The chloroplast genome sequencing of two important annual *Trifolium* species *T. alexandrinum* and *T. resupinatum* and comparative analysis with other congeneric species

yanli xiong
Sichuan Agricultural University
yi xiong
Sichuan Agricultural University
jun he
Southwest University
qingqing yu
Sichuan Agricultural University
junming zhao
Sichuan Agricultural University
xiong lei
Sichuan Agricultural University
zhixiao dong
Sichuan Agricultual University
jian yang
Sichuan Agricultural University
yan peng
Sichuan Agricultural University
xinquan zhang
Sichuan Agricultural University
Xiao Ma (maroar@126.com)
Sichuan Agricultural University https://orcid.org/0000-0002-4491-3528

Research article

Keywords: chloroplast genome, divergence time, IR lacking, rearrangement, repetitive events, Trifolium

Posted Date: January 9th, 2020

DOI: https://doi.org/10.21203/rs.2.20410/v1

Abstract

Background: Chloroplast (cp) genome of most plant species has two typical inverted-repeats (IRs) regions. However, in some species this IR structure is lost for unknown reasons and the consequence still needs to be revealed . Here, we present whole cp genome sequencing of *Trifolium* alexandrinum (Egyptian clover) and *T. resupinatum* (Persian clover) from the IR lacking clade (IRLC) . Results: Global aligning of *T. alexandrinum* and *T. resupinatum* to other eight *Trifolium* species revealed a large amount of rearrangement and repetitive events in these ten species. We found that IR lacking species have lower GC content and higher percentage of repetition than IR containing species. Abundant single nucleotide polymorphisms (SNPs) and insertions/deletions (In/Dels) were discovered between those two species. As hypothetical cp open reading frame (ORF) and RNA polymerase subunits severally, two genes ycf1 and rpoC2 in the cp genomes, which both contain vast repetitive sequences and high Pi values (0.6656, 0.455) between *T. alexandrinum* and *T. resupinatum* , possessed highly variation among ten *Trifolium* species. Thus they could greatly influence evolutionary process of *Trifolium* species. In addition, IR containing and IR lacking *Trifolium* species were estimated to split during the upper Cretaceous period, which was potentially related to the violent crustal movement and sea-land changes. Conclusions: Cp genomes of *T. alexandrinum* and *T. resupinatum*, which belong to IRLC were sequenced and annotated in present study, and compared with cp genomes of other eight *Trifolium* species reported previously. This valuable information will provide insight into the evolution of IR lacking species. Nevertheless, further investigating of the detailed reason of IR lacking is still challenging, but it may be related to the violent crustal movement and sea-land changes of the Cretaceous period presented in this study.

Background

Trifolium L. (Papilionoideae), one of the largest genera in the Leguminosae, contains several important fodder species, such as *T. repens* (white clover), *T. pretense* (red clover), *T. alexandrinum* (Egyptian clover), *T. resupinatum* (Persian clover) and so on [1]. *Trifolium* species are also widely grown as green manure crops, and about 11 species including *T. alexandrinum* and *T. resupinatum* were introduced to subtropical zone of east Asia and have been reported to be excellently adapted to saline-alkali soil thus useful for agricultural production [2, 3]. *T. alexandrinum* is generally grown as an annual winter legume fodder crop in the Middle East, Mediterranean and the Indian subcontinent. Its aerial part can be used for cattle feed and the seeds are used as an antidiabetic treatment [4]. Furthermore, *T. alexandrinum* also contributes to soil fertility and improves soil physical characteristics [4]. *T. resupinatum*, an annual, prostrate or semi-erect branched legume, can supply highly palatable and nutritive pasture and hay [5]. What's more, it is also very important as a park, garden and green place plant [1].

As an important part of plant organelles and photosynthetic organ, chloroplast (cp) has played an irreplaceable role in the plants [6]. The cp genomes are not only essential for the study of plants light system for potentially improving the photosynthetic capacity and thus increase plant yield, but also suitable for phylogenic study for their maternal inheritance and highly conserved genomic structure [7]. The cp genome has a typically covalently closed circular molecule structure including a small single-copy

region (SSC), a large single-copy region (LSC) and two almost identical inverse repeats (IRs) regions [6]. A typical cp genome contains approximately 130 genes. Many of them participate in photosynthesis, some others also encode proteins or function in regulating gene transcription [7].

The IR regions with the average length between 10 Kb to 76 Kb were found in all families in the angiosperm plants and some gymnosperm and fern genus [8]. However, there are some exceptions like the clover genus (*Trifolium*) of the legume family (Leguminosae). Some species in *Trifolium* have a normal cp gnome and in some others the IR is lacking [8, 9]. IR lacking cp genome only exists in species belong to "refractory clade" of *Trifolium* L., which includes the *Trifolium* sections Lupinaster, Trifolium, Tricocephalum, Vesicastrum and Trifoliastrum [9]. Large quantities of repeated DNA, which are the main resource of sequence variation and allowed to appropriately assess the phylogenetic relationships, are found to typically exist in IR lacking subclover (*T. subterraneum*) based on the previous studies [10]. Those repetitive structure might cause sequence rearrangement via intra homologous recombination [8]. Leguminosae are accepted to have flourished since the Cretaceous period and IR is considered as the major feature of the cp genome in plants since 400 million of years ago (Mya). However, the time of divergence between IR lacking and IR containing species has not been reported yet. Phylogenetic relationships between IR containing and IR lacking species of *Trifolium* were well estimated using 58 protein-coding genes in cp genomes [10]. However, *T. alexandrinum* and *T. resupinatum* have not been studied yet. Variation among different species could provide a fascinating glimpse into the understanding of plant biology and diversity [7].

Here, cp genomes of *T. alexandrinum* and *T. resupinatum* were sequenced and annotated using next generation sequencing (NGS). We compared the sequence differences caused by nucleic acid polymorphism (Pi), In/Del and repetitive sequences, as well as the evolution pressure reflected by non-synonymous/synonymous (Ka/Ks) between these two species. Furthermore, these two species were compared with other eight (four IR containing species and four IR lacking species of *Trifolium*) congeneric species and divergent times were estimated. This study provides insights into evolution of IR lacking cp genomes.

Results

Features of the *T. alexandrium* and *T. resupinatum* cp genomes

More than 20 million ReadSum (pair-end reads) were yielded for *T. alexandrium* and *T. resupinatum*, with the Q20 and Q30 (the percentage of bases whose mass value is greater than or equal to 20, 30) were higher than 94% and 87%, respectively. We assembled them successfully based on the alignment of paired-end sequences to the reference of *T. medeseum* (Fig 1). The cp genomes of *T. alexandrium* and *T. resupinatum* were detected with IR lacking and have a size of 148,545 bp and 149,026 bp, respectively (Table 1). The GC content in the two cp genomes was about 34.09% and 33.80% overall, and 37.05% and 36.64% in coding sequences (CDS). A total of 112 and 109 genes were consisted in the complete cp genomes of *T. alexandrinum* and *T. resupinatum*, which were contained 31 and 37 tRNA, 75 and 66 mRNA, and 6 rRNA, and 13 and five genes possessing introns, respectively.

Table 1 Comparison of the ten species of *Trifolium* genus

Species	Genome length(bp)	GC content (%)		Gene density	tRNA	rRNA	mRNA	genes	genes with exons	GenBank number
	repetitive %	cp Genome	CDS							
T. alexandrinum	148545	34.09	37.05	7.54E-04	31	6	75	112	13	MN857160
	2.85%									
T. resupinatum	149026	33.8	36.64	7.31E-04	37	6	66	109	5	MN857161
	2.69%									
T. subterraneum	144763	34.83	37.1	7.60E-04	30	4	76	110	16	NC011828
	20.71%									
T. meduseum	142595	34.87	37.34	7.78E-04	30	4	77	111	10	NC476730.1
	12.83%									
T. pratense	121178	34.63	36.94	7.43E-04	28	4	58	90	9	KJ788290
	NA									
T. repens	132120	34.53	36.96	8.10E-04	31	4	72	107	10	KC894706.1
	20.70%									
*T.strictum**	125834	34.98	36.7	8.82E-04	31	5	75	111	11	NC025745.1
	0.71%									
*T.aureum**	126970	34.86	36.81	8.51E-04	30	4	74	108	9	KC894708.1
	5.60%									
*T.boissieri**	125740	35.24	36.83	8.75E-04	31	5	74	110	9	NC025743.1
	1.05%									
*T.glanduliferum**	126149	34.9	36.7	8.72E-04	30	5	75	110	11	NC025744.1
	0.78%									
Mean of IR lacking species	126173.25	34.45	37.01	7.63E-04	31.17	4.67	70.67	106.5	10.5	
	11.96%									
Mean of IR containing species	139704.5	35	36.76	8.70E-04	30.5	4.75	74.5	109.75	10	
	2.04%									

There were 46 genes related to photosynthesis in cp genomes of *T. alexandrinum* and *T. resupinatum* (Table 2), of which four genes *psbN*, *atpF*, *ndhA* and *ndhB* were specific for *T. alexandrinum*. These genes include the ones encoding subunit of Rubisco, subunits of photosystem I, subunits of photosystem II, subunits of ATP synthase, cytochrome b/f complex, c-type cytochrome synthesis and subunits of NADH dehydrogenase. Thirty-one genes were related to self-replication, including four ribosomal RNA genes and 27 transfer RNA genes, in which *trnT-CGU* was unique in *T. alexandrinum*. Besides, ten genes encoding ribosomal proteins and twelve were associated with transcription. Among them, *rps18*, *rpl2* and *rpoC1* were unique in *T. alexandrinum*. Furthermore, two genes *clpP* and *ycf3* with other functions were particular for *T. alexandrinum* (Additional file1: Table S1).

Introns are not subject to natural selection thus theoretically accumulate more mutations than exons. In this study, a total of seven genes (*atpF*, *clpP*, *ndhA*, *ndhB*, *rpoC1*, *rps18* and *tRNA-CGU*) only contained an intron in *T. alexandrinum* (Additional file1: Table S1). Other five genes *tRNA-UAA*, *tRNA-UAC*, *tRNA-UGC*, *tRNA-UUC* and *tRNA-UUU* all had an intron in *T. alexandrinum* and *T. resupinatum.* The exons length of

those five genes was more conserved compared with intron. In particular, *ycf3* had two introns in *T. alexandrinum*.

Table 2 List of genes annotated in the cp genomes of *T. alexandrinum* and *T. resupinatum.*

Category	Function	Name of gene					
Self-replication (31)	Ribosomal RNA Genes	*rrn4.5*	*rrn5*	*rrn16*	*rrn23*		
	Transfer RNA genes	*trnA-ACG*	*trnA-GUC*	*trnA-GUU*	*trnA-UCU*	*trnA-UGC**	*trnC-GCA*
		trnG-GCC	*trnG-UUC**	*trnG-UUG*	*trnH-GUG*	*trnL-CAA*	*trnL-UAA**
		trnL-UAG	*trnL-UUU**	*trnM-CAU*	*trnP-GAA*	*trnP-UGG*	*trnS-GCU*
		trnS-GGA	*trnS-UGA*	*trnT-CCA*	*trnT-CGU** (ale)	*trnT-GGU*	*trnT-GUA*
		trnT-UGU	*trnV-GAC*	*trnV-UAC**			
Ribosomal proteins (10)	Small subunit of ribosome (SSU)	*rps2*	*rps3*	*rps4*	*rps7*	*rps8*	*rps11*
		rps14	*rps15*	*rps18**/ale	*rps19*		
Transcription (12)	Large subunit of ribosome (LSU)	*rpl2* (ale)	*rpl14*	*rpl16*	*rpl20*	*rpl23*	*rpl32*
		rpl33	*rpl36*				
	RNA polymerase subunits	*rpoA*	*rpoB*	*rpoC1** (ale)	*rpoC2*		
Photosynthesis related genes (46)	Large subunit of Rubisco	*rbcL*					
	Subunits of Photosystem I	*psaA*	*psaB*	*psaC*	*psaI*	*psaJ*	*ycf4*
	Subunits of Photosystem II	*psbA*	*psbB*	*psbC*	*psbD*	*psbE*	*psbF*
		psbH	*psbI*	*psbJ*	*psbK*	*psbL*	*psbM*
		psbN (ale)	*psbT*	*psbZ*			
	Subunits of ATP synthase	*atpA*	*atpB*	*atpE*	*atpF** (ale)	*atpH*	*atpI*
	Cytochrome b/f complex	*petA*	*petB*	*petD*	*petG*	*petL*	*petN*
	C-type cytochrome synthesis gene	*ccsA*					
	Subunits of NADH dehydrogenase	*ndhA** (ale)	*ndhB** (ale)	*ndhC*	*ndhD*	*ndhE*	*ndhF*
		ndhG	*ndhH*	*ndhI*	*ndhJ*	*ndhK*	
Other genes (7)	Maturase	*matK*					
	Protease	*clpP** (ale)					
	Chloroplast envelope membrane protein	*cemA*					
	Subunit of acetyl-CoA	*accD*					
	Hypothetical open reading frames	*ycf1*	*ycf2*	*ycf3*** (ale)			

Note: *, Genes containing a single intron; **, Genes containing two introns; (ale), Genes that are particular for *Trifolium alexandrinum*; */ale, Genes that only have an intron in *Trifolium alexandrinum*.

Repeats analysis

Scattered repeating sequences (palindrome repeats and direct repeats) and simple sequencing repeats (SSRs) were analyzed respectively. Over all, percentage of repetitive sequences in IR containing species (2.035%, Table 1) was less than those IR lacking species (11.956%). A total of 1941 scattered repetitive sequences in *T. alexandrinum* cp genome were annotated, which was greater than *T. resupinatum* (1250).

The percentages of palindrome repeats (type P, 50.49%, Fig 2B) of *T. alexandrinum* were slightly larger than *T. resupinatum* (46.4%). A total of 370 (Fig 2A) and 383 SSRs (sizes ranged from 8 – 81 bp and 8 – 36 bp) were predicted in *T. alexandrinum* and *T. resupinatum* and 30.54% and 23.24% of them were distributed in genic regions. In particular, the majority of SSRs were located in *ycf1* (18 for *T. alexandrinum* and *T. resupinatum*), followed by *rpoC2* (11 for *T. resupinatum* and 9 for *T. alexandrinum*). The mononucleotide repeats were dominant (65.41% in *T. alexandrinum* and 74.93% in *T. resupinatum*), followed by trinucleotide repeats (25.68% in *T. alexandrinum* and 22.19% in *T. resupinatum*), in which the repeats of the polyadenine (poly A, 37.34% for *T. resupinatum* and *35.95%* for *T. alexandrinum*) and polythymine (poly T, 36.55% for *T. resupinatum* and 37.84% for *T. alexandrinum*) were much more than guanine (G) or cytosine (C) repeats (less than 1.35%). A total of 24 SSRs were identified to be shared by *T. alexandrinum* and *T. resupinatum (*Additional file2: Table S2; Fig 2A*).* The common repeat sequences larger than 30 bp with the longest length of 117 bp was showed in Fig 2C.

The relative synonymous codon usage analysis

The relative synonymous codon usage analysis (RSCU), which is considered to be a combination result of natural selection, species mutation and genetic drift, was analyzed (Fig 3; Additional file3: Table S3). The value for initiation codon AUG (RSCU = 2.9745) was much greater than GUG (RSCU = 0.0129). The values of three termination codons UAA, UAG and UGA were 1.6215, 0.5676 and 0.8109. A total of 46.97% (31 of 66, include three termination codons) of the codons were with the greater codon frequency (RSCU values more than one), in which 93.55% (29 of 31) were prefer A or U in the third sites. In the other codons with RSCU values less than one (including one), C or G were more general in the third position (88.57%, 31 of 35).

Ka/Ks, single nucleotide polymorphisms (SNPs) and insertions/deletions (In/Dels)

Single nucleotide polymorphisms (SNPs), mainly including transversion (Tv) and transition (Tn), along with insertions/deletions (In/Dels) could lead the non-synonymous (Ka) or synonymous (Ks) substitution. The total SNPs and In/Dels in every gene varied from 1 (*ndhE* and *psaC*) to 677 (*atpB*) with the total of 8560. Additionally, more In/Dels, Tn and Tv were detected in intergenic regions (5.66%, 17.11% and 38.70%) than genic regions (3.05%, 10.40% and 25.08%) (Fig 4; Additional file4: Table S4). The 62 shared protein-coding genes with variations were used to calculate the Ka/Ks (Additional file5: Table S5). The values of Ka and Ks were ranged from 0 (*ndhE*, *petD*, *psaI*, *psbA*, *psbB* and so on) to 3.0151 (*rps4*) and 0 (*petG*, *petN*, *trnR-ACG*, *trnL-CAA*, *trnI-CAU*, *trnI-CAU* and so on) to 2.9415 (*rps8*) and Ka/Ks varied from 0 (*ndhE*, *psbZ*, *psbA*, *psbJ* and so on) to 3.7723 (*rps4*, Fig 5), respectively. Nine genes including *rpoC2*, *ndhG*, *trnK-UUU*, *ccsA*, *ndhF*, *ycf1*, *rps4*, *psaC* and *rrn4.5* have Ka/Ks values above one, implying positive selection on these genes. The Pi values calculated by 88 common genes of *T. alexandrinum* and *T. resupinatum* were from 0 to 0.7867 (*trnI-CAU*). Twenty-one genes had a Pi values of 0. Ninteen of them were tRNA. The nine genes with Ka/Ka above one also possessed relatively high Pi values (Fig 6; Additional file6: Table S6).

Whole-cp genome comparison with other *Trifolium* species

In order to excavate the sequence divergence of *Trifolium* genus and further shed light on the evolutionary events, such as pseudogenization, gene mutation, rearrangement and gene loss, cp genomes of ten species including four IR containing species (*T. aureum*, *T. boissieri*, *T. glanduliflerum* and *T. strictum*) and six IR lacking species (*T. alexandrinum*, *T. resupinatum*, *T. repens*, *T. pratense*, *T. subterraneum* and *T. meduseum*) were compared. The results indicated that the average size of cp genomes of these IR lacking species (126173.25 bp, Table 1) and genic density (7.63E-04) were lower than the ones containing IR (139704.5 bp, 8.70E-04), and the latter held higher mean GC content of whole cp genome (35.00%). Only minor variations were detected in the total numbers of genes, tRNA and mRNA among the selected species. *T. pratense* possessed the smallest numbers of tRNA (28), mRNA (58) and total number of genes (90).

Furthermore, abundant gene rearrangements were detected among ten *Trifolium* species using MAUVE program and the *T. subterraneum* as the reference sequence (Fig 7). Compared with CDS, non-coding sequences (CNS) showed most significant variation among selected *Trifolium* species (Fig 8). In other words, sequences variation in genic regions was lower than intergenic regions in the cp genomes of those ten species.

Phylogenetic divergence time estimation

The CDS of 76 genes shared in cp genomes of the 20 species (18 of Papilionoideae, one of Caesalpinioideae and one of Mimosaceae) were subjected to phylogeny analysis and divergence times estimation (Fig 9). The topological structure of phylogenetic tree was almost consistent with the classification of Leguminosae with strong bootstrap support. Three subfamilies Papilionoideae, Caesalpinioideae and Mimosaceae were clearly separated. It is worth noting that *Glycine max* and *Lotus japonicus,* belonging to Papilionoideae, were evolutionally grouped with *Ceratonia siliqua* (Caesalpinioideae) and *Albizia odoratissima* (Mimosaceae). *Trifolium* species split from *Medicago* species during the Early Cretaceous (127.2288 Mya). It seems that during the Late Cretaceous (83.5049 Mya), the IR lacking *Trifolium* species diverged with IR containing *Trifolium* species.

Discussion

Base mutation of *T. alexandrinum* and *T. resupinatum* to IR containing species

Point mutation was generally more common than frame shift for natural mutation [19]. As expected, more SNPs (21963, 6618 Tn and 15345 Tv; Additional file4: Table S4) than In/Del (2097) were found between *T. alexandrinum* and *T. resupinatum*. What's more, 60% of them occurred in intergenic regions, which was consistent with the conservatism of CDS displayed in mVISTA (Fig 6). This discovery was in agreement with the hypothesis that CDS had a slower rate of evolution compared with CNS [20]. Furthermore, minuscule SNPs (159 between *Oryza sativa* and *O. nivara* [21], 330 between *Citrus sinensis* and *C. aurantiifolia* [22] and 231 between *Machilus yunnanensis* and *M. balansae* [23]) were identified in IR containing species, which were exceptionally smaller than the SNPs between two IR lacking species *T. alexandrinum* and *T. resupinatum* calculated in present study. As an important structure in stabilizing cp

genome, IR region can hold from deviating by selective force [24]. Thus the observed abundant SNPs/Indels in *T. alexandrinum* and *T. resupinatum* are not surprising.

The comparison between the Ka and Ks of genes is an important content of molecular evolution [25]. Most genes were subjected to the neutral selection and purification selection, however, there are also limited genes whose rate of Ka is higher than that of Ks because the function of the gene has been dramatically changed, called Darwinian positive selection [26]. Lacking one IR region is believed to directly enhance the nucleotide substitution rate of the single repeat sequence. Previous studies have shown that in the IR lacking cp genome, the nucleotide substitution rate in the remaining repeat region is comparable to that of the single repeat region, which is 2.3 times higher than that in the IR containing cp genome [27]. Here, seven protein-coding genes in the cp genome of *T. alexandrinum* and *T. resupinatum* (*rps4*, *rpoC2*, *ndhG*, *ccsA*, *ndhF*, *ycf1* and *psaC*) have high ratio of Ka to Ks, which led by high values of Ka but extremely low values of Ks, could imply that they are under positive selection and played an important role in enhancing individual adaptability in evolutionary process. *rps4* [28] and *rpoC2* [29] have been reported to be under positive selection in previous studies. However, beneficial mutations might be fixed in those genes and, thus, reduce genetic polymorphism at selected sites [30].

Global alignment between IR containing and IR lacking *Trifolium* species

Comparing the cp genomes of *T. alexandrinum* and *T. resupinatum* sequenced in present study to other eight (four IR containing species) *Trifolium* species showed relatively high conserved genome length and gene content though *T. pretense* possessed only 90 genes (Table 1). Furthermore, the high average GC content, which forms a more stable structure of genome, were observed in four IRs containing species *(T. strictum*, *T. aureum*, *T. boissieri* and *T. glanduliferum*) compared with IR lacking species (Table 1). According to Millen et al. [31], the vast of angiosperms held in shared 74 coding-protein genes but other five genes (*accD*, *ycf1*, *ycf2*, *rpl23* and *infA*) were only existed in some specific species. *Ycf1*, with the premature stop codons in the CDS thus always be defined as pseudogene or ORF in other angiosperm [32], was present in all the ten *Trifolium* species and with the biggest value of Pi (0.6656; Additional file6: Table S6) among 76 common genes in ten species. This is closely related to the fact that most pseudogenes have undergone the processes of accelerated mutation rate, decreased GC content, and decreased secondary structure stability [33]. Therefore, *ycf1* possesses considerable variation among different species thus could be considered as the good candidate gene for phylogenic study among *Trifolium* species.

As the driving force of evolution, repetitive sequences indicate that the genetic material of a species is continuously self-replicating in the process of evolution, thus greatly expanding and enriching the genetic information [34]. Present study revealed higher repetitive percentage (11.96%, Table 1) in IR lacking species than IR containing species (2.04%). The number of repeated sequences in cp genome are associated with rearrangement in some species [35]. However, the driving force of repetitive sequence was seemly related to nuclear genes and genomic recombination [10], such as homologous recombination and microhomology-mediated break-induced replication acting on more than 50 bp and less than 30 bp repeats, respectively. Known as “hotspots” for variation [36], *ycf1* and *rpoC2*, highly varied among ten

Trifolium species (Fig 8), possessed the majority of repetitive sequences in *T. alexandrinum* and *T. resupinatum* and high values of Pi, thus could have essential function in the evolutionary process of *Trifolium* species.

In general, there is a strong correlation between the presence of IR and structurally stabilization of cp genomes. Substantial rearrangement was usually found in cp genomes lacking IR [9]. Among those IR lacking species of Leguminosae such as alfalfa, subclover, pea, etc, some are structurally stable and have not been rearranged, some undergo intermediate rearrangements, while others experienced a series of complex rearrangements [9]. This study found abundant rearrangements within six IR lacking species or four IR containing cp genomes of *Trifolium* (Fig 7). According to Palmer and Thompson [8], IR could prevent the rearrangement of cp sequence to some extent, so the rearrangement probability will be increased in the species lacking IR sequence. However, some species of legumes were reported having obtained the ability to increase rearrangement, this could be why there are many rearrangement events detected among those ten *Trifolium* species [8]. However, lacking IR leads to increased rearrangement is only one of the explains.

Phylogeny analysis and divergent time

The topological structure of other eight *Trifolium* species using 76 protein coding genes in this study was generally in agreement with the report of Sveinsson and Cronk [9] by 58 protein coding genes. In addition, the phylogenetic location of tested *T. alexandrinum* and *T. resupinatum* was confirmed (Fig 9). Furthermore, *T. alexandrinum* and *T. pratense*, both belonging to *Trifolium* Sect. Trifolium, were clustered together and *T. resupinatum* was grouped with *T. repens* though those two species were in a separate group in Malaviya's study based on isozyme data [37]. Two IR containing species *T. boissieri* and *T. aureum* were predicted to differentiate with other eight species in late Cretaceous period, then another two IR containing species *T. strictum* and *T. glanduliferum* were further diverted with IR lacking species at about 14 Mya. In late Cretaceous period, violent crustal movement and sea-land changes led to a flourished development of angiosperms and IR lacking species might form at the same time. It looks as if the ancestor of some IR lacking species had gone through a battery of evolutionary alternation (including high rearrangement and repetition) and the precise mechanism of such evolutionary pattern is underway to illuminate.

Conclusion

Cp genomes of *T. alexandrinum* and *T. resupinatum,* which belong to inverted-repeat-lacking clade (IRLC) were sequenced and annotated in present study, and compared with cp genomes of other eight *Trifolium* species reported previously. The results revealed a high variation in CDS and abundant rearrangement within *Trifolium* genus. Compared to IR containing species, IR lacking species held lower GC content, higher SNP, In/Del and repeats. This valuable information will provide insight into the evolution of IR lacking species. Nevertheless, further investigating of the detailed reason of IR lacking is still challenging, but it may be related to the violent crustal movement and sea-land changes of the Cretaceous period presented in this study.

Methods

Plant material, DNA isolation and sequencing

Plant seeds of *T. alexandrinum* (cv 'Elite II') and *T. resupinatum* (cv 'Laser') were kindly provided by Barenbrug (Australia) then germinated in growth chamber (25°C, 300 μmols·m^2·s^{-1}; 16-h photoperiod). Total DNA was extracted from 50 mg of fresh leaves following the Plant DNA Isolation Kit (Tiangen, Beijing). Sequencing was performed on Illumina Novaseq sequencing platform (Illumina, San Diego, CA) according to protocol of Illumina manual (San Diego, CA, USA). Sheared low molecular weight DNA fragments were used to construct paired-end (PE) libraries then sequenced by Genepioneer Biotechnologies (Nanjing, China).

Cp genome assembly, annotation and visualization

Both the raw data for two *Trifolium* species were filtered according to the following criterion: reads of less than 5% unidentified nucleotides and more than 50% of its bases with the quality score of ≥20 were retained after adapter trimming. With the reference genome of *T. meduseum* [10] (National Center for Biotechnology Information, NCBI number KJ 788288), the cp DNA were assembled as follows. In order to decrease the difficulty of sequences assemble, filtered reads (clean data) were aligned to the cp genome database built by Genepioneer Biotechnologies (Nanjing, China) using bowtie2 v 2.2.4 and SPAdes v3.10.1 to acquire SEED sequences then obtained contigs by kmer iterative extend seed. Scaffolds connected by contigs using SSPACE v 2.0 were filled gap by Gapfiller v 2.1.1 until gain the integrated cp genome.

The results of CDS and rRNA were obtained using BLAST V 2.2.25 and HMMER V3.1 b2 via aligned to cp genome database of NCBI. ARAGORN V 1.2.38 and tRNAscan-SE search server (http://lowelab.ucsc.edu/tRNAscan-SE/) were used to predict and further check tRNA. Finally, the result of genome annotation was performed via Geneious (https://www.geneious.com, [11]) and visualized in OGDRAW (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html).

The relative synonymous codon usage analysis (RSCU) and simple repeating sequences (SSRs) prediction

The RSCU was analyzed using MEGA v7.0 to reflect the relative preference of a particular codon encoding the corresponding amino acid codon [12]. The values of RSCU more than one was considered as greater codon frequency. SSRs with the same repeats units and times and distributed in the genic regions were considered as shared repeats, the repetitive sequences were distinguished using VMATCH V2.3.0 (http://www.vmatch.de/) and MISA v1.0 (http://pgrc.ipk-gatersleben.de/misa/misa.html) based on the genomic data, which was also utilized to determine the mono-, di-, tri-, tetra-, penta- and hexa- nucleotides.

Sequence variation analysis and Ka/Ks

Whole cp genome alignment and collinearity analysis of sequenced species herein along with eight *Trifolium* species, namely *T. strictum* (NC025745.1, [9]), *T. aureum* (KC894708.1, [10]), *T. boissieri* (NC025743.1, [9]), *T. glanduliferum* (NC025744.1, [9]), *T. subterraneum* (NC011828, [13]), *T. meduseum*

(NC476730.1, [9]), *T. pratense* (KJ788290, [9]) and *T. repens* (KC894706.1, [10]) was implemented using Mauve [14] and mVISTA [15], respectively. Among those species, the former four contain IR and the latter four lack IR. Furthermore, the common genes of *T. alexandrium* and *T. resupinatum* tested in present study were utilized for Ka/Ks and nucleotide diversity (Pi) calculation. Ka/Ks, which was generally considered as a reflection of selection pressures, was computed via KaKs_Calculator v2.0 [16]. Pi, which could be used to estimate the degree of nucleotide sequences variation and further provide potential molecular markers for population genetics, was calculated using VCFTOOLS by sequences comparison of the CDS of the common genes of different species by MAFFT version 7.017 [17]. Finally, single nucleotide polymorphisms (SNPs) and insertions/deletions (In/Dels) of *T. alexandrium* and *T. resupinatum* were also identified using Mafft program [17].

Divergence time estimates

A total of 20 Leguminosae species including 18 Papilionoideae, one Caesalpinioideae and one Mimosaceae were utilized to assess the divergence time using BEAST v 1.7.3 package [18] with the Bayesian method. GTR+G+I substitution model with strict clock model and Yule model for Priors tree were applied for BEAUti along with MCMC analysis setting as follows, 10,000,000 of Chain length, 1,000 of Tracelog, 1,000 of screenlog, 1,000 of treelog.t: tree. The assessment of results was executed in Tracer v 1.5 (http://www.beast.bio.ed.ac.uk/) to confirm that the value of effective sample size (ESS) was greater than 200. Finally, the document of “tree” obtained from TreeAnnotator was visualized in Figtree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

Abbreviations

Cp: Chloroplast; IRs inverted-repeats; IRLC: IR lacking clade; SNP: single nucleotide polymorphisms; In/Dels: insertions/deletions; ORF: open reading frame; SSC: small single-copy region; LSC :large single-copy region; NGS: next generation sequencing; Pi: nucleic acid polymorphism; Ka: non-synonymous; Ks: synonymous; SSRs: simple sequencing repeats; RSCU: the relative synonymous codon usage analysis; Tv: transversion; Tn: transition; CDS: coding sequences; CNS: non-coding sequences; UTR: untranslated regions

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The annotated chloroplast genomes of *T. alexandrinum* and *T. resupinatum* have been deposited in the NCBI GenBank with the accession numbers MN857160 and MN857161.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the earmarked fund for Modern Agro-industry Technology Research System (No. CARS-34) and National Natural Science Foundation of China (3177131276).

Authors' contributions

XM, YP and YLX conceived and designed the study. YLX and YX participated in assembly and sequence analysis and wrote the draft paper. QQY and JMZ cultured and maintained the experimental material. ZXD and JY participated in the assembly of the genome sequences. XQZ and YX performed sequence analyses and generated the figures. JH and XL analyzed the data and prepared tables. All authors read and approved the final manuscript.

Acknowledgements

We thank test support of laboratory staff in the Department of Grassland Science, Animal Science and Technology College, Sichuan Agricultural University.

Authors' information

[1]College of Animal science and Technology, Sichuan Agricultural University, Chengdu, China. [2]State Key Laboratory of Exploration and Utilization of Crop Gene Resources in 10 Southwest China, Key Laboratory of Biology and Genetic Improvement of Maize in 11 Southwest Region, Ministry of Agriculture, Maize Research Institute of Sichuan 12 Agricultural University, Chengdu 600031, China.

References

1. Sabudak T, Guler N. *Trifolium* L. - A review on its phytochemical and pharmacological profile. Phytotherapy Research. 2009;23(3):439-446. doi:10.1002/ptr.2709.
2. Steiner JJ. Molecular phylogenetics of the clover genus (*Trifolium*–Leguminosae). Molecular Phylogenetics & Evolution. 2006;39(3):688-705. doi:10.1016/j.ympev.2006.01.004.
3. Turpin JE, Herridge DF, Robertson MJ. Nitrogen fixation and soil nitrate interactions in field-grown chickpea (*Cicer arietinum*) and fababean (*Vicia faba*). Crop & Pasture Science. 2002;53(5):599-608.

doi: 10.1071/AR01136.

4. Bakheit RB. Egyptian clover (*Trifolium alexandrinum* L.) breeding in Egypt. Asian Journal of Crop Sci. 2013;5:325-337. doi:10.3923/ajcs.2013.
5. Nazir M, Shah FH. Studies on persian clover (*Trifolium resupinatum*). Plant Foods for Human Nutrition. 1985;1(37):3-8. doi: 10.1007/BF01092017.
6. Douglas SE. Chloroplast origins and evolution. Advances in Photosynthesis. 1994;91-118. doi: 10.1007/978-94-011-0227-8_5.
7. Daniell H, Lin C, Yu M, Chang W. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biology. 2016;17(1):134. doi: 10.1186/s13059-016-1004-2.
8. Palmer JD, Osorio B, Aldrich J, Thompson WF. Chloroplast DNA evolution among legumes: Loss of a large inverted repeat occurred prior to other sequence rearrangements. Current Genetics. 1987;11(4):275-286. doi: 10.1007/BF00355401.
9. Sveinsson S, Cronk Q. Evolutionary origin of highly repetitive plastid genomes within the clover genus (*Trifolium*). BMC Evolutionary Biology. 2014;14(1):228. doi: 10.1186/s12862-014-0228-6.
10. Barrett CF, Freudenstein JV, Li J, Mayfield-Jones DR, Perez L, Pires JC, Santos C. Investigating the path of plastid genome degradation in an early-transitional clade of heterotrophic Orchids, and implications for heterotrophic angiosperms. Molecular Biology & Evolution. 2014;31(12):3095-112. doi: 10.1093/molbev/msu252.
11. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647-1649. doi: 10.1093/bioinformatics/bts199.
12. Kumar S, Nei M, Dudley J, Tamura K. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings in Bioinformatics. 2008;9(4):299-306. doi:10.1093/bib/bbn017.
13. Cai Z, Guisinger M, Kim H, Ruck E, Blazier JC, Mcmurtry V, et al. Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. Journal of Molecular Evolution. 2008;67(6):696-704. doi: 10.1007/s00239-008-9180-7.
14. Darling ACE, Mau B, Blattner FR, Perna ANT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Research. 2004;14(7):1394-1403. doi:10.1101/gr.2289704.
15. Cheng H, Li J, Zhang H, Cai B, Mi L. The complete chloroplast genome sequence of strawberry (*Fragaria* × *ananassa* Duch.) and comparison with related species of Rosaceae. Peerj. 2017;5(10):e3919. doi:10.7717/peerj.3919.
16. Wong KS. KaKs_calculator: calculating Ka and Ks through model selection and model averaging. Genomics,Proteomics & Bioinformatics. 2006;4:60-64. doi:CNKI:SUN:GPBI.0.2006-04-007
17. Kazutaka K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology & Evolution. 2013;4(30):772-780. doi: 10.1093/molbev/mst010.

18. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology & Evolution. 2012; 29(8):1969-1973. doi:10.1093/molbev/mss075.
19. Raes J, Peer YVD. Functional divergence of proteins through frameshift mutations. Trends in Genetics Tig. 2005;21(8): 428-431. doi:10.1016/j.tig.2005.05.013.
20. Small RL, Cronn RC, Wendel JF. Use of nuclear genes for phylogeny reconstruction in plants. Australian Systematic Botany, 2004;17(2):145-170. doi:10.1071/sb03015.
21. Masood MS, Nishikawa T, Fukuoka S, Njenga PK, Tsudzuki T, Kadowaki K. The complete nucleotide sequence of wild rice (*Oryza nivara*) chloroplast genome: first genome wide comparative sequence analysis of wild and cultivated rice. Gene. 2004;340(1):0-139. doi:10.1016/j.gene.2004.06.008.
22. Su HJ, Hogenhout SA, Al-Sadi AM, Kuo CH. Complete chloroplast genome sequence of omani lime (*Citrus aurantiifolia*) and comparative analysis within the Rosids. Plos One. 2014;11(9):e113049. doi:10.1371/journal.pone.0113049.
23. Yu S, Wenpan D, Bing L, Chao X, Xin Y, Jie G, et al. Comparative analysis of complete chloroplast genome sequences of two tropical trees *Machilus yunnanensis* and *Machilus balansae* in the family Lauraceae. Frontiers in Plant Science. 2015;6:662-670. doi:10.3389/fpls.2015.00662.
24. Doorduin L, Gravendeel B, Lammers Y, Ariyurek Y, Chin-A-Woeng T, Vrieling K. The complete chloroplast genome of 17 individuals of pest species *Jacobaea vulgaris*: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. DNA Research. 2011;18(2):93-105. doi:10.1093/dnares/dsr002.
25. Nei M, Kumar S. Molecular evolution and phylogenetics. Oxford University Press; 2000.
26. Yi L. Comparing and analyzing on models of calculation and statistical testing of nosynonymous substitution rate and synonymous substitution rate during gene evolution. Journal of Qujing Normal University. 2006;25(6):1-6.
27. Perry AS, Wolfe KH. Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. Journal of Molecular Evolution. 2002;55(5):501-508. doi:10.1007/PL00020998.
28. Bittner-Eddy PD, Crute IR, Holub EB, Beynon JL. RPP13 is a simple locus in *Arabidopsis thaliana* for alleles that specify downy mildew resistance to different avirulence determinants in *Peronospora parasitica*. Plant Journal for Cell & Molecular Biology. 2010;21(2):177-188. doi:10.1046/j.1365-313x.2000.00664.x.
29. Dong WL, Wang RN, Zhang NY, Fan WB, Fang MF, Li ZH. Molecular evolution of chloroplast genomes of Orchid species: insights into phylogenetic relationship and adaptive evolution. International Journal of Molecular Sciences. 2018;19(3):694-716. doi:10.3390/ijms19030716.
30. Zhang L. Collection and annotation of Suinong14 full-length transcripts and gene diversity analysis of Glyma13g21630. Chinese academy of agricultural sciences, 2011;37(10)：1724-1734. doi:10.3724/SP.J.1006.2011.01724.
31. Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, et al. Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. The Plant Cell, 2001;13(3):645-658. doi:10.2307/3871412.

32. Curci PL, De Paola D, Donatella D, Vendramin GG, Sonnante G. Complete chloroplast genome of the multifunctional crop globe artichoke and comparison with other Asteraceae. Plos One;10(3):e120589. doi:10.1371/journal.pone.0120589.
33. Xiao L. Intra-genomic polymorphism in the internal transcribed spacer (ITS) regions of *Cycas revoluta*: evidence of incomplete concerted evolution. 2009;17(5):476-481. doi:10.3724/SP.J.1003.2009.09100.
34. Flavell RB, Bennett MD, Smith JB, Smith DB. Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochemical genetics. 1974;12(4):257-269. doi:10.1007/BF00485947.
35. Haberle RC, Fourcade HM, Boore JL, Jansen RK. Extensive rearrangements in the chloroplast Genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. Journal of Molecular Evolution. 2008;66(4):350-361. doi:10.1007/s00239-008-9086-4.
36. Wei YL, Wen ZF, Liu F, et al. Bioinformatics analysis of *ycf1* gene in *Corylus*. Journal of Shanxi Agricultural Science. 2018;46(8):1244-1247. doi:10.3969/j.issn.1002-2481.2018.08.04.
37. Malaviya DR, Roy AK, Kaushal P, Kumar B, Tiwari A. Genetic similarity among *Trifolium* species based on isozyme banding pattern. Plant Systematics & Evolution. 2008; 276(1-2):125-136. doi:10.1007/s00606-008-0070-7.

Additional Files

Additional file1: Table S1. Location and length of intron-containing genes in the chloroplast genomes of *T. alexandrinum* and *T. resupinatum*.

Additional file2: Table S2. The shared repeats of *T. alexandrinum* and *T. resupinatum*. * means the shared location for *T. alexandrinum* and *T. resupinatum*, ^{res} means locations particular for *T. resupinatum*, the numbers of “Number” mean number of repeats in *T. alexandrinum* and *T. resupinatum*, respectively.

Additional file3: Table S3. The relative synonymous codon usage (RSCU) analyzed using CodonW.

Additional file4: Table S4. Transversion (Tv) and transition (Tn) were detected between *T. alexandrinum* and *T. resupinatum*.

Additional file5: Table S5. The synonymous/synonymous substitution rates (Ka/Ks) calculated using 62 shared genes in *T. alexandrinum* and *T. resupinatum*.

Additional file6: Table S6. The nucleic acid polymorphism (Pi) computed using 88 common genes of *T. alexandrinum* and *T. resupinatum*.

Figures

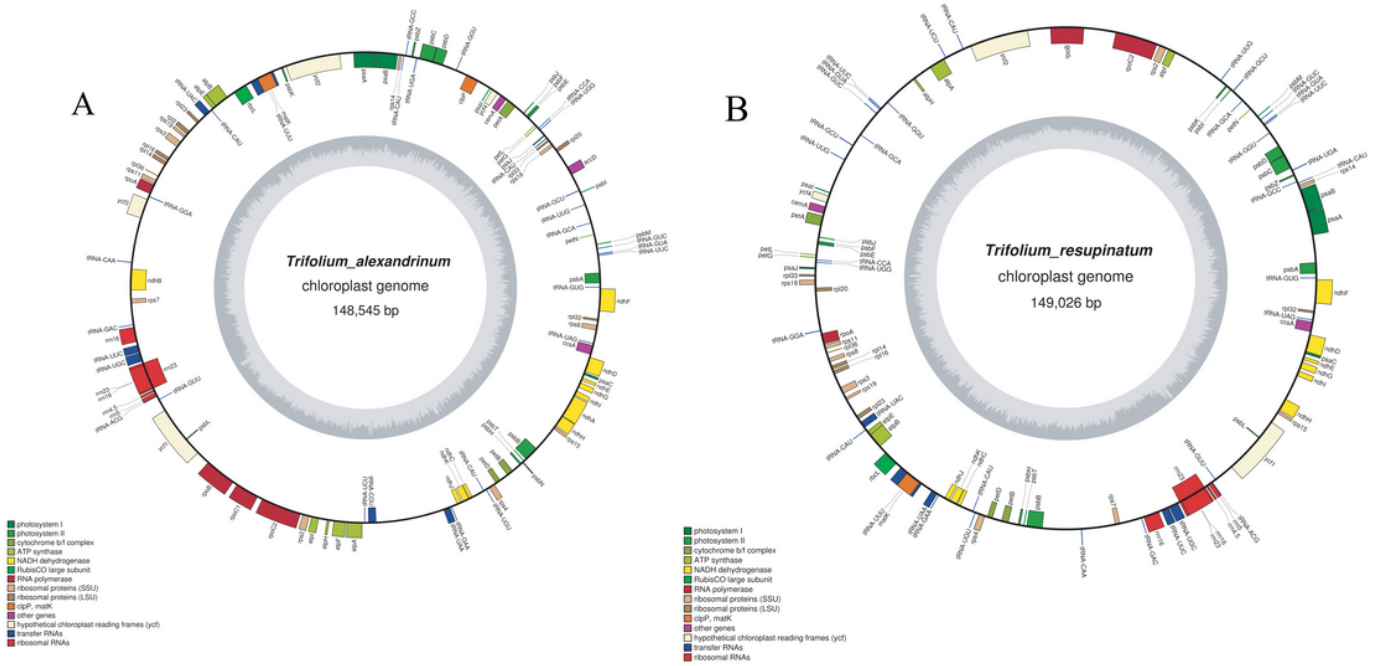

Figure 1

Gene maps of T. alexandrinum and T. resupinatum. Notes: Genes drawn inside and outside of the circle are transcribed clockwise and counterclockwise, respectively. Genes belonging to different functional groups are color coded. The darker gray color and lighter gray color in the inner circle corresponds to the GC content and the AT content, respectively.

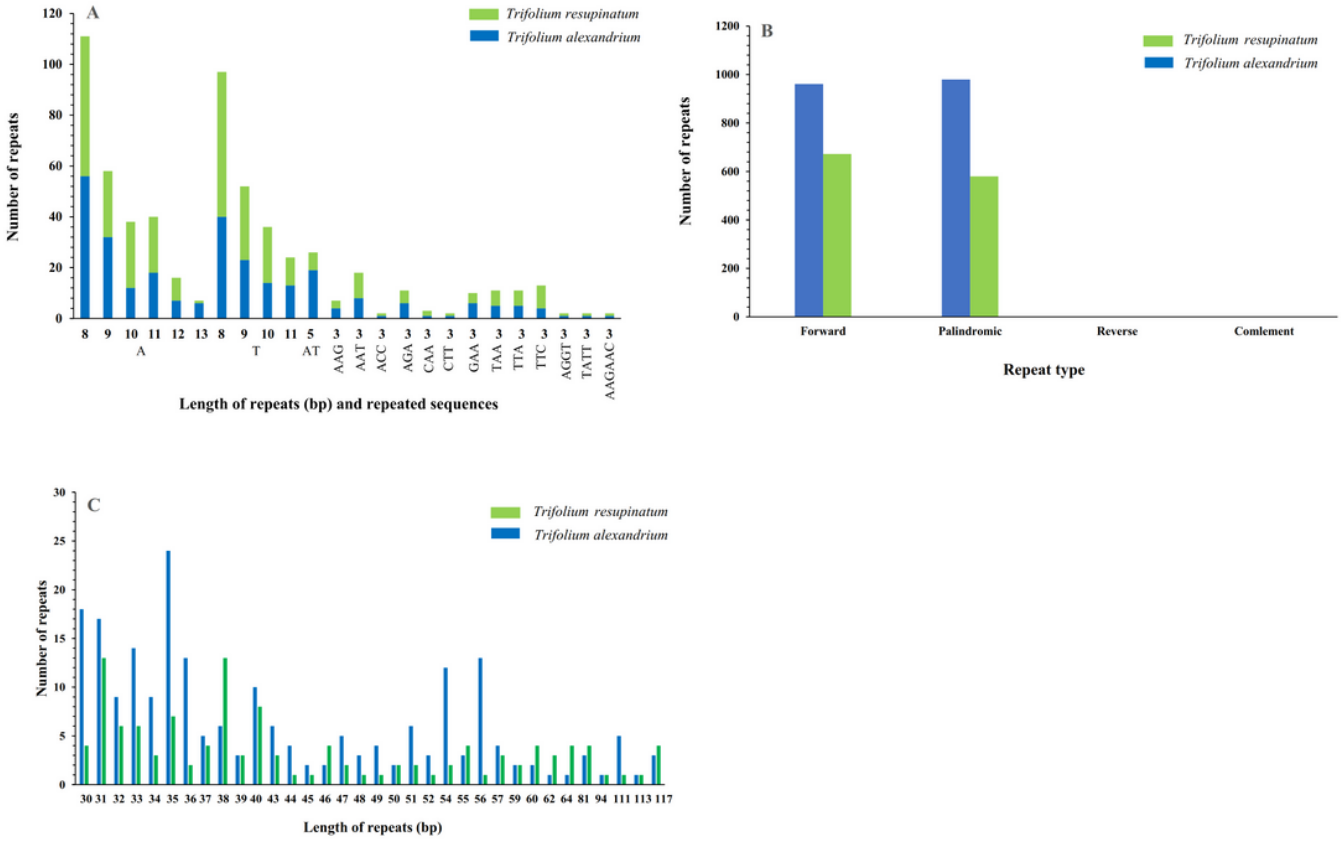

Figure 2

Repeating events of shared genes of T. alexandrinum and T. resupinatum. (A) Shared length of repeats and repeated sequences of T. alexandrinum and T. resupinatum; (B) Repeat type predicted in T. alexandrinum and T. resupinatum and (C) Listing of shared repetitive sequences with more than 30 bp.

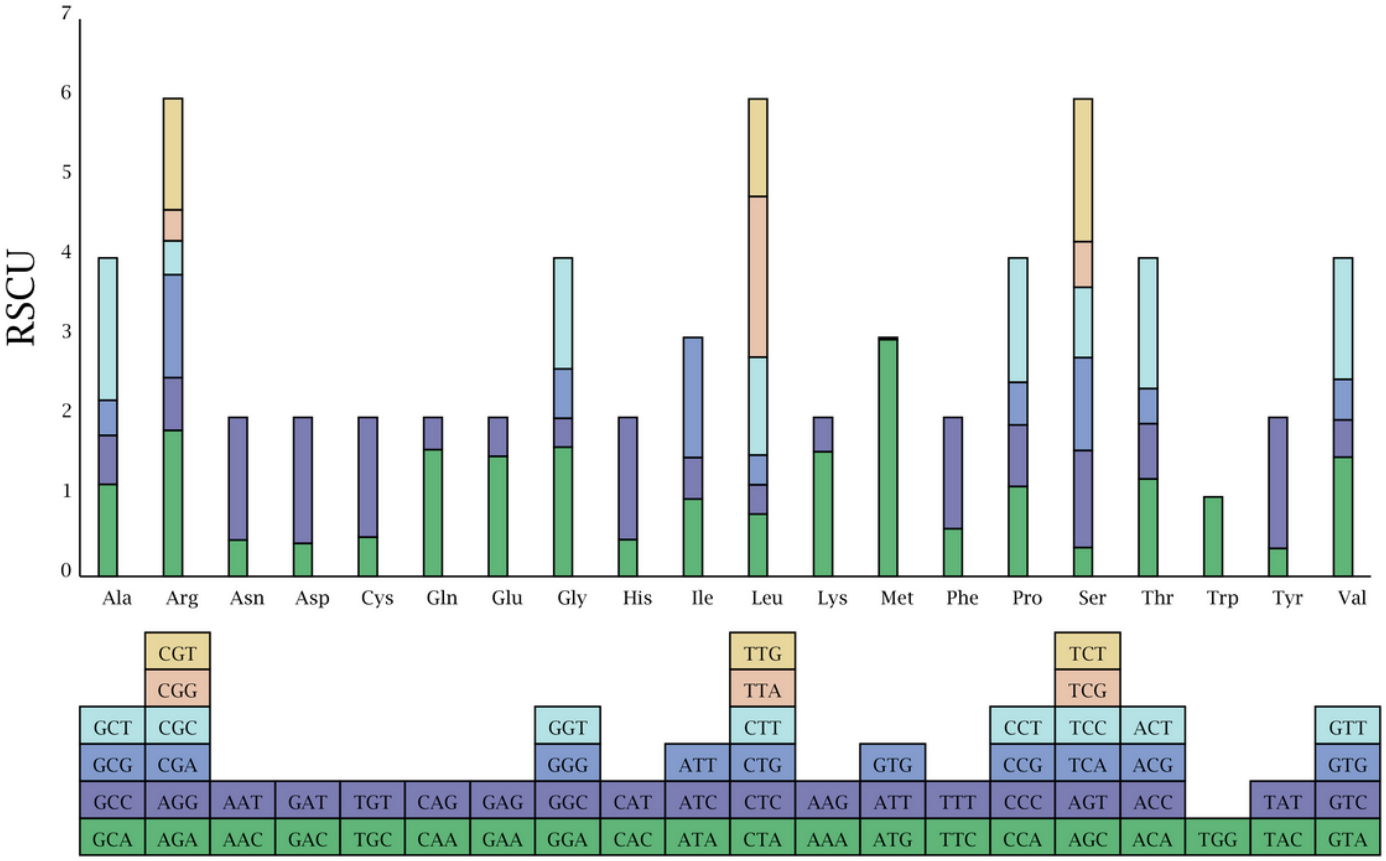

Figure 3

The relative synonymous codon usage (RSCU) for the T. alexandrinum and T. resupinatum.

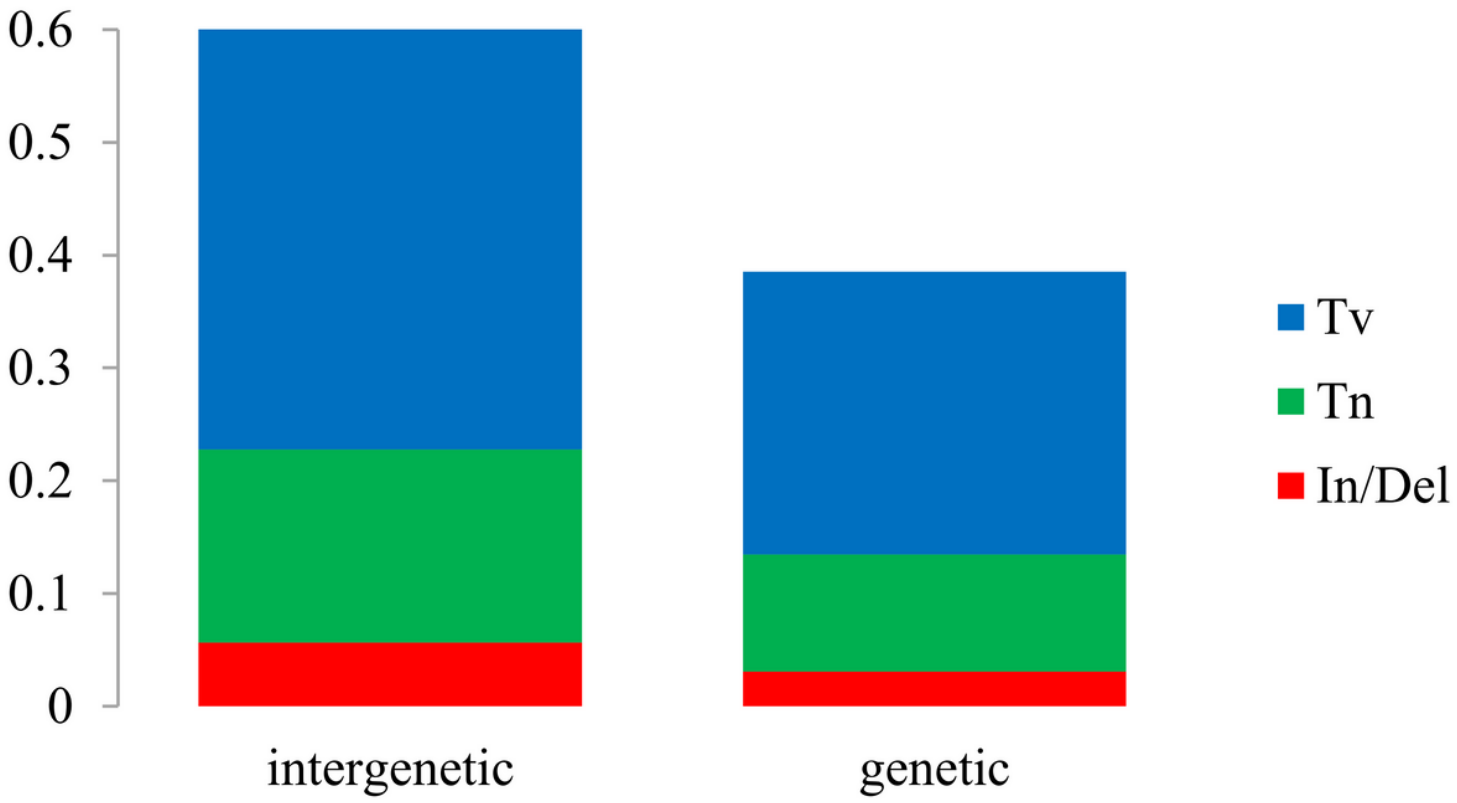

Figure 4

Transversion (Tv), transition (Tn) and Insert/Deletion (In/Del) were showed in intergenic and genic regions.

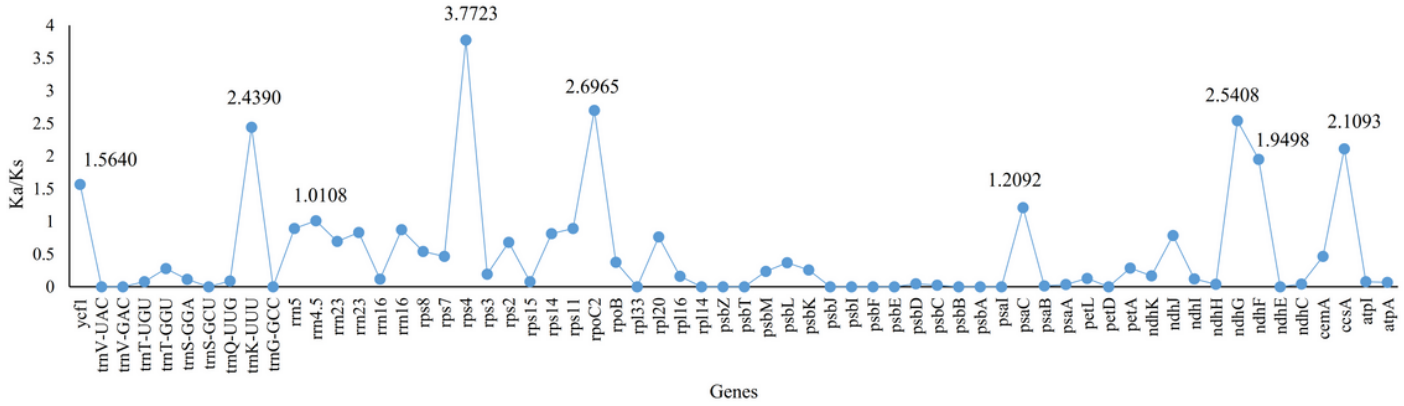

Figure 5

The synonymous/synonymous substitution rates (Ka/Ks) calculated using 62 shared genes in T. alexandrinum and T. resupinatum.

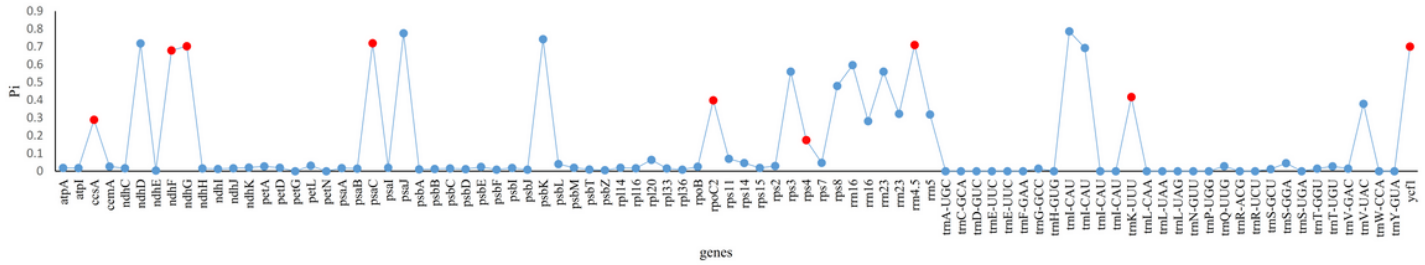

Figure 6

The nucleic acid polymorphism (Pi) computed using 88 common genes of T. alexandrinum and T. resupinatum. Genes with more than one Ka/Ks were red coded.

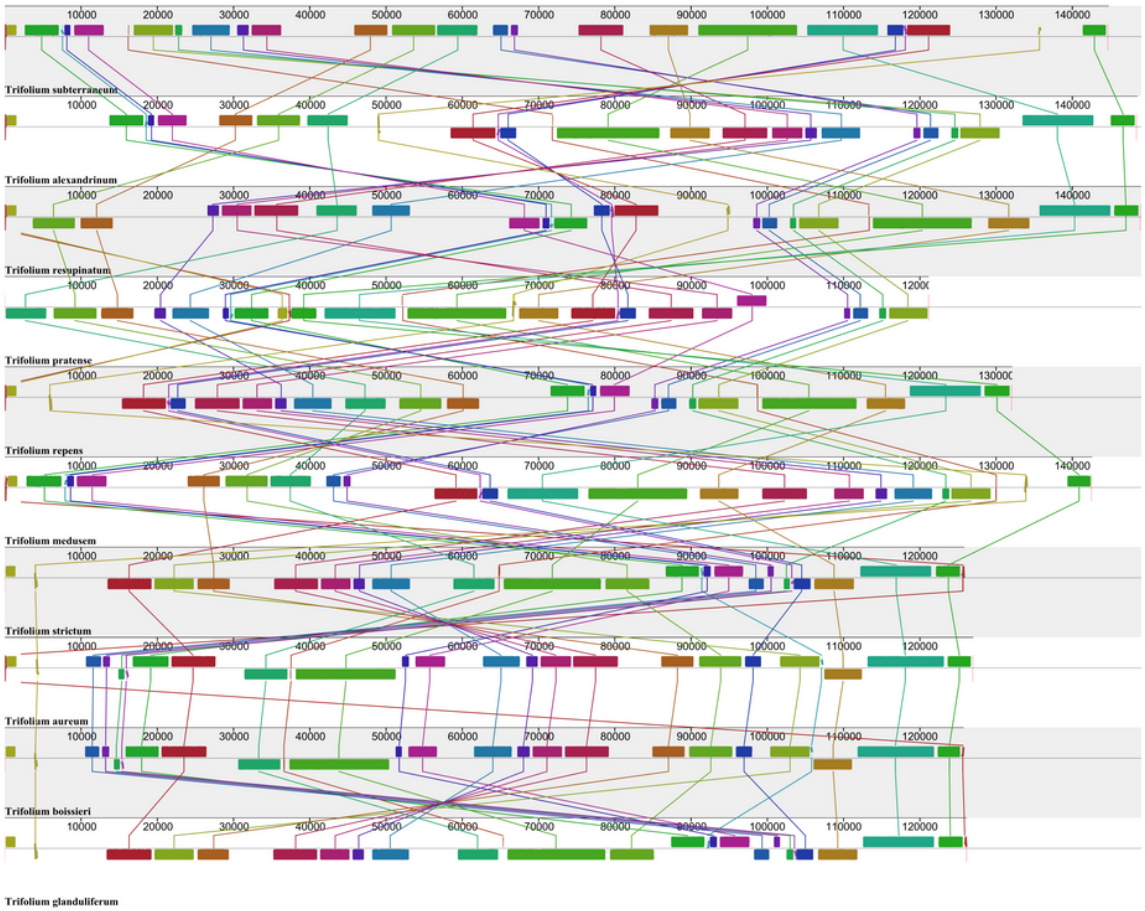

Figure 7

Synteny comparison of ten Trifolium chloroplast genomes with the reference of T. subterraneum using Mauve. Rectangular blocks with the same color indicate collinear regions.

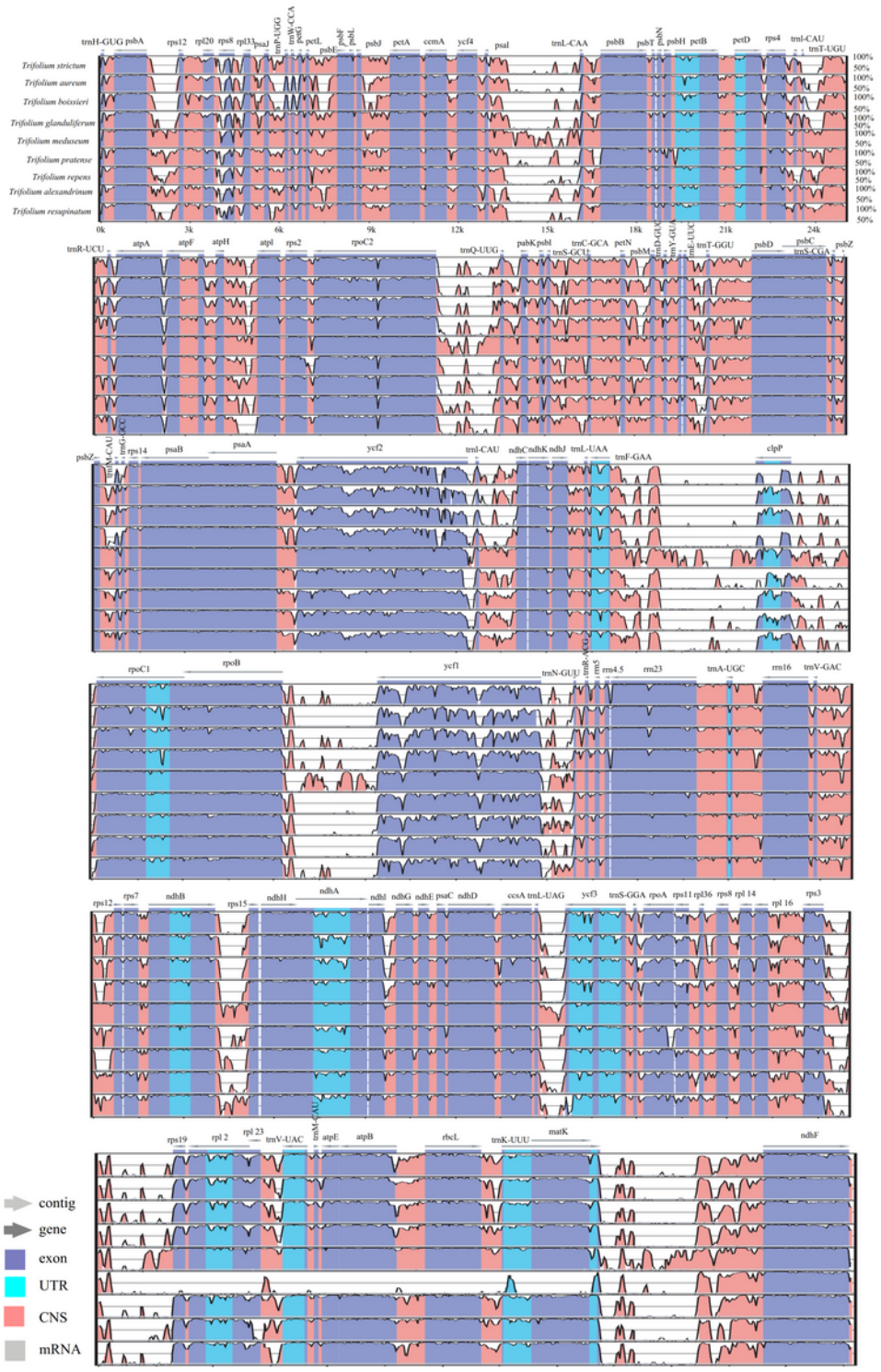

Figure 8

mVISTA alignment of the ten Trifolium cp genome sequences with T. subterraneum as the reference. The horizontal axis indicated the coordinates within the cp genome. The vertical scale ranging from 50 to 100% indicated the percentage of identity. Genome regions are color coded as protein coding, exon, mRNA, untranslated regions (UTR) and conserved non-coding sequences (CNS).

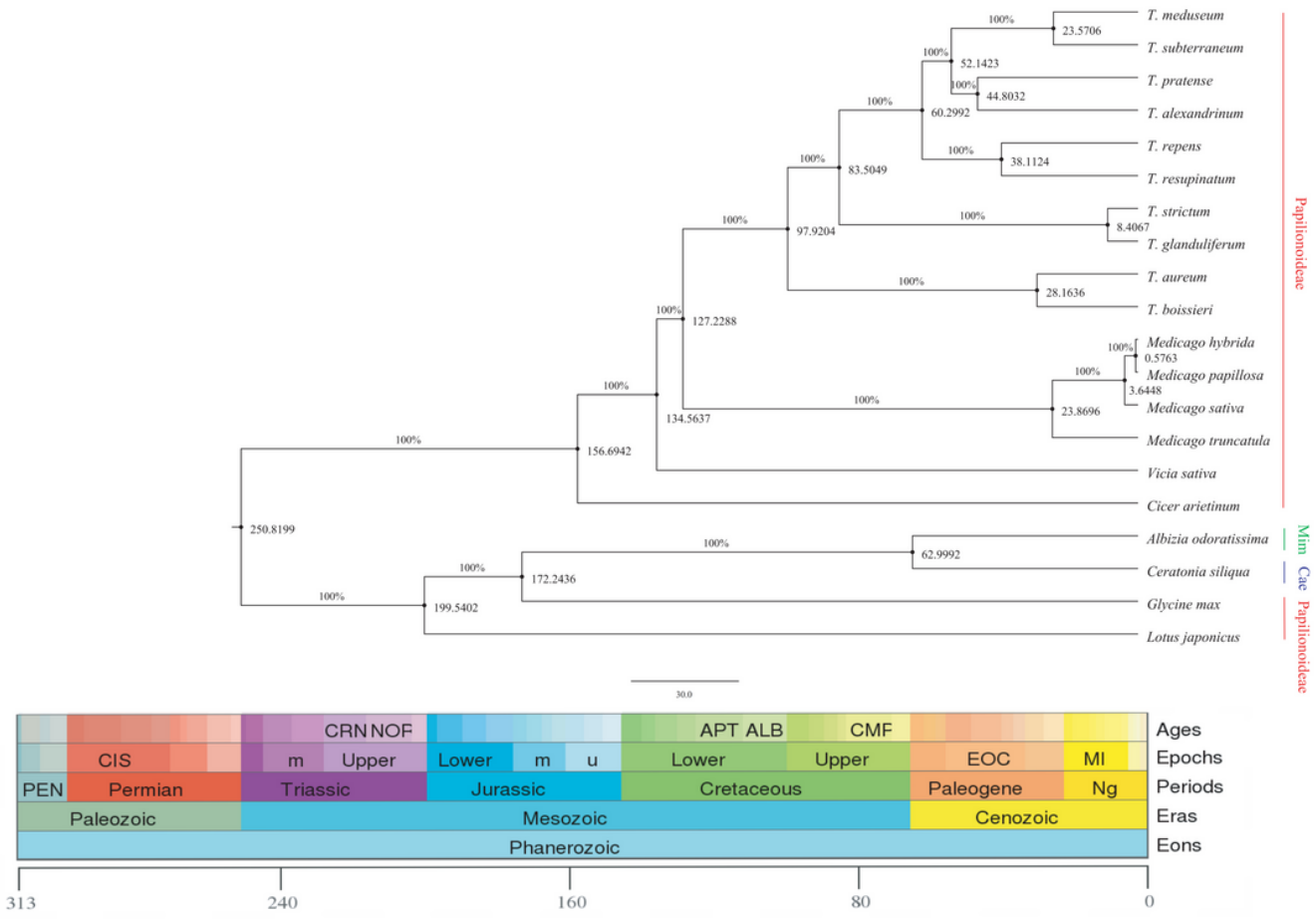

Figure 9

BEAST chronogram of the 20 Leguminosae species based on the common protein-coding genes. Geologic timescale was obtained from TIMETREE, time is shown in millions of years (MYA). Min, Mimosaceae; Cas, Caesalpinioideae.

Supplementary Files

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

- TableS4.xlsx
- TableS5.xlsx
- TableS1.docx
- TableS3.docx
- TableS2.docx
- TableS6.xlsx