Plant Gene xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Plant Gene

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

Toward phylogenomics of Lauraceae: The complete chloroplast genome sequence of *Litsea glutinosa* (Lauraceae), an invasive tree species on Indian and Pacific Ocean islands

Damien Daniel Hinsinger, Joeri Sergej Strijk *

Plant Ecophysiology & Evolution Group, State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, College of Forestry, Guangxi University, Nanning, Guangxi 530005, PR China

ARTICLE INFO

Article history:
Received 15 June 2016
Received in revised form 15 August 2016
Accepted 19 August 2016
Available online xxxx

Keywords:
Litsea glutinosa
Lauraceae
Complete chloroplast genome
Invasive species
Phylogenomics
Islands

ABSTRACT

Litsea glutinosa (Lour.) C.B.Rob. is a native species from Indo-Malesia and China. Due to its invasive behaviour on Indian and Pacific Ocean islands, a better understanding of *L. glutinosa* is becoming increasingly urgent to improve management practices and our understanding of its dispersal ability throughout its native range. Indeed, knowledge on the distribution of genetic diversity in native populations is crucial to understand the factors that drive the invasive character of *L. glutinosa*. Here we assemble and analyze its complete chloroplast genome sequence, the first in the genus *Litsea*. The total genome size was 152,618 bp in length, containing a pair of inverted repeats (IRs) of 20,063 bp, which were separated by a large single copy (LSC) and small single copy (SSC) of 93,690 bp and 18,802 bp, respectively. The overall GC content of the plastid genome was 39.2%. 127 genes were annotated, including 83 protein-coding genes, 36 tRNA genes and 8 rRNA genes. In these genes, eighteen contained one or two introns. Nine repeated sequences (5 palindromic and 4 forward) and 56 simple sequence repeats were identified in the plastid genome of *L. glutinosa*. Comparing our sequence with available complete chloroplast genomes in Lauraceae, five intergenic spacers (including *trnH-psbA* and *rpl32-trnL*) and one intron showed promising levels of variations for application in DNA-barcoding or intrageneric studies. In addition, phylogenetic analysis of complete Magnoliid plastid genomes highlighted affinities between *Litsea* and *Cinnamomum*. These results are expected to be useful to both our understanding of the characteristics and evolution of the invasive behaviour, as well as to efficiently manage these pest species in their introduced areas.

1. Introduction

Lauraceae currently comprise about 50 genera, holding ±3000 economically and ecologically important species found in tropical and subtropical tropical regions of both hemispheres (Van Der Werff et al., 1996), with important centres of endemism in Southeast Asia and the Neotropics. They comprise one of the most widespread and species rich families, with local ecological dominance accounting for up to 20% of the species present, and general morphological features of the family leading to the designation of a unique habitat type (Laurasilva) where they co-occur with families that have similar morphotypes (Magnoliaceae, Winteraceae, Fagaceae, Myrtaceae). The vast species diversity and geographic range, lack of diagnostic characters and insufficient taxonomic attention have led to a general poor understanding of the family to this day. Full plastome data for only five species (three genera) in the family Lauraceae are currently deposited in GenBank, preventing any phylogenomic analyses of the family or understanding the pattern and process in the family's historical (and extant) distribution and evolutionary diversification. The urgency of increasing the availability of genomic information on this large relict family by adding whole chloroplast genome data of new genera is a fundamental step in clarifying its internal evolutionary connections and ecological roles as a key forest component.

In Lauraceae, the genus *Litsea* contains about 130 species, and represents a middle-size genus between several monotypic and highly speciose genera (e.g. *Cryptocarya* – >350 sp., *Ocotea* – >320 sp., *Cinnamomum* – >300 sp.) (Chanderbali et al., 2001; Van der Werff, 1996; Van Der Werff et al., 1996). To gain information for further studies on Lauraceae phylogenomics, we assembled the complete chloroplast genome of *L. glutinosa* (Lour.) C.B. Rob., a tree native in India, Southern China to Malaysia, Australia and the western Pacific islands. Due to the occurrence of both paraphyletic and polyphyletic genera in Lauraceae (in their current circumscription), and because of the extensive time- and resource-consumption associated with sequencing the complete chloroplasts of all species in the family (±3000), careful selection of representative species in each genus is crucial in any attempt to redefine

Abbreviations: Cp, chloroplast; LSC, large single-copy region; SSC, small single-copy region; IR, inverted repeat; bp, base pair(s); rRNA, ribosomal RNA; tRNA, transfer RNA; IGS, intergenic spacer; CDS, coding sequence; SSR, simple sequence repeat; ML, maximum likelihood.

* Corresponding author at: Plant Ecophysiology & Evolution Group, Room 124, State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, College of Forestry, Guangxi University, Nanning, Guangxi 530005, PR China.
E-mail address: jsstrijk@gxu.edu.cn (J.S. Strijk).

http://dx.doi.org/10.1016/j.plgene.2016.08.002
2352-4073/

and assess the evolutionary structure and relationships within Lauraceae. Therefore, we selected *L. glutinosa* in a wider approach involving the selection of multiple core generic representatives (sensu Li et al., 2004), in Laureae (sensu Chanderbali et al., 2001) and we use this approach to initiate the construction of a solid phylogenomic framework to finally elucidate the paraphyletic and polyphyletic complexities hampering our understanding of the family. *L. glutinosa* is of ecological interest, as it is considered a severe invasive element in Indian and Pacific Ocean islands (Jacq et al., 2005; but see US Forest Service, 2013). It has been introduced as structural support in vanilla agroforestry and for charcoal production (Kueffer et al., 2004) but has since spread outside plantations. The tree has a strong ability to displace regenerating native plant species in disturbed environments and its management is causing severe conflicts of interests (Kueffer et al., 2004; Macdonald et al., 1991). The IUCN Red List (IUCN, 2015) currently holds 281 species, with the majority (72%) in the top three categories (CR - critically endangered, EN - endangered and VU - vulnerable) and the most important threat reported being logging and wood harvesting. Within the Red List, six species of threatened Lauraceae are recorded as being invasive in other habitats, but *L. glutinosa* is not among them, as it is not threatened in its native range. The Global Invasive Species Database (www.issg.org) lists three Lauraceae as invasive species, while the Pacific Island Ecosystems at Risk project (PIER - www.hear.org) lists a total of seven invasive species. In recent years, the applications for use of sequenced plastomes have broadened from purely genomic to ecological and economic interests. The use of genomes in attempts to understand the ecological behaviour of invasive species is now fairly commonplace (Doorduin et al., 2011; Dowell et al., 2016; Huotari and Korpelainen, 2012; Leseberg and Duvall, 2009; Nie et al., 2012; Wood et al., 2016). More importantly, genomic information of invasive species has been demonstrated to have tremendous potential to understand and predict species' taxonomic diversity in their host ranges and the differential environmental responses and threats that these unique lineages may affect a newly invaded habitat. This may prove crucial in attempts to contain and treat existing invasive populations, and to prevent future potentially catastrophic invasive elements from establishing themselves by increasing quarantine and border screenings (Dowell et al., 2016; Wood et al., 2016).

In angiosperms, the chloroplast genome is a circular molecule (76–217 kb), with a conserved structure of two inverted repeats (IR) separated by small (SSC) and large (LSC) single-copy regions (Jansen and Ruhlman, 2012). Chloroplast DNA loci have been widely used in plant studies, both for evolutionary studies and for identification purposes, due to their natural abundance in plant cells (≈3–5% of the cell DNA content), compared to nuclear DNA, resulting in easier PCR amplification. Moreover, comparison of complete chloroplasts in a species, genus or a family have been shown to be useful in identifying loci for population genetics (SSRs, repeats) (Doorduin et al., 2011), DNA barcoding (intergenic and coding regions) (Coissac et al., 2016; Li et al., 2014; Nock et al., 2011) species delimitation and to delineate evolutionary history and taxonomy (Parks et al., 2012). In invasive species, complete organelles have been shown to be a powerful tool to unlock research in both animals (e.g. Liao et al., 2010) and plants (Doorduin et al., 2011; Huotari and Korpelainen, 2012; Nie et al., 2012), they allow discrimination of invasive populations (e.g. Dowell et al., 2016; Wood et al., 2016), despite reduced genetic diversity (e.g. Tsutsui et al., 2000), by the identification of variations (SNPs, indels, SSRs) in previously overlooked regions.

The goals of this study are to assemble the complete chloroplast genome of *L. glutinosa*, and to compare its features (SSRs, repeats) with other available Lauraceae chloroplast sequences. We expect that these genomic resources will allow future evolutionary and demographic studies to better manage invasive Lauraceae species, to understand the effects of different points of origin, and to serve as a starting point for selecting and analysing genomes of other core members within Lauraceae to elucidate the current complex paraphyletic and polyphyletic status of many of its genera.

2. Materials and methods

2.1. DNA extraction and sequencing

Genomic DNA was extracted from 0.1 g of silicagel-dehydrated leaves, from an individual collected in Yunnan (22°35′24″N, 99°30′01″E) in 2014 (voucher deposited at our research group herbarium, STRIJK_1654), using a protocol modified from Healey et al. (2014). The modifications were as follows: genomic DNA was extracted in 15 mL tubes, using 6 mL of extraction buffer, incubated at 65 °C for 60 min, and two volumes of temperate absolute ethanol were added for precipitation, without incubation. The library construction and sequencing were performed by Novogene (Beijing, China), using NEBNext Ultra II DNA Library Prep Kit (Ipswich, Massachusetts, USA) and an Illumina HiSeq2500 platform (San Diego, California, USA), respectively, according to the manufacturer's specifications. 1 Gb of raw data (paired-end 125 bp reads, 500 bp fragments length) were generated.

2.2. Chloroplast genome reconstruction

Raw data were imported in Geneious R9 v.9.0.5 (Biomatters Ltd., Auckland, New Zealand). Raw reads were trimmed according to their quality, removing bases from 5′ and 3′ –ends until no base with $Q < 20$ was found. Reads with <10 low quality bases and/or ambiguities were then assembled using an iterative reference-guided assembly approach, as implemented in Geneious R9 and successfully used in different taxonomic groups (Hinsinger and Strijk, 2015; Raman and Park, 2016). Using the available chloroplast of *Machilus balansae* (KT348517), the algorithm iteratively mapped the reads against the reference, starting with the most conserved regions. These first contigs were then used as a "pseudo-reference" and refined or extended with the partially overlapping reads newly mapped. This approach is very similar to MITObim (Hahn et al., 2013), which was successfully used in several chloroplast reconstructions (Du et al., 2015; Mariac et al., 2014). Such approaches that iteratively extend and improve the contigs, and merge them when two adjacent contigs are overlapping, benefit from the advantage of mapping, namely a low sensitivity to artifactual low coverage areas, without its main disadvantage, by allowing the reconstructed sequence to be significantly different from the provided reference sequence. 1000 iterations were performed with gaps allowed (up to 15% of the reads length), a word length of 14 bp and an index word length of 12 bp. The maximum mismatch per read and maximum ambiguities were set to 30% and 4, respectively. The "Accurately map reads with errors to repeat regions" option was checked, only reads assembled to the correct distance (i.e. ≈500 bp) were considered by the Geneious algorithm, and this information was used for scaffolding.

Positions under 10× coverage were masked (replaced by Ns) for the generation of a consensus sequence (Ripma et al., 2014; Straub et al., 2012). We individually checked these positions to verify the base calling accuracy and their identity to the reference. We chose this conservative approach to not include in the final sequence positions from the *M. balansae* reference and these positions were accounted for in the length calculations. The inverted repeat borders were carefully checked by eye. No evidence for any structural change of these IR borders were found, as the mapping depth and base calling were without ambiguity, as previously demonstrated in other groups (Kremer et al., 2012; Lu et al., 2015). Raw Illumina reads and the complete plastid genome were submitted to GenBank under accession number SRP072142 and KU382356, respectively.

2.3. Genome annotation and sequence statistics

We determined annotations using cpGAVAS (Liu et al., 2012a) and validated their boundaries using ORF Finder (NCBI), followed by manual adjustments. We used tRNAscan-SE v1.21 (Schattner et al., 2005) to

Please cite this article as: Hinsinger, D.D., Strijk, J.S., Toward phylogenomics of Lauraceae: The complete chloroplast genome sequence of *Litsea glutinosa* (Lauraceae), an invasive tree species on Indian and Pacific Ocean islands..., (2016), http://dx.doi.org/10.1016/j.plgene.2016.08.002