ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/330781961

# The complete chloroplast genome sequence of the endangered species Syringa pinnatifolia (Oleaceae)

**Article** *in* Nordic Journal of Botany · January 2019

DOI: 10.1111/njb.02201

CITATION
1

READS
176

**5 authors**, including:

Han Zhao
Northwest A & F University
**3** PUBLICATIONS **1** CITATION
SEE PROFILE

Jing Cai
Northwest A & F University
**5** PUBLICATIONS **6** CITATIONS
SEE PROFILE

**Some of the authors of this publication are also working on these related projects:**




tree water physiology View project

All content following this page was uploaded by Han Zhao on 17 December 2019.

The user has requested enhancement of the downloaded file.

NORDIC JOURNAL OF
BOTANY

*Research*

# The complete chloroplast genome sequence of the endangered species *Syringa pinnatifolia* (Oleaceae)

**Jingwen Zhang, Zaimin Jiang, Hao Su, Han Zhao and Jing Cai**

*J. Zhang (https://orcid.org/0000-0002-8121-5040), H. Su, H. Zhao and J. Cai (https://orcid.org/0000-0001-7594-0783) ✉ (cjcaijing@163.com), College of Forestry, Northwest A&F University, CN-712100 Yangling, PR China. – Z. Jiang, College of Life Science, Northwest A&F University, Yangling, PR China.*

**Nordic Journal of Botany**
**2019: e02201**
doi: 10.1111/njb.02201

Subject Editor and
Editor-in-Chief: Torbjörn Tyler
Accepted 22 January 2019

*Syringa pinnatifolia* is an endangered endemic species in China with important ornamental and medicinal value, and it needs urgent protection. Here, we report the complete chloroplast (cp) genome structure of *S. pinnatifolia* and its evolution is inferred through comparative studies with related species. The *S. pinnatifolia* cp genome was 155 326 bp and contained a large single copy region (LSC) of 86 167 bp and a small single copy region (SSC) of 17 775 bp, as well as a pair of inverted repeat regions (IRs) of 25 692 bp. A total of 113 unique genes were annotated, including 79 protein-coding genes, 30 tRNA genes and four rRNA genes. The GC content of the *S. pinnatifolia* cp genome was 37.9%, and the corresponding values in the LSC, SSC and IR regions were 36.0, 32.1, 43.2% respectively. Repetitive sequences analysis revealed that the *S. pinnatifolia* cp genome contained 38 repeats. Microsatellite marker detection analysis identified 253 simple sequence repeats (SSRs), which provides opportunities for future studies of the population genetics and phylogenetic relationships of *Syringa*. Phylogenetic analysis of 29 selected cp genomes revealed that *S. pinnatifolia* is closely related to *Syringa vulgaris* and all 27 Lamiales species formed a clade separate from the two outgroup species. This newly characterized *S. pinnatifolia* chloroplast genome will provide a useful genomic resource of phylogenetic inference and the development of more genetic markers for species discrimination and population studies in the genus *Syringa*.

Keywords: conservation, phylogenetic analysis

## Introduction

*Syringa* (Oleaceae), is a tree genus with important ornamental and economic value, that includes approximately 35 species distributed in East Asia and southeastern Europe (Ming et al. 2007). Due to its diverse habitat, wide adaptability and unique ornamental characters, *Syringa* has a cultivation history of nearly 1000 years (Zang and Liu 1999, Cui et al. 2004). *Syringa pinnatifolia* is the only plant with pinnately compound leaves in the genus (Jin et al. 2008). It is also a National Grade III Endangered Species, because of its poor fecundity, slow growth and destruction of habitats caused

© 2019 The Authors. Nordic Journal of Botany © 2019 Nordic Society Oikos

NORDIC SOCIETY OIKOS

www.nordicjbotany.org

by deforestation. In the wild, *S. pinnatifolia* has been continuously reduced and will soon be in danger of extinction (Jiang et al. 1999).

*Syringa pinnatifolia* has potential economic value as a valuable medicinal material for traditional Chinese medicine and Mongolian medicine, known as 'Shanchenxiang' in Chinese. The roots and branches warm the kidney, and act to cure asthma, stomach pain and myocardial ischemia (Burie 2012, Su et al. 2015a, b). Thus, it may be worthwhile to conduct more in-depth research on the medicinal value of *S. pinnatifolia*. What's more, it can also be used as a garden ornamental tree with good adaptability. However, due to the rariety and endangered state of the species, there are only few studies of this species and our knowledge of it remain fragmentary.

Chloroplasts are the photosynthesis organelles of plants and an important unit of plant cell genetics. Chloroplasts carry autonomous genetic information referred to as the chloroplast genome or chloroplast DNA (cpDNA). It is 120–160 kb long in higher plants (Maier and Schmitz-Linneweber 2004). Most cpDNAs are double-stranded covalently closed circular molecules, with a typical quadripartite structure consisting of a large single copy (LSC) region and a small single copy (SSC) region interspersed between two inverted repeats (IRa/IRb) (Wicke et al. 2011).

The photosynthesis of green plants is the fundamental source of the oxygen needed for survival, reproduction and development of organisms on the planet, and photosynthesis in the chloroplasts is strictly genetically controlled. Therefore, studies of the chloroplast genome are important to reveal how photosynthesis efficiency may be improved. At the same time, the angiosperm cp genome sequences are highly conserved and generally maternally inherited (Kuroiwa et al. 1982). Thus, studying the structure and sequences of the cp genome is useful for revealing the origin of species, their evolution and the relationships among different species. In recent years, with the development of molecular biology and sequencing technology, as well as improvements of sequence splicing software and reduced sequencing costs, the acquisition of cp genome data has become efficient and rapid (Zhang and Li 2011). These advances have enabled the chloroplast genome database to be enriched rapidly, and research on cp genomes has increased. Up to now, approximately 2800 plant cp genomes have been completely sequenced and added to the National Center for Biotechnology Information (NCBI) database (NCBI 2018).

At present, no reports are available on the chloroplast genome of *S. pinnatifolia*, and only seven nucleotide sequences from this species have been published in GenBank. Here, we report the complete cp genome sequence of *S. pinnatifolia*, together with a characterization of its long repetitive sequences and simple sequence repeats (SSR). We also compared the complete cp genome of *S. pinnatifolia* with other Oleaceae plants. Our results will provide a basic genetic resource that will help to establish the accurate phylogenetic relationship and reasonable protection measures for *S. pinnatifolia* and other species of the genus.

## Material and methods

### Plant materials, DNA extraction and sequencing

The plant material used in this research was intact, fresh, young leaves collected from the *Syringa pinnatifolia* germplasm nursery of North West Agriculture and Forestry University. About 30 g of fresh leaf material was used to exact cpDNA using the modified high salt–low pH method (Diekmann et al. 2008, Liu 2012, Chen and Chen 2014). DNA quality was checked using an ultra-micro spectrophotometer and agarose gel electrophoresis. Cp genome sequencing was performed on an Illumina HiSeq 2500-PE125 platform with MPS (massively parallel sequencing) Illumina technology.

### Genome assembly and annotation

The low quality data were trimmed from the raw sequence data using the NGSQC toolkit (Patel and Jain 2012) (quality values <38, length >40%). After trimming, the high quality clean reads were assembled using SOAPdenovo (Li et al. 2008) with default settings (K-mer = 55). The output contigs were aligned using the *Sesamum indicum* cp genome as reference (GenBank accession number KC569603) to assemble the cp genome in GENEIOUS ver. 8.0.2 (Kearse et al. 2012). Next, the online web-based server BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) was used to verify the inverted repeat (IR) and the single copy (SC) junctions. The assembled genome was annotated using the dual organellar genome annotator (DOGMA) (Wyman et al. 2004), and using default parameters to predict protein-coding genes, transfer RNA (tRNA) genes, and ribosome RNA (rRNA) genes. Then, BLAST was used to check the annotation, followed by manual correction through comparison with other closely related cp genomes of Oleaceae in GENEIOUS 8.0.2 (Kearse et al. 2012). The tRNAscanSE (<http://lowelab.ucsc.edu/tRNAscan-SE/>) (Schattner et al. 2005) was used to identify tRNA genes. Finally, the gene map of the cp genome of *Syringa pinnatifolia* was generated using OGDraw (<https://chlorobox.mpimp-golm.mpg.de/OGDraw.html>) (Lohse et al. 2013).

### Genome comparison

The boundaries between the inverted repeat (IR) region and the single copy (SC) region of *S. pinnatifolia* and four other species of Oleaceae, including *Syringa vulgaris*, *Forsythia suspense*, *Olea europaea* and *Hesperelaea palmeri* were compared. Additionally, the whole cp genome of these five species with the *S. pinnatifolia* annotation as the reference was compared and plotted using the mVISTA program (Mayor et al. 2000, Frazer et al. 2004).

## Repeat analysis

We identified the location and size of the repeats, including forward, reverse, palindromic and complement sequences in the *S. pinnatifolia* cp genome using the REPuter program (<https://bibiserv.cebitec.uni-bielefeld.de/reputer>) (Kurtz et al. 2001) with a minimal repeat size of 20 bp. SSRhunter (Li and Wan 2005) was used to detect SSRs, with the minimum number of repeats set to eight repeat units for mononucleotide, four repeat units for dinucleotide and three repeat units for trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide SSRs.

## Codon usage

The codon usage frequency and RSCU was investigated using the CodonW 1.4.2 software (<http://codonw.sourceforge.net/>) (John 1999). All protein coding genes of the *S. pinnatifolia* cp genome were selected for the codon usage bias analysis. Relative synonymous codon usage (RSCU) refers to the relative probability of a synonymous codon for a particular codon. When synonymous codons are used less frequently than expected, the RSCU value <1, otherwise the value >1 (Sharp and Li 1987).

## Phylogenetic analysis

Phylogenetic analysis was preformed using the complete *S. pinnatifolia* chloroplast genome and 26 Lamiales species plus two outgroup species (*Coffea canephora*, *Catharanthus roseus*) (Supplementary material Appendix 1 Table A1). The complete cp genomes of these species were downloaded from GenBank. All cp genome sequences were aligned using ClustalW (Larkin et al. 2007). MEGA7.0 (Kumar et al.

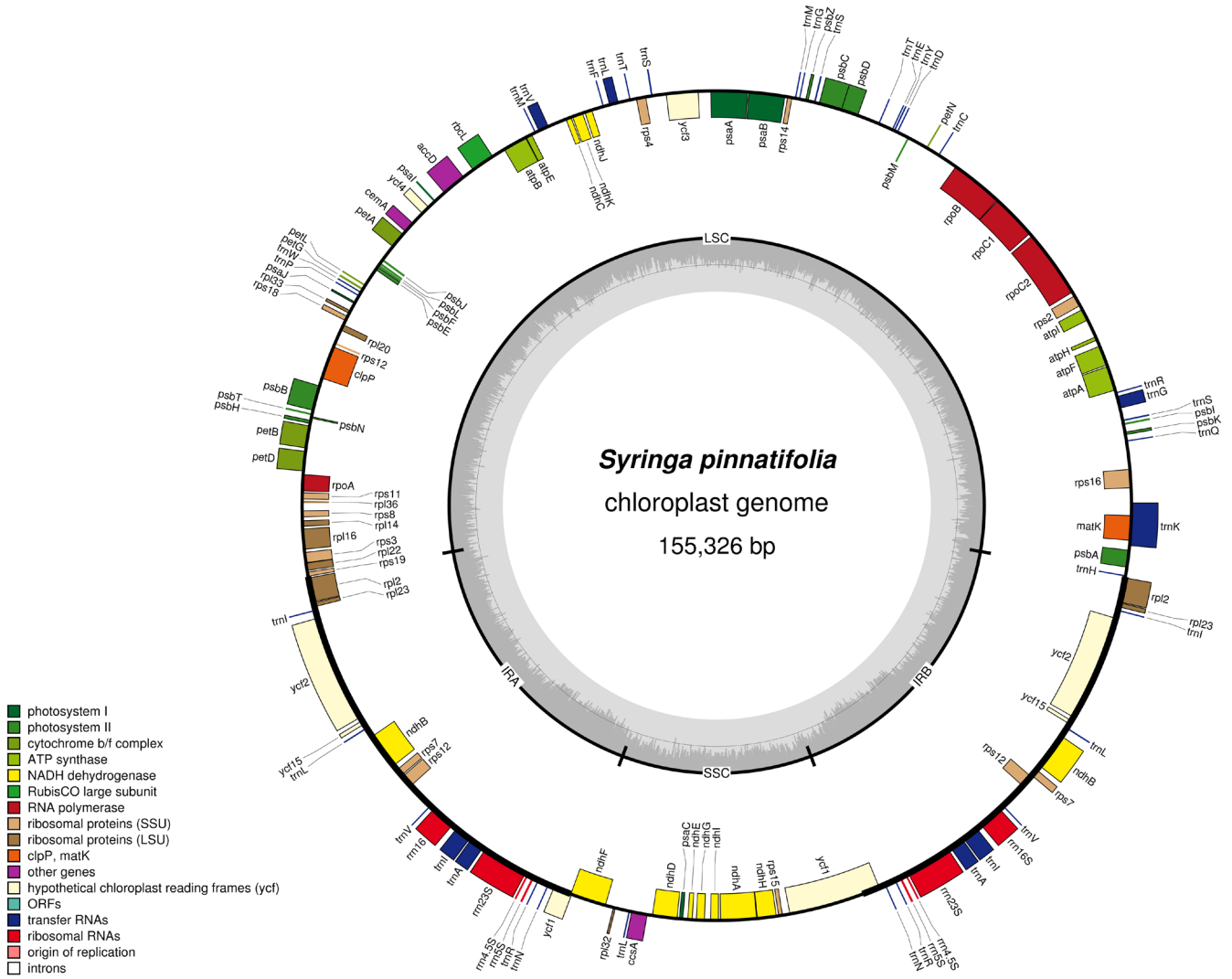


Figure 1. Gene map of the *Syringa pinnatifolia* complete chloroplast genome. Genes drawn inside the circle are transcribed clockwise, and those outside are transcribed counterclockwise. Genes belonging to different functional groups are represented by different colors. The darker gray in the inner circle corresponds to GC content, and the lighter gray corresponds to AT content.

Table 1. Comparsion of chloroplast genome size among five Oleaceae species.

| Region | Features | *S. pinnatifolia* | *S. vulgaris* | *F. suspense* | *O. europaea* | *H. palmeri* |
|---|---|---|---|---|---|---|
| LSC | length (bp) | 86 167 | 86 242 | 87 159 | 86 611 | 86 615 |
| | G+C (%) | 36.0 | 35.9 | 35.8 | 35.8 | 35.8 |
| | length (%) | 55.5 | 55.4 | 55.7 | 55.6 | 55.6 |
| SSC | length (bp) | 17 775 | 17 908 | 17 811 | 17 787 | 17 779 |
| | G+C (%) | 32.1 | 32.1 | 31.8 | 31.9 | 32.0 |
| | length (%) | 11.4 | 11.5 | 11.4 | 11.4 | 11.4 |
| IR | length (bp) | 25 692 | 25 733 | 25 717 | 25 732 | 25 713 |
| | G+C (%) | 43.2 | 43.2 | 43.2 | 43.2 | 43.2 |
| | length (%) | 16.5 | 16.5 | 16.4 | 16.5 | 16.5 |
| Total | length (bp) | 155 326 | 155 616 | 156 404 | 155 862 | 155 820 |
| | G+C (%) | 37.9 | 37.9 | 37.8 | 37.8 | 37.8 |

LSC, large single copy region; SSC, small single copy region; IR, inverted repeats.

2016) was used to estimate the maximum likelihood (ML) phylogenetic tree. Bootstrap analysis was executed with 1000 replicates and tree bisection and reconnection (TBR) branch swapping. We used 1000 replicates and TBR branch exchange to complete the bootstrap analysis.

## Results and discussion

### Features of *Syringa pinnatifolia* cpDNA

Illumina paired-end sequencing generated 1.56 Gb raw reads for *Syringa pinnatifolia*, After assembly, the complete cp genome sequence of *S. pinnatifolia* was obtained and submitted to the NCBI database with the GenBank accession number MG917095. The complete cp genome was 155 326 bp long and composed of a large single copy (LSC) region of 86 167 bp, a small single copy (SSC) region of 17 775 bp, and two inverted repeats (IRa/IRb) of 25 692 bp (Fig. 1, Table 1). The gene content, order and orientation of the *S. pinnatifolia* cp genome were similar to that of other Oleaceae plants (Wang et al. 2017). The GC content of the cpDNA was 37.9%, and the corresponding values in the LSC, SSC and IR regions were 36.0, 32.1 and 43.2% respectively.

The *S. pinnatifolia* chloroplast genome contained 113 different genes, including 79 protein-coding genes, 30 tRNA genes and four rRNA genes (Table 2). In addition, 17 genes including six protein-coding, seven tRNA and four rRNA genes were duplicated in the IR regions. The LSC region contained 60 protein-coding and 22 tRNA genes, and the SSC region contained 11 protein-coding and one tRNA genes. Among the 113 genes, two genes crossed adjacent

Table 2. List of genes found in the *Syringa pinnatifolia* chloroplast genome.

| Functional category | Group of genes | Name of genes |
|---|---|---|
| Self-replication | tRNA genes | *trnA-UGC*[*a], trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnfM-CAU, trnG-UCC*, trnG-GCC, trnH-GUG, trnI-CAU*[a], trnI-GAU*[*a], trnK-UUU*, trnL-UAA*[*a], trnL-UAG, trnL-CAA, trnM-CAU, trnN-GUU*[a], trnP-UGG, trnQ-UUG, trnR-UCU, trnR-ACG*[a], trnS-GCU, trnS-UGA, trnS-GGA, trnT-GGU, trnT-UGU, trnV-UAC*[*a], trnV-GAC, trnW-CCA, trnY-GUA* |
| | rRNA genes | *rrn16*[a], *rrn23*[a], *rrn4.5*[a], *rrn5*[a] |
| | ribosomal proteins (LSU) | *rpl2*[*a], rpl14, rpl16*, rpl20, rpl22, rpl23*[a], *rpl32, rpl33, rpl36* |
| | ribosomal proteins (SSU) | *rps2, rps3, rps4, rps7*[a], *rps8, rps11, rps12, rps14, rps15, rps16*[**], *rps18, rps19* |
| | DNA-dependent RNA polymerase | *rpoA, rpoB, rpoC1*, rpoC2* |
| Photosynthesis | photosystem I | *psaA, psaB, psaC, psaI, psaJ* |
| | photosystem II | *psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ* |
| | NADH dehydrogenase | *ndhA*, ndhB*[*a], ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK* |
| | cytochrome b/f complex | *petA, petB*, petD*, petG, petL, petN* |
| | ATP synthase | *atpA, atpB, atpE, atpF*, atpH, atpI* |
| | large subunit of rubisco | *rbcL* |
| Other genes | maturase | *matK* |
| | envelop membrane protein | *cemA* |
| | subunit of acetyl-CoA-carboxylase | *accD* |
| | c-type cytochrome synthesis gene | *ccsA* |
| | protease | *clpP*** |
| | conserved open reading frames | *ycf1, ycf2*[a], *ycf3*, ycf4, ycf15*[a] |

* Genes containing one introns; ** genes containing two introns; [a] duplicated genes (genes present in the IR regions).

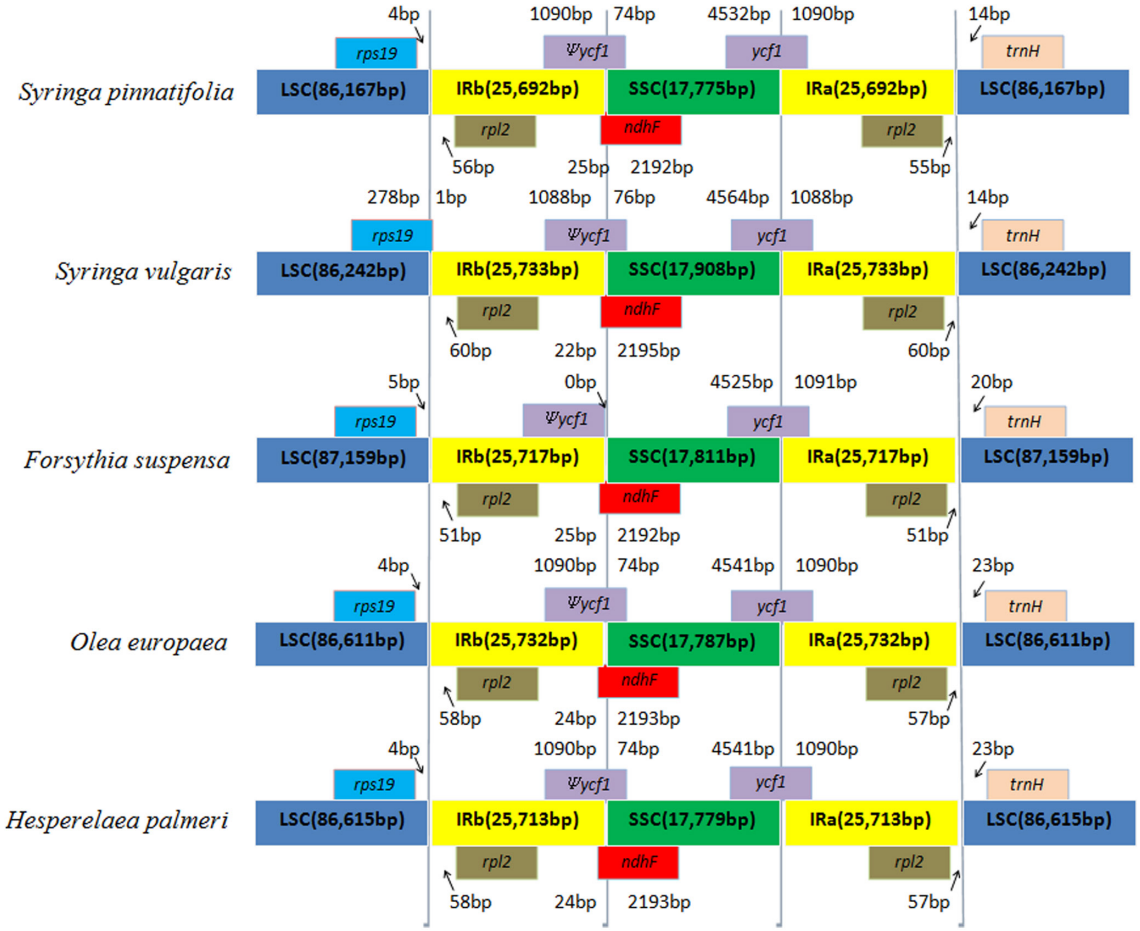


Figure 2. Comparison of the LSC/IRb/SSC/IRa junctions among the chloroplast genomes of five Oleaceae species. Colored boxes represent the adjacent border genes. Numbers above the gene indicates the distance in bp between the ends of genes and the junction sites. Ψ indicates a pseudogene. These features are not to scale.

regions: the *ycf1* gene crossed the IRb/SSC and SSC/IRa junction, and *ndhF* crossed the IRb/SSC junction. A total of 15 genes (*rps16*, *atpF*, *rpoC1*, *petB*, *petD*, *rpl2*, *rpl16*, *ndhA*, *ndhB*, *trnK-UUU*, *tanG-UCC*, *trnL-UAA*, *trnV-UAC*, *trnI-GAU* and *trnA-UGC*) contained one intron and three genes (*ycf3*, *clpP* and *rps16*) contained two introns. The protein-coding region of *S. pinnatifolia* accounted for the highest proportion of the whole cp genome, which was 49%, followed by gene spacers and introns accounting for 46%, tRNA accounting for 5% and rRNA accounting for 1%.

## IR expansion and contraction

The chloroplast genome sequences of angiosperms are highly conserved, but maintain subtle differences (Kim and Lee 2004). Contraction and expansion of the borders of the IR regions are the main reason for differences in chloroplast genome structure (Goulding et al. 1996, Ravi et al. 2008). Therefore, we compared the boundaries of the LSC/IRb/SSC/IRa junctions of five Oleaceae cp genomes, viz *Syringa vulgaris*, *S. pinnatifolia*, *Forsythia suspense*, *Olea europaea* and *Hesperelaea palmeri* (Fig. 2).

The SSC/IRa junction was located in the *ycf1* genes of all Oleaceae species cp genomes but extended for different lengths (*S. pinnatifolia*, 4532; *S. vulgaris*, 4564; *F. suspense*, 4525; *O. europaea*, 4541; *H. palmeri*, 4541) into the SSC region. The IRb/SSC border extended into the *ycf1* genes to create long *ycf1* pseudogenes, and overlapped with the *ndhF* gene in all five Oleaceae cp genomes. The *rps19* genes of four Oleaceae species were all completely located in the LSC region, except 1 bp of the *S. vulgaris rps19* gene that crossed into the IRb region. The *trnH* genes were located in the SSC region, and their distance from the LSC/IRa junction was 14–23 bp. Taken together, the LSC/IRb/SSC/IRa junctions of the Oleaceae cp genomes were similar.

## Codon usage analysis

Codons play the important role of transmitting genetic information in organisms as they link nucleic acids and proteins. Codon usage bias means that each gene of a species has its own preferred codon for the same amino acid (Wu et al. 2007). Many factors in the evolution of a species affect codon bias, which is the result of gene mutation and

Table 3. The relative synonymous codon usage in the *Syringa pinnatifolia* chloroplast genome.

| Amino acid | Codon | No. | RSCU* | Proportion (%) | Amino acid | Codon | No. | RSCU* | Proportion (%) |
|---|---|---|---|---|---|---|---|---|---|
| Phe | UUU | 1041 | 1.15 | 7.18 | Gln | CAA | 530 | 1.34 | 3.14 |
| Phe | UUC | 773 | 0.85 | | Gln | CAG | 263 | 0.66 | |
| Leu | UUA | 584 | 1.24 | 11.21 | Asn | AAU | 849 | 1.38 | 4.86 |
| Leu | UUG | 616 | 1.30 | | Asn | AAC | 380 | 0.62 | |
| Leu | CUU | 518 | 1.10 | | Lys | AAA | 1033 | 1.35 | 6.07 |
| Leu | CUC | 333 | 0.71 | | Lys | AAG | 501 | 0.65 | |
| Leu | CUA | 462 | 0.98 | | Asp | GAU | 534 | 1.46 | 2.90 |
| Leu | CUG | 320 | 0.68 | | Asp | GAC | 198 | 0.54 | |
| Ile | AUU | 916 | 1.17 | 9.30 | Glu | GAA | 662 | 1.44 | 3.64 |
| Ile | AUC | 606 | 0.77 | | Glu | GAG | 257 | 0.56 | |
| Ile | AUA | 830 | 1.06 | | Cys | UGU | 343 | 1.17 | 2.32 |
| Met | AUG | 539 | 1.00 | 2.13 | Cys | UGC | 243 | 0.83 | |
| Val | GUU | 483 | 1.41 | 5.43 | Stop | UAA | 47 | 1.69 | 0.33 |
| Val | GUC | 207 | 0.60 | | Stop | UAG | 13 | 0.46 | |
| Val | GUA | 457 | 1.33 | | Stop | UGA | 24 | 1.85 | |
| Val | GUG | 225 | 0.66 | | Trp | UGG | 364 | 1.00 | 1.44 |
| Pro | CCU | 356 | 1.04 | 5.41 | Ser | UCU | 607 | 1.46 | 9.86 |
| Pro | CCC | 315 | 0.92 | | Ser | UCC | 468 | 1.13 | |
| Pro | CCA | 507 | 1.48 | | Ser | UCA | 527 | 1.27 | |
| Pro | CCG | 189 | 0.55 | | Ser | UCG | 314 | 0.76 | |
| Thr | ACU | 315 | 1.10 | 4.55 | Ser | AGU | 316 | 0.76 | |
| Thr | ACC | 294 | 1.02 | | Ser | AGC | 261 | 0.63 | |
| Thr | ACA | 363 | 1.26 | | Arg | CGU | 183 | 0.67 | 6.47 |
| Thr | ACG | 178 | 0.62 | | Arg | CGC | 119 | 0.44 | |
| Ala | GCU | 313 | 1.43 | 3.47 | Arg | CGA | 302 | 1.11 | |
| Ala | GCC | 190 | 0.87 | | Arg | CGG | 206 | 0.76 | |
| Ala | GCA | 277 | 1.26 | | Arg | AGA | 552 | 2.02 | |
| Ala | GCG | 97 | 0.44 | | Arg | AGG | 274 | 1.00 | |
| Tyr | UAU | 566 | 1.27 | 3.52 | Gly | GGU | 319 | 1.14 | 4.38 |
| Tyr | UAC | 323 | 0.73 | | Gly | GGC | 167 | 0.60 | |
| His | CAU | 417 | 1.40 | 2.35 | Gly | GGA | 405 | 1.44 | |
| His | CAC | 178 | 0.60 | | Gly | GGG | 231 | 0.82 | |

selection (Wong et al. 2002). Relative synonymous codon usage (RSCU) refers to the relative probability of a synonymous codon for a particular codon that encodes the corresponding amino acid. If the use of a codon is not biased, the RSCU value is 1. RSCU value >1 indicates that the codon is used more frequently than expected, and vice versa (Sharp and Li 1987).

In this study, we analyzed codon usage frequency and RSCU in *S. pinnatifolia*. All protein coding genes presented 25 280 codons in the *S. pinnatifolia* chloroplast genome. Leucine (Leu) and tryptophan (Trp) with 11.21% and 1.44%, respectively, were the most and least abundant amino acids (Table 3). The RSCU values revealed that 32 codons showed codon usage bias (RSCU values >1) in the *S. pinnatifolia* cp genome. Of these 32 codons, 29 were A/U-ending codons. Conversely, of the 29 codons with RSCU values <1, 27 were C/G-ending codons, indicating that the amino acid codons preferentially end with A/U (RSCU values >1) in the *S. pinnatifolia* cp genome. The two amino acids of methionine (Met) and tryptophan (Trp) exhibited no codon bias (RSCU = 1). Natural selection and variation direction play an important role influencing codon usage bias in the chloroplast genome (Liu and Xue 2005, Pechmann and Frydman 2013), and is also related to the structure and function of the protein encoded by the gene. In 2008, Zhou et al. reported that codons in the cp genomes of plants tend to end with A/U (Zhou et al. 2008). In this study, our analysis of codon usage in the *S. pinnatifolia* cp genome revealed similar results.

## Long repetitive sequences and simple sequence repeats analysis

Long repetitive sequences of the chloroplast genomes of five species, including *S. pinnatifolia* and four other Oleaceae species were analyzed using the REPuter program. A total of 193 repeats containing forward, palindromic, reverse and complement repeats were found in these five species (Fig. 3). Among them, 38 were detected in the *S. pinnatifolia* chloroplast genome, including 15 forward repeats, 18 palindromic repeats, four reverse repeats and one complement repeat. Most of them were 20–29 bp long and accounted for 92.1% of the total. In the *S. vulgaris*, *F. suspense*, *O. eurpaea* and *H. palmeri* cp genomes, 43, 40, 37 and 35 repeats were found, respectively (Fig. 3A–B). These results suggest that *S. pinnatifolia* is more similar to *O. eurpaea* than to the other species with respect to the number of repeats.

Simple sequence repeats (SSRs) are highly efficient molecular markers that refer to the genetic site of repeated

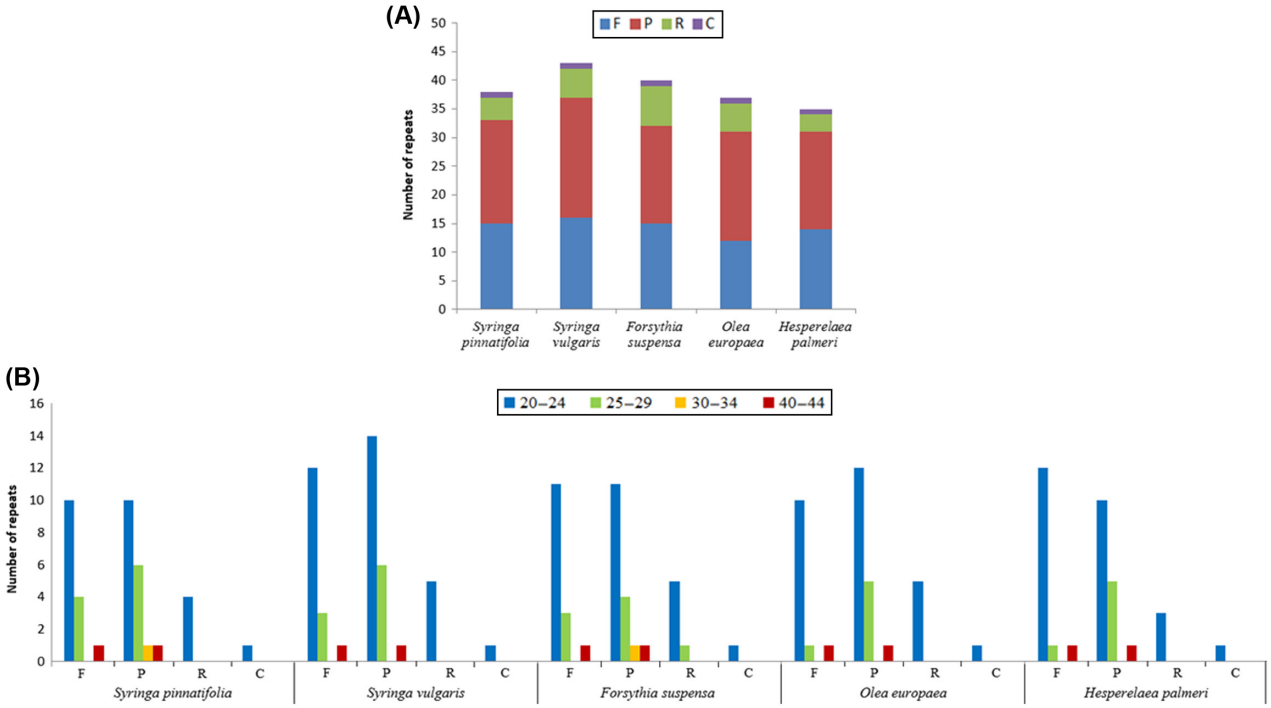


Figure 3. Long repetitive sequences of five chloroplast genomes. F, P, R and C indicate the repeat types F (forward), P (palindrome), R (reverse) and C (complement), respectively. (A) Number of identified repeats in five chloroplast genomes. (B) Number of different repeat types of different lengths in five chloroplast genomes.

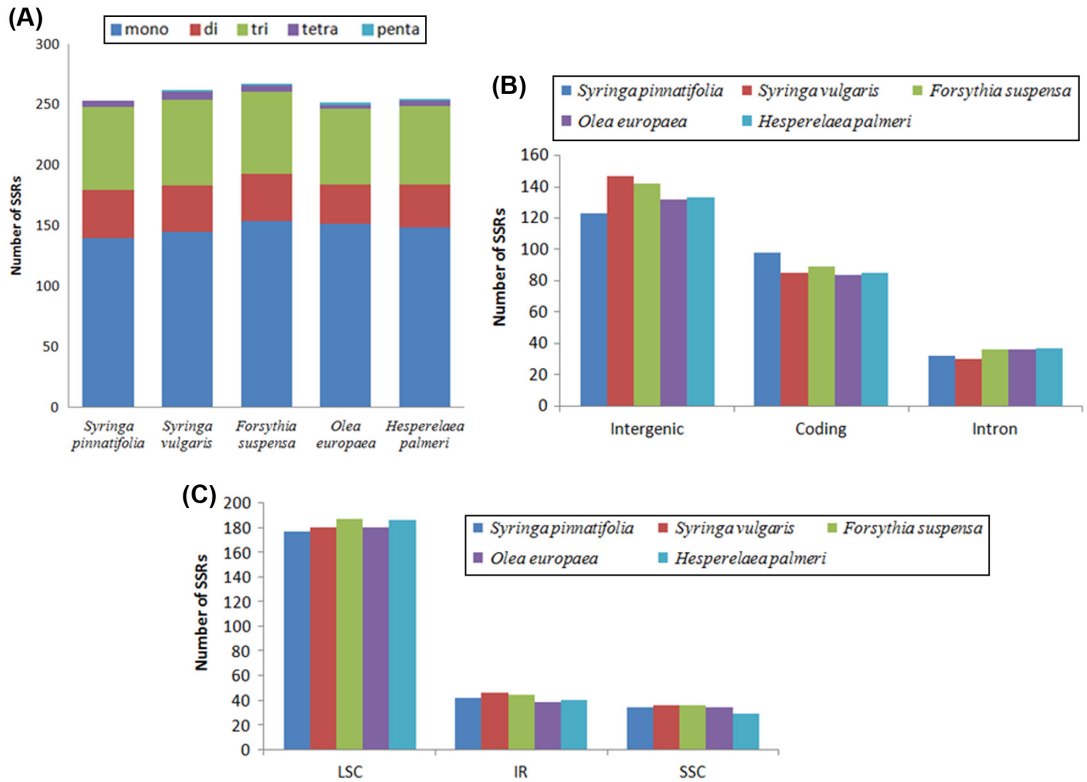


Figure 4. Analysis of simple sequence repeats (SSRs) in the five Oleaceae chloroplast genomes. (A) Number of different SSR types in five chloroplast genomes. (B) Frequency of SSRs in the protein-coding regions, intergenic regions and intron. (C) Frequency of SSRs in the LSC, SSC and IR regions.

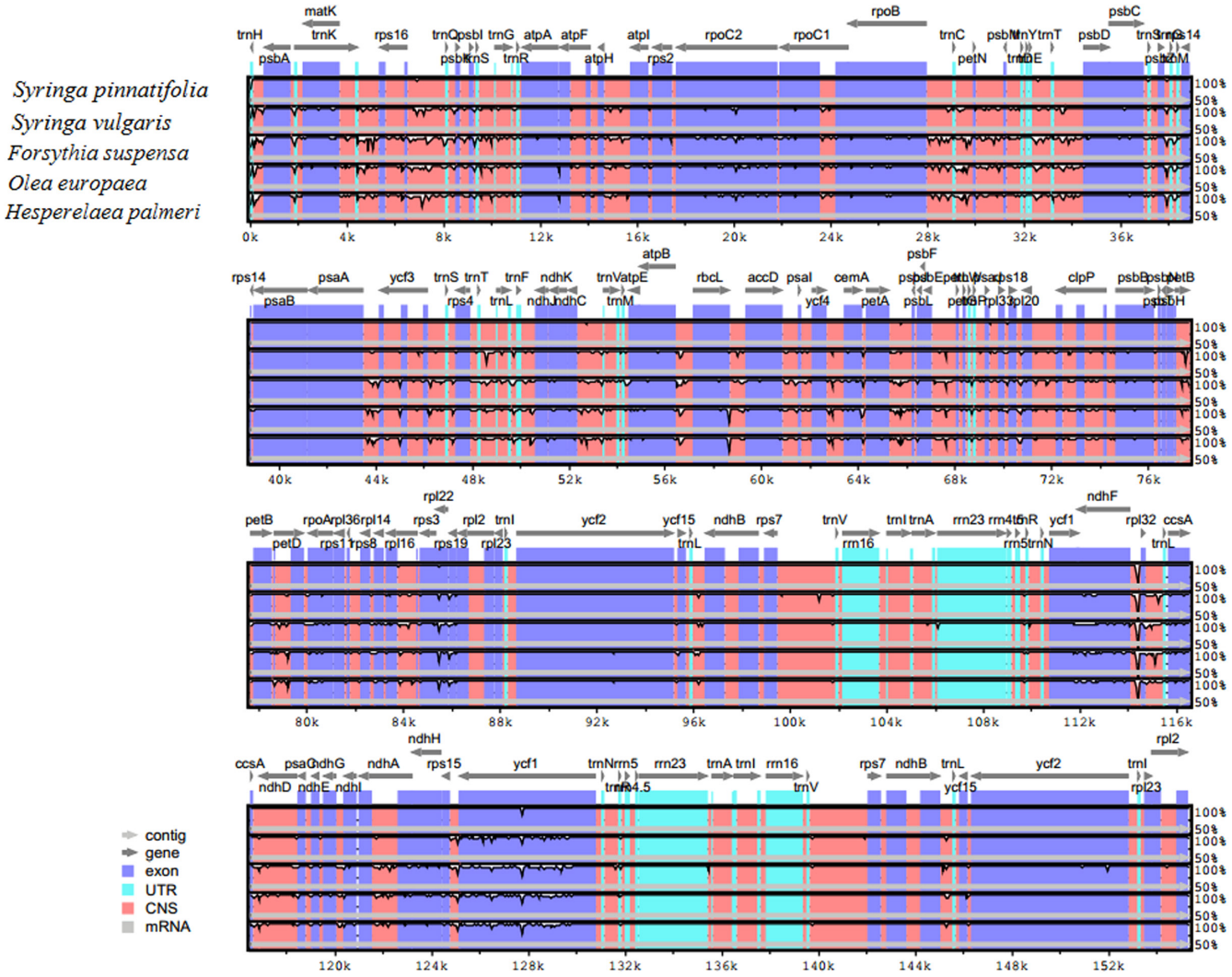


Figure 5. Comparsion of chloroplast genomes of five Oleaceae species using mVISTA. Gray arrows above the alignment indicate the direction of the gene. The dark blue regions represent exons, the light-blue regions represent untranslated regions (UTRs), the pink regions represent conserved non-coding sequences (CNS), the gray regions represent mRNA, and white peaks represent differences of genomics. The y-axis represents the percent identity ranging from 50 to 100%.

formation of short sequence nucleotides (1–6 bp). SSRs have been widely used to identify species, analyze genetic differences at population and individual levels, and for phylogenetic investigations (Kaundun and Matsumoto 2002, Jiao et al. 2012). A total of 253 SSRs were found in the *S. pinnatifolia* chloroplast genome (Fig. 4), and these may act as molecular markers in future genetic and phylogenetic studies. These included 140 mononucleotide (55.3%), 40 dinucleotide (15.8%), 68 trinucleotide (26.9%) and five tetranucleotide (2%) SSRs. Generally, tetranucleotide repeats tend to exceed the trinucleotides in number (Qian et al. 2013). However, in this study, the trinucleotides (26.9%) were second to mononucleotides (55.3%) and exceeded the tetranucleotides (2%). Among the 140 mononucleotide repeats in *S. pinnatifolia*, 135 (96.4%) were of the AT-type, Only five mononucleotides were of the GC-type, a results similar to those of other species (Kuang et al. 2011). The SSRs were non-uniformly distributed within the cp genome: 177 in the LSC, 42 in the IR and 34 in the SSC regions. Analyses of function-related location revealed that 98 were detected in protein-coding regions, 32 in introns and 123 in intergenic regions.

The numbers of SSRs in the other four Oleaceae species were 145, 154, 152 and 149 in the *S. vulgaris*, *F. suspense*, *O. eurpaea* and *H. palmeri* cp genomes, respectively (Supplementary material Appendix 1 Table A2). Among those genes, *S. pinnatifolia* had the fewest SSRs and *O. eurpaea* had the most SSRs. Additionally, only a few tetranucleotide and pentanucleotide SSRs were found in these cp genomes, and no hexanucleotide existed. The SSRs from this study can be used to conduct evolutionary and genetic diversity studies among Oleaceae species.

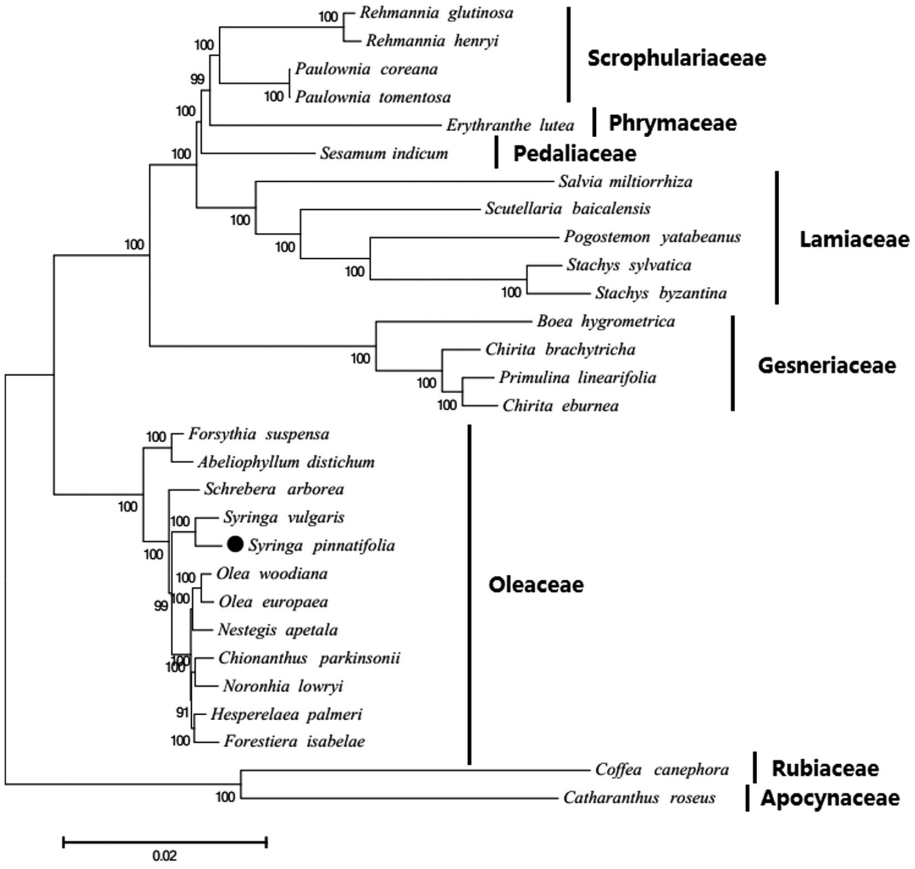


Figure 6. Phylogenetic tree of 27 Lamiales species and two outgroup species based on the maximum likelihood (ML) method based and complete chloroplast genome sequences. Numbers at nodes are values of bootstrap support.

## Comparative chloroplast genomic analysis

With the increase in the number of published complete cpDNA sequences of species, comparative analysis of chloroplast genomics has become increasingly important to examine interspecific relationships, differences and evolutionary processes (Li 2016). We compared the whole cp genome of *S. pinnatifolia* to those of *S. vulgaris*, *F. suspense*, *O. eurpaea* and *H. palmeri* using the mVISTA program (Fig. 5). The most conserved regions were the tRNA and rRNA coding regions, and the IR regions were more conserved than the LSC and SSC. Additionally, the protein-coding, tRNA and rRNA coding regions were relatively conserved, and non-coding regions exhibit higher divergence than the coding regions. The most divergent regions among the five Oleaceae cp genomes were localized in the intergenic spacers. These highly divergent regions include the *ycf1* gene and the intergenic regions of *trnH-psbA*, *trnK-rps16*, *rps16-trnQ*, *atpB-rbcL* and *ndhF-rpl32*. These regions are hypervariable regions, which can be used as specific DNA barcodes and provide useful phylogenetic information.

## Phylogenetic analysis

The rapid development of DNA sequencing technology has much increased the available plant chloroplast genome information, and this data hs been used to determine the phylogenetic relationships of major plant groups, and to detect the evolutionary rate and modes of individual genes (Bock and Knoop 2012). Therefore, the complete chloroplast genome plays an important role in the evolutionary analysis of plant systems, and has been applied to phylogenetic studies in angiosperms (Jansen et al. 2007, Moore et al. 2007).

In this study, 26 complete cp genomes of Lamiales species and two outgroup species were obtained from GenBank to determine the phylogenetic position of *S. pinnatifolia*. The data were analyzed with the maximum likelihood (ML) method to build a phylogenetic tree (Fig. 6). In the ML tree, 23 of the total 26 nodes had bootstrap values of 100%, and the remaining three nodes were >90%. The result indicates that *S. pinnatifolia* is the closest sister species to *S. vulgaris*. Notably, *Syringa* diverged relatively early from other members of the Oleaceae. In general, all 27 Lamiales species formed a lineage separate from the two outgroup species. Among the Oleaceae genera, *Syringa* is the most polymorphic and the phylogenetic distances among its members is relatively large. Indeed, some different views on the subgeneric and specific taxonomy of *Syringa* have been presented in the past. At present, in the genus *Syringa* only the cp genomes of *S. pinnatifolia* and *S. vulgaris* have been published, so the phylogenetic relationship in the genus

*Syringa* needs further study, but our results provide a basis for future phylogenetic analyses.

## Conclusions

In this study, we reported the first complete chloroplast genome of *S. pinnatifolia*. It was found to have a circular and quadripartite structure, as is common to most land plant cp genomes. We compared the *S. pinnatifolia* cp genome with other Oleaceae species, and the results revealed that gene size, gene content and organization were highly similar among four Oleaceae species. Repeat sequences and SSRs were analyzed in the *S. pinnatifolia* cp genome, and these may provide genetic information for the development of molecular markers and future research on the genetic diversity of *Syringa*. The phylogenetic analysis based on the complete cp genome indicated that the closest sister species of *S. pinnatifolia* is *S. vulgaris*. This newly characterized *S. pinnatifolia* cp genome is a useful genomic resource for phylogenetic inference and develop more genetic markers for species discrimination in the genus *Syringa*. It will provide a useful genetic resource for further study on resource conservation and genetic engineering of *S. pinnatifolia*.

*Acknowledgments* – The authors would like to express my gratitude to all those who have helped during the writing of this thesis. We express our sincere appreciation to Peng Zhao, Yiheng Hu for their valuable assistance in the field work. We would like to thank Dr. Jing Cai for her valuable comments on the manuscript.
*Funding* – This study was funded by the Special Fund for Forest Scientific Research in the Public Welfare (201204308).
*Conflicts of interest* – The authors alone are responsible for the content and writing of the paper.
*Author contributions* – JZ, JC, ZJ, HS and HZ conceived and designed the experiment; JZ performed the experiments; JZ and JC analyzed the data and wrote the manuscript. All the authors read and approved submission of the final manuscript.

## References

Bock, R. and Knoop, V. 2012. Genomics of chloroplast and mitochondria (advances in photosynthesis and respiration). – Springer, pp. 115–117.
Burie. 2012. Study on resources situation and medicinal national botany of endangered plants Helan lilac (*Syringe pinnatifolia Hemsl.* Var. alashanensis Ma et. Q. zhou) in the Helan mountains. – Asia–Pacific Ethnobotany Forum, Yin Chuan, pp. 198–202.
Chen, C. M. and Chen, L. 2014. Extraction method of chloroplast DNA of tea plant (*Camellia sinensis*). – Mol. Plant Breed. 12: 562–566.
Cui, H. X. et al. 2004. The distribution, origin and evolution of *Syringa*. – Bull. Bot. Res. 24: 141–145.
Diekmann, K. et al. 2008. An optimized chloroplast DNA extraction protocol for grasses (Poaceae) proves suitable for whole plastid genome sequencing and SNP detection. – PLoS One 3: e2813.
Frazer, K. A. et al. 2004. VISTA: computational tools for comparative genomics. – Nucleic Acids Res. 32: 273.
Goulding, S. E. et al. 1996. Ebb and flow of the chloroplast inverted repeat. – Mol. Gen. Genet. 252: 195–206.
Jansen, R. K. et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. – Proc. Natl Acad. Sci. USA 104: 19369–19374.
Jin, S. et al. 2008. Species diversity of *Syringe pinnatifolia* in Helan mountains. – J. West China For. Sci. 37: 40–44.
Jiang, Z. M. et al. 1999. The anatomical features of the seed for the precious, rare and endangered plants – *Corylus chinensis* and *Syringa pinnatifolia*. – Shaanxi For. Sci. Technol. 3: 14–16.
Jiao, Y. et al. 2012. Development of simple sequence repeat (SSR) markers from a genome survey of Chinese bayberry (*Myrica rubra*). – BMC Genomics 13: 201.
John, F. 1999. Analysis of codon usage. – PhD thesis, Univ. of Nottingham.
Kaundun, S. S. and Matsumoto, S. 2002. Heterologous nuclear and chloroplast microsatellite amplification and variation in tea, *Camellia sinensis*. – Genome 45: 1041–1048.
Kearse, M. et al. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. – Bioinformatics 28: 1647–1649.
Kim, K. J. and Lee, H. L. 2004. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng nees*) and comparative analysis of sequence evolution among 17 vascular plants. – DNA Res. 11: 247–261.
Kuroiwa, T. et al. 1982. Epifluorescent microscopic evidence for maternal inheritance of chloroplast DNA. – Nature 298: 481–483.
Kurtz, S. et al. 2001. Reputer: the manifold applications of repeat analysis on a genomic scale. – Nucleic Acids Res. 29: 4633–4642.
Kumar, S. et al. 2016. MEGA7: molecular evolutionary genetics analysis ver. 7.0 for bigger datasets. – Mol. Biol. Evol. 33: 1870–1874.
Kuang, D. Y. et al. 2011. Complete chloroplast genome sequence of magnolia kwangsiensis (magnoliaceae): implication for dna barcoding and population genetics. – Genome 54: 663–673.
Larkin, M. et al. 2007. Clustal W and Clustal X ver. 2.0. – Bioinformatics 23: 2947–2948.
Liu, J. 2012. Analysis of chloroplast genome of *Smilax china* L. and comparative studies of cpDNA genome of monocots. – PhD thesis, Zhejiang Univ.
Li, M. Z. 2016. Comparative studies on the extraction methods of cpDNA and analyses of two medical plants. – PhD thesis, Zhejiang Univ.
Li, Q. and Wan, J. M. 2005. SSRHunter: development of a local searching software for SSR sites. – Hereditas 27: 808–810.
Liu, Q. P. and Xue, Q. Z. 2005. Comparative studies on codon usage pattern of chloroplast and their host nuclear genes in four plant species. – J. Genet. 84: 55–62.
Li, R. et al. 2008. SOAP: short oligonucleotide alignment program. – Bioinformatics 24: 713–714.
Lohse, M. et al. 2013. Organellar genome DRAW-a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. – Nucleic Acids Res. 41: 575–581.

Maier, R. M. and Schmitz-Linneweber, C. 2004. Molecular biology and biotechnology of the plant organelles: chloroplasts and mitochondria. – Kluwer Academic Publisher, pp. 115–150.
Mayor, C. et al. 2000. VISTA: visualizing global DNA sequence alignments of arbitrary length. – Bioinformatics 16: 1046–1047.
Ming, J. et al. 2007. Advances in germplasm resources of *Syringa* Linn. Research. – World For. Res. 20: 20–26.
Moore, M. J. et al. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. – Proc. Natl Acad. Sci. USA 104: 19363–19368.
NCBI 2018. Organelle genome resources. – <www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=2759&opt=plastid>, accessed 18 May 2018.
Patel, R.K and Jain, M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. –PLoS One 7: e30619.
Pechmann, S. and Frydman, J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. – Nat. Struct. Mol. Biol. 20: 237–243.
Qian, J. et al. 2013. The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. – PLoS One 8: e57607.
Ravi, V. et al. 2008. An update on chloroplast genomes. – Plant Syst. Evol. 271: 101–122.
Schattner, P. et al. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. – Nucleic Acids Res. 33: 686–689.

Supplementary material (available online as Appendix njb-02201 at <www.nordicjbotany.org/appendix/njb-02201>). Appendix 1.

Sharp, P. M. and Li, W. H. 1987. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. – Nucleic Acids Res. 15: 1281–1295.
Su, G. Z. et al. 2015a. Phytochemical and pharmacological progress on peeled stem of *Syringe pinnatifolia*, a Mongolian medicine. – China J. Chin. Mater. Med. 40: 4334–4338.
Su, G. Z. et al. 2015b. Phytochemical and pharmacological progress on the genus *Syringa*. – Chem. Cent. J. 9: 1–12.
Wang, W. B. et al. 2017. The complete chloroplast genome sequences of the medicinal plant *Forsythia suspensa* (Oleaceae). – Int. J. Mol. Sci. 18: 2288.
Wicke, S. et al. 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. – Plant Mol. Biol. 76: 273–297.
Wong, G. K. S. et al. 2002. Compositional gradients in Gramineae genes. – Genome Res. 12: 851–856.
Wu, X. M. et al. 2007. The analysis method and progress in the study of codon bias. – Hereditas 29: 420–426.
Wyman, S. K. et al. 2004. Automatic annotation of organellar genomes with DOGMA. – Bioinformatics 20: 3252–3255.
Zang, S. Y. and Liu, G. X. 1999. Lilac. – China Forestry Publishing House.
Zhang, Y. J. and Li, D. Z. 2011. Advances in phylogenomics based on complete chloroplast genomes. – Plant Divers. 33: 365–375.
Zhou, M. et al. 2008. Patterns of synonymous codon usage bias in chloroplast genomes of plants. – For. Ecosys. 10: 235–242.



View publication stats