RESEARCH ARTICLE Legume Research- An International Journal, Volume 49 Special Issue: 77-88 (2026)

Assembly and Comparative Analysis of the *Onobrychis cyri* Mitochondrial Genome

Yansong Gou[1], Haijun Ding[2], Fengyan Yi[2], Xingyue Li[2], Chengyu Guo[2], Ruxue He[2], Chao Wen[2], Zhidan Shi[2], Cuiping Gao[1]

10.18805/LRF-981

ABSTRACT

Background: Mitochondrial genomic resources remain limited for *Onobrychis*. This study aimed to assemble and characterize the mitochondrial genome of *Onobrychis cyri* and assess its variation and relationships within a sampled set of Fabaceae mitogenomes.

Methods: Leaf DNA was sequenced using Illumina and Oxford Nanopore platforms. A hybrid assembly was annotated and analysed for repeats, codon usage, predicted RNA editing and putative plastid-derived sequences. Comparative analyses evaluated interspecific nucleotide variability, whole-genome similarity, mitochondrial phylogeny and exploratory pairwise Ka/Ks ratios.

Result: The circularized reference assembly was 359,147 bp with 44.67% GC content and contained 52 intact genes and two pseudogenes. It harboured 252 repeat elements and 474 predicted RNA-editing sites in 30 protein-coding genes. Among 34 shared loci, *rps7* showed the greatest interspecific nucleotide variability. Putative plastid-derived regions covered 5,173 bp (1.44%) of the mitochondrial assembly. *Onobrychis cyri* grouped with *Caragana spinosa* with 77% bootstrap support. Most pairwise Ka/Ks estimates (795/908; 87.6%) did not exceed one. The assembly expands mitochondrial genomic resources for *Onobrychis* and identifies candidate variable loci for further study. Physical genome configurations, predicted RNA-editing sites and candidate marker or selection signals require targeted validation.

Key words: Comparative genomics, Fabaceae, Mitochondrial genome, *Onobrychis*, Organelle DNA transfer, RNA editing.

INTRODUCTION

The genus *Onobrychis* comprises important perennial forage legumes whose members are valued in grassland production and ecological applications in temperate and semiarid regions. Morphological characters, nuclear ribosomal DNA internal transcribed spacer sequences and molecular markers have revealed complex relationships within the genus and substantial genetic variation among cultivated sainfoin landraces (Avci *et al*., 2016; Ertuş, 2021).

Differences in chromosome number and rDNA-site organization further indicate a complex history of nuclear-genome evolution in *Onobrychis* (Yucel *et al*., 2022). In contrast to the available nuclear and conventional marker studies, mitochondrial genomic resources for the genus remain limited, restricting analyses of cytoplasmic inheritance and evolutionary relationships.

Mitochondria are central to cellular respiration, oxidative phosphorylation and energy conversion in plants. Their genomes encode components associated with the respiratory chain and ATP synthesis and contribute to nucleo-cytoplasmic interactions, cytoplasmic inheritance and plant development. Plant mitochondrial genomes are usually maternally inherited and generally show relatively low nucleotide-substitution rates. They provide a source of phylogenetic and organellar evolutionary information that is independent of the nuclear genome (Gualberto and Newton, 2017; Mower *et al*., 2012).

Unlike the compact mitochondrial genomes of animals, plant mitogenomes vary markedly in size, gene order and physical configuration and contain extensive non-coding and repetitive DNA. Repeat-mediated recombination can produce genome rearrangements and structural isoforms, while DNA movement among the nucleus, plastids and mitochondria adds further compositional complexity (Gandini and Sanchez-Puerta, 2017; Stern and Palmer, 1984; Wang *et al*., 2007). C-to-U RNA editing can alter start or stop codons and restore conserved amino acids, making it an important post-transcriptional process in plant mitochondria (Takenaka *et al*., 2013). Gene content,

[1]Key Laboratory of Grassland Resources, Ministry of Education, College of Grassland Science, Inner Mongolia Agricultural University, Hohhot 010018, China.

[2]Grassland Research Institute, Inner Mongolia Academy of Agricultural and Animal Husbandry Sciences, Hohhot 010031, China.

Corresponding Author: Cuiping Gao, Key Laboratory of Grassland Resources, Ministry of Education, College of Grassland Science, Inner Mongolia Agricultural University, Hohhot 010018, China. Email: gcp1978@126.com

ORCIDs: 0009-0009-9132-8039, 0009-0002-5943-5036, 0009-0007-4141-3505, 0000-0003-1701-6757, 0009-0005-6411-735X, 0009-0002-4446-2365, 0009-0001-5487-3726, 0009-0008-4767-4666, 0000-0002-5776-3784

How to cite this article: Gou, Y., Ding, H., Yi, F., Li, X., Guo, C., He, R., Wen, C., Shi, Z. and Gao, C. (2026). Assembly and Comparative Analysis of the *Onobrychis cyri* Mitochondrial Genome. *Legume Research*. **49(Special Issue):** 77-88. doi: 10.18805/LRF-981.

Submitted: 31-08-2026 **Accepted:** 04-11-2026 **Online:** 18-11-2026

repeat composition, RNA editing, codon usage and interorganellar homology are therefore integral to the characterization of a plant mitochondrial genome.

Advances in high-throughput and long-read sequencing have accelerated mitochondrial genomics in Fabaceae. Studies of *Vigna radiata*, *Vicia faba* and *Glycine max* have revealed pronounced differences among legume mitogenomes in size, repeat content and shared non-coding DNA (Alverson *et al.*, 2011; Chang *et al.*, 2013; Choi *et al.*, 2019; Negruk, 2013). Using combined NovaSeq and PromethION data, Chen *et al.* (2025) assembled the *Lotus corniculatus* mitochondrial genome and analyzed gene content, RNA editing, codon usage, plastid-mitochondrial homology and phylogenetic relationships, providing a comprehensive framework for mitogenomic analysis of forage legumes. Long-read-supported analysis of *Caragana spinosa* also demonstrated multiple repeat-mediated organellar configurations, showing that a circularized reference sequence does not necessarily represent the only physical molecule present *in vivo* (Zhou *et al.*, 2023).

Mitochondrial genome characterization of *Onobrychis cyri* Grossh. can help address the limited organellar resources available for this genus. Short reads alone may not bridge long repetitive regions, whereas long reads provide continuity but still benefit from correction with highly accurate short reads. Combining Illumina and Oxford Nanopore data can therefore generate a mitochondrial reference suitable for annotation and comparative analysis. Comparison with related Fabaceae mitogenomes can further assess gene conservation, interspecific sequence variability, structural rearrangement and the phylogenetic position of *O. cyri*.

In the present study, we combined short- and long-read data to assemble a mitochondrial reference genome of *O. cyri* and characterized its gene content, repeat composition, codon usage and predicted RNA-editing sites. We further compared the assembly with representative Fabaceae mitogenomes, assessed whole-genome homology and interspecific nucleotide variability of homologous genes, identified putative plastid-derived sequences, reconstructed phylogenetic relationships from mitochondrial coding sequences and explored selective constraints using pairwise Ka/Ks ratios. The results expand organellar genomic resources for *Onobrychis* and provide a basis for comparative genomics, evolutionary studies and the development of candidate molecular markers.

MATERIALS AND METHODS

Plant material, DNA library construction and sequencing

Leaf tissue of *O. cyri* Grossh. was collected in Hohhot, Inner Mongolia Autonomous Region, China. Total genomic DNA was isolated from the leaf tissue using a cetyltrimethylammonium bromide (CTAB)-based method (Doyle and Doyle, 1987). DNA integrity, purity and concentration were assessed before construction of Illumina short-read and Oxford Nanopore long-read libraries.

For the short-read library, genomic DNA was randomly fragmented by ultrasonication, followed by fragment purification, end repair, 32 adenylation, adapter ligation, agarose-gel size selection, PCR enrichment and library quality assessment. Paired-end sequencing was performed on an Illumina NovaSeq 6000 platform. Raw reads were processed with fastp v0.23.4 (Chen *et al.*, 2018) to remove adapter and primer sequences and to discard reads with an average quality score below Q5 or more than five undetermined bases. For the long-read library, genomic DNA was randomly sheared, large fragments were enriched and purified with magnetic beads and size-selected by gel recovery, followed by damage repair, end repair, A-tailing and SQK-LSK109 adapter ligation. After library quantification, real-time single-molecule sequencing was performed on an Oxford Nanopore PromethION platform. Reads were filtered with Filtlong v0.2.1 using -min_length 1000 and -min_mean_q 7.

After filtering, the Illumina dataset contained 59,755,144 read pairs (17,926,543,200 bp), with Q20 and Q30 proportions of 97.21% and 92.20%, respectively and a GC content of 35.80%. The Nanopore dataset contained 2,626,129 reads (15,847,551,652 bp), with a mean read length of 6,034 bp and a read N50 of 17,693 bp.

Mitochondrial genome assembly and structural inspection

Conserved plant mitochondrial protein-coding and ribosomal RNA (rRNA) sequences were used to enrich candidate Nanopore reads. Raw long reads were first aligned to the reference gene sequences with minimap2 v2.1 (Li, 2018). Alignments longer than 50 bp were screened to identify seed reads with broad, high-quality coverage of core genes; raw reads overlapping the expanding seed set by more than 1 kb were then added iteratively. The enriched long reads were corrected with Canu (Koren *et al.*, 2017). Illumina reads were aligned to the corrected sequences with Bowtie2 v2.3.5.1 (Langmead and Salzberg, 2012) and the aligned short reads and corrected long reads were jointly assembled with Unicycler v0.4.8 under default settings (Wick *et al.*, 2017). The assembly graph was inspected with Bandage v0.8.1 (Wick *et al.*, 2015). Corrected long reads were subsequently remapped to the assembled contigs with minimap2 and branch directions were examined manually to obtain the final circularized reference sequence.

Genome annotation

Protein-coding and rRNA genes were initially annotated by BLAST comparison with published plant mitochondrial gene sequences, followed by manual adjustment of gene boundaries using related species as references. Transfer RNA (tRNA) genes were detected with tRNAscan-SE (Chan and Lowe, 2019). Candidate open reading frames (ORFs) of at least 102 bp were identified with NCBI ORF Finder;

redundant ORFs and those overlapping known genes were removed. Candidate ORFs longer than 300 bp were compared with the nr database according to the archived analytical workflow. ORF predictions were summarized separately from the manually curated set of intact genes and pseudogenes. The mitochondrial genome map was generated with OGDRAW (Lohse *et al.*, 2013).

Repeat-sequence and codon-usage analyses

Simple sequence repeats were detected with MISA v1.0 using minimum repeat numbers of 10, 5, 4, 3, 3 and 3 for mono-, di-, tri-, tetra-, penta- and hexanucleotide motifs, respectively (Thiel *et al.*, 2003). Tandem repeats were identified with TRF v4.09 using the parameters 2 7 7 80 10 50 2000 -f -d -m (Benson, 1999). Dispersed repeats were detected with BLASTN v2.10.1 using -word_size 7 and an E-value threshold of 1×10^{-5}; redundant hits and tandem-repeat matches were removed (Camacho *et al.*, 2009). The genomic distribution of repeat sequences was visualized with Circos v0.69-5 (Krzywinski *et al.*, 2009).

A project analysis script was used to retain non-redundant protein-coding sequences and tabulate codon frequencies. Relative synonymous codon usage (RSCU) was calculated as the observed frequency of a codon divided by its expected frequency under equal use of synonymous codons. Stop codons were excluded from comparisons of sense-codon bias and the single-codon amino acids Met and Trp were assigned an RSCU value of 1.

Prediction of RNA-editing sites

Potential C-to-U RNA-editing sites in mitochondrial protein-coding genes were predicted with the PmtREP web service (http://cloud.genepioneer.com:9929/#/tool/alltool/detail/336) using a multiple-sequence-alignment-based approach. Predicted codon and amino-acid changes were classified according to transitions between hydrophilic and hydrophobic residues and the formation of stop codons. All RNA-editing sites reported here are computational predictions and were not validated using transcriptome or cDNA data.

Plastid-mitochondrial homologous-sequence analysis

Homologous sequences between the chloroplast and mitochondrial assemblies generated in the same project were identified with BLAST using an E-value threshold of 1×10^{-5} and a minimum nucleotide identity of 70%. Homologous fragments were visualized with Circos v0.69-5 (Krzywinski *et al.*, 2009). Chloroplast protein sequences were also aligned to the mitochondrial genome at an E-value threshold of 1×10^{-5} to identify coding regions represented completely or partially within homologous fragments. These matches were classified as putative plastid-derived sequences and were not considered independent evidence of transfer direction.

Comparative genomics and interspecific nucleotide variability

To compare mitochondrial genome structure and sequence variation across selected evolutionary lineages within Papilionoideae, the *O. cyri* assembly was compared with publicly available mitochondrial genomes from eight species: *Vicia faba*, *Glycyrrhiza uralensis*, *G. glabra*, *Phyllolobium chinense*, *Phaseolus vulgaris*, *Medicago sativa*, *Oxytropis splendens* and *Caragana spinosa*. These species span seven genera and include several lineages within the inverted repeat-lacking clade (IRLC), to which *Onobrychis* also belongs, together with *P. vulgaris* outside the IRLC. The two Glycyrrhiza species additionally provide a congeneric comparison alongside comparisons among genera. Their GenBank accession numbers were KC189947.1, MW013322.1, MW448466.1, MZ127833.1, NC_045135.1, ON782580.1, OP345444.1 and OQ785640.1, respectively. Whole-genome alignments were generated with nucmer v4.0.0beta2 (-maxmatch) and forward and reverse homologous fragments were displayed in dot plots. Homologous genes from the comparative taxon set were aligned with MAFFT v7.427 (-auto; Katoh and Standley, 2013) and nucleotide diversity (Pi) and the number of variable sites were calculated with DnaSP v5 (Librado and Rozas, 2009). Because the dataset comprised different species rather than multiple individuals within a population, Pi was interpreted only as a measure of interspecific sequence variability.

Phylogenetic and selective-constraint analyses

The phylogenetic analysis included 24 taxa and 30 mitochondrial protein-coding genes present in at least 70% of the sampled taxa. Individual genes were aligned with MAFFT v7.427 (-auto) and concatenated. The concatenated alignment was trimmed with trimAl v1.4.rev15 (-gt 0.7), as described by Capella-Gutiérrez *et al.* (2009). After substitution-model selection with jModelTest v2.1.10 (Darriba *et al.*, 2012), a maximum-likelihood tree was reconstructed in RAxML v8.2.10 under the GTRGAMMA model with 1,000 bootstrap replicates (Stamatakis, 2014). For selective-constraint analysis, homologous protein-coding genes were aligned with MAFFT and Ka, Ks and Ka/Ks values were estimated using the MLWL method in KaKs_Calculator v2.0 (Wang *et al.*, 2010). Values of Ka/Ks > 1 were treated only as candidate signals, not as independent statistical evidence of positive selection.

Results and Discussion

Sequencing, assembly and genome organization

The Illumina dataset contained 59.76 million paired reads and 17.93 Gb of sequence. The Q20 and Q30 proportions were 97.21% and 92.20%, respectively and the overall GC content was 35.80%. Long-read sequencing yielded 2.63 million reads and 15.85 Gb, with an N50 read length of 17.69 kb. The mitochondrial assembly was presented as a single 359,147-bp circularized reference sequence with a GC content of 44.67% (Table 1; Fig 1 and 2). The Nanopore read-length distribution was strongly right-skewed, with most reads concentrated in shorter size classes and a tail extending beyond 90 kb. This dataset

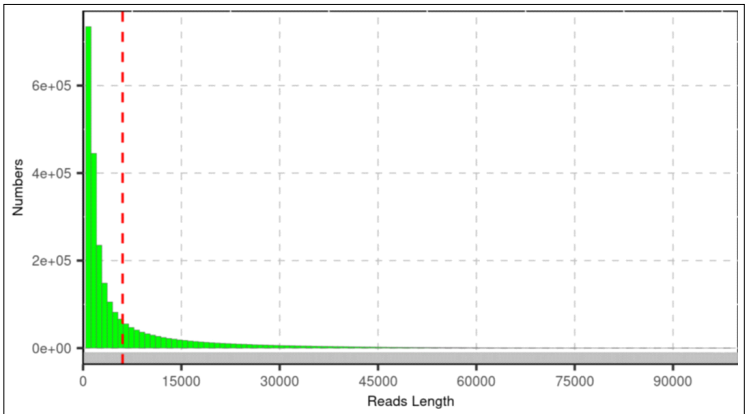
supplied long-range information for assembly, but physical genome configurations require direct junction-spanning validation.

The annotation comprised 52 intact genes (31 protein-coding, 18 tRNA and three rRNA genes) and two pseudogenes, representing 54 annotated loci in total.

Table 1: Sequencing and assembly characteristics of *Onobrychis cyri.*

Dataset/feature	Read count /assembly	Bases	Quality/length	GC (%)
Illumina paired-end	59,755,144	17.93 Gb	Q20 97.21%; Q30 92.20%	35.80
Long-read sequencing	2,626,129	15.85 Gb	Mean 6,034 bp; N50 17,693 bp	n.a.
Mitochondrial assembly	1 reference sequence	359,147 bp	Circularized representation	44.67

Illumina counts are read pairs; Oxford Nanopore counts are individual reads. Gb, gigabases; bp, base pairs; Q20 and Q30, percentages of bases with Phred quality scores of at least 20 and 30, respectively; N50, read length at which half of the total sequenced bases occur in reads of that length or longer; n.a., not available. The assembly is a circularized reference representation.



The red dashed line marks the mean read length of 6,034 bp.

Fig 1: Length distribution of filtered Oxford Nanopore reads.



Genes outside and inside the circle are transcribed clockwise and counterclockwise, respectively. The circle is a reference representation and does not establish the physical configuration of mitochondrial DNA *in vivo.*

Fig 2: Circular map of the Onobrychis cyri mitochondrial reference assembly.

The protein-coding complement included five ATP synthase genes, four cytochrome c biogenesis genes, three cytochrome c oxidase genes, nine NADH dehydrogenase genes, six small-subunit ribosomal protein genes, *rpl5*, *cob*, *matR* and *mttB*. The *trnM-CAU* gene occurred in four copies. Introns were annotated in *ccmFc*, *nad1*, *nad2*, *nad4*, *nad5*, *nad7*, *rps1*, *rps3* and *trnL-UAG*. Two ribosomal protein genes, *rpl16* and *sdh4*, were annotated as pseudogenes. The minimum-102-bp ORF scan returned 1,836 candidates, which were excluded from the curated gene inventory because they lacked independent expression or conservation evidence.

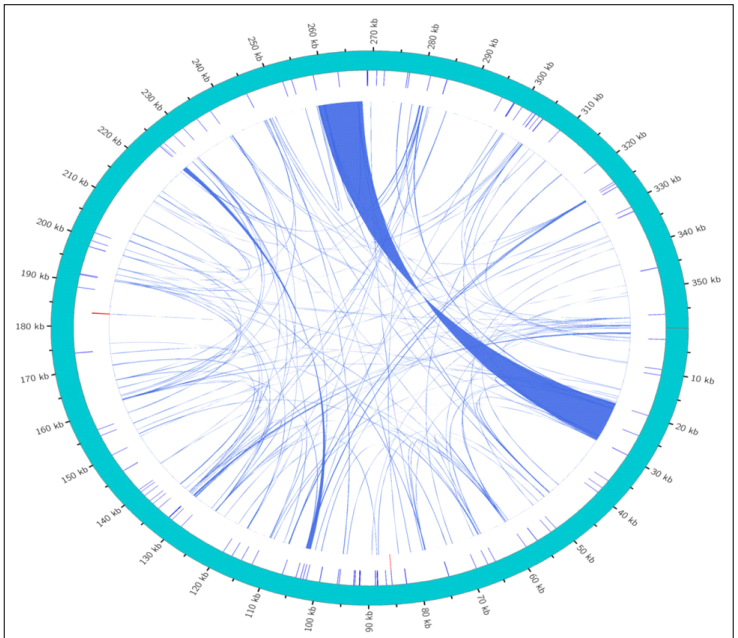
Most protein-coding genes used canonical ATG start codons. The *nad1* and *nad4L* annotations contained ACG start codons predicted to be converted to AUG through RNA editing. Several genes terminated with TGA, while *ccmFc* contained a genomic CGA codon predicted to become TGA after editing.

The 359.15 kb *O. cyri* mitochondrial genome assembly falls within the broad size range reported for Fabaceae, whose mitochondrial genomes show ancient and recent shifts in size and content (Choi *et al*., 2019). Its core coding complement is consistent with other legume mitogenomes, whereas pseudogenization of *rpl16* and *sdh4* and the copy number of *trnM* may reflect lineage-specific gene loss, transfer and duplication (Choi *et al*., 2019; Gualberto and Newton, 2017). The circularized sequence is a convenient reference representation rather than proof of one genome-sized molecule: plant mitochondrial DNA frequently occurs as a dynamic mixture of recombination-derived configurations (Kozik *et al*., 2019). Multiple repeat-mediated configurations have also been validated in the closely compared legume *C. spinosa* (Zhou *et al*., 2023).

The hybrid sequencing strategy combined Nanopore long-range information with high-Q30 Illumina reads. Reference-guided read enrichment may under-represent highly divergent or lineage-specific mitochondrial sequences. Current guidance on land-plant mitogenome assembly emphasizes hybrid short- and long-read evidence together with explicit evaluation of alternative configurations (Štorchová and Krüger, 2024). Detailed assembly-graph evidence, coverage summaries and junction-spanning read counts are not presented here. Consequently, the circularized sequence should be interpreted as a reference representation; the physical configurations present in vivo remain to be resolved.

Repeat sequences and structural variation

The genome contained 252 repeat elements, including 99 SSRs, four tandem repeats and 149 dispersed repeats (Table 2). Among SSRs, tetranucleotide motifs were the most frequent class (30), followed by mononucleotide (27), dinucleotide (19), trinucleotide (17), pentanucleotide (4) and hexanucleotide motifs (2). The four tandem repeats had repeat units of 6-36 bp and approximately 1.9-4.2 copies. Dispersed-repeat alignments ranged from 29 to 9,778 bp; the longest was a 9,778-bp direct repeat with 100% identity. Repeat features occurred throughout the circularized reference and several dense dispersed-repeat link bundles identify regions that could participate in repeat-mediated recombination (Fig 3).



Internal links connect dispersed repeat pairs; short peripheral marks denote local repeat features.

Fig 3: Distribution of repeat sequences across the Onobrychis cyri mitochondrial reference assembly.

The abundance of dispersed repeats provides a substrate for homologous recombination and structural isoforms, a major source of plant mitochondrial genome dynamism (Alverson *et al.*, 2011; Gualberto and Newton, 2017; Kozik *et al.*, 2019). The 99 SSRs also define candidate mitochondrial marker loci, although organellar specificity must be established against the nuclear and chloroplast genomes before these loci are used in marker-assisted genetic improvement programmes.

The repeat map shows connections distributed across the reference sequence, including several dense bundles. These regions identify candidates for targeted examination of alternative configurations. Repeat occurrence alone does not demonstrate active recombination or establish the relative abundance of physical isoforms.

Codon usage

Codon usage was biased toward several A- or U-ending codons. The highest RSCU values among synonymous sense codons included UUA for leucine (1.5494), GCU for alanine (1.5380), UAU for tyrosine (1.5369), CAU for histidine (1.4825) and CAA for glutamine (1.4786). Codon-level values grouped by amino-acid family are shown in Fig 4.

Codon usage in *O. cyri* favored several A- and U-ending synonymous codons, consistent with an influence of nucleotide composition on mitochondrial coding sequences. RSCU describes relative use within synonymous families; it does not by itself distinguish mutation pressure from selection or demonstrate translational optimization.

The RSCU profile shows broad within-family usage differences for synonymous codons of arginine, leucine and serine, whereas the one-codon families Met and Trp carry no information about synonymous preference. Because RNA editing can alter codons after transcription, these values describe genomic codon usage rather than edited transcripts.

Predicted RNA editing

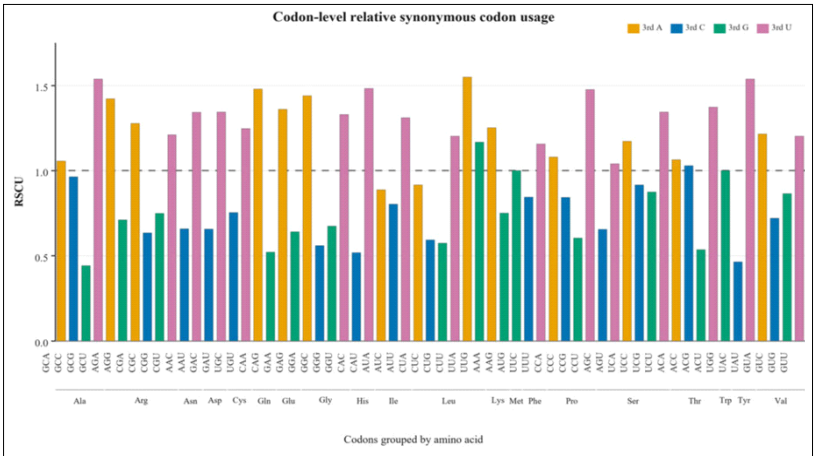
A total of 474 RNA-editing sites were predicted in 30 protein-coding genes. The largest numbers occurred in *nad4* (41 sites), *ccmFn* (35), *nad7* (30), *ccmB* and *nad2* (29 each) and *ccmC* (28). Edits predicted to convert hydrophilic amino acids to hydrophobic amino acids formed the largest functional class (227 sites; 47.89%). Hydrophobic-to-hydrophobic changes accounted for 150 sites (31.65%), hydrophilic-to-hydrophilic changes for 62 (13.08%) and hydrophobic-to-hydrophilic changes for 34 (7.17%). One predicted edit converted CGA to the stop codon UGA. The distribution among genes is shown in Fig 5.

The 474 predicted editing sites are consistent with extensive post-transcriptional modification of mitochondrial transcripts in flowering plants (Takenaka *et al.*, 2013). NADH dehydrogenase and cytochrome c biogenesis genes carried particularly high editing loads and many changes

Table 2: Genome annotation, repeat content and predicted RNA-editing summary.

Category	Number	Category	Number
Intact protein-coding genes	31	SSR loci	99
tRNA genes	18	Tandem repeats	4
rRNA genes	3	Dispersed repeats	149
Pseudogenes	2	Total repeats	252
Genes with predicted editing	30	Predicted editing sites	474

Counts refer to the curated gene annotations and repeat features reported in this study. SSR: Simple sequence repeat; tRNA: Transfer RNA; rRNA: Ribosomal RNA. RNA-editing sites are computational predictions.



Codons are grouped by amino-acid family and colored by the third nucleotide. The dashed line indicates RSCU = 1; stop codons were excluded.

Fig 4: Codon-level relative synonymous codon usage in mitochondrial protein-coding genes.

increased amino-acid hydrophobicity, consistent with restoration of conserved residues in membrane-associated respiratory proteins. The predicted stop codon in *ccmFc* and ACG-to-AUG start-codon changes in *nad1* and *nad4L* are functionally notable; without matched RNA-sequencing or cDNA data, they remain computational predictions.

Interspecific nucleotide variability

Interspecific nucleotide variability differed among the 34 evaluated loci (mean Pi = 0.01706). The *rps7* locus had the highest Pi value (0.17157), followed by *nad9* (0.04445), *atp6* (0.03532), *atp1* (0.03434) and *atp8* (0.02795) (Table 3; Fig 6). In contrast, *atp9* was the most conserved locus (Pi = 0.00198), followed by *nad7*, *nad4*, *mttB* and *nad4L*. These estimates describe divergence among sampled species and should not be interpreted as within-population diversity.

The interspecific Pi analysis identified *rps7*, *nad9*, *atp6*, *atp1* and *atp8* as the most variable loci in the sampled taxa. The exceptional *rps7* value may be influenced by its short alignment and by gene absence or divergence in some plant lineages. The more moderately variable *nad9*, *atp6*, *atp1* and *atp8* loci are candidates for comparative marker development, but their utility for population-level studies requires validation in multiple *O. cyri* individuals.

Whole-genome similarity and rearrangement

Whole-genome alignments showed that 35.48-51.63% of the *O. cyri* mitochondrial reference was represented in homologous regions with the eight comparison genomes. The greatest reference coverage was observed with *G. glabra* (51.63%) and *G. uralensis* (51.62%), followed by *C. spinosa* (49.99%) and *V. faba* (47.89%). The lowest coverage occurred with *P. vulgaris* (35.48%), illustrating extensive divergence in non-coding and structurally variable regions despite conservation of core mitochondrial genes. The pairwise dot plots contained many short, discontinuous matches in both orientations rather than a single genome-length diagonal (Fig 7), consistent with rearrangement and lineage-specific intergenic turnover.

The fragmented forward- and reverse-orientation matches in the collinearity plots provide structural context for the 35.48-51.63% aligned proportions. Conserved mitochondrial blocks have been reordered against a background of rapidly changing intergenic DNA, allowing core coding genes to retain a coherent phylogenetic signal despite poor whole-genome collinearity. The dot plots describe the resulting structural differences but do not resolve a unique rearrangement pathway.

Putative plastid-derived mitochondrial sequences

The *O. cyri* chloroplast genome used for comparison was 121,932 bp. Homologous chloroplast regions totaled 5,208 bp (4.27% of the chloroplast genome), corresponding to 5,173 bp or 1.44% of the mitochondrial assembly (Fig 8). Five tRNA genes-*trnW-CCA*, *trnD-GUC*, *trnN-GUU*, *trnH-GUG* and *trnM-CAU*- were completely located within

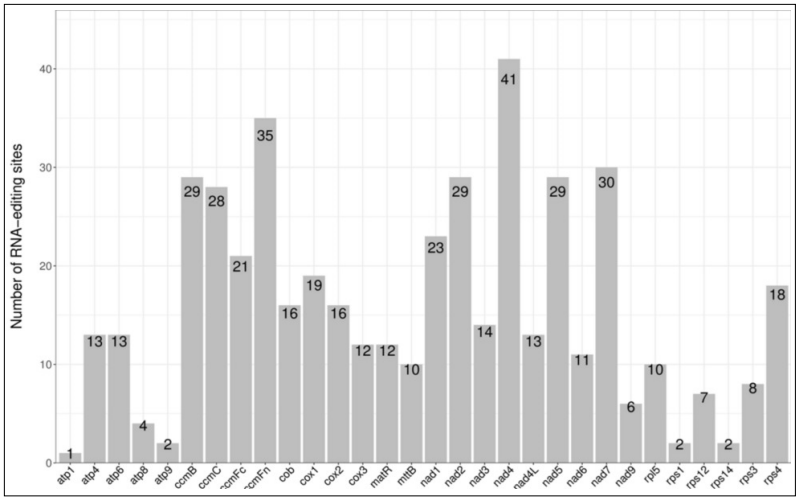


Fig 5: Number of predicted RNA-editing sites in mitochondrial protein-coding genes of *Onobrychis cyri*.

Table 3: Mitochondrial loci with the highest interspecific nucleotide variability.

Rank	Gene	Pi	Variable sites	Alignment length (bp)
1	rps7	0.17157	35	309
2	nad9	0.04445	57	675
3	atp6	0.03532	73	888
4	atp1	0.03434	177	1,527
5	atp8	0.02795	47	489

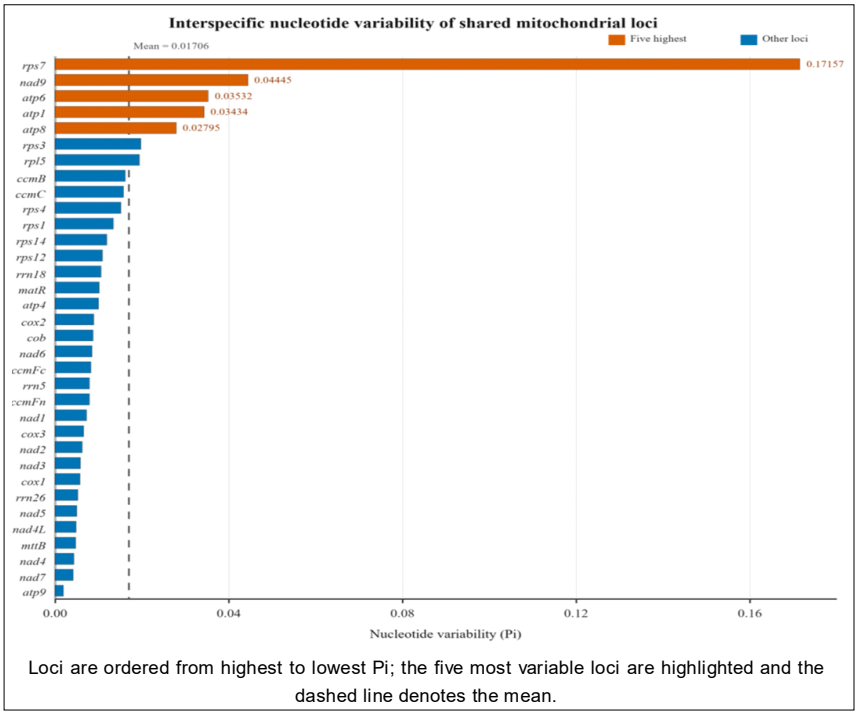
Pi, interspecific nucleotide variability calculated from the sampled species alignments. Variable sites and alignment lengths refer to the same locus alignments. These five loci are a subset of the 34 loci analysed and do not represent within-population diversity.

homologous chloroplast regions. Partial chloroplast-gene matches included *psaA*, *psaB*, *trnI-GAU*, *trnA-UGC*, *rrn16*, *petG*, *psaJ* and *rrn23*. On the mitochondrial side, the corresponding homologous regions included complete copies of the five tRNA genes and partial matches involving *rrn18* and *rrn26*. These similarities identify putative plastid-derived regions but do not by themselves establish transfer direction or functionality.

Approximately 1.44% of the *O. cyri* mitochondrial assembly was homologous to chloroplast DNA. Complete plastid-like *trnW*, *trnD*, *trnN*, *trnH* and *trnM* genes are consistent with the persistence of plastid-derived tRNA genes in seed-plant mitochondria (Gandini and Sanchez-Puerta, 2017; Stern and Palmer, 1984; Wang *et al*., 2007). Partial matches to *psaA* and *psaB* are more consistent with non-functional remnants than intact mitochondrial genes. At candidate transfer boundaries, reliable organellar assignment requires long reads anchored uniquely in mitochondrial flanking sequence; sequence similarity alone does not identify transfer direction.

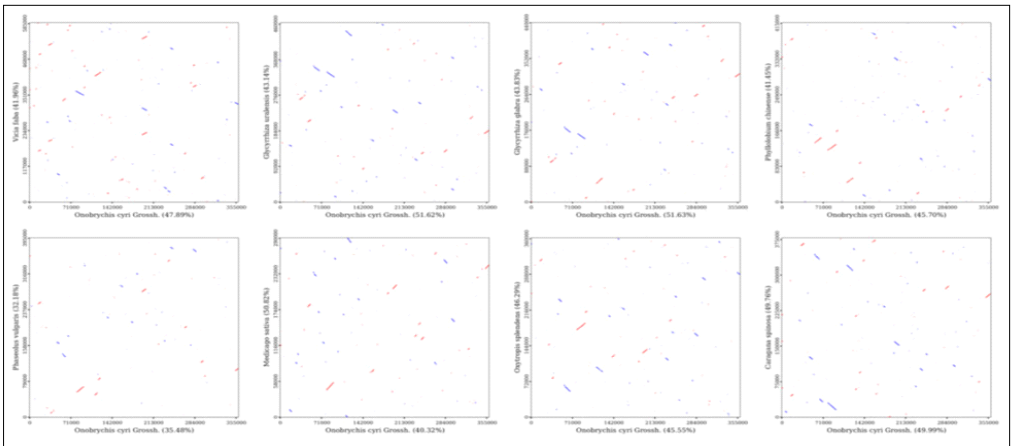
Phylogenetic relationships

The maximum-likelihood analysis recovered the major relationships represented by the sampled Fabaceae



Loci are ordered from highest to lowest Pi; the five most variable loci are highlighted and the dashed line denotes the mean.

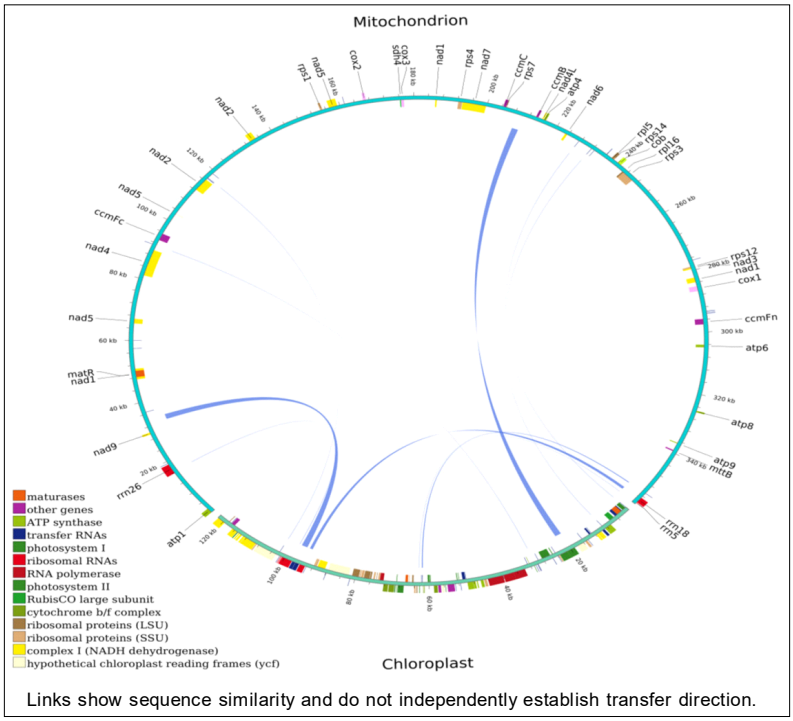
Fig 6: Interspecific nucleotide variability (Pi) of 34 mitochondrial loci in representative Fabaceae species.



Red and blue segments denote forward- and reverse-orientation matches, respectively. Panels follow the comparison species listed in the Materials and Methods, from left to right across the top row and then the bottom row.

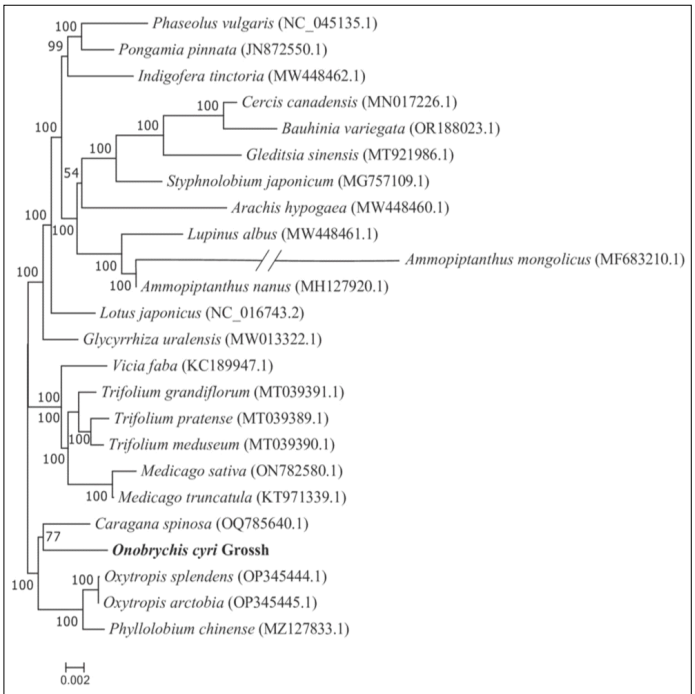
Fig 7: Pairwise mitochondrial genome collinearity between *Onobrychis cyri* and eight Fabaceae species.

mitogenomes (Fig 9). *O. cyri* was placed as sister to *C. spinosa* with 77% bootstrap support and this pair was associated with *Phyllolobium* and *Oxytropis* lineages within a strongly supported broader clade. The 77% support identifies the *O. cyri*-*C. spinosa* relationship as moderately supported within the current taxon sample.



Links show sequence similarity and do not independently establish transfer direction.

Fig 8: Homologous sequence links between the chloroplast and mitochondrial genomes of *Onobrychis cyri.*



Values at nodes are bootstrap percentages from 1,000 replicates. The focal *O. cyri* sequence has been assigned GenBank accession number QB199901. The scale bar indicates substitutions per site.

Fig 9: Maximum-likelihood phylogeny inferred from concatenated mitochondrial coding sequences.

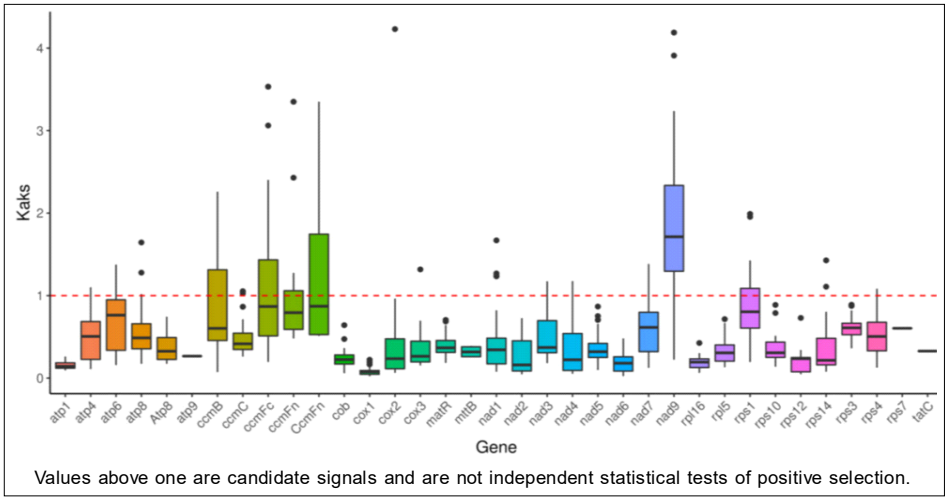


Fig 10: Distribution of exploratory pairwise Ka/Ks ratios among mitochondrial protein-coding genes.

The placement of *O. cyri* with *C. spinosa* is supported by the sampled mitochondrial coding sequences. The 77% bootstrap value and limited representation of close relatives define this relationship as provisional rather than a definitive genus-level taxonomic result.

Exploratory selective-constraint analysis

The Ka/Ks dataset comprised 908 pairwise gene estimates. The mean ratio was 0.544 and 795 estimates (87.6%) did not exceed one, consistent with predominantly constrained evolution in this descriptive dataset. A total of 113 estimates exceeded one, most frequently in *nad9* (29 comparisons), *ccmB* (14), *ccmFn* and *ccmFc* (13 each), *rps1* (10) and *atp6* (7). The largest value was 4.23084 for *cox2*. Values above one were treated as candidate signals because estimates can be inflated by low Ks values, short sequences or local alignment uncertainty (Fig 10).

Most pairwise Ka/Ks ratios did not exceed one, a pattern consistent with broad functional constraint on mitochondrial protein-coding genes. Ratios above one in *nad9*, ccm genes, *rps1* and *atp6* identify candidate deviations from this background, but pairwise ratios alone do not establish adaptive evolution and remain sensitive to small Ks values, short alignments, non-independence among comparisons and alignment uncertainty.

Conclusion

The *O. cyri* mitochondrial assembly was represented by a 359,147-bp circularized sequence containing 52 intact genes and two pseudogenes. The assembly contained 252 repeat elements, 474 predicted RNA-editing sites and 5.17 kb of putative plastid-derived sequence. Comparative analyses identified candidate variable mitochondrial genes and placed *O. cyri* with *C. spinosa* in the sampled phylogeny. The sequence expands organellar genomic resources for *Onobrychis*; structural configurations, predicted edits and candidate marker or selection signals require targeted validation.

Acknowledgement

The authors thank all colleagues who assisted with sample collection and data organization.

Author contributions

Yansong Gou prepared the original draft and organized the genomic analyses. Haijun Ding, Fengyan Yi, Xingyue Li, Chengyu Guo, Ruxue He, Chao Wen and Zhidan Shi contributed to research coordination, data interpretation and manuscript revision. Cuiping Gao supervised the study, acquired funding and critically revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the 2025 Project of the Key Laboratory of Grass Seed Innovation and Sustainable Utilization of Grassland Resources of Inner Mongolia Autonomous Region, “Research on Breeding High-quality Forage Varieties and High-yield, Stress-resistant Cultivation Technologies” (2025KYPT0033) and the Inner Mongolia Autonomous Region “Open Competition Mechanism to Select the Best Candidates” Project (2022JBGS00400204).

Disclaimers

The views and conclusions expressed in this article are solely those of the authors and do not necessarily represent the views of their affiliated institutions. The authors are responsible for the accuracy and completeness of the information provided.

Ethics and informed consent

This study involved plant material and did not involve human participants or animals. Informed consent was therefore not applicable.

Availability of data and materials

The mitochondrial genome assembly generated in this study has been assigned GenBank accession number QB199901. Additional data supporting the findings are available from the corresponding author upon reasonable request.

Use of artificial intelligence

No artificial intelligence tools were used for writing, editing, data analysis, or formatting of this manuscript. All content was prepared solely by the author.

Conflict of interest

The authors declare no conflict of interest.

REFERENCES

Alverson, A.J., Zhuo, S., Rice, D.W., Sloan, D.B. and Palmer, J.D. (2011). The mitochondrial genome of the legume *Vigna radiata* and the analysis of recombination across short mitochondrial repeats. *PLoS ONE*. **6(1):** e16404. https://doi.org/10.1371/journal.pone.0016404.

Avci, S., Tekin, N., Sancak, C., Özcan, S. and Marangi, A.O. (2016). Phylogenetic relationship of some *Onobrychis* taxa naturally grown in Turkey based on morphology and nuclear ribosomal DNA ITS sequences. *Legume Research*. **39(5):** 665-673. doi: 10.18805/lr.v0iof.11184.

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

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*. **10:** 421. https://doi.org/10.1186/1471-2105-10-421.

Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. **25(15):** 1972-1973. https://doi.org/10.1093/bioinformatics/btp 348.

Chan, P.P. and Lowe, T.M. (2019). tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. In: Gene Prediction: Methods and Protocols. [Kollmar, M. (ed.)], Methods in Molecular Biology, Vol. 1962 (pp. 1-14). New York, NY: Humana Press. https://doi.org/10.1007/978-1-4939-9173-0_1.

Chang, S., Wang, Y., Lu, J., Gai, J., Li, J., Chu, P., Guan, R. and Zhao, T. (2013). The mitochondrial genome of soybean reveals complex genome structures and gene evolution at intercellular and phylogenetic levels. *PLoS ONE*. **8(2):** e56502. https://doi.org/10.1371/journal.pone.0056502.

Chen, S., Zhou, Y., Chen, Y. and Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. **34(17):** i884-i890. https://doi.org/10.1093/bioinformatics/bty560.

Chen, X., Wu, Z., Yang, Y., Tao, Q., Na, N., Wan, W., Tian, C., Gong, W. and Li, Z. (2025). The complete mitochondrial genome and phylogenetic analysis of *Lotus corniculatus* (Fabaceae, Papilionoideae). *Frontiers in Plant Science*. **16:** 1555595. https://doi.org/10.3389/fpls.2025.1555595.

Choi, I.S., Schwarz, E.N., Ruhlman, T.A., Khiyami, M.A., Sabir, J.S.M., Hajarah, N.H., Sabir, M.J., Rabah, S.O. and Jansen, R.K. (2019). Fluctuations in Fabaceae mitochondrial genome size and content are both ancient and recent. *BMC Plant Biology*. **19:** 448. https://doi.org/10.1186/s12870-019-2064-8.

Darriba, D., Taboada, G.L., Doallo, R. and Posada, D. (2012). jModel Test 2: More models, new heuristics and parallel computing. *Nature Methods*. **9(8):** 772. https://doi.org/10.1038/nmeth. 2109.

Doyle, J.J. and Doyle, J.L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*. **19(1):** 11-15.

Ertuş, M.M. (2021). Genetic relationship in sainfoin (*Onobrychis viciifolia*) landraces cultivated East Anatolia by using RAPD and ISSR markers. *Legume Research*. **44(8):** 888-893. doi: 10.18805/lr-614.

Gandini, C.L. and Sanchez-Puerta, M.V. (2017). Foreign plastid sequences in plant mitochondria are frequently acquired *via* mitochondrion-to-mitochondrion horizontal transfer. *Scientific Reports*. **7:** 43402. https://doi.org/10.1038/srep43402.

Gualberto, J.M. and Newton, K.J. (2017). Plant mitochondrial genomes: Dynamics and mechanisms of mutation. *Annual Review of Plant Biology*. **68:** 225-252. https://doi.org/10.1146/annurev-arplant-043015-112232.

Katoh, K. and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*. **30(4):** 772-780. https://doi.org/10.1093/molbev/mst010.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*. **27(5):** 722-736. https://doi.org/10.1101/gr.215087.116.

Kozik, A., Rowan, B.A., Lavelle, D., Berke, L., Schranz, M.E., Michelmore, R.W. and Christensen, A.C. (2019). The alternative reality of plant mitochondrial DNA: One ring does not rule them all. *PLoS Genetics*. 15(8): e1008373. https://doi.org/10.1371/journal.pgen.1008373.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*. **19(9):** 1639-1645. https://doi.org/10.1101/gr.092759.109.

Langmead, B. and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*. **9(4):** 357-359. https://doi.org/10.1038/nmeth.1923.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*. **34(18):** 3094-3100. https://doi.org/10.1093/bioinformatics/bty191.

Librado, P. and Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*. **25(11):** 1451-1452. https://doi.org/10.1093/bioinformatics/btp 187.

Lohse, M., Drechsel, O., Kahlau, S. and Bock, R. (2013). Organellar GenomeDRAW-a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research*. **41(W1):** W575-W581. https://doi.org/10.1093/nar/gkt289.

Mower, J.P., Sloan, D.B. and Alverson, A.J. (2012). Plant mitochondrial genome diversity: The genomics revolution. In: *Plant Genome Diversity Volume 1: Plant Genomes, their Residents and their Evolutionary Dynamics* [Wendel, J.F., Greilhuber, J., Doležel, J. and Leitch, I.J. (eds.)] (pp. 123-144). Vienna: Springer. https://doi.org/10.1007/978-3-7091-1130-7_9

Negruk, V. (2013). Mitochondrial genome sequence of the legume *Vicia faba*. *Frontiers in Plant Science*. **4:** 128. https://doi.org/10.3389/fpls.2013.00128.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **30(9):** 1312-1313. https://doi.org/10.1093/bioinformatics/btu033.

Stern, D.B. and Palmer, J.D. (1984). Extensive and widespread homologies between mitochondrial DNA and chloroplast DNA in plants. *Proceedings of the National Academy of Sciences of the United States of America*. **81(7):** 1946-1950. https://doi.org/10.1073/pnas.81.7.1946.

Štorchová, H. and Krüger, M. (2024). Methods for assembling complex mitochondrial genomes in land plants. *Journal of Experimental Botany*. **75(17):** 5169-5174. https://doi.org/10.1093/jxb/erae034.

Takenaka, M., Zehrmann, A., Verbitskiy, D., Härtel, B. and Brennicke, A. (2013). RNA editing in plants and its evolution. *Annual Review of Genetics*. **47:** 335-352. https://doi.org/10.1146/annurev-genet-111212-133519.

Thiel, T., Michalek, W., Varshney, R.K. and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics*. **106(3):** 411-422. https://doi.org/10.1007/s00122-002-1031-0.

Wang, D., Wu, Y.W., Shih, A.C.C., Wu, C.S., Wang, Y.N. and Chaw, S.M. (2007). Transfer of chloroplast genomic DNA to mitochondrial genome occurred at least 300 MYA. *Molecular Biology and Evolution*. **24(9):** 2040-2048. https://doi.org/10.1093/molbev/msm133.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J. and Yu, J. (2010). KaKs_Calculator 2.0: A toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, Proteomics and Bioinformatics*. **8(1):** 77-80. https://doi.org/10.1016/s1672-0229(10)60008-3.

Wick, R.R., Judd, L.M., Gorrie, C.L. and Holt, K.E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*. **13(6):** e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

Wick, R.R., Schultz, M.B., Zobel, J. and Holt, K.E. (2015). Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics*. **31(20):** 3350-3352. https://doi.org/10.1093/bioinformatics/btv383.

Yucel, G., Betekhtin, A., Cabi, E., Tuna, M., Hasterok, R. and Kolano, B. (2022). The chromosome number and rDNA loci evolution in *Onobrychis* (Fabaceae). *International Journal of Molecular Sciences*. **23(19):** 11033. https://doi.org/10.3390/ijms231911033.

Zhou, Q., Ni, Y., Li, J., Huang, L., Li, H., Chen, H. and Liu, C. (2023). Multiple configurations of the plastid and mitochondrial genomes of *Caragana spinosa*. *Planta*. **258(5):** 98. https://doi.org/10.1007/s00425-023-04245-6.