Gene 540 (2014) 201–209

Contents lists available at ScienceDirect

Gene

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

The complete chloroplast genome sequence of *Taxus chinensis* var. *mairei* (Taxaceae): loss of an inverted repeat region and comparative analysis with related species

Yanzhen Zhang [1], Ji Ma [1], Bingxian Yang, Ruyi Li, Wei Zhu, Lianli Sun, Jingkui Tian, Lin Zhang *

College of Biomedical Engineering & Instrument Science, Zhejiang University, 38 Zheda Road, Hangzhou, Zhejiang, China

ARTICLE INFO

Article history:
Accepted 24 February 2014
Available online 26 February 2014

Keywords:
Taxus chinensis var. *mairei*
Chloroplast genome
Large inversion
IR loss
SSRs

ABSTRACT

Taxus chinensis var. *mairei* (Taxaceae) is a domestic variety of yew species in local China. This plant is one of the sources for paclitaxel, which is a promising antineoplastic chemotherapy drugs during the last decade. We have sequenced the complete nucleotide sequence of the chloroplast (cp) genome of *T. chinensis* var. *mairei*. The *T. chinensis* var. *mairei* cp genome is 129,513 bp in length, with 113 single copy genes and two duplicated genes (trnI-CAU, trnQ-UUG). Among the 113 single copy genes, 9 are intron-containing. Compared to other land plant cp genomes, the *T. chinensis* var. *mairei* cp genome has lost one of the large inverted repeats (IRs) found in angiosperms, fern, liverwort, and gymnosperm such as *Cycas revoluta* and *Ginkgo biloba* L. Compared to related species, the gene order of *T. chinensis* var. *mairei* has a large inversion of ~110 kb including 91 genes (from rps18 to accD) with gene contents unarranged. Repeat analysis identified 48 direct and 2 inverted repeats 30 bp long or longer with a sequence identity greater than 90%. Repeated short segments were found in genes rps18, rps19 and clpP. Analysis also revealed 22 simple sequence repeat (SSR) loci and almost all are composed of A or T.

1. Introduction

Since the first report of the complete chloroplast (cp) genome sequences of the tobacco and the liverwort (Shinozaki et al., 1986), a number of land plant chloroplast genomic sequences have been determined. These recent determination of complete cp genomic sequence of various plant species have enabled numerous comparative analyses as well as advancements in plant and genome evolutionary studies, including transcriptome analysis and pangenomes that are based on these data (Medini et al., 2005). Although the published complete cp genome sequence of gymnosperm species were few in number, unique characteristics such as genome-scale genomic rearrangement and a more frequent gene lost and gain events were found in them (Jansen et al., 2007). The probability of genomic rearrangements and gene loss events of a land plant cp genome during evolutionary progress was thought to have intimate relationship with the size of IRs (Wu et al., 2007). Large IRs can help stabilize the cp genome and reduce the possibility of gene loss and rearrangements (Xiao et al., 2008). In most angiosperms such as date palm (*Phoenix dactylifera* L.), the relative size of LSC, SSC and IRs remains constant, the gene order and organization are almost the same with inferred ancestral angiosperm cp genomes (Yang et al., 2010). However, some clades of gymnosperm such as Pinaceae and Cupressaceae have lost one of the large inverted repeats, which lead to more gene loss and structural rearrangements in their cp genomes (Kolodner and Tewari, 1979).

Taxus chinensis var. *mairei* is a variety of the *Taxus* genus, yew family (Taxaceae) in domestic China. Its secondary metabolite paclitaxel (taxol) is a chemotherapy drug given to treat ovarian, breast and non-small cell lung cancer, which is one of the most promising antineoplastic agents of the last decade, with demonstrated activity in advanced and refractory ovarian, breast, lung, and head and neck cancers (Rowinsky et al., 1993). Paclitaxel was first isolated from the bark of pacific yew tree in 1970s, but leaves of *Taxus* were also examined as a source of paclitaxel and related toxoids (Ketchum et al., 1999). As the breast cancer rate increases, the unique medicinal value of *Taxus* was gradually recognized. The access to plastid genome information of *T. chinensis* var.

Abbreviations: aa, amino acid; ATP, adenosine triphosphate; BLAT, The BLAST-Like Alignment Tool; BLAST, Basic Local Alignment Search Tool; CDS, coding sequence; cp, chloroplast; cpDNA, chloroplast DNA; CRA, comparative repeat analysis; CST, chloroplast SSR type; DNase, Deoxyribonuclease; DOGMA, Dual Organellar GenoMe Annotator; F, forward; I, invariable sites; IGS, intergenic spacer; indel, insertion/deletion; I, inverted; IR, inverted repeat; JLA, JLB, junctions between the two IRs and the LSC region; JSA, JSB, junctions between the two IRs and the SSC region; Kb, Kilobase; LSC, large single-copy region; NADH, nicotinamide adenine dinucleotide; nt, nucleotide(s); ORF, open reading frame; P, palindromic; PCR, polymerase chain reaction; RNase, ribonuclease; SOAP, Short Oligonucleotide Analysis Package; SolexaQA, The Solexa sequencing data quality assessment package; SSC, small single-copy region; SSR, simple sequence repeat; tRNA, transfer ribonucleic acid.

* Corresponding author.
E-mail address: zhanglin@zju.edu.cn (L. Zhang).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.gene.2014.02.037
0378-1119/

mairei will provide usage of information for further transcriptomic and proteomic analysis, and pave the way to study the enzymes that catalyze the biosynthesis of the natural compounds in chloroplast.

Currently, the gene content and genomic structure of some species of gymnosperms are still little known, because there are only 3 published complete cp genome sequences of Taxaceae in GenBank (http://www.ncbi.nlm.nih.gov). Here, we report the complete cp genome sequence of *Taxus c.* var. *mairei*, the first reported cp genome in the *Taxus* genus. In this report, we described details of the genome assembly, annotation, and simple sequence repeats (SSRs). Dot-plot analyses and genomic comparative analyses were also performed in order to better understand the unique structure of the cp genome of *T. c.* var. *mairei*.

2. Materials and methods

2.1. DNA sequencing and genome assembly

Fresh leaves of *T. chinensis* var. *mairei* were collected for the preparation of genomic DNA extraction. 5 μg purified DNA was used for the construction of cp DNA libraries. Solexa high-throughput sequencing system (Illumina Genome Analyzer II) was used to generate raw sequence reads for this project.

Since the original sequence reads are a mixture of DNA from nucleus and organelles, BLAT (Kent, 2002) software was used to isolate chloroplast-related 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). Contigs were then manually assembled into the complete circular genome sequence based on the reference chloroplast genome *Cephalotaxus oliveri* (NC_021110).

Gaps between contigs were filled up with a method like this: we use BLAT to map the raw sequenced reads onto both ends of the assembled contigs and scaffolds and elongated the contig by joining overlapping reads with it, and the two steps were taken repeatedly until the gaps between contigs were filled. To avoid assembly errors, we designed 4 pairs of primers used for PCR amplifications to validate the possible SC/IR boundary regions and polymer-abundant segments (Supplementary Table 2).

2.2. Genome annotation

The *T. chinensis* var. *mairei* cp genome was annotated using DOGMA (Dual Organellar GenoMe Annotator, Wyman et al., 2004). The predicted gene sequences were rechecked by BLAST online tool (http://blast.ncbi.nlm.nih.gov/Blast.cgi), and open reading frames (ORFs) were identified by ORF finder on NCBI website (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). The predicted tRNA genes were rechecked by tRNAscan-SE (Lowe and Eddy, 1997). Codon usage was calculated by CodonW (http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html). The circular map of *T. chinensis* var. *mairei* was drawn by OGdraw online tool (Lohse et al., 2007).

2.3. Comparative analysis of genomic structure and gene diversification

The Harr-plot analysis was performed using PipMaker (Schwartz et al., 2000), and was used to provide comparative analysis of genomic architecture between *T. chinensis* var. *mairei* and its related species (represented by *Cephalotaxus wilsoniana* and *Cycas taitungensis*). The GRIMM web server (Tesler, 2002) was used to identify the minimum number of rearrangements by inversion in pairwise comparisons of the cp genome to estimate the genome rearrangements. Eight cp genomes were used to explore the diversified genes: *C. taitungensis* (NC_009618), *G. biloba* (JN867583), *Cedrus deodara* (NC_014575), *Podocarpus totara* (NC_020361), *Nicotiana tabacum* (NC_001879), *Oryza sativa* (JN861110), *Zea mays* L. (NC_001666), *Cryptomeria japonica* (AP009377), and *Taiwania cryptomerioides* (NC_016065). ClustalX (Larkin et al., 2007) were used to indicate the variable sites within the diversified genes by aligning gene sequences of *T. chinensis* var. *mairei* with other species.

2.4. Repeat structure

The REPuter program (Kurtz et al., 2001) was used to assess both direct (forward) and inverted (palindrome) repeats within the cp genome of *T. chinensis* var. *mairei*. 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. MISA (Thiel et al., 2003) was used to detect simple sequence repeats (SSRs), with thresholds of mononucleotide repeats ≥10 bases, dinucleotide repeats ≥12 bases, tri- and tetranucleotide repeats ≥15 bases, and hexanucleotide or greater repeats ≥24 bases.

3. Results

3.1. Genome assembly and validation

Using the Illumina Hiseq 2000 system, 49,743,352 paired-end reads were generated to assemble the cp genome of *T. chinensis* var. *mairei*. After filtering low-quality reads (≤Q20 bases) and aligning with reference cp genomes, we collected 1,802,286 reads (3.62% of total) reaching 95× coverage over the cp genome (Supplementary Table 1). The unassembled reads (~96.38%) were mostly from the nuclear genome due to the raw reads which was a collection of DNA from nucleus and organelles. We have manually corrected ambiguous nucleotide sites in the cp genome, which were produced in the scaffold extension step during de novo assembly. We also corrected errors associated with heterogeneous insertions/deletions (Indels), which were arisen from homopolymeric repeats in the genome. To test the assembly of the cp genome of *T. chinensis* var. *mairei*, we designed 4 pairs of primers for PCR amplification to validate the possible boundary regions and polymer-abundant regions (Supplementary Table 2 and Fig. S1). The final complete chloroplast genome sequence of *T. chinensis* var. *mairei* has been deposited into GenBank under the accession number KJ123824.

3.2. General features of the T. chinensis var. mairei cp genome

The complete cp genome of *T. chinensis* var. *mairei* is 129,513 bp in length (Fig. 1), shorter than that of *C. wilsoniana* (136,196 bp) (Wu et al., 2011) and *C. oliveri* (134,337 bp) (Yi et al., 2013), which are all in the Cephalotaxaceae family; and also shorter than that of *T. cryptomerioides* (132,588 bp) (Wu et al., 2011) and *C. japonica* (131,810 bp) (Hirao et al., 2008), which belong to the Cupressaceae family. Among gymnosperms, *C. taitungensis* has the largest cp genome size (163,403 bp) (Wu et al., 2007), and *G. biloba* (156,945 bp) ranks second (Lin et al., 2012). Similar to plum-yews and most pines, the cp genome of *T. chinensis* var. *mairei* has no typical IR regions, resulting in the boundary of LSC and SSC regions unable to be defined.

The *T. chinensis* var. *mairei* cp genome encodes a total of 116 unique genes (Table 1). Among them, 113 genes are single copy, except for two transfer RNA genes (trnI-CAU and trnQ-UUG), which are duplicated to have double copies in the cp genome. Among the 113 single copy genes, there are 4 ribosomal RNA genes (3.36%), 31 individual transfer RNA genes (26.05%), 21 gene encoding large and small ribosomal subunits (18.58%), four gene encoding DNA-dependent RNA polymerases (3.36%), 47 gene encoding photosynthesis related proteins (40.52%), and 8 gene encoding other proteins, including four encoding proteins with unknown functions (6.89%).

The overall AT content of *T. chinensis* var. *mairei* cp genome is 65.37%, which is similar to that of *C. oliveri* (64.76%) and *C. wilsoniana* (64.92%),