

Article

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

Yanli Xiong ^{1,†}, Yi Xiong ^{1,†} , Jun He ², Qingqing Yu ¹ , Junming Zhao ¹, Xiong Lei ¹, Zhixiao Dong ¹, Jian Yang ¹, Yan Peng ¹, Xinquan Zhang ^{1,*} and Xiao Ma ^{1,*} 

¹ College of Animal science and Technology, Sichuan Agricultural University, Chengdu 611130, China; yanlimaster@126.com (Y.X.); xiongyi95@126.com (Y.X.); yuqinggz93@126.com (Q.Y.); junmingzhao163@163.com (J.Z.); lxforage@126.com (X.L.); dongzhixiao94@126.com (Z.D.); iwanwin@126.com (J.Y.); pengyanlee@163.com (Y.P.)

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; hejunyer@outlook.com

* Correspondence: zhangxq@sicau.edu.cn (X.Z.); maroar@126.com (X.M.); Tel.: +86-28-86291010 (X.Z. & X.M.)

† These authors contributed equally to this work.

Received: 3 March 2020; Accepted: 4 April 2020; Published: 9 April 2020



Abstract: *Trifolium* L., which belongs to the IR lacking clade (IRLC), is one of the largest genera in the Leguminosae and contains several economically important fodder species. Here, we present whole chloroplast (cp) genome sequencing and annotation of two important annual grasses, *Trifolium alexandrinum* (Egyptian clover) and *T. resupinatum* (Persian clover). Abundant single nucleotide polymorphisms (SNPs) and insertions/deletions (In/Dels) were discovered between those two species. Global alignment of *T. alexandrinum* and *T. resupinatum* to a further thirteen *Trifolium* species revealed a large amount of rearrangement and repetitive events in these fifteen species. As hypothetical cp open reading frame (ORF) and RNA polymerase subunits, *ycf1* and *rpoC2* in the cp genomes both contain vast repetitive sequences and observed high Pi values (0.7008, 0.3982) between *T. alexandrinum* and *T. resupinatum*. Thus they could be considered as the candidate genes for phylogenetic analysis of *Trifolium* species. In addition, the divergence time of those IR lacking *Trifolium* species ranged from 84.8505 Mya to 4.7720 Mya. This study will provide insight into the evolution of *Trifolium* species.

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

1. Introduction

Trifolium L. (Leguminosae, Fabaceae), one of the largest genera in the Leguminosae, contains several important fodder species, such as *T. repens* (white clover), *T. pratense* (red clover), *T. alexandrinum* (Egyptian clover), *T. resupinatum* (Persian clover) amongst others [1]. *Trifolium* species are also widely grown as green manure crops, and about 11 species, including *T. alexandrinum* and *T. resupinatum*, were introduced to the subtropical zone of east Asia and have been reported to be excellently adapted to saline-alkali soil thus useful for agricultural production [2]. *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 [3]. Furthermore, *T. alexandrinum* also contributes to soil fertility and improves soil physical characteristics [3]. *T. resupinatum*, an annual, prostrate or semi-erect branched legume, can supply highly palatable and nutritive pasture and hay [4]. What's more, it also has economic significance for the ornamental and landscape industries [4].

As an important part of plant organelles and photosynthetic organ, chloroplast (cp) has played an irreplaceable role in plants [5]. The cp genomes are not only essential for the study of plants light system for potentially improving the photosynthetic capacity and thus increasing plant yield, but are also commonly employed for phylogenetic study for their maternal inheritance and highly conserved genomic structure [6]. 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 [5]. 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 [6].

Possessing the function of replication initiation, genome stabilization and gene conservation [7], the IR regions with the average length between 10 Kb to 76 Kb were found in almost families in the angiosperm plants and some gymnosperm and fern genus [8]. However, a lack of IRs was found in the clover genus (*Trifolium*) of the legume family (Leguminosae) and coniferous cp genomes, such as *Medicago* L., *Melilotus* Miller, *Ononis* L. and so on [2]. A previous study has shown that differential elimination of genes from the IR regions may result in IR lacking in Ulvophyceae [9]. Meanwhile, gene loss including *ycf2* and *psbA* lead to IR lacking from cp genomes of *Cedrus wilsoniana* and *C. deodara* [10]. However, what causes the IR lack of cp genomes in *Trifolium* species still remains to be elucidated.

The lack of IR generates a highly rearranged cp genome, thus experiencing gene losses and inversions in the SSC or LSC regions [11]. Furthermore, quadruple of repeated DNA compared to related legume species is found to typically exist in IR lacking subclover (*T. subterraneum*) based on previous studies [12]. However, this unusual repeat-rich cp genome structure was also reported in other IR lacking *Trifolium* species, such as *T. repens*, *T. meduseum*, *T. semipilosum* and so on [13]. Those repetitive structures might be related to the sequence rearrangement of cp genomes via intra homologous recombination [8] in some angiosperm lineages like the Campanulaceae [14]. However, what causes the repeat-rich cp genomes structures of the *Trifolium* species is still worth studying, although previous studies have suggested that it may be related to the nuclear genes [13].

Insights gained from cp genome sequences and structure have improved the revelation of variation among plant species and made significant contributions to phylogenetic analysis [6]. Leguminosae are accepted to have flourished since the Cretaceous period and the phylogenetic relationships among some *Trifolium* species were well estimated using 58 protein-coding genes in cp genomes [13]. However, two important annual *Trifolium* species, *T. alexandrinum* and *T. resupinatum* have not been included in any previous study. Variation among different species could provide a fascinating glimpse into the understanding of plant biology and diversity [6]. Here, cp genomes of *T. alexandrinum* and *T. resupinatum* were sequenced and annotated. We compared the sequence differences caused by nucleotide diversity (Pi), In/Del and repetitive sequences, as well as the evolution pressure reflected by non-synonymous/synonymous (Ka/Ks) between these two species. Furthermore, they were compared with further thirteen IR lacking congeneric species and divergent times were estimated. This study provides insights into the evolution of IR lacking cp genomes.

2. Results

2.1. Features of the *T. alexandrinum* and *T. resupinatum* cp Genomes

More than 20 million ReadSum (pair-end reads) were yielded from *T. alexandrinum* and *T. resupinatum*, with the Q20 and Q30 (the percentage of bases whose mass value is greater than or equal to 20, 30) higher than 94% and 87%, respectively. We assembled them successfully based on the alignment of

paired-end sequences to the reference of *T. medeseum* (Figure 1). The cp genomes of *T. alexandrinum* and *T. resupinatum* were detected with a lack of IR and have a size of 148,545 bp and 149,026 bp, respectively (Table 1). The GC content observed in the two cp genomes was 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 contains 31 and 37 tRNA, 75 and 66 mRNA, and 6 rRNA, and 13 and five genes possessing introns, respectively. In particular, there were two *rrn* 16 genes in each of the *Trifolium* species with the unidentical sequences.

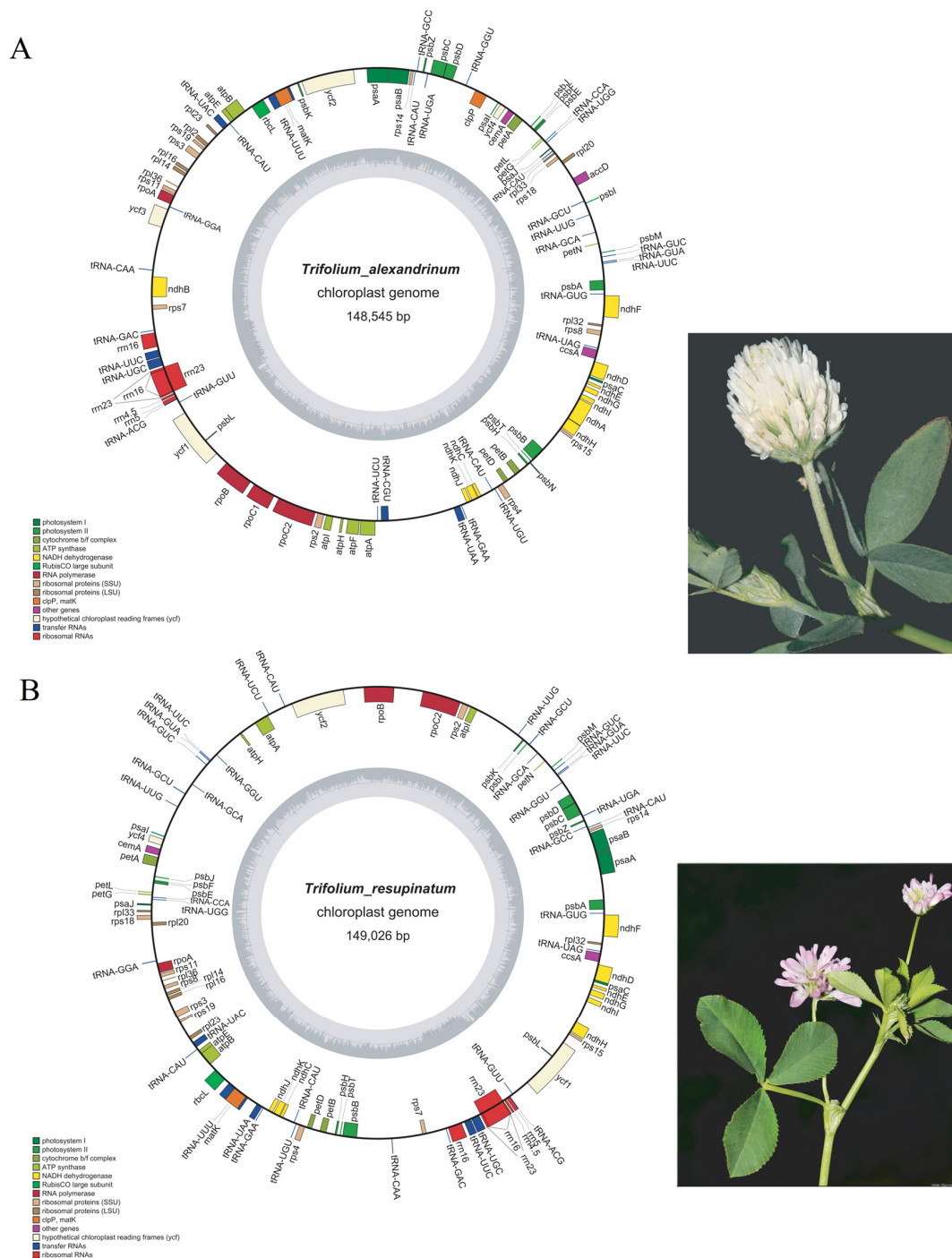


Figure 1. Gene maps of *T. alexandrinum* (A), *T. resupinatum* (B) cp genomes and pictures of *T. alexandrinum* and *T. resupinatum* were shown right in the gene maps. Pictures were cited from the open website <http://plants-of-styria.uni-graz.at>. Genes drawn inside and outside of the circle are transcribed 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.

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 subunits 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 encoded ribosomal proteins and twelve were associated with transcription. Among them, *rpl2* and *rpoC1* were unique in *T. alexandrinum*. Furthermore, three genes *clpP*, *accD* and *ycf3* with other functions were particular for *T. alexandrinum* (Table S1).

Introns are generally 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* (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 the intron. In particular, *ycf3* had two introns in *T. alexandrinum*.

Table 1. Comparison of the fifteen *Trifolium* species.

Species	Genome Length(bp) Repetitive %	GC Content (%)		Gene Density	tRNA	rRNA	mRNA	Genes	Genes with Introns	GenBank Number
		cp Genome	CDS							
<i>T. alexandrinum</i>	148545/2.85%	34.09	37.05	7.54×10^{-4}	31	6	75	112	13	MN857160
<i>T. resupinatum</i>	149026/2.69%	33.80	36.64	7.31×10^{-4}	37	6	66	109	5	MN857161
<i>T. subterraneum</i>	144763/20.71%	34.83	37.10	7.60×10^{-4}	30	4	76	110	16	NC011828
<i>T. meduseum</i>	142595/12.83%	34.87	37.34	7.78×10^{-4}	30	4	77	111	15	NC476730.1
<i>T. pratense</i>	121178/NA *	34.63	36.94	7.43×10^{-4}	28	4	58	90	11	KJ788290
<i>T. repens</i>	132120/20.70%	34.53	36.96	8.10×10^{-4}	31	4	72	107	16	KC894706.1
<i>T. strictum</i>	125834/0.71%	34.98	36.70	8.82×10^{-4}	31	5	75	111	18	NC025745.1
<i>T. aureum</i>	126970/5.60%	34.86	36.81	8.51×10^{-4}	30	4	74	108	15	KC894708.1
<i>T. boissieri</i>	125740/1.05%	35.24	36.83	8.75×10^{-4}	31	5	74	110	17	NC025743.1
<i>T. glanduliferum</i>	126149/0.78%	34.90	36.70	8.72×10^{-4}	30	5	75	110	17	NC025744.1
<i>T. grandiflorum</i>	126149/5.60%	37.20	35.82	8.80×10^{-4}	30	4	77	111	16	NC_024034
<i>T. hybridum</i>	134831/7.86%	34.33	35.97	8.08×10^{-4}	31	4	74	109	17	KJ788286
<i>T. lupinaster</i>	135049/5.98%	33.97	35.71	8.15×10^{-4}	30	4	76	110	15	KJ788287
<i>T. occidentale</i>	133780/4.64%	36.34	35.91	8.00×10^{-4}	29	4	74	107	17	KJ788289
<i>T. semipilosum</i>	138194/10.55%	36.31	35.81	7.89×10^{-4}	31	4	74	109	17	KJ788291

* The cp genome annotation of *Trifolium pratense* is incomplete, so the percentage of repetitive cannot be calculated [13].

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

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

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

2.2. Repeat Sequences Analysis

Scattered repetitive sequences (palindrome repeats and direct repeats) and simple sequence repeats (SSRs) were analyzed respectively. A total of 1941 scattered repetitive sequences in the *T. alexandrinum* cp genome were annotated, which was greater than *T. resupinatum* (1250). The percentages of palindrome repeats (type P, 50.49%, Figure 2B) of *T. alexandrinum* were slightly larger than *T. resupinatum* (46.4%). A total of 370 (Figure 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*). 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 polyadenine repeats (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; Figure 2A). The common repeat sequences larger than 30 bp with the longest length of 117 bp was showed in Figure 2C.

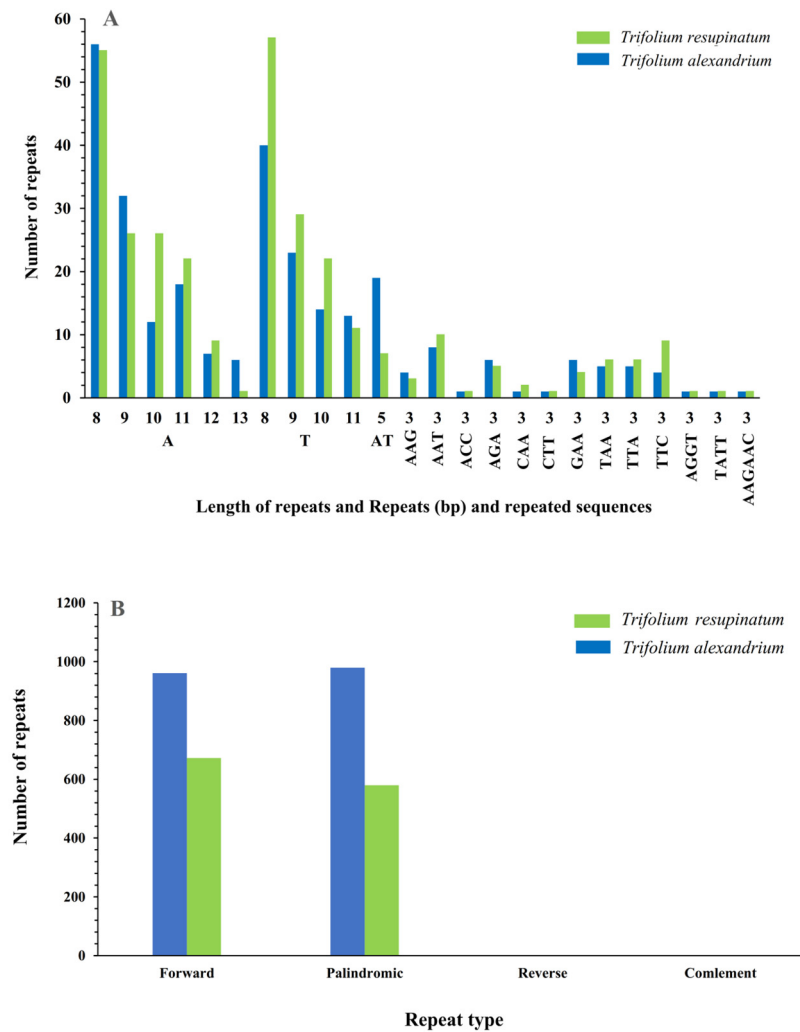


Figure 2. Cont.

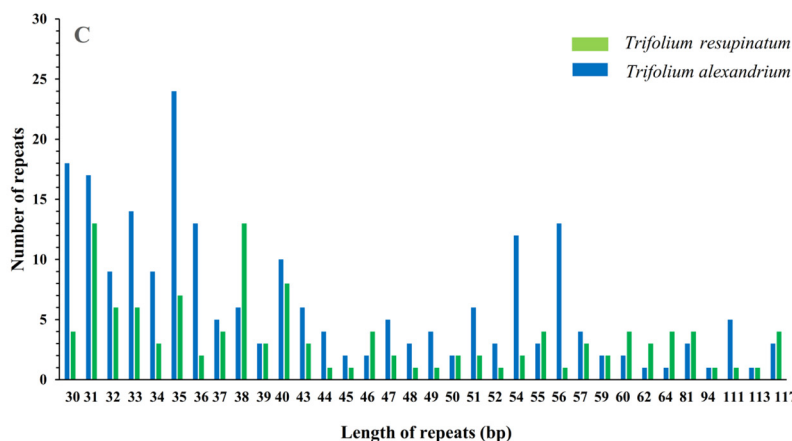


Figure 2. Repeating events of shared genes of *T. alexandrinum* and *T. resupinatum*. (A) Shared length of repeats and repeated sequences of *T. alexandrinum* and *T. resupinatum*; (B) Repeat type predicted in *T. alexandrinum* and *T. resupinatum* and (C) Listing of shared repetitive sequences with more than 30 bp.

2.3. Relative Synonymous Codon Usage Analysis

Relative synonymous codon usage analysis (RSCU), which is considered to be a combination result of natural selection, species mutation and genetic drift, was analyzed (Additional file1: Figure S1; Additional file3: Table S3). The RSCU value for initiation codon AUG was 2.9745 in *T. alexandrinum* and 2.9721 in *T. resupinatum*. The values of three termination codons UAA, UAG and UGA were 1.6215, 0.5676 and 0.8109 in *T. alexandrinum*, and 1.5909, 0.5454 and 0.8637 in *T. resupinatum*. The codons with an RSCU value greater than one were considered to be a greater codon frequency. 46.97% (31 of 66, include three termination codons) of the codons were with the greater codon frequency both in *T. alexandrinum* and *T. resupinatum*, in which 93.55% (29 of 31) prefers A or U in the third sites. In the other codons with RSCU values less than one (including one), C or G were more common in the third position (88.57%, 31 of 35).

2.4. 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. SNPs and In/Dels in every gene varied from 1 (*ndhE* and *psaC*) to 677 (*atpB*) with a 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%) (Figure 3; Additional file4: Table S4). The 66 shared protein-coding genes with variations were used to calculate the Ka and Ks (Additional file5: Table S5). The values of Ka and Ks ranged from 0 (*ndhE*, *petD*, *psaI*, *psbA*, *psbB* and so on) to 3.0151 (*rps4*) and 0 (*petG*, *petN*, *ndhD*, *psaJ*, *pabK*, *rpl23* and *rpl36*) to 2.9415 (*rps8*). Except for seven genes with Ks = 0, the 59 shared genes were used to calculate Ka/Ks, which varied from 0 (*ndhE*, *psbZ*, *psbA*, *psbJ* and so on) to 3.7723 (*rps4*, Figure 4), respectively. Seven genes including *rps4*, *rpoC2*, *ndhG*, *ccsA*, *ndhF*, *rpoA* and *psaC* have Ka/Ks values above one, implying positive selection on these genes. The Pi values calculated by 96 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, among which nineteen were tRNA. What's more, the nine genes with Ka/Ks above one also possessed relatively high Pi values (Figure 5; Additional file6: Table S6).

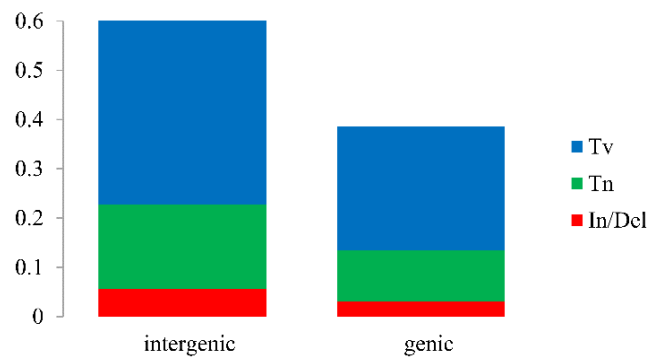


Figure 3. Transversion (Tv), transition (Tn) and Insert/Deletion (In/Del) were showed in intergenic and genic regions.

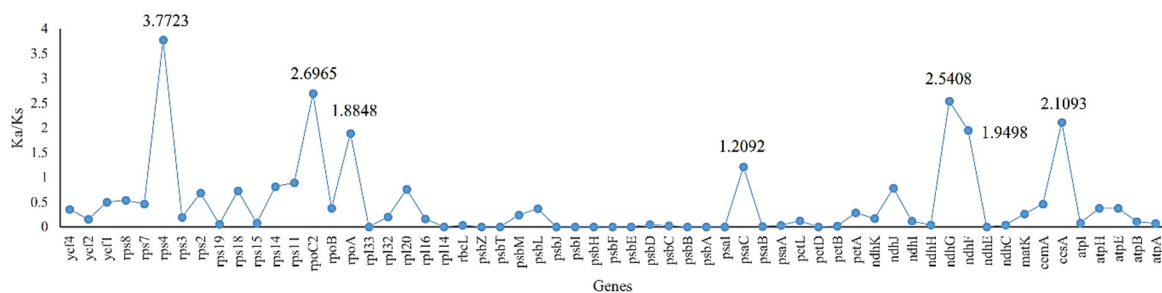


Figure 4. The synonymous/substitution rates (Ka/Ks) calculated using 59 shared genes in *T. alexandrinum* and *T. resupinatum*.

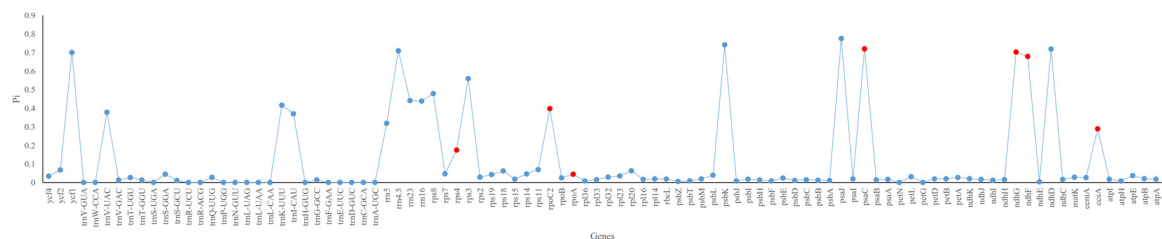


Figure 5. The nucleotide diversity (Pi) computed using 96 common genes of *T. alexandrinum* and *T. resupinatum*. Genes with Ka/Ks values more than one were red coded.

2.5. Whole cp Genome Comparison with Other *Trifolium* Species

In order to examine the sequence divergence of *Trifolium* genus and further shed light on the evolutionary events, such as gene mutation, rearrangement and gene loss, cp genomes of fifteen IR lacking species (*T. grandiflorum*, *T. hybridum*, *T. lupinaster*, *T. occidentale*, *T. semipilosum*, *T. aureum*, *T. boissieri*, *T. glanduliferum*, *T. strictum*, *T. repens*, *T. pratense*, *T. subterraneum*, *T. meduseum*, *T. alexandrinum* and *T. resupinatum*) were compared. The results showed that the size of cp genomes of these IR lacking species ranged from 121,178 bp (*T. pratense*) to 149,026 bp (*T. resupinatum*), with an average of 134,062 bp (Table 1). The GC content of those fifteen species changed from 33.80% to 37.20% in whole cp genome with the mean value of 34.99%, and 35.71% to 37.34% in CDS with the mean of 36.55%. 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 at the cp genome level were detected among fifteen *Trifolium* species using MAUVE program and the *T. resupinatum* as the reference sequence (Figure 6).

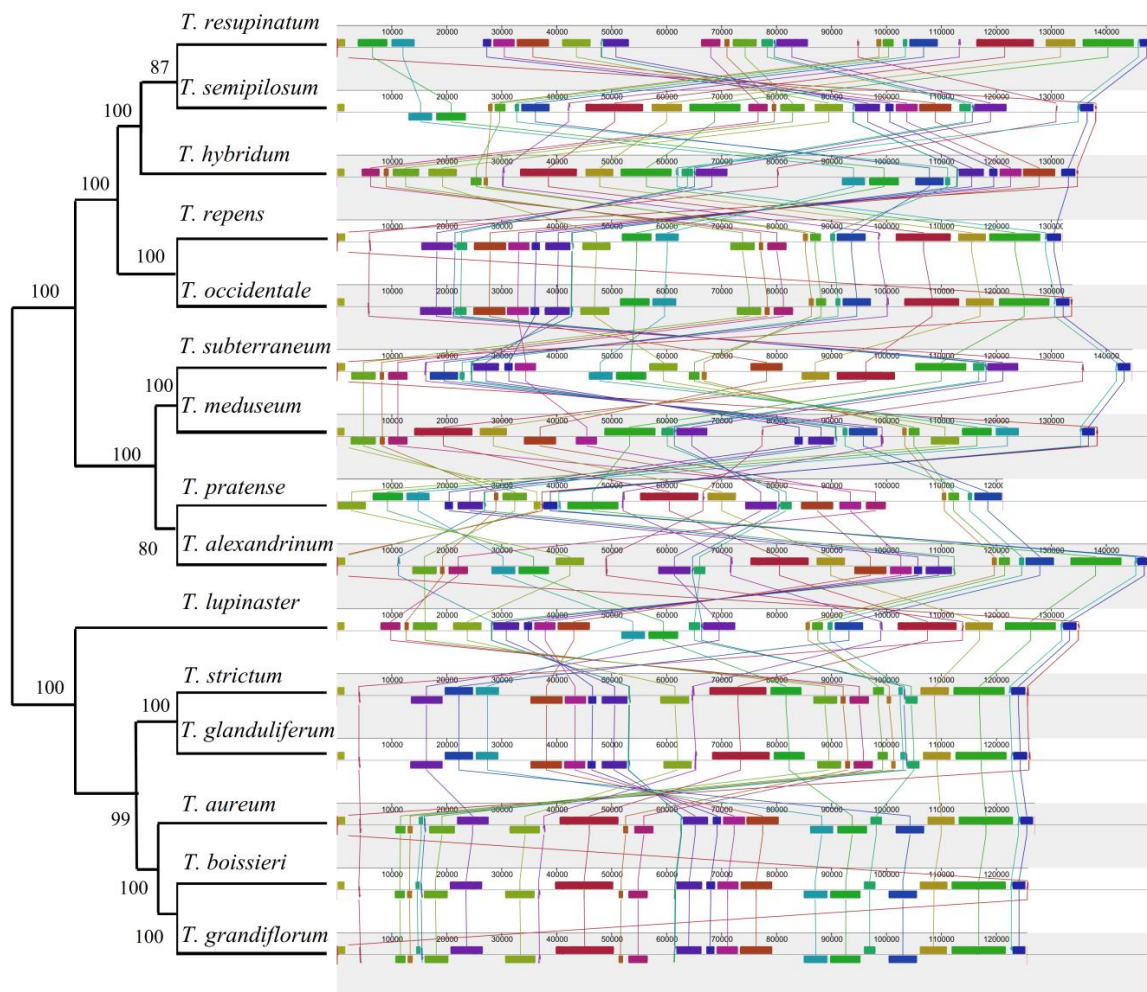


Figure 6. Synteny comparison of fifteen *Trifolium* chloroplast genomes with the reference of *T. resupinatum* using Mauve. Rectangular blocks with the same color indicate collinear regions.

2.6. Phylogenetic Divergence Time Estimation

The 41 protein-coding genes shared in cp genomes of the 25 species (23 of Papilionoideae, one of Caesalpinoideae and one of Mimosaceae) were subjected to phylogeny analysis and divergence times estimation (Figure 7A). The topological structure of phylogenetic tree was almost consistent with the classification of Leguminosae with strong bootstrap support. Three subfamilies of Leguminosae, Papilionoideae, Caesalpinoideae and Mimosaceae were clearly separated. Furthermore, two genes *ycf1* and *rpoC2*, which both contain vast repetitive sequences and high Pi values (0.7008, 0.3982) between *T. alexandrinum* and *T. resupinatum*, were also used to construct the phylogenetic trees. The result showed that the topological structure of the phylogenetic relationship of *Trifolium* species based on *ycf1* (Figure 7B) and *rpoC2* (Figure 7C) was almost in accordance with the phylogenetic tree constructed using 41 shared genes. Each Section of *Trifolium* was clearly divided based on *ycf1* and *rpoC2* genes. The *Trifolium* species of “refractory clade” (including Section Lupinaster, Trifolium, Tricocephalum, Vesicastrum and Trifoliasstrum) were grouped together in Figure 7C. In Figure 7B, however, Section Trifolium, Tricocephalum, Vesicastrum and Trifoliasstrum were grouped into one clade, and another clade consisted of Section Lupinaster, Paramesus and Subg. Chronosemium. *Trifolium* species split from *Medicago* species during the Early Cretaceous (116.1575 Mya) and the divergence time of those fifteen IR lacking *Trifolium* species ranged from 84.8505 Mya to 4.7720 Mya (Figure 7).

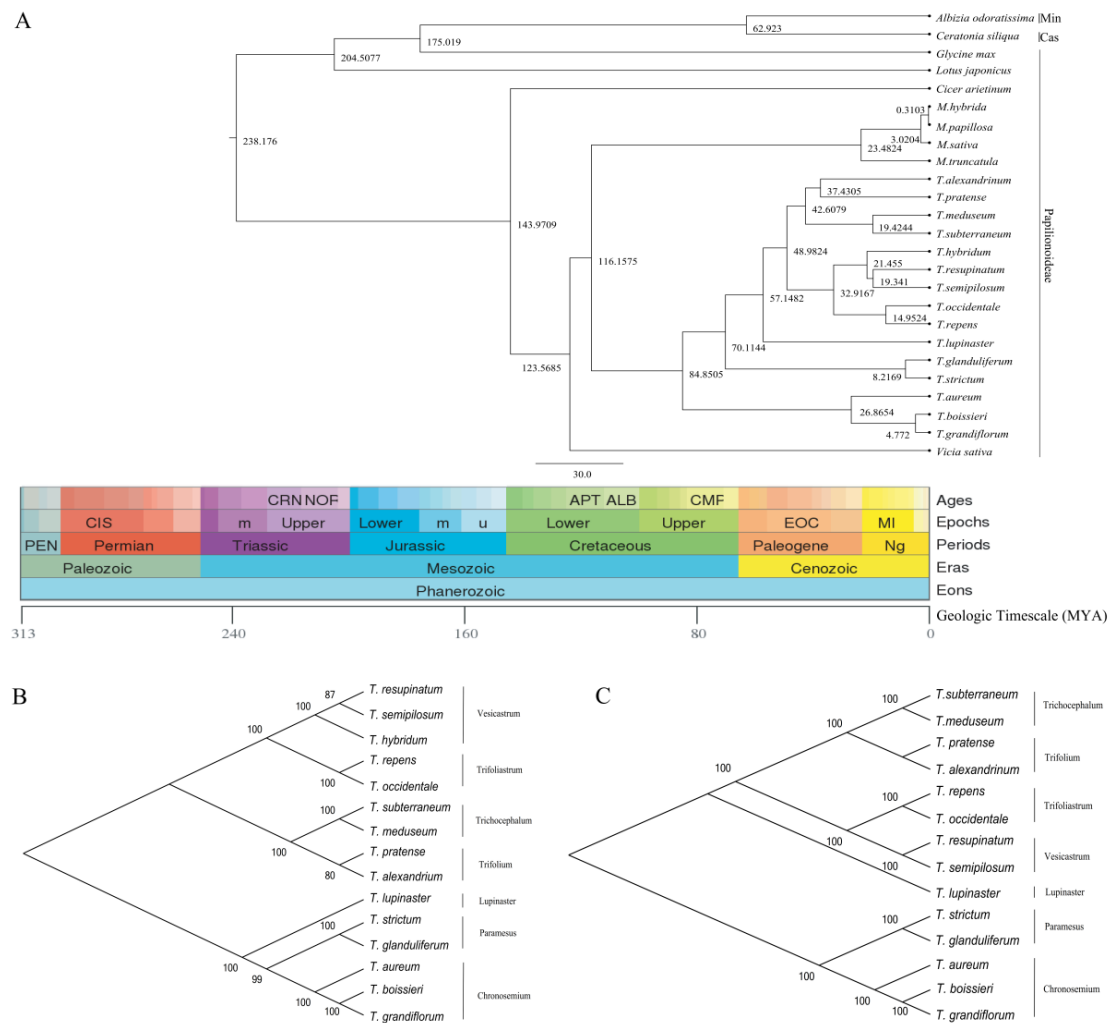


Figure 7. (A) BEAST chronogram of the 25 Leguminosae species based on the 41 common protein-coding genes and the phylogenetic trees constructed using *rpoC2* gene (B) and *ycf1* gene (C, *ycf1* gene was absent in *T. hybridum*). Geologic timescale was obtained from TIMETREE, time is shown in millions of years (MYA). Min, Mimosaceae; Cas, Caesalpinioideae. The over 50% bootstrap support values were showed in each main branch.

3. Discussion

3.1. Genome Feature of *T. alexandrinum* and *T. resupinatum* and Comparison with Other IR Lacking *Trifolium* Species

In this study, we sequenced and annotated cp genomes of two IR lacking species *T. alexandrinum* and *T. resupinatum*, identified the repeats and hotspot genes within the cp genomes, and constructed the phylogenetic tree along with other cp genomes. Our results added information about cp genomes of *Trifolium* species and provided new insights into the evolution study of IR lacking species.

The two cp genomes sequenced in this study revealed 75 and 66 protein-coding genes in *T. alexandrinum* and *T. resupinatum*, respectively (Table 1). There were nine unique genes in the *T. alexandrinum* cp genome which were absent in *T. resupinatum*, which might have been transferred from the cp genome to the nucleus genome of *T. resupinatum* in the evolutionary process of the species. A similar result was revealed in tobacco with an absent *accD* gene, which was essential for leaf development [15,16]. The *rpl2* gene, which was absent in *T. resupinatum*, has been completely or partially transferred to the nucleus genome of some legumes like soybean and *Medicago* [17]. Besides, the loss of two *ndh* genes (*ndhA* and *ndhB*) was usually related to the nutritional status and rearrangement in

most angiosperm species [18]. Although gene loss of *atpF* [19], *psbN* [19], *rpoC1* [19] and *ycf3* [20] was also found in other species, however, the detailed reason for that gene loss remained to be explained.

Compared to other IR containing cp genomes of angiosperms, the number of protein-coding genes of *Trifolium* species is less conserved [21,22]. It might be related to the fact that IR lacking will lead to an extensively arranged cp genome thus causing diverse genes loss [11]. According to Millen et al. [23], the vast majority of cp genomes of angiosperms held in shared 74 coding-protein genes but other five genes (*accD*, *ycf1*, *ycf2*, *rpl23* and *infA*) only existed in some specific species. *infA* gene, which codes for translation initiation factor 1 (IF1), was defunct in all the listed fifteen *Trifolium* species. Considered as the most transferable gene in cp genome, *infA* was in existence in about 24 angiosperm lineages including *Trifolium* species, and related *Medicago* species [24,25]. It is worth noting that there are two *rrn16* gene in each of *Trifolium* species sequenced in this study, and this phenomenon was also reported in cp genomes of *T. strictum* [13] and *T. glanduliferum* [13]. Furthermore, we also found that one of the *rrn16* and *rrn23* was partially overlapped. Indeed, no similar phenomenon was found in other *Trifolium* species. Previous studies have shown that *Cyanobacteria*, *Chlamydomonas reinhardtii*, *Cyanophora paradoxa*, *Zea mays*, *Oryza sativa* and *Arabidopsis thaliana* can start and stop transcription anywhere in the cp genome [26]. Chloroplasts are known to originate from ancient *Cyanobacteria*, this transcription property may be conserved in some species like *T. alexandrinum* and *T. resupinatum*, so that the rRNA genes of those two species may gain potential transcription ability. On the other hand, the cp genomes of most plant species are small and in order to avoid costs, genes overlapping happened. The *Ycf1* (hypothetical chloroplast reading frame no. 1) gene, generally with the premature stop codons in the CDS thus be defined as pseudogene in other angiosperm [27], has undergone processes of accelerated mutation rate, decreased GC content, and decreased secondary structure stability [28]. In this study, the *ycf1* gene detected in all the fourteen *Trifolium* species (except *T. hybridum*) contains a normal stop codon and is able to code protein. Two genes *ycf4* (hypothetical chloroplast reading frame no. 4) and *rps16* (ribosomal protein S16), which were found in most cp genomes of angiosperm and some relic plant like *Amborella trichopoda* [29], are not present in the two *Trifolium* species cp genomes sequenced in this study, and *rps16* is not present in all the fifteen *Trifolium* species. The *rps8* (ribosomal protein S8) gene was found without stop codons in *Medicago truncatula* and *M. sativa* [25], while it possesses a stop codon in the fifteen *Trifolium* cp genomes. In the gene *ndhB*, no internal stop codon was detected in the fourteen (except *T. resupinatum*) *Trifolium* cp genomes, which is inconsistent with many legume cp genomes whose *ndhB* gene contains an internal stop codon [25].

3.2. Relative Synonymous Codon Usage Analysis (RSCU)

The unequal using frequencies of synonymous codons detected in most sequenced genomes was termed synonymous codons usage bias [30], and is now considered crucial in shaping gene expression and cellular function [31]. RSCU indicates the relative probability of a particular codon encoding the synonymous codon of the corresponding amino acid. In this study, 93.5% of the codons were found prefer A/U in third position, similar to that of *M. sativa* [25] and *Gossypium* [32]. This phenomenon could be attributed to the fact that dicotyledon prefers to the A/U-end codons and manifests a potential force in molecular evolution of *Trifolium* species: mutation and natural selection.

3.3. SSRs and Large Repeat Sequences

Given a matrilineal inheritance feature, rich number of tags and low frequency of genetic recombination, cpSSRs (chloroplast simple sequence repeats) are considered an efficient molecular marker in genetic variety analyzing, population structure studying, species identification and phylogeny analysis [33]. The SSRs identified in *T. alexandrinum* and *T. resupinatum* were poly(A)/(T), which was consistent with the majority of plant family [34–36]. Although there have many studies reported on the application of cpSSRs in plant genetic diversity analysis, the important potential of cpSSRs in studying the ecological and evolutionary processes of wild materials of plants and their related species still need to be recognized [37]. Therefore, the cpSSRs of the two *Trifolium* species detected in this study can be

used to evaluate genetic relationships among different species and to detect polymorphism of *Trifolium* species and their relatives at the population level.

Repetitive sequences related to the continuously self-replicating of genetic material in the process of evolution, thus indicating the greatly expanded and enriched genetic information [38]. The present study revealed a relatively high repetitive percentage (7.325%, Table 1) in the cp genomes of fifteen IR lacking species, which was higher than IR lacking *Tydemania expeditionis* (0.4%) and *Bryopsis plumose* (2.4%) [7]. The number of repeated sequences in the cp genome are associated with rearrangement in some species [14]. However, the driving force of the repetitive sequence was predominantly related to nuclear genes and genomic recombination [12], 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 [39], *ycf1* and *rpoC2* possessed high values of Pi (Figure 5) and the majority of repetitive sequences in *T. alexandrinum* and *T. resupinatum*. Therefore, these two genes could have suffered from selection pressure and could be used for phylogenetic analysis and population genetic study of *Trifolium* species.

3.4. Sequence Divergence and Hotspots

Point mutation was generally more common than frame shift for natural mutation [40]. 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 hypothesis that CDS had a slower rate of evolution compared with CNS [41]. Furthermore, minor SNPs (159 between *Oryza sativa* and *O. nivara* [42], 330 between *Citrus sinensis* and *C. aurantiifolia* [43] and 231 between *Machilus yunnanensis* and *M. balansae* [44]) 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 the present study. As an important structure in stabilizing cp genome, the IR region can prevent the genome mutating by selective force [45]. Thus the observed abundant SNPs/Indels in *T. alexandrinum* and *T. resupinatum* are not surprising.

The comparison between the K_a and K_s of genes is an important measure of molecular evolution [46]. Most genes were subjected to neutral selection and purification selection; however, there are also limited genes whose rate of K_a is higher than that of K_s because the function of the gene has been dramatically changed, called Darwinian positive selection [47]. 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 [48]. Here, seven protein-coding genes in the cp genomes of *T. alexandrinum* and *T. resupinatum* (*rps4*, *rpoC2*, *ndhG*, *ccsA*, *ndhF*, *ycf1* and *psaC*) have a high ratio of K_a to K_s , which is led by high values of K_a but extremely low values of K_s , and could imply that they are under positive selections. *rps4* [49] and *rpoC2* [50] 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 [51].

In general, there is a strong correlation between the presence of IR and structural stabilization of cp genomes. Substantial rearrangement was usually found in cp genomes lacking IR [13]. Among those IR lacking species of Leguminosae such as alfalfa, subclover, pea, and so forth, some are structurally stable and have not been rearranged, some undergo intermediate rearrangements, while others experienced a series of complex rearrangements [13]. This study found abundant rearrangements within fifteen cp genomes of IR lacking species of *Trifolium* (Figure 6). 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, this could be why there are many rearrangement events detected among those fifteen *Trifolium* species [8]. However, lacking IR leading to increased rearrangement is only one of the explanations. The repeats, acting as a locus of recombination within

homologous genes, along with transposable elements (TEs), have been suggested as a reasonable mechanism for highly-rearranged cp genomes of *Trifolium* species [8].

3.5. Phylogeny Analysis and Divergence Time

The topological structure of other thirteen *Trifolium* species using 41 protein coding genes in this study was generally in agreement with the report of Sveinsson and Cronk [13] by 58 protein coding genes. In addition, the phylogenetic location of tested *T. alexandrinum* and *T. resupinatum* was confirmed (Figure 7). Furthermore, *T. alexandrinum* and *T. pratense*, both belonging to *Trifolium* Sect. *Trifolium*, were clustered together though *T. alexandrinum* and *T. subterraneum* were grouped together in Malaviya's study based on isozyme data [52]. Three IR lacking species *T. boissieri*, *T. grandiflorum* and *T. aureum* were predicted to differentiate with other twelve species in the late Cretaceous period, then another two IR lacking species *T. strictum* and *T. glanduliferum* were further diverted at about 8 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 an evolutionary pattern is underway to illuminate.

4. Methods

4.1. Plant Material, DNA Isolation and Sequencing

Plant seeds of *T. alexandrinum* (cv 'Elite II') and *T. resupinatum* (cv 'Laser') were kindly provided by Barenbrug (Queensland, Australia) then germinated in a growth chamber (25 °C, 300 $\mu\text{mol}\cdot\text{m}^{-2}\cdot\text{s}^{-1}$; 16-h photoperiod). Total DNA was extracted from 50 mg of fresh leaves following the Plant DNA Isolation Kit (Tiangen, Beijing, China). Sheared low molecular weight DNA fragments were used to construct paired-end (PE) libraries according to protocol of Illumina manual (San Diego, CA, USA). Completed libraries were pooled and sequenced in the Illumina NovaSeq platform with PE150 sequencing strategy and 350 bp insert size.

4.2. cp Genome Assembly, Annotation and Visualization

The raw read data for two *Trifolium* species were filtered according to the following criteria—reads of less than 5% unidentified nucleotides and more than 50% of their bases with a quality score of >20 were retained. With the reference genome of *T. meduseum* [12] (National Center for Biotechnology Information, NCBI number KJ 788288), the cp DNA were assembled as follows. In order to decrease the difficulty of sequences assembly, filtered reads (clean data) were aligned to the cp genome database built by Genepioneer Biotechnologies (Nanjing, China) using Bowtie2 v 2.2.4 [53] and SPAdes v3.10.1 [54] to acquire SEED sequences then obtained contigs by kmer iterative extend seed. Contigs, whose E-values were less than 1×10^{-5} , Identities values were close to 100% and Gaps were close to 0 by a BLAST research in NCBI with the data set including *T. meduseum* [12], *T. pratense* [13], *T. repens* [12] and *T. subterraneum* [20], were obtained. Then the contigs were connected as scaffolds using SSPACE v 2.0 [55] followed by gap filling using Gapfiller v 2.1.1 [56] until the complete chloroplast genome sequence was recovered.

The results of CDS and rRNA were obtained using BLAST V 2.2.25 and HMMER V3.1 b2 and the cp genome database of NCBI. ARAGORN V 1.2.38 [57] and tRNAscan-SE search server (<http://lowelab.ucsc.edu/tRNAscan-SE/>, [58]) were used to predict and further check tRNA. Finally, the consensus annotation was obtained via Geneious (<https://www.geneious.com>, [59]) and visualized in OGDRAW (<https://chlorobox.mpimp-golm.mpg.de/OGDraw.html>, [60]).

4.3. The Relative Synonymous Codon Usage Analysis (RSCU) and Simple Sequence Repeats (SSRs) Prediction

The RSCU was analyzed using MEGA v7.0 to reflect the relative preference of a particular base encoding the corresponding amino acid codon [61]. Values of RSCU over one were considered to be a 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.

4.4. Sequence Variation Analysis and Ka/Ks

Whole cp genome alignment and collinearity analysis of sequenced species herein along with further thirteen IR-lacking *Trifolium* species, namely *T. grandiflorum* (NC_024034, [13]), *T. hybridum* (KJ788286, [13]), *T. lupinaster* (KJ788287, [13]), *T. occidentale* (KJ788289, [13]), *T. semipilosum* (KJ788291, [13]), *T. strictum* (NC025745.1, [13]), *T. aureum* (KC894708.1, [12]), *T. boissieri* (NC025743.1, [13]), *T. glanduliferum* (NC025744.1, [13]), *T. subterraneum* (NC011828, [20]), *T. meduseum* (NC476730.1, [13]), *T. pratense* (KJ788290, [13]) and *T. repens* (KC894706.1, [12]) was implemented using Mauve [62]. Furthermore, the common genes and shared protein-coding genes of *T. alexandrinum* and *T. resupinatum* tested in the present study were utilized for nucleotide diversity (Pi) and Ka/Ks calculation. Ka/Ks, which was generally considered to be a reflection of selection pressures, was computed via KaKs_Calculator v2.0 [63]. 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 [64] after sequences alignment of the common genes by MAFFT version 7.017 [65]. Finally, single nucleotide polymorphisms (SNPs) and insertions/deletions (In/Dels) of *T. alexandrinum* and *T. resupinatum* were also identified using Mafft program [65].

4.5. Divergence Time Estimates

The 41 common genes sequence of fifteen *Trifolium* species and another ten Leguminosae species, including *Lotus japonicus* (AP002983.1), *Glycine max* (NC_007942.1), *Cicer arietinum* (NC_011163.1), *Ceratonia siliqua* (NC_026678.1), *Albizia odoratissima* (NC_034987.1), *Medicago truncatula* (KF241982.1), *M. sativa* (KU321683.1), *M. papillosa* (NC_027154.1), *M. hybrida* (NC_027153.1) and *Vicia sativa* (NC_027155.1) were first blasted using MEGA [61] then the alignment file was utilized to assess the divergence time using BEAST v 1.7.3 package [66] with the Bayesian method. GTR + G + I substitution model with a 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, 1000 of Tracelog, 1000 of screenlog, 1000 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 tree file obtained from TreeAnnotator was visualized in Figtree v1.4.3 (<http://tree.bio.ed.ac.uk/software/figtree/>). Furthermore, phylogenetic trees of the fifteen *Trifolium* species based on *ycf1* gene and *rpoC2* gene were constructed under Maximum Composite Likelihood method with 1000 bootstrap replications using MEGA v7.0 [61].

5. Conclusions

cp genomes of *T. alexandrinum* and *T. resupinatum*, which belong to inverted-repeat-lacking clade (IRLC), were sequenced and annotated in present study and were compared with the cp genomes of other thirteen IR lacking *Trifolium* species reported previously. The results revealed abundant SNP and In/Del in *T. alexandrinum* and *T. resupinatum* cp genomes and high variation in CDS and abundant rearrangement within *Trifolium* genus. This valuable information will provide insight into the evolution of IR lacking species.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2223-7747/9/4/478/s1>, Figure S1. The relative synonymous codon usage (RSCU) for the *T. alexandrinum* and *T. resupinatum*. Figure S2. The assembly sequence genomic coverage maps of *T. alexandrinum* and *T. resupinatum*. Table S1. Location and length of intron-containing genes in the chloroplast genomes of *T. alexandrinum* and *T. resupinatum*. 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. Table S3. The relative synonymous codon usage (RSCU) analyzed using CodonW. Table S4. Transversion (Tv) and transition (Tn) were detected between *T. alexandrinum* and *T. resupinatum*. Table S5. The synonymous/synonymous substitution rates (Ka/Ks) calculated using 59 shared genes in *T. alexandrinum* and *T. resupinatum*. Table S6. The nucleotide diversity (Pi) computed using 96 common genes of *T. alexandrinum* and *T. resupinatum*.

Author Contributions: Conceptualization, X.L. and Y.P.; Data curation, Y.X. (Yanli Xiong) and Y.X. (Yi Xiong); Formal analysis, Y.X. (Yanli Xiong), Y.X. (Yi Xiong) and J.Y.; Funding acquisition, X.M.; Investigation, J.H.; Methodology, J.H., J.Y. and Y.P.; Project administration, X.L.; Resources, Q.Y., Z.D. and X.Z.; Supervision, Q.Y.; Validation, J.Z.; Visualization, J.Z. and Z.D.; Writing—original draft, Y.X. (Yanli Xiong) and Y.X. (Yi Xiong); Writing—review & editing, X.Z. and X.M. All authors have read and agreed to the published version of the manuscript.

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).

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

Conflicts of Interest: The authors declare that they have no competing interests.

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.

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	nucleotide diversity;
Ka	non-synonymous;
Ks	synonymous;
SSRs	simple sequence repeats;
RSCU	the relative synonymous codon usage analysis;
Tv	transversion;
Tn	transition;
CDS	coding sequences;
CNS	non-coding sequences;
UTR	untranslated regions

References

1. Sabudak, T.; Guler, N. *Trifolium* L.—A review on its phytochemical and pharmacological profile. *Phytother. Res.* **2009**, *23*, 439–446. [[CrossRef](#)] [[PubMed](#)]
2. Ellison, N.W.; Liston, A.; Steiner, J.J.; Williams, W.M.; Taylor, N.L. Molecular phylogenetics of the clover genus (*Trifolium*–Leguminosae). *Mol. Phylogenet. Evol.* **2006**, *39*, 688–705. [[CrossRef](#)] [[PubMed](#)]
3. Badr, A.; El-Shazly, H.H.; Watson, L.E. Origin and ancestry of egyptian clover (*Trifolium alexandrinum* L.) as revealed by AFLP markers. *Genet. Resour. Crop Evol.* **2008**, *55*, 21–31. [[CrossRef](#)]
4. Nazir, M.; Shah, F.H. Studies on persian clover (*Trifolium resupinatum*). *Plant Food Hum. Nutr.* **1985**, *35*, 57–62. [[CrossRef](#)]

5. Douglas, S.E. Chloroplast origins and evolution. In *The Molecular Biology of Cyanobacteria*; Springer: Dordrecht, the Netherlands, 1994; pp. 91–118.
6. Daniell, H.; Lin, C.S.; Yu, M.; Chang, W. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Biol.* **2016**, *17*, 130–134. [[CrossRef](#)] [[PubMed](#)]
7. Turmel, M.; Otis, C.; Lemieux, C. Divergent copies of the large inverted repeat in the chloroplast genomes of ulvophyceean green algae. *Sci. Rep.* **2017**, *7*, 1–14. [[CrossRef](#)]
8. Palmer, J.D.; Osorio, B.; Aldrich, J.; Thompson, W.F. Chloroplast DNA evolution among legumes: Loss of a large inverted repeat occurred prior to other sequence rearrangements. *Curr. Genet.* **1987**, *11*, 275–286. [[CrossRef](#)]
9. Turmel, M.; Otis, C.; Lemieux, C. Mitochondrion-to-chloroplast DNA transfers and intragenomic proliferation of chloroplast group II introns in *Gloeotilopsis* green algae (Ulotrichales, Ulvophyceae). *Genome Biol. Evol.* **2016**, *8*, 2789–2805. [[CrossRef](#)]
10. Wu, C.S.; Wang, Y.N.; Hsu, C.Y.; Lin, C.P.; Chaw, S.M. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and Cupressophytes and influence of heterotachy on the evaluation of Gymnosperm phylogeny. *Genome Biol. Evol.* **2011**, *3*, 1284–1295. [[CrossRef](#)]
11. Kaila, T.; Chaduvla, P.K.; Rawal, H.C.; Saxena, S.; Tyagi, A.; Mithra, S.V.; Gaikwad, K. Chloroplast genome sequence of clusterbean (*Cyamopsis tetragonoloba* L.): Genome structure and comparative analysis. *Genes* **2017**, *8*, 212. [[CrossRef](#)]
12. Barrett, C.F.; Freudenstein, J.V.; Li, J.; Mayfield-Jones, D.R.; Perez, L.; Pires, J.C.; Santos, C. Investigating the path of plastid genome degradation in an early-transitional clade of heterotrophic orchids, and implications for heterotrophic angiosperms. *Mol. Biol. Evol.* **2014**, *31*, 3095–3112. [[CrossRef](#)]
13. Sveinsson, S.; Cronk, Q. Evolutionary origin of highly repetitive plastid genomes within the clover genus (*Trifolium*). *BMC Evol. Biol.* **2014**, *14*, 218–228. [[CrossRef](#)] [[PubMed](#)]
14. Haberle, R.C.; Fourcade, H.M.; Boore, J.L.; Jansen, R.K. Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. *J. Mol. Evol.* **2008**, *66*, 350–361. [[CrossRef](#)] [[PubMed](#)]
15. Rousseau-Gueutin, M.; Huang, X.; Higginson, E.; Ayliffe, M.; Day, A.; Timmis, J.N. Potential functional replacement of the plastidic acetyl-CoA carboxylase subunit (*accD*) gene by recent transfers to the nucleus in some angiosperm lineages. *Plant Physiol.* **2013**, *161*, 1918–1929. [[CrossRef](#)] [[PubMed](#)]
16. Kode, V.; Mudd, E.A.; Lamtham, S.; Day, A. The tobacco plastid *accD* gene is essential and is required for leaf development. *Plant J. Cell Mol. Biol.* **2005**, *44*, 237–244. [[CrossRef](#)] [[PubMed](#)]
17. Adams, K.L.; Chuan, O.H.; Palmer, J.D. Mitochondrial gene transfer in pieces: Fission of the ribosomal protein gene *rpl2* and partial or complete gene transfer to the nucleus. *Mol. Biol. Evol.* **2001**, *18*, 2289–2297. [[CrossRef](#)]
18. Kim, H.T.; Kim, J.S.; Moore, M.J.; Neubig, K.M.; Williams, N.H.; Whitten, W.M.; Kim, J.H. Seven new complete plastome sequences reveal rampant independent loss of the *ndh* gene family across orchids and associated instability of the inverted repeat/small single-copy region boundaries. *PLoS ONE* **2015**, *10*, e0142215. [[CrossRef](#)]
19. Xi, L.; Zhang, T.C.; Qiao, Q.; Ren, Z.M.; Zhao, J.; Yonezawa, T.; Hasegawa, M.; Crabbe, M.J.; Li, J.Q.; Zhong, Y. Complete chloroplast genome sequence of holoparasite *Cistanche deserticola* (Orobanchaceae) reveals gene loss and horizontal gene transfer from its host *Haloxylon ammodendron* (Chenopodiaceae). *PLoS ONE* **2013**, *8*, e58747.
20. Cai, Z.Q.; Guisinger, M.; Kim, H.G.; Ruck, E.; Blazier, J.C.; McMurtry, V.; Kuehl, J.V.; Boore, J.; Jansen, R.K. Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. *J. Mol. Evol.* **2008**, *67*, 696–704. [[CrossRef](#)]
21. Yang, Y.; Tao, Z.; Dong, D.; Jia, Y.; Li, F.; Zhao, G.F. Comparative analysis of the complete chloroplast Genomes of five *Quercus* species. *Front. Plant Sci.* **2016**, *7*, 944–959. [[CrossRef](#)] [[PubMed](#)]
22. Sajjad, A.; Muhammad, W.; Abdul, L.K.; Muhammad, A.K.; Sang-Mo, K.; Qari, M.I.; Raheem, S.; Saqib, B.; Byung-Wook, Y.; In-Jung, L. The complete chloroplast genome of wild rice (*Oryza minuta*) and its comparison to related species. *Front. Plant Sci.* **2017**, *8*, 289–304.
23. Millen, R.S.; Olmstead, R.G.; Adams, K.L. Many parallel losses of *infA* from chloroplast DNA during Angiosperm evolution with multiple independent transfers to the nucleus. *Cochrane Database Syst. Rev.* **2001**, *13*, 645–658.

24. Jansen, R.K.; Wojciechowski, M.F.; 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*, 1204–1217. [[CrossRef](#)]
25. Tao, X.L.; Ma, L.C.; Zhang, Z.S.; Liu, W.X.; Liu, Z.P. Characterization of the complete chloroplast genome of alfalfa (*Medicago sativa*) (Leguminosae). *Gene Rep.* **2017**, *6*, 67–73. [[CrossRef](#)]
26. Shi, C.; Wang, S.; Xia, E.H.; Jiang, J.J.; Zeng, F.C.; Gao, L.Z. Full transcription of the chloroplast genome in photosynthetic eukaryotes. *Sci. Rep.* **2016**, *6*, 30135. [[CrossRef](#)]
27. Pasquale, L.C.; Domenico, D.P.; Donatella, D.; Giovanni, G.V.; Gabriella, S. Complete chloroplast genome of the multifunctional crop globe artichoke and comparison with other Asteraceae. *PLoS ONE* **2015**, *10*, e120589.
28. Xiao, L. Intra-genomic polymorphism in the internal transcribed spacer (ITS) regions of *Cycas revoluta*: Evidence of incomplete concerted evolution. *Biodivers. Sci.* **2009**, *17*, 476–481.
29. Goremykin, V.V.; Hirsch-Ernst, K.I.; Wölfel, S.; Hellwig, F.H. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal Angiosperm. *Mol. Biol. Evol.* **2003**, *9*, 1499–1505. [[CrossRef](#)]
30. Wright, F. The 'effective number of codons' used in a gene. *Gene* **1990**, *87*, 23–29. [[CrossRef](#)]
31. Li, Y.F.; Sylvester, S.P.; Li, M.; Zhang, C.; Li, X.; Duan, Y.F.; Wang, X.R. The complete plastid genome of *agnolia zenii* and genetic comparison to Magnoliaceae species. *Molecules* **2019**, *24*, 261. [[CrossRef](#)]
32. Talat, F.; Wang, K. Comparative bioinformatics analysis of the chloroplast genomes of a wild diploid *Gossypium* and two cultivated allotetraploid species. *Iran. J. Biotechnol.* **2015**, *13*, 47–56. [[CrossRef](#)] [[PubMed](#)]
33. Wheeler, G.L.; Dorman, H.E.; Buchanan, A.; Challagundla, L.; Wallace, L.E. A review of the prevalence, utility, and caveats of using chloroplast simple sequence repeats for studies of plant biology. *Appl. Plant Sci.* **2014**, *2*, 1400059. [[CrossRef](#)] [[PubMed](#)]
34. Saski, C.; Lee, S.; Daniell, H.; Wood, T.C.; Tomkins, J.; Kim, H.; Jansen, R.K. Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. *Plant Mol. Biol.* **2005**, *59*, 309–322. [[CrossRef](#)] [[PubMed](#)]
35. Morgante, M.; Pfeiffer, A.; Costacurta, A.; Olivieri, A.M.; Rafalski, J.A. *Polymorphic Simple Sequence Repeats in Nuclear and Chloroplast Genomes: Applications to the Population Genetics of Trees*; Springer: Dordrecht, the Netherlands, 1996.
36. Hu, Y.; Woeste, K.E.; Zhao, P. Completion of the chloroplast genomes of five Chinese *Juglans* and their contribution to chloroplast phylogeny. *Front. Plant Sci.* **2017**, *7*, 1939–1955. [[CrossRef](#)]
37. Ebert, D.; Peakall, R.O. Chloroplast simple sequence repeats (cpSSRs): Technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. *Mol. Ecol. Resour.* **2009**, *9*, 673–690. [[CrossRef](#)]
38. Tachida, H. Evolution of repeated sequences in non-coding regions of the genome. *Jpn. J. Genet.* **1993**, *68*, 549–565. [[CrossRef](#)]
39. Wei, Y.L.; Wen, Z.F.; Liu, F.; Zhang, J.W.; Huang, W.G.; Lan, Y.P.; Cheng, L.L.; Cao, Q.C.; Hu, G.G.; Yun, C. Bioinformatics analysis of *ycf1* gene in *Corylus*. *J. Shanxi Agric. Sci.* **2018**, *46*, 1244–1247.
40. Raes, J.; Van, D.P.Y. Functional divergence of proteins through frameshift mutations. *Trends Genet.* **2005**, *21*, 428–431. [[CrossRef](#)]
41. Bobrova, V.K.; Troitsky, A.V.; Ponomarev, A.G.; Antonov, A.S. Low-molecular-weight rRNAs sequences and plant phylogeny reconstruction: Nucleotide sequences of chloroplast 4.5S rRNAs from *Acorus calamus* (Araceae) and *Ligularia calthifolia* (Asteraceae). *Plant Syst. Evol.* **1987**, *156*, 13–27. [[CrossRef](#)]
42. Masood, M.S.; Nishikawa, T.; Fukuoka, S.; Njenga, P.K.; 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*, 133–139. [[CrossRef](#)]
43. Su, H.J.; Hogenhout, S.A.; Al-Sadi, A.M.; Kuo, C.H. Complete chloroplast genome sequence of Omani lime (*Citrus aurantiifolia*) and comparative analysis within the Rosids. *PLoS ONE* **2014**, *11*, e113049. [[CrossRef](#)]
44. Song, Y.; Dong, W.P.; Liu, B.; Xu, C.; Yao, X.; Gao, J.; Richard, T.C. Comparative analysis of complete chloroplast genome sequences of two tropical trees *Machilus yunnanensis* and *Machilus balansae* in the family Lauraceae. *Front. Plant Sci.* **2015**, *6*, 662–670. [[CrossRef](#)]

45. Doorduyn, 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 Res.* **2011**, *18*, 93–105. [[CrossRef](#)]
46. Nei, M.; Kumar, S. *Molecular Evolution and Phylogenetics*; Oxford University Press: Oxford, UK, 2000.
47. Yi, L. Comparing and analyzing on models of calculation and statistical testing of nonsynonymous substitution rate and synonymous substitution rate during gene evolution. *J. Qujing Norm. Univ.* **2006**, *25*, 1–6.
48. Perry, A.S.; Wolfe, K.H. Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. *J. Mol. Evol.* **2002**, *55*, 501–508. [[CrossRef](#)]
49. Bittner-Eddy, P.D.; Crute, I.R.; Holub, E.B.; Beynon, J.L. RPP13 is a simple locus in *Arabidopsis thaliana* for alleles that specify downy mildew resistance to different avirulence determinants in *Peronospora parasitica*. *Plant J.* **2010**, *21*, 177–188. [[CrossRef](#)]
50. Dong, W.L.; Wang, R.N.; Zhang, N.Y.; Fan, W.B.; Fang, M.F.; Li, Z.H. Molecular evolution of chloroplast genomes of *Orchid* species: Insights into phylogenetic relationship and adaptive evolution. *Int. J. Mol. Sci.* **2018**, *19*, 716. [[CrossRef](#)]
51. Zhang, L. *Collection and Annotation of Suinong14 Full-Length Transcripts and Gene Diversity Analysis of Glyma13g21630*; Chinese Academy of Agricultural Science: Beijing, China, 2011.
52. Malaviya, D.R.; Roy, A.K.; Kaushal, P.; Kumar, B.; Tiwari, A. Genetic similarity among *Trifolium* species based on isozyme banding pattern. *Plant Syst. Evol.* **2008**, *276*, 125–136. [[CrossRef](#)]
53. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [[CrossRef](#)]
54. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Pyshkin, A.V. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [[CrossRef](#)]
55. Acemel, R.D.; Tena, J.J.; Irastorza-Azcarate, I.; Marlétaz, F.; Gómez-Marín, C.; Calle-Mustienes, E.; Mangenot, S. A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. *Nat. Genet.* **2016**, *48*, 336–341. [[CrossRef](#)]
56. Nadalin, F.; Vezzi, F.; Policriti, A. GapFiller: A de novo assembly approach to fill the gap within paired reads. *BMC Bioinform.* **2012**, *13*, S8. [[CrossRef](#)]
57. Laslett, D.; Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **2004**, *32*, 11–16. [[CrossRef](#)]
58. Lowe, T.M.; Chan, P.P. tRNAscan-SE On-line: Integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **2016**, *44*, W54–W57. [[CrossRef](#)]
59. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Thierer, T. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **2012**, *28*, 1647–1649. [[CrossRef](#)]
60. Wyman, S.K.; Jansen, R.K.; Boore, J.L. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **2004**, *20*, 3252–3255. [[CrossRef](#)]
61. Kumar, S.; Nei, M.; Dudley, J.; Tamura, K. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief. Bioinform.* **2008**, *9*, 299–306. [[CrossRef](#)]
62. Darling, A.C.E.; Mau, B.; Blattner, F.R.; Perna, N.T. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **2004**, *14*, 1394–1403. [[CrossRef](#)]
63. Zhang, Z.; Li, J.; Zhao, X.Q.; Wang, J.; Wong, G.K.S.; Yu, J. KaKs_Calculator: Calculating Ka and Ks through model selection and model averaging. *Genom. Proteom. Bioinform.* **2006**, *4*, 259–263. [[CrossRef](#)]
64. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; McVean, G. The variant call format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [[CrossRef](#)]
65. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [[CrossRef](#)]
66. Drummond, A.J.; Suchard, M.A.; Xie, D.; Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **2012**, *29*, 1969–1973. [[CrossRef](#)]

