Gene 528 (2013) 120–131

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

The complete chloroplast genome sequence of *Mahonia bealei* (Berberidaceae) reveals a significant expansion of the inverted repeat and phylogenetic relationship with other angiosperms

Ji Ma [a], Bingxian Yang [a], Wei Zhu [a], Lianli Sun [a], Jingkui Tian [a,*], Xumin Wang [b,*]

[a] College of Biomedical Engineering & Instrument Science, Zhejiang University, 38 Zheda Road, Hangzhou, Zhejiang, China
[b] Beijing Institute of Genomics, Chinese Academy of Science, Beijing, China

ARTICLE INFO

Article history:
Accepted 18 July 2013
Available online 27 July 2013

Keywords:
Mahonia bealei
Chloroplast
IR expansion
SSRs
Molecular dating
RNA editing

ABSTRACT

Mahonia bealei (Berberidaceae) is a frequently-used traditional Chinese medicinal plant with efficient anti-inflammatory ability. This plant is one of the sources of berberine, a new cholesterol-lowering drug with anti-diabetic activity. We have sequenced the complete nucleotide sequence of the chloroplast (cp) genome of *M. bealei*. The complete cp genome of *M. bealei* is 164,792 bp in length, and has a typical structure with large (LSC 73,052 bp) and small (SSC 18,591 bp) single-copy regions separated by a pair of inverted repeats (IRs 36,501 bp) of large size. The *Mahonia* cp genome contains 111 unique genes and 39 genes are duplicated in the IR regions. The gene order and content of *M. bealei* are almost unarranged which is consistent with the hypothesis that large IRs stabilize cp genome and reduce gene loss-and-gain probabilities during evolutionary process. A large IR expansion of over 12 kb has occurred in *M. bealei*, 15 genes (*rps19*, *rpl22*, *rps3*, *rpl16*, *rpl14*, *rps8*, *infA*, *rpl36*, *rps11*, *petD*, *petB*, *psbH*, *psbN*, *psbT* and *psbB*) have expanded to have an additional copy in the IRs. The IR expansion rearrangement occurred via a double-strand DNA break and subsequence repair, which is different from the ordinary gene conversion mechanism. Repeat analysis identified 39 direct/inverted repeats 30 bp or longer with a sequence identity ≥90%. Analysis also revealed 75 simple sequence repeat (SSR) loci and almost all are composed of A or T, contributing to a distinct bias in base composition. Comparison of protein-coding sequences with ESTs reveals 9 putative RNA edits and 5 of them resulted in non-synonymous modifications in *rpoC1*, *rps2*, *rps19* and *ycf1*. Phylogenetic analysis using maximum parsimony (MP) and maximum likelihood (ML) was performed on a dataset composed of 65 protein-coding genes from 25 taxa, which yields an identical tree topology as previous plastid-based trees, and provides strong support for the sister relationship between Ranunculaceae and Berberidaceae. Molecular dating analyses suggest that Ranunculaceae and Berberidaceae diverged between 90 and 84 mya, which is congruent with the fossil records and with recent estimates of the divergence time of these two taxa.

1. Introduction

Chloroplast (cp) is a photosynthetic organelle that provides essential energy for plant cells (Dyall et al., 2004). Except for photosynthesis, lots of biosynthetic activities take place in chloroplast, such as the production of certain amino acids and lipids, as well as certain pigments in flowers and several key pathways of sulphur and nitrogen metabolism. It is also considered to have association with the plant's immune response (Pyke, 1999). In angiosperms, most cp genomes are circular DNA molecules ranging from 120 to 160 kb in length and highly conserved in gene content and orders (Wicke et al., 2011). However, large-scale genomic rearrangement and gene loss-and-gain events are also identified in some species (Bausher et al., 2006). The probability of gene loss-and-gain events and genomic rearrangements is closely related to the size of IRs (Wu et al., 2007), and large IRs can help stabilize the rest of the part of the cp genome and prevent gene loss-and-gain rearrangements to take place (Xiao et al., 2008; Yang et al., 2010). Because of the existence of large IRs (~23–26 kb), the relative sizes of LSC, SSC and IRs of most angiosperms remain constant, and gene order and content are well conserved (Shinozaki et al., 1986). However, cpDNAs which have lost some of its inverted repeat segments tend to have more rearrangements and gene loss conditions, the cp genome of gymnosperms such as pines and cypresses has lost one of the relevant large inverted repeats (Kolodner and Tewari, 1979), and gene loss-and-gain events and structural rearrangements were more frequently found in these species (Yi et al., 2013).

Mahonia bealei (Berberidaceae) is a traditional Chinese herb that deals with all sorts of inflammation, ranging from teeth inflammation,

Abbreviations: Cp, chloroplast; cpDNA, chloroplast DNA; IGS, intergenic spacer; Indel, insertion/deletion; IR, inverted repeat; LSC, large single-copy region; SSC, single-copy region.

* Corresponding authors. Tel.: +86 571 87951301; fax: +86 571 8795 1091.
E-mail addresses: tjk@zju.edu.cn (J. Tian), wangxm@big.ac.cn (X. Wang).

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.07.037

pneumonia to mastitis (Ferris and Zheng, 1999). This plant is a great medicinal value because it is one of the sources of berberine, which is considered to be a new cholesterol-lowering drug (Kong et al., 2004). Recently, there has been extensive research focused on the biochemical activities and mechanisms of actions of this anti-diabetic compound as berberine has shown to have anti-diabetic properties (Lee et al., 2006). As the population with hyperglycemia and hyperlipidemia is increasing rapidly, the unique medicinal value of this plant is gaining more and more attention. It is sure that access to genetic information of *M. bealei* will not only improve the genetic large-scale breeding and selection of germplasm process of this species, but also facilitate further usage of sequence data, such as phylogenetic and transplastomic analysis for the family of Berberidaceae.

Currently, there are only 46 sequences available for *M. bealei*, including 34 mitochondrial DNA sequences and 12 chloroplast DNA sequences, listed in GenBank (http://www.ncbi.nlm.nih.gov/genbank/nuccore/?term=mahonia+bealei). Due to the lack of chloroplast genome information of this medicinal plant, there is an urgent need to develop further genetic resources for it. Here, we report the complete cp genome sequence of *M. bealei*, using one of the next-generation sequencing (NGS) technologies — pyrosequencing (Illumina Genome Analyzer II), to mainly address two particular questions. First, we characterized the unique structure of this genome, and analyzed the differences comparatively with its relative species. Second, we performed phylogenetic analyses based on 65 protein-coding genes from 25 taxa, and molecular dating analyses to determine the divergent time of Ranunculaceae and Berberidaceae, which may contribute to a better understanding of the evolution progress of the *Mahonia* species. Additionally, we also described details of the genome assembly, annotation, and nucleotide polymorphisms of *M. bealei*.

2. Materials and methods

2.1. DNA sequencing and genome assembly

Fresh green leaves from adult plants were collected for the preparation of genomic DNA extraction. 5 μg purified DNA was used for the construction of cpDNA libraries.

Since the original sequence reads are a mixture of DNA from nucleus and organelles, BLAT (Kent, 2002) software was used to isolate chloroplast reads from the raw reads based on known reference cp genomes. SolexaQA (Cox et al., 2010) was used to filter low quality reads with the options −h 25 (quality cutoff p = 0.01) and −l 40 (length cutoff = 40 bp). SOAPdenovo (Luo et al., 2012) was carried out for the assembly with an option −K 57 (Kmer size = 57 bp). Then, all contigs were mapped to the reference chloroplast genome of *Nadina domestica* (NC_008336) using BLATZ (Schwartz et al., 2003) to identify the position and direction of the contigs. Gaps between contigs were filled up with a method like this: first, we used BLAT to map chloroplast reads onto both ends of the assembled contigs and then elongated the contigs by joining the reads, which were partly overlapped (≥95% identical) with the contigs. The two steps were taken repeatedly until the gaps were filled; we thus acquired the complete cp genome of *M. bealei*. To avoid assembly errors, we tested the length distribution of all assembled paired-end reads (Supplementary Fig. 1), and 3 pairs of primers for PCR amplifications were designed based on the variation of the genome to validate the sequence of the assembly (Supplemetary Table 1).

2.2. Genome annotation and analysis

The *M. bealei* cp genome was annotated using DOGMA (Dual Organellar GenoMe Annotator, Wyman et al., 2004). Predicted annotations of genes and open reading frames (ORFs) were identified with BLAST tools and ORF finder at NCBI website (http://www.ncbi.nlm.nih.gov/). The transfer RNA genes were identified by using DOGMA paralleled with using tRNAscan-SE1.23 (Lowe and Eddy, 1997). The functional classification of the annotated genes was referred to DOGMA. The circular map of the *M. bealei* cp genome was drawn by the OGdraw online tool (Lohse et al., 2007).

2.3. Repeat structure

The REPuter program (Kurtz et al., 2001) was used to assess both direct (forward) and inverted (palindrome) repeats within the cp genome. The identity and the size of the repeats were limited to no less than 90% (hamming distance equal to 3) and 30 bp in unit length, respectively (−f −p −l 30 −h 3 −best 10000). We ran the same REPuter analyses against the other three cp genomes of Ranunculales species which were also used for mVISTA to assess the relative number of the repeat sequences in *M. bealei* cp genome. MISA (Thiel et al., 2003) were used to detect simple sequence repeats (SSRs), with thresholds of mononucleotide repeats ≥10 bases, dinucleotide repeats ≥12 bases, tri- and tetranucleotide repeats ≥15 bases, hexanucleotide or greater repeats ≥24 bases.

2.4. RNA editing

Positional determination of potential RNA edits was accomplished by aligning publicly available *M. bealei* EST sequences obtained from GenBank to our assembly. BioEdit (Hall, 1999) was used to identify genome loci that have more than one nucleotide type as candidates for base changes. We only considered those loci with nucleotide sequencing quality value greater than 20 and the limit of the number of aligned EST reads for each selected locus is set to be above 4.

2.5. Phylogenetic and divergence time analysis

The data set consisted of concatenated nucleotide sequences of 65 protein coding genes including *atpA*, *B*, *E*, *F*, *H*, *I*, *ccsA*, *cemA*, *ndhA*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, *I*, *J*, *K*, *petA*, *B*, *D*, *G*, *L*, *N*, *psaA*, *B*, *C*, *I*, *J*, *B*, *C*, *psbD*, *E*, *F*, *H*, *I*, *J*, *K*, *L*, *M*, *N*, *T*, *Z*, *rbcL*, *rpl2*, *16*, *20*, *23*, *32*, *33*, *36*, *rps2*, *3*, *4*, *7*, *8*, *11*, *12*, *14*, *15*, *16*, *18*, *19* and *matK* extracted from 25 chloroplast genome sequences representing all lineages of angiosperms. These genes were extracted from complete cp genome sequences in the GenBank database with local perl scripts. Sequences were aligned using MAFFT (Katoh and Standley, 2013) and manually adjusted.

Maximum likelihood (ML) trees were constructed using RAxML 7.0.4 (Stamatakis et al., 2008) with a general time reversible (GTR+I+G) model. Maximum parsimony (MP) trees were performed using PAUP* 4.10 (Swofford, 2003). Phylogenetic analyses excluded gap regions. All MP searches included 100 random addition replicates and TBR branch swapping with the Multrees option. Non-parametric bootstrap analyses were performed for MP analyses with 1000 replicates.

Divergence times were estimated by the program BEAST version 1.7.5 (Drummond et al., 2012), using the same dating strategies adopted by Moore et al. (2010). Three time constraints were applied for molecular dating, in which the minimum ages of angiosperms were set to be 131.8 mya, the minimum ages of eudicots were set to be 125 mya, and 85 mya for the most recent common ancestor of *Quercus* and *Cucumis*. Additionally, we set the age of the root node to be 308 mya as reports of conifer leaves and shoots are known from 309.2 to 307.1 mya.

3. Results

3.1. Genome assembly and validation

Using the Illumina Hiseq 2000 system, 16,139,329 paired-end reads were generated for this project. After filtering low-quality reads ($\leq Q_{20}$ bases) and aligning with reference cp genomes, we collected 562,270 reads (3.48% of total) reaching 266X coverage over the cp genome (Table 1). The unassembled reads (~96.52%) were mostly from the