RESEARCH ARTICLE

The Complete Chloroplast Genome of Ye-Xing-Ba (*Scrophularia dentata*; Scrophulariaceae), an Alpine Tibetan Herb

Lianghong Ni1, Zhili Zhao1 *, Gaawe Dorje2, Mi Ma2

1 School of Pharmacy, Shanghai University of Traditional Chinese Medicine, Shanghai, China, 2 Tibetan Traditional Medical College, Lhasa, China

* zhilzhao@163.com

Abstract

*Scrophularia dentata* is an important Tibetan medicinal plant and traditionally used for the treatment of exanthema and fever in Traditional Tibetan Medicine (TTM). However, there is little sequence and genomic information available for *S. dentata*. In this paper, we report the complete chloroplast genome sequence of *S. dentata* and it is the first sequenced member of the Sect. Tomiophyllum within *Scrophularia* (Scrophulariaceae). The gene order and organization of the chloroplast genome of *S. dentata* are similar to other Lamiales chloroplast genomes. The plastome is 152,553 bp in length and includes a pair of inverted repeats (IRs) of 25,523 bp that separate a large single copy (LSC) region of 84,058 bp and a small single copy (SSC) region of 17,449 bp. It has 38.0% GC content and includes 114 unique genes, of which 80 are protein-coding, 30 are transfer RNA, and 4 are ribosomal RNA. Also, it contains 21 forward repeats, 19 palindrome repeats and 41 simple sequence repeats (SSRs). The repeats and SSRs within *S. dentata* were compared with those of *S. takesimensis* and present certain discrepancies. The chloroplast genome of *S. dentata* was compared with other five publicly available Lamiales genomes from different families. All the coding regions and non-coding regions (introns and intergenic spacers) within the six chloroplast genomes have been extracted and analysed. Furthermore, the genome divergent hotspot regions were identified. Our studies could provide basic data for the alpine medicinal species conservation and molecular phylogenetic researches of Scrophulariaceae and Lamiales.

Introduction

Ye-Xing-Ba (Tibetan name), a common Tibetan herb, is traditionally used for the treatment of exanthema and fever in Traditional Tibetan Medicine (TTM). As a medicinal alpine species [1], *Scrophularia dentata* Royle ex Benth. (Scrophulariaceae) is one of the original plants of the Tibetan herb [2]. In recent years, our research team has been concentrating on the ethnobotanical survey of Tibetan herbs since 1992 [3, 4]. Based on our specimen collection and taxonomic identification, some chemical and pharmacological studies on *S. dentata* have been conducted by our team [5, 6, 7]. Because of the harsh living environment and increasing demand for medicinal use of the alpine plant, it is of utmost urgency to develop a conservation strategy for...
the species. Also, selection of excellent germplasm requires more efficient genetic data and molecular markers [8]. However, there is little sequence and genomic information available for the species in the GenBank, except for nrDNA ITS sequences, some chloroplast and mitochondrial segment sequences submitted by our team (http://www.ncbi.nlm.nih.gov/nuccore/?term=Scrophularia+dentata).

Chloroplast contains its own independent genome encoding a specific set of proteins [9]. The plastid genomes typically are composed of a large single copy (LSC) region and a small single copy (SSC) region, which are separated by two copies of inverted repeats (IRs) [10]. The small size of the chloroplast genome, about 115 to 165 kb in length, makes it suitable for complete sequencing and the data can be further applied to DNA bar coding and phylogeny construction [11, 12]. As complete chloroplast genome sequences contain sufficient information, sequencing of whole chloroplast genomes is essential for the analyses of plant species [13, 14].

*S. dentata* belongs to Scrophulariaceae, which is a large family of Lamiales, consisting of over 3,000 species [15]. Also, the genus *Scrophularia* has ca. 200 species and is divided into two sections, i.e. Sect. Scrophularia and Sect. Tomiophyllum [16]. However, the complete chloroplast genome of *Scrophularia takesimensis* within Sect. Scrophularia is the first published whole sequence in *Scrophularia*. This is the first sequenced member of the Sect. Tomiophyllum in the paper. Details of the *S. dentata* chloroplast genome structure and organization are reported and compared with previously annotated chloroplast genomes of other Lamiales species. Our studies could expand understanding of the diversity of *Scrophularia* cp genomes and provide basic data for the alpine medicinal species conservation and molecular phylogenetic researches of Scrophulariaceae and Lamiales.

**Materials and Methods**

**Plant material**

Samples of *S. dentata* were collected in Lhasa, Tibet, China. The voucher specimens were deposited in the herbarium of Shanghai University of Traditional Chinese Medicine (field number: XZ201416). Also, the location of the specimens is not within any protected area.

**DNA extraction, genome sequencing and validation**

Total chloroplast DNA was extracted from 100 g of fresh leaves using a sucrose gradient centrifugation method improved by Li et al. [17]. The genome was sequenced and assembled on an Illumina MiSeq platform following Gogniashvili et al. [18]. The four junctions between the SSC/LSC and IRs were validated by PCR amplification and Sanger sequencing, and nine other fragments were selected to validate the genome sequence further (S1 Table) [19].

**Genome annotation**

The chloroplast genome was annotated using the annotator DOGMA [20], coupled with manual corrections. For all of the tRNA genes, the annotations were verified using tRNAscan-SE1.21 [21]. The genome map was drawn by GenomeVX [22].

**Genome comparison**

The complete cp genomes of *S. dentata* and five other Lamiales species were compared using the mVISTA program [23]. The boundaries between IR and SC regions of these species were also compared and analysed.
Repeat structure
The REPuter program [24] was used to identify repeats (forward, palindrome, complement and reverse sequences). The size and identity of the repeats were limited to no less than 30 bp and 90%, respectively, with Hamming distance equal to 3 [25, 26]. Meanwhile, simple sequence repeats (SSRs) were detected using MISA [27] by setting the minimum number of repeats to 10, 5, 4, 3, 3 and 3 for mono-, di-, tri-, tetra-, penta- and hexanucleotides. The repeats and SSRs of \textit{S. dentata} were compared with those of \textit{S. takesimensis}, the only one complete chloroplast genome available in Scrophulariaceae.

Results and Discussion

Genome organization and features
The junction regions between IRs and SSC/LSC and nine additional regions were confirmed by PCR amplification and Sanger sequencing. We compared these sequences to the assembled genome and no mismatch or indel was observed, which validated the accuracy of genome sequencing and assembly. The chloroplast genome sequence of \textit{S. dentata} has been submitted to GenBank (accession number: KT428154).

The complete chloroplast genome of \textit{S. dentata} has a total length of 152,553 bp, with a pair of inverted repeats (IRs) of 25,523 bp that separate a large single copy (LSC) region of 84,058 bp and a small single copy (SSC) region of 17,449 bp (Fig 1). The total GC content is 38.0%, which is similar to the published asterids cp genomes [28, 29]. And the GC content is unevenly distributed in the genome. The GC content of IRs (43.1%) is higher than that of LSC and SSC region (36.0 and 32.29%). The high GC content in the IR regions is due to the increased presence of GC nucleotides in the four rRNA genes: \textit{rrn5}, \textit{rrn4.5}, \textit{rrn23} and \textit{rrn16}, which is congruent to what has been found in other chloroplast genomes [30, 31].

The chloroplast genome of \textit{S. dentata} encodes a total of 114 unique genes, of which 18 are duplicated in IR regions. Among the 114 genes, there are 80 protein-coding genes (70.2%), 30 transfer genes (26.3%) and 4 rRNA genes (3.5%) (Table 1). 18 genes contain introns, 15 (nine protein-coding and six tRNA genes) of which contain one intron and three (\textit{clpP}, \textit{ycf3} and \textit{rps12}) contain two introns (Table 2). The \textit{rps12} gene is a trans-spliced gene, with the first exon located in LSC region and the other two exons duplicated in IR regions.

Protein-coding regions account for 59.6% of the whole genome, while tRNA and rRNA regions account for 4.6% and 5.9%, respectively. The remaining regions are noncoding sequences, including intergenic spacers, introns and pseudogenes. There are two pseudogenes identified: \textit{ycf1} and \textit{rps19}. They are located in the boundary regions between IRb/SSC and IRA/LSC, respectively. The lack of their protein-coding ability is due to the partial gene duplication.

Comparison to other Lamiales species
The chloroplast genome of \textit{S. dentata} was compared with other five publicly available Lamiales genomes from different families, i.e., \textit{Boea hygrometrica} (Gesneriaceae), \textit{Hesperelaea palmeri} (Oleaceae), \textit{Salvia miltiorrhiza} (Lamiaceae), \textit{Andrographis paniculata} (Acanthaceae) and \textit{Sesamum indicum} (Pedaliaceae) (Table 3). Organization of the Lamiales chloroplast genome is conserved; neither translocations nor inversions were detected in the analyses. However, there are differences in terms of genome size. The \textit{A. paniculata} chloroplast genome is the shortest (150,249 bp), while that of \textit{H. palmeri} (155,820 bp) is longer than other species. The genome size variation is attributed mainly to the difference in the length of the LSC region, similar to the features in Asteraceae chloroplast genomes [36]. Also, the average size of the six Lamiales chloroplast genomes is 152,794 bp.
The overall sequence identity of the six Lamiales chloroplast genomes was plotted using the software mVISTA with S. dentata as reference (Fig 2). The results show that the chloroplast genomes within Lamiales are conservative, although divergent regions could be detected. As expected, the IR region is more conserved than the LSC and SSC regions in these species. Meanwhile, non-coding regions reveal a higher divergence than coding regions. In order to check divergent hotspot regions further, all the coding regions and non-coding regions (introns and intergenic spacers) within the six chloroplast genomes were extracted and analysed (S2 Table, Fig 3). The most divergent regions are localized in the intergenic spacers, including ccsA-ndhD, rps16-trnQ-UGU, ndhf-ndhK, ndhE-ndhG, ndhC-trnV-UAC, trnH-GUG-psbA and ndhG-ndhI. These intergenic regions could be used in assessing phylogenetic relationships within Lamiales species. For the coding regions, the most divergent regions are ycf1, matK, rpl22, rpl32, rps15 and ccsA. These genes are all located in single copy regions.

The expansion and contraction at the borders of IR/SC are common in chloroplast genomes, which is the main reason for the size variation within angiosperm chloroplast genomes [37]. The junctions of LSC/IRb/SSC/IRa of the six Lamiales chloroplast genomes, which is the main reason for the size variation within angiosperm chloroplast genomes, range from 25,141 bp to 25,712 bp, some differences in IR expansions and contractions were observed.

Table 1. Genes present within *Scrophularia dentata* chloroplast genome.

<table>
<thead>
<tr>
<th>Group of genes</th>
<th>Gene names</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Photosystem I</td>
<td>psaA, psaB, psaC, psal, psAJ</td>
</tr>
<tr>
<td>3 Cytochrome b/f complex</td>
<td>petA, petB*, petD*, petG, petL, petN</td>
</tr>
<tr>
<td>4 ATP synthase</td>
<td>atpA, atpB, atpE, atpF*, atpH, atpI</td>
</tr>
<tr>
<td>5 NADH dehydrogenase</td>
<td>ndhA*, ndhB*(x2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhK</td>
</tr>
<tr>
<td>6 RubisCO large subunit</td>
<td>rbcL</td>
</tr>
<tr>
<td>7 RNA polymerase</td>
<td>rpoA, rpoB, rpoC1*, rpoC2</td>
</tr>
<tr>
<td>8 Ribosomal proteins (SSU)</td>
<td>rps2, rps3, rps4, rps7(x2), rps8, rps11, rps12**, rps14, rps15, rps16, rps18, rps19</td>
</tr>
<tr>
<td>9 Ribosomal proteins (LSU)</td>
<td>rpl2*(x2), rpl14, rpl16*, rpl20, rpl22, rpl23(x2), rpl32, rpl33, rpl36</td>
</tr>
<tr>
<td>10 Other genes</td>
<td>cipP**, matK, accD, ccsA, infA, cemA</td>
</tr>
<tr>
<td>11 Proteins of unknown function</td>
<td>ycf1, ycf2(x2), ycf3**, ycf4, ycf15(x2)</td>
</tr>
<tr>
<td>12 Ribosomal RNAs</td>
<td>rnr4.5(x2), rnr5(x2), rnr16(x2), rnr23(x2)</td>
</tr>
</tbody>
</table>

*Gene containing one intron.
**Gene containing two introns.
(x2) Two gene copies in the IRs.

doi:10.1371/journal.pone.0158488.g001
The rps19 genes of *H. palmeri* and *A. paniculata* are located in the LSC region completely, and the IR region is expanded to rps19 gene in the other four genomes, with rps19 pseudogene created at the IRa/LSC border. In *S. dentata*, a short rps19 pseudogene of 40 bp is created. The same pseudogene is 30 bp in *S. indicum*, 43 bp in *S. miltiorrhiza* and 37 bp in *B. hygrometrica*, respectively. The border between IRb and SSC extends into the ycf1 genes, with ycf1 pseudogenes created in all of the six species. The length of ycf1 pseudogene is 910 bp in *S. dentata*, 1,012 bp in *S. indicum*, 1,091 bp in *H. palmeri*, 983 bp in *A. paniculata*, 1,056 bp in *S. miltiorrhiza* and 816 bp in *B. hygrometrica*, respectively. Overlaps are detected between the ycf1 pseudogene and ndhF gene in *H. palmeri*, *A. paniculata*, *S. miltiorrhiza* and *B. hygrometrica*, with lengths of 25 bp, 40 bp, 32 bp and 118 bp, respectively. The trnH-GUG genes are all located in the LSC region, with 4–57 bp apart from the LSC/IRa boundary.

**Table 2. Genes with introns within Scrophularia dentata chloroplast genome and the length of exons and introns.**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Location</th>
<th>Exon I (bp)</th>
<th>Intron I (bp)</th>
<th>Exon II (bp)</th>
<th>Intron II (bp)</th>
<th>Exon III (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>petB</td>
<td>LSC</td>
<td>6</td>
<td>720</td>
<td>642</td>
<td></td>
<td></td>
</tr>
<tr>
<td>petD</td>
<td>LSC</td>
<td>8</td>
<td>735</td>
<td>475</td>
<td></td>
<td></td>
</tr>
<tr>
<td>atpF</td>
<td>LSC</td>
<td>144</td>
<td>723</td>
<td>411</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ndhA</td>
<td>SSC</td>
<td>552</td>
<td>1067</td>
<td>531</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ndhB</td>
<td>IR</td>
<td>777</td>
<td>679</td>
<td>756</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rpoC1</td>
<td>LSC</td>
<td>435</td>
<td>756</td>
<td>1626</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rps12*</td>
<td>LSC</td>
<td>114</td>
<td>-</td>
<td>232</td>
<td>537</td>
<td>26</td>
</tr>
<tr>
<td>rps16</td>
<td>LSC</td>
<td>42</td>
<td>848</td>
<td>225</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rpl2</td>
<td>IR</td>
<td>393</td>
<td>664</td>
<td>435</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rpl16</td>
<td>LSC</td>
<td>9</td>
<td>886</td>
<td>399</td>
<td></td>
<td></td>
</tr>
<tr>
<td>clpP</td>
<td>LSC</td>
<td>69</td>
<td>727</td>
<td>291</td>
<td>626</td>
<td>228</td>
</tr>
<tr>
<td>ycf3</td>
<td>LSC</td>
<td>129</td>
<td>702</td>
<td>228</td>
<td>720</td>
<td>153</td>
</tr>
<tr>
<td>trnA-UGC</td>
<td>IR</td>
<td>38</td>
<td>813</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trnG-UCC</td>
<td>LSC</td>
<td>23</td>
<td>683</td>
<td>47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trnL-GAU</td>
<td>IR</td>
<td>42</td>
<td>943</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trnK-UUU</td>
<td>LSC</td>
<td>37</td>
<td>2517</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trnL-UAA</td>
<td>LSC</td>
<td>37</td>
<td>488</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trnV-UAC</td>
<td>LSC</td>
<td>38</td>
<td>585</td>
<td>37</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The rps12 is a trans-spliced gene with the 5’ end located in the LSC region and the duplicated 3’ end in the IR regions.

doi:10.1371/journal.pone.0158488.t002

**Table 3. Comparison of the base composition of six Lamiales chloroplast genomes.**

<table>
<thead>
<tr>
<th></th>
<th>Scrophularia dentata</th>
<th>Boea hygrometrica</th>
<th>Hesperelaea palmeri</th>
<th>Salvia miltiorrhiza</th>
<th>Andrographis paniculata</th>
<th>Sesamum indicum</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSC</td>
<td>84,058 bp</td>
<td>84,692 bp</td>
<td>86,616 bp</td>
<td>82,695 bp</td>
<td>82,459 bp</td>
<td>85,170 bp</td>
</tr>
<tr>
<td>SSC</td>
<td>17,449 bp</td>
<td>17,901 bp</td>
<td>17,780 bp</td>
<td>17,555 bp</td>
<td>17,190 bp</td>
<td>17,872 bp</td>
</tr>
<tr>
<td>IR</td>
<td>25,523 bp</td>
<td>25,450 bp</td>
<td>25,712 bp</td>
<td>25,539 bp</td>
<td>25,300 bp</td>
<td>25,141 bp</td>
</tr>
<tr>
<td>Total</td>
<td>152,553 bp</td>
<td>153,493 bp</td>
<td>155,820 bp</td>
<td>151,328 bp</td>
<td>150,249 bp</td>
<td>153,324 bp</td>
</tr>
<tr>
<td>CG content</td>
<td>38.0%</td>
<td>37.6%</td>
<td>37.8%</td>
<td>36.2%</td>
<td>38.3%</td>
<td>38.2%</td>
</tr>
<tr>
<td>Coding gene</td>
<td>80</td>
<td>86</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>rRNA</td>
<td>30</td>
<td>29</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>tRNA</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>References</td>
<td>this study</td>
<td>[32]</td>
<td>[33]</td>
<td>[34]</td>
<td>NC_022451</td>
<td>[35]</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0158488.t003
Repeat and SSR analysis

Repeat sequences in the *S. dentata* chloroplast genome were analysed by REPuter and the results showed that 21 forward repeats and 19 palindrome repeats were at least 30 bp long per repeat unit with a sequence identity greater than 90% (Table 4). The complement repeats and reverse repeats had not been detected. Of all the repeats found, 33 repeats (82.5%) are 30 to 39 bp long, 7 repeats (17.5%) are 40 to 49 bp long, and the longest repeat is 44 bp. We present a comparative analysis of repeats between *S. dentata* and *S. takesimensis*, the only one complete chloroplast genome available in Scrophulariaceae [38] (S3 Table, Fig 5A and 5B). Most (92.5%)
of the repeats in *S. dentata* are conserved and could be found in *S. takesimensis*, and there are another 3 complement repeats and 2 reverse repeats found in *S. takesimensis*.

In general, repeats are mostly distributed in noncoding regions [39, 40]. However, most of the repeats (62.5%) in the *S. dentata* chloroplast genome are located in coding regions (CDS),

![Fig 3. Percentages of variable sites in homologous regions across the six Lamiales chloroplast genomes.](A) coding regions; (B) introns and intergenic spacers (non-coding regions).](doi:10.1371/journal.pone.0158488.g003)

![Fig 4. Comparison of the borders of LSC, SSC and IR regions in six Lamiales chloroplast genomes.](Ψ indicates a pseudogene. This figure is not to scale.](doi:10.1371/journal.pone.0158488.g004)
<table>
<thead>
<tr>
<th>No.</th>
<th>Size(bp)</th>
<th>Type</th>
<th>Repeat 1 start (location)</th>
<th>Repeat 2 start (location)</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>F</td>
<td>9597(trnG-UCC*-trnR-UCU)</td>
<td>36418(trnG-GCC*-tmfM-CAU)</td>
<td>LSC</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>F</td>
<td>38641(psaB)</td>
<td>40865(psaA)</td>
<td>LSC</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>F</td>
<td>43669(ycf3 intron1)</td>
<td>119061(ndhA intron)</td>
<td>LSC; SSC</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>F</td>
<td>88876(ycf2)</td>
<td>88918(ycf2)</td>
<td>IRb</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>F</td>
<td>91289(ycf2)</td>
<td>91343(ycf2)</td>
<td>IRb</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>F</td>
<td>91318(ycf2)</td>
<td>91336(ycf2)</td>
<td>IRb</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>F</td>
<td>145238(ycf2)</td>
<td>145292(ycf2)</td>
<td>IRa</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>F</td>
<td>145258(ycf2)</td>
<td>145294(ycf2)</td>
<td>IRa</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
<td>F</td>
<td>147663(ycf2)</td>
<td>147705(ycf2)</td>
<td>IRa</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>F</td>
<td>43669(ycf3 intron1)</td>
<td>98305(rps12-trnV-GAC)</td>
<td>LSC; IRb</td>
</tr>
<tr>
<td>11</td>
<td>31</td>
<td>F</td>
<td>12909(atpF-atpH)</td>
<td>69566(rpl20-rps12)</td>
<td>LSC</td>
</tr>
<tr>
<td>12</td>
<td>32</td>
<td>F</td>
<td>8053(psbl-trnS-GCU*)</td>
<td>35525(psbc-trnS-UGA*)</td>
<td>LSC</td>
</tr>
<tr>
<td>13</td>
<td>34</td>
<td>F</td>
<td>91308(ycf2)</td>
<td>91326(ycf2)</td>
<td>IRb</td>
</tr>
<tr>
<td>14</td>
<td>34</td>
<td>F</td>
<td>145251(ycf2)</td>
<td>145294(ycf2)</td>
<td>IRa</td>
</tr>
<tr>
<td>15</td>
<td>36</td>
<td>F</td>
<td>91288(ycf2)</td>
<td>91306(ycf2)</td>
<td>IRb</td>
</tr>
<tr>
<td>16</td>
<td>36</td>
<td>F</td>
<td>145269(ycf2)</td>
<td>145287(ycf2)</td>
<td>IRa</td>
</tr>
<tr>
<td>17</td>
<td>39</td>
<td>F</td>
<td>43657(ycf3 intron1)</td>
<td>98293(rps12-trnV-GAC)</td>
<td>LSC; IRb</td>
</tr>
<tr>
<td>18</td>
<td>39</td>
<td>F</td>
<td>43657(ycf3 intron1)</td>
<td>119049(ndhA intron)</td>
<td>LSC; SSC</td>
</tr>
<tr>
<td>19</td>
<td>40</td>
<td>F</td>
<td>91290(ycf2)</td>
<td>91326(ycf2)</td>
<td>IRb</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>F</td>
<td>145245(ycf2)</td>
<td>145294(ycf2)</td>
<td>IRa</td>
</tr>
<tr>
<td>21</td>
<td>41</td>
<td>F</td>
<td>98291(trnR-UCU)</td>
<td>119047(ndhA intron)</td>
<td>IRb; SSC</td>
</tr>
<tr>
<td>22</td>
<td>30</td>
<td>P</td>
<td>8055(psbl-trnS-GCU*)</td>
<td>45358(trnS-GGA)</td>
<td>LSC</td>
</tr>
<tr>
<td>23</td>
<td>30</td>
<td>P</td>
<td>43669(ycf3 intron1)</td>
<td>138276(trnV-GAC-rps12)</td>
<td>LSC; IRa</td>
</tr>
<tr>
<td>24</td>
<td>30</td>
<td>P</td>
<td>75205(petB intron)</td>
<td>75205(petB intron)</td>
<td>LSC</td>
</tr>
<tr>
<td>25</td>
<td>30</td>
<td>P</td>
<td>35527(psbc-trnS-UGA*)</td>
<td>45358(trnS-GGA)</td>
<td>LSC</td>
</tr>
<tr>
<td>26</td>
<td>30</td>
<td>P</td>
<td>88876(ycf2)</td>
<td>147663(ycf2)</td>
<td>IRb; IRa</td>
</tr>
<tr>
<td>27</td>
<td>30</td>
<td>P</td>
<td>88918(ycf2)</td>
<td>147705(ycf2)</td>
<td>IRb; IRa</td>
</tr>
<tr>
<td>28</td>
<td>30</td>
<td>P</td>
<td>91289(ycf2)</td>
<td>145238(ycf2)</td>
<td>IRb; IRa</td>
</tr>
<tr>
<td>29</td>
<td>30</td>
<td>P</td>
<td>91318(ycf2)</td>
<td>145245(ycf2)</td>
<td>IRb; IRa</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>P</td>
<td>91336(ycf2)</td>
<td>145263(ycf2)</td>
<td>IRb; IRa</td>
</tr>
<tr>
<td>31</td>
<td>30</td>
<td>P</td>
<td>91343(ycf2)</td>
<td>145292(ycf2)</td>
<td>IRb; IRa</td>
</tr>
<tr>
<td>32</td>
<td>34</td>
<td>P</td>
<td>91308(ycf2)</td>
<td>145251(ycf2)</td>
<td>IRb; IRa</td>
</tr>
<tr>
<td>33</td>
<td>34</td>
<td>P</td>
<td>91326(ycf2)</td>
<td>145269(ycf2)</td>
<td>IRb; IRa</td>
</tr>
<tr>
<td>34</td>
<td>36</td>
<td>P</td>
<td>91288(ycf2)</td>
<td>145269(ycf2)</td>
<td>IRb; IRa</td>
</tr>
<tr>
<td>35</td>
<td>36</td>
<td>P</td>
<td>91306(ycf2)</td>
<td>145287(ycf2)</td>
<td>IRb; IRa</td>
</tr>
<tr>
<td>36</td>
<td>39</td>
<td>P</td>
<td>43657(ycf3 intron1)</td>
<td>138279(trnV-GAC-rps12)</td>
<td>LSC; IRa</td>
</tr>
<tr>
<td>37</td>
<td>40</td>
<td>P</td>
<td>91290(ycf2)</td>
<td>145245(ycf2)</td>
<td>IRb; IRa</td>
</tr>
<tr>
<td>38</td>
<td>40</td>
<td>P</td>
<td>91326(ycf2)</td>
<td>145281(ycf2)</td>
<td>IRb; IRa</td>
</tr>
<tr>
<td>39</td>
<td>41</td>
<td>P</td>
<td>119047(ndhA intron)</td>
<td>138279(trnV-GAC-rps12)</td>
<td>SSC; IRa</td>
</tr>
<tr>
<td>40</td>
<td>44</td>
<td>P</td>
<td>74375(psbt-psbN)</td>
<td>74375(psbt-psbN)</td>
<td>LSC</td>
</tr>
</tbody>
</table>

* part in the gene.

Repeat sequences conserved in the *S. takesimensis* chloroplast genome are highlighted in bold.

F = forward, P = palindrome, IGS = intergenic spacer.

doi:10.1371/journal.pone.0158488.t004

mainly in *ycf2*; and it is similar to that of *S. takesimensis* (Fig 5B). Meanwhile, 30% repeats are located in intergenic spacers (IGS) and introns, and 7.5% repeats in parts of the IGS and CDS.
Also, simple sequence repeats (SSRs) exert significant influence on genome rearrangement and recombination [41]. A total of 44 SSRs were detected in the *S. dentata* chloroplast genome, accounting for 500 bp of the total sequence (ca.0.33%), and there were 33, 7, 1 and 3 mono-, di-, tri-, and tetra- nucleotide repeats, respectively (Table 5). No penta- or hexa- nucleotide repeats were found. Most of the SSRs are mononucleotide repeats. 43 SSRs (97.7%) are composed of A and T nucleotides, whereas only one is composed of "GTCT" repeat. The high content of AT in SSRs contributes to the AT richness of the chloroplast genome [42, 43]. Among the SSRs, 33 are located in IGS and introns, 10 are found in coding genes, including ycf1, rpoC2, atpB, rpoA, atpA and ndhD, and 1 is located in parts of the IGS and CDS. Compared with *S. takesimensis*, 12 SSRs were identity, 30 exhibited length polymorphisms and 2 had not been detected (S4 Table, Fig 5C and 5D). These repeat sequences may be useful in developing lineage-specific markers, which could be widely used in genetic diversity and evolutionary studies of Scrophularia.

**Conclusions**

In this paper, we reported the complete sequence of *S. dentata* cp genome, as the first whole cp genome in the Sect. Tomiophyllum of the genus Scrophularia. This cp genome sequence was compared to other five genomes from the Lamiales species (i.e., *B. hygrometrica*, *H. palmeri*, *S. miltiorrhiza*, *A. paniculata* and *S. indicum*). No significant structural changes were detected among the chloroplast genomes. All the coding regions and non-coding regions (introns and intergenic spacers) within the six chloroplast genomes were extracted and analysed. The most
Table 5. Simple sequence repeats within *Scrophularia dentata* chloroplast genome.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Length</th>
<th>No.</th>
<th>SSR start</th>
<th>SSR-containing region</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>16</td>
<td>1</td>
<td>44287</td>
<td>ycf3 intron1</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>1</td>
<td>12158</td>
<td>atpF intron</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>2</td>
<td>8024</td>
<td>IGS (psbI-tmS-GCU)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16175</td>
<td>IGS (rps2-rpoC2)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>1</td>
<td>75204</td>
<td>petB intron</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>8</td>
<td>22732</td>
<td>rpoC1 intron</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>44557</td>
<td>IGS (ycf3-tmS-GGA)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59913</td>
<td>IGS (accD-psal)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>63314</td>
<td>IGS (petA-psbJ)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>111914</td>
<td>IGS (ndhF-rp32)</td>
<td>SSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>112458</td>
<td>IGS (rp32-trnL-UAG)</td>
<td>SSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>137924</td>
<td>IGS (tmV-GAC-rps12)</td>
<td>IRa</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>152482</td>
<td>IGS (rip2-Vps19)</td>
<td>IRa</td>
</tr>
<tr>
<td>T</td>
<td>15</td>
<td>1</td>
<td>12921</td>
<td>IGS (atpF-atpH)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>1</td>
<td>60099</td>
<td>IGS (psal-ycf4)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>1</td>
<td>69580</td>
<td>IGS (rpl20-rps12)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>1</td>
<td>124137</td>
<td>ycf1</td>
<td>SSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8632</td>
<td>IGS (tmS-GCU-trnG-UCC)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>14323</td>
<td>IGS (atpH-atpI)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18378</td>
<td>rpoC2</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>57534</td>
<td>IGS (rbcl-accD)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>80092</td>
<td>IGS (infA-rps8)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>80606</td>
<td>IGS (rps8-rpl14)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>112115</td>
<td>IGS (ndhF-rp32)</td>
<td>SSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>124952</td>
<td>ycf1</td>
<td>SSC</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>8</td>
<td>52007</td>
<td>tmV-UAC intron</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>54874</td>
<td>atpB</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>54959</td>
<td>IGS (atpB-rbcl)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>78182</td>
<td>rpoA</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>84121</td>
<td>IGS (rps19-rp32)</td>
<td>IRb</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>98679</td>
<td>IGS (rps12-trnV-GAC)</td>
<td>IRb</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>124603</td>
<td>ycf1</td>
<td>SSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>124880</td>
<td>ycf1</td>
<td>SSC</td>
</tr>
<tr>
<td>AT</td>
<td>12</td>
<td>1</td>
<td>42432</td>
<td>IGS (psaA-ycf3)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2</td>
<td>19751</td>
<td>rpoC2</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13822</td>
<td>IGS (atpH-atpI)</td>
<td>LSC</td>
</tr>
<tr>
<td>TA</td>
<td>18</td>
<td>1</td>
<td>46524</td>
<td>IGS (rps4-trnT-UGU)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>3</td>
<td>8212</td>
<td>IGS (tmS-GCU-trnG-UCC)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>31456</td>
<td>IGS (tmE-UUC-trnT-GGU)</td>
<td>LSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>42433</td>
<td>IGS (psaA-ycf3)</td>
<td>LSC</td>
</tr>
<tr>
<td>ATA</td>
<td>12</td>
<td>1</td>
<td>54928</td>
<td>IGS (atpB-rbcl)</td>
<td>LSC</td>
</tr>
<tr>
<td>GTCT</td>
<td>12</td>
<td>1</td>
<td>10993</td>
<td>atpA</td>
<td>LSC</td>
</tr>
<tr>
<td>TTTA</td>
<td>12</td>
<td>1</td>
<td>49645</td>
<td>IGS (ndhJ-ndhK)</td>
<td>LSC</td>
</tr>
<tr>
<td>AATA</td>
<td>12</td>
<td>1</td>
<td>114611</td>
<td>ndhD</td>
<td>SSC</td>
</tr>
</tbody>
</table>

* part in the gene.

SSRs that are identical in the *S. takesimensis* chloroplast genome are highlighted in bold.

doi:10.1371/journal.pone.0158488.t005
divergent regions are localized in the intergenic spacers, including ccsA-ndhD, rps16-trnQ-UUG, ndhJ-ndhK, ndhE-ndhG, ndhC-trnV-UAC, trnH-GUG-psbA and ndhG-ndhI. For the coding regions, the most divergent regions are ycf1, matK, rpl22, rpl32, rps15 and ccsA. These genes are all located in single copy regions. There were some differences in the genome size and IR expansion or contraction; and the divergent regions were analysed. Tandem repeats and SSRs within S. dentata were compared with those of S. takesimensis, which may provide markers for the analyses of infraspecific genetic differentiation within Scrophularia. In addition, our studies could provide basic data for the alpine medicinal species conservation and molecular phylogenetic researches of Scrophulariaceae and Lamiales.

Supporting Information
S1 Table. List of all pairs of primers used for assembly validation. (DOC)
S2 Table. Percentages of variable sites across the six Lamiales chloroplast genomes. (XLS)
S3 Table. Repeat sequences within Scrophularia takesimensis chloroplast genome. (XLS)
S4 Table. Simple sequence repeats within Scrophularia takesimensis chloroplast genome. (XLS)

Author Contributions
Conceived and designed the experiments: ZZ.Performed the experiments: LN. Analyzed the data: LN. Contributed reagents/materials/analysis tools: ZZ GD MM. Wrote the paper: LN ZZ.

References


