J. For. Res. (2018) 29(4):1003–1011
https://doi.org/10.1007/s11676-017-0531-3

ORIGINAL PAPER

Analysis and characterization of the *Salix suchowensis* chloroplast genome

Congrui Sun[1] · Jie Li[1] · Xiaogang Dai[1] · Yingnan Chen[1]

Received: 2 September 2016 / Accepted: 19 October 2017 / Published online: 9 November 2017

Abstract By screening sequence reads from the *Salix suchowensis* chloroplast (cp) genome that were generated by next-generation sequencing platforms, we assembled a complete circular pseudomolecule for the cp genome. This pseudomolecule is 155,508 bp long and has a typical quadripartite structure that contains two single copy regions, a large single copy region (LSC, 84,385 bp), and a small single copy region (SSC, 16,209 bp) separated by inverted repeat regions (IRs, 27,457 bp). Gene annotation revealed that the *S. suchowensis* cp genome encoded 119 unique genes, including four ribosome RNA genes, 30 transfer RNA genes, 82 protein-coding genes, and three pseudogenes. Analysis of the repetitive sequences revealed 31 tandem repeats, 16 forward repeats, and five palindromic repeats. In addition, a total of 148 perfect microsatellites, which were characterized as A/T dominant in nucleotide composition, were detected. Significant shifting of the IR/SSC boundaries was revealed by comparing this cp genome with those of other rosid plants. We also constructed phylogenetic trees to demonstrate the phylogenetic position of *S. suchowensis* in Rosidae based on 66 orthologous protein-coding genes present in the cp genomes of 32 species. Sequencing 30 amplicons based on the pseudomolecule for experimental verification revealed 99.88% accuracy for the *S. suchowensis* cp genome assembly. Therefore, we assembled a high-quality pseudomolecule of the *S. suchowensis* cp genome, which is a useful resource for facilitating development of this shrub willow into a more productive bioenergy crop.

Keywords *Salix suchowensis* · Chloroplast · Genome structure · Gene content · Phylogenetic tree

Introduction

Chloroplast (cp) genomes provide essential information for the study of biological processes in plant cells (Raubeson and Jansen 2005), such as biosynthesis of starch, fatty acids, pigments, and amino acids (Neuhaus and Emes 2000). It is generally accepted that cps originated from endosymbiosis of cyanobacteria (Timmis et al. 2004). Cp genomes are typically paternally or biparentally inherited in gymnosperms (Reboud and Zeyl 1994). In contrast, cp genomes are maternally inherited in most angiosperms (Palmer et al. 1988).

The chloroplast genomes of angiosperms have a typical quadripartite structure that contains a large single copy region (LSC) and a small single copy region (SSC) separated by two inverted repeat regions (IRs), and range from 120 to 160 kb in length with closed circular DNA (Sugiura

Project Funding: This work was supported by the Key Forestry Public Welfare Project (201304102), and the Natural Science Foundation of China (31400564 and 315005533). It was also enabled by the Innovative Research Team of the Educational Department of China and the PAPD (Priority Academic Program Development) program at Nanjing Forestry University.

The online version is available at http://www.springerlink.com

Corresponding Editor: Yu Lei.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11676-017-0531-3) contains supplementary material, which is available to authorized users.

✉ Yingnan Chen
chenyingnan@njfu.edu.cn

[1] Co-Innovation Center for Sustainable Forestry in Southern China, College of Forestry, Nanjing Forestry University, Nanjing 210037, People's Republic of China

1995). Moreover, cp genomes are more conserved in genome structure and organization than nuclear and mitochondrial genomes (Raubeson and Jansen 2005). A study by Pyke (1999) revealed that approximately 400–1600 copies of cp genomes in each cell, which leads to high expression of cp genes.

In recent years, cp transformation has emerged as an environmentally friendly approach for plant genetic engineering (Daniell et al. 2002). Foreign genes in transformed cps cannot be disseminated by pollen, because this plastid is maternally inherited in most flowering plants, thus posing significantly lower environmental risks. Cp transformation also has many other unique advantages over nuclear transformation, such as permitting the introduction of thousands of copies of foreign genes per plant cell, which allows uniformly and extraordinarily high expression levels of foreign genes, and eliminates gene silencing and the "position effect" (Qian et al. 2013; Daniell 2007; Verma and Daniell 2007). With the development of next-generation sequencing technologies, almost 1174 cp genomes in Viridiplantae have been completely sequenced and deposited in the NCBI Organelle Genome Resources (http://www.ncbi.nlm.nih.gov/genome/organelle/) to date.

Salix suchowensis (subgenus *Vetrix*) is a small, early-flowering shrub willow endemic to China (Wang et al. 1984) and mainly distributed in Jiangsu, Shandong, Zhejiang, and Henan Provinces in China (Fang et al. 1999). For thousands of years, it has been used as basket-weaving material, but currently, it is being considered as a promising crop for bioenergy because of its high biomass yield (Smart and Cameron 2008). Because biomass yield is highly correlated with plant photosynthetic efficiency, analyzing and characterizing the cp genome of this shrub willow will provide essential information to improve productivity and facilitate the development of a plastid transformation system in this woody crop.

In 2014, the whole genome of *S. suchowensis* was sequenced by using a whole-genome shotgun strategy that incorporated Roche/454 and Illumina/HiSeq-2000 sequencing technologies, which produced 10.1 Gb 454 GS FLX reads and 230.2 Gb Illumina reads (Dai et al. 2014). Because the sequencing libraries were constructed with leaf tissue, the generated reads included numerous sequence reads from the willow cp genome and provided sufficient sequence information for assembling the cp genome. In this study, our goals were to assemble and characterize the *S. suchowensis* cp genome by screening the organelle reads from the willow genome sequencing project and experimentally assessing the assembly quality of the cp genome.

Materials and methods

Sequence reads and cp genome assembly

Sequence reads were selected from the sequence database from the *S. suchowensis* genome sequencing generated by Dai et al. (2014). By mapping the raw reads to 660 cp genomes of terrestrial plants in the NCBI Organelle Genome Resources database (http://www.ncbi.nlm.nih.gov/genome/organelle/), we screened the willow cp sequence reads using BLASTN with an E value of $1e^{-50}$ following the protocol described by Ma et al. (2016). The obtained reads were further assembled using Amos (Treangen et al. 2011). Finally, a complete circular cp genome was established using Phrap (Ewing et al. 1998) according to the reference cp genomes of *S. purpurea* (Wu 2015), *Populus trichocarpa* (Tuskan et al. 2006), and *Arabidopsis thaliana* (Sato et al. 1999). The complete circular cp genome of *S. suchowensis* was deposited in GenBank under accession no. KU341117. The sequencing depth was estimated using the modified formula of Zhang et al. (2010): the total size of the sampling reads divided by the size of the assembled cp genome.

Gene annotation and comparative analysis

First, the *S. suchowensis* cp genome was annotated using the online program Dual Organellar GenoMe Annotator (DOGMA, Wyman et al. 2004). Genes that could not been annotated by DOGMA were manually identified by referring to the annotation of *P. trichocarpa* (Tuskan et al. 2006). In addition, all tRNA genes were predicted with the online program tRNAscan-SE 1.21 (Schattner et al. 2005). Then, a circular cp genome map was generated using the OrganellarGenomeDRAW tool (OGDRAW) (http://ogdraw.mpimp-golm.mpg.de/).

Besides the sequenced *S. suchowensis* cp genome, eight complete cp genome sequences of Salicaceae species, including five poplars (*Populus alba*, AP008956; *P. balsamifera*, KJ664927; *P. fremontii*, KJ664926; *P. tremula*, KP861984; *P. trichocarpa*, EF489041;) and three willows (*Salix babylonica*, KT449800; *S. interior*, KJ742926; *S. purpurea*, KP019639), were obtained from NCBI Organelle Genome Resources database (http://www.ncbi.nlm.nih.gov/genome/organelle/). Comparative cp genome analysis among the nine Salicaceae species was performed by using blast 2.3.0 (ftp://ncbi.nlm.nih.gov/blast/executables/blast+/2.3.0/).

cp genome structure and sequence analysis

Tandem repeats were evaluated using Tandem Repeat Finder 4.09 (Benson 1999) with default settings. Forward repeats and palindromic repeats were identified using REPuter (http://bibiserv.techfak.uni-bielefeld.de/reputer/), and the minimal repeat size setting was greater than 30 bp with a Hamming distance of 3. Microsatellite or simple sequence repeats (SSRs) of one to six nucleotides were detected using the Perl script MISA (http://pgrc.ipk-gatersleben.de/misa/), and thresholds of nine, five, five, three, three, and three repeat units were set for mono-, di-, tri-, tetra-, penta- and hexanucleotide SSRs, respectively.

Phylogenetic analysis

For phylogenetic analysis, besides *S. suchowensis*, we selected 31 rosid lineages that have complete cp genomes available. These lineages were from six rosid families (Salicaceae, Rosaceae, Moraceae, Fagaceae, Chrysobalanaceae, and Fabaceae), and *Ginkgo biloba* was used as the outgroup species. Based on the functional annotation from NCBI Organelle Genome Resources database, we identified 66 protein-coding genes that were commonly present in the analyzed cp genomes. These 66 orthologous genes were selected to construct the phylogenetic tree. The protein sequences were aligned with ClustalW 2.0 (Larkin et al. 2007), and a matrix consisting of 83,072 amino acids (aa) in default length was obtained. Optimal maximum likelihood (ML) and neighbor-joining (NJ) trees were constructed using MEGA 6.0 (Tamura et al. 2013). For the ML analyses, the Nearest-Neighbor-Interchange model was used with 1000 bootstrap replicates. The NJ tree was constructed using the Poisson model with 1000 bootstrap replicates.

Experimental verification of the cp genome assembly

To verify the assembly, we randomly designed 30 primer pairs (Table S1) based on the derived *S. suchowensis* cp genome using Primer Premier 5.0 (Lalitha 2000). The cpDNA was extracted according to the method described by Mcpherson et al. (2013). Following amplification with these primers against the extracted DNA templates, the generated amplicons were sequenced on an ABI 3730 sequencer by Genscript Biology Company (Nanjing, Jiangsu, China). PCRs were performed as follows: each 20-μL PCR mixture consisted of 2.0 μL genomic DNA (100 ng), 2.0 μL 10 × PCR buffer, 0.2 μL Taq DNA polymerase (TaKaRa, Japan), 1.6 μL $MgCl_2$ (25 mM), 4.0 μL dNTP (2.5 mM each), 1.0 μL of each primer (10 mmol/L), and 8.2 μL ddH2O. PCR amplification conditions included initial denaturation at 94 °C for 4 min, followed by 30 cycles of 94 °C for 1 min, 58 °C for 30 s, and 72 °C for 1 min, followed by a final extension at 72 °C for 5 min and storage at 4 °C.

Results and discussion

cp genome assembly and structure analysis

By mapping the raw reads of the *S. suchowensis* genome sequencing project to the NCBI Organelle Genome Resources database (http://www.ncbi.nlm.nih.gov/genome/organelle/), a total of 1,171,821 reads (approximately 533 Mb) from the willow cp genome were obtained. De novo assembly by Amos (Treangen et al. 2011) yielded 3773 contigs. Referring to the cp genomes of *S. purpurea* (Wu 2015), *P. trichocarpa* (Tuskan et al. 2006), and *A. thaliana* (Sato et al. 1999), these contigs were integrated into a complete circular pseudomolecule that was 155,508 bp long (GenBank accession no. KU341117). Thus, the sequencing depth of the cp genome was expected to be more than 3000×. The physical map of the derived cp genome (Fig. 1) showed that it possessed a typical quadripartite structure that contained a pair of IRs (27,457 bp) separated by LSCs (84,385 bp) and SSCs (16,209 bp).

When we compared the cp genomes across nine Salicaceae species, we found that, although the cp genomes were more conserved in structure and organization than the nuclear and mitochondrial genomes are (Raubeson and Jansen 2005), the length of certain regions of the cp genomes varied among these closely related species. In these species, the length of IRs, LSCs, and SSCs ranged from 27,167 to 27,838 bp, 84,377 to 85,979 bp, and 16,209 to 16,600 bp, respectively (Table 1), with very high sequence similarities.

The GC content, an important characteristic of the cp genome that affects genome stability (Yap et al. 2015), in the cp genomes of Salicaceae species ranged from 36.65% to 37.00%, with an average of 36.73% (Table 1). The global GC content in the *S. suchowensis* cp genome was 36.73%, which was the same as the average of the closely related Salicaceae species, but higher than those of *Wollemia nobilis* (36.5%) (Yap et al. 2015) and *Metasequoia glyptostroboides* (35.3%) (Chen et al. 2015), and lower than those of *Actinidia chinensis* (37.2%) (Yao et al. 2015), *Macadamia integrifolia* (38.1%) (Nock et al. 2014), and *Hyoscyamus niger* (37.6%) (Sanchezpuerta and Abbona 2014). These species were more diverged from *S. suchowensis* than those listed in Table 1.

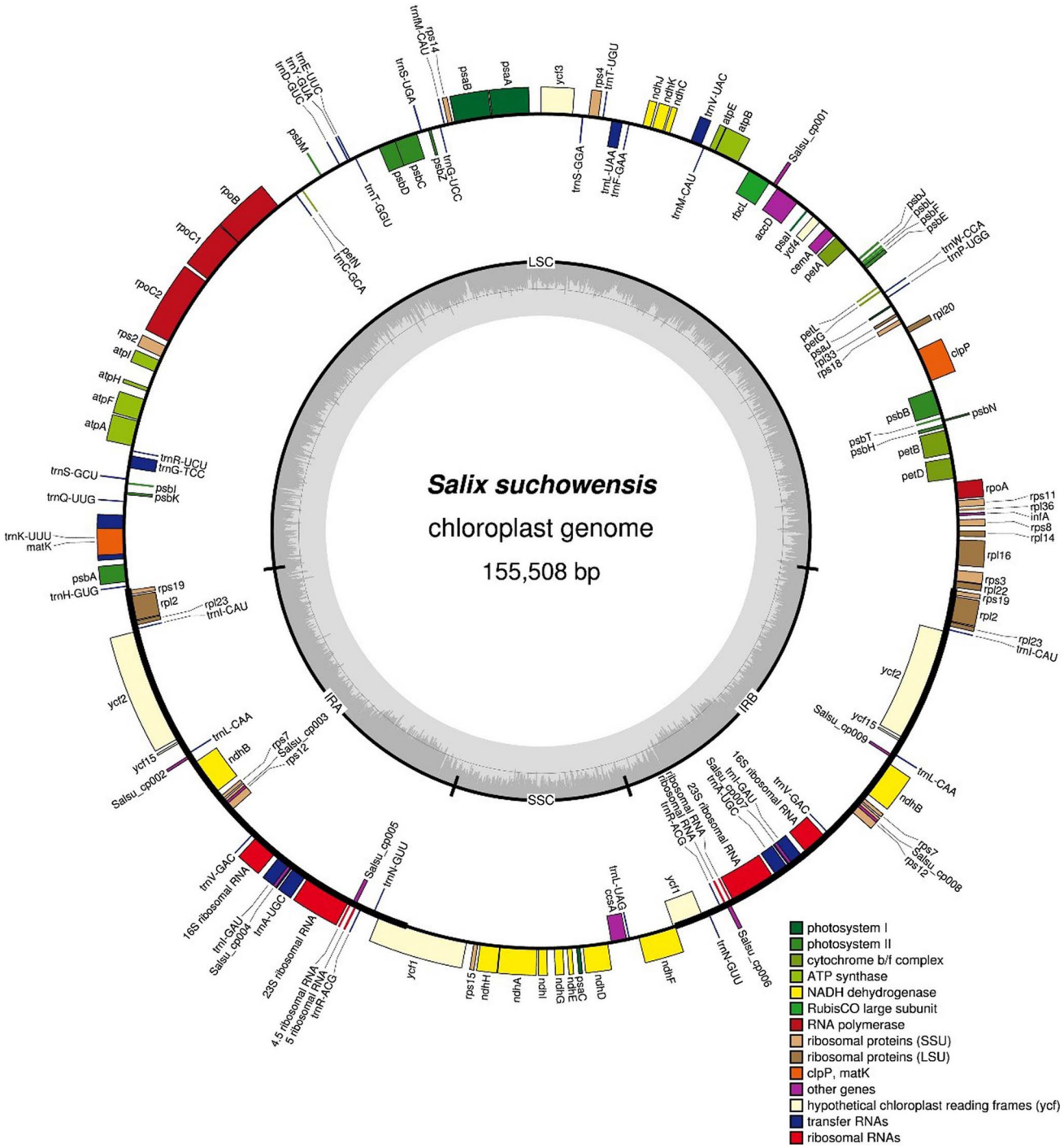

Fig. 1 Physical map of *Salix suchowensis* complete chloroplast genome. Genes outside the circle are transcribed counterclockwise, and genes inside the circle are transcribed clockwise. Genes in the same color are in the same functional group. Internal circle of darker gray and lighter gray indicate GC content and AT content, respectively

Gene annotation

Annotation of the *S. suchowensis* cp genome revealed 143 genes. In the gene function analysis, the 143 genes were classified into four categories, including genes associated with self-replication, photosynthesis, and other functions, and genes of unknown function (Table 2). Among these genes, 119 were unique, including four rRNA genes, 30 tRNA genes, 82 protein-coding genes, and three pseudogenes. Moreover, four rRNA, seven tRNA, and 13 protein-coding genes were duplicated in the IRs. Most of the unique genes contained no introns, but one intron was

Table 1 Comparative analysis of cp genomes across nine Salicaceae species

Species	IR (bp)	SSC (bp)	LSC (bp)	GC content (%)	No. of genes
Populus alba	27,660	16,567	84,618	36.74	109 + 1
P. balsamifera	27,836	16,499	84,921	36.65	109 + 3
P. fremontii	27,838	16,316	85,454	36.67	106 + 3
P. tremula	27,600	16,490	84,377	36.76	111
P. trichocarpa	27,652	16,600	85,129	36.68	119 + 1
Salix babylonica	27,646	16,273	85,255	36.65	109 + 1
S. interior	27,167	16,307	85,979	37.00	106 + 2
S. purpurea	27,459	16,220	84,455	36.69	110 + 1
S. suchowensis	27,457	16,209	84,385	36.73	116 + 3

The number after "+" is the number of pseudogenes. *IR* inverted repeat, *SSC* small single copy, *LSC* large single copy

Table 2 Summary of gene annotation for the cp genome of *Salix suchowensis*

Gene category	Group of genes	Name of genes
Self replication	Transfer RNA genes	30 tRNAs (6 contain an intron)
	Ribosomal RNA genes	*rrn4.5*, *rrn5*, *rrn16*, *rrn23*
	DNA dependent RNA polymerase	*rpoA*, *rpoB*, *rpoC1**, *rpoC2*
	Small subunit of ribosome	*rps2*, *rps3*, *rps4*, *rps7*, *rps8*, *rps11*, *rps12***, *rps14*, *rps15*, *rps18*, *rps19*
	Large subunit of ribosome	*rpl2**, *rpl14*, *rpl16**, *rpl20*, *rpl22*, *rpl23*, *rpl33*, *rpl36*
Large subunit of ribosome	Subunits of photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*
	Subunits of photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*
	Subunits of cytochrome	*petA*, *petB**, *petD**, *petG*, *petL*, *petN*
	Subunits of ATP synthase	*atpA*, *atpB*, *atpE*, *atpF**, *atpH*, *atpI*
	Subunits of NADH dehydrogenase	*ndhA**, *ndhB**, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
	Large subunit of Rubisco	*rbcL*
	ATP-dependent protease subunit p gene	*clpP***
Other genes	Subunit of acetyl-CoA-carboxylase	*accD*
	c-type cytochrome synthesis gene	*ccsA*
	Envelop membrane protein	*cemA*
	Maturase	*matK*
Genes of unknown function	Pseudogene	*Pseudo-ycf68*, *Pseudo-ycf1*, *Pseudo-infA*
	Conserved open reading frames	*ycf1*, *ycf2*, *ycf3***, *ycf4*, *ycf15*, *cp001*, *cp002*, *cp003*, *cp004*, *cp005*

* Contains one intron, ** Contains two introns

found in each of six tRNA genes and eight protein-coding genes, and two introns were found in each of three protein-coding genes (Table 2).

ycf1 is one of the longest open reading frames in cp genomes and has been found in nearly all plastid genomes sequenced to date (Raubeson and Jansen 2005). Vries et al. (2015) assumed that *ycf1* encodes the translocon on the inner envelope of chloroplasts (TIC). Drescher et al. (2000) predicted that *ycf1* is involved in essential pathways in cellular metabolism or serves a structural function for the plastid compartment. The function of *ycf1* has not been clearly resolved; nevertheless, it is considered essential to plant survival (Drescher et al. 2000).

In the sequenced cp genome, *ycf1* usually spans the boundary of the IR and SSC regions (Raubeson and Jansen 2005). Based on the common location of *ycf1* in the plastid genome, a copy of the *ycf1* gene (5424 bp) was found at the IRa/SSC border (Fig. 1), and a truncate copy of a *ycf1* pseudogene (1878 bp) was present at the IRb/SSC border (Fig. 1) in the *S. suchowensis* cp genome. *ycf1* is highly

variable and approximately 5500 bp in plant plastid genomes. Compared with Chlorophyta species, the length of the *S. suchowensis* YCF1 protein (1807 aa) is much longer than that of *Nephroselmis olivacea* (956 aa; NC_0000927), and much shorter than that of *Schizomeris leibleinii* (3212 aa; NC_015645).

Recent studies demonstrated that genes can transfer from the cp genome to the nuclear genome at a relatively high frequency (Huang et al. 2003; Stegemann and Bock 2006). *infA*, which encodes the plastid translation initiation factor 1, provides a striking example of gene transfer events (Millen et al. 2001). We found a parallel of *infA* gene with an uncommon initiation codon of AGA in the *S. suchowensis* cp genome that was located in the LSC, and the length of this gene was 165 bp. Sequence alignment detected a fragment with high similarity (92.73%) on chromosome II of the *S. suchowensis* nuclear genome (Fig. 2). This *infA*-like fragment might be transferred from the cp genome to the nuclear genome.

Repeat sequence analysis

Previous studies revealed that gene duplication, gene expansion, and cpDNA rearrangement seemed to be associated with repetitive sequences (Cavalier-Smith 2002). We identified 31 tandem repeats, 16 forward repeats, and five palindromic repeats in the (Table S2). The tandem repeat units were 7–26 bp long, and almost all the tandem repeats were located at intergenic spacer (IGS) regions except for one located in an intron region. Additionally, the forward repeat units were 30–76 bp long. The majority of these repeats were distributed in IGS regions, with some of them detected in protein-coding and tRNA gene regions. Alternatively, for the palindromic repeats, the repeat units were 30–42 bp long. Four repeats were detected in IGS regions, and one was located in a tRNA gene region. Overall, 21 repeats $\geq$ 30 bp were detected in the *S. suchowensis* cp genome, and most (84.6%) were distributed in the intergenic spacer region. These repeat motifs can be selected for population studies since they are an informative source for developing markers.

Microsatellites or SSRs are common in the plant cp genome. The MISA output revealed 148 perfect SSRs in *S. suchowensis* cp genome. Among these, 126 SSRs were mononucleotide repeats, 10 were dinucleotide repeats, 11 were tetranucleotide repeats, and one was a pentanucleotide repeat (Table S3). Among the monomers, 121 consisted of A/T repeats, and only five consisted of G/C repeats. The A/T content of monomers was similar to that in the *M. glyptostroboides* cp genome (96.03%) (Chen et al. 2015). All dinucleotides in the *S. suchowensis* cp genome were AT/TA repeats, and A/T contents in tetramers and pentamers were 86.36% and 80%, respectively. When analyzed with the same parameters in MISA, the average SSR length and SSR density in *S. suchowensis* cp genome (10.23 bp, 9.74/1000 bp) were found to be lower than those of the *W. nobilis* (16.97 bp, 14.65/1000 bp) (Yap et al. 2015) and *M. glyptostroboides* cp genomes (11.01 bp, 9.85/1000 bp) (Chen et al. 2015).

Chloroplast SSRs (cpSSR) represent ideal complementary molecular tools for nuclear genetic markers. In combination with nuclear SSR markers, cpSSR markers have a high capability in differentiating among closely related taxa, e.g., grapes (Arroyo-Garcia et al. 2006). In the *S. suchowensis* cp genome, most SSRs were AT-rich, and the mononucleotides were found to be the dominant repeats. These results are consistent with the previous contention that cpSSRs are generally composed of short polyA or polyT repeats (Kuang et al. 2011).

IR contraction and expansion

IRs are prominent features of most angiosperm cp genomes. During the evolutionary process of angiosperms, IR

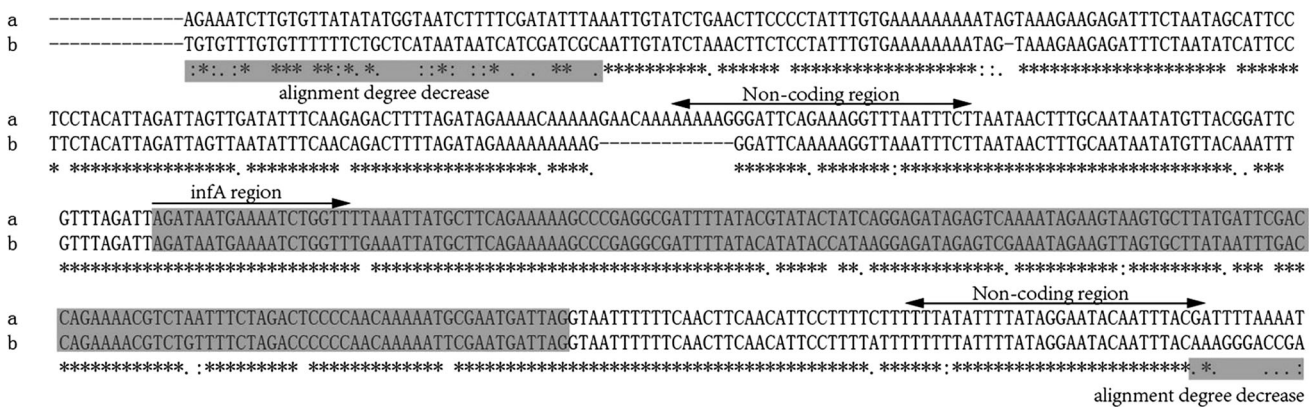

Fig. 2 Sequence alignment of *infA* from cp genome and that from nuclear genome. **a** cp genome of *S. suchowensis*; **b** nuclear genome of *S. suchowensis*

contraction and expansion might influence cp genome size (Goulding et al. 1996; Wang et al. 2008) and could create pseudogenes that cannot be transcribed (Wang et al. 2008). Here, we compared in detail the IR/single copy (SC) boundaries of four rosid plants: *S. suchowensis*, *S. integra*, *Prunus padus* and *Morus notabilis* (Fig. 3).

In the cp genome of *S. suchowensis* and *S. integra*, the IRb/LSC junction was found within the *rpl22* gene, and the *rpl22* pseudogene (52 bp) was detected at the IRa/LSC boundary, whereas the *rps19* pseudogene (39 bp) was found at the IRa/LSC border in *P. padus*. As for *M. notabilis*, IRb was found to be immediately adjacent to the *rps19* gene, and no pseudogene was observed at the IRa/LSC boundary. The *trnH* genes were all located within the LSC region in these four species, but varied in being between 16 and 36 bp from the IRa/LSC junctions.

In most land plant cp genomes, the *ycf1* pseudogene and *ndhF* are located at the LSC/IR border, such as in *P. trichocarpa* (Tuskan et al. 2006), *Glycine stenophita* (Sherman-Broyles et al. 2014), and *S. miltiorrhiza* (Qian et al. 2013). At the IRa/SSC border of the four cp genomes, the IR expanded into the *ycf1* gene, creating the *ycf1* pseudogene at the IRb/SSC border. The length of the *ycf1* pseudogene was 1878 bp in *S. suchowensis*, 1713 bp in *S. integra*, 1036 bp in *P. padus*, and 1002 bp in *M. notabilis*. In addition, the *ycf1* pseudogene and *ndhF* gene overlapped in *S. suchowensis*, *P. padus* and *M. notabilis* by 140 bp, 19 bp and 26 bp, respectively.

Phylogenetic trees

To elucidate the phylogenetic position of *S. suchowensis* among rosids, we analyzed 66 orthologous protein-coding genes present in the cp genomes of 33 species (Table S4). The ML bootstrap analysis resolved 29 nodes, of which 25 had bootstrap values ≥ 90%, and 18 of these had bootstrap support of 100% (Fig. 4). With the NJ tree, we obtained a sum of branch lengths of 0.61439509. The NJ bootstrap analysis was similar to that of the ML tree that resolved into 29 nodes, because 24 nodes had bootstrap values ≥ 90%, and 20 of these had 100% bootstrap support (Fig. S1). Both the ML and NJ trees showed that these species were evident in three rosid categories: Rosids I (Salicaceae and Chrysobalanaceae), Rosids II (Fagaceae, Moraceae, and Rosaceae), and Rosids III (Fabaceae). In Rosids I, *S. suchowensis* and *S. babylonica* were the closest relatives. The topology of the derived ML tree was very similar to that of the established NJ tree; the only incongruence was the position of *P. fremontii* and of *P. balsamifera* relative to *P. trichocarpa*. It is noteworthy that the bootstrap supports for grouping *P. fremontii* or *P. balsamifera* with *P. trichocarpa* are relatively low in both the ML and NJ trees. The relationships of these three species might not be able to be properly resolved merely based on plastid-level information.

Verification of the *S. suchowensis* cp genome assembly

In this study, the raw reads were generated by next-generation sequencing platforms, and the screening of reads and cp genome assembly were conducted merely by using bioinformatics tools. To verify the quality of the assembly, we sequentially selected 30 sites from the cp genome.

All synthesized primers succeeded in PCR amplification. Sequenced using a Sanger sequencer, the 30 amplicons covered a total physical length of 18,639 bp. Alignment of the amplicon sequences to the genome assembly produced sequence errors in seven of the tested sites, whereas a 100% match was revealed with amplicons

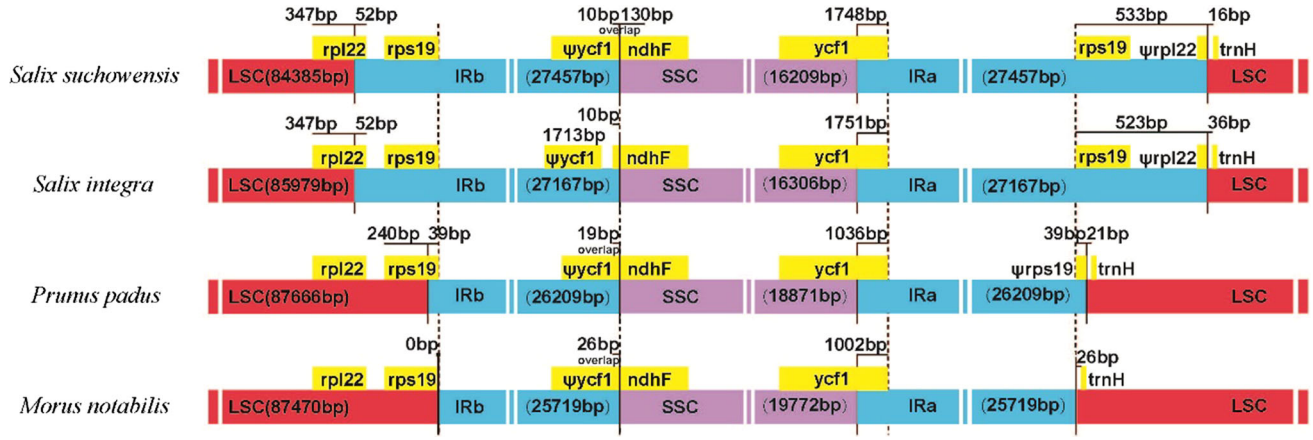

Fig. 3 Comparison of the borders of IR regions among the cp genomes of four rosid plants. Three colors were used to indicate the LSC, IR and SSC regions, respectively. The figure mainly indicates the shift of the genes located in the IR border. "ψ" means pseudogene, "overlap" means overlap of *ψycf1* and *ndhF*

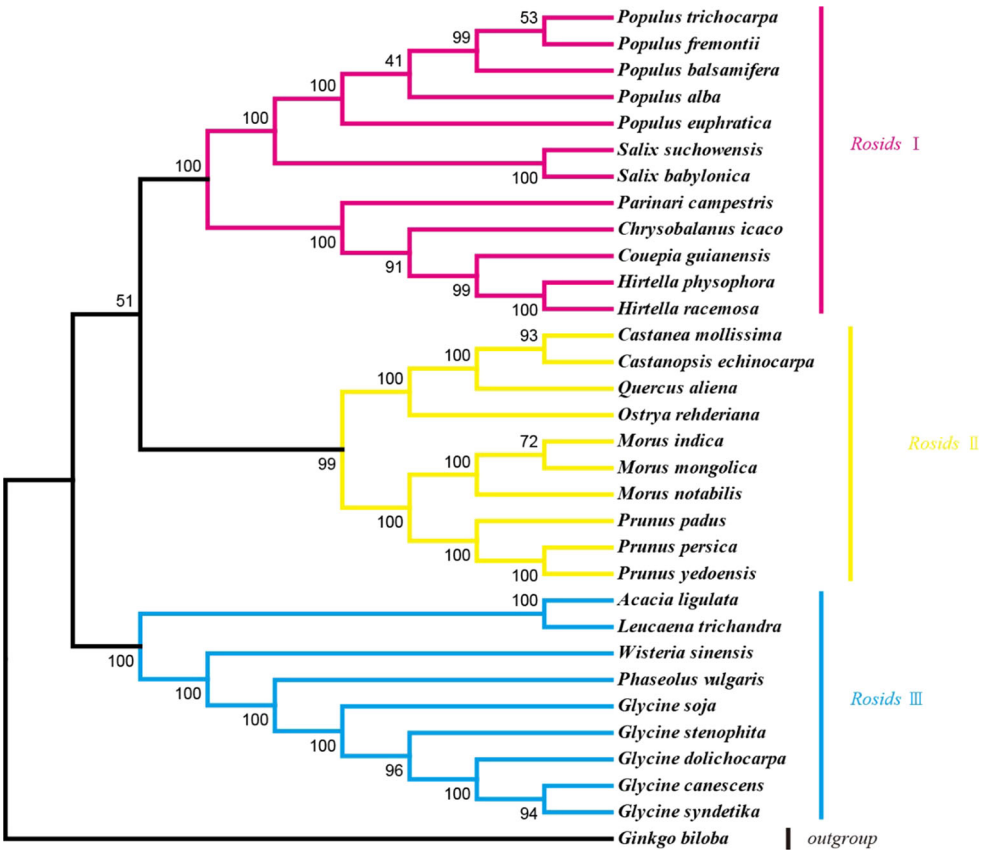

Fig. 4 Maximum likelihood (ML) phylogenetic tree

at the other 23 sites in the cp genome assembly. The overall accuracy of the derived assembly was estimated to be 99.88%. Therefore, the cp genome obtained in this study was high quality. Moreover, as mentioned above, we detected raw reads that covered over 3000× sequencing depth of the *S. suchowensis* cp genome. The high sequencing depth ensures accuracy and integrity of the obtained pseudomolecule of the cp plastid.

Conclusion

We revealed a pseudomolecule of the *S. suchowensis* cp genome with high reliability based on raw reads generated by next-generation sequencing platforms. The genome structure and organization of the *S. suchowensis* cp genome were similar to those in other Salicaceae species. IR expansion was observed in *S. suchowensis* in comparison with those of other rosid plants. The repeats and SSRs identified here will be informative sources for developing markers for evolutionary and population-genetic studies. The *S. suchowensis* cp genome obtained in this study is highly desirable for facilitating the biological study of this promising biofuel plant and will facilitate more extensive studies such as cpSSR development, endosymbiotic gene transfer and plastid genetic engineering in willows.

References

Arroyo-Garcia R, Ruiz-Garcia L, Bolling L, Ocete R, Lopez MA et al (2006) Multiple origins of cultivated grapevine (*Vitis vinifera* L. *ssp. sativa*) based on chloroplast DNA polymorphisms. Mol Ecol 15(12):3707–3714

Benson G (1999) Tandem repeats finder: a program to analyze dna sequences. Nucleic Acids Res 27(2):573–580

Cavalier-Smith T (2002) Chloroplast evolution: secondary symbiogenesis and multiple losses. Curr Biol Cb 12(2):62–64

Chen J, Hao Z, Xu H, Yang L, Liu G et al (2015) The complete chloroplast genome sequence of the relict woody plant *Metasequoia glyptostroboides* hu et cheng. Front Plant Sci 6:447

Dai X, Hu Q, Cai Q, Feng K, Ye N et al (2014) The willow genome and divergent evolution from poplar after the common genome duplication. Cell Res 24(10):1274–1277

Daniell H (2007) Transgene containment by maternal inheritance: effective or elusive? Proc Natl Acad Sci 104(17):6879–6880

Daniell H, Khan MS, Allison L (2002) Milestones in chloroplast genetic engineering: an environmentally friendly era in biotechnology. Trends Plant Sci 7(2):84–91

Drescher A, Ruf S, Calsa T, Carrer H, Bock R (2000) The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J 22(2):97–104

Ewing B, Hillier LD, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. i. accuracy assessment. Genome Res 8(3):175–185

Fang Z, Zhao S, Skvortsov AK (1999) Saliceae. In: Zheng-yi W, Raven PH (eds) Flora of China. Missouri Botanical Garden Press, St. Louis, pp 139–274

Goulding SE, Wolfe KH, Olmstead RG, Morden CW (1996) Ebb and flow of the chloroplast inverted repeat. Mol Gen Genet MGG 252(1–2):195–206

Huang CY, Ayliffe MA, Timmis JN (2003) Direct measurement of the transfer rate of chloroplast dna into the nucleus. Nature 422(6927):72–76

Kuang D, Wu H, Wang Y, Gao L, Zhang S et al (2011) Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. Genome 54:663–673

Lalitha S (2000) Primer premier 5. Biotech Softw Internet Rep 1(6):270–272

Larkin MA, Blackshields G, Brown NP, Chenna RM, McGettigan PA et al (2007) Clustal w and clustal x version 2.0. Bioinformatics 23(21):2947–2948

Ma Q, Li S, Bi C, Hao Z, Sun C, Ning Y (2016) Complete chloroplast genome sequence of a major economic species, *Ziziphus jujuba* (Rhamnaceae). Curr Genet 63:1–13

Mcpherson H, Merwe MVD, Delaney SK, Edwards MA, Henry RJ et al (2013) Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. BMC Ecol 13(1):53–65

Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT et al (2001) Many parallel losses of infa from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell 13(3):645–658

Neuhaus HE, Emes MJ (2000) Nonphotosynthetic metabolism in plastids. Annu Rev Plant Biol 51(4):111–140

Nock CJ, Baten A, King GJ (2014) Complete chloroplast genome of *Macadamia integrifolia* confirms the position of the gondwanan early-diverging eudicot family proteaceae. BMC Genom 15(Suppl 9):S13

Palmer JD, Jansen RK, Michaels HJ, Chase MW, Manhart JR (1988) Chloroplast DNA variation and plant phylogeny. Ann Mo Bot Gard 75(4):1180–1206

Pyke KA (1999) Plastid division and development. Plant Cell 11(4):549–556

Qian J, Song J, Gao H, Zhu Y, Xu J et al (2013) The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. PLoS ONE 8(2):e57607

Raubeson LA, Jansen RK (2005) Chloroplast genomes of plants. In: Henry RJ (eds) Plant diversity and evolution: genotypic and phenotypic variation in higher plants. CABI Publishing, Cambridge, MA, pp 45–68

Reboud X, Zeyl C (1994) Organelle inheritance in plants. Heredity 72(2):132–140

Sanchezpuerta MV, Abbona CC (2014) The chloroplast genome of *Hyoscyamus niger* and a phylogenetic study of the tribe hyoscyameae (solanaceae). PLoS ONE 9(5):e98353

Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S (1999) Complete structure of the chloroplast genome of Arabidopsis thaliana. DNA Res 6(5):283–290

Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 33(suppl 2):W686–W689

Sherman-Broyles S, Bombarely A, Grimwood J, Schmutz J, Doyle J (2014) Complete plastome sequences from glycine syndetika and six additional perennial wild relatives of soybean. G3: genes Genomes. Genetics 4(10):2023–2033

Smart LB, Cameron KD (2008) Genetic improvement of Willow (*Salix* spp.) as a dedicated bioenergy crop. Genet Improv Bioenergy Crops 2:377–396

Stegemann S, Bock R (2006) Experimental reconstruction of functional gene transfer from the tobacco plastid genome to the nucleus. Plant Cell 18(11):2869–2878

Sugiura M (1995) The chloroplast genome. Essays Biochem 30(1):49–57

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) Mega6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30(12):2725–2729

Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5(2):123–135

Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M (2011) Next generation sequence assembly with AMOS. Curr Protoc Bioinform 33:11.8.1–11.8.18

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I et al (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313(5793):1596–1604

Verma D, Daniell H (2007) Chloroplast vector systems for biotechnology applications. Plant Physiol 145(4):1129–1143

Vries JD, Sousa FL, Bölter B, Soll J, Gould SB (2015) YCF1: a green Tic? Plant Cell 27(7):1827–1833

Wang C, Fang CF, Zhao SD (1984) Salicaceae. Flora Reipublicae Pop Sin 20(2):79–403

Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM (2008) Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. BMC Evol Biol 8(1):1

Wu Z (2015) The new completed genome of purple Willow (*Salix purpurea*) and conserved chloroplast genome structure of Salicaceae. J Nat Sci 1:e49

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20(17):3252–3255

Yao X, Tang P, Li Z, Li D, Liu Y, Huang H (2015) The first complete chloroplast genome sequences in actinidiaceae: genome structure and comparative analysis. PLoS ONE 10(6):e0129347

Yap JYS, Rohner T, Greenfield A, Merwe MVD, Mcpherson H et al (2015) Complete chloroplast genome of the wollemi pine (*Wollemia nobilis*): structure and evolution. PLoS ONE 10(6):e0128126

Zhang G, Guo G, Hu X, Zhang Y, Li Q et al (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res 20(5):646–654