

RESEARCH

Open Access



# A comparative plastome approach enhances the assessment of genetic variation in the *Melilotus* genus

Pan Xu<sup>1†</sup>, Minghui Meng<sup>1†</sup>, Fan Wu<sup>1</sup> and Jiyu Zhang<sup>1\*</sup>

## Abstract

**Background** *Melilotus*, a member of the Fabaceae family, is a pivotal forage crop that is extensively cultivated in livestock regions globally due to its notable productivity and ability to withstand abiotic stress. However, the genetic attributes of the chloroplast genome and the evolutionary connections among different *Melilotus* species remain unresolved.

**Results** In this study, we compiled the chloroplast genomes of 18 *Melilotus* species and performed a comprehensive comparative analysis. Through the examination of protein-coding genes, we successfully established a robust phylogenetic tree for these species. This conclusion is further supported by the phylogeny derived from single-nucleotide polymorphisms (SNPs) across the entire chloroplast genome. Notably, our findings revealed that *M. infestus*, *M. siculus*, *M. sulcatus*, and *M. speciosus* formed a distinct subgroup within the phylogenetic tree. Additionally, the chloroplast genomes of these four species exhibit two shared inversions. Moreover, inverted repeats were observed to have reemerged in six species within the IRLC. The distribution patterns of single-nucleotide polymorphisms (SNPs) and insertions/deletions (InDels) within protein-coding genes indicated that *ycf1* and *ycf2* accumulated nonconservative alterations during evolutionary development. Furthermore, an examination of the evolutionary rate of protein-coding genes revealed that *rps18*, *rps7*, and *rpl16* underwent positive selection specifically in *Melilotus*.

**Conclusions** We present a comparative analysis of the complete chloroplast genomes of *Melilotus* species. This study represents the most thorough and detailed exploration of the evolution and variability within the genus *Melilotus* to date. Our study provides valuable chloroplast genomic information for improving phylogenetic reconstructions and making biogeographic inferences about *Melilotus* and other Papilionoideae species.

**Keywords** *Melilotus*, Chloroplast genome, Evolution, Inversion

<sup>†</sup>Pan Xu and Minghui Meng contributed equally.

\*Correspondence:

Jiyu Zhang  
zhangjy@lzu.edu.cn

<sup>1</sup>State Key Laboratory of Grassland Agro-ecosystems, Key Laboratory of Grassland Livestock Industry Innovation, Ministry of Agriculture and Rural Affairs, Engineering Research Center of Grassland Industry, College of Pastoral Agriculture Science and Technology, Ministry of Education, Lanzhou University, Lanzhou 730000, China



## Background

With the extensive evolution of plants, each plant has a distinct origin and diverse evolutionary history that has changed its genetic makeup compared to that of related species, leading to various physiological and phenotypic differences [1, 2]. With advancements in technology, plant genome sequencing has become more affordable and accessible. By analysing sequences and using comparative genomics methods [3], researchers can assess biological events such as positive selection, genetic diversity, chromosome structure variation and polyploidy, which have become hotspots of biological research in recent years [4, 5]. The maternal inheritance of chloroplast genomes, characterized by their unique attributes [6], offers a convenient and dependable avenue for elucidating plant evolution and genetic interconnections among closely related species [7]. In contrast to intricate and extensive nuclear genomes, chloroplast genomes, which are typically smaller than 200 kb, exhibit moderate nucleotide substitution rates [8], rendering them amenable to sequencing using both next-generation sequencing and single-molecule long-read sequencing technologies [9]. The analysis of complete chloroplast genome sequences offers a substantial amount of valuable information, including insights into structural variations, gene losses, and single-base mutations. This information can be utilized to enhance our understanding of evolutionary distinctions, investigate genetic diversity, and construct detailed phylogenetic trees [10, 11]. Comparative chloroplast genome examination has proven to be an effective method for identifying evolutionary relationships at the species level [12–14].

Notably, the family Fabaceae is the third largest family among angiosperms [15]. Species of this family are grown for food and feed, as well as are ideal models for studying classification, diversity and genetic evolution. Based on the analysis of *matK* sequences and a comprehensive sampling approach, it is evident that the legume family can be distinctly categorized into six monophyletic subfamilies [16]. Notably, Papilionoideae, which comprises a vast assemblage of more than 14,000 species, stands out as the subfamily with the highest species diversity. Furthermore, within Papilionoideae, a more intricate classification reveals six distinct clades, with the largest one being IRLC (inverted repeat lacking clade) [17]. This particular clade encompasses approximately 4000 species spanning 52 genera, including *Melilotus*. At present, the genomes of multiple species of Fabaceae, such as soybean [18], alfalfa [19], pea [20], and red clover [21], have been sequenced, providing a basis for comparative genome studies and evolutionary history examination [22]. However, while *Melilotus* is an important Papilionoideae genus with 19 species [23], information about the genome of the *Melilotus* spp (also called sweet clover)

remains scarce [24–26]; however, *Melilotus* members are widely planted in global livestock areas due to their high yield and resistance to abiotic stress [27, 28].

*Melilotus*, a significant member of the Leguminosae family, consists of 19 annual or biennial species. *Melilotus* spp serves as a rotational crop primarily utilized for forage production, soil enhancement, and a nectar source. *Melilotus* spp. are renowned for their remarkable resilience to drought, cold, and high-salinity conditions [29, 30]. Despite the application of various molecular marker-based techniques to evaluate genetic diversity in *Melilotus* species, the evolutionary connections among these species remain ambiguous [31–34]. Here, we report the assemblies and gene annotations of the chloroplast genomes of 18 *Melilotus* species and compare their genome sequences to identify structural variants. Phylogenetic trees are constructed to illustrate the evolutionary relationships of *Melilotus* by based on single-copy genes and single-nucleotide polymorphisms (SNPs). The chloroplast genomes of the 18 *Melilotus* species assembled in this study will be a useful resource for genetic studies and taxonomy.

## Results

### Sequencing of plant samples

We downloaded the whole-genome sequencing data of 18 species of sweet clover from the National Center for Biotechnology Information Short Read Archive, including *M. albus*, *M. officinalis*, *M. altissimus*, *M. dentatus*, *M. elegans*, *M. hirsutus*, *M. indicus*, *M. infestus*, *M. italicus*, *M. polonicus*, *M. segetalis*, *M. siculus*, *M. speciosus*, *M. spicatus*, *M. suaveolens*, *M. sulcatus*, *M. tauricus* and *M. wolgicus*. The short reads from whole-genome sequencing of all species were generated using the Illumina HiSeq 2500 platform. The total number of clean bases ranged from 5.0 Gb to 9.7 Gb, with GC contents ranging from 34.75 to 37.14%. This sequencing depth represented approximately 10-fold coverage of the entire genome. To verify the reliability of the assembly, 17 out of the 18 species were selected and newly sequenced, with the total number of clean bases ranging from 14.0 Gb to 58.0 Gb (Additional file 2: Table S1). Subsequently, the chloroplast genomes were assembled again based on the new data and compared with the previous assembly for each species.

### Chloroplast genome assembly and annotation

The lengths of the chloroplast genomes of the assembled species ranged from 122,620 to 146,150 base pairs (bp). Additionally, the GC content of these genomes varied between 33.56% and 34.53%. Notably, no gaps were detected in any of the genomes. Each of the 18 species underwent comprehensive annotation, resulting in the identification of approximately 110 functional

genes. These genes included 76–79 protein-coding genes (PCGs), 30–35 transfer RNA (tRNA) genes, and 4–8 ribosomal RNA (rRNA) genes (as presented in Table 1). Furthermore, the majority of these species possessed a single copy of the inverted repeat (IR). The arrangement of genes on the chromosomes exhibited remarkable consistency across all chloroplast genomes, with only a few potential inversions observed (as depicted in Fig. 1). Inverted repeat regions were detected only in *M. altissimus*, *M. dentatus*, *M. speciosus*, *M. infestus*, *M. siculus*, and *M. sulcatus*, and the length of the IR region varied greatly from 3,518 bp ~ 18,439 bp (Table 1). Due to duplication and the emergence of IR regions, 5 genes were found to have two copies, including one protein-coding gene (*psbM*), two tRNA genes (*trnN-GUU* and *trnR-ACG*), two rRNA genes (*rrn5* and *rrn4.5*) in *M. infestus*, *M. siculus*, *M. sulcatus*, and *M. speciosus*; 10 genes were found to have two copies in *M. dentatus*, including one protein-coding gene (*ycf1*), five tRNA genes (*trnA-UGC*, *trnI-GAU*, *trnN-GUU*, *trnR-ACG*, and *trnV-GAC*), and four rRNA genes (*rrn16*, *rrn23*, *rrn4.5*, and *rrn5*); and 13 genes were found to have two copies in *M. altissimus*, including four protein-coding genes (*ycf1*, *rps15*, *ndhH*, and *ndhA*), five tRNA genes (*trnA-UGC*, *trnI-GAU*, *trnN-GUU*, *trnR-ACG*, and *trnV-GAC*), and four rRNA genes (*rrn16*, *rrn23*, *rrn4.5*, and *rrn5*).

#### Comparison of chloroplast genomes within *Melilotus*

According to the collinearity between *M. albus* and other *Melilotus* chloroplast genomes (Additional file 1: Fig. S1), three main inversions (INV1, INV2 and INV3) were detected at locations near 9,890 to 13,774 bp, 61,784 to

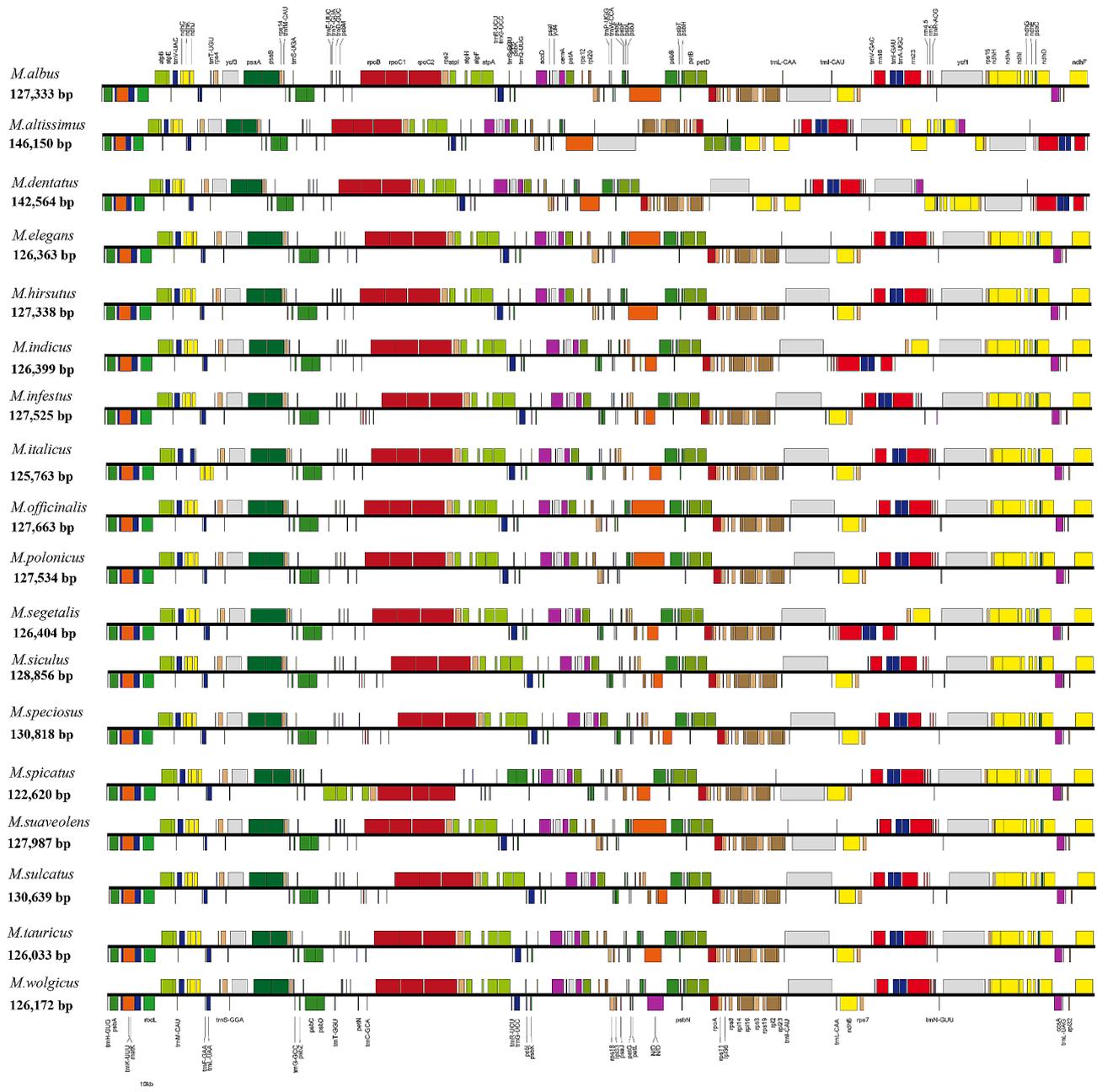
67,994 bp and 87,643 to 94,511 bp, with fragment lengths of 3,884 bp, 6,210 bp and 6,868 bp, respectively (Fig. 2a). The decline in mapping depths near the start and end of potential inversion regions during the mapping of short reads from whole-genome resequencing data to the *M. albus* genome may support the presence of an inversion (Fig. 2b). We found that the *ndhC*, *ndhK* and *ndhJ* genes were contained in INV1; the *rps12*, *rpl20*, *rps18*, *rpl33*, *psaI*, *petG*, *petL*, *psbE*, *psbF*, *psbL* and *psbJ* genes were contained in INV2; and only the *ycf2* gene was contained in INV3. The species *M. italicus* had all three INVs; *M. infestus*, *M. speciosus*, *M. siculus*, and *M. sulcatus* had both INV2 and INV3; and no INVs were detected in *M. elegans* or *M. suaveolens* (Table 2).

#### Phylogenetic analysis

We constructed a phylogenetic tree of 18 *Melilotus* species utilizing 76 protein-coding genes through IQ-TREE and MrBayes. *Medicago truncatula* was employed as an outgroup in the analysis (Fig. 3 and Additional file 1: Fig. S2). Additionally, we constructed a phylogenetic tree based on the complete chloroplast genome sequences (Additional file 1: Fig. S3). The topologies of the three phylogenetic trees were consistent, revealing that all 18 species could be categorized into two distinct subgroups. One subgroup consisted of *M. infestus*, *M. siculus*, *M. sulcatus* and *M. speciosus*. All species within this subgroup exhibited INV2 and INV3, in comparison to the *M. albus* reference. Of particular significance, they all possessed inverted repeat regions. All remaining species, including two important ones, *M. albus* and *M. officinalis*, constitute the other subgroup. The interspecific relationships

**Table 1** Features summary of of the 18 *Melilotus* species chloroplast genomes

Species	Length	IR length	GC%	Number of genes			
				PCGs	tRNA	rRNA	Total
<i>M. albus</i>	127,333	-	33.65	76	30	4	110
<i>M. altissimus</i>	146,150	18,439	34.53	79	35	8	122
<i>M. dentatus</i>	142,564	15,494	34.50	77	35	8	120
<i>M. elegans</i>	126,363	-	33.67	76	30	4	110
<i>M. hirsutus</i>	127,338	-	33.57	76	30	4	110
<i>M. indicus</i>	126,399	-	33.86	76	30	4	110
<i>M. infestus</i>	127,525	3,528	34.23	77	32	6	115
<i>M. italicus</i>	125,763	-	33.99	76	30	4	110
<i>M. officinalis</i>	127,663	-	33.63	76	30	4	110
<i>M. polonicus</i>	127,534	-	33.83	76	31	4	111
<i>M. segetalis</i>	126,404	-	33.86	76	30	4	110
<i>M. siculus</i>	128,856	3,670	34.33	77	32	6	115
<i>M. speciosus</i>	130,818	4,317	34.10	76	32	6	114
<i>M. spicatus</i>	122,620	-	33.94	76	30	4	110
<i>M. suaveolens</i>	127,987	-	33.56	76	30	4	110
<i>M. sulcatus</i>	130,639	3,876	34.09	77	32	6	115
<i>M. tauricus</i>	126,033	-	33.79	76	30	4	110
<i>M. wolgicus</i>	126,172	-	33.80	76	30	4	110



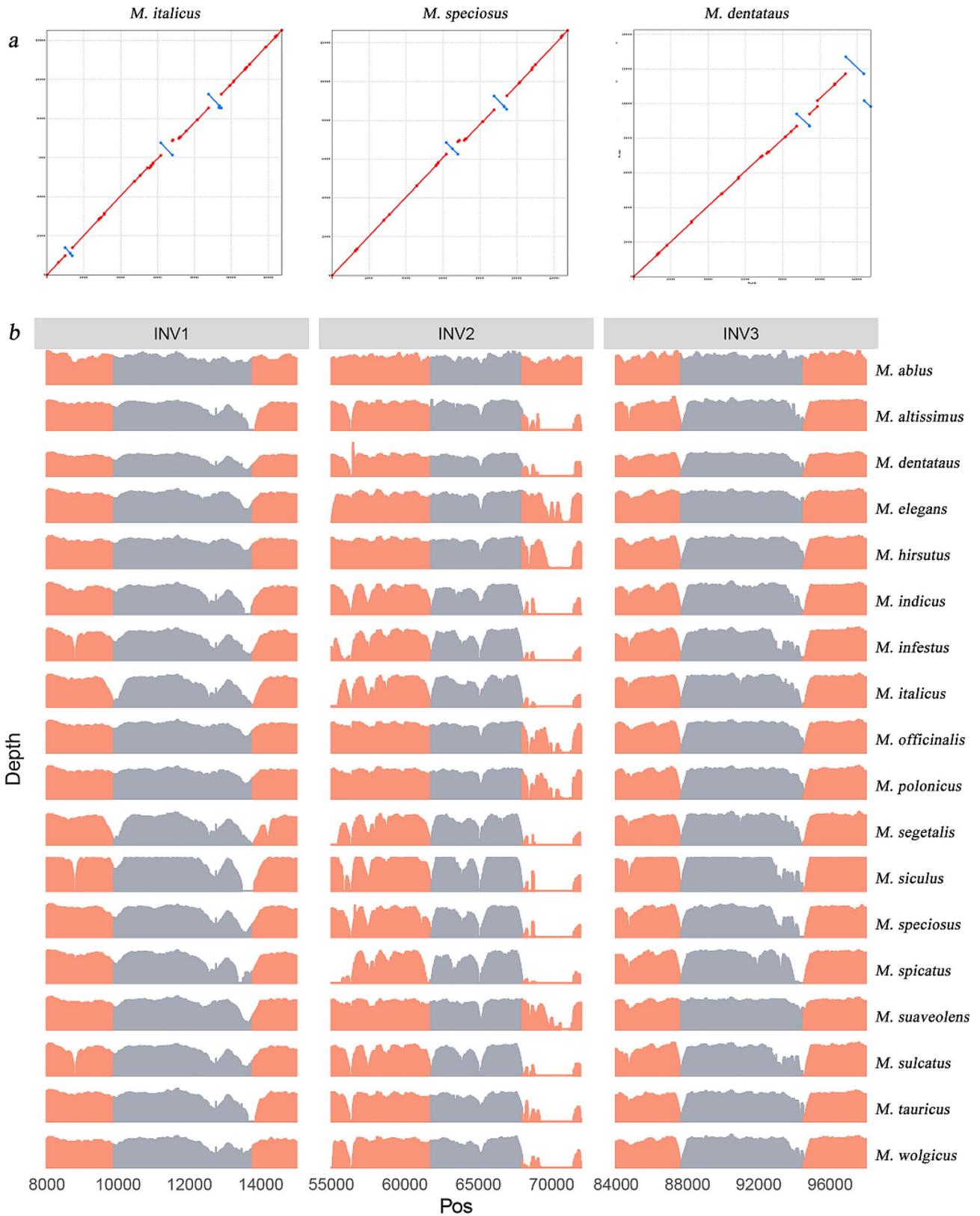
**Fig. 1** Gene features of 18 assembled chloroplast genomes. The genes drawn above the line are on the positive strand, and the genes drawn under the line are located on the negative strand

revealed that *M. officinalis* diverged earlier than *M. albus*. Among the other species, *M. indicus* and *M. segetalis* shared a common ancestor and exhibited the most robust collinearity among the 18 chloroplast genomes (Fig. 1).

**Nucleotide substitution rates**

Substitution rates, such as synonymous substitution (Ds), nonsynonymous substitution (Dn) and Dn/Ds, were estimated for the 76 PCGs to detect evolutionary rate heterogeneity and to represent different selection

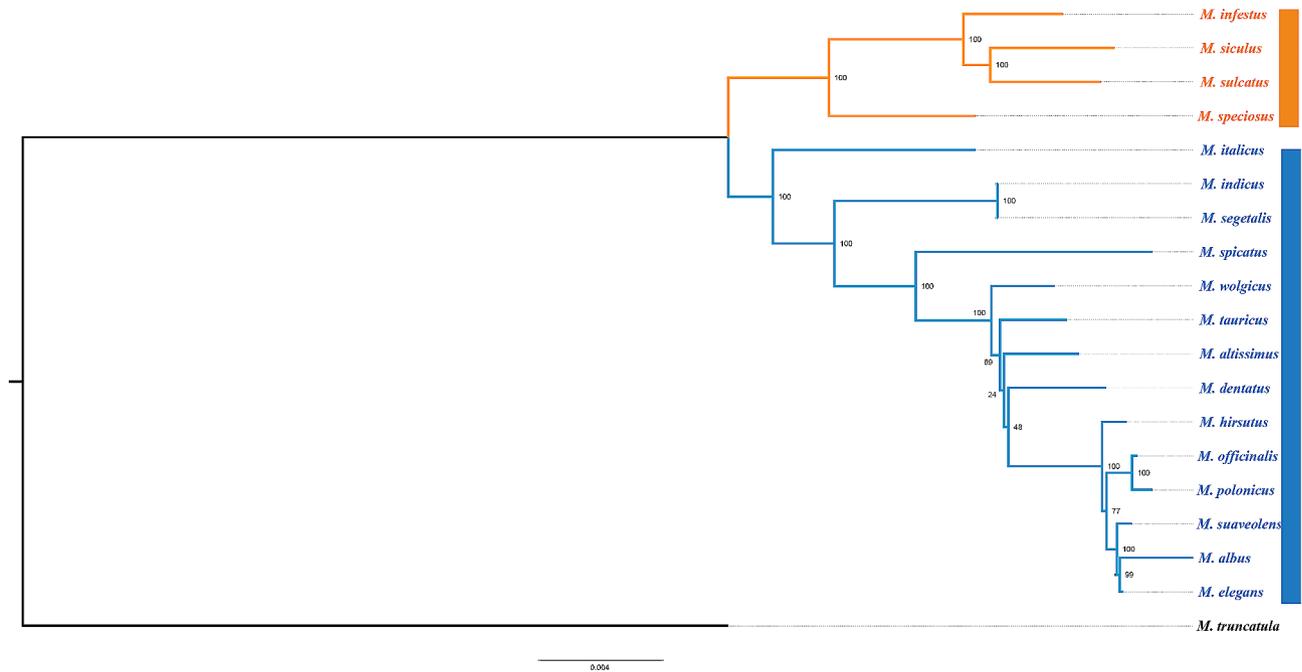
regimes acting on PCGs (Table 3). Among the 76 PCGs, *accD*, *clpP*, and *ycf3* had relatively high Ds and Dn values simultaneously, while *rps18*, *rps7*, and *rpl16* exhibited high Dn/Ds values (>1), indicating that they have experienced positive selection. These three genes were located within the inversion regions, suggesting that these specific regions may have played a role in the elevated Dn/Ds values in *Melilotus* (Fig. 4a). These genes were further categorized into ten groups, comprising nine functional groups and one group of other genes (OG, Table 3).



**Fig. 2** Structural variations between *M. albus* and other species. Collinear plot between *M. italicus*, *M. speciosus*, *M. dentataus* and *M. albus* (a). Depth of Illumina short reads mapped to *M. albus* near three INVs (b)

**Table 2** Three main inversions among the species of sweet clovers

ID	Start	End	Species	Inversion PCGs
INV1	9,890	13,774	<i>M. italicus</i>	<i>ndhC, ndhK, ndhJ</i>
INV2	61,784	67,994	<i>M. indicus, M. infestus, M. italicus, M. segetalis, M. siculus, M. speciosus, M. spicatus, M. sulcatus</i>	<i>rps12, rpl20, rps18, rpl33, psal, petG, petL, psbE, psbF, psbL, psbJ</i>
INV3	87,643	94,511	<i>M. dentatus, M. hirsutus, M. indicus, M. infestus, M. italicus, M. polonicus, M. officinalis, M. segetalis, M. siculus, M. speciosus, M. sulcatus, M. tauricus, M. wolgicus</i>	<i>ycf2</i>



**Fig. 3** Phylogenetic tree was constructed using single copy genes of 18 species by IQ-Tree, the size of green circle on the branch represented the value of bootstrap

**Table 3** Plastid genes and their functional groups included in analyses

Functional groups	Genes
Photosystem I (PSA)	<i>psaA, psab, psac, psal, psaj</i>
Photosystem II (PSB)	<i>psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbL, psbM, psbN, psbT, psbZ</i>
Cytochrome B6f complex (PET)	<i>petA, petB, petD, petG, petF, petN</i>
ATP synthase (ATP)	<i>atpA, atpB, atpE, atpF, atpH, atpI</i>
Rubisco large subunit (Rubisco)	<i>rbcl</i>
RNA polymerase (RPO)	<i>rpoA, rpoB, rpoC1, rpoC2</i>
Ribosomal proteins large subunit (RPL)	<i>rpl2, rpl14, rpl16, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36</i>
Ribosomal proteins small subunit (RPS)	<i>rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps14, rps15, rps18, rps19</i>
NADH dehydrogenase (NDH)	<i>ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK</i>
Other genes (OG)	
Conserved coding frame	<i>ycf1, ycf2, ycf3, ycf4</i>
Acetyl-CoA-carboxylase	<i>accD</i>
ATP-dependent protease	<i>clpP</i>
Cytochrome c biogenesis	<i>ccsA</i>
Membrane protein	<i>cemA</i>
Maturase	<i>matK</i>



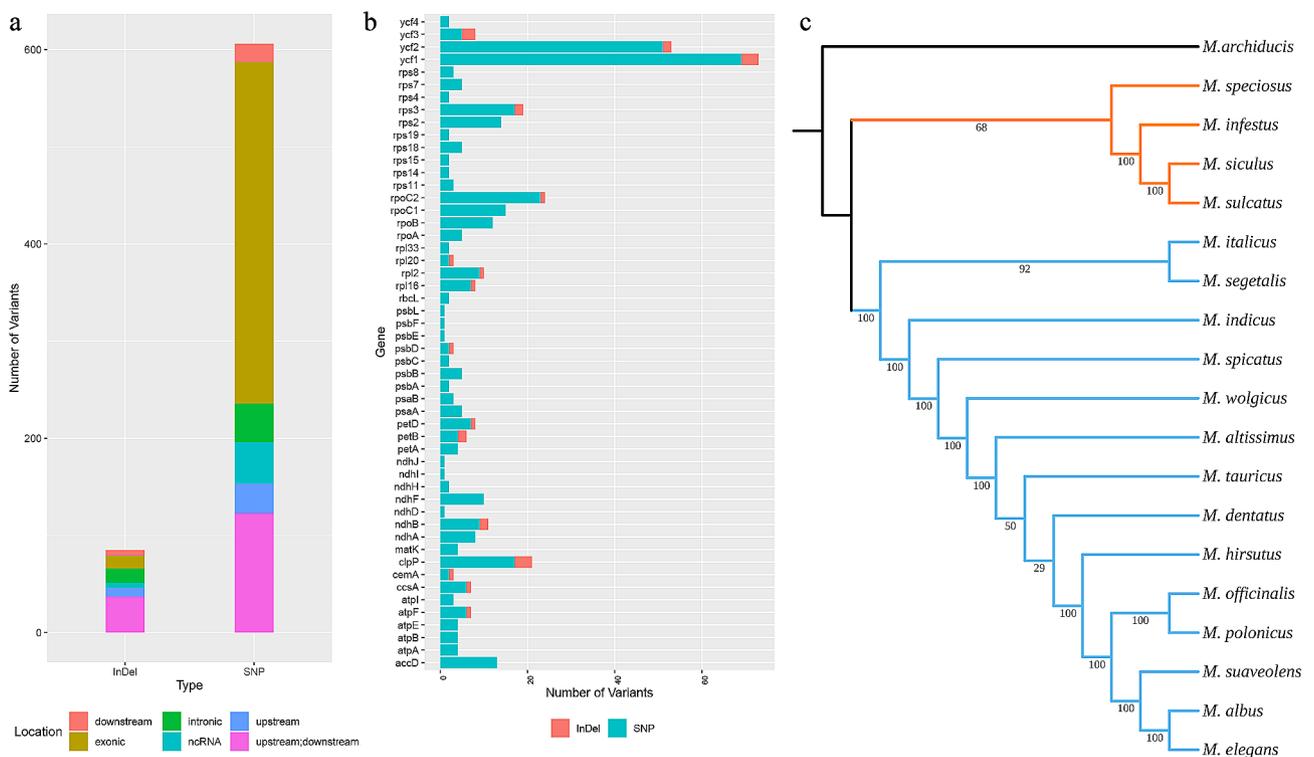
### Variant calling and construction of the SNP phylogenetic tree

Variants were identified using the reference genome of *M. albus*, ultimately revealing a total of 606 SNPs and 85 InDels. There were 364 mutations located in coding sequences (351 SNPs and 13 InDels), including 117 synonymous and 233 nonsynonymous mutations. Among the other mutations, 172 SNPs and 52 InDels were located upstream and downstream of genes, and 83 SNPs and 20 InDels were located in noncoding RNA or introns (Fig. 5a and Additional file 2: Table S2). The distribution of SNPs and InDels on PCGs suggested that genes such as *ycf1* and *ycf2* have undergone nonconservative evolution. (Fig. 5b and Additional file 3: Table S3). Only the SNPs from these mutations were used to construct the phylogenetic tree, which had a similar structure to the tree constructed using single-copy genes. *M. speciosus*, *M. infestus*, *M. siculus*, and *M. sulcatus* formed a distinct group with the greatest genetic distance from *M. albus*. However, some differences were detected between the relationships of *M. atissimus*, *M. dentatus* and *M. tauricus* in the two trees (Fig. 5c).

### Discussion

The results presented in this study mark the first instance of utilization of the entire chloroplast genome to investigate genetic variation across 18 *Melilotus* species. Our

study revealed that the genotypes and gene numbers were basically the same among the studied species, and no inverted repeats (IRs) were detected in the 12 chloroplast genomes. However, the presence of IRs was identified in another six species. This may be an uncommon occurrence in the conserved chloroplast genomes of the IRLC [35]. Recently, a reemergence of approximately 15 kb of novel IRs in *M. dentatus* and a parallel reappearance of approximately 9 kb of IRs in another species within the IRLC, *Medicago minima*, have been reported [36, 37]. Previous studies have shown that the IR region of chloroplasts is not as conserved as other regions and that the size of the IR could be highly variable [38]. Furthermore, some species have lost one copy of IR as same as *Melilotus*, including *Erodium texanum* (GenBank accession number HM125536) [38], *Cicer arietinum* (GenBank accession number NC\_011163.1) [39], *Glycyrrhiza inflata* (GenBank accession number NC\_042146.1) [40], *Medicago truncatula* (GenBank accession number NC\_003119) [41], and three species from the genus *Astragalus* [42]. Our findings suggest that the differences in gene order in the chloroplast genomes among the 18 species are not unique but rather a common (Fig. 1). The inversion regions in these genomes were found to have a positive selection effect on the *rps18*, *rps7*, and *rpl16* genes, which may be the main cause of variation among the species. The *rps18* gene is crucial for chloroplast



**Fig. 5** The distributions of mutations located in the chloroplast genome (a). Numbers of SNPs and InDels in each PCG (b). Phylogenetic tree was constructed using SNPs (c)

translation during plant development [43, 44]. The intergenomic inversion of chloroplasts and IR reemergence contributed to changes in the gene order, that is, chloroplast genome rearrangement and obvious chloroplast genome rearrangements have been observed in *Campanulaceae*, *Oleaceae* and *Geraniaceae* [38, 45, 46]. Genome rearrangement is often related to repetitive sequences, which can induce recombination, especially in some gymnosperms; because of IR deletion, gene recombination and deletion occur frequently [47]. The inverted fragments play a very important role in evolution; inverted fragments have resulted in the recombination of the chloroplast genome and contribute to the diversity of the chloroplast genome structure in *Pinaceae* [48, 49].

The analysis of entire chloroplast genomes allowed us to derive a highly reliable phylogenetic tree based on a set of 606 high-quality SNPs rooted with *M. truncatula* as an outgroup (Fig. 5c). Our phylogenomic framework is largely congruent with the phylogenetic framework of using ITS+*rbcL*+*matK*, *rbcL*+*matK*+*trnL-F*+ITS and the supermatrix of 70 EST-SSR markers [31–33]. These analyses all support the classification of species within the *Melilotus* genus into two groups. Apart from the analysis based on the concatenation of the four genes *rbcL*+*matK*+*trnL-F*+ITS, all other analyses endorse the clustering of *M. infestus*, *M. siculus*, *M. sulcatus*, and *M. speciosus*. In fact, the topology obtained based on *rbcL*+*matK*+*trnL-F*+ITS represents a multifurcating structure, indicating that the clustering of *M. speciosus* with the other three species cannot be rejected [32]. Notably, reports based on 47 L support these four species forming a monophyletic group. This finding is consistent with our results based on the complete chloroplast genome, although there may be differences in the specific interspecies relationships [50]. In contrast, in analyses based on EST-SSR markers and a small number of chloroplast genes, these four species clustered together with a few other species within the *Melilotus* genus [31]. In a phylogenetic tree based on LTRs, researchers also identified a monophyletic clade (clade II). This clade is composed of three species: *M. indicus*, *M. italicus*, and *M. spicatus* [50]. The clustering of these three species is also supported by analysis of EST-SSR markers, although the results indicate the inclusion of an additional species, *M. segetalis* [31]. The clustering of these four species, *M. indicus*, *M. italicus*, *M. spicatus*, and *M. segetalis*, is also supported to some extent by phylogenetic analysis based on a small number of chloroplast genes [32]. According to our study results, these four species are positioned at the base of this subgroup, forming paraphyletic groups along with other species within this subgroup. This finding highlights the distinctiveness of these four species. However, determining their precise phylogenetic position may necessitate the analysis of additional markers, such as

whole-genome sequences. Comparing trees constructed from genomic SNPs and coding genes revealed good consistency, with only some differences among *M. atisimius*, *M. dentatus*, and *M. tauricus*, suggesting that it is possible to use both methods to identify the evolutionary relationships of species in *Melilotus* (Figs. 3 and 5c). The INVs we identified in this research could also be used as indicators to study the relationships among germplasms. In other words, species that have a close evolutionary relationship, such as *M. speciosus*, *M. infestus*, *M. siculus* and *M. sulcatus*. In these inversion regions, we also found that four specific genes (*rpoC2*, *clpP*, *ycf2*, and *ycf1*) simultaneously accumulated high numbers of SNPs and InDels ( $\geq 20$  bp) in the coding region (Fig. 5b). The *ycf2* gene around the INV is a putative ATPase with unknown function. This gene is expressed in many plants, including nonphotosynthetic plants. Previous experiments in tobacco indicated that this putative ATPase plays an essential role in cell survival via activity in the tobacco chloroplast [51]. These genes may serve as common hotspots of genetic variation in *Melilotus*, as indicated by this observation. Our research on the *Melilotus* panplastome shed light on the maternal inheritance in this genus and can be used as a basis for studying the phylogenetic degeneration of *Melilotus* with other species, constructing phylogenetic trees in the Fabaceae family, and even the Papilionoideae subfamily.

## Conclusions

This work is based on a comparative analysis of the chloroplast genomes of 18 *Melilotus* species and presents a comprehensive study of their evolutionary relationships and nucleotide substitution rates. The comparative genomic analysis was used to identify the genomic SNPs, InDels, main inversion positions, and evolutionary rate heterogeneity occurring in the chloroplast genomes of the studied *Melilotus* species. In total, 391 SNPs and 28 InDels located in exons and introns of 52 PCGs were found, and the results indicated that four specific genes (*rpoC2*, *clpP*, *ycf2*, and *ycf1*) simultaneously accumulated high numbers of SNPs and indels in the coding regions, while the other three specific genes (*rps18*, *rps7*, and *rpl16*) exhibited positive selection effects in the inversion regions; these distinctions may be the main source of variation among the 18 studied species.

## Materials and methods

### Plant materials and sequences

The raw sequence reads of the *Melilotus* accessions were download from National Center for Biotechnology Information [52]. To verify the accuracy of the chloroplast genome assembly, we also selected 16 *Melilotus* accessions for sequencing (Additional file 2: Table S1). These accessions were cultivated in a greenhouse on Yuzhong

Campus, Lanzhou, China. Genomic DNA was extracted from fresh young leaves using the CTAB method. Sixteen paired-end libraries with an average insert size of 150 bp were constructed and sequenced using the BGISEQ-500 platform.

#### Assembly and annotation of the chloroplast genome of *Melilotus*

GetOrganelle-1.7.1 [53] was employed to perform *de novo* assembly of the chloroplast genome, and the sequence was reordered to determine the assembly according to the alignment result on the basis of the comparison between the sequence and the *M. albus* reference chloroplast genome performed by MUMmer4.0 [54]. The gene features of the circular chloroplast DNA were annotated by GeSeq packages [55] and revised manually. Finally, all circular chloroplast genomes were constructed with OGRAW [56].

#### Phylogenetic analysis

OrthoFinder was used to identify the gene families, and only the amino acid sequences of single-copy genes were used to construct the phylogenetic tree [57]. Multiple sequence alignments were performed with MUSCLE, and then a phylogenetic tree was constructed with both the maximum likelihood and Bayesian methods using IQ-TREE [58] and MrBayes [59] with 1000 times bootstrap replicates. *Medicago truncatula* (JX512023.1) was used as an outgroup to root the tree. All chloroplast genome sequences were aligned using Cactus [60] with the -auto option, and conserved blocks of the alignment were used to construct a phylogenetic tree with IQ-TREE. The tool iTOL was used to visualize and modify the tree file [61].

#### Comparison and variant calling of the chloroplast genome

Chloroplast sequences of 17 other *Melilotus* species were aligned to the reference chloroplast genome of *M. albus* to perform genome comparisons using MUMmer4.0 [54], and structural variants were identified via smartie-sv [62]. The R package ggplot2 was used to visualize the mapping depths near the breakpoints.

#### Nucleotide substitution rates

The nucleotide substitution rates, nonsynonymous rates (Dn), synonymous rates (Ds), and the ratio of nonsynonymous to synonymous rates (Dn/Ds), were determined using PAML v.4.9 [63]. Codon substitution models and likelihood ratio tests (codeml) were conducted based on the branch site model. The phylogeny generated was using the concatenated method. The “model=0” option was used to allow single Dn/Ds values to vary among branches. The 76 PCGs were consolidated into nine groups (Table 2) to compare the different functions among the groups, such as photosystem

I (PSA), photosystem II (PSB), the cytochrome B6f complex (PET), ATP synthase (ATP), rubisco large subunit (Rubisco), RNA polymerase (RPO), ribosomal proteins large subunit (RPL), ribosomal proteins small subunit (RPS) and NADH dehydrogenase (NDH), and other genes (OG, including *ycf1*, *ycf2*, *ycf3*, *ycf4*, *accD*, *clpP*, *ccsA*, *cemA* and *matK*).

#### Variant calling and construction of the phylogenetic tree

The data for the *Medicago archiducis* sample were downloaded from NCBI (SRX9404272), and bwa [64] was used to align the clean reads, which included 18 *Melilotus* and *Medicago archiducis* reads, to the *M. albus* chloroplast genome. The alignment files were sorted and indexed with SAMtools [65]. Variant calling was performed using GATK-4.0 [66] (<https://gatk.broadinstitute.org>) with default parameters, and variants were filtered by VCFtools (<http://vcftools.sourceforge.net>) with the parameters “--min-alleles 2 --max-alleles 2 --min-meanDP 5 --maf 0.05 --max-missing 0.5”. The filtered mutations were annotated using ANNOVAR [67]. The filtered SNPs were used to analyse the phylogenetic tree, and IQ-TREE [58] was used to select the best model and construct a phylogenetic tree using the maximum likelihood method with 1000 times bootstrap. The online iTOL [61] was used to visualize the phylogenetic tree.

#### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-024-10476-y>.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

#### Acknowledgements

Not applicable.

#### Author contributions

P.X. and M.M. performed the data analyses and writing the manuscript; F.W. participated in materials collection and revised the manuscript; J.Z. designed the research, supervised the project and contributed to the writing of the manuscript.

#### Funding

This research was supported by the National Natural Science Foundation of China (Grant No. 32061143035), Gansu Provincial Science and Technology Major Projects (Grant No. 19ZD2NA002), and the China Postdoctoral Science Foundation (2022M711453).

#### Data availability

The raw sequence reads of the *Melilotus* accessions were obtained from National Center for Biotechnology Information under the BioProject accession numbers PRJNA781345 and PRJNA759778. *Melilotus* accessions were applied from the National Plant Germplasm System (NPGS, United States Department of Agriculture, USA). The application procedure is legal. Plant germplasm is distributed to scientists, educators, producers and other bona fide research

and education entities from U.S. National Plant Germplasm System (NPGS) active collection sites.

## Declarations

### Ethics approval and consent to participate

All methods with the use of plants were carried out in accordance with relevant institutional, national, and international guidelines and legislation.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

Received: 29 November 2023 / Accepted: 29 May 2024

Published online: 03 June 2024

## References

- Doebley JF, Gaut BS, Smith BD. The molecular genetics of crop domestication. *Cell*. 2006;127(7):1309–21.
- Diamond J. Evolution, consequences and future of plant and animal domestication. *Nature*. 2002;418(6898):700–7.
- Haubold B, Wiehe T. Comparative genomics: methods and applications. *Die Naturwiss*. 2004;91(9):405–21.
- Paterson AH, Bowers JE, Burrow MD, Draye X, Elsik CG, Jiang C-X, Katsar CS, Lan T-H, Lin Y-R, Ming R. Comparative genomics of plant chromosomes. *Plant Cell*. 2000;12(9):1523–39.
- Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ, Wortman JR, Hine EE, Althoff R, Arbogast TS, Tallon LJ. Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell*. 2006;18(6):1348–59.
- Flavell R. Mitochondria and chloroplasts as descendants of prokaryotes. *Biochem Genet*. 1972;6(4):275–91.
- Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol*. 2016;17(1):134.
- Wicke S, Schneeweiss GM, Depamphilis CW, Müller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol Biol*. 2011;76(3):273–97.
- Yao S, Liang F, Gill RA, Huang X, Cheng X, Liu Y, Tong C, Liu S. A global survey of the transcriptome of allopolyploid *Brassica napus* based on single-molecule long-read isoform sequencing and Illumina-based RNA sequencing data. *Plant J*. 2020;103(2):843–57.
- Gao C, Wu C, Zhang Q, Zhao X, Wu M, Chen R, Zhao Y, Li Z. Characterization of chloroplast genomes from two *Salvia* medicinal plants and gene transfer among their mitochondrial and chloroplast genomes. *Front Genet*. 2020;11:1267.
- Xiong Y, Xiong Y, He J, Yu Q, Zhao J, Lei X, Dong Z, Yang J, Peng Y, Zhang X. The complete chloroplast genome of two important Annual Clover species, *Trifolium alexandrinum* and *T. resupinatum*: genome structure, comparative analyses and phylogenetic relationships with relatives in Leguminosae. *Plants*. 2020;9(4):478.
- Gu L, Su T, An M-T, Hu G-X. The complete chloroplast genome of the vulnerable *Oreocharis esquirolii* (Gesneriaceae): structural features, comparative and phylogenetic analysis. *Plants*. 2020;9(12):1692.
- Wu Z, Liao R, Yang T, Dong X, Lan D, Qin R, Liu H. Analysis of six chloroplast genomes provides insight into the evolution of *Chrysosplenium* (Saxifragaceae). *BMC Genomics*. 2020;21(1):621.
- Sobreiro MB, Vieira LD, Nunes R, Novaes E, Coissac E, Silva-Junior OB, Gratapaglia D, Collevatti RG. Chloroplast genome assembly of *Handroanthus Impetiginosus*: comparative analysis and molecular evolution in Bignoniaceae. *Planta*. 2020;252(5):1–16.
- Mabberley DJ. *The plant-book*. Cambridge: Cambridge University Press; 1997.
- Azani N, Babineau M, Bailey CD, Banks H, Barbosa AR, Pinto RB, Boatwright JS, Borges LM, Brown GK, Bruneau A, et al. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon*. 2017;66(1):44–77.
- Wojciechowski MF, Lavin M, Sanderson MJ. A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades within the family. *Am J Bot*. 2004;91(11):1846–62.
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J. Genome sequence of the palaeopolyploid soybean. *Nature*. 2010;463(7278):178–83.
- Shen C, Du H, Chen Z, Lu H, Zhu F, Chen H, Meng X, Liu Q, Liu P, Zheng L. The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasm provide genomic resources for alfalfa research. *Mol Plant*. 2020;13(9):1250–61.
- Kreplak J, Madoui M-A, Cápál P, Novák P, Labadie K, Aubert G, Bayer PE, Gali KK, Syme RA, Main D. A reference genome for pea provides insight into legume genome evolution. *Nat Genet*. 2019;51(9):1411–22.
- Ištvánek J, Jaroš M, Křenek A, Řepková J. Genome assembly and annotation for red clover (*Trifolium pratense*; Fabaceae). *Am J Bot*. 2014;101(2):327–37.
- Zhu H, Choi H-K, Cook DR, Shoemaker RC. Bridging model and crop legumes through comparative genomics. *Plant Physiol*. 2005;137(4):1189–96.
- Ou LJ, Li D, Lv JH, Chen WC, Zhang ZQ, Li XF, Yang BZ, Zhou SD, Yang S, Li WG, et al. Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence-absence variation analyses. *New Phytol*. 2018;220(2):360–3.
- Luo K, Jahufer M, Zhao H, Zhang R, Wu F, Yan Z, Zhang J, Wang Y. Genetic improvement of key agronomic traits in *Melilotus albus*. *Crop Sci*. 2018;58(1):285–94.
- Luo K, Wu F, Zhang D, Dong R, Fan Z, Zhang R, Yan Z, Wang Y, Zhang J. Transcriptomic profiling of *Melilotus albus* near-isogenic lines contrasting for coumarin content. *Sci Rep*. 2017;7(1):1–14.
- Popoola J, Ojuederie O, Omonhinmin C, Adegbite A. Neglected and underutilized legume crops: Improvement and future prospects. In: *Recent Advances in Grain Crops Research* IntechOpen; 2019.
- Al Sherif EA. *Melilotus indicus* (L.) All., a salt-tolerant wild leguminous herb with high potential for use as a forage crop in salt-affected soils. *Flora-Morphology Distribution Funct Ecol Plants*. 2009;204(10):737–46.
- Smith W, Gorz H. Sweetclover improvement. *Advances in agronomy*. Volume 17. Elsevier. 1965;pp. 163–231.
- Zabala JM, Marinoni L, Giavedoni JA, Schrauf GE. Breeding strategies in *Melilotus albus* Desr., a salt-tolerant forage legume. *Euphytica*. 2018;214(2):22.
- Duan Z, Wang S, Zhang Z, Yan Q, Zhang C, Zhou P, Wu F, Zhang J. The MabHLH11 transcription factor interacting with MaMYB4 acts additively in increasing plant scopolin biosynthesis. *Crop J*. 2023;11(6):1675–85.
- Yan Z, Wu F, Luo K, Zhao Y, Yan Q, Zhang Y, Wang Y, Zhang J. Cross-species transferability of EST-SSR markers developed from the transcriptome of *Melilotus* and their application to population genetics research. *Sci Rep*. 2017;7(1):17959.
- Wu F, Ma J, Meng Y, Zhang D, Pascal Muvunyi B, Luo K, Di H, Guo W, Wang Y, Feng B, et al. Potential DNA barcodes for *Melilotus* species based on five single loci and their combinations. *PLoS ONE*. 2017;12(9):e0182693.
- Di H, Duan Z, Luo K, Zhang D, Wu F, Zhang J, Liu W, Wang Y. Interspecific Phylogenetic relationships within Genus *Melilotus* based on Nuclear and Chloroplast DNA. *PLoS ONE*. 2015;10(7):e0132596.
- Wu F, Zhang D, Ma J, Luo K, Di H, Liu Z, Zhang J, Wang Y. Analysis of genetic diversity and population structure in accessions of the genus *Melilotus*. *Ind Crops Prod*. 2016;85:84–92.
- Wojciechowski MF, Sanderson MJ, Hu J-M. Evidence on the Monophyly of *Astragalus* (Fabaceae) and its major subgroups based on Nuclear ribosomal DNA ITS and chloroplast DNA trnL Intron Data. *Syst Bot*. 1999;24(3):409–37.
- Choi IS, Jansen R, Ruhlman T. Lost and found: return of the inverted repeat in the Legume Clade defined by its absence. *Genome Biol Evol*. 2019;11(4):1321–33.
- Wu S, Chen J, Li Y, Liu A, Li A, Yin M, Shrestha N, Liu J, Ren G. Extensive genomic rearrangements mediated by repetitive sequences in plastomes of *Medicago* and its relatives. *BMC Plant Biol*. 2021;21(1):421.
- Guisinger M, Kuehl J, Boore J, Jansen R. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. *Mol Biology Evol*. 2011;28(1):583–600.
- Jansen RK, Wojciechowski MF, Sanniyasi E, Lee S-B, Daniell H. Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). *Mol Phylogenet Evol*. 2008;48(3):1204–17.
- Jiang W, Tan W, Gao H, Yu X, Zhang H, Bian Y, Wang Y, Tian X. Transcriptome and complete chloroplast genome of *Glycyrrhiza inflata* and comparative analyses with the other two licorice species. *Genomics*. 2020;112(6):4179–4188.

41. Csanad G, Pal, Maliga: two distinct plastid genome configurations and unprecedented intraspecies length variation in the accD coding region in *Medicago truncatula*. *DNA Research*. 2014;21(4):417–27.
42. Tian C, Li X, Wu Z, Li Z, Hou X, Li FY. Characterization and comparative analysis of complete chloroplast genomes of three species from the Genus *Astragalus* (Leguminosae). *Front Genet*. 2021;12:705482.
43. Rogalski M, Ruf S, Bock R. Tobacco plastid ribosomal protein S18 is essential for cell survival. *Nucleic Acids Res*. 2006;34(16):4537–45.
44. Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, Janell D, Bashan A, Bartels H, Agmon I, Franceschi F, et al. Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell*. 2000;102(5):615–23.
45. Cosner ME, Raubeson LA, Jansen RK. Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. *BMC Evol Biol*. 2004;4(1):1–17.
46. Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK. The complete chloroplast genome sequence of *Pelargonium × hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Mol Biol Evol*. 2006;23(11):2175–2190.
47. Chaw SM, Jansen RK. *Plastid Genome Evolution*: Academic Press; 2018.
48. Wu CS, Chaw SM. Large-scale comparative analysis reveals the mechanisms driving Plastomic Compaction, reduction, and inversions in Conifers II (Cupressophytes). *Genome Biol Evol*. 2016;8(12):3740–50.
49. Hsu CY, Wu CS, Chaw SM. Birth of four chimeric plastid gene clusters in Japanese Umbrella Pine. *Genome Biol Evol*. 2016;8(6):1776–84.
50. Ouyang Z, Wang Y, Ma T, Kanzana G, Wu F, Zhang J. Genome-wide identification and development of LTR retrotransposon-based molecular markers for the *Melilotus* Genus. *Plants (Basel)* 2021, 10(5):890.
51. Drescher A, Ruf S, Calsa T Jr, Carrer H, Bock R. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *Plant Journal: Cell Mol Biology*. 2000;22(2):97–104.
52. Wu F, Duan Z, Xu P, Yan Q, Meng M, Cao M, Jones CS, Zong X, Zhou P, Wang Y et al. Genome and systems biology of *Melilotus albus* provides insights into coumarins biosynthesis. *Plant Biotechnol J* 2022; 20(3):592–609.
53. Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi TS, Li DZ. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol*. 2020;21(1):241.
54. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol*. 2018;14(1):e1005944.
55. Michael T, Pascal L, Tommaso P, Ulbricht-Jones ES, Axel F, Ralph B, Stephan G. GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res* 2017(W1):W6–W11.
56. Lohse M, Drechsel O, Bock R. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr Genet*. 2007;52(5):267–74.
57. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;20(1):238.
58. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020;37(5):1530–4.
59. Huelsenbeck JP, Ronquist F. MRBAYES: bayesian inference of phylogenetic trees. *Bioinformatics*. 2001;17(8):754–5.
60. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*. 2020;587(7833):246–51.
61. Letunic I, Bork P. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. 2019;47(W1):W256–9.
62. Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML et al. High-resolution comparative analysis of great ape genomes. *Science*. 2018;360(6393):eaar6343.
63. Yang Z, Yang ZH. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24: 1586–1591. *Mol Biol Evol*. 2007;24(8):1586–1591.
64. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–60.
65. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
66. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The genome analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–303.
67. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.