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

fengming Ren
Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Medicinal Plant Development
Liqiang Wang
Heze University
Ying Li
Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Medicinal Plant Development
Wei Zhuo
Chongqing Insititute of Medicinal Plant Cultivation
Zhichao Xu
Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Medicinal Plant Development
Haojie Guo
Wuhu Institute of Technology
Yan Liu
Chongqing Institute of Medicinal Plant Cultivation
Ranran Gao
Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Medicinal Plant Development
Jingyuan Song (jysong@implad.ac.cn)

Research article

Keywords: Corydalis saxicola, Corydalis tomentella, Papaveraceae, taxonomic study, cp genome

Posted Date: April 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-18411/v1

Version of Record: A version of this preprint was published at Ecology and Evolution on March 19th, 2021. See the published version at https://doi.org/10.1002/ece3.7312.

Abstract

Backgroud: *Corydalis DC.*, the largest genus of Papaveraceae, is recognized as one of the most taxonomically challenging plant taxa. However, no complete chloroplast (cp) genome for this genus has been reported to date.

Results: We sequenced four complete cp genomes of two affinities *Corydalis saxicola* and *C. tomentellav* of the genus *Corydalis*, compared these cp genomes with each other and others from Papaveraceae, and analyzed the phylogenetic relationships based on the sequences of common CDS. The cp genomes are 189,029 to 190,247 bp in length, possessing a quadripartite structure and with two highly expanded inverted repeat (IR) regions (length: 41,955 to 42,350 bp). Comparison between the cp genomes of *C. tomentella*, *C. saxicola* and Papaveraceae species revealed high variability in genome sizes, genome structures, gene content, and gene arrangements. Five NADH dehydrogenase-like genes 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. In addition, the *acc*D gene was found to be absent. The *ycf1* gene has shifted from the IR/SSC border to the SSC region as a single copy. Phylogenetic analysis showed that genus *Corydalis* is quite distantly related to the other genera of Papaveraceae, supporting for recent advocacy to establish a separate Fumariaceae family.

Conclusions: Our results provide a useful resource for classification of this taxonomically complicated genus, and will be valuable for understanding Papaveraceae evolutionary relationships.

Background

Chloroplasts (cp), generally considered to have originated from ancient cyanobacteria, are the main site of photosynthesis and energy conversion in plant cells, containing all necessary enzyme systems for photosynthesis and a complete, highly conserved genome [1–2]. With the development of high-throughput sequencing technology, cp genomics has made rapid progress [3]. 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 [4].

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 [5]. Compared with traditional molecular markers, the cp genomes provide specific advantages for establishing plant phylogenetic relationships and taxonomic research [6]. The length of cp genomes is usually 115–165 kb, a modest size that is easily sequenced, and an appropriate nucleotide substitution frequency has produced sufficient mutagenesis for analysis. Relatively conserved gene sequences allow produce co-linearity among plant groups, and the evolution rates of coding regions and non-coding regions are significantly different to be suited for phylogenetic analysis of different ranks [7]. Taxonomists have widely used cp genomes to study plant phylogenetics and advocated for use of cp genomes as a super DNA barcode for species identification [6].

A large number of cp genome sequences 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 taxa. Rorbert et al. clarified long-confusing phylogenetic relationships in rosids by comparing the whole cp genomes of 28 rosid species [8]. Zhang et al. analyzed plastomes from 130 species of 87 genera and offered new insights into deep phylogenetic

relationships and the diversification history of Rosaceae [9]. On the basis of more abundant information sites, cp genomes have been successfully used as a "super barcode" to identify several taxonomically difficult species. Guo et al. found that the *rpl*32 gene in the *Epimedium wushanense* cp genome was deleted, and they successfully identified 11 *Epimedium* species by studying the evolutionary relationship of *rpl*32 genes [10]. Cui et al. analyzed eight cp genomes from *Amomum* to accurately identify *Amomum villosum* and related species [11]. Analysis of *Lycium barbarum, L. chinense*, and *L. rutenicum* cp genomes showed that these three plants could be successfully identified by cp genomics [12]. With the reduced cost of sequencing and the development of bioinformatics technology, cp genome analysis will be extensively used in future studies of plant systematic relationships and taxonomy.

Corydalis is the largest genus in Papaveraceae [13]. There are more than 400 *Corydalis* species that are widely distributed in the North Temperate Zone [14]. *Corydalis* has extremely complex morphological variation because of typical reticulate evolution and intense differentiation during evolution [13], including extensive interspecific hybridization and gene introgression [13–15]. Taxonomic study of the genus on the basis of morphological characteristics and DNA barcoding has been very difficult and slow, and a complete classification system has not yet been established in *Corydalis* [16]. Consequently, it is considered to be one of the most taxonomically complex taxa. Though cp genomes have been proven effective for phylogenetic research of many taxonomically complex taxa [2, 8–12], there currently are no published reports about cp genomes of genus *Corydalis*.

In this study, high-throughput sequencing and comparative genomics were used to study the cp genomes of two important medicinal plants in genus *Corydalis*: *C. saxicola* and *C. tomentella*. They have a very special habitat (Fig. 1), grow in dry cracks of limestone and are critically endangered. We sequenced four complete cp genome sequences from these two species, described their genomic characteristics, conducted comparisons between these genomes and other Papaveraceae cp genomes, analyzed the phylogenetic relationships on the basis of common CDS in the cp genomes.

Results

Organization and features of tomentella and saxicola genomes

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) (Fig. 1). The GC content of the genome and each genomic region was also typical of the angiosperm cp style. Specific lengths and contents are shown in Fig. 1 and Table 1. The lengths of the two complete *C. saxicola* genomes were 189,029 bp 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).

Table 1

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

Species	Voucher No.	Genbank No.	Total	Length (bp)			GC content (%)			
				IR	LSC	SSC	Total	IR	LSC	SSC
Corydalis tomentella	MHJ1	MT093187	190247	41955	96701	9636	40.3	42.2	39.2	35.4
	MHJ2	MT077878	190198	42002	96530	9664	40.2	42.2	39.0	35.4
Corydalis saxicola	YHL1	MT077877	189155	42350	94744	9711	40.2	42.2	39.1	35.1
	YHL2	MT077879	189029	42164	94993	9708	40.3	42.2	39.1	35.1

CPGAVAS2 was used to annotate the cp genomes of *C. tomentella* and *C. saxicola*. Removing duplicate genes, a total of 119 annotated genes (Fig. 2, Table 2 and 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 *ycf3* 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 *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 (Fig. 2, Table 2 and S1).

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)

*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.

Codon usage bias, SSRs analysis, 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.

SSRs are short tandem repeats of 1–6 bp DNA sequences that are widely distributed throughout the cp genome [17]. 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 genomes. A total of 172 SSRs were identified in the whole *C. tomentella* cp genome (take MHJ1 as an example), *in*cluding 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 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 S3).

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

T: tandem repeats, F: Forward repeats and P: palindromic repeats.

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 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.

IR contraction and expansion

IR 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 [18–19]. We selected four phylogenetically close species (*Papaver rhoeas*, *Papaver orientale*, *Papaver somniferum*, and *Coreanomecon hylomeconoides*) and two model species (*N. tabacum* and *Arabidopsis thaliana*) as references for cp genome structure comparisons. Figure 3 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 bp to 1,200 bp. However, the *ycf*1 genes in *C. tomentella* and *C. saxicola* cp genomes 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 bp to 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.

Comparative genomic analysis and genome sequence divergence

VISTA software was used to make multiple comparisons of the *C. tomentella* and *C. saxicola* cp genome sequences, and results show that intra-specific variation was small but there were still some inter-specific differences (Fig. 4). The coding and non-coding 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 *rps12-clp*P, *clp*P-*psb*B, and *pet*B-*psb*H. Comparing *C. tomentella* and *C. saxicola*, the most highly divergent regions mainly was 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 and difficult to identify accurately. 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 genome structure (Fig. 5). 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 (Fig. 1). A few *rpl*19 and *rpl*2 genes migrated from the IR region to the LSC region. In 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 non-coding regions of *C. tomentella* and *C. saxicola* cp genomes differ greatly from those of other Papaveraceae and tobacco.

Phylogenetic analysis of Papaveraceae

With *C. chinensis* and *N. tabacum* as outgroups, common protein coding sequences from 13 cp genome sequences were extracted from *C. saxicola*, *C. tomentella*, and six Papaveraceae species (*P. somniferum*: NC029434, *P. orientale*: NC037832, *P. rhoeas*: MF943221, *Coreanomecon hymenoides*: NC031446, *Macleaya microcarpa*: NC039623, and *Meconopsis racemosa*: MH394401 NC039625) to build a Maximum Likelihood (ML) phylogenetic tree (Fig. 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* (*Corydalis saxicola* and *Corydalis 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* genera, which had only one sample, species in the same genera 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. 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. On one hand, this shows that *C. saxicola* and *C. tomentella*, both from Sect. Thalictrifoliae in *Corydalis*, have a close genetic relationship. On the other hand, it also shows that *Corydalis* has a relatively distant genetic relationship with the other Papaveraceae genera included in this study.

Discussion

High variability of genome size and the expansion of IRs

C. saxicola and *C. tomentella* cp genomes are among the largest cp genomes due to 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 the LSC region [19]. The sizes of *C. saxicola* and *C. tomentella* cp genomes are larger than those of most flowering plants, such as *N. tabacum* [20–22], 30–40 kb larger than those reported genomes in Papaveriaceae, such as *P. somniferum* [23] and *C. hymenoides* [24]. Distinctions between different cp genomes mainly result from the variability of the length and direction of IR regions [25]. In terms of length, IR regions of the genus *Taxodium* (*T. distichum, T. mucronatum and T. ascendens*) contracted to about 282 bp [26], while IR regions were entirely absent in *Pisum sativum* and *Cryptomeria japonica* [27–28]. In contrast, the length of *Pelargonium hortorum* IR regions expanded to 76 kb [25]. Numerous studies have shown that IR region lengths are the main factor influencing cp genome size [29]. In our study, IR region lengths for the two newly sequenced species were 41,955 bp 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.

Gene inversions, duplications, and deletions

Inversions usually serve as useful phylogenetic markers [30–32]. 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 emerged in some other flowering plants. The 22.8 kb inversion is present in all Asteraceae, except *Barnadesioideae* [32–33], the 36 kb and 78 kb inversions have been detected in core genistoid legumes and Fabaceae subtribe *Phaseolinae*, respectively [34–35]. 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 [26, 30, 36–37]. In Geraniaceae, Campanulaceae and some Fabaceae species, a mass of short inverted repeats have been found to be present at their inversion endpoints [29–30]. 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 [38–39]. 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 of plastid NDH genes are dispensable under optimal growth conditions, which have been lost in a number of autotrophic and heterotrophic lineages, although they are widely retained across land plants [29, 37–39]. 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) [38], even they have been completely lost in *Selaginella tamariscina* [40]. 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 [38–39]. The non-photochemical quenching ability of NDH deficient mutants decreased under mild drought [43]. NDH deficient mutants grow slowly at low humidity [42]. Under strong light, tobacco *ndh*B mutants were more susceptible to photobleaching [43]. 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 [44]. *C. tomentella* and *C. saxicola* mainly grow in dry cracks of limestone, a unique environment with little available soil and water [45] (Fig. 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 [29, 46], but in a very few plants, for example, legumes and Circaeasteraceae [40, 47], 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 [29, 30, 37]. The *acc*D gene encodes an acetyl-CoA carboxylase subunit and is an important regulator of carbon flow entering the fatty acid biosynthesis pathway [48]. It is known to be essential for leaf development in angiosperms [49–50]. Recent research has shown that the *acc*D gene present in the plastome of most angiosperms is functional [48, 50]. 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 [37–48, 50–51]. 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.

Papaveraceae phylogenetic relationships and potential application of cp genomics in Corydalis

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. Genus *Corydalis*, is considered to be one of the most taxonomically complex taxa [13]. 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* [14, 45, 52]. Cp genomes, exhibiting many advantages, including a moderate size and an appropriate frequency of nucleotide substitutions that can provide sufficient mutation sites [29], have been successfully used in the identification of various taxa, such as genera *Epimedium* [10] *Fritillaria* [53], *Epipremnum* [54], and *Papaver* [55]. In this study, *C. saxicola* and *C. tomentella*, two closely related species from *Section Thalictrifoliae* in *Corydalis*, are clustered into two branches in

the phylogenetic tree, which indicates they could be accurately distinguished by cp genome analysis. While, these two affinities were not monophyletic in the phylogenetic analysis based on short sequences of DNA barcodes and couldn't been 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 [3]. 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 genomics also efficiently identified Papaveraceae genera. 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 [7]. Genus *Corydalis* belongs to Papaveraceae Fumarioideae (Corydaleae) and the phylogenetic relationships between this genus and its relatives remain controversial [13]. Recent studies have tended to treat genus *Corydalis* and closely related genera as an independent Fumariaceae family because the morphological characteristics of this family constitute a unique evolutionary series [15–16, 56–58]. In this study, a Papaveriaceae phylogenetic tree, built using common protein coding sequences, shows that each genera is clustered into one branch. However, *Corydalis* species are clustered into a distinct clade that is quite distant from other Papaveriaceae genera. Combined with the substantial differences in cp genome structures between *Corydalis* and other Papaveriaceae genera, these results provide preliminary molecular evidence that supports classifying *Corydalis* as a separate Fumariaceae family. However, because there were few species included in this study, it will be necessary to analyze additional representative species in further studies.

Methods

Materials, DNA extraction and sequencing

Plant materials were provided by the Chongqing Institute of Medicinal Plant Cultivation and identified by researcher Zhengyu Liu as *C. tomentella Franch*. and *C. saxicola Bununting*. 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, and the DNA quality and concentration were detected using 1% agarose electrophoresis and a Nanodrop2000. Qualified total DNA was sent to Shanghai Biotech for sequencing with the Illumina HiSeq4000 platform.

Genome assembly and annotation

Referring to the method of Li et al. [4], cp genomes were assembled on a Linux system using BLAST+, capturereads.py, soapdenovo-127mer, and SSPACE-Basic-v2.0.pl software [4, 55]. The completed genomes were annotated using CPGAVAS2 [59], and the results were modified for starter and terminator revisions by Apollo software [60]. 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 [4]. Physical maps of the cp genomes were drawn by GenomeDRAW [61] using a GB format file exported from the sqn file by sequin software.

Genome structure analyses and genome comparison

GC content was analyzed using Mega 6.0 software [55, 62]. The distribution of codon usage was investigated using CodonW software with the RSCU ratio [55, 63]. MISA software (http://pgrc.ipk

-gatersleben.de/misa/) was used to detect simple sequence repeats (SSRs) [64]. 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 [65]. Tandem Repeats Finder v4.0.4 software [66] was used to detect tandem repeat sequences, and the default parameter was set to 2–7 -7 -80- 10 -50- 500-f-d-m [4]. REPuter software (http://bibiserv.techfak.uni-bielefeld.de/reputer) was used to detect scattered repeating sequences (> 30 bp) using the parameter: hamming distance = 3 [4, 67]. VISTA software was used to compare multiple cp genomes [68].

Phylogenetic analysis

A total of 15 cp whole genome sequences were used in cluster analysis. Thirteen genomes were from Papaveraceae (four *Corydalis* genomes, five *Papaver* genomes, two *Meconopsis* genomes, one *Macleaya*, and one *Coreanomecon* genome), and *Coptis chinensis* and *Nicotiana tabacum* 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 [4], and multiple global alignments of the protein coding sequences was performed using the Clustalw module in MEGA6.06 software [62]. The maximum-likelihood tree (ML) was constructed on the basis of the alignments of common protein sequences and full-length cp genome sequences by MEGA6.06 software [62].

Conclusions

In this study, we sequenced and characterized four complete cp genomes for *C. tomentella* and *C. saxicola* from Papaveraceae. Comparisons of these cp genomes indicated high similarity. In comparisons of these cp genomes with those from other Papaveraceae species, *C. tomentella* and *C. saxicola* cp genomes exhibited large variations in genome sizes, genome structures, gene content and gene arrangements. *C. saxicola* and *C. tomentella* have large cp genome sizes of 189,029 bp to 190,247 bp, which are the results of IR expansion to sizes as large as 42,350 bp. Genes (*ndh*F, *ndh*D, *ndh*L, *ndh*G, *ndh*E, *psa*C, *ccs*A, *rpl*32 and *trn*L-UAG genes) normally located in the SSC region have migrated to IRs resulting in IR expansion and gene duplication. Duplication of the NDH genes may contribute to *C. saxicola* and *C. tomentella* adaptation to harsh environments. An up to 9 kb inversion containing five genes (*rpl*23, *ycf*2, *ycf*15, *trn*I-CAU and *trn*L-CAA) was found in IR regions. In addition, *acc*D was found to have been deleted. The *ycf*1 gene has shifted from the IR/SSC border to the SSC region as a single copy. In its place, *ndh*A and *rpl*2 were located at the IR/SSC border and as partial genes (pseudogenes) in other IR regions. Phylogenetic analysis showed that genus *Corydalis* is quite distantly related to the other Papaveraceae genera. Thus, these cp genome characteristics provide preliminary molecular evidence that supports recent suggestion to create a separate Fumariaceae family. Our results provide a useful resource for classification and identification of this taxonomically complicated genus, and will be valuable for understanding Papaveraceae evolutionary relationships.

Abbreviations

cp: chloroplast; IR: inverted repeat; NCBI: National Center for Biotechnology Information; LSC: large single-copy region; SSC: small single-copy region; IR: inverted repeats; RSCU: relative synonymous codon usage; PSI: photosystem I; CEF: cyclic electron flow; T: tandem repeats; F: forward repeats; P: palindromic repeats; ML: Maximum Likelihood; NDH: NADH dehydrogenase-like; ML: maximum-likelihood; SSR: simple sequence repeats.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The collection of plant material was complied with institutional and national guidelines. The field of the study was conducted in accordance with local legislation. Four plastomes sequenced in this study have been deposited in the National Center for Biotechnology Information (NCBI) genome database (https://www.ncbi.nlm.nih.gov/) (Accession numbers: see Table 1). All sequences used in phylogenetic analysis of Papaveraceae are available from NCBI (Accession numbers: see the “Phylogenetic analysis of Papaveraceae” in Result section).

Competing interests

The authors declare no conflict of interest.

Funding

This research was funded 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.

Author’s Contributions

Conceptualization: S.J.-Y. and R.F.-M.; Formal analysis: L.Y.(Li), Z.W. and X.Z-C.; Funding acquisition: S.J.-Y.; Software: W.L.-Q., L.Y.(Li), X.Z-C. and G.H-J.; Validation: G.R.-R; Writing-original draft: R.F.-M. and Z.W.; Writing-review & editing: W.L.-Q., L.Y. and S.J.-Y. All co-authors approved the manuscript before submission. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We are grateful to Jianguo Zhou and Yuanyao Xin for support and troubleshooting.

Authors' information

[1]Key Lab of Chinese Medicine Resources Conservation, State Administration of Traditional Chinese Medicine of the People's Republic of China, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100193, China. [2]Chongqing Institute of Medicinal Plant Cultivation, Research and Utilization on Characteristic Biological Resources of Sichuan and Chongqing Co-construction Lab, Chinese Medicine Breeding and Evaluation Engineering Technology Research Center of Chongqing, Chongqing 400010, China. [3]College of Pharmacy, Heze University, Heze 274015, Shandong Province, China. [4]Wuhu Institute of Technology, Wuhu 241000, Anhui Province, China. [5]Engineering Research Center of Chinese Medicine Resource, Ministry of Education, Beijing 100193, China

References

1. Ahlert D, Ruf S, Bock R. Plastid protein synthesis is required for plant development in tobacco. Proc Natl Acad Sci USA. 2003;100(26):15730–35.
2. Moore MJ, Soltis S, Bell CD, Burleigh JG, Soltis DE. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc Natl Acad Sci USA. 2010;107(10):4623–28.
3. Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S. Plant DNA barcoding: from gene to genome. Biol Rev. 2015;90:157–66.
4. LI QS, Li Y, Song JY, Xu HB, Xu J, Zhu YJ, Li XW, Gao HH, Dong LL, Qian J, Sun C, Chen SL. High-accuracy *de novo* assembly and SNP detection of chloroplast genomes using a SMRT circular consensus sequencing strategy. The New Phytol. 2014;204:1041–49.
5. Chen ZD, Lu AM, Zhang SZ, Wang QF, Liu ZJ, Li DZ, Ma H, Li JH, Soltis DE, Soltis PS, Wen J. The tree of life: China project. J Syt Evol. 201;54:273–76.
6. Guo H, Liu J, Luo L, Wei X, Zhang J, Qi Y, Zhang B, Liu H, Xiao P. Complete chloroplast genome sequences of *Schisandra chinensis*: genome structure, comparative analysis, and phylogenetic relationship of basal angiosperms. Sci China Life Sci. 2017;60:1286–90.
7. Clegg MT, Gaut BS, Learn GH, Morton BRL. Rates and patterns of chloroplast DNA evolution. Proc Natl Acad Sci USA. 1994;91(15):6795–801.
8. Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Henry D. Phylogenetic analyses of Vitis(Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. Bmc Evol Biol. 2006;6(1):32.
9. Zhang SD, Jin JJ, Chen SY, Chase MW, Soltis DE, Li HT, JunBY, De ZL, Ting SY. Diversification of Rosaceae since the late cretaceous based on plastid phylogenomics. New Phytol. 2017;214(3):1355–67.
10. Guo M, Ren L, Xu Y, Liao B, Song J, Li Y, Mantri N, Guo B, Chen S, Pang X. Development of plastid genomic resources for discrimination and classification of *Epimedium wushanense* (Berberidaceae). Int J Mol Sci. 2019;20(16):4003.
11. Cui Y, Chen X, Nie L, Sun W, Hu H, Lin Y, Li H, Zheng X, Song J, Yao H. Comparison and phylogenetic analysis of chloroplast genomes of three medicinal and edible *Amomum* species. Int J Mol Sci. 2019;20(16):4040.
12. Ying XC, Jian GZ, Xin LC, Zhi DX, Yu W, Wei S. JingYS, HuiY. Complete chloroplast genome and comparative analysis of three *Lycium* (Solanaceae) species with medicinal and edible properties. Gene Rep. 2019;17:100464.

13. Wu ZY, Xuan Z, Yun SZ. The systematic evolution of corydalis in relation to florogenesis and floristic regionalization in the world. Acta Botanica Yunnanica. 1996;18(3):241–67.
14. Magnus L, Fukuhara T, Axberg T. Phylogeny of corydalis, ITS and morphology. Plant Syst Evol. 1996;9:183–88.
15. Lidén M, Fukuhara T, Rylander J, Oxelman B. Phylogeny and classification of Fumariaceae, with emphasis on Dicentra s. l., based on the plastid gene *rps*16 intron. Plant Syst Evol. 1997;206:411–20.
16. Lu J, Mei HL, Feng XZ, Shan SC, Liang PZ, Tao X, Hua SP, Wei Z. Molecular identification and taxonomic implication of herbal species in genus *Corydalis* (Papaveraceae). Molecules. 2018;23(6):1393.
17. Lee KJ, Raveendar S, Choi J. Development of chloroplast microsatellite markers for identification of *Glycyrrhiza* species. Plant Genetic Res. 2019;17:95–9.
18. Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL. The complete chloroplast genome sequence of *Pelargonium x hortorum*: or ganization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol Biol Evol. 2006;23(11):2175–90.
19. Xin TY, Zhang Y, Pu XD, Gao RR, Xu ZC, Song JY. Trends in Herbgenomics. Sci China Life Sci. 2019;062(3):288–308.
20. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayshida N, Matsubayasha T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki J, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kata A, Tohdoh N, Shimada H, Sugiura M. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 1986;5(9):2043–49.
21. Sajjad A, Khan AL, Khan AR, Muhammad W, Sang-Mo K, Khan MA. Complete chloroplast genome of *Nicotiana otophora* and its comparison with related species. Front Plant Sci. 2016;7(14):843.
22. Yukawa M, Tsudzuki T, Sugiura M. 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*. Mol Genet Genomics. 2006;275(4):367–73.
23. Sun Y, Moore MJ, Zhang S, Soltis PS, Soltis DE, Zhao TMeng A, Li XJ, Li JQ, Wang HC. 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. Mol Phylogenet Evol. 2016;96:93–101.
24. Kim HW, Kim KJ. Complete plastid genome sequences of *Coreanomecon hylomeconoides* Nakai (Papaveraceae), a korea endemic genus. Mitochondrial DNA Part B. 2016;1(1):601–2.
25. Duan H, Guo J, Xuan L, Wang Z, Li M, Yin Y, Yang Y. Comparative chloroplast genomics of the genus *Taxodium*. BMC Genomics. 2020;21(1):114.
26. Saski C, Lee SB, Daniell H, Wood TC, Tomkins J, Kim HG, Jansen RK. Complete chloroplast genome sequence of glycine max and comparative analyses with other legume genomes. Plant Mol Biol. 2005;59(2):309–22.
27. Hirao T, Watanabe A, Kurita M, Kondo T, Takata K. Complete nucleotide sequence of *Thecryptomeria japonicad*. don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. BMC Plant Biol. 2008;8(1):70.
28. Ki JK, Hae LL. Complete chloroplast genome sequences from Korean Ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res. 2005;11(4):247–61.
29. Yanxia S, Moore MJ, Nan L, Adelalu KF, Aiping M, Shuguang J, YangL, Li JQ, Wang HC. Complete plastome sequencing of both living species of Circaeasteraceae (Ranunculales) reveals unusual rearrangements and the loss of the *ndh* gene family. BMC Genom. 2017;18(1):592.

30. Cosner ME, Raubeson LA, Jansen RK. Chloroplast DNA rearrangements in campanulaceae: phylogenetic utility of highly rearranged genomes. BMC Evol Biol. 2004;4(1):27.
31. Kim KJ. Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). Mol Biol Evol. 2005;22(9):1783–92.
32. Jansen RK, Palmer JD. A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). Proc Natl Acad Sci USA. 1987;84(16):5818–22.
33. Martin GE, Rousseau-Gueutin M, Cordonnier S, Lima O, Michon-Coudouel S, Naquin D, Carvalho JF, Aïnouche M, SalmonA, Aïnouche A. 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. Ann Bot. 2014;113(7):1197–210.
34. Bruneau A, Palmer DJ. A chloroplast DNA inversion as a subtribal character in the Phaseoleae (Leguminosae). Syst Bot. 1990;15(3):378–86.
35. Jansen RK. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol. 2011;28(1):583–600.
36. Yi X, Gao L, Wang B, Su YJ, Wang T. 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 Biol Evol. 2013;5(4):688–98.
37. Joachim R, Wicke S, Weinl S, Jörg K, Kai FM. Genus-wide screening reveals four distinct types of structural plastid genome organization in *Pelargonium* (Geraniaceae). Genome Biol Evol. 2017;9(1):64–76.
38. Mercedes M, Bartolomé S. Plastid ndh genes in plant evolution. Plant Physiol Bioch. 2010;48(8):636–45.
39. Ruhlman TA, Chang WJ, Chen JJ, Huang YT, Chan MT, Zhang J, Liao DC, Blazier JC, Jin XH, Shih MC, Jansen RK, Lin CS. NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. BMC Plant Biol. 2015;15(1):100.
40. Xu ZC, Xin TY, Bartels D, Li Y, Gu W, Yao H, LiuS, Yu HY, Pu XD, Zhou JG, Xu J, Xi CC, Lei HT, Song JY, Chen SL. Genome analysis of the ancient tracheophyte, *Selaginella tamariscina*, reveals evolutionary features relevant to the acquisition of desiccation tolerance. Mol Plant. 2018;11(7):983–94.
41. Burrows PA. Identification of a functional respiratory complex in chloroplasts through analysis of tobacco mutants containing disrupted plastid ndh genes. EMBO J. 1998;17(4):868–76.
42. Horvath EM. Targeted inactivation of the plastid *ndh*B gene in tobacco results in an enhanced sensitivity of photosynthesis to moderate stomatal closure. Plant Physiol. 2000;123(4):1337–50.
43. Sergi MB, Shikanai T, Asada K. Enhanced ferredoxin-dependent cyclic electron flow around photosystem I and α-tocopherol quinone accumulation in water-stressed ndhb-inactivated tobacco mutants. Planta. 2005;222(3):502–11.
44. Ju YZ, Yu YJ, Hua MI. Stimulation of activity of chloroplast NADPH dehydrogenase complex by elevated temperature in tobacco. Acta Photophysiologica Sinica. 2003;29(5):395–400.
45. Ren FM, Wang YW, Xu ZC, Li Y, Xin T Y Zhou JG, Qi YD, Wei XP, Yao H, Song JY. DNA barcoding of *Corydalis*, the most taxonomically complicated genus of Papaveraceae. Ecol Evol. 2019;9(4):1934–45.
46. Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, Chumley TW, Lee SB, Peery R, McNeal JR, Kuehl JV, Boore JL. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA. 2007;104(49):19369–74.

47. Magee AM, Aspinall S, Rice DW, Cusack BP, Semon M, Perry AS, Stefanović S, Milbourne D, Barth S, Palmer JD, Gray JC. Kavanagh TAWolfe KH. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res. 2010;20(12):1700–10.
48. Rousseau-Gueutin M, Huang X, Higginson E, Ayliffe M, Timmis DJ. Potential functional replacement of the plastidic acetyl-coa carboxylase subunit (*acc*D) gene by recent transfers to the nucleus in some angiosperm lineages. Plant Physiol. 2013;161(4):1918–29.
49. Kode V, Mudd EA, Iamtham S, Day A. The tobacco plastid *acc*D gene is essential and is required for leaf development. Plant J. 2005;44(2):237–44.
50. Pyo HC, Jihye P, Yi L, Minjee L, Gi PS, Yurry U, Jungho L, Chang KK. *Acc*D nuclear transfer of *Platycodon grandiflorum* and the plastid of early Campanulaceae. BMC Genom. 2017;18(1):607.
51. Liu TJ, Zhang CY, Yan HF, Zhang L, Ge XJ, Hao J. Complete plastid genome sequence of *Primula sinensis* (Primulaceae): structure comparison, sequence variation and evidence for *acc*D transfer to nucleus. PeerJ. 2016;4:2101.
52. Wang YW. PhD. thesis. Institute of botany, the Chinese academy of sciences; Beijing, China: 2006. Systematics of *Corydalis* DC. (Fumariaceae).
53. Yan L, Zhi RZ, Jun BY, Guang HL, TzenYC. Complete chloroplast genome of seven *Fritillaria* species, variable DNA markers identification and phylogenetic relationships within the genus. Plos One. 2018;13(3):e0194613.
54. Tian N, Han L, Chen C, Wang Z. The complete chloroplast genome sequence of *Epipremnum aureum* and its comparative analysis among eight Araceae species. Plos One. 2018;13(3):e0192956.
55. Zhou JG, Yao H, Chen XL, Li Y, Song JY. Molecular structure and phylogenetic analyses of complete chloroplast genomes of two *Aristolochia* medicinal species. Int J Mol Med. 2017;18(9):1839.
56. Wu ZY, Lu AM. The families and genera of angiosperms in China. Beijing: Science Press; 2003. p. 647.
57. Zhang ML, Su ZY, Magnus L, Corydalis. Flora of China. In: Wu ZY, Peter HR, editors. Beijing: Science Press, China: Missouri Botanical Garden Press, St. Louis, MO, USA: 2008.Volume 7. p. 295–428.
58. Pérez-Gutiérrez MA, Romero-García AT, Salinas MJ, Blanca G, Fernández MC, Suárez-Santiago VN. Phylogeny of the tribe Fumarieae (Papaveraceae S.L.) based on chloroplast and nuclear DNA sequences: evolutionary and biogeographic implications. Am J Bot. 2012;99(3):517–28.
59. Shi LC, Chen HM, Jiang M, Wang LQ, Wu X, Huang LH, Liu C. CPGAVAS2, an integrated plastome sequence annotator and analyzer. Nucleic Acids Res. 2019;47(W1):W65–73.
60. Lee E, Harris N, Gibson M, Chetty R, Lewis S. Apollo: a community resource for genome annotation editing. Bioinformatics. 2009;25(14):1836–37.
61. Marc L, Oliver D, Sabine K, Ralph B. Organellar genome DRAW-a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 2013;41(W1):W575-81.
62. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30:2725–29.
63. Sharp PM, Li WH. The codon Adaptation Index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15:1281–95.
64. MISA-Microsatellite Identification Tool. . Accessed on 21 September 2017.
65. Huang J, Yu Y, Liu YM, Xie. Comparative chloroplast genomics of *Fritillaria* (Liliaceae), inferences for phylogenetic relationships between *Fritillaria* and *Lilium* and plastome evolution. Plants. 2020;9:133.

66. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80.
67. Stefan K, Jomuna VC, Enno O, Chris S, Jens S, Robert G. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29(22):4633–42.
68. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: Computational tools for comparative genomics. Nucleic Acids Res. 2004;32:W273.

Figures

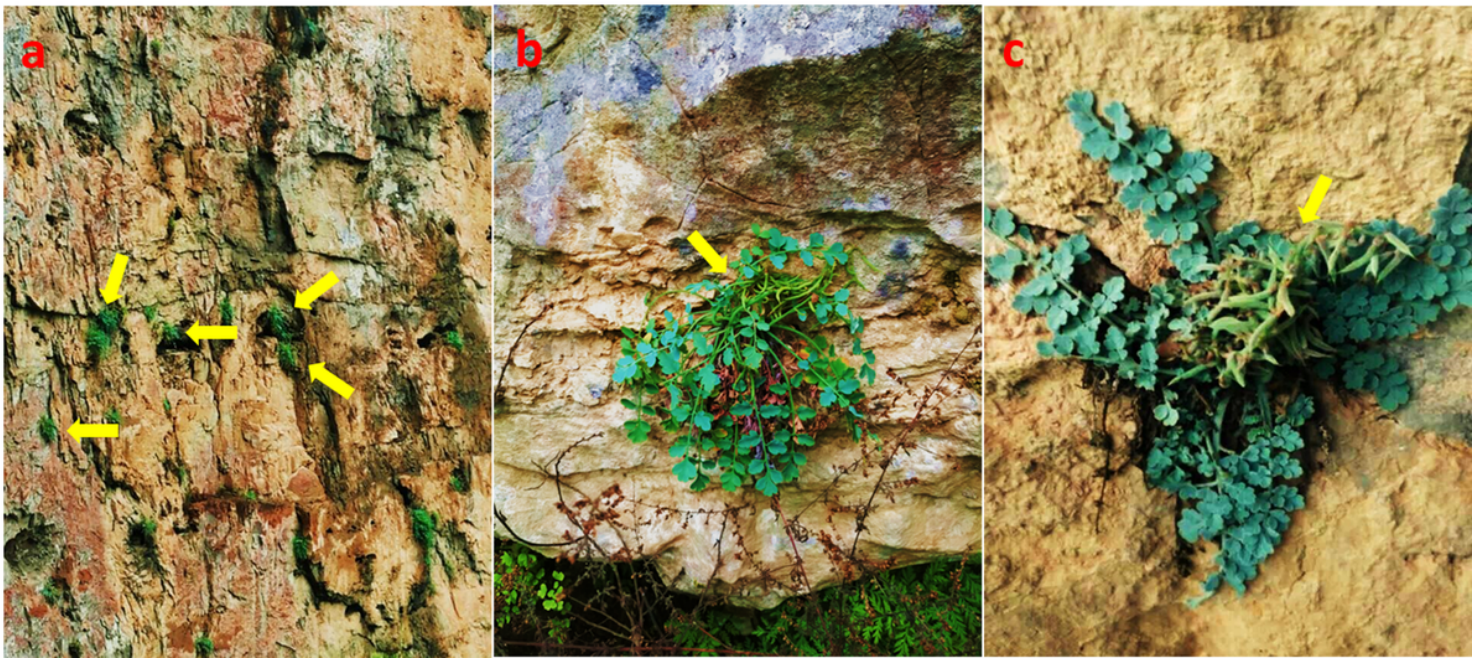

Figure 1

The habitat of C. saxicola and C. tomentella. aThe distant view of steep cliff growing C. saxicola; b the close shot of C. saxicola , and c the close shot of C. tomentella. The yellow arrows indicated the Corydalis plants.

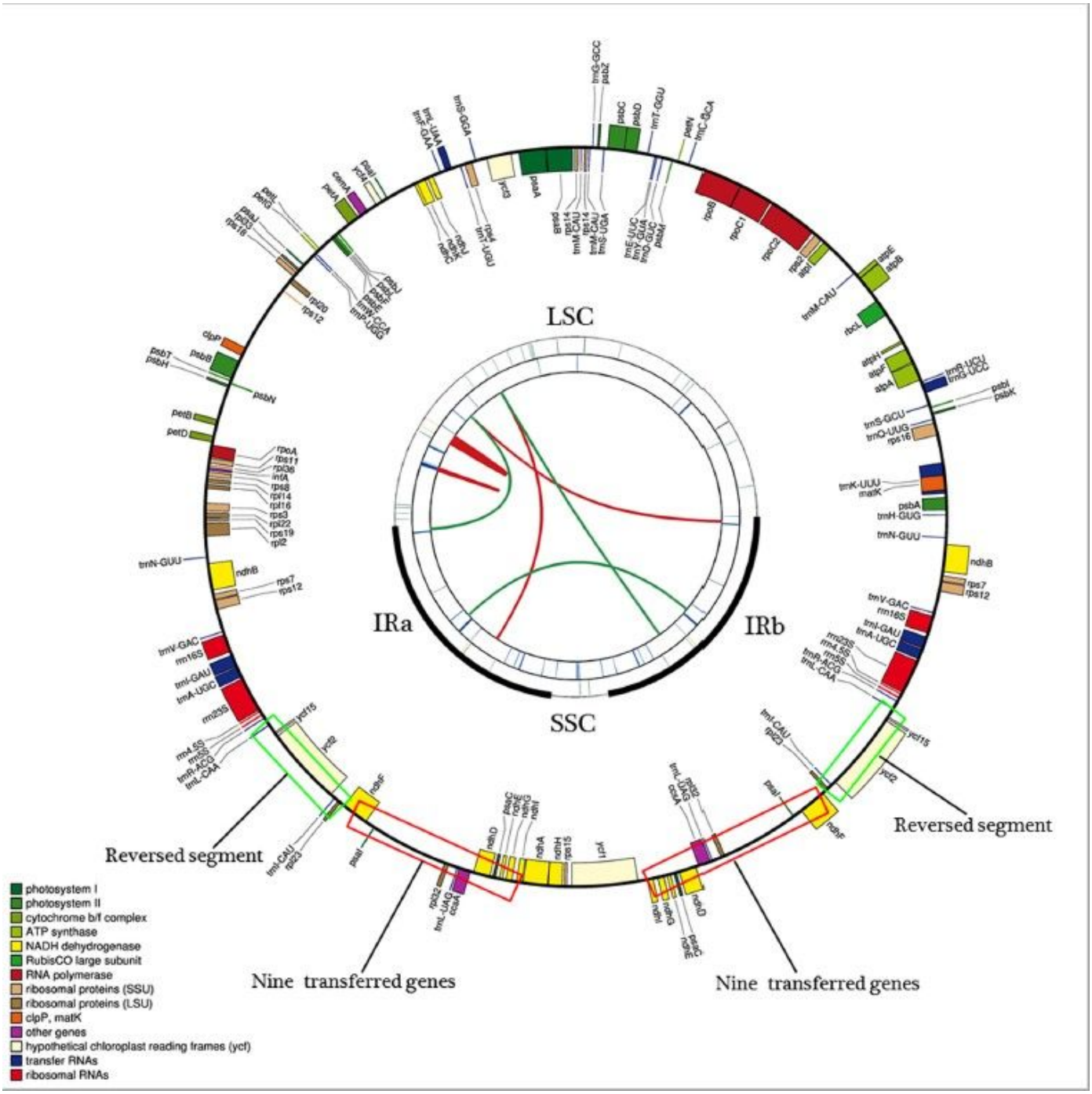

Figure 2

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. LSC: large single copy; SSC: small single copy; IR: inverted repeat. The red rectangles indicated the nine gens(ndhF, ndhD, ndhL, ndhG, ndhE, psaC, ccsA, rpl32 and trnL-UAG) normally located in the SSC region have migrated to IRs; the green rectangles indicated the reversed segment involving five genes (rpl23, ycf2, ycf15, trnI-CAU and trnL-CAA).

Figure 3

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.

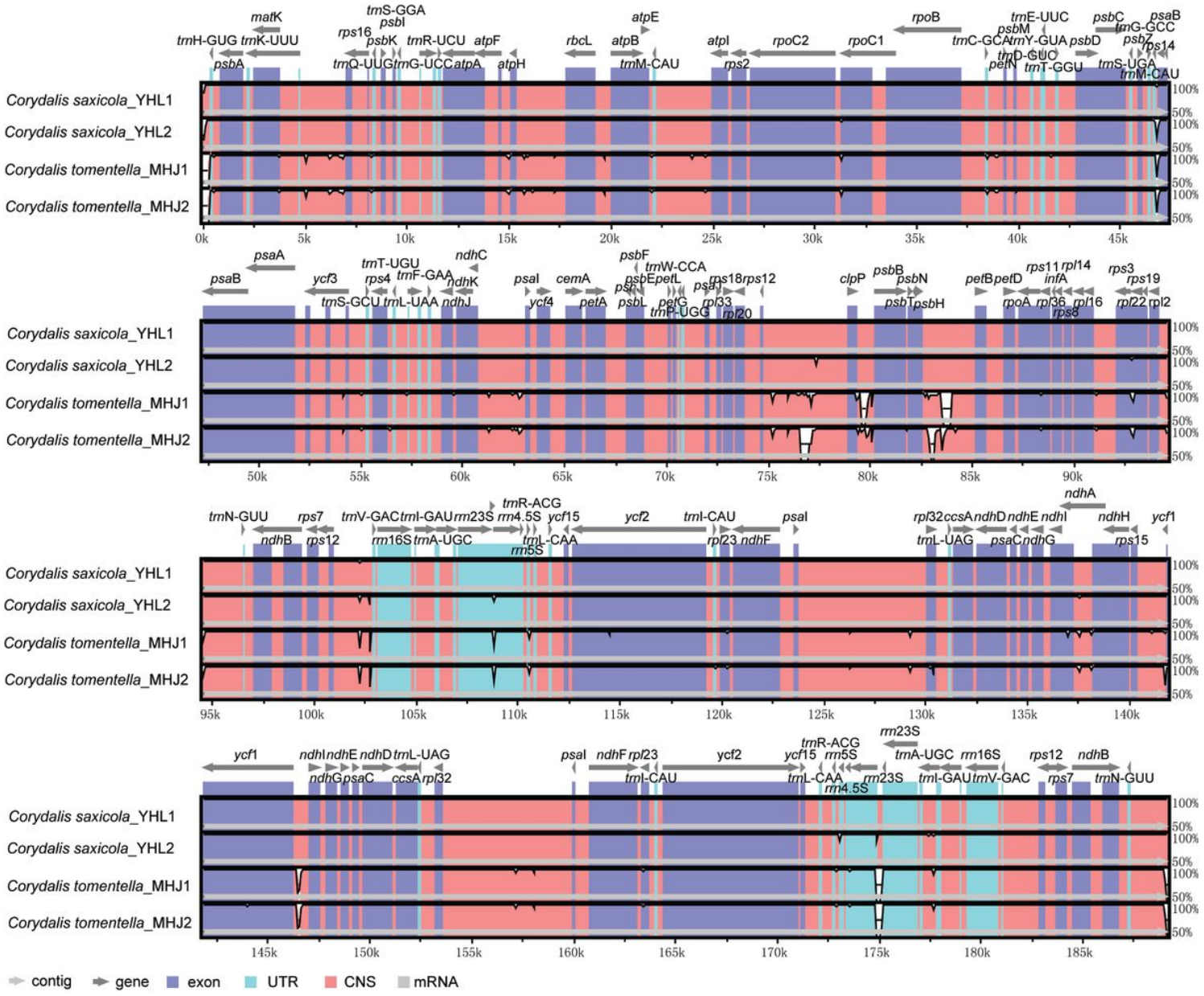

Figure 4

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 cut-off of 70% identity was used for the plots, and the Y-scale represents the percent identity ranging from 50 to 100%.

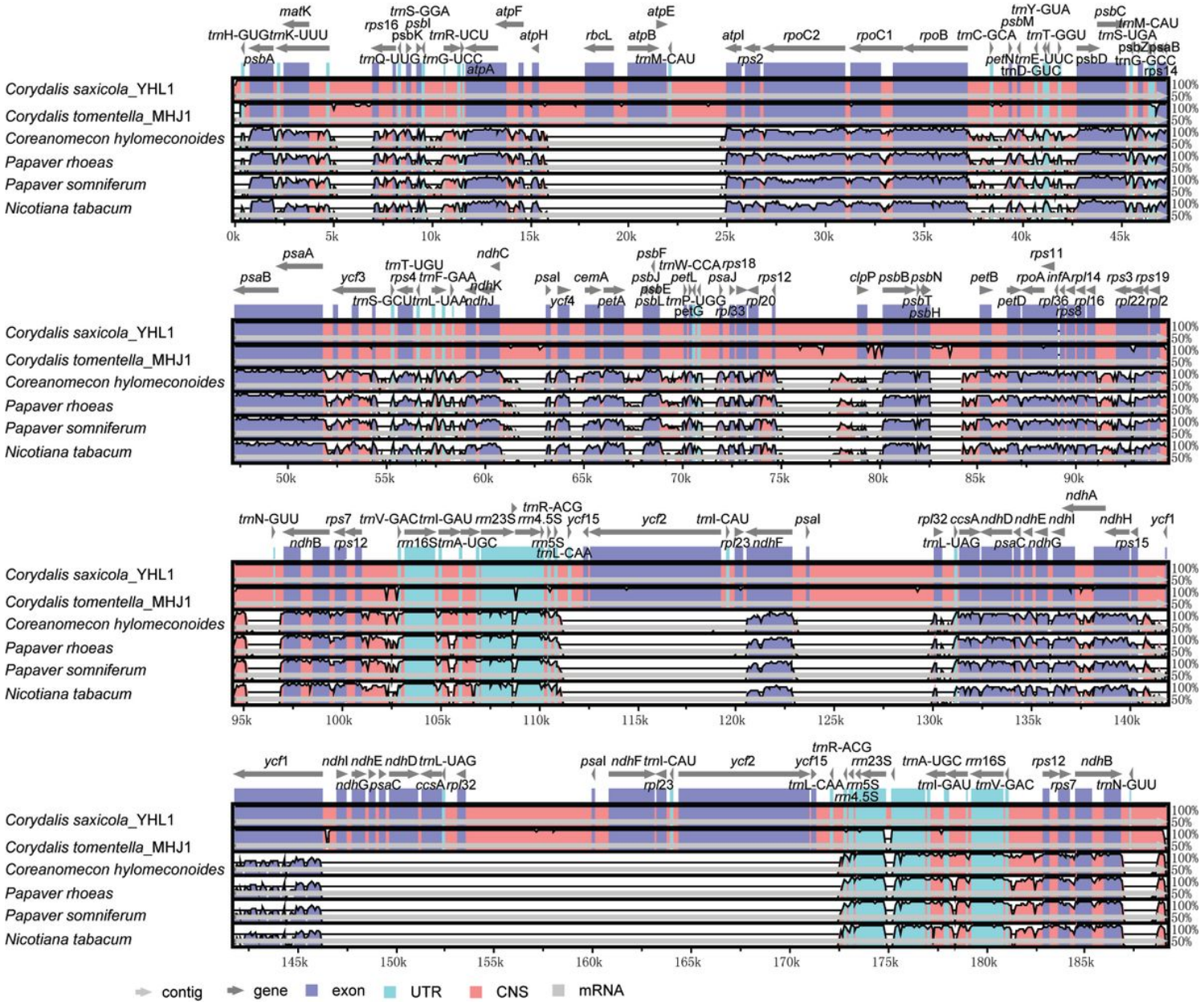

Figure 5

Sequence identity plot comparison of the cp genomes of C. tomentella, C. saxicola, Papaver somniferum, P. rhoeas, Coreanomecon 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 cut-off of 70% identity was used for the plots, and the Y-scale represents the percent identity ranging from 50 to 100%.

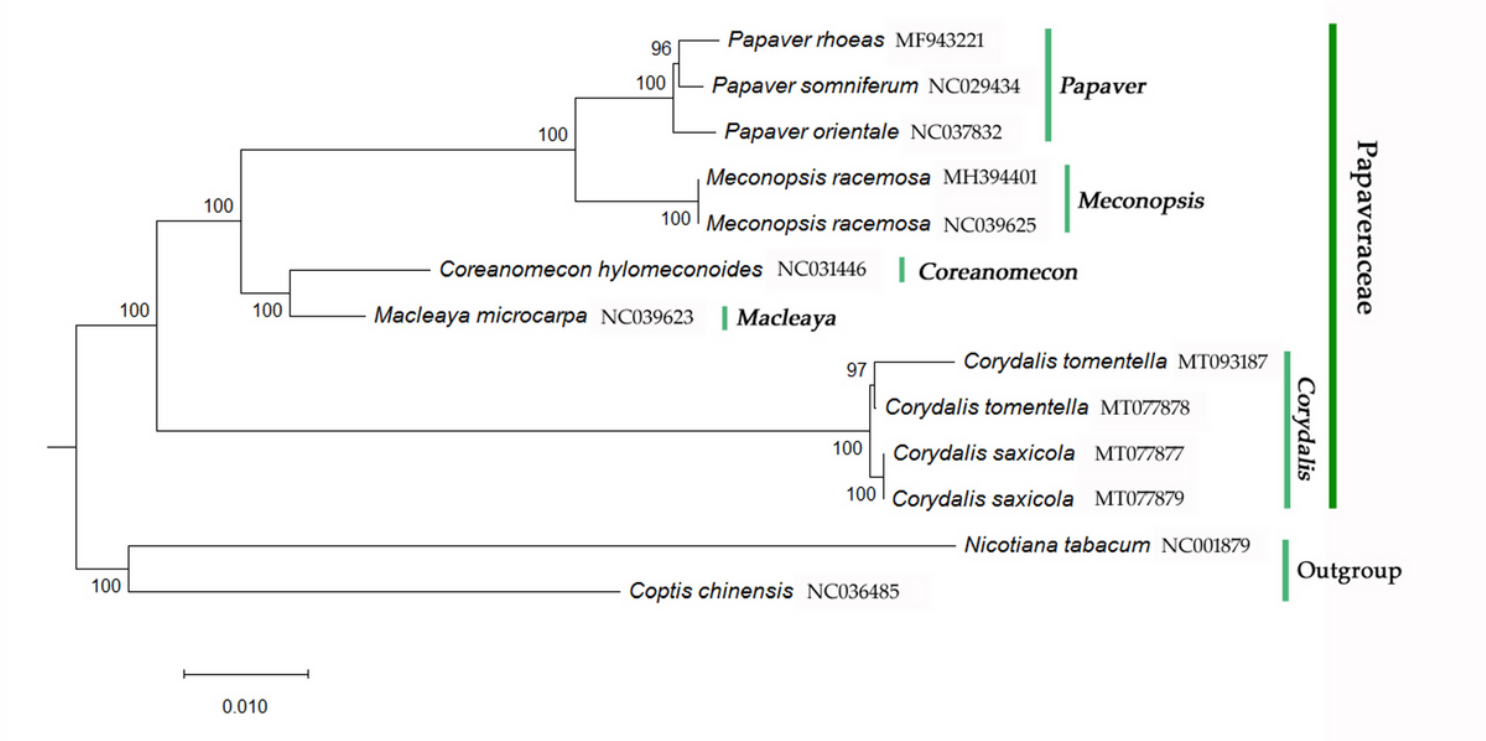

Figure 6

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

Supplementary Files

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

- TableS1.docx
- FigureS1.docx
- TableS3.docx
- TableS2.docx