Plastome characteristics of Cannabaceae

Huanlei Zhang, Jianjun Jin, Michael J. Moore, Tingshuang Yi, Dezhu Li

Article Info

Article history:
Received 17 March 2018
Accepted 18 April 2018
Available online 23 April 2018

Keywords:
Plastome
IR expansion/contraction
Repeats
SSR
Sequence divergence
Phylogenomics

Abstract

Cannabaceae is an economically important family that includes ten genera and ca. 117 accepted species. To explore the structure and size variation of their plastomes, we sequenced ten plastomes representing all ten genera of Cannabaceae. Each plastome possessed the typical angiosperm quadripartite structure and contained a total of 128 genes. The Inverted Repeat (IR) regions in five plastomes had experienced small expansions (330–983 bp) into the Large Single-Copy (LSC) region. The plastome of Chaetachme aristata has experienced a 942-bp IR contraction and lost rpl22 and rps19 in its IRs. The substitution rates of rps19 and rpl22 decreased after they shifted from the LSC to IR. A 270-bp inversion was detected in the Parasponia rugosa plastome, which might have been mediated by 18-bp inverted repeats. Repeat sequences, simple sequence repeats, and nucleotide substitution rates varied among these plastomes. Molecular markers with more than 13% variable sites and 5% parsimony-informative sites were identified, which may be useful for further phylogenetic analysis and species identification. Our results show strong support for a sister relationship between Gironniera and Lozannell (BS = 100). Celtis, Cannabis-Humulus, Chaetachme-Pteroceltis, and Trema-Parasponia formed a strongly supported clade, and their relationships were well resolved with strong support (BS = 100). The availability of these ten plastomes provides valuable genetic information for accurately identifying species, clarifying taxonomy and reconstructing the intergeneric phylogeny of Cannabaceae.

1. Introduction

Cannabaceae sensu APG IV (Byng et al., 2016) comprise ten genera (Lipton, 1997; Sytsma et al., 2002; Haston et al., 2007, 2009; Mabberley, 2008; Bell et al., 2010) and ca. 117 species (Jin et al., unpublished). Most Cannabaceae species are trees and shrubs, while some are herbs (Cannabis L.) or vines (Humulus L.). The family has a cosmopolitan distribution; Aphananthe (Thunb.) Planch., Celtis L. and Trema Lour. are widely distributed in tropical and temperate regions (Yang et al., 2013; Jin et al., unpublished); the remaining genera have restricted distributions. A few species of this family are of great economic importance. Cannabis sativa L. (hemp) is one of earliest and most important domesticated food and fiber crops, and an increasingly important drug used for its anesthetic and antipsychotic properties (Measham et al., 1994; Kostic et al., 2008; Marks et al., 2009). Humulus lupulus L. (hops) is a key ingredient for brewing beer (Wilson, 1975; Murakami et al., 2006), and the phloem fiber of Pteroceltis tatarinowii Maxim. is the sole raw material for manufacturing traditional Chinese Xuan paper (Cao, 1993).

There are long-standing controversies over the circumscription and phylogenetic position of Cannabaceae. Cannabaceae was first separated from Moraceae by Rendle (1925). The circumscription of this family has been expanded significantly to include most former members of Ulmaceae subfam. Celtidoideae sensu Engler and Prantl (1893) or Celtidaceae sensu Link (1829) (Yang et al., 2013). A series of molecular studies elucidated the phylogenetic position of this family, which was supported to be a member of Rosales and sister to Moraceae and Urticaceae (Sytsma et al., 2002; Van Velzen et al., 2006; Wang et al., 2009; Zhang et al., 2011a,b). Multiple molecular studies have also helped to clarify intergeneric relationships of the family (Yang et al., 2013; Jin et al., unpublished). However, a few
nodes among genera have remained unresolved with weak support (Yang et al., 2013).

The plastome of angiosperms is usually conserved in gene content and structure, typically featuring two ~25 kb Inverted Repeat (IR) regions separating the remainder of the genome into Large and Small Single-Copy regions (LSC, SSC). Size variation among plastomes is mostly due to the expansion or contraction of the IR and/or larger indels, as for example caused by the loss of genes (especially the ndh genes) (Downie and Jansen, 2015). Plastomes have proved highly valuable in resolving difficult phylogenetic relationships at both deeper taxonomic levels (e.g. Jansen et al., 2007; Moore et al., 2007, 2010), as well as at more shallow levels (e.g. Zhang et al., 2011a; Givnish et al., 2015; Wysocki et al., 2015; Duvaill et al., 2016).

In this article, we report the complete plastome sequences of ten species representing all ten genera of Cannabaceae. We annotated the plastomes in detail, identified structure and size variation, and determined the distribution and location of microsatellites (SSRs) and repeats. We demonstrate that the resulting plastome information will be widely useful for understanding phylogenetic relationships, population genetics and breeding programs across the family.

2. Materials and methods

2.1. Chloroplast DNA extraction and sequencing

We used about 100 mg of fresh leaf material of each species (see Table S1 for voucher specimens). Total genomic DNA was extracted with a modified CTAB (Cetyl Trimethyl Ammonium Bromide) method (Doyle and Doyle, 1987), in which 4% CTAB with approximately 1% polyvinyl polypyrrolidone (PVP) and 0.2% Dithiothreitol (DTT) was included (Yang et al., 2014). Long-range polymerase chain reaction (PCR) was used for DNA amplification of the plastome using 15 universal primers pairs and methods described by Zhang et al. (2016). Illumina Nextera XT libraries (Illumina, San Diego, CA, USA) with 500 bp inserts were constructed following the manufacturer’s protocol. Paired-end (PE) sequencing was performed on an Illumina Hiseq 2500 instrument at the Beijing Genomics Institute (BGI, Shenzhen, Guangdong, China) or on a Hiseq 2000 instrument at the Plant Germplasm and Genomics Center (Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China).

2.2. Plastome assembly and annotation

Raw reads were filtered using NGSQC Toolkit (Patel and Jain, 2012; cut-off value for percentage of read length = 80, cut-off value for PHRED quality score = 30) to obtain high quality reads that were free of vector and adaptor sequences. Filtered reads were then assembled into contigs using the software CLC Genomics Workbench 8, via the de novo method using a k-mer of 63 and a minimum contig length of 1 kb. Using BLAST (Altschul et al., 1990) with default search parameters, all contigs were aligned to the Morus mongolica Schneid. plastome (NC025772.2) as a reference. We mapped the paired reads to the assembled plastomes using Bowtie 2 (Langmead and Salzberg, 2012), as implemented in Geneious v9.5 (Kearse et al., 2012), to verify the IR boundaries, correct some biased bases brought in by the CLC assembler, and detect the number of matched paired-end (PE) reads and the depth of coverage. Lastly, we filled the remaining gaps using long-range PCR and Sanger sequencing. We designed primers based on previous incomplete plastomes (Table S2). Each amplification was performed in 25 μL reaction volume containing 12.5 μL Taq DNA polymerase, 0.5 μL each of forward and reverse primers (dissolved in 10× ddH2O), and 1 μL (30 ng/μL) template DNA. The amplification was conducted using 94 °C for 3 min, 35 cycles of 94 °C for 50 s, 50 °C for 2 min, and 72 °C for 1 min, followed by a final extension step at 72 °C for 8 min. PCR products were sequenced at the Kunming Sequencing Department of Biosune Biotechnology Limited Company (Shanghai, China).

Assembled genomes were annotated using DOGMA (Wyman et al., 2004) along with manual correction of start and stop codons and intron/exon boundaries in Geneious. Transfer RNA (tRNA) genes were further annotated using tRNAscAn-SE (Schattner et al., 2005). Genome maps were created in OGDraw 1.2 (Lohse et al., 2013). All annotated plastomes were deposited in GenBank; accession numbers are MH118117–MH118121 that provided in Table S1.

2.3. Phylogenetic analysis

Phylogenetic analyses included all ten genera of Cannabaceae as ingroups, two species of M. mongolica (Moraceae) and Ulmus macrocarpa (Ulmaceae) representing closely related families as outgroups (Table S1). A total of 237 loci (112 coding and 125 noncoding regions) were extracted from each plastome (exons were joined) for phylogenetic analysis. Loci shared by less than 6 taxa or with length <30 bp were excluded (Table S3). Sequences were aligned using MAFFT version 7 (Katoh and Standley, 2013) with default parameters. Maximum likelihood analysis was performed with RAxMLv8.2.10 (Stamatakis, 2006), by using the ‘-f a’ option, GTR+GAMMA model, and 1000 bootstrap replicates, with data partitioned by locus.

2.4. Analysis of sequence divergence

To characterize sequence divergence among all sequenced plastomes of Cannabaceae, we extracted 133 coding and 129 non-coding regions (including intergenic spacers and introns), each of them treated as a separate locus. These regions were aligned using MEGA v6.06 (Tamura et al., 2013). For each alignment, the number of invariant sites, variable but parsimony-uninformative sites, and parsimony-informative sites were calculated, as was pairwise sequence divergence (uncorrected “p” distance), all using PAUP* 4.0a147 (Swoford, 2002). Gaps were treated as missing data. Using the Humulus scandens plastome as a reference, sequence identity was also plotted using mVISTA (Frazer et al., 2004) in LAGAN mode.

2.5. Repeat analysis

REPutter (Kurtz et al., 2001) was used to locate sequence repeats including forward, reverse, and palindromic repeats. The minimal repeat size was set to 30 bp and repeat identity was set to ≥90% (hamming distance equal to 3). Before using REPutter to detect repeats, to avoid redundancy we removed the IR region from each plastome. However, IR repeats were treated twice (to represent both copies) when summarizing repeats across the genome. Tandem repeats were analyzed using TRF (Tandem Repeat Finder program) web interface (Benson, 1999) with the parameters setting as 2, 7 and 7 for match, mismatch and indel respectively. The minimum alignment score and maximum period size were set as 50 and 500. After analysis, tandem repeats <15 bp in length and the redundant results of REPutter were manually removed (Wang et al., 2017). We also tallied the total number of repeats, measured repeat lengths, and calculated the proportion of repeats in the LSC, SSC, and IR.
2.6. SSR analysis

Microsatellite detection was performed using MISA with minimum number of repeats of 8, 5, 4, 3, 3, and 3 respectively for mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats. One copy of the IR was removed prior to microsatellite detection. All of the repeats were manually verified, and redundant results were removed.

3. Results and discussion

3.1. Conservation of Cannabaceae plastomes

Illumina sequencing produced from 289,464 (Celtis blondii) to 4,807,452 (Trema orientalis) paired-end reads, among which 257,965 (Celtis blondii) to 4,346,229 (T. orientalis) reads were mapped to their respective assembled genomes. De novo and reference-guided assembly produced full coverage for all plastomes, with mean coverages ranging from 120.3 \( \times \) (Celtis blondii) to 2569.3 \( \times \) (T. orientalis) (Table 1).

All sequenced plastomes displayed the typical quadripartite structure of most angiosperms (Wang et al., 2013; Li et al., 2014). The ten plastomes ranged in size from 153,776 bp (H. scandens) to 159,001 bp (Celtis blondii). The length of their LSC region varied from 83,885 bp (H. scandens) to 87,620 bp (P. tatarinowii), that of the SSC region from 17,751 bp (H. scandens) to 20,064 bp (Chaetachne aristata), and their IR region from 25,512 bp (T. orientalis) to 26,879 bp (Celtis blondii) (Table 1). The overall GC content was approximately 37.3% across all ten sampled plastomes. The gene content and structural organization of all ten sequenced plastomes were also highly conserved (Fig. 1, Fig. S1). Most plastomes harbored 112 unique genes, including 78 protein-coding genes, 30 transfer RNA (tRNA) genes, and four ribosomal RNA (rRNA) genes. The exceptions were the plastomes of Celtis blondii and C. aristata; the former had a pseudogenic rpl22 and the latter lost rpl22 (Table 2). All plastomes lost infA, which was consistent with those of most euroids (Millen et al., 2001).

The IR, LSC, and SSC gene content, as well as intron content, for most of the Cannabaceae plastomes matched the typical content for angiosperms, with some differences in IR gene content (Fig. 2, Table S4). The plastomes of Aphananthe aspera, Lozannella enantiophylla, Parasponia rugosa and T. orientalis possessed canonical IRs ranging from 25,512 bp in T. orientalis to 26,015 bp in A. aspera. Their IRs contained 17 complete genes (including six protein-coding genes, seven tRNAs, and all four rRNAs) as well as the 5’ ends of ycf1 (1037–1076 bp) and rps19 (0–100 bp). The plastomes of C. sativa, H. scandens, P. tatarinowii, Celtis blondii and Gironniera subaequalis had longer IRs, ranging from 26,011 bp (C. sativa) to 26,879 bp (Celtis blondii), caused by 330-bp (C. sativa) to 983-bp (Celtis blondii) IR expansions into the LSC; specifically, IRs expanded into all of rps19 and all or part of rpl22 (25–408 bp). In contrast, C. aristata had the shortest IR at 25,566 bp, due to a 942-bp IR contraction. Its IRs lost rps19 and rpl22, but rps19 was found before trnH–GUG in LSC near the IRA/LSC junction (Fig. S1). IRs of C. aristata may have experienced more than a 942-bp IR expansion into LSC firstly to include rps19 and rpl22, followed by the loss of rps19 (279 bp) and rpl22 (408 bp) from IRb and rpl22 from IRA. In contrast, the IR/SC junctions showed little variation, including 0 (A. aspera) to 45 bp (L. enantiophylla) of the 3’ end of ndhF.

IR expansion and contraction are common, especially small contractions and expansions of <100 base pairs (bp), and the positions of four IR/single-copy junctions can vary even among closely related species (Goulding et al., 1996; Plunkett and Downie, 2000). Large IR expansions occur less frequently and sometimes accompany structural rearrangements elsewhere in the plastid genome (Guisinger et al., 2011; Wicke et al., 2011). Cannabaceae plastomes provide yet another example of moderate to small IR expansion and contraction. IR expansion has been suggested to start with double-strand breaks followed by strand invasion and recombination (Goulding et al., 1996; Wang et al., 2008). Regions with a high content of short repeats or “poly A” tracts were inferred to be associated with the dynamics of IR-LSC junctions and expansions of IR (Wang et al., 2008; Dugas et al., 2015). In Cannabaceae plastomes with expanded IRs, a region ca. 100 bp upstream of the IR-LSC junctions was found to be extremely AT-rich (>90%), including many poly A tracts and short repeats, which could explain the IR expansion of Cannabaceae plastomes. Large IR contractions have been rarely reported, and illegitimate recombination has been considered as the most plausible explanation (Goulding et al., 1996; Downie and Jansen, 2015; Blazier et al., 2016), which may also account for the IR contraction in C. aristata.

Nucleotide substitution rates of most plastome coding genes have been demonstrated to decrease after translocation from SC regions to the IR (Lin et al., 2012; Li et al., 2016; Zhu et al., 2016; but see exceptions in Lin et al., 2012; Wang et al., 2017). In this study, we also found a decrease of substitution rates for rps19 (0.0154) and rpl22 (0.0229) after their shifts from LSC into IR.

Finally, an interesting 270-bp inversion between petN and psbM was detected in the plastome of P. rugosa, representing the first known reasonably long inversion in Cannabaceae plastomes. A pair of 18-bp inverted repeats resided at the boundaries of this inversion, and it is likely that these repeats helped mediate this inversion, as seen for other smaller inversions (Kim et al., 2005; Qu et al., 2017a,b). Likewise, short repeats have also been inferred to associated with large inversions, such as the association of 29-kb repeats with a 36-kb inversion in legumes (Martin et al., 2014); the association >20-bp repeats with a 45-kb inversion of Medicago truncatula (Gurdon and Maliga, 2014); and the association of 11-bp repeats with a 36-kb inversion in Calocedrus macrolepis (Qu et al., 2017a,b).

| Species               | Total PE reads | Matched PE reads | Mean coverage \( \times \) | Genome length (bp) | LSC length (bp) | SSC length (bp) | IR length (bp) | GC content (%) |
|----------------------|----------------|------------------|-----------------------------|--------------------|----------------|----------------|----------------|----------------|
| Aphananthe aspera    | 1,695,716      | 374,611          | 583.7                       | 157,687            | 86,135         | 19,442         | 26,015         | 36.4           |
| Cannabis sativa      | 2,040,500      | 1,880,700        | 1351.8                      | 153,910            | 84,059         | 17,829         | 26,011         | 36.7           |
| Celtis blondii       | 289,464        | 295,965          | 120.3                       | 159,001            | 86,072         | 19,171         | 26,879         | 36.3           |
| Chaetachne aristata  | 1,142,608      | 1,045,891        | 1415.4                      | 157,939            | 86,743         | 20,064         | 25,566         | 36.1           |
| Gironniera subaequalis| 396,352        | 374,583          | 583.6                       | 157,807            | 86,215         | 18,942         | 26,325         | 36.3           |
| Humulus scandens      | 1,010,646      | 839,251          | 1436.6                      | 153,776            | 83,885         | 17,751         | 26,070         | 36.9           |
| Lozannella enantiophylla| 1,077,002      | 1,026,115        | 1573.4                      | 156,711            | 85,928         | 19,133         | 25,825         | 36.6           |
| Parasponia rugosa    | 586,024        | 498,328          | 627.5                       | 157,434            | 86,961         | 19,313         | 25,580         | 36.3           |
| Pieroceltis tatarinowii| 1,051,432      | 992,380          | 1711.1                      | 158,904            | 87,620         | 18,856         | 26,014         | 36.1           |
| Trema orientalis     | 4,807,452      | 4,346,229        | 2569.3                      | 157,192            | 86,859         | 19,309         | 25,512         | 36.3           |

PE = paired-end; LSC = Large Single-Copy region; SSC = Small Single-Copy region; IR = Inverted Repeat region.
3.2. Phylogenetic relationships

The monophyly of Cannabaceae was strongly supported (BS = 100). Relationships among the ten genera of Cannabaceae were also fully resolved with high bootstrap support (BS) (Fig. 3). Complete plastome sequences have also been used to successfully resolve intergeneric relationships in many other vascular plants (e.g., Givnish et al., 2015; Qu et al., 2017a,b; Zhang et al., 2017; Wang et al., 2018), and our study provides yet another example. Some previously resolved intrafamilial relationships were strongly supported in this study (Fig. 3): Aphananthe was sister to other genera of Cannabaceae (Song et al., 2001; Sytsma et al., 2002; Van Velzen et al., 2006; Yang et al., 2013); Gironniera, Lozanella and the clade B together formed a monophyletic group (Yang et al., 2013); Chae- tachne and Pteroceltis were sisters (Van Velzen et al., 2006; Yang et al., 2013); Cannabis and Humulus were sisters (Song et al., 2001; Song and Li, 2002; Sytsma et al., 2002); Parasponia was nested within Trema (Zavada and Kim, 1996; Sytsma et al., 2002; Yesson et al., 2004; Van Velzen et al., 2006; Yang et al., 2013). However, our study supported some new relationships. Our results show strong support (BS = 100) for a sister relationship between Gironniera and Lozanella. Celtis was strongly supported to be sister

**Fig. 1. Gene maps of the plastome of Humulus scandens.** Genes are indicated by boxes on the inside (clockwise transcription) and outside (counterclockwise transcription) of the outermost circle. The inner circle identifies the major structural components of the plastome (LSC, IR, and SSC). Genes belonging to different functional groups are color-coded. Dashed area in the inner circle indicates the GC content of the plastome. * represents the tRNA with an intron.
| Category | Gene groups | Name of genes |
|----------|-------------|---------------|
| Self-replication | Large subunit of ribosomal proteins | rpl2b (×2), rpl14, rpl16b, rpl20, rpl22 (×2)², rpl23 (×2), rpl32, rpl33, rpl36 |
| | Small subunit of ribosomal proteins | rps2, rps3, rps4, rps7 (×2), rps8, rps11, rps12, rps13, rps14, rps15, rps16b, rps18, rps19 (×2)³ |
| | DNA-dependent RNA polymerase | rpoA, rpoB, rpoC1, rpoC2 |
| | Ribosomal RNA genes | rm4.5 (×2), rm5 (×2), rm16 (×2), rm23 (×2) |
| | Transfer RNA genes | trnA-UGC (×2)², trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnfM-CAU, trnG-GCC, trnG-GUG, trnH-GUG, trnI-CAU (×2)², trnL-CAU (×2), trnL-CAU (×2)², trnK-UGP, trnL-CAU (×2), trnL-UGA, trnL-UGA, trnN-CU2 (×2), trnT-UUA, trnT-UUA, trnR-UGC, trnR-UGC, trnW-CGA, trnW-GUA |
| Photosynthesis | Photosystem I | psaA, psaB, psaC, psaD, psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbO, psbZ |
| | Photosystem II | psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbO, psbZ |
| | NADH dehydrogenase | ndhA, ndhB² (×2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK |
| | Cytochrome b/f complex | petA, petB, petD², petG, petL, petN |
| | ATP synthase | atpA, atpB, atpE, atpF, atpH, atpI |
| | Rubisco large subunit | rbcL |
| | Envelope membrane proteins | cemA |
| | Subunit of acetyl-CoA carboxylase | accD |
| | c-type cytochrome synthesis gene | ccsA |
| | Protease | clpP |
| | Proteins of unknown function | ycf1, ycf2 (×2), ycf3², ycf4 |

(×2) — gene present twice due to position within the IR; ² Contains two introns; ³ Contains one intron; ⁴ Exons separated and joined by trans-splicing; ⁵ gene present in the IRs in the IR-expanded species; ⁶ Gene present in the IR of Celtis blondii; ⁷ Gene present in the IR of Chaetachme aristata.

**Fig. 2. Comparison of IR/SC boundaries among Cannabaceae plastomes.** JSB, JSA and JLA refer to junctions of SSC/IRB, SSC/IRA, and LSC/IRA, respectively. ² indicates a pseudogene copy of a gene partially duplicated in the IR.
of clade A (BS = 100). The *Humulus*-Cannabis clade and the Trema-Parasponia clade were sisters with strong support (BS = 100). Morphologically, they all have persistent tepals and stigmas. The *Chaeatohme*-Pteroceltis clade was sister to the *Humulus*-Cannabis-Trema-Parasponia with relatively low support (BS = 80).

### 3.3. Sequence divergence and phylogenetic informativeness

Sequence alignments and the mVISTA plot (Fig. 4) revealed high sequence similarity among Cannabaceae plastomes. Aligned lengths of 133 coding and 129 noncoding regions ranged from 9 bp (psbF-psbE intergenic spacer) to 6828 bp (ycf2). The number of variable sites ranged from 0 (for 20 loci) to 943 (ycf1). The number of parsimony-informative sites ranged from 0 (for 26 loci) to 392 (ycf1). Percentages of variable and parsimony-informative sites in coding and noncoding regions are provided in Fig. 5A and Table S5. Among coding regions, *matK*, *rps8*, *rpl22*, *ndhF* and *ycf1* had the highest percentages of variable and parsimony-informative sites, with *matK* having an especially high percentage of variable sites (14.05%) and *rpl22* having a high percentage of parsimony-informative sites (6.70%). The percentages of variable sites in noncoding regions ranged from 0 to 28.93% with a mean value of 9.43%, which was nearly twice that of coding regions (5.24% on average). The five noncoding regions with highest percentages of variable sites were trnM-CAU-rps14, psal-ycf4, petD-2-rpoA, rpl36-rps8 and rps15-ycf1, with rpl36-rps8 having the highest percentage of variable (28.93%) and parsimony-informative sites (10.85%). The five noncoding regions with highest percentage of parsimony-informative sites in noncoding regions ranged from 0 to 10.85% with a mean value of 2.99%, which was higher than that of the coding regions (2.19% on average). In IRs, both of the percentages of variable sites and informative sites ranged from 0 to 2.78% with a mean value of 0.88% in coding regions. Among noncoding regions, the percentages of variable sites ranged from 0 to 6.93% with a mean value of 2.65%, which was similar low to the percentages of PIS (0—2.97% and mean of 1.00%). These findings all showed that fewer mutations were observed within IR regions, including coding and non-coding regions, than LSC and SSC regions. Those with no mutations were mostly tRNAs and rrn5, illustrating that tRNAs are more conserved than other genes.

Plastomes supply many valuable loci for reconstructing phylogenetic relationships at multiple taxonomic scales. A number of plastid coding and noncoding loci have been used in phylogenetic studies among genera in the same family, including for example *atpB*, *atpB-rbcL*, *matK*, *ndhF*, *rbcL*, *rpl16*, *rps4-trnS*, *rps16*, *trnH-psbA*, *trnL-F*, and *trnS-G* (Kim and Jansen, 1995; Gao et al., 2008; Hilu et al., 2008; Wilson, 2009; Peterson et al., 2010). Some plastome regions, such as *atpF-H*, *matK*, *psbK-I*, *rbcl*, *rpoB*, *rpoC1*, *trnH-psbA*, etc., have been relied upon heavily for development of candidate markers for plant DNA barcoding (Kress et al., 2005; Newmaster et al., 2006; Chase et al., 2007; Hollingsworth et al., 2011; Dong et al., 2012). The fast-evolving loci we identified, such as *rpl36-rps8*, *rpl22*, *rpl33-rps18*, *rps15-ycf1*, *matK* and *rps8* could be applied to resolve inter- or intraspecific relationships.

### 3.4. Repetitive sequences

Repeat regions are thought to play an important role in genome recombination and rearrangement (Smith, 2002). In this study, a total of 431 repeats were detected across all Cannabaceae plastomes, including 116 dispersed repeats and 314 tandem repeats (Table S6). Among all ten plastomes, *T. orientalis* had the most repeats (56) and *C. sativa* had the fewest (29). After excluding overlapped repeats detected by REPuter and accounting for both IR copies, 7 (G. subaequalis) – 19 (C. aristata) pairs of dispersed repeats were identified. Plastomes of *C. aristata*, *P. rugosa*, and *T. orientalis* had three repeat types—direct, reverse and palindromic repeats (Fig. 6). Among these, 61% were direct, 33% were palindromic and 6% were reverse. The lengths of repeats ranged from 30 to 55 bp. The total length of dispersed repeats ranged from 541 (G. subaequalis) to 1229 bp (C. aristata), and their proportion of the whole plastome ranged from 0.34% (G. subaequalis) to 0.77% (C. aristata). We detected 20 (C. sativa)—42 (T. orientalis) tandem repeats with a size ≥ 15 bp, of which 184 were 15–20 bp in size, 112 were 21–30 bp, 13 were 31–40 bp, four were 41–50 bp, and one was 61 bp (in A. aspera). The total length of tandem repeats ranged from...
Fig. 4. mVISTA-based identity plot showing sequence identity among Cannabaceae plastomes. *Humulus scandens* is set as the reference. Coding and noncoding regions are colored in blue and red, respectively.
Fig. 5. Percentages of variable (blue, top line) and parsimony-informative (red, bottom line) sites across coding and non-coding loci. A coding regions; B noncoding regions. Regions are oriented according to their genome locations.

Fig. 6. Analyses of repeated sequences in Cannabaceae plastomes. A Numbers of the three dispersed repeat types; B Numbers of tandem repeats; C Frequency of dispersed repeats by length; D Frequency of tandem repeats by length; E The locations of repeats.
3.5. Simple sequence repeat (SSR) polymorphisms

SSRs, including mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats, were detected in all plastomes, although hexanucleotide repeats were absent from the plastomes of Celtis blondii, H. scandens, and P. rugosa. (see Table S7 for a comprehensive list of SSRs, including their positions within the plastome). In total, 221, 186, 193, 229, 210, 172, 195, 250, 209 and 228 SSRs were found in the plastomes of Aphananthe spera, C. sativa, Celtis blondii, C. aristata, G. subaequalis, H. scandens, L. enantiophylla, P. rugosa, P. tatarinowii and T. orientalis, respectively. The majority of mono-nucleotide repeat units were A/T, ranging from 8 to 23 bp in length (Fig. 7; the longest was present in T. orientalis). This finding is consistent with previous observations that cpSSRs are dominated by A/T mononucleotide repeats (Kuang et al., 2011). SSR loci were mainly located within intergenic spacers, followed by coding sequences and introns. Most SSRs were located in the LSC region, followed by the IR and SSC regions. SSRs have been used to understand evolutionary relationships among some closely related plant taxa, and are also effective genetic markers for studying plant breeding, population genetics, biological conservation, mating systems, and uniparental lineages (Terrab et al., 2006; Cardle et al., 2000; Peakall et al., 1998). The SSRs characterized in this study may prove useful for understanding phyleogeography and genetic structure of populations.

4. Conclusion

We reported ten complete plastomes in Cannabaceae using Illumina sequencing technology via a combination of de novo and reference-guided assembly. These plastomes were relatively conserved, but the IR regions in some plastomes experienced small expansions and contractions. Substitution rates were calculated after the genes shifted from the LSC to IR. We investigated the variation of repeat sequences, SSRs, and sequence divergence among the ten complete plastomes. Molecular markers with rapid evolution rates were identified, which may be useful for further phylogenetic analysis and species identification. Phylogenies were constructed using the entire genomes. The availability of these ten plastomes provided valuable genetic information for accurately identifying species, clarifying taxonomy and reconstructing the intergeneric phylogeny of Cannabaceae.

Acknowledgments

This study was supported by grants from the National Natural Science Foundation of China, key international (regional) cooperative research project (31720103903), The Strategic Priority Research Program of the Chinese Academy of Sciences (XDPB0201). We would like to thank the Beijing Botanical Garden, Shanghai Chen Shan Botanical Garden, Wuhu Botanical Garden, Missouri Botanical Garden, and San Francisco Botanical Garden for permission to sample fresh leaves, Shudong Zhang, Jie Cai for providing samples, Yinhuan Wang, Rong Zhang for experimental assistance, Xiaojian Qu, Siyun Chen, Yingying Yang for data analysis and their valuable comments. This study was conducted in the Key Laboratory of the Southwest China Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.pld.2018.04.003.

References

Aitchison, S.F., Gish, W., Miller, W., et al., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Bell, C.D., Soltis, D.E., Soltis, P.S., 2010. The age and diversification of the angiosperms re-revisited. Am. J. Bot. 97, 1296–1303.
Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580.
Blazier, J.C., Jansen, R.K., Mower, J.P., et al., 2016. Variable presence of the inverted repeat and plastome stability in Erodium. Ann. Bot. 117, 1209–1220.
Byng, J.W., Chase, M.W., Christenhusz, M.J.M., et al., 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot. J. Linn. Soc. 181, 1–20.
Cao, T.S., 1993. Xuan Paper of China. China Light Industry, Beijing, pp. 20–34.
Cardle, L., Ramsay, L., Milbourne, D., et al., 2000. Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156, 847–854.
Chase, M.W., Cowan, R.S., Hollingsworth, P.M., et al., 2007. A proposal for a standardised protocol to barcode all land plants. Taxon 56, 295–299.
Dong, W.P., Liu, J., Yu, J., et al., 2012. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. PLoS One 7, e33071.
Downie, S.R., Jansen, R.K., 2015. A comparative analysis of whole plastid genomes from the Apiaceae: expansion and contraction of the inverted repeat, mitochondrial to plastid transfer of DNA, and identification of highly divergent noncoding regions. Syst. Bot. 40, 336–351.
Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19, 11–15.
Yang, J.B., Li, D.Z., Li, H.T., 2014. Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. Mol. Ecol. Resour. 14, 1024–1031.

Yang, M.Q., van Velzen, R., Bakker, T.F., et al., 2013. Molecular phylogenetics and character evolution of Cannabaceae. Taxon 62, 473–485.

Yesson, C., Russell, S.J., Parrish, T., et al., 2004. Phylogenetic framework for Trema (Celtidaceae). Plant Syst. Evol. 248, 85–109.

Zavada, M.S., Kim, M., 1996. Phylogenetic analysis of Ulmaceae. Plant Syst. Evol. 200, 13–20.

Zhang, S.D., Jin, J.J., Chen, S.Y., et al., 2017. Diversification of Rosaceae since the late Cretaceous based on plastid phylogenomics. New Phytol. 214, 1355–1367.

Zhang, S.D., Soltis, D.E., Yang, Y., et al., 2011a. Multi-gene analysis provides a well-supported phylogeny of Rosales. Mol. Phylogen. Evol. 60, 21–28.

Zhang, Y.J., Ma, P.F., Li, D.Z., 2011b. High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). PLoS One 6, e20596.

Zhu, A., Guo, W., Gupta, S., et al., 2016. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytol. 209, 1747–1756.