1 The complete chloroplast genome of 2 Chinese medicine (*Psoralea corylifolia* 3 *Linn*): Molecular Structures, barcoding 4 analysis, and phylogenetic Analysis

5 Wei Tan^{1¶}, Han Gao^{1¶}, Huanyu Zhang¹, Xiaolei Yu¹, Xiaoxuan
6 Tian^{1*}, Weiling Jiang^{1, 2}, Kun Zhou^{1*}

7 ¹Tianjin State Key Laboratory of Modern Chinese Medicine, Tianjin University of Traditional
8 Chinese Medicine, Poyang lake road 10, Tianjin 301617, China

9 ²Research and Development Center of Traditional Chinese Medicine, Tianjin International
10 Joint Academy of Biomedicine, Tianjin 300457, China

11 * Corresponding author:

12 E-mail: tian_xiaoxuan@126.com (TXT);

13 z.k.ken@263.net (KZ).

14 [¶]These authors contributed equally to this work.

15 Abstract

16 *Psoralea corylifolia* is one kind of traditional Chinese medicine used
17 in China widely. In this study, we sequence the complete
18 chloroplast genome of *P. corylifolia*, which is 153,114 bp in size and
19 includes a pair of inverted repeats regions of 25,557 bp
20 interspersed by a small single copy of 17,885 bp and a large single
21 copy of 84,115 bp region. Approximately 98 simple sequence
22 repeats, 14 forward, 2 reverse, 2 complement, 32 palindromic and

49 tandem repeats are identified in the *P. corylifolia* chloroplast genome. The chloroplast genomes of *P. corylifolia* and three *Glycine* species are conserved in gene order and content, but show high diversity within intergenic spacers. *P. corylifolia* with three *Glycine* species in Papilionoideae fall into the same clade based on 75 conserved coding-protein genes phylogenomic analysis. Moreover, four chloroplast DNA regions (*ycf1*, *matK*, *accD*, *ndhF*) can serve as the barcodes. In general, our findings will dedicate to better comprehension of the genome aspect as well as evolutionary status of *P. corylifolia*.

Introduction

Leguminosae is known as the third-largest family of angiosperm in terms of 751 genera, consisting of 19,000 species[1, 2]. The subfamily Papilionoideae of Leguminosae compose of several tribes, including Fabeae, Galegeae, Indigoferae, Loteae, Millettieae, Phaseoleae, Psoraleae, etc[3]. Among these tribes, nucleotide substitution rates of Psoraleae have been found elevated relative to other tribes, suggestive of rapid evolution or diversification[4]. The genus *Psoralea* of Psoraleae contains about 120 species mainly distributed in southern Africa, North and South America, Australia and only one species lives in China[5]. *Psoralea Corylifolia* (Chinese name Buguzhi) is an annual plant, generating

45 pale-purple flowers. Since the occurrence of minute brown glands,
 46 the plant has a distinctive and pleasant fragrance[6]. Furthermore,
 47 the seed of *P. Corylifolia*, which can traditionally be used for the
 48 treatment of menopause, depression, kidney deficiency[7, 8].
 49 Previous pharmacological studies indicate its antioxidant,
 50 antimicrobial, antiinflammatory and chemoprotective properties.
 51 Thus, *P. Corylifolia* has been widely used in traditional Chinese
 52 medicine for its pharmacological effects to treat multiple
 53 diseases[9]. Based on its medicinal value, *P. Corylifolia* deserves
 54 utilization and development. Recently, with the rapid growth of
 55 next-generation sequencing (NGS) technologies, more and more
 56 genomic resources at reasonable schedules and prices have been
 57 provided[10]. However, despite *P. Corylifolia* is very important, the
 58 studies regarding its genetic character are scarce. Therefore, the
 59 genetic variety and phylogenetic status of this species needs
 60 analyzing by molecular techniques.

61 Chloroplast genomes, which comprise a typical structure
 62 consisting of two duplicate inverted repeat (IRs) isolated by the
 63 large and small single copy (LSC and SSC) regions, frequently
 64 have conserved annular double-stranded structure, ranging from
 65 120 to 160 kb in length[11, 12]. However, previous studies have
 66 revealed that the chloroplast genomes have shown notable

67 structural variation among Papilionoideae subfamily[13].
 68 Meanwhile, most species in Papilionoideae possess a large (50-kb)
 69 inversion in their chloroplast genomes[14]. Not only that, the loss of
 70 IRs have been detected in many species of Papilionoideae, such as
 71 *Glycyrrhiza glabra*, *Lens culinaris*, *Vicia faba*[15]. The genes
 72 obtained from the chloroplast genome have been applied to
 73 molecular identification and phylogenetic evolution analysis, as a
 74 result of its maternal inheritance pattern and comparatively
 75 independent evolution[16]. Besides that, although previous studies
 76 have recommended some loci as the plant barcode, including *rbcL*,
 77 *matK*, *trnH-psbA*, *trnL-trnF*[17, 18], the complete chloroplast
 78 genomes might be properer to be used as a super-barcode for
 79 species identification[19].

80 For such an important group, the classification and phylogenetic
 81 relationships of Psoraleae remain poorly known, even no
 82 chloroplast genomes of this tribe have been reported. Therefore, in
 83 this research, we confirm the first complete chloroplast genome
 84 sequence of *Psoralea Corylifolia* and compare *P. Corylifolia* with
 85 the published chloroplast genome of the related genus of *Glycine*,
 86 including *Glycine max*[20], *Glycine soja*[21] and *Glycine gracilis*[22].
 87 Our aim is to construct and characterize the structure of the
 88 complete chloroplast genome of *Psoralea Corylifolia* and provide

89 vital phylogenetic and genetic information for future studies in
90 Psoraleeae and legume plastomes.

91 **Materials and methods**

92 **Chloroplast genome sequencing and assembly**

93 The standard sample was bought from National Institutes for Food
94 and Drug Control, which batch number was 121056-200904. And
95 the voucher was deposited in Tianjin State Key Laboratory of
96 Modern Chinese Medicine. Total genomic DNA was detached using
97 Extract Genomic DNA Kit following the protocol of the manufacturer.
98 Then, Illumina Animal and Plant Small Fragments Libraries were
99 prepared from the sample with an average insert size of 800 bp.
100 The libraries were sequenced with 2×150 bp on the Hiseq
101 instrument. The quality of raw data was assessed with FastQC. The
102 complete chloroplast genome of *Glycine soja* (GenBank Accession:
103 NC_022868) was used as seed in the assembly of *Psoralea*
104 *corylifolia* and the paired-end reads were assembled by
105 NOVOPlasty 2.6.3[23] with k-mer length 39. The contigs were
106 subsequently oriented to construct the complete chloroplast
107 genome by mapping to reference (NC_022868). Where required,
108 the raw data was mapped to the chloroplast genome with CLC
109 workbench v11.0.1 to resolve degenerate bases. We used
110 DOGMA[24] (<http://dogma.ccbb.utexas.edu/>) to annotate the

complete plastid genome, and besides that, tRNAscan-SE[25] (<http://lowelab.ucsc.edu/tRNAscan-SE/>) was used to confirm the annotations of tRNA. Finally, all of the annotations were carefully manually checked against the reference (NC_022868, NC_030329, NC_007942), and confirmed in Sequin v15.50. OGDRAW[26] (<https://chlorobox.mpimp-golm.mpg.de/OGDraw.html>) was used to create the physical map of the circular chloroplast genome of this species. The annotated chloroplast genome sequence of *Psoralea corylifolia* was deposited in GenBank (Accession Number: MK069582).

Comparative genome analysis

The complete chloroplast genome of *Psoralea corylifolia* was compared with three species (*Glycine max*, *Glycine gracilis*, *Glycine soja*), using the mVISTA[27] (<http://genome.lbl.gov/vista/mvista/submit.shtml>) program in a LAGAN mode, while *Glycine max* was set as a reference. IRscope[28] (<https://irscope.shinyapps.io/irapp/>) was used to visualize the genes on the boundaries of LSC, SSC, and IRs according to their annotations.

Codon usage

Codon usage was determined for all protein-coding genes, the relative synonymous codon usage (RSCU) was calculated using

MEGA X[29]. Besides, the ggplot2 package was used to visualize codon usage frequency.

Analysis of single sequence repeats and tandem repeats.

REPuter[30] (http://bibiserv2.cebitec.unibielefeld.de/reputer?id=reputer_view_submission) was used to identify forward, reverse, complement and palindromic repeats in *Psoralea corylifolia* chloroplast genome. The minimal repeat size was set as 30 bp and the Hamming distance was 3. In addition to REPuter, MISA[31] was used to detect single sequence repeats (SSR), with the following setting: >10 for mononucleotide, >5 for dinucleotide, >4 for trinucleotide, and >3 for tetra-nucleotide, pentanucleotide, and hexanucleotide SSRs. Furthermore, Tandem Repeats Finder[32] (http://tandem.bu.edu/trf/trf_submit.options.html) was used to find tandem repeats with the default setting.

RNA editing site prediction

The online program Predictive RNA Editor for Plants (PREP) suite[33] (<http://prep.unl.edu/>) was used to predict potential RNA editing sites using 35 protein-coding genes (*matK*, *atpB*, *ycf3*, *psaB*, *rps14*, *rpoB*, *rpoC1*, *rpoC2*, *rps2*, *atpI*, *atpF*, *atpA*, *rps16*, *accD*, *psaI*, *psbL*, *psbF*, *psbE*, *petL*, *petG*, *rpl20*, *clpP*, *psbB*, *petB*, *petD*, *rpoA*,

155 *rps8, rpl2, rpl23, ndhB, ndhA, ndhG, ndhD, ccsA, ndhF*) of *Psoralea*
156 *corylifolia* chloroplast genome with default parameters.

157 **Synonymous and nonsynonymous substitution rate** 158 **analysis.**

159 The genes of the species (*P. corylifolia*, *G. max*, *G. gracilis*, and *G.*
160 *soja*) with the same functions were grouped following previous
161 studies[34-36]. The singular genes (*clpP*, *cemA*, *matK*, *ccsA*, and
162 *ycf1*) and the same functions genes such as ATP synthase (*atp*),
163 Cytochrome b6/f complex (*pet*), Subunits of NADH-dehydrogenase
164 (*ndh*), Photosystem I (*psa*), Photosystem II (*psb*), Large subunit of
165 ribosomal (*rpl*), DNA-dependent RNA polymerase (*rpo*) and Small
166 subunit of ribosomal proteins (*rps*) were analyzed. PAML v4.9
167 package[37] was used to analyze synonymous and
168 nonsynonymous substitution rate, and *Psoralea corylifolia* was set
169 as outgroup. Within the PAML package, the yn00 program was
170 utilized to calculate dN, dS, and dN/dS. Each sequence was set as
171 3×4 (codon position×base) table, and we used the Nei–Gojobori
172 (1986) method[38] to confirm the value of dN/dS.

173 **Divergent Hotspots**

174 MAFFT v7[39] was applied to align the four complete chloroplast
175 genomes. Furthermore, the nucleotide polymorphism of the
176 chloroplast genomes were estimated by DnaSP v6.11.01[40]. The

177 following conditions for sliding window option were set as: window
178 length of 600 bp, step size of 200 bp.

179 **Phylogenetic and DNA barcoding analysis.**

180 Previous studies indicated that the chloroplast genome was
181 qualified to solve the phylogenetic relationship of Fabaceae[13]. In
182 this study, we compared 36 species of Papilionoideae, including *P.*
183 *corylifolia*, *G.max*, *G.gracilis*, *G.soja*. With the exception of these
184 four species, the remaining 32 species were downloaded from the
185 GenBank (S1 Table). It's worth noting that the 36 species
186 represented 36 genera in Papilionoideae subfamily. The 75
187 protein-coding genes were extracted manually and aligned
188 separately using MAFFT v7, and then these matrices were
189 concatenated to form a final data matrix. The phylogenetic
190 relationship of Papilionoideaea was rebuilt while *Pararchidendron*
191 *pruinsum* was set as the outgroup. RAxML version 8.2.4[41] was
192 used to perform Maximum likelihood (ML) analysis, using 1000
193 replicates of rapid bootstrap with the GTRGAMMAI substitution
194 model. Bayesian inference (BI) was implemented using MrBayes
195 v3.2.6[42] with default parameters (Lset nst=6, rates=invgamma,
196 Prset statefreqpr=dirichlet(1,1,1,1)). In our study, we found
197 Psoraleeae was more closely related to the Phaseoleae than to the
198 other tribes in the Papilionoideae. Therefore, 9 species of two tribes

199 were collected to test the potential barcodes. Four various regions
200 in the chloroplast genomes have been proposed as DNA barcodes,
201 including *ycf1*, *matK*, *accD*, *ndhF*, genes. The ML trees were
202 reconstructed using the four genes with GTRGAMMAI model,
203 respectively.

204 **Results and discussion**

205 **Chloroplast genome assembly**

206 By using the Illumina sequencing platform, we obtained a total of
207 12,330,274 reads with an average read length of 150 bp. The reads
208 were re-mapped to the chloroplast genome, and the coverage of
209 the chloroplast genome was 3,241×. The size of the whole genome
210 was 153,114 bp.

211 **Organization and gene content.**

212 The complete plastid genome of *Psoralea corylifolia* displayed a
213 typical quadripartite structure (Fig 1), which comprised of a pair of
214 inverted repeats (each 25,557 bp in length) separated by the SSC
215 and the LSC region (17,885 and 84,115 bp in length, respectively).
216 All the above values were similar to three *Glycine* species which
217 can be seen in the table (Table 1). The GC content of the complete
218 chloroplast genome was 35.5%. The GC content of the IR (41.9%)
219 was higher than that in LSC (32.7%) and SSC regions (28.8%). The
220 complete chloroplast genome consisted of 111 unique genes,

including 77 protein-coding genes, 30 tRNA genes, 4 rRNA genes. Among these genes, 16 genes (*petB*, *petD*, *rpl2*, *rpl16*, *rps16*, *rpoC1*, *trnI-GAU*, *trnA-UGC*, *trnG-UCC*, *trnL-UAA*, *trnV-UAC*, *trnK-UUU*, *ndhA*, *ndhB*, *rps12*, *atpF*, *clpP* and *ycf3*) included one intron, excepted *ycf3* and *clpP* genes who contained two. In the functional classification of genes, 39 genes were involved in photosynthesis, 5 for Photosystem I, 15 for Photosystem II, 6 for Cytochrome b6/f complex, 6 for different subunits of ATP synthase and 1 for RuBisCO. 5 genes were associated with different functions, and besides, the function of *ycf* genes (*ycf1*, *ycf2*, *ycf3*) existing in *Psoralea corylifolia* was unknown (Table 2).

Fig 1. Physical map of the chloroplast genome of the *Psoralea corylifolia*.

The genes drawn inside and outside of the circle are transcribed clockwise and counterclockwise, respectively. Different colors represented different functional genes. The GC content of the genome is plotted on the inner circle as dark gray, whereas the light gray corresponded to AT content.

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

Species	<i>P.corylifolia</i>	<i>G.max</i>	<i>G.gracilis</i>	<i>G.soja</i>
Size (bp)	153,114	152,218	152,218	152,217
LSC (bp)	84,115	83,175	83,175	83,174
IR (bp)	25,557	25,574	25,574	25,574

SSC (bp)	17,885	17,895	17,895	17,895
genes	111	111	111	111
CDS	77	77	77	77
tRNA	30	30	30	30
rRNA	4	4	4	4
GC%	35.3%	34%	35.37%	35.4%

238 **Table 2. Genes in the sequenced *Psoralea corylifolia* genome.**

Category	Group of genes	Name of genes
Self-replication	Large subunit of ribosomal proteins	<i>rpl2^a</i> (2), <i>rpl14</i> , <i>rpl16^a</i> , <i>rpl20</i> , <i>rpl23</i> (2), <i>rpl32</i> , <i>rpl33</i> , <i>rpl36</i>
	Small subunit of ribosomal proteins	<i>rps2</i> , <i>rps3</i> , <i>rps4</i> , <i>rps7</i> (2), <i>rps8</i> , <i>rps11</i> , <i>rps12^a</i> (3), <i>rps14</i> , <i>rps15</i> , <i>rps16^a</i> , <i>rps18</i>
	DNA-dependent RNA polymerase	<i>rpoA</i> , <i>rpoB</i> , <i>rpoC1^a</i> , <i>rpoC2</i>
	rRNA genes	<i>rrn4.5</i> (2), <i>rrn5</i> (2), <i>rrn16</i> , <i>rrn23</i>
	tRNA genes	<i>trnI-CAU</i> (2), <i>trnL-CAA</i> (2), <i>trnV-GAC</i> (2), <i>trnI-GAU^a</i> (2), <i>trnA-UGC^a</i> (2), <i>trnR-ACG</i> (2), <i>trnN-GUU</i> (2), <i>trnL-UAG</i> , <i>trnP-UGG</i> , <i>trnW-CCA</i> , <i>trnQ-UUG</i> , <i>trnS-GCU</i> , <i>trnR-UCU</i> , <i>trnC-GCA</i> , <i>trnD-GUC</i> , <i>trnY-GUA</i> , <i>trnE-UUC</i> ,

		<i>trnT</i> -GGU, <i>trnS</i> -UGA, <i>trnG</i> -UCC ^a , <i>trnG</i> -GCC, <i>trnM</i> -CAU, <i>trnS</i> -GGA, <i>trnH</i> -GUG, <i>trnT</i> -UGU, <i>trnL</i> -UAA ^a , <i>trnF</i> -GAA, <i>trnV</i> -UAC ^a , <i>trnM</i> -CAU, <i>trnK</i> -UUU ^a
Photosynthesis	Photosystem I	<i>psaA</i> , <i>psaB</i> , <i>psaC</i> , <i>psaI</i> , <i>psaJ</i>
	Photosystem II	<i>psbA</i> , <i>psbB</i> , <i>psbC</i> , <i>psbD</i> , <i>psbE</i> , <i>psbF</i> , <i>psbH</i> , <i>psbI</i> , <i>psbJ</i> , <i>psbK</i> , <i>psbL</i> , <i>psbM</i> , <i>psbN</i> , <i>psbT</i> , <i>psbZ</i>
	Cytochrome b6/f complex	<i>petA</i> , <i>petB</i> ^a , <i>petD</i> ^a , <i>petG</i> , <i>petL</i> , <i>petN</i>
	ATP synthase	<i>atpA</i> , <i>atpB</i> , <i>atpE</i> , <i>atpF</i> ^a , <i>atpH</i> , <i>atpI</i>
	RuBisCO	<i>rbcL</i>
	Subunits of NADH-dehydrogenase	<i>ndhA</i> ^a , <i>ndhB</i> ^a (2), <i>ndhC</i> , <i>ndhD</i> , <i>ndhE</i> , <i>ndhF</i> , <i>ndhG</i> , <i>ndhH</i> , <i>ndhI</i> , <i>ndhJ</i> , <i>ndhK</i>
Other genes	Maturase	<i>matK</i>
	Protease	<i>clpP</i> ^b
	Envelop membrane protein	<i>cemA</i>
	Subunit acetyl-CoA-carboxylate	<i>accD</i>
	c-Type cytochrome synthesis gene	<i>ccsA</i>

Unknown	Conserved open reading frames	<i>ycf1</i> , <i>ycf2</i> (2), <i>ycf3</i> ^b
---------	-------------------------------	---

239 a: Gene containing a single intron

240 b: Gene containing two introns.

241 **Comparative genome analysis**

242 The mVISTA program was utilized to analyze the whole sequence
243 identity of the four chloroplast genomes, with *G.max* as a reference.

244 As expected, the alignment of the chloroplast genome showed high
245 sequence similarity, which indicated that they were highly
246 conserved at the scale of genome level. The IR regions were less
247 divergent than LSC and SSC regions, and besides that, more
248 sequence variations were detected to be located in the non-coding
249 regions rather than coding regions. Furthermore, the divergent
250 coding regions were within *ycf1*, *matK*, *accD*, *ndhF*, which could be
251 used as specific DNA barcodes. The *trnH-psbA*, *trnK-rbcL*,
252 *trnG-psbZ*, *psbD-trnT-trnE*, *petN-trnC*, *trnG-trnS*, *trnR-trnG-trnS*,
253 *ndhI-ndhG*, *trnL-rpl32* loci and *petD*, *rpl16*, *ndhA* introns were most
254 divergent regions, which were located in the intergenic spacers
255 (IGSs) and introns, respectively. The boundary of IR regions
256 indicated hypervariable regions with many nucleotide changes in
257 chloroplast genomes of closely related species[43], simultaneously,
258 the IR border expansion/contraction was considered as an

259 evolutionary event[36]. In this study, we compared genes on the
 260 boundaries of the junction sites of the *Psoralea corylifolia* with other
 261 three species of *Glycine* (*Glycine max*, *Glycine gracilis*, *Glycine*
 262 *soja*) (Fig 2). The IRb/LSC (JLB) border of *Psoralea corylifolia* was
 263 located on *rps19* gene, with a size of pseudogene as 61 bp,
 264 showing the length 7 bp shorter than other species. The *ycf1* gene
 265 crossed the IRb/SSC (JSB) region, and the size of the *ycf1*
 266 pseudogene was 5,378 bp in all three *Glycine* species and 5,405
 267 bp in *P. corylifolia*. The SSC/IRA (JSA) border extended into *ndhF*
 268 genes in *P. corylifolia* by 9 bp, whereas the *ndhF* genes of the other
 269 3 species were all found in the SSC region. The *trnH-GUG* gene
 270 was at the IRa/LSC (JLA) border among the three *Glycine* species
 271 and *P. corylifolia* by 2 bp and 16 bp, respectively (Fig 3).

272 **Fig 2. Sequence identity plots among chloroplast genomes of *P.***
 273 ***corylifolia*, *G. max*, *G. gracilis*, *G. soja*, with *G. max* as a reference.** Grey
 274 arrows above the alignment indicate the position and orientation of genes.
 275 The vertical scale represents the percent identity among 50-100%. Exon,
 276 intron, and conserved non-coding sequences (CNS) are colored purple, blue
 277 and red, respectively.

278 **Fig 3. Comparison of LSC, IR, and SSC junction positions among**
 279 **chloroplast genomes of *P. corylifolia*, *G. max*, *G. gracilis*, *G. soja*.** The JLB,
 280 JSB, JSA and JLA refer to junctions of IRb/LSC, IRb/SSC, SSC/IRa and

281 IRa/LSC, respectively.

282 **Codon usage frequency analysis**

283 The codon usage frequency was generally linked with gene
284 function both in the level of individual amino acids and
285 nucleotides[44], and besides that, previous evidence indicated that
286 the chloroplast genome might display particular features of codon
287 usage[45]. Based on the nucleic acid sequence in the
288 protein-coding region, MEGA X was used to calculate codon usage
289 frequency and RSCU in the *P. corylifolia* chloroplast genome. The
290 protein-coding genes were composed of 77,916 bp which encoded
291 25,972 codons. In addition, the most frequently used amino acids
292 were leucine encoded 2,747 (10.6%), whereas only 306 (1.2%)
293 encoded cysteine was the least frequently used amino acids.
294 Almost all the chloroplast genomes strongly biased A- and/or
295 U-ending codons. The result also revealed that all bias favoring
296 synonymous codons ($RSCU > 1$) with an A or U at the third codon
297 position except for UUG, CUA, UGA. Usage of the initial codon
298 AUG and tryptophan UGG, liked other land plant chloroplast
299 genomes, demonstrated no bias ($RSCU = 1$) (S2 Table) (Fig 4).

300 **Fig 4. Percentage of codon usage of each amino acid in the *Psoralea***
301 ***corylifolia* chloroplast genome.** The x-axis represents codon families.

302 **Repeat sequence**

There existed numerous tandem repeats in the genome of eukaryotes with extremely high repetition frequency and very low order complexity. Simple sequence repeats (SSRs), consisting of 1-6 repeat nucleotide units, are frequently used as molecular markers, taking advantage of their highly reproducible, rich in polymorphism, co-dominance, and highly reliable[46]. We detected a total of 98 SSRs in *P. corylifolia* including 63 repeats for mononucleotide, 22 repeats for dinucleotide, 2 repeats for trinucleotide, 11 repeats for tetranucleotide, respectively. Because almost all chloroplast genomes had high AT content[47], and almost all SSR loci contained A or T, especially in the mononucleotide (Fig 5). In addition to the SSRs, we identified four type repeats in *P. corylifolia*, including forward, reverse, complement and palindromic repeats by REPuter. These repeats contained 14 forward repeats, 2 reverse repeats, 2 complement repeats, 32 palindromic repeats. The length of repeats varied between 30 and 257, even up to 25,557. Among these repeats, great majority repeats located in intergenic regions (S3 Table). As for tandem repeats, 49 repeats were identified in *P. corylifolia* chloroplast genome (S4 Table).

Fig 5. Analysis of simple sequence repeat (SSR) in the *Psoralea corylifolia* chloroplast genome. (A) Number of different SSRs types. (B)

325 Proportion of SSRs in the LSC, IR and SSC regions, respectively. (C)

326 Frequency of SSR loci identified in different repeat types.

327 **Potential RNA editing sites**

328 It was known that RNA editing involved various procedures which
 329 could alter the protein coding sequence of a transcribed RNA by
 330 inserting, deleting or modifying nucleotides in the transcript[48]. In
 331 this study, 35 genes were employed to predict potential RNA editing
 332 sites , and 10 genes (*atpB*, *ycf3*, *atpI*, *atpA*, *psaI*, *petL*, *petG*, *psbB*,
 333 *petD*, *rpl2*) had been identified no sites. As a result, a total of 90
 334 potential RNA editing sites were predicted. S to L (26%) of amino
 335 acids change occurred most, whereas P to S (2.2%) and R to C
 336 (1.1%) occurred least. Furthermore, all the type of RNA editing sites
 337 were C- to U-. We also observed that 25 (27.7%) and 65 (72.3%)
 338 editing sites were located at the first and the second codon position,
 339 respectively (S5 Table).

340 **Synonymous and nonsynonymous substitution rate**

341 In genetics, dN/dS represented the ratio between a synonymous
 342 substitution (dN) and a nonsynonymous substitution (dS), which
 343 could be used to determine whether there is selective pressure on
 344 the protein-coding gene. Previous studies had estimated that,
 345 dN/dS ratios <1, =1, >1 indicating purifying selection, neutral
 346 evolution, and positive selection, respectively[49]. The *psa*, *psb*,

347 and *pet* genes exhibited the lowest dN values, while the *ycf1* gene
348 presented the highest dN values. As for dS values, the *clpP*, *ycf1*
349 genes occupied the lowest and highest dS values, respectively.
350 Furthermore, the *ccsA* and *ycf1* genes ($dN/dS > 0.5$) were
351 observed to have highest the ratios of dN/dS, followed by *clpP*,
352 *matK*, *cemA* genes (Fig 6).

353 **Fig 6. Nonsynonymous substitution (dN), synonymous substitution (dS),**
354 **and dN/dS values for groups of gene or individual genes.**

355 **Divergent Hotspots Analysis**

356 Some highly variable regions existed in chloroplast genome
357 sequences, which were usually clustered and described as
358 “hotspots”[50, 51]. The nucleotide variability (Pi) was calculated by
359 the DnaSP v6.11 software. The values ranged from 0 to 0.07028,
360 meanwhile, the IR region was more conserved than the LSC and
361 SSC regions. Seven mutational hotspots ($P_i > 0.04$), including
362 *psbD-trnT-trnE*, *atpI-atpH*, *trnG-trnS-psbI*, *accD-psaI-cemA*,
363 *rps11-rpl36-rps8*, *ycf1*, *trnL-rpl32-ndhF*, as well as two loci ($P_i > 0.05$)
364 including *trnK-rbcL*, *rps16-accD*, exhibited remarkably higher Pi
365 values, which was similar to the result of comparative genome
366 analysis discovered by mVISTA program (Fig 7). What’s more,
367 these hotspots regions would provide potential molecular marker
368 candidates for plant identification and accelerate subsequent

369 studies on phylogenetic analysis.

370 **Fig 7. Sliding window analysis among the whole chloroplast genomes of**
 371 ***P. corylifolia*, *G. max*, *G. gracilis*, *G. soja*.** The x-axis represents the position
 372 of the midpoint of a window, the y-axis represents the nucleotide diversity of
 373 each window.

374 **Phylogenetic and DNA barcoding analysis.**

375 We used 75 chloroplast protein-coding genes of 37 species to
 376 analyze the phylogenetic position of *P. corylifolia*. The phylogenetic
 377 trees generated by ML and BI alignment had similar topologies. In
 378 the maximum likelihood (ML) and Bayesian inference (BI) trees,
 379 almost all of nodes had a 100% bootstrap value and 1.0 Bayesian
 380 posterior probability. Furthermore, the result of *P. corylifolia* formed
 381 a single clade with three *Glycine* species strongly supported the
 382 sister relationship between *P. corylifolia* and *Glycine* species (Fig 8),
 383 which confirmed the relationship between *Psoralea* and *Glycine*
 384 genus inferred from the phylogenetic work conducted based on
 385 *trnK/matK* genes[52]. More importantly, the ML trees performed
 386 using four candidate barcodes (*ycf1*, *matK*, *accD*, *ndhF*) was
 387 consistent with the 75 protein-coding-based tree (S6 Fig).

388 **Fig 8. Phylogenetic trees of Papilionoideae subfamily.** Phylogenetic trees
 389 reconstructed with 75 protein-coding genes of 37 species using maximum
 390 likelihood (ML) and bayesian inference (BI) methods. Numbers at nodes are

391 values for bootstrap and Bayesian posterior probability.

392 **Conclusion and discussion**

393 In this study, we present the complete chloroplast genome of *P.*
 394 *corylifolia* using Illumina sequencing platforms, and this is the first
 395 reported chloroplast genome assembly in the tribe Psoraleeae. *P.*
 396 *corylifolia* chloroplast genome (153,114 bp) is fully characterized
 397 and compare to the chloroplast genomes of related species
 398 previously reported. The chloroplast genome of *P. corylifolia*
 399 contains 111 unique genes, including 77 protein-coding genes, 4
 400 rRNA genes and 30 tRNA genes. The ML and BI phylogenetic trees
 401 suggest that *Psoralea* is a sister genus of *Glycine*. The 16 highest
 402 divergent regions (*ycf1*, *matK*, *accD*, *ndhF*, *petD*, *rpl16*, *ndhA*,
 403 *trnH-psbA*, *trnK-rbcL*, *trnG-psbZ*, *psbD-trnT-trnE*, *petN-trnC*,
 404 *trnG-trnS*, *trnR-trnG-trnS*, *ndhI-ndhG*, *trnL-rpl32*) are detected.

405 In our study, the four highly variable protein-coding genes (*ycf1*,
 406 *matK*, *accD*, *ndhF*) are selected as the potential markers. Previous
 407 evidence supported that *ycf1* gene was one of the core plastid DNA
 408 barcode of land plants[53]. Meanwhile, the phylogenetic
 409 relationship of Papilionoideae based on *matK* gene had been
 410 reported[54]. Moreover, not only the *ycf1* and *matK* genes but also
 411 the rest two genes show good performance in distinguishing
 412 Phaseoleae and Psoraleeae species. These genes can be taken as

413 promising chloroplast DNA barcodes for species identification and
414 phylogenetic studies.

415 Across 36 species (genome size 176,692~121,020) of
416 Papilionoideae subfamily, we find that all species hold 50-kb
417 inversion. Moreover, we also find that 13 species from 6 tribes lose
418 their IR (IRL), including tribes Galegeae, Millettieae, Caraganeae,
419 Trifolieae, Fabeae, as well as Cicereae. Interestingly, though
420 *Wisteria* was very similar in appearance to *Millettia* and they
421 classified into the same tribe (Millettieae)[52], *Wisteria* and *Millettia*
422 are not supported as monophyletic in our result. According to this
423 phenomenon, we speculate that it might be attributed in part to the
424 different genome structure of *Wisteria* (IRL). In addition to these
425 findings, a distinctive characteristic of Papilionoideae is the lack of
426 *rpl22* genes in all species. It was quite possible that the genes loss
427 are previously functionally transferred to the nucleus, as previously
428 noted by researchers[55]. At the same time, it was likely that the
429 *rpl33* gene contained stop codons within its coding region,
430 therefore the gene presented as a pseudogene in *Vigna* and
431 *Phaseolus*[56, 57].

432 In China, *P. corylifolia* is a vital traditional Chinese medicine. We
433 believe that the complete chloroplast genome can provide helpful
434 information for further research. The selected candidate barcodes

(*ycf1*, *matK*, *accD*, *ndhF*) from the highest divergent regions are able to enrich the barcode library and increase the efficiency of species identification, and it is a big step forward.

Acknowledgments

We are thankful to the National Natural Science Foundation of China (Grant no. 81673826) for its financial support to our study. We also acknowledge Ms. Liran Sun for her previous work.

Author Contributions

Conceptualization: Xiaoxuan Tian, Wei Tan, Kun Zhou.

Data Curation: Wei Tan, Han Gao.

Formal Analysis: Wei Tan, Han Gao, Huanyu Zhang .

Funding Acquisition: Kun Zhou.

Resources: Wei Tan, Han Gao.

Software: Wei Tan, Han Gao.

Supervision: Xiaoxuan Tian, Kun Zhou.

Validation: Wei Tan, Han Gao, Xiaolei Yu.

Writing – Original Draft Preparation: Wei Tan.

Writing – Review & Editing: Xiaoxuan Tian, Weiling Jiang, Han Gao, Huanyu zhang.

References

1. CHRISTENHUSZ MJM, BYNG JW. The number of known plants species in the world and its annual increase. 2016. 2016;261(3):17. Epub 2016-05-20. doi:

- 10.11646/phytotaxa.261.3.1.
2. Stevens PF. Angiosperm Phylogeny Website 28 April 2008. Available from: <http://www.mobot.org/MOBOT/Research/APweb/orders/fabalesweb.htm#Fabaceae>.
3. Wojciechowski M, J. Sanderson M, Steele K, Liston A. Molecular phylogeny of the “Temperate Herbaceous Tribes” of papilionoid legumes 2000.
4. Egan AN, Crandall KA. Divergence and diversification in North American Psoraleeae (Fabaceae) due to climate change. BMC biology. 2008;6:55-. doi: 10.1186/1741-7007-6-55. PubMed PMID: 19091055.
5. Committee FOCE. Flora of China: Science Press; 1995. 344 p.
6. G. Miller AM, Miranda. Plants of Dhofar. Oman. 174 ed 1998.
7. Wang Y, Hong C, Zhou C, Xu D, Qu H-b. Screening Antitumor Compounds Psoralea and Isopsoralea from Psoralea corylifolia L. Seeds. Evidence-Based Complementary and Alternative Medicine. 2011;2011. doi: 10.1093/ecam/nen087.
8. Zhao L, Huang C, Shan Z, Xiang B, Mei L. Fingerprint analysis of Psoralea corylifolia L. by HPLC and LC–MS. Journal of Chromatography B Analytical Technologies in the Biomedical & Life Sciences. 2005;821(1):67-74. doi: 10.1016/j.jchromb.2005.04.008.
9. Zhang X, Zhao W, Wang Y, Lu J, Chen X. The Chemical Constituents and Bioactivities of Psoralea corylifolia Linn.: A Review. The American journal of Chinese medicine. 2016;44(1):35-60. Epub 2016/02/27. doi: 10.1142/s0192415x16500038. PubMed PMID: 26916913.
10. Mardis ER. Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics. 2008;9(1):387-402. doi: 10.1146/annurev.genom.9.081307.164359.

- 479 PubMed PMID: 18576944.
- 480 11. Bock R. Structure, function, and inheritance of plastid genomes. In: Bock R, editor. Cell
481 and Molecular Biology of Plastids. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p.
482 29-63.
- 483 12. Jansen RK, Ruhlman TA. Plastid Genomes of Seed Plants. In: Bock R, Knoop V, editors.
484 Genomics of Chloroplasts and Mitochondria. Dordrecht: Springer Netherlands; 2012. p.
485 103-26.
- 486 13. Wang YH, Qu XJ, Chen SY, Li DZ, Yi TS. Plastomes of Mimosoideae: structural and size
487 variation, sequence divergence, and phylogenetic implication. Tree Genetics & Genomes.
488 2017;13(13):41. doi: 10.1007/s11295-017-1124-1.
- 489 14. Doyle JJ, Doyle JL, Ballenger JA, Palmer JD. The Distribution and Phylogenetic
490 Significance of a 50-kb Chloroplast DNA Inversion in the Flowering Plant Family Leguminosae.
491 Molecular Phylogenetics and Evolution. 1996;5(2):429-38. doi:
492 <https://doi.org/10.1006/mpev.1996.0038>.
- 493 15. Sabir J, Schwarz E, Ellison N, Zhang J, Baeshen NA, Mutwakil M, et al. Evolutionary and
494 biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of
495 legumes. Plant biotechnology journal. 2014;12(6):743-54. Epub 2014/03/13. doi:
496 10.1111/pbi.12179. PubMed PMID: 24618204.
- 497 16. Zhou J, Chen X, Cui Y, Sun W, Li Y, Wang Y, et al. Molecular Structure and Phylogenetic
498 Analyses of Complete Chloroplast Genomes of Two Aristolochia Medicinal Species.
499 International Journal of Molecular Sciences. 2017;18(9):1839. doi: 10.3390/ijms18091839.
- 500 17. Ferri G, Corradini B, Ferrari F, Santunione AL, Palazzoli F, Alu M. Forensic botany II,

501 DNA barcode for land plants: Which markers after the international agreement? Forensic
502 science international Genetics. 2015;15:131-6. Epub 2014/12/03. doi:
503 10.1016/j.fsigen.2014.10.005. PubMed PMID: 25457632.

504 18. Hollingsworth PM, Graham SW, Little DP. Choosing and using a plant DNA barcode. PloS
505 one. 2011;6(5):e19254. Epub 2011/06/04. doi: 10.1371/journal.pone.0019254. PubMed PMID:
506 21637336; PubMed Central PMCID: PMC3102656.

507 19. Xia Y, Hu Z, Li X, Wang P, Zhang X, Li Q, et al. The complete chloroplast genome
508 sequence of *Chrysanthemum indicum*. Mitochondrial DNA Part A. Epub 4669. doi:
509 10.3109/19401736.2015.1106494.

510 20. Saski C, Lee SB, Daniell H, Wood TC, Tomkins J, Kim HG, et al. Complete chloroplast
511 genome sequence of *Glycine max* and comparative analyses with other legume genomes.
512 Plant molecular biology. 2005;59(2):309-22. Epub 2005/10/26. doi:
513 10.1007/s11103-005-8882-0. PubMed PMID: 16247559.

514 21. Gao CW, Gao LZ. The complete chloroplast genome sequence of wild soybean, *Glycine*
515 *soja*. Conservation Genetics Resources. 2017;9(2):1-3.

516 22. Gao CW, Gao LZ. The complete chloroplast genome sequence of semi-wild soybean,
517 *Glycine gracilis* (Fabales: Fabaceae). Conservation Genetics Resources. 2017;9(2):1-3. doi:
518 10.1007/s12686-016-0683-z.

519 23. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle
520 genomes from whole genome data. Nucleic Acids Research. 2017;45(4):e18. doi: doi:
521 10.1093/nar/gkw955. PubMed Central PMCID: PMC28204566.

522 24. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with

523 DOGMA. Bioinformatics (Oxford, England). 2004;20(17):3252-5. Epub 2004/06/08. doi:
524 10.1093/bioinformatics/bth352. PubMed PMID: 15180927.

525 25. Lowe TM, Chan PP. tRNAscan-SE On-line: integrating search and context for analysis of
526 transfer RNA genes. Nucleic Acids Res. 2016;44(W1):W54-7. Epub 2016/05/14. doi:
527 10.1093/nar/gkw413. PubMed PMID: 27174935; PubMed Central PMCID: PMC4987944.

528 26. Lohse M, Drechsel O, Kahlau S, Bock R. OrganellarGenomeDRAW--a suite of tools for
529 generating physical maps of plastid and mitochondrial genomes and visualizing expression
530 data sets. Nucleic Acids Res. 2013;41(Web Server issue):W575-81. Epub 2013/04/24. doi:
531 10.1093/nar/gkt289. PubMed PMID: 23609545; PubMed Central PMCID: PMC3692101.

532 27. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for
533 comparative genomics. Nucleic Acids Res. 2004;32(Web Server issue):W273-9. Epub
534 2004/06/25. doi: 10.1093/nar/gkh458. PubMed PMID: 15215394; PubMed Central PMCID:
535 PMCPMC441596.

536 28. Amiryousefi A, Hyvönen J, Poczar P. IRscope: an online program to visualize the junction
537 sites of chloroplast genomes. Bioinformatics (Oxford, England). 2018;34(17):3030-1. doi:
538 10.1093/bioinformatics/bty220.

539 29. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary
540 Genetics Analysis across Computing Platforms. Molecular biology and evolution.
541 2018;35(6):1547-9. Epub 2018/05/04. doi: 10.1093/molbev/msy096. PubMed PMID:
542 29722887; PubMed Central PMCID: PMC5967553.

543 30. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter:
544 the manifold applications of repeat analysis on a genomic scale. Nucleic acids research.

545 2001;29(22):4633-42. PubMed PMID: 11713313.

546 31. Thiel T, Michalek W, Varshney RK, Graner A. Exploiting EST databases for the
547 development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare*
548 L.). TAG Theoretical and applied genetics Theoretische und angewandte Genetik.
549 2003;106(3):411-22. Epub 2003/02/18. doi: 10.1007/s00122-002-1031-0. PubMed PMID:
550 12589540.

551 32. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids*
552 *Res.* 1999;27(2):573-80. Epub 1998/12/24. PubMed PMID: 9862982; PubMed Central PMCID:
553 PMCPMC148217.

554 33. Mower JP. The PREP suite: predictive RNA editors for plant mitochondrial genes,
555 chloroplast genes and user-defined alignments. *Nucleic Acids Res.* 2009;37(Web Server
556 issue):W253-9. Epub 2009/05/13. doi: 10.1093/nar/gkp337. PubMed PMID: 19433507;
557 PubMed Central PMCID: PMCPMC2703948.

558 34. Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, et al. The chloroplast genome
559 of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that
560 of grasses and its phylogenetic implications. *Molecular biology and evolution.*
561 2006;23(2):279-91. Epub 2005/10/07. doi: 10.1093/molbev/msj029. PubMed PMID:
562 16207935.

563 35. Dong W, Xu C, Cheng T, Zhou S. Complete chloroplast genome of *Sedum sarmentosum*
564 and chloroplast genome evolution in Saxifragales. *PloS one.* 2013;8(10):e77965. Epub
565 2013/11/10. doi: 10.1371/journal.pone.0077965. PubMed PMID: 24205047; PubMed Central
566 PMCID: PMCPMC3799696.

567 36. Li B, Lin F, Huang P, Guo W, Zheng Y. Complete Chloroplast Genome Sequence of
568 *Decaisnea insignis*: Genome Organization, Genomic Resources and Comparative Analysis.
569 *Scientific reports*. 2017;7(1):10073. Epub 2017/09/01. doi: 10.1038/s41598-017-10409-8.
570 PubMed PMID: 28855603; PubMed Central PMCID: PMC5577308.

571 37. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular biology and*
572 *evolution*. 2007;24(8):1586-91. doi: 10.1093/molbev/msm088.

573 38. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and
574 nonsynonymous nucleotide substitutions. *Molecular biology and evolution*. 1986;3(5):418-26.
575 doi: 10.1093/oxfordjournals.molbev.a040410.

576 39. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7:
577 Improvements in Performance and Usability. *Molecular biology and evolution*.
578 2013;30(4):772-80. doi: 10.1093/molbev/mst010.

579 40. Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins
580 SE, et al. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Molecular*
581 *biology and evolution*. 2017;34(12):3299-302. Epub 2017/10/14. doi: 10.1093/molbev/msx248.
582 PubMed PMID: 29029172.

583 41. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
584 large phylogenies. *Bioinformatics (Oxford, England)*. 2014;30(9):1312-3. Epub 2014/01/24. doi:
585 10.1093/bioinformatics/btu033. PubMed PMID: 24451623; PubMed Central PMCID:
586 PMC3998144.

587 42. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes
588 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space.

589 Systematic biology. 2012;61(3):539-42. Epub 2012/02/24. doi: 10.1093/sysbio/sys029.
590 PubMed PMID: 22357727; PubMed Central PMCID: PMC3329765.

591 43. Li Z, Long H, Zhang L, Liu Z, Cao H, Shi M, et al. The complete chloroplast genome
592 sequence of tung tree (*Vernicia fordii*): Organization and phylogenetic relationships with other
593 angiosperms. Scientific reports. 2017;7(1):1869-. doi: 10.1038/s41598-017-02076-6. PubMed
594 PMID: 28500291.

595 44. Liu Q, Dou S, Ji Z, Xue Q. Synonymous codon usage and gene function are strongly
596 related in *Oryza sativa*. Bio Systems. 2005;80(2):123-31. Epub 2005/04/13. doi:
597 10.1016/j.biosystems.2004.10.008. PubMed PMID: 15823411.

598 45. Liu Q, Xue Q. Comparative studies on codon usage pattern of chloroplasts and their host
599 nuclear genes in four plant species. Journal of genetics. 2005;84(1):55-62. Epub 2005/05/07.
600 PubMed PMID: 15876584.

601 46. Powell W, Morgante M, McDevitt R, Vendramin GG, Rafalski JA. Polymorphic simple
602 sequence repeat regions in chloroplast genomes: applications to the population genetics of
603 pines. Proceedings of the National Academy of Sciences of the United States of America.
604 1995;92(17):7759-63. Epub 1995/08/15. PubMed PMID: 7644491; PubMed Central PMCID:
605 PMCPMC41225.

606 47. Yang Y, Dang Y, Li Q, Lu J, Li X, Wang Y. Complete chloroplast genome sequence of
607 poisonous and medicinal plant *Datura stramonium*: organizations and implications for genetic
608 engineering. PloS one. 2014;9(11):e110656. Epub 2014/11/05. doi:
609 10.1371/journal.pone.0110656. PubMed PMID: 25365514; PubMed Central PMCID:
610 PMCPMC4217734.

- 611 48. Tsudzuki T, Wakasugi T, Sugiura M. Comparative Analysis of RNA Editing Sites in Higher
612 Plant Chloroplasts 2001. 327-32 p.
- 613 49. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under
614 realistic evolutionary models. *Molecular biology and evolution*. 2000;17(1):32-43. Epub
615 2000/02/10. doi: 10.1093/oxfordjournals.molbev.a026236. PubMed PMID: 10666704.
- 616 50. Kim KA, Cheon KS, Jang SK, Yoo KO. Complete chloroplast genome sequence of
617 *Adenophora remotiflora* (Campanulaceae). Mitochondrial DNA Part A, DNA mapping,
618 sequencing, and analysis. 2016;27(4):2963-4. Epub 2015/06/30. doi:
619 10.3109/19401736.2015.1060461. PubMed PMID: 26119125.
- 620 51. Liu L, Wang Y, He P, Li P, Lee J, Soltis DE, et al. Chloroplast genome analyses and
621 genomic resource development for epilithic sister genera *Oreotrophe* and *Mukdenia*
622 (*Saxifragaceae*), using genome skimming data. *BMC genomics*. 2018;19(1):235. Epub
623 2018/04/06. doi: 10.1186/s12864-018-4633-x. PubMed PMID: 29618324; PubMed Central
624 PMCID: PMCPMC5885378.
- 625 52. Hu JM, Lavin M, Wojciechowski MF, Sanderson MJ. Phylogenetic systematics of the tribe
626 *Millettieae* (Leguminosae) based on chloroplast trnK/matK sequences and its implications for
627 evolutionary patterns in *Papilionoideae*. *Am J Bot*. 2000;87(3):418-30. doi: 10.2307/2656638.
628 PubMed PMID: 10719003.
- 629 53. Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, et al. ycf1, the most promising plastid DNA
630 barcode of land plants. *Scientific reports*. 2015;5:8348. Epub 2015/02/13. doi:
631 10.1038/srep08348. PubMed PMID: 25672218; PubMed Central PMCID: PMCPMC4325322.
- 632 54. de Queiroz LP, Pastore JF, Cardoso D, Snak C, de CLAL, Gagnon E, et al. A multilocus

633 phylogenetic analysis reveals the monophyly of a recircumscribed papilionoid legume tribe
634 Diocleae with well-supported generic relationships. *Mol Phylogenet Evol.* 2015;90:1-19. Epub
635 2015/05/03. doi: 10.1016/j.ympev.2015.04.016. PubMed PMID: 25934529.

636 55. Martin GE, Rousseau-Gueutin M, Cordonnier S, Lima O, Michon-Coudouel S, Naquin D,
637 et al. The first complete chloroplast genome of the Genistoid legume *Lupinus luteus*: evidence
638 for a novel major lineage-specific rearrangement and new insights regarding plastome
639 evolution in the legume family. *Annals of botany.* 2014;113(7):1197-210. Epub 2014/04/29. doi:
640 10.1093/aob/mcu050. PubMed PMID: 24769537; PubMed Central PMCID:
641 PMC4030815.

642 56. Guo X, Castillo-Ramirez S, Gonzalez V, Bustos P, Fernandez-Vazquez JL, Santamaria
643 RI, et al. Rapid evolutionary change of common bean (*Phaseolus vulgaris* L) plastome, and
644 the genomic diversification of legume chloroplasts. *BMC genomics.* 2007;8:228. Epub
645 2007/07/12. doi: 10.1186/1471-2164-8-228. PubMed PMID: 17623083; PubMed Central
646 PMCID: PMC1940014.

647 57. Tangphatsornruang S, Sangsrakru D, Chanprasert J, Uthapaisanwong P, Yoocha T,
648 Jomchai N, et al. The chloroplast genome sequence of mungbean (*Vigna radiata*) determined
649 by high-throughput pyrosequencing: structural organization and phylogenetic relationships.
650 *DNA research : an international journal for rapid publication of reports on genes and genomes.*
651 2010;17(1):11-22. Epub 2009/12/17. doi: 10.1093/dnares/dsp025. PubMed PMID: 20007682;
652 PubMed Central PMCID: PMC2818187.

653 **Supporting information**

654 **S1 Table. NCBI GenBank accession numbers used in this**

655 **study.**

656 **S2 Table. Codon usage frequency and RSCU in the Psoralea**
657 **corylifolia chloroplast genome.**

658 **S3 Table. Repeat sequences in the Psoralea corylifolia**
659 **chloroplast genome.**

660 **S4 Table. Tandem repeat sequences in the Psoralea corylifolia**
661 **chloroplast genome.**

662 **S5 Table. RNA editing predicted in Psoralea corylifolia**
663 **chloroplast genome.**

664 **S6 Fig. Phylogenetic relationship between Phaseoleae and**
665 **Psoraleeae based on *ycf1*, *matK*, *accD*, *ndhF* genes.**

- photosystem I
- photosystem II
- cytochrome b/f complex
- ATP synthase
- NADH dehydrogenase
- RubisCO large subunit
- RNA polymerase
- ribosomal proteins (SSU)
- ribosomal proteins (LSU)
- clpP, matK
- other genes
- hypothetical chloroplast reading frames (ycf)
- transfer RNAs
- ribosomal RNAs

Inverted Repeats

A

B

C

dN

dS

dN/dS