NDH complex participates in photosystem I (PSI) cyclic electron flow (CEF), chlororespiration. NDH-dependent CEF provides additional pH change and ATP for CO_2 assimilation and alleviates oxidative stress caused by stromal over-reduction under stress conditions (Ruhlman et al., 2015). The nonphotochemical quenching ability of NDH deficient mutants decreased under mild drought (Sergi et al., 2005). NDH deficient mutants grow slowly at low humidity (Horvath, 2000). Under strong light, tobacco *ndh*B mutants were more susceptible to photobleaching (Sergi et al., 2005). Under heat stress conditions, NDH-mediated cyclic and chlororespiratory electron transport are accelerated, mitigating photo-oxidative damage, and inhibition of CO2 assimilation caused by high temperature (Ju et al., 2003). *Corydalis tomentella* and *C. saxicola* mainly grow in dry cracks of limestone, a unique environment with little available soil and water (Ren et al., 2019) (Figure 1). So they have long been subjected to extreme environmental conditions, such as high temperature, drought, and low light. In view of NDH gene functions in plant defense against various environmental stresses, the doubling of NDH genes those results from IR expansion could lead to overexpression of these doubled genes, which would be helpful for adaptation to harsh environmental conditions. The special structure of the *C. tomentella* and *C. saxicola* cp genomes provides a clue that could explain their robust adaptation to harsh environments.

The *acc*D gene was absent in *C. saxicola* and *C. tomentella* cp genomes. Usually, gene content is highly conserved among photosynthetic angiosperm cp genomes (Jansen et al., 2007; Yan et al., 2017), but in a very few plants, for example, legumes and Circaeasteraceae (Magee et al., 2010; Xu et al., 2018), a number of genes have been lost or pseudogenized. The loss of *acc*D in the cp genome is mirrored in other plant taxa, such as grasses, Circaeasteraceae, and Oleaceae (Joachim et al., 2017; Yan et al., 2017). The *acc*D gene encodes an acetyl-CoA carboxylase subunit and is an important regulator of carbon flow entering the fatty acid biosynthesis pathway (Rousseau-Gueutin et al., 2013). It is known to be essential for leaf development in angiosperms (Hong et al., 2017; Kode et al., 2005). Recent research has shown that the *acc*D gene present in the plastome of most angiosperms is functional (Hong et al., 2017; Rousseau-Gueutin et al., 2013). Furthermore, several studies have shown that the *acc*D gene has been transferred into the nucleus, and the proteins it encodes are transported from the nucleus to the chloroplast to function in the form of a transfer peptide (Joachim et al., 2017; Liu et al., 2016). Whether the *C. tomentella* and *C. saxicola acc*D genes have been lost or transferred to the nucleus, the effects on development are currently unknown.

4.3 | Potential application of cp genome in phylogenetic research of *Corydalis* and Papaveraceae

By exhibiting high species identification power that accurately distinguished two closely related species (*C. saxicola* and *C. tomentella*),

cp genomes have demonstrated a great potential for use as a super-barcode to discriminate *Corydalis* species. *Corydalis*, is considered to be one of the most taxonomically complex taxa (Wu et al., 1996). It is extremely difficult to depend on morphological characteristics for *Corydalis* species identification. Single-locus DNA barcodes lack adequate variation in closely related taxa. Researches using short sequence gene fragments and DNA barcodes showed that both nuclear genome (ITS/ITS2) sequence and cp genome (*mat*K/*rbc*L/*rps*16) sequence produced unsatisfactory taxonomic identifications within *Corydalis* (Ren et al., 2019; Wang, 2006). Cp genomes, exhibiting many advantages, including a moderate size and an appropriate frequency of nucleotide substitutions that can provide sufficient mutation sites (Yan et al., 2017), have been successfully used in the identification of various taxa, such as genera *Epimedium* (Guo et al., 2019), *Fritillaria* (Yan et al., 2018), *Epipremnum* (Tian et al., 2018), and *Papaver* (Zhou et al., 2017). In this study, *C. tomentella* and *C. saxicola*, two closely related species from Sect. *Thalictrifoliae* in *Corydalis*, are clustered into two branches in the phylogenetic tree, which indicates they could be accurately distinguished by cp genome analysis. While, in the phylogenetic analysis based on short sequences of DNA barcodes, these two related species were not monophyletic and couldn't be effectively distinguished. Recent barcoding studies have placed a greater emphasis on the use of whole-cp genome sequences, which are now more readily available as a consequence of improving sequencing technologies (Li et al., 2015). The demonstrated use of cp genomics in *Corydalis* species identification suggests that it has a great potential for taxonomic identification of this genus.

The cp genome also efficiently identified every genus of Papaveraceae in this study. The evolution rates of coding and non-coding regions are significantly different in cp genomes, enabling cp genome use for systematic analysis of different phylogenetic ranks (Clegg et al., 1994). The genus *Corydalis* belongs to Papaveraceae Fumarioideae (Corydaleae) and the phylogenetic relationships of this genus remain controversial (Wu et al., 1996). Recent studies have tended to treat the genus *Corydalis* as an independent Fumariaceae family because the morphological characteristics of this genus constitute a unique evolutionary series (Pérez-Gutiérrez et al., 2012; Wu & Lu, 2003; Zhang et al., 2008). In this study, a Papaveraceae phylogenetic tree, built using common protein CDS, shows that every genus is clustered into one separate clade. However, the clade of *Corydalis* is far from the other genera of Papaveraceae. Combined with the substantial differences in cp genome structures between *Corydalis* and the other Papaveraceae genera, it will be necessary to analyze more representative species to reveal the phylogenetic relationship of *Corydalis*.

ACKNOWLEDGMENTS

We are grateful to Jianguo Zhou and Yuanyao Xin for support and troubleshooting. This work was supported by National Natural Science Foundation of China (81874339); National Science and Technology Major Project for "Significant New Drugs Development" (2019ZX09201005-006-003); Science and Technology Project of Traditional Chinese Medicine in Chongqing (ZY201702143) and 2019 Basic Scientific Research Project of Chongqing (19KF10-2012). The funders did not play any roles in the design of the study, collection, analysis and interpretation of the relevant data, and writing the manuscript.

CONFLICT OF INTEREST

None declared.

AUTHOR CONTRIBUTION

Fengming Ren: Conceptualization (equal); Writing-original draft (equal). **Liqiang Wang:** Data curation (equal); Software (equal); Writing-review & editing (equal). **Ying Li:** Formal analysis (equal); Software (equal). **Wei Zhuo:** Formal analysis (equal); Writing-original draft (equal). **Zhichao Xu:** Formal analysis (equal); Software (equal). **Haojie Guo:** Software (equal). **Yan Liu:** Writing-review & editing (equal). **Ranran Gao:** Validation (equal). **Jingyuan Song:** Conceptualization (equal); Funding acquisition (equal); Writing-review & editing (equal).

DATA AVAILABILITY STATEMENT

The DNA sequences reported in this study have been deposited in the National Center for Biotechnology Information (NCBI) genome database, and Genbank accessions: MT093187, MT07787-MT07789. All sequences used in phylogenetic analysis of Papaveraceae are available from NCBI (Accession numbers: see the "Phylogenetic analysis of Papaveraceae" in Section 3).

ORCID

Wei Zhuo https://orcid.org/0000-0003-4579-4433

REFERENCES

Ahlert, D., Ruf, S., & Bock, R. (2003). Plastid protein synthesis is required for plant development in tobacco. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(26), 15730–15735. https://doi.org/10.1073/pnas.2533668100

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, *27*(4), 578–579. https://doi.org/10.1093/bioinformatics/btq683

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bruneau, A., Doyle, J. J., & Palmer, J. D. (1990). A chloroplast DNA inversion as a subtribal character in the Phaseoleae (Leguminosae). *Systematic Botany*, *15*(3), 378–386. https://doi.org/10.2307/2419351

Chen, Z. D., Lu, A. M., Zhang, S. Z., Wang, Q. F., Liu, Z. J., Li, D. Z., Ma, H., Li, J., Soltis, D. E., Soltis, P. S., & Wen, J. (2016). The tree of life: China project. *Journal of Sytematics and Evolution*, *54*, 273–276. https://doi.org/10.1111/jse.12215

Chumley, T. W., Palmer, J. D., Mower, J. P., Fourcade, H. M., Calie, P. J., & Boore, J. L. (2006). The complete chloroplast genome sequence of *Pelargonium x hortorum*: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Molecular Biology and Evolution*, *23*(11), 2175–2190. https://doi.org/10.1093/molbev/msl089

Clegg, M. T., Gaut, B. S., Learn, G. H., & Morton, B. R. L. (1994). Rates and patterns of chloroplast DNA evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(15), 6795–6801. https://doi.org/10.1073/pnas.91.15.6795

Cosner, M. E., Raubeson, L. A., & Jansen, R. K. (2004). Chloroplast DNA rearrangements in campanulaceae: Phylogenetic utility of highly rearranged genomes. *BMC Evolutionary Biology*, *4*(1), 27. https://doi.org/10.1186/1471-2148-4-27

Cui, Y., Chen, X., Nie, L., Sun, W., Hu, H., Lin, Y., Li, H., Zheng, X., Song, J., & Yao, H. (2019). Comparison and phylogenetic analysis of chloroplast genomes of three medicinal and edible *Amomum* species. *International Journal of Molecular Sciences*, *20*(16), 4040. https://doi.org/10.3390/ijms20164040

Duan, H., Guo, J., Xuan, L., Wang, Z., Li, M., Yin, Y., & Yang, Y. (2020). Comparative chloroplast genomics of the genus *Taxodium*. *BMC Genomics*, *21*(1), 114. https://doi.org/10.1186/s12864-020-6532-1

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., & Dubchak, I. (2004). VISTA: Computational tools for comparative genomics. *Nucleic Acids Research*, *32*, W273. https://doi.org/10.1093/nar/gkh458

Guo, H., Liu, J., Luo, L. I., Wei, X., Zhang, J., Qi, Y., Zhang, B., Liu, H., & Xiao, P. (2017). Complete chloroplast genome sequences of *Schisandra chinensis*: Genome structure, comparative analysis, and phylogenetic relationship of basal angiosperms. *Science China-Life Sciences*, *60*, 1286–1290. https://doi.org/10.1007/s11427-017-9098-5

Guo, M., Ren, L., Xu, Y., Liao, B., Song, J., Li, Y., Mantri, N., Guo, B., Chen, S., & Pang, X. (2019). Development of plastid genomic resources for discrimination and classification of *Epimedium wushanense* (Berberidaceae). *International Journal of Molecular Sciences*, *20*(16), 4003. https://doi.org/10.3390/ijms20164003

Hirao, T., Watanabe, A., Kurita, M., Kondo, T., & Takata, K. (2008). Complete nucleotide sequence of *Thecryptomeria japonica* D. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. *BMC Plant Biology*, *8*(1), 70. https://doi.org/10.1186/1471-2229-8-70

Hong, C. P., Park, J., Lee, Y. I., Lee, M., Park, S. G., Uhm, Y., Lee, J., & Kim, C.-K. (2017). *Acc*D nuclear transfer of *Platycodon grandiflorum* and the plastid of early Campanulaceae. *BMC Genomics*, *18*(1), 607. https://doi.org/10.1186/s12864-017-4014-x

Horvath, E. M. (2000). Targeted inactivation of the plastid *ndh*B gene in tobacco results in an enhanced sensitivity of photosynthesis to moderate stomatal closure. *Plant Physiology*, *123*(4), 1337–1350.

Huang, J., Yu, Y., & Liu, Y. M. (2020). Comparative chloroplast genomics of *Fritillaria* (Liliaceae), inferences for phylogenetic relationships between *Fritillaria* and *Lilium* and plastome evolution. *Plants*, *9*, 133. https://doi.org/10.21203/rs.2.14263/v1

Jansen, R. K. (2011). Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: Rearrangements, repeats, and codon usage. *Molecular Biology and Evolution*, *28*(1), 583–600. https://doi.org/10.1093/molbev/msr037

Jansen, R. K., Cai, Z., Raubeson, L. A., Daniell, H., DePamphilis, C. W., Leebens-Mack, J., Muller, K. F., Guisinger-Bellian, M., Haberle, R. C., Hansen, A. K., Chumley, T. W., Lee, S.-B., Peery, R., McNeal, J. R., Kuehl, J. V., & Boore, J. L. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(49), 19369–19374. https://doi.org/10.1073/pnas.0709121104

Jansen, R. K., Kaittanis, C., Saski, C., Lee, S. B., Tomkins, J., Alverson, A. J., & Henry, D. (2006). Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: Effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Molecular Biology*, *6*(1), 32. https://doi.org/10.1186/1471-2148-6-32

Jansen, R. K., & Palmer, J. D. (1987). A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). *Proceedings of the National Academy of Sciences of the United States of America*, *84*(16), 5818–5822. https://doi.org/10.1073/pnas.84.16.5818

Joachim, R., Wicke, S., Weinl, S., Jörg, K., & Kai, F. M. (2017). Genus-wide screening reveals four distinct types of structural plastid genome organization in *Pelargonium* (Geraniaceae). *Genome Biology and Evolution*, *9*(1), 64–76. https://doi.org/10.1093/gbe/evw271

Ju, Y. Z., Yu, Y. J., & Hua, M. I. (2003). Stimulation of activity of chloroplast NADPH dehydrogenase complex by elevated temperature in tobacco. *Acta Photophysiologica Sinica*, *29*(5), 395–400.

Kanwal, N., Zhang, X., Afzal, N., Yang, J., Li, Z., & Zhao, G. (2019). Complete chloroplast genome of a Chinese endemic species *Corydalis trisecta* Franch. (Papaveraceae). *Mitochondrial DNA Part B-Resources*, *4*(2), 2291–2292. https://doi.org/10.1080/23802359.2019.1627930

Ki, J. K., & Hae, L. L. (2005). Complete chloroplast genome sequences from Korean Ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Research*, *11*(4), 247–261. https://doi.org/10.1093/dnares/11.4.247

Kim, H. W., & Kim, K. J. (2016). Complete plastid genome sequences of *Coreanomecon hylomeconoides* Nakai (Papaveraceae), a Korea endemic genus. *Mitochondrial DNA Part B-Resources*, *1*(1), 601–602. https://doi.org/10.1080/23802359.2016.1209089

Kim, K.-J., Choi, K.-S., & Jansen, R. K. (2005). Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). *Molecular Biology and Evolution*, *22*(9), 1783–1792. https://doi.org/10.1093/molbev/msi174

Kode, V., Mudd, E. A., Iamtham, S., & Day, A. (2005). The tobacco plastid *acc*D gene is essential and is required for leaf development. *Plant Journal*, *44*(2), 237–244. https://doi.org/10.1111/j.1365-313X.2005.02533.x

Lee, E., Harris, N., Gibson, M., Chetty, R., & Lewis, S. (2009). Apollo: A community resource for genome annotation editing. *Bioinformatics*, *25*(14), 1836–1837. https://doi.org/10.1093/bioinformatics/btp314

Lee, K. J., Raveendar, S., & Choi, J. (2019). Development of chloroplast microsatellite markers for identification of *Glycyrrhiza* species. *Plant Genetic Resources*, *17*, 95–99. https://doi.org/10.1017/S1479262118000308

Li, Q. S., Li, Y., Song, J. Y., Xu, H. B., Xu, J., Zhu, Y. J., Li, X., Gao, H., Dong, L., Qian, J., Sun, C., & Chen, S. (2014). High-accuracy de novo assembly and SNP detection of chloroplast genomes using a SMRT circular consensus sequencing strategy. *New Phytologist*, *204*, 1041–1049. https://doi.org/10.1111/nph.12966

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., & Chen, S. (2015). Plant DNA barcoding: From gene to genome. *Niological Reviews*, *90*, 157–166. https://doi.org/10.1111/brv.12104

Liu, T. J., Zhang, C. Y., Yan, H. F., Zhang, L., Ge, X. J., & Hao, J. (2016). Complete plastid genome sequence of *Primula sinensis* (Primulaceae): Structure comparison, sequence variation and evidence for *acc*D transfer to nucleus. *Peer J*, *4*, 2101. https://doi.org/10.7717/peerj.2101

Lu, J., Mei, H. L., Feng, X. Z., Shan, S. C., Liang, P. Z., Tao, X., Huasheng, P., & Wei, Z. (2018). Molecular identification and taxonomic implication of herbal species in genus *Corydalis* (Papaveraceae). *Molecules*, *23*(6), 1393. https://doi.org/10.3390/molecules23061393

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q. I., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B. O., Lu, Y., Han, C., … Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(1), 18. https://doi.org/10.1186/2047-217X-1-18

Magee, A. M., Aspinall, S., Rice, D. W., Cusack, B. P., Semon, M., Perry, A. S., Stefanovic, S., Milbourne, D., Barth, S., Palmer, J. D., Gray, J. C., Kavanagh, T. A., & Wolfe, K. H. (2010). Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Research*, *20*(12), 1700–1710. https://doi.org/10.1101/gr.111955.110

Magnus, L., Fukuhara, T., & Axberg, T. (1996). Phylogeny of *corydalis*, ITS and morphology. *Plant Systematics and Evolution*, *9*, 183–188. https://doi.org/10.1007/978-3-7091-6612-3_17

Marc, L., Oliver, D., Sabine, K., & Ralph, B. (2013). Organellar genome DRAW-a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research*, *41*(W1), W575–W581. https://doi.org/10.1093/nar/gkt289

Martin, G. E., Rousseau-Gueutin, M., Cordonnier, S., Lima, O., Michon-Coudouel, S., Naquin, D., de Carvalho, J. F., Aïnouche, M., Salmon, A., & Aïnouche, A. (2014). The first complete chloroplast genome of the Genistoid legume *Lupinus luteus*: Evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. *Annals of Botany*, *113*(7), 1197–1210. https://doi.org/10.1093/aob/mcu050

Mercedes, M., & Bartolomé, S. (2010). Plastid *ndh* genes in plant evolution. *Plant Physiology and Biochemistry*, *48*(8), 636–645. https://doi.org/10.1016/j.plaphy.2010.04.009

MISA-Microsatellite Identification Tool (2017). Retrieved from http://pgrc.ipk-gatersleben.de/misa/

Moore, M. J., Soltis, S., Bell, C. D., Burleigh, J. G., & Soltis, D. E. (2010). Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(10), 4623–4628. https://doi.org/10.1073/pnas.0907801107

Pérez-Gutiérrez, M. A., Romero-García, A. T., Salinas, M. J., Blanca, G., Fernández, M. C., & Suárez-Santiago, V. N. (2012). Phylogeny of the tribe Fumarieae (Papaveraceae S.L.) based on chloroplast and nuclear DNA sequences: Evolutionary and biogeographic implications. *American Journal of Botany*, *99*(3), 517–528. https://doi.org/10.3732/ajb.1100374

Ren, F. M., Wang, Y. W., Xu, Z. C., Li, Y., Xin, T. Y., Zhou, J. G., Qi, Y.-D., Wei, X.-P., Yao, H., & Song, J. Y. (2019). DNA barcoding of *Corydalis*, the most taxonomically complicated genus of Papaveraceae. *Ecology and Evolution*, *9*(4), 1934–1945. https://doi.org/10.1002/ece3.4886

Rousseau-Gueutin, M., Huang, X., Higginson, E., Ayliffe, M., & Timmis, D. J. (2013). Potential functional replacement of the plastidic acetyl-coa carboxylase subunit (*acc*D) gene by recent transfers to the nucleus in some angiosperm lineages. *Plant Physiology*, *161*(4), 1918–1929. https://doi.org/10.1104/pp.113.214528

Ruhlman, T. A., Chang, W.-J., Chen, J. J. W., Huang, Y.-T., Chan, M.-T., Zhang, J., Liao, D.-C., Blazier, J. C., Jin, X., Shih, M.-C., Jansen, R. K., & Lin, C.-S. (2015). NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. *BMC Plant Biology*, *15*(1), 100. https://doi.org/10.1186/s12870-015-0484-7

Sajjad, A., Khan, A. L., Khan, A. R., Muhammad, W., Sang-Mo, K., & Khan, M. A. (2016). Complete chloroplast genome of *Nicotiana otophora* and its comparison with related species. *Frontiers in Plant Science*, *7*(14), 843. https://doi.org/10.3389/fpls.2016.00843

Saski, C., Lee, S. B., Daniell, H., Wood, T. C., Tomkins, J., Kim, H. G., & Jansen, R. K. (2005). Complete chloroplast genome sequence of glycine max and comparative analyses with other legume genomes. *Plant Molecular Biology*, *59*(2), 309–322. https://doi.org/10.1007/s11103-005-8882-0

Sergi, M. B., Shikanai, T., & Asada, K. (2005). Enhanced ferredoxin-dependent cyclic electron flow around photosystem I and α-tocopherol quinone accumulation in water-stressed ndhb-inactivated tobacco mutants. *Planta*, *222*(3), 502–511. https://doi.org/10.1007/s11103-005-8882-0

Sharp, P. M., & Li, W. H. (1987). The codon Adaptation Index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, *15*, 1281–1295. https://doi.org/10.1093/nar/15.3.1281

Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X. I., Huang, L., & Liu, C. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Research*, *47*(W1), W65–W73. https://doi.org/10.1093/nar/gkz345

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K., Ohto, C., Torazawa, K., Meng, B. Y., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusuda, J., Takaiwa, F., … Sugiura, M. (1986). The complete nucleotide sequence of the tobacco chloroplast genome: Its gene organization and expression. *Embo Journal*, *5*(9), 2043–2049. https://doi.org/10.1007/BF02669253

Stefan, K., Jomuna, V. C., Enno, O., Chris, S., Jens, S., & Robert, G. (2001). REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*, *29*(22), 4633–4642. https://doi.org/10.1093/nar/29.22.4633

Sun, Y., Moore, M. J., Zhang, S., Soltis, P. S., Soltis, D. E., Zhao, T., Meng, A., Li, X., & Wang, H. C. (2016). Phylogenomic and structural analyses of 18 complete plastomes across all families of early-diverging eudicots, including an angiosperm-wide analysis of IR gene content evolution. *Molecular Phylogenetics and Evolution*, *96*, 93–101. https://doi.org/10.1016/j.ympev.2015.12.006

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*, 2725–2729. https://doi.org/10.1093/molbev/mst197

Tian, N., Han, L., Chen, C., & Wang, Z. (2018). The complete chloroplast genome sequence of *Epipremnum aureum* and its comparative analysis among eight Araceae species. *PLoS One*, *13*(3), e0192956. https://doi.org/10.1371/journal.pone.0192956

Wang, Y. W. (2006). *Systematics of Corydalis DC. (Fumariaceae)*. Dissertation for Doctoral Degree, Institute of Botany, the Chinese Academy of Sciences, Beijing.

Wu, J., Lin, P., Guo, Y., & Liu, M. (2020). The complete chloroplast genome of *Corydalis conspersa*. *Mitochondrial DNA Part B-Resources*, *5*(2), 1977–1978. https://doi.org/10.1080/23802359.2020.1756944

Wu, Z. Y., & Lu, A. M. (2003). Ranunculaceae. In J. F. Z. China (Ed.), *The families and genera of angiosperms in China* (*9*, pp. 392–340). Science Press.

Wu, Z. Y., Xuan, Z., & Yun, S. Z. (1996). The systematic evolution of *Corydalis* in relation to florogenesis and floristic regionalization in the world. *Plant Diversity*, *18*(3), 1–3.

Xin, T. Y., Zhang, Y., Pu, X. D., Gao, R. R., Xu, Z. C., & Song, J. Y. (2019). Trends in herbgenomics. *Science China-Life Sciences*, *62*(3), 288–308. https://doi.org/10.1007/s11427-018-9352-7

Xu, Z., Xin, T., Bartels, D., Li, Y., Gu, W., Yao, H., Liu, S., Yu, H., Pu, X., Zhou, J., Xu, J., Xi, C., Lei, H., Song, J., & Chen, S. (2018). Genome analysis of the ancient tracheophyte, *Selaginella tamariscina*, reveals evolutionary features relevant to the acquisition of desiccation tolerance. *Molecular Plant*, *11*(7), 983–994. https://doi.org/10.1016/j.molp.2018.05.003

Yan, L., Zhi, R. Z., Jun, B. Y., Guang, H. L., & Tzen, Y. C. (2018). Complete chloroplast genome of seven *Fritillaria* species, variable DNA markers identification and phylogenetic relationships within the genus. *PLoS One*, *13*(3), e0194613. https://doi.org/10.1371/journal.pone.0194613

Yan, X. S., Moore, M. J., Nan, L., Adelalu, K. F., Aiping, M., Shuguang, J., Yang, L., Li, J., & Wang, H. C. (2017). Complete plastome sequencing of both living species of Circaeasteraceae (Ranunculales) reveals unusual rearrangements and the loss of the *ndh* gene family. *BMC Genomics*, *18*(1), 592. https://doi.org/10.1186/s12864-017-3956-3

Yi, X., Gao, L., Wang, B., Su, Y. J., & Wang, T. (2013). The complete chloroplast genome sequence of *Cephalotaxus oliveri* (Cephalotaxaceae): Evolutionary comparison of *Cephalotaxus* chloroplast DNAs and insights into the loss of inverted repeat copies in gymnosperms. *Genome Biology and Evolution*, *5*(4), 688–698. https://doi.org/10.1093/gbe/evt042

Ying, X. C., Jian, G. Z., Xin, L. C., Zhi, D. X., Yu, W., Wei, S., Jingyuan, S., & Hui, Y. (2019). Complete chloroplast genome and comparative

analysis of three *Lycium* (Solanaceae) species with medicinal and edible properties. *Gene Reports*, *17*, 100464. https://doi.org/10.1016/j.genrep.2019.100464

Yukawa, M., Tsudzuk, I. T., & Sugiura, M. (2006). The chloroplast genome of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*: Complete sequencing confirms that the *Nicotiana sylvestris* progenitor is the maternal genome donor of Nicotiana tabacum. *Molecular Genetics and Genomics*, *275*(4), 367–373. https://doi.org/10.1007/s00438-005-0092-6

Zhang, M. L., Su, Z., & Magnus, L. (2008). Corydalis. In Z. Y. Wu, & H. R. Peter (Ed.), *Flora of China* (Vol. *7*. pp. 295–428). Science Press and Missouri Botanical Garden Press.

Zhang, S.-D., Jin, J.-J., Chen, S.-Y., Chase, M. W., Soltis, D. E., Li, H.-T., Yang, J.-B., Li, D.-Z., & Yi, T.-S. (2017). Diversification of Rosaceae since the late cretaceous based on plastid phylogenomics. *New Phytologist*, *214*(3), 1355–1367. https://doi.org/10.1111/nph.14461

Zhou, J., Chen, X., Cui, Y., Sun, W., Li, Y., Wang, Y. U., Song, J., & Yao, H. (2017). Molecular structure and phylogenetic analyses of complete chloroplast genomes of two *Aristolochia* medicinal species. *International Journal of Molecular Medicine*, *18*(9), 1839. https://doi.org/10.3390/ijms18091839

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Ren F, Wang L, Li Y, et al. Highly variable chloroplast genome from two endangered Papaveraceae lithophytes *Corydalis tomentella* and *Corydalis saxicola*. *Ecol Evol*. 2021;11:4158–4171. https://doi.org/10.1002/ece3.7312