Wu *et al. BMC Plant Biology* (2021) 21:421
https://doi.org/10.1186/s12870-021-03202-3

BMC Plant Biology

RESEARCH Open Access

Extensive genomic rearrangements mediated by repetitive sequences in plastomes of *Medicago* and its relatives

Shuang Wu[1], Jinyuan Chen[1], Ying Li[1], Ai Liu[1], Ao Li[1], Mou Yin[1], Nawal Shrestha[1], Jianquan Liu[1,2] and Guangpeng Ren[1*]

Abstract

Background: Although plastomes are highly conserved with respect to gene content and order in most photosynthetic angiosperms, extensive genomic rearrangements have been reported in Fabaceae, particularly within the inverted repeat lacking clade (IRLC) of Papilionoideae. Two hypotheses, i.e., the absence of the IR and the increased repeat content, have been proposed to affect the stability of plastomes. However, this is still unclear for the IRLC species. Here, we aimed to investigate the relationships between repeat content and the degree of genomic rearrangements in plastomes of *Medicago* and its relatives *Trigonella* and *Melilotus*, which are nested firmly within the IRLC.

Results: We detected abundant repetitive elements and extensive genomic rearrangements in the 75 newly assembled plastomes of 20 species, including gene loss, intron loss and gain, pseudogenization, tRNA duplication, inversion, and a second independent IR gain (IR ~ 15 kb in *Melilotus dentata*) in addition to the previous first reported cases in *Medicago minima*. We also conducted comparative genomic analysis to evaluate plastome evolution. Our results indicated that the overall repeat content is positively correlated with the degree of genomic rearrangements. Some of the genomic rearrangements were found to be directly linked with repetitive sequences. Tandem repeated sequences have been detected in the three genes with accelerated substitution rates (i.e., *accD*, *clpP*, and *ycf1*) and their length variation could be explained by the insertions of tandem repeats. The repeat contents of the three localized hypermutation regions around these three genes with accelerated substitution rates are also significantly higher than that of the remaining plastome sequences.

Conclusions: Our results suggest that IR reemergence in the IRLC species does not ensure their plastome stability. Instead, repeat-mediated illegitimate recombination is the major mechanism leading to genome instability, a pattern in agreement with recent findings in other angiosperm lineages. The plastome data generated herein provide valuable genomic resources for further investigating the plastome evolution in legumes.

Keywords: *Medicago*, *Trigonella*, *Melilotus*, IRLC, Plastome evolution, Genomic rearrangement, Repeat

* Correspondence: rengp@lzu.edu.cn
[1]State Key Laboratory of Grassland Agro-Ecosystems, Institute of Innovation Ecology & School of Life Sciences, Lanzhou University, Lanzhou, China
Full list of author information is available at the end of the article

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

In photosynthetic angiosperms, plastid genomes (plastomes) are highly conserved in gene content and structure and exhibit quadripartite structure with a pair of inverted repeats (IRs), which separate large single copy (LSC) and small single copy (SSC) regions [1, 2]. The plastomes of angiosperms typically consist of approximately 80 protein-coding genes, which play roles in photosynthesis and housekeeping along with 30 tRNA and 4 rRNA genes [3], of which approximately 17 genes are duplicated in the IR region. Because of the advent of high-throughput sequencing technologies, over 4500 land plant plastomes have been sequenced since the first tobacco plastome published in 1986 [4] and are publicly available in NCBI (accessed November 19, 2020). The size of these plastomes ranges from 16 to 242 kb. The majority of land plant plastomes range from approximately 110 to 170 kb, and the variation in plastome size is often attributed to IR expansion, contraction, or loss [5, 6]. The most enormous IR expansion is found in *Pelargonium transvaalense* (Geraniaceae) [7], where the IR expanded more than three times (87.7 kb) compared with the usual size of IR (~ 25 kb). On the opposite extreme, the IR loss, which causes reduction in plastome size, have been documented in many independent lineages, including two lineages of *Erodium* (Geraniaceae) [8, 9], *Carnegiea gigantean* (Cactaceae) [10], *Tahina spectabilis* (Arecaceae) [11], the Putranjivoid clade of Malpighiales [12] and the IR-lacking clade (IRLC) of Papilionoideae (Fabaceae) [13].

Although IR loss seems to be more common than previously thought, the presence of IR across angiosperms is still predominant, suggesting its functional importance in angiosperms. Early findings, which suggest that lineages lacking the IR have undergone more frequent genomic rearrangement than those that have retained the IR, support the hypothesis that IR plays a role in stabilizing plastome structure [13–15]. However, recent studies on the plastomes of Oleaceae [16], *Erodium* [8], *Pelargonium* [5–7], and *Plantago* [17] showed that the presence of IR does not ensure genome stability. Instead, the genome stability is more correlated with the overall repeat content in *Erodium* [8]. Other functions of IR, such as conservation of genes encoding the translational machinery [13] have also been suggested, with regard to the fact that substitution rates of genes in the IR are approximately three-fold slower than those in the single-copy (SC) regions [17–19]. However, this pattern of reduced IR substitution rates does not apply universally to many other plants. The IR genes from species in the genera *Pelargonium*, *Plantago,* and *Silene* have different levels of add substitution rates compared with the SC genes, which result from a mixture of locus-specific, lineage-specific, and IR-dependent effects [7, 17]. Increased locus-specific rates have been observed in plastomes of many plants [13, 14, 17, 20, 21], and such mutation hotspots are suggested to be linked to increased recombinational activities, which are likely driven by the proliferation of repeats. Consequently, repetitive DNA may have played an important role in structural variations of plastomes.

Previous studies have suggested that some species of the IRLC have acquired dramatic variations in plastome structures, including abundant inversions, mutation hotspots, gene transfers of *rpl22, infA*, substitution of *rps16*, losses or pseudogenization of *accD* and *ycf4*, IR reemergence, and losses of two *clpP* introns during their evolution [11, 20, 22–24]. Many of these rearrangement events have been also reported in *Passiflora* plastomes [25–28]. Illegitimate recombination between homologous and/or homeologous sequences within and between unit genome copies is proposed to yield structural variations in plastomes [29, 30]. Despite the IR losses, there are no clear signals of illegitimate recombination found in the IRLC species, or at least none remains as the IR loss is not recent [8]. A direct link between recombination and mutation hotspots was reported from the IR-lacking plastome of *Lathyrus*, in which *c.* 1.5 kb localized hypermutation region around *ycf4* was caused by repeated DNA breakage and repair [20]. In addition, repeat-mediated recombination-dependent replication has caused a ~ 9 kb IR reemergence in *Medicago minima* [11, 31]. Notably, plastomes of the IRLC species have abundant repetitive DNA and this seems to be rare in angiosperms' chloroplast DNA (cpDNA) [22]. However, whether the content of repetitive DNA is correlated with structural variations, and if yes, how repetitive DNA affect plastomes variations in the IRLC species remains understudied.

In this study, we focus on *Medicago* (*M.*) L. and its relatives *Trigonella* (*T.*) L. and *Melilotus* Miller, all belonging to the tribe Trifolieae, which is nested firmly within the IRLC. Species of these three genera are very important legume forage with significant ecological and economic values, including the widely cultivated major forage crop species *M. sativa*, the legume model species *M. truncatula* and the widely cultivated medicinal species *T. foenum-graecum*. Choi et al. [11] completed the plastomes of 19 *Medicago* species and one *Trigonella* species and revealed modest structural variations among them, but their discussion focused mainly on the IR reemergence in *M. minima*. Here, we took advantage of whole-genome resequencing data and assembled the plastomes of 75 individuals representing 20 species in the three genera. We aimed to characterize plastome structural variations of the 20 species at multiple individual levels and investigate the correlations between structural variations and repetitive elements.

Results

Plastome features

The plastome sequences were assembled and annotated for 75 individuals representing 20 species. Sizes of the plastomes ranged from 121,043 bp (*M. orthoceras*) to 142,713 bp (*Melilotus dentata*), and numbers of unique annotated genes from 110 to 111 (see Table S1). The total genes included 75–76 unique protein-coding genes (PCGs), 30 unique transfer RNA (tRNA) genes and 4 unique ribosomal RNA (rRNA) genes (Fig. S1). We found that the GC content of all plastomes ranged from 33.6 to 34.1% (Table 1; Table S1).

IR reemergence in *Melilotus dentata*

Assembly and annotation of plastomes of *Melilotus dentata* using a series of parameters (see Materials and methods) suggested the presence of a large inverted repeat (~ 15 kb, ranging in size from 15,336 bp to 15,553 bp), which contained 10 coding genes, including *ycf1*, the conserved four rRNA genes (4.5S, 5S, 16S, and 23S rRNA) and five tRNA genes (*trnR*-ACG, *trnN*-GUU, *trnA*-UGC, *trnI*-GAU, and *trnV*-GAC) in the seed plants (Fig. 1a; Fig. S2). We preformed read mapping to confirm the novel ~ 15 kb IR assembled in *Melilotus dentata*. The whole genome resequencing reads of *Melilotus dentata* were mapped to the assembled plastome sequence. Visualization of the mapping result (Fig. 1b) showed even distribution of reads over the assembled plastome sequence when both copies of the IR are included and ~ two-fold higher depth of coverage over the IR region compared with SC regions when only one IR copy is considered. This adds another evidence to the IR reemergence in the IRLC in addition to previous findings in *M. minima* [11, 31]. The IR reemergence in *Melilotus dentata* was further confirmed by multiple individuals, but with some INDELs (insertions and deletions) between the two IR copies in *Melilotus dentata* 02 and *Melilotus dentata* 03 (Fig. S2). We found that the size

Table 1 Information of plastome assembly, annotation, number of genomic rearrangements, and percent repetitive elements

Genus	Species	No. of individuals	Entire plastome size (bp)	Overall GC content (%)	No. of genes	No. of Protein-coding genes	No. of rRNA genes	No. of tRNA genes	No. of genomic rearrangements	Percent repetitive DNA (%)[b]
Medicago	*M. polymorpha*	5	124,247-124,445	34.0, 34.1	110	75	4	30	5	4.41–4.56
	M. truncatula	2	123,391-123,767	34.0	110	76	4	30	4, 5	4.00–4.16
	M. sativa	3	125,523-125,623	33.8, 33.9	110	76	4	30	4	4.13–4.55
	M. lupulina[a]	3	122,194-122,310	34.1	110	76	4	30	9	5.93–6.02
	M. minima[a]	5	132,071-132,219	34.2	110	75, 76	4	30	7, 8	4.99–5.06
	M. ruthenica	5	127,065-127,674	34.2	112	76	4	32	8	5.53–6.07
	M. archiducis-nicolai	4	126,635-126,810	34.1	112	76	4	32	7	5.20–5.38
	M. platycarpos	2	125,502-125,528	34.1	110	76	4	30	6	4.27
	M. falcata	3	125,357-125,555	33.8, 33.9	110	76	4	30	5	4.17–4.22
	M. edgeworthii	5	122,454-122,549	33.9, 34.0	110	76	4	30	4	4.01–4.09
	M. monantha	3	121,336-121,358	34.1	110	76	4	30	5	3.43–3.46
	M. orthoceras	5	121,043-121,065	34.1	110	76	4	30	5	2.98–3.11
	M. arcuata	2	121,728-121,777	34.0	110	76	4	30	5	3.54–3.57
	M. cancellata	5	121,889-121,953	34.0	110	76	4	30	5	3.57–3.63
Trigonella	*T. cachemiriana*	3	125,555	34.0	110	76	4	30	6	3.96–4.04
	T. emodi	5	128,493-128,643	33.8	110	76	4	30	9	6.19–6.29
Melilotus	*Melilotus dentata*[a]	4	141,922-142,713	33.7, 33.8	111	76	4	30	6	4.62–4.94
	Melilotus indicus	3	127,703-128,044	33.6	112	76	4	31	7	6.07–6.39
	Melilotus officinalis	4	126,534-127,451	33.7	111,112	76	4	31	5, 6	5.43–6.13
	Melilotus alba	4	127,293-127,694	33.6, 33.7	111, 112	76	4	31	6, 7	5.96–6.25
Total	**20**	**75**	**121,043-132,219**	**33.6–34.1**	**110–112**	**75–76**	**4**	**30–32**	**4–9**	**2.98–6.39**

[a] GC content, No. of genes, protein-coding genes, rRNA genes, tRNA genes, and Percent repetitive DNA were calculated using only one IR copy

[b] Tandem repeats and dispersed repeats ≥30 bp

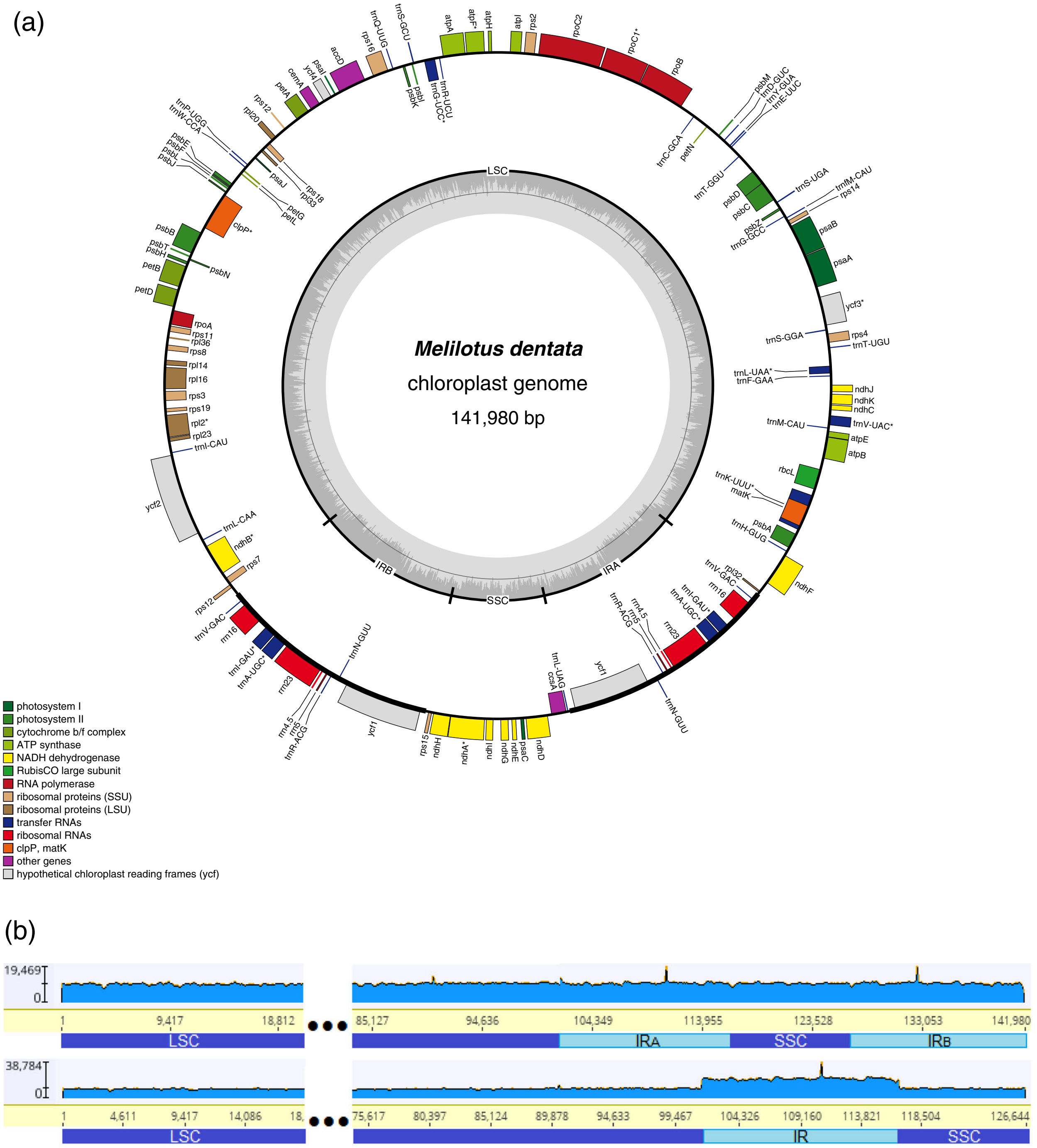

Fig. 1 Circular representation of *Melitotus dentata* plastome and confirmation of IR extent in it. **a** Circular representation of *Melitotus dentata* plastome. Genes shown outside the circle are transcribed clockwise and those inside are transcribed counter clockwise. Genes belonging to different functional groups are color-coded. The dark gray area in the inner circle indicates GC content and the thick black line shows the extent of different regions. LSC: large single copy; SSC: small single copy; IR: inverted repeat; IRA: inverted repeat A; IRB: inverted repeat B. **b** Confirmation of IR extent in *Melilotus dentata*. Whole genome resequencing reads were mapped to the assembled *Melilotus dentata* plastome containing both two copies (upper) or a single copy (lower) of the novel IR. The scale at the left reports the depth of reads, which is indicated graphically by the blue histogram

variation of the two IR copies was due to the difference in the copy number of tandem repeats (Fig. S2).

Plastome structural variations

The plastome structural variations, including three IR regain, two pseudogenization, three gene loss, four intron loss, two intron gain, two tRNA duplicate, and 10 inversions (using *Wisteria floribunda* as the reference, see Fig. S3), were detected among the 20 species (Fig. 2; Table S1). Structural variations among individuals within each species were mostly the same, except for the pseudogenization of *ycf2* (presence of premature stop codons within the gene) in *M. minima* 01, duplicate of *trnN*-GUU in some individuals of *Melilotus officinalis* and *Melilotus alba* (Table S1), and two distinct plastome configurations in *M. truncatula* (Fig. S1k; Table S1). Most of the variations were shared by two to multiple species, while some of them were specific to certain species. The *rpl22* and *infA* were absent in all the 20 species, whereas *rps16* was lost in the *Medicago* and *Trigonella* clades, and pseudogenized (presence of premature stop codons within the gene) in the *Melilotus* clade. Two kinds of duplicated tRNA were present in three *Melilotus* and two *Medicago* species. The tRNA duplicates were confirmed by read mapping as shown in Fig. S4. The intron loss of *clpP*, *atpF* and *rpoC1* in *M. lupulina* and *M. minima* was consistent with previous findings [11]. The intron 1 of *clpP* was lost in all the 20 species, consistent with previous studies [23, 32]. We further found that the intron loss of *rpoC1* was shared by six *Medicago* species (Fig. 2; Table S1). After validation based on transcriptomic data (see more details in Methods), the intron gain of *ycf1* and *accD* were specific to *T. emodi* and *M. falcata*, respectively. Only one gene pseudogenization (*accD*, truncated sequence) was unique to *M. polymorpha*, and others were shared by multiple

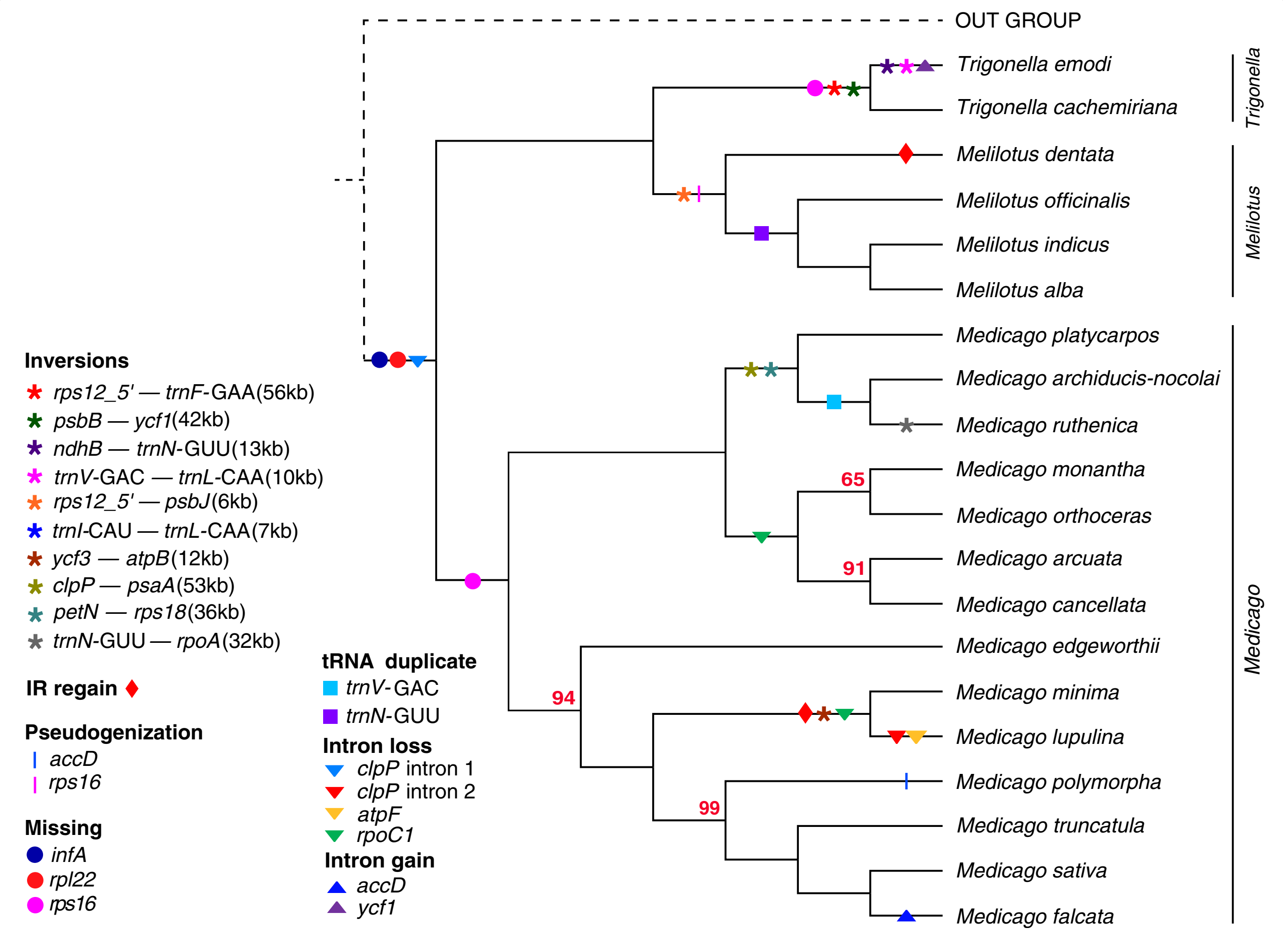

Fig. 2 Phylogenetic relationships and distributions of plastome structural variations. Structural features of all 20 taxa are plotted on the branches of the cladogram inferred using PhyML and indicated in the key (inset) and inversions are relative to *Wisteria floribunda*. The red number at the node indicated bootstrap values < 100. The branch of the outgroup was indicated by dotted line and did not participate in the statistics and labeling of structural variation. Structural features for all individuals are listed in Table S1

species, mostly by closely related species. For inversions, the large ones (> 30 kb) occurred mainly in two groups: the *Trigonella* clade and the clade containing *M. platycarpos*, *M. ruthenica*, and *M. archiducis-nocolai* (section *Platycarpae*). For the remaining inversions, two were unique to *T. emodi*, two were present in the *Melilotus* clade, and one was shared by *M. lupulina* and *M. minima*.

Repeat analysis

The analysis of the 75 plastomes recognized 64–597 pairs of dispersed repeats (including forward, reverse, palindromic and complement repeat), ranging in size from 2550 to 7315 bp (Table S2). The most abundant repeat type was forward repeats with the number ranging from 40 pairs in *M. orthoceras* to 571 pairs in *Melilotus indicus*. The second abundant repeat type was palindromic repeats ranging from 17 pairs in *M. edgeworthii, M. truncatula* 02, and *T. cachemiriana* to 47 pairs in *M. truncatula* 01. Then a pair of reverse repeats in *M. polymorpha*, *M. edgeworthii*, *M. monantha*, and *Melilotus dentata* to 50 pairs in *Melilotus officinalis*. Lastly, the complimentary repeats were infrequent among the species ranging from 1 to 7 while others did not have such as *M. polymorpha*, *M. truncatula*, *M. sativa*, *M. archiducis-nicolai*, and *M. falcata*. A total of 38–110 tandem repeats were recognized across the 75 plastomes with their sizes ranging from 2033 to 5951 bp (Table S2).

Overall, we identified 3604 (*M. orthoceras*) to 8179 bp (*Melilotus indicus*) repetitive sequences across the 75 plastomes, accounting from 2.98 to 6.39% of their full plastomes, respectively (Table 1, Table S2). We found that the overall repeat content showed significant positive correlation with the degree of genomic rearrangement ($R = 0.77$, $P < 6.1e\text{-}16$; Fig. 3a). Among the repeats, both the dispersed repeats and tandem repeats also showed significant positive correlation with these structural variations (Fig. S5). Some of the genomic rearrangements were found to be directly linked to repetitive sequences (Fig. 3b-d; Figs. S6-S8; Tables S3-S6). For example, the three copies of *trnV*-GAC in *M. archiducis-nicolai* were linked to two forward repeats and one tandem repeat (Fig. 3b; Fig. S6a; Table S3). The detected 56 kb inversion (*rps12_5'—trnF*-GAA) in *T. cachemiriana* had a pair of 41 bp inverted repeats flanked to its two endpoints (Fig. 3c; Fig. S7a; Table S4; Table S5). Such short inverted repeats were also present in the 32 kb inversion (*trnN*-GUU—*rpoA*) in *M. ruthenica* and the 36 kb inversion (*petN—rps18*) in *M. archiducis-nicolai, M. ruthenica* and *M. platycarpos* (Fig. S7i, j; Table S4; Table S5). Furthermore, *M. truncatula* 02 had a ~ 44-kb inversion compared to the *M. truncatula* 01, mediated by a short, imperfect repeat (Fig. S1k; Table S1), which is consistent with the findings of Gurdon and Maliga [33]. In addition, the gained intron mainly consisted of repetitive sequences (Fig. 3d; Fig. S8; Table S6).

Comparative plastome analysis

To display interspecific variations graphically, the sequence variations of the 20 species was generated using mVISTA with plastome of *M. falcata* as the reference (Fig. S9). The most highly divergence regions among the 20 species appeared mostly in the intergenic spacers, while the coding regions showed relative conservatism except for the genes *clpP*, *atpF*, *rpoC1*, *accD*, *ycf1* and *ycf2*. As a result, we identified eight highly divergent coding regions ($\pi > 0.04$) and 16 highly divergent noncoding regions ($\pi > 0.1$) (Fig. 4; Table S7).

Acceleration of substitution rates in *accD, clpP,* and *ycf1*

We found that there were significant increases ($P < 0.0001$) in the substitution rates (dN and dS) of *clpP*, *accD*, and *ycf1* compared to *matK* and *rbcL* for all the comparisons within both IRLC taxa and our own 20 species (pink branch) (Tables S8-S12; Fig. 5; Fig. S10). Furthermore, all of the investigated IRLC species including our 20 species showed signs of various degrees of elevated branch lengths in the dN and dS trees of *accD*, *clpP*, and *ycf1*, but again no similar pattern was seen in *matK* and *rbcL* (Fig. 5; Fig. S11). The three genes also exhibited high variations in terms of their coding sequence (CDS) length compared with *matK* and *rbcL* in all the 20 species (Table S13). The CDS length of *accD* varied from 1299 bp (*M. falcata*) to 2190 bp (*M. minima*); *clpP* from 582 bp (*M. falcata*) to 684 bp (*M. ruthenica*) and *ycf1* from 4893 bp (*Melilotus alba* and *Melilotus indicus*) to 5352 bp (*M. lupulina*), while the length of the other two genes did not show much variation (1512 or 1521 bp in *matK* and 1428 bp in *rbcL*) (Table S13). To investigate the mechanisms of length-associated mutation for CDS of these three genes, we counted the number and overall length of repetitive elements in the CDS of all the five genes (Table S13). As expected, *accD* and *ycf1* in 20 species all included tandem repeated sequences whose numbers were positively correlated with gene length. The longest length of *clpP* was found in *M. ruthenica*, which also included a tandem repeated sequence. However, we did not detect any tandem repeated sequence in *matK* and *rbcL*. We also found that the repeat contents within these genes and their neighboring hypermutable regions ($\pi > 0.1$) (Fig. 4) were significantly higher ($P < 0.0001$) than that of the remaining plastome sequences (Fig. 6; Table S14). Therefore, the acceleration of substitution rates of these three genes may be explained by the insertions of repetitive sequences.

Discussion

IR reemergence in *Melilotus dentata*

The existence of a pair of inverted repeats is a feature of plastomes throughout the 400 million years of land plant

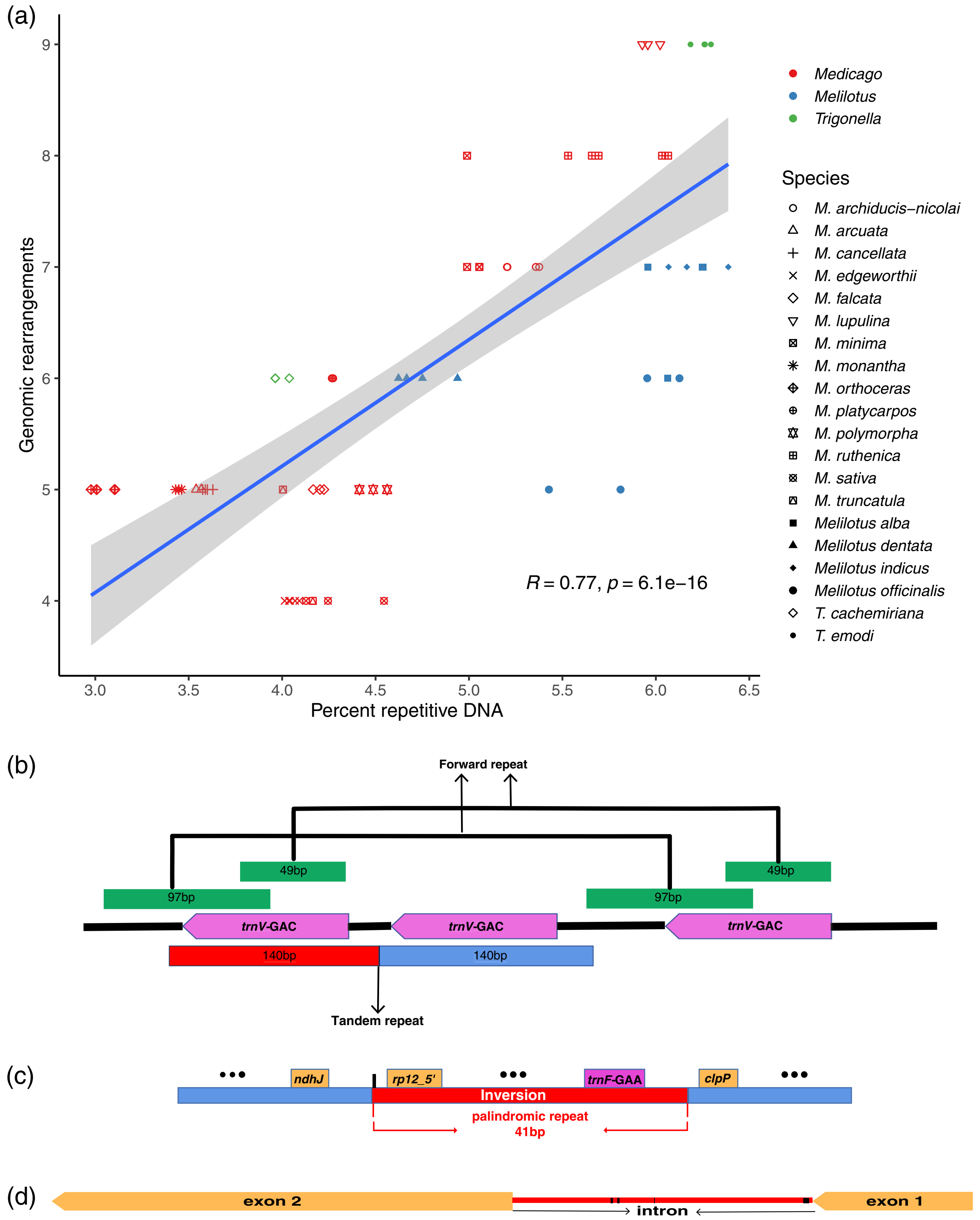

Fig. 3 Relationships between repetitive sequences and plastome structural variations. **a** The overall repeat content shows significant positive correlation with the degree of genomic rearrangements ($R = 0.77$, $P < 6.1e\text{-}16$). **b**-**d** Examples of repeat-mediated tRNA duplicate, inversion and intron gain. The gained intron is constituted mostly by repeats (red)

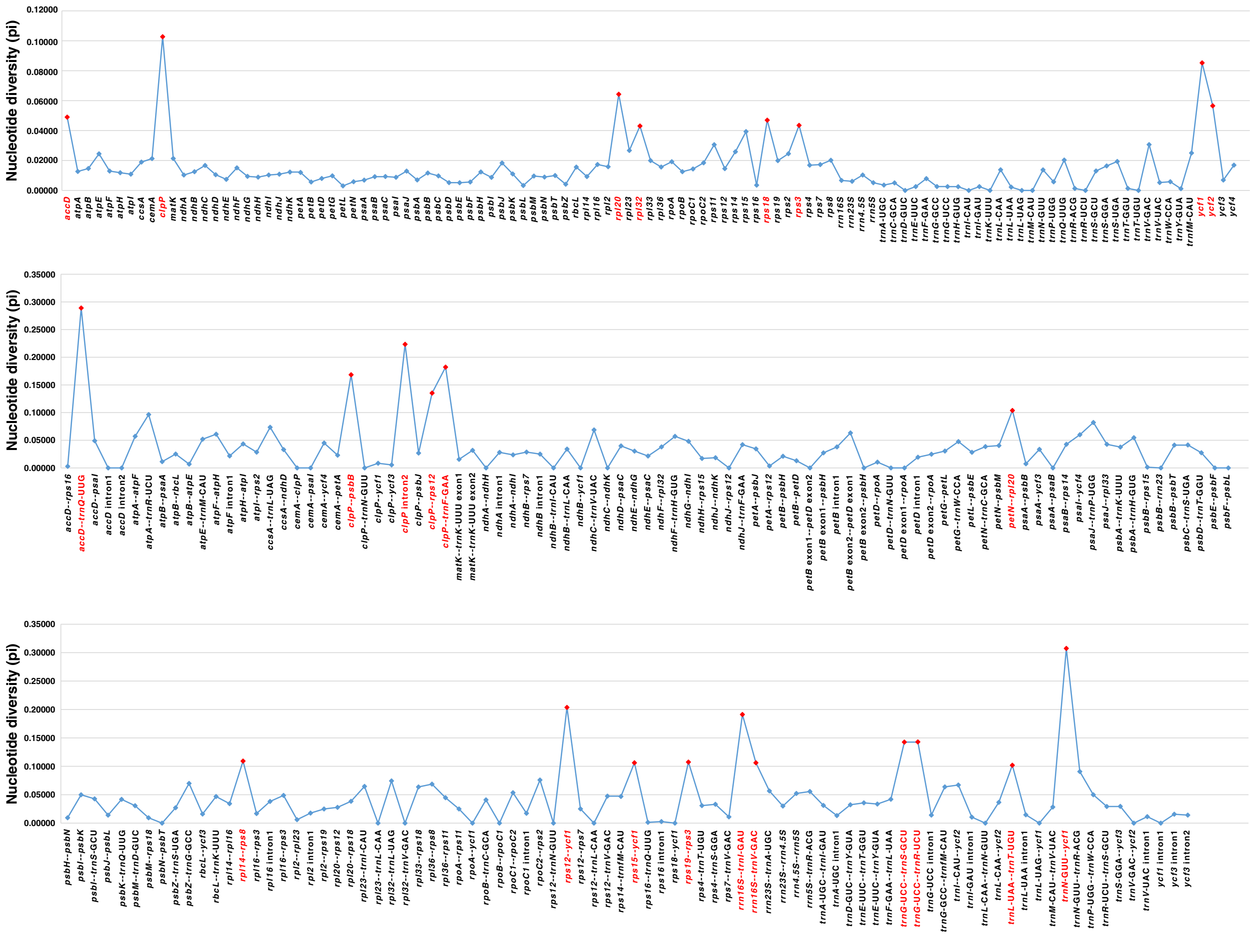

Fig. 4 Comparison of nucleotide diversity (Pi) among the 20 species. X-axis: names of genes or intergenic regions, Y-axis: nucleotide diversity (Pi) of each region. We identified 8 genes (Pi > 0.04) and 16 intergenic regions (Pi > 0.1) as highly divergent regions and colored them with red

evolution [34] with the exception that a copy of IR is absent in some lineages, such as the IRLC [35], some species of *Erodium* (Geraniaceae) [8, 14], Pinaceae [36], and *Carnegiea gigantea* [10]. In addition to the first ~ 9 kb IR reemergence (containing 7 coding sequences) found in *M. minima* in the IRLC [11], we reported here the second case in *Melilotus dentata*, in which ~ 15 kb of a large inverted repeat is regained, containing 10 coding sequences compared to the ~ 17 in the typical IR of angiosperms. Choi et al. [11] speculated that the IR reemergence in *M. minima* occurs via synthesis-dependent strand annealing or the formation and resolution of Holiday junctions during recombination-dependent DNA repair. Both processes need to be mediated by repeats. Choi et al. [11] detected unique variations (i.e., two pairs of inverted repeats flanked by *rps12_5'* and between *trnN*-GUU and *ycf1*, and a series of tandem repeats between *trnN*-GUU and *ycf1*) retained in *Medicago suffruticosa,* which is closely related to *M. lupulina* and *M. minima*. The position of repeats in *M. suffruticosa* flanked the endpoints corresponding to the novel IRs in *M. minima* suggests that their ancestor may have experienced repeat-mediated perturbations, which may have caused IR regain in the two species. Repeat-mediated and recombination-dependent replication play a significant role during the IR reemergence [11]. However, we did not detect a similar pattern in *Melilotus dentata* and its relatives, which may be due to incomplete sampling in our study. Dynamic characteristics of IR boundary migration process [17, 37] and the IR regain in *M. minima* and *Melilotus dentata* supports the hypothesis that the novel IRs are quite possible to expand continuously and include more adjacent canonical IR genes gradually [11]. A deeper and wider sampling of the IRLC species is necessary to further investigate how many species of the IRLC have regained novel IR and the mechanism behind this process.

Acceleration of substitution rates in *accD*, *clpP,* and *ycf1* are related to repetitive sequences

Gene-specific rate acceleration has been frequently reported in plastome evolution. Among legumes, *clpP* and

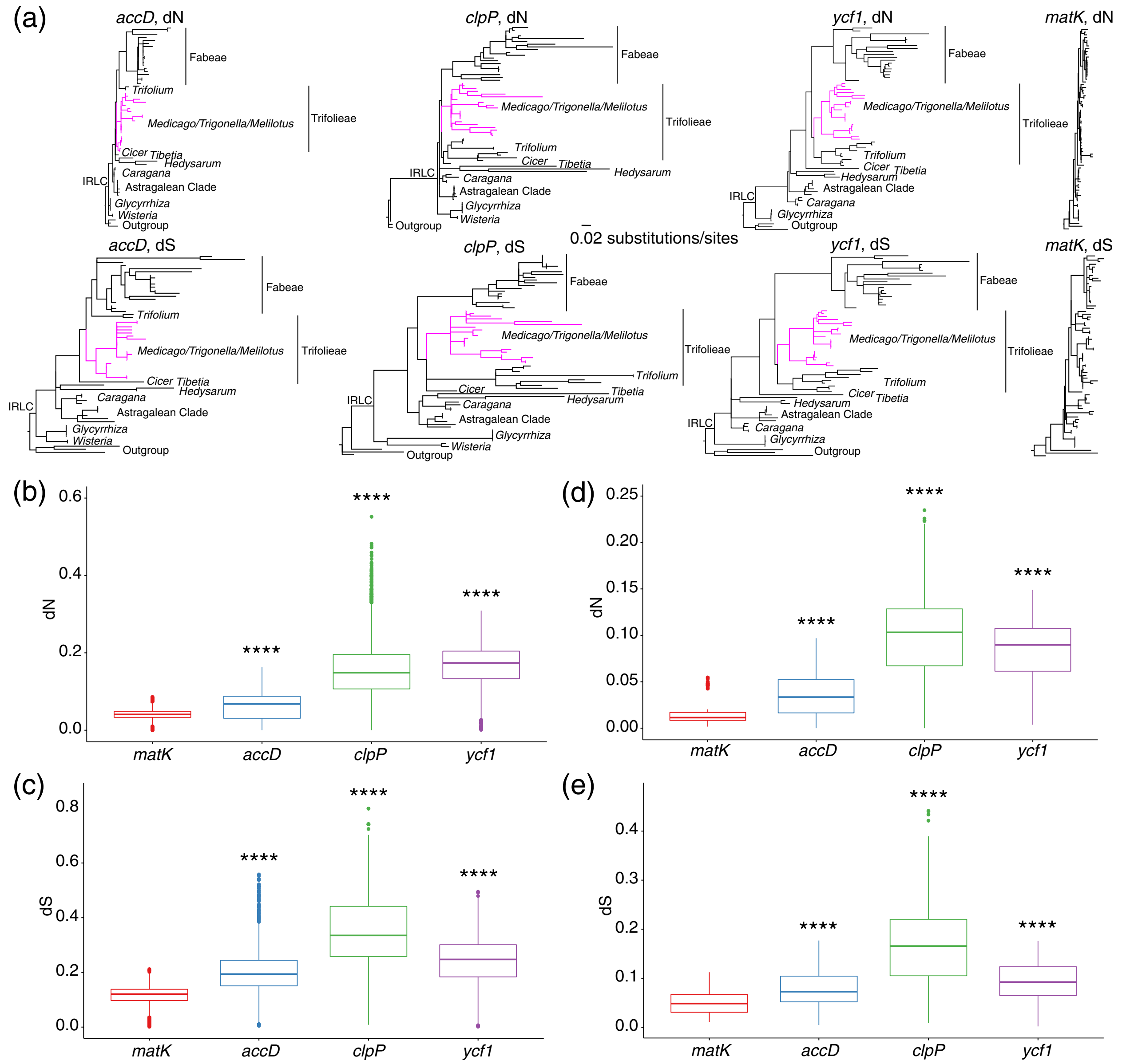

Fig. 5 Acceleration of substitution rates in *accD*, *clpP* and *ycf1*. **a** Synonymous (dS) and nonsynonymous (dN) divergence in the IRLC species for four chloroplast genes: *accD*, *clpP*, *ycf1* and *matK*. Shown are dN (upper) and dS (lower) trees resulting from a codon-based maximum likelihood (ML) analyses performed by RAxML, rooted using three Robinioid sequences (non-IRLC species): *Lotus japonicus*, *Sesbania grandiflora* and *Robinia pseudoacacia*. All trees are drawn to the same scale. The species are in the same order from top to bottom in dN and dS trees of each gene, to the greatest extent possible, and are named in full in Fig. S11. Pink branches in the trees for *accD*, *clpP*, and *ycf1* represent species in our study. Trees for *matK* do not show comparable rate heterogeneity at either synonymous or nonsynonymous sites. **b**-**e** Significance test for nonsynonymous (dN) (upper) and synonymous (dS) (lower) substitution rates of *clpP*, *accD* and *ycf1* compared to *matK* for all the comparisons within both IRLC taxa used in this study (left) and our own 20 species (right). ****, $P < 0.0001$ (T-test). Detailed information can be found in Tables S8-S11

ycf1 have been reported as rapidly evolving genes in Papilionoids [38]. The gene *clpP* is also accelerated in Mimosoid [24] and *ycf4* is accelerated in most legumes especially in the tribe Fabeae [20, 39]. Accelerated evolutionary rate of *accD* in *Jasminum* [16], *Silene* [38], and *Plantago* [17] has also been described. In agreement with previous findings, our results indicate that the three genes exhibit different degrees of substitution rate acceleration (Fig. 5; Figs. S10, S11) with high variations in point mutations and length as compared to *matK* and *rbcL* (Fig. 4; Tables S7-S13). The gene *accD* encodes acetyl-CoA carboxylase, which acts in fatty acid biosynthesis [40] and *clpP* encodes a protein, which is a part of multimeric protease [41]. For *ycf1,* recent experiments

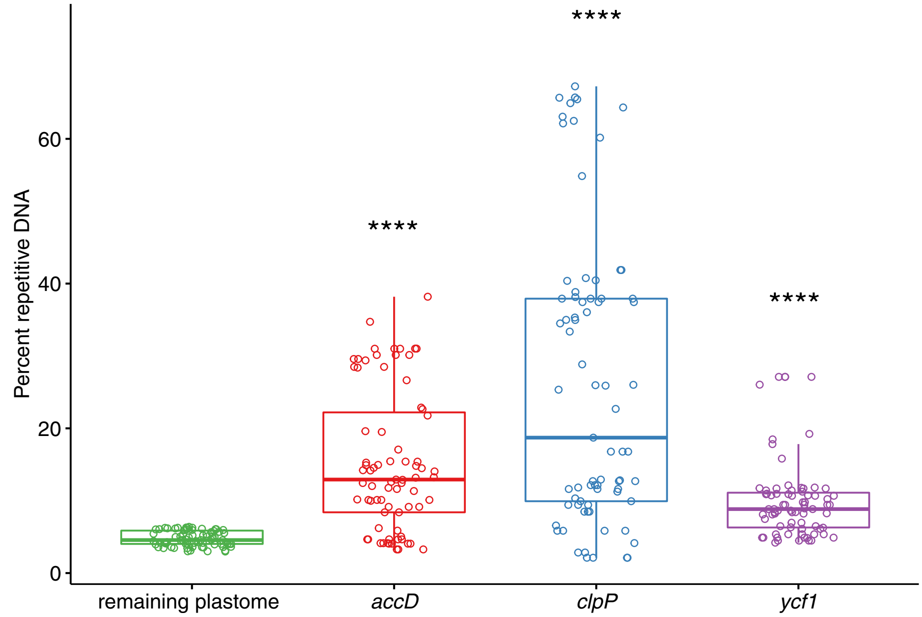

Fig. 6 Repeat content of the three localized hypermutation regions around the three genes (*accD, clpP,* and *ycf1*) compared to the remaining part of plastome sequences across the 75 individuals. ****, *P* < 0.0001 (T-test). Detailed information can be found in Table S14

suggest that it encodes Tic214, which is a vital component of *Arabidopsis* translocon on the inner chloroplast membrane (TIC) complex, thus it is essential for plant viability [42]. Notably, tandem repeated sequences have been detected in these three genes and their length variation could be explained by the insertions of tandem repeats (Table S13). Interestingly, our results suggest that the overall repeat contents within the three genes with accelerated substitution rates and their neighboring hypermutable regions (π > 0.1) were significantly higher than that of the remaining plastome sequences (Fig. 6; Table S14). It is possible that the repetitive elements inserted into or near the three genes might have promoted them to become more variable, resulting in the acceleration of the substitution rate.

Plastome structural variations are mediated by repetitive sequences

Although the gene content and order of plastome is highly conserved in most seed plants [1], extensive genomic rearrangements have mostly been reported in Fabaceae [22, 23, 43, 44], particularly within the IRLC [23]. Indeed, we detected abundant plastome structural variations, including three IR regains, three gene losses, two pseudogenization, four intron losses, two intron gains, two tRNA duplications, and ten inversions in the 20 species (Fig. 2; Table S1). Among these variations, some are reported in previous studies too. For example, the loss of *rpl22* and *infA* found in the 20 species and also in other legumes [45] and almost all rosids [46] have been suggested as successful gene transfers from the plastome to the nuclear genome. Similarly, the loss or pseudogenization of *rps16* in the 20 species could be explained by gene substitution as suggested for *M. truncatula* [47].

With these extensive genomic rearrangements, we must consider what factors might contribute to the plastome instability in these species. The hypothesis that IR plays an important role in stabilizing plastome structure may be one possible explanation because all our 20 species are defined as IR-lacking species within the IRLC [13–15]. However, relatively higher structural variations were detected in the two IR reemergence species (*M. minima* and *Melilotus dentata*, IR > 9 kb) compared with other species (Fig. 2), such as *M. truncatula, M. sativa*, and *M. edgeworthii*, do not support this hypothesis. With more available plastome sequences, the recent findings in a few unrelated lineages suggested that neither loss of the IR has destabilized the plastome nor the presence of the IR has ensured genome stability [7, 8, 16, 17]. Instead, the accumulation of repeat sequences may be more important to plastome stability as suggested by many recent studies [8, 20, 48].

We did find abundant repetitive elements in the 20 species and a significant positive correlation between the overall repeat content and the degree of genomic rearrangements (Fig. 3a). Moreover, the dispersed repeats likely contributed more to these variations than the tandem repeats (*R* = 0.73 vs. 0.48; Fig. S5). Specifically, we found a high frequency of repeats flanked the endpoints of most inversions (Fig. 3c; Fig. S7; Tables S4, S5), and the gained introns were constituted mainly by repeats (Fig. 3d; Fig. S8; Table S6). Furthermore, the two tRNA duplicates were mediated by tandem repeats and/or forward repeats (Fig. 3b; Fig. S6; Table S3). Many previous studies have also reported that repetitive sequences are commonly found in the flanking regions of inversions, losses, and tRNA duplicates in the plastomes of angiosperms [14, 33, 48–50]. Repeat-mediated illegitimate recombination is thought to be one of the major mechanisms leading to these genomic rearrangements [51, 52]. As recombinogenic substrates, when a plastome's repeat content increases, differential resolution of recombination events (including illegitimate recombination) between highly similar regions (e.g., repeats) within and between unit genome copies become more likely, which may ultimately result in a rearranged plastome.

Although it is unclear how repetitive elements are generated, our results suggest that their appearances and losses are a dynamic process and are random on an evolutionary timescale. Given the fact that most of the genomic rearrangements are species-specific or only shared by closely related species, if all these different variations are due to repeat-mediated illegitimate recombination, then those repeats must be different and their appearances should be random during the evolutionary histories of these species. Similarly, repeats may be lost randomly. One contrary assumption is that if losses of repeats are progressively on a timescale, then we would expect a similar repeat pattern flanked the endpoints of genomic rearrangements on the same evolutionary scale. However, the two inversions (Fig. S7a, b) shared by *T. cachemiriana* and *T. emodi*, and another two inversions (Fig. S7h, i) shared by *M. ruthenica*, *M. archiducis-nicolaiand*, and *M. platycarpos*, show different distribution patterns of repeats, which do not support the assumption. It should be noted that the mechanisms behind the generation and loss of repeats are very complex and our data is not suitable to gain a detailed knowledge of the processes at play here. Future studies involving dense taxon sampling and/or proteins and pathways implicated in recombination, selection, and repair in plastomes (e.g. [52]) should improve our understanding of plastome disruption [48, 53].

Conclusions

In this study, we completed 75 plastomes representing 14 species of *Medicago* and 6 species from its two closely related genera (*Melilotus* and *Trigonella*) and reported a second independent IR gain in *Melilotus dentata* for the IRLC species. We detected abundant repetitive elements and extensive genomic rearrangements in these plastomes. Notably, we found that the overall repeat content is positively correlated with the degree of genomic rearrangements. Moreover, the overall repeat contents within localized hypermutation regions around genes with accelerated substitution rates were also significantly higher than that of the remaining plastome sequences. Our findings highlight the role of repetitive sequences in affecting plastome stability in the IRLC species. The plastome data generated herein provide valuable genomic resources for further investigating the plastome evolution in legumes.

Methods

Plant materials

We sampled 75 individuals representing 20 species of *Medicago* and its relatives *Trigonella* and *Melilotus,* including 14 *Medicago,* 2 *Trigonella*, and 4 *Melilotus* species. Our sampling included 19 species of the three genera that are distributed in China and the model species of *Medicago*, i.e. *Medicago truncatula*. For *M. truncatula,* plastomes were assembled using whole-genome resequencing data, which were downloaded from NCBI (SRR1524305 and SRR965443, https://www.ncbi.nlm.nih.gov/). For *Medicago sativa*, seeds were germinated in the greenhouse and emergent leaves from a single plant from each accession were collected. For other species, leaves were collected from the wild (Table S15). No specific permissions were required for the relevant locations/activities. We followed the Flora of China (http://www.iplant.cn/foc/) for the nomenclature system in this study, but treated *T. arcuata* and *T. cancellata* as *M. arcuata* and *M. cancellata*, respectively, based on a recent study [54]. All voucher specimens were deposited in Lanzhou University. We selected 2–5 individuals from each species for whole genome sequencing and used the available plastome of *Trifolium subterraneum* (NC_011828.1) as an outgroup (Table S15).

DNA sequencing, assembly, and annotation

The total genomic DNA was extracted using the modified CTAB procedure from the dried leaves [55], which were then sent to the BGI Genomics for sequencing. Paired-end libraries (2 × 150 bp) were constructed and sequenced using the Illumina Hiseq X-Ten Platform (Illumina, San Diego, CA). The raw whole genome reads (Table S1) were first quality checked by FASTQC [56], and the results showed that all the raw data were clean without adapter contamination. We, therefore, used raw genome data to de novo assemble the plastome of each individual using NOVOPlasty v.3.8.3 [57]. We used the relatively conserved *rbcL* chloroplast gene sequence as a seed to assemble the plastomes. For the four *Melilotus* species, we selected the *rbcL* sequence of *Melilotus albus* (GenBank accession: NC_041419.1) as seed. The *rbcL* sequence of *Trigonella foenum-graecum* voucher I.S.

(GenBank accession: NC_042857.1) was used as seed for *Trigonella cachemiriana* and *Trigonella emodi*. For *Medicago polymorpha, M. truncatula, M. sativa, M. lupulina, M. minima*, and *M. falcata,* their seeds were provided by their own *rbcL* sequences (GenBank accession: NC_042848.1, NC_003119.8, NC_042841.1, NC_042847.1, NC_042849.1, and NC_032066.1). The *rbcL* sequence of *M. falcata* (GenBank accession: NC_032066.1) was used as seed for other *Medicago* species. We left other parameters as the default values (see NOVOPlasty README.md). We annotated the assembled plastomes using GeSeq [58] with MPI-MP chloroplast references and HMMER profile search. We further confirmed all tRNAs by tRNAscan-SE v.2.0.5 [59]. For confirmation, all annotations were compared with previously published plastomes of the three genera available in NCBI, and exon boundaries were manually corrected in Geneious v.10.2.6 [60]. In addition, we found an extra sequence inserted within *ycf1* in the four *Melilotus* species and *T. emodi*, and two extra sequences inserted within *accD* in *M. falcata* when comparing with the sequences of the two genes in other species. In order to validate whether these insertions were introns or not, we downloaded transcriptomic data of *M. falcata* (SRR1823822), *Melilotus albus* (SRR5115455), and *Trigonella foenum-graecum L.* (SRR8281660) from NCBI. For the other three *Melilotus* species and *T. emodi,* transcriptomic data are not available and the validation of *ycf1* in these species was based on the transcriptomic data of *Melilotus albus* and *Trigonella foenum-graecum*, respectively. We assembled these transcriptomes to get fasta sequences using trinityrnaseq-Trinity-v2.8.5 [61], then converted them to coding sequence (CDS) using TransDecoder-v5.5.0 (http://transdecoder.sf.net) and aligned the coding sequence (CDS) with our gene sequences (both exons and introns) via ncbi-blast-2.10.0+ [62]. In this way, we confirmed the authenticity of intron gain and corrected intron boundaries. The results showed that the intron gain of *ycf1* was true in *T. emodi* (intron boundaries were also true) but false in the four *Melilotus* species. In addition, the intron gain of *accD* in *M. falcata* was true (it does have two introns) but the intron boundaries were different from the original result. The visual images of the annotations of all species were generated by OGDRAW v.1.3.1 [63] (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html). All the plastomes of the 75 individuals were newly assembled and deposited in the Genome Warehouse of CNCB-NGDC (under BioProject accession PRJCA005341).

Inversion inference and repeat content estimate

Inversions were identified according to the arrangement of locally colinear blocks (LCB) among the newly assembled plastomes of 20 species that were estimated using progressiveMauve v.2.4.0 [64]. In progressiveMauve alignments, we selected *Wisteria floribunda,* an early diverging IRLC taxon, as the reference to identify inversions.

We calculated the repeat content of all the 75 plastomes. Tandem Repeats Finder v.4.09 [65] was used to characterize tandem repeats with the following parameters: Match of 2, Mismatch and Delta of 7, PM of 80, PI of 10, Minscore of 50, and MaxPeriod of 500. We used REPuter program [66] to identify dispersed repeats with a minimum repeat size of 30 bp and identity of no less than 90% (hamming distance equal to 3). Furthermore, we wrote a custom Perl script to eliminate the influence of nested or overlapping repeats in subsequent analysis. We only considered one IR copy for the species containing two copies of IR, both in the inversion inference and repeat content estimation.

Comparative plastome analysis

Because structural variations among individuals within each species were almost the same (see Results), we selected 20 individuals representing all the 20 species (see the selected individuals in Table S15) for comparative analysis. To display interspecific variations graphically, full alignments with annotations of the 20 species were plotted using mVISTA [67] in the Shuffle-LAGAN mode with the annotation of *M. falcata* as a reference. To detect the sequence divergence and determine highly divergent regions of the 20 species, we used the python script "get_annotated_regions_from_gb.py" (https://github.com/Kinggerm/PersonalUtilities/) [68] to automatically extract all annotated regions and regions between annotations of the 20 plastomes, and aligned all regions using MAFFT v.7.453 [69]. We used the python script "concatenate_fasta.py" (https://github.com/Kinggerm/PersonalUtilities/) [68] to concatenate the alignments of the separate loci. Finally, we constructed three datasets, which included concatenated coding regions (PC), the concatenated noncoding loci (PN), and whole plastome (PCN). Subsequently, the nucleotide diversity (pi) of each coding gene and non-coding regions (i.e. PC, PN, and PCN) was calculated using DnaSP v.6 [70].

Phylogenetic inference

To reconstruct the phylogenetic relationships among *Medicago* and its relatives *Trigonella* and *Melilotus,* we selected the same 20 individuals as mentioned above and used *Trifolium subterraneum* (NC_011828.1) as an outgroup. The coding sequences (CDS) of 73 protein-coding genes (PCGs) shared across the 21 species (Table S16) were extracted from each plastome using the python script "get_annotated_regions_from_gb.py" (https://github.com/Kinggerm/PersonalUtilities/) [68]. Each region was individually aligned using MAFFT v7.453 [69]. Then we used the python script "concatenate_fasta.py" (https://github.com/Kinggerm/PersonalUtilities/) [68] to concatenate the alignments. FASTA files were converted to PHYLIP format using ClustalW v.2.1 [71]. Molecular phylogenetic analysis was conducted by Maximum likelihood (ML) analysis using PhyML 3.1 [72] with 100 bootstrap replicates based on the best-fit model 012003 + I + G + F estimated by jModeltest v.2.1.7 [73].

dN and dS analysis

Nonsynonymous (dN) and synonymous (dS) substitution rates were calculated using "yn00" from the PAML4.8 package [74] for five coding regions: *accD*, *clpP*, *ycf1*, *rbcL*, and *matK*. We then downloaded the sequences of these five genes of the IRLC species from NCBI (taxon names and accession numbers are listed in Table S17) and constructed ML trees using RAxML [75] with a general time-reversible model GTR + G + I and 100 bootstrap replicates for each of the five genes. The ML trees, which were generated using RAxML, were used as the constraint trees for five genes during the branch-specific dN and dS rate estimation in PAML4.8/codeml free-ratio model (model = 1).

Abbreviations

M.: *Medicago*; *T.*: *Trigonella*; IRLC: Inverted repeat lacking clade; cpDNA: Chloroplast DNA; SC: Single-copy; LSC: Large single copy; SSC: Small single copy; IR: Inverted repeat; INDELs: Insertions and deletions; CDS: Coding sequence; PCGs: Protein-coding genes; tRNA: Transfer RNA; rRNA: Ribosomal RNA; PC: Plastome coding sequence; PN: Plastome noncoding sequence; PCN: Whole plastome sequence; pi: Nucleotide diversity; dN: Nonsynonymous substitution rate; dS: Synonymous substitution rate; P: *P*-value in statistical tests; TIC: Translocon on the inner chloroplast membrane

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-021-03202-3.

Additional file 1 : **Table S1.** The plastome assembly, annotation information, and distributions of genomic rearrangements for the 75 individuals. **Table S2.** Detailed information of repeat content for the 75 individuals. Repeat content for the three IR regained plastomes were calculated using only one IR copy. **Table S3.** Repeats mediated tRNA duplicates. For dispersed repeats, F: forward repeat; C: complement repeat; P: palindromic repeat; R: reverse repeat, and the numbers after the colon represent length of dispersed repeats. For tandem repeats, the numbers before the colon represent length of tandem repeats, the content after the colon represent unit size × copy number. **Table S4.** Repeats around endpoints of inversions. For dispersed repeats, F: forward repeat; C: complement repeat; P: palindromic repeat; R: reverse repeat, and the numbers after the colon represent length of dispersed repeats. For tandem repeats, the numbers before the colon represent length of tandem repeats, the content after the colon represent unit size × copy number. Palindromic repeat (P) are marked in red. **Table S5.** Palindromic repeat sequences around endpoints of inversions. **Table S6.** Repetitive DNA in the acquired introns. **Table S7.** Nucleotide diversity (pi) for different genes, intergenic regions, and datasets. PC, plastid coding regions; PN, plastid noncoding regions; PCN, the whole plastome. Eight highly divergent coding regions ($\pi > 0.04$) and 16 highly divergent noncoding regions ($\pi > 0.1$) are marked in red. **Table S8.** Sequence divergence in accD among the IRLC species. **Table S9.** Sequence divergence in clpP among the IRLC species. **Table S10.** Sequence divergence in ycf1 among the IRLC species. **Table S11.** Sequence divergence in matK among the IRLC species. **Table S12.** Sequence divergence in rbcL among the IRLC species. **Table S13.** Information of repetitive elements for the coding sequence (CDS) of three genes (accD, clpP and ycf1) with accelerated substitution rates and two relatively conserved genes (matK and rbcL). There is no coding sequence (CDS) in accD for *M. polymorpha* because it is a pseudogene (truncated sequence). **Table S14.** Percent repetitive DNA of the three localized hypermutation regions around the three genes (accD, clpP, and ycf1) with accelerated substitution rates and the remaining plastome sequences. **Table S15.** Locations of the 75 individuals representing 20 Medicago, Trigonella, and Melilotus species. The individuals for which were planted in laboratory are marked by asterisks. Plastomes of *Medicago truncatula* were assembled from whole-genome resequencing data downloaded from NCBI (SRR1524305 and SRR965443). The outgroup was downloaded from NCBI (NC_011828.1). The individuals for which were chosen as the representatives of each species are marked in red. **Table S16.** The 73 protein-coding genes (CDS) shared across 21 taxa included in the phylogenetic analysis. **Table S17.** Taxa included in the synonymous and nonsynonymous divergence analyses of accD, clpP, ycf1, matK, and rbcL. (√) adopt in analysis, (–) not available in NCBI and not adopt in analysis.

Additional file 2 : **Figure S1.** Plastome maps for the 20 *Medicago*, *Trigonella*, and *Melilotus* species. Genes shown outside the circle are transcribed clockwise and those inside are transcribed counter clockwise. Genes belonging to different functional groups are color-coded. The dark gray area in the inner circle indicates GC content and the thick black line shows the extent of different regions. LSC: large single copy; SSC: small single copy; IRA: inverted repeat A; IRB: inverted repeat B. The two areas enclosed by red boxes in (d) indicate small IRs in *Medicago lupulina*. In (k), red arrows No. 1 and 2 outside the circle point to the breakpoints of the inverted region involving all genes from *ycf1* to *rpl20*. Gene order in the *M. truncatula* 02 ptDNA between the arrows is in the reverse orientation and its length is ~ 44-kb (44,228 bp). Below the map are shown the alignments of 24-bp incomplete inverted repeats in the inversion endpoints in the *M. truncatula* 02 and cognate sequences in *M. truncatula* 01. **Figure S2.** IRs alignment within the four *Melilotus dentata* individuals and the repetitive elements of indel regions in *Melilotus dentata* 02 and *Melilotus den*tata 03. Yellow bar represents protein coding gene; red bar represents rRNA gene; pink bar represents tRNA gene. Numerals above indicate nucleotide positions within the repeat alignment indicate the length of an indel within the IR. Mismatches are indicated by colored blocks and identical bases are gray. Mean pairwise identity over all pairs in each alignment column is indicated by the histogram: green 100%. IRA: inverted repeat A, IRB: inverted repeat B. Different repetitive elements were marked with different colored boxes. **Figure S3.** Mauve (Multiple Alignment of Conserved Genomic Sequence with Rearrangements) alignment of the plastomes of the 20 species using plastome of *Wisteria floribunda* as a reference. **Figure S4.** Confirmation of tRNA duplication in five species. Plastome sequences were mapped to themselves which contain all copies of the replicated tRNA (lower) and plastome sequences which contain a single copy of the replicated tRNA (upper). The scale at the left reports the depth of sequences, which is indicated graphically by the blue histogram. **Figure S5.** The degree of genomic rearrangements shows significant positive correlation with (a) dispersed repeats and (b) tandem repeats. **Figure S6.** The mechanism of tRNA duplication detected in five species. Both (a) and (b) contain a tandem repeat and two forward repeats that duplicate *trnV*-GAC twice. The three species in (c) contain a tandem repeat that duplicates the *trnN*-GUU gene. Thick black lines represent double stranded DNA. Pink boxes represent gene sequences, red and blue boxes represent tandem repeats, green boxes represent forward repeats and numbers indicate the length of repeats. **Figure S7.** Repeats around endpoints of all inversions. Red boxes represent regions of inversion, orange boxes represent protein-coding genes, and pink boxes represent tRNA genes. Black lines above double stranded DNA represent tandem repeats. Black lines below double stranded DNA represent dispersed repeats. The palindromic repeats are indicated by red lines. **Figure S8.** Repetitive DNA in the acquired introns. Yellow boxes represent exons. The lines between the yellow boxes represent acquired introns. The red content represents repetitive sequences. **Figure S9.** Comparison of the 20 plastomes using the annotation of *Medicago* *falcata* as a reference. The vertical scale indicates the percentage of identity, ranging from 50 to 100%. The horizontal axis indicates the coordinates within the plastomes. Genomic regions are color-coded as conserved non-coding sequences (CNS), exons, and tRNA or rRNA. **Figure S10.** Significance test for nonsynonymous (dN) (upper) and synonymous (dS) (lower) substitution rates of *clpP*, *accD*, and *ycf1* compared to *rbcL* for all the comparisons within both IRLC taxa we have found (left) and our own 20 species (right). ****, $P < 0.0001$ (T-test). Detailed information can be found in Tables S8–10 and Table S12. **Figure S11.** Synonymous and nonsynonymous divergence in the IRLC species for five chloroplast genes: *matK* (a), *rbcL* (b), *accD* (c), *clpP* (d), and *ycf1* (e). Shown are dN (left) and dS (right) trees resulting from a codon-based maximum likelihood (ML) analyses using RAxML, rooted using three Robinioid sequences (non-IRLC species): *Lotus japonicus*, *Sesbania grandiflora,* and *Robinia pseudoacacia*. The species are in the same order from top to bottom in the dN and dS trees of each gene.

Acknowledgements

The paper benefited greatly from the comments received from two anonymous reviewers. We would like to thank the support for computational work from the Big Data Computing Platform for Western Ecological Environment and Regional Development and Supercomputing Center of the Lanzhou University.

Authors' contributions

G.P.R. and S.W. conceived the project; J.Y.C. collected the samples and extracted the total genomic DNA; S.W. analyzed the data; Y.L., Ai.L., Ao.L., and M.Y. participated in discussions and provided valuable advice; S.W. and G.P.R. wrote the manuscript; N.S. and J.Q.L. revised the manuscript. All the authors read and approved the manuscript.

Funding

This work was supported equally by the Second Tibetan Plateau Scientific Expedition and Research (STEP) program (2019QZKK0502) to J.Q.L. and the National Natural Science Foundation of China (31971391), as well as by the National Natural Science Foundation of China (41901056) to G.P.R.

Availability of data and materials

The following data were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/): whole-genome resequencing data of *Medicago truncatula* (SRR1524305 and SRR965443); transcriptomic data of *M. falcata* (SRR1823822), *Melilotus albus* (SRR5115455) and *Trigonella foenum-graecum L.* (SRR8281660); plastome sequence of *Trifolium subterraneum* (NC_011828.1); *rbcL* chloroplast gene sequences of *Medicago polymorpha, M. truncatula, M. sativa, M. lupulina, M. minima, M. falcata, Melilotus albus* and *Trigonella foenum-graecum* (NC_042848.1, NC_003119.8, NC_042841.1, NC_042847.1, NC_042849.1, NC_032066.1, NC_041419.1, and NC_042857.1); coding sequences of *accD*, *clpP*, *ycf1*, *rbcL,* and *matK* of the IRLC species included in our study (taxon names and accession numbers are listed in Table S17). All the plastomes of the 75 individuals were newly assembled and deposited into the Genome Warehouse of CNCB-NGDC (under BioProject accession PRJCA005341). All of the parameters, program versions, and perl scripts in this study are available at https://github.com/ShuangWu888/plastomes_of_Medicago_and_its_relatives.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]State Key Laboratory of Grassland Agro-Ecosystems, Institute of Innovation Ecology & School of Life Sciences, Lanzhou University, Lanzhou, China. [2]Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education &State Key Lab of Hydraulics & Mountain River Engineering, College of Life Sciences, Sichuan University, Chengdu, China.

Received: 1 April 2021 Accepted: 31 August 2021
Published online: 14 September 2021

References

1. Jansen RK. Plastid genomes of seed plants: advances in photosynthesis and respiration 35. In: Genomics of chloroplasts and mitochondria; 2012.
2. Palmer JD. Chloroplast DNA exists in two orientations. Nature. 1983; 301(5895):92–3. https://doi.org/10.1038/301092a0.
3. Bock R. Structure, function, and inheritance of plastid genomes. In: Bock R, editor. Book: Structure, function, and inheritance of plastid genomes. In: Edited by. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p. 29–63. https://doi.org/10.1007/4735_2007_0223.
4. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, et al. The complete nucleotide-sequence of the tobacco chloroplast genome - its gene organization and expression. EMBO J. 1986;5(9):2043–9. https://doi.org/10.1002/j.1460-2075.1986.tb04464.x.
5. Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, et al. The complete chloroplast genome sequence of Pelargonium x hortorum: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol Biol Evol. 2006;23(11):2175–90. https://doi.org/10.1093/molbev/msl089.
6. Wicke S, Schneeweiss GM, de Pamphilis CW, Muller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 2011;76(3–5):273–97. https://doi.org/10.1007/s11103-011-9762-4.
7. Weng ML, Blazier JC, Govindu M, Jansen RK. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Mol Biol Evol. 2014;31(3):645–59. https://doi.org/10.1093/molbev/mst257.
8. Blazier JC, Jansen RK, Mower JP, Govindu M, Zhang J, Weng ML, et al. Variable presence of the inverted repeat and plastome stability in Erodium. Ann Bot-London. 2016;117(7):1209–20. https://doi.org/10.1093/aob/mcw065.
9. Ruhlman TA, Zhang J, Blazier JC, Sabir JSM, Jansen RK. Recombination-dependent replication and gene conversion homogenize repeat sequences and diversify plastid genome structure. Am J Bot. 2017;104(4):559–72. https://doi.org/10.3732/ajb.1600453.
10. Sanderson MJ, Copetti D, Burquez A, Bustamante E, Charboneau JLM, Eguiarte LE, et al. Exceptional reduction of the plastid genome of saguaro cactus (Carnegiea gigantea): loss of the ndh gene suite and inverted repeat. Am J Bot. 2015;102(7):1115–27. https://doi.org/10.3732/ajb.1500184.
11. Choi IS, Jansen R, Ruhlman T. Lost and found: return of the inverted repeat in the legume clade defined by its absence. Genome Biol Evol. 2019;11(4): 1321–33. https://doi.org/10.1093/gbe/evz076.
12. Jin DM, Wicke S, Gan L, Yang JB, Jin JJ, Yi TS. The loss of the inverted repeat in the putranjivoid clade of Malpighiales. Front Plant Sci. 2020;11. https://doi.org/10.3389/fpls.2020.00942.
13. Palmer JD, Thompson WF. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. Cell. 1982;29(2):537–50. https://doi.org/10.1016/0092-8674(82)90170-2.
14. Guisinger MM, Kuehl JV, Boore JL, Jansen RK. Extreme reconfiguration of plastid genomes in the angiosperm family geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol. 2011;28(1):583–600. https://doi.org/10.1093/molbev/msq229.
15. Hirao T, Watanabe A, Kurita M, Kondo T, Takata K. Complete nucleotide sequence of the Cryptomeria japonicia D. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. BMC Plant Biol. 2008;8:70. https://doi.org/10.1186/1471-2229-8-70.
16. Lee HL, Jansen RK, Chumley TW, Kim KJ. Gene relocations within chloroplast genomes of Jasminum and Menodora (Oleaceae) are due to multiple, overlapping inversions. Mol Biol Evol. 2007;24(5):1161–80. https://doi.org/10.1093/molbev/msm036.
17. Zhu AD, Guo WH, Gupta S, Fan WS, Mower JP. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytol. 2016;209(4):1747–56. https://doi.org/10.1111/nph.13743.
18. Wolfe KH, Li WH, Sharp PM. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear Dnas. P Natl Acad Sci USA. 1987;84(24):9054–8. https://doi.org/10.1073/pnas.84.24.9054.

19. Perry AS, Wolfe KH. Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. J Mol Evol. 2002;55(5): 501–8. https://doi.org/10.1007/s00239-002-2333-y.
20. Magee AM, Aspinall S, Rice DW, Cusack BP, Semon M, Perry AS, et al. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res. 2010;20(12):1700–10. https://doi.org/10.1101/gr.111955.110.
21. Mower JP, Touzet P, Gummow JS, Delph LF, Palmer JD. Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol Biol. 2007;7(1). https://doi.org/10.1186/1471-2148-7-135.
22. Cai ZQ, Guisinger M, Kim HG, 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. J Mol Evol. 2008;67(6):696–704. https://doi.org/10.1007/s00239-008-9180-7.
23. Sabir J, Schwarz E, Ellison N, Zhang J, Baeshen NA, Mutwakil M, et al. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. Plant Biotechnol J. 2014;12(6): 743–54. https://doi.org/10.1111/pbi.12179.
24. Dugas DV, Hernandez D, Koenen EJM, Schwarz E, Straub S, Hughes CE, et al. Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in clpP. Sci Rep-Uk. 2015;5(1). https://doi.org/10.1038/srep16958.
25. Shrestha B, Gilbert LE, Ruhlman TA, Jansen RK. Rampant nuclear transfer and substitutions of plastid genes in Passiflora. Genome Biol Evol. 2020;12(8): 1313–29. https://doi.org/10.1093/gbe/evaa123.
26. Cauz-Santos LA, da Costa ZP, Callot C, Cauet S, Zucchi MI, Berges H, et al. A repertory of rearrangements and the loss of an inverted repeat region in Passiflora chloroplast genomes. Genome Biol Evol. 2020;12(10):1841–57. https://doi.org/10.1093/gbe/evaa155.
27. Shrestha B, Weng ML, Theriot EC, Gilbert LE, Ruhman TA, Krosnick SE, et al. Highly accelerated rates of genomic rearrangements and nucleotide substitutions in plastid genomes of Passiflora subgenus Decaloba. Mol Phylogenet Evol. 2019;138:53–64. https://doi.org/10.1016/j.ympev.2019.05.030.
28. Rabah SO, Shrestha B, Hajrah NH, Sabir MJ, Alharby HF, Sabir MJ, et al. Passiflora plastome sequencing reveals widespread genomic rearrangements. J Syst Evol. 2019;57(1):1–14. https://doi.org/10.1111/jse.12425.
29. Goulding SE, Olmstead RG, Morden CW, Wolfe KH. Ebb and flow of the chloroplast inverted repeat. Mol Gen Genet. 1996;252(1–2):195–206. https://doi.org/10.1007/BF02173220.
30. Day A, Madesis P. DNA replication, recombination, and repair in plastids. In: Bock R, editor. Book: DNA replication, recombination, and repair in plastids. In: Edited by. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p. 65–119. https://doi.org/10.1007/4735_2007_0231.
31. Choi IS, Jansen R, Ruhlman T. Caught in the act: variation in plastid genome inverted repeat expansion within and between populations of Medicago minima. Ecol Evol. 2020;10(21):12129–37. https://doi.org/10.1002/ece3.6839.
32. Jansen RK, Wojciechowski MF, Sanniyasi E, Lee SB, Daniell H. Complete plastid genome sequence of the chickpea (Cicer arietinum) and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). Mol Phylogenet Evol. 2008;48(3):1204–17. https://doi.org/10.1016/j.ympev.2008.06.013.
33. Gurdon C, Maliga P. Two distinct plastid genome configurations and unprecedented intraspecies length variation in the accD coding region in Medicago truncatula. DNA Res. 2014;21(4):417–27. https://doi.org/10.1093/dnares/dsu007.
34. 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. Curr Genet. 1987;11(4):275–86. https://doi.org/10.1007/BF00355401.
35. Wojciechowski MF, Lavin M, Sanderson MJ. A phylogeny of legumes (Leguminosae) based on analyses of the plastid matK gene resolves many well-supported subclades within the family. Am J Bot. 2004;91(11):1846–62. https://doi.org/10.3732/ajb.91.11.1846.
36. Wu CS, Wang YN, Hsu CY, Lin CP, Chaw SM. 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;3:1284–95. https://doi.org/10.1093/gbe/evr095.
37. Ruhlman TA, Jansen RK. The plastid genomes of flowering plants. Methods Mol Biol. 2014;1132:3–38. https://doi.org/10.1007/978-1-62703-995-6_1.
38. Schwarz EN, Ruhlman TA, Weng ML, Khiyami MA, Sabir JSM, Hajarah NH, et al. Plastome-wide nucleotide substitution rates reveal accelerated rates in Papilionoideae and correlations with genome features across legume subfamilies. J Mol Evol. 2017;84(4):187–203. https://doi.org/10.1007/s00239-017-9792-x.
39. Moghaddam M, Kazempour-Osaloo S. Extensive survey of the ycf4 plastid gene throughout the IRLC legumes: Robust evidence of its locus and lineage specific accelerated rate of evolution, pseudogenization and gene loss in the tribe Fabeae. Plos One. 2020;15(3):e0229846. https://doi.org/10.1371/journal.pone.0229846.
40. Kode V, Mudd EA, Iamtham S, Day A. The tobacco plastid accD gene is essential and is required for leaf development. Plant J. 2005;44(2):237–44. https://doi.org/10.1111/j.1365-313X.2005.02533.x.
41. Peltier JB, Ripoll DR, Friso G, Rudella A, Cai Y, Ytterberg J, et al. Clp protease complexes from photosynthetic and non-photosynthetic plastids and mitochondria of plants, their predicted three-dimensional structures, and functional implications. J Biol Chem. 2004;279(6):4768–81. https://doi.org/10.1074/jbc.M309212200.
42. Kikuchi S, Bedard J, Hirano M, Hirabayashi Y, Oishi M, Imai M, et al. Uncovering the protein translocon at the chloroplast inner envelope membrane. Science. 2013;339(6119):571–4. https://doi.org/10.1126/science.1229262.
43. Sveinsson S, Cronk Q. Evolutionary origin of highly repetitive plastid genomes within the clover genus (Trifolium). BMC Evol Biol. 2014;14(1):228. https://doi.org/10.1186/s12862-014-0228-6.
44. Schwarz EN, Ruhlman TA, Sabir JSM, Hajrah NH, Alharbi NS, Al-Malki AL, et al. Plastid genome sequences of legumes reveal parallel inversions and multiple losses of rps16 in papilionoids. J Syst Evol. 2015;53(5):458–68. https://doi.org/10.1111/jse.12179.
45. Gantt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD. Transfer of Rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. EMBO J. 1991;10(10):3073–8. https://doi.org/10.1002/j.1460-2075.1991.tb07859.x.
46. 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. Plant Cell. 2001;13(3): 645–58. https://doi.org/10.1105/tpc.13.3.645.
47. Ueda M, Nishikawa T, Fujimoto M, Takanashi H, Arimura S, Tsutsumi N, et al. Substitution of the gene for chloroplast RPS16 was assisted by generation of a dual targeting signal. Mol Biol Evol. 2008;25(8):1566–75. https://doi.org/10.1093/molbev/msn102.
48. Sinn BT, Sedmak DD, Kelly LM, Freudenstein JV. Total duplication of the small single copy region in the angiosperm plastome: rearrangement and inverted repeat instability in Asarum. Am J Bot. 2018;105(1):71–84. https://doi.org/10.1002/ajb2.1001.
49. Ansell SW, Schneider H, Pedersen N, Grundmann M, Russell SJ, Vogel JC. Recombination diversifies chloroplast trnF pseudogenes in Arabidopsis lyrata. J Evolution Biol. 2007;20(6):2400–11. https://doi.org/10.1111/j.1420-9101.2007.01397.x.
50. Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. BMC Evol Biol. 2008;8(1). https://doi.org/10.1186/1471-2148-8-36.
51. Ogihara Y, Ohsawa T. Molecular analysis of the complete set of length mutations found in the plastomes of Triticum-Aegilops species. Genome. 2002;45(5):956–62. https://doi.org/10.1139/g02-046.
52. Marechal A, Parent JS, Veronneau-Lafortune F, Joyeux A, Lang BF, Brisson N. Whirly proteins maintain plastid genome stability in Arabidopsis. P Natl Acad Sci USA. 2009; 106(34):14693–8. https://doi.org/10.1073/pnas.0901710106.
53. Kwon T, Huq E, Herrin DL. Microhomology-mediated and nonhomologous repair of a double-strand break in the chloroplast genome of Arabidopsis. P Natl Acad Sci USA. 2010;107(31):13954–9. https://doi.org/10.1073/pnas.1004326107.
54. Chen JY, Wu GL, Shrestha N, Wu S, Guo W, Yin M, et al. Phylogeny and species delimitation of Chinese Medicago (Leguminosae) and its relatives based on molecular and morphological evidence. Front Plant Sci. 2021;11. https://doi.org/10.3389/fpls.2020.619799.
55. Doyle JJ. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 1987;19:11-15.
56. Andrews S. FastQC: a quality control tool for high throughput sequence data; 2010.

57. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 2017;45(4): e18. https://doi.org/10.1093/nar/gkw955.
58. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45(W1):W6–W11. https://doi.org/10.1093/nar/gkx391.
59. Lowe TM, Chan PP. tRNAscan-SE on-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. 2016;44(W1):W54–7. https://doi.org/10.1093/nar/gkw413.
60. 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): 1647–9. https://doi.org/10.1093/bioinformatics/bts199.
61. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–U130. https://doi.org/10.1038/nbt.1883.
62. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004;32(Web Server):W20–5. https://doi.org/10.1093/nar/gkh435.
63. Lohse M, Drechsel O, Kahlau S, Bock R. OrganellarGenomeDRAW-a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 2013;41(W1):W575–81. https://doi.org/10.1093/nar/gkt289.
64. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. Plos One. 2010;5(6): e11147. https://doi.org/10.1371/journal.pone.0011147.
65. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80. https://doi.org/10.1093/nar/27.2.573.
66. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29(22):4633–42. https://doi.org/10.1093/nar/29.22.4633.
67. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004;32(Web Server): W273–9. https://doi.org/10.1093/nar/gkh458.
68. Zhang R, Wang YH, Jin JJ, Stull GW, Bruneau A, Cardoso D, et al. Exploration of plastid phylogenomic conflict yields new insights into the deep relationships of Leguminosae. Syst Biol. 2020;69(4):613–22. https://doi.org/10.1093/sysbio/syaa013.
69. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4):772–80. https://doi.org/10.1093/molbev/mst010.
70. Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol. 2017;34(12):3299–302. https://doi.org/10.1093/molbev/msx248.
71. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and clustal X version 2.0. Bioinformatics. 2007;23(21):2947–8. https://doi.org/10.1093/bioinformatics/btm404.
72. Guindon S, Dufayard JF, Hordijk W, Lefort V, Gascuel O. PhyML: fast and accurate phylogeny reconstruction by maximum likelihood. Infect Genet Evol. 2009;9(3):384–5.
73. Posada D. jModelTest: phylogenetic model averaging. Mol Biol Evol. 2008; 25(7):1253–6. https://doi.org/10.1093/molbev/msn083.
74. Yang ZH. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91. https://doi.org/10.1093/molbev/msm088.
75. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3. https://doi.org/10.1093/bioinformatics/btu033.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC