Ann. For. Res. 65(1): 3-16, 2022
https://doi.org/10.15287/afr.2022.2153

ANNALS OF FOREST RESEARCH
www.afrjournal.org

Phylogenetic analysis of *Chosenia arbutifolia* (Pall.) A. Skv. in Salicaceae using complete chloroplast genome sequence

Xudong He[1,2]✉, Yu Wang[1,3], Jiwei Zheng[1,2], Zhongyi Jiao[1,2], Jie Zhou[1,2], Baosong Wang[1,2], Qiang Zhuge[3]

He X., Wang Y., Zheng J., Jiao Z., Zhou J., Wang B., Zhuge Q., 2022. Phylogenetic analysis of *Chosenia arbutifolia* (Pall.) A. Skv. in Salicaceae using complete chloroplast genome sequence. Ann. For. Res. 65(1): 3-16.

Abstract As a unique and endangered species in the family Salicaceae, *Chosenia arbutifolia* (Pall.) A. Skv. has great potential for use in ornamental and industrial purposes. Despite its comprehensive importance, the phylogenetic position of *C. arbutifolia* within Salicaceae is still ambiguous. In the present study, the whole chloroplast genome of *C. arbutifolia* was sequenced and compared with the genome of other Salicaceae species. A phylogenetic tree was established based on the maximum-likelihood (ML) methods. The *de novo* assemblies generated 155684 bp in length for the completed cp genome of *C. arbutifolia*, including a large single-copy region of 84551 bp, a small single-copy region of 16217 bp, and two inverted repeat regions of 27458 bp each. In total, 130 genes were predicted, of which 85 protein-coding genes were annotated in at least one of the five reference databases. In the repeat analysis, 23 forward, 15 palindromic, one complement, one reverse long repeats, and 221 putative SSRs were identified. The results of genome comparison showed that the large single copy region (LSC) region was more divergent than the small single copy region (SSC) and inverted repeated (IR) regions, and a higher divergence occurred in non-coding regions than in coding regions. Significant contractions or expansions were also observed at the IR-LSC/SSC boundaries. Phylogenetic analysis of 20 Salicaceae species confirmed that *C. arbutifolia* is closely related to *Salix* species and may therefore be treated as a member of the genus *Salix*. The complete *C. arbutifolia* chloroplast genome will provide insight into the chloroplast architecture, function, and evolution of this species and provide additional resources for future research.

Keywords: chloroplast genome, comparative analysis, phylogeny, *Chosenia arbutifolia*, *Salix*.

Addresses: [1]Department of Tree Genetics and Breeding, Jiangsu Academy of Forestry, Nanjing, China.| [2]Willow Nursery of the Jiangsu Provincial Platform for Conservation and Utilization of Agricultural Germplasm, Jiangsu Academy of Forestry, Nanjing, China.| [3]College of Biology and the Environment, Nanjing Forestry University, Nanjing, China.

✉ **Corresponding Author:** Xudong He (hxd_519@163.com).

Manuscript: received March 03, 2021; revised January 05, 2022; accepted January 28, 2022.

Introduction

The species *Chosenia arbutifolia* is placed the monotypic genus *Chosenia* belongings to the family Salicaceae, alongside some common plant genera such as *Salix* and *Populus* (Wang & Fang 1984). *C. arbutifolia* has a wide geographic range from the broadleaf forest zone of Honshu in southern Japan to the tundra in the lower reaches of the Anadyr and Lena rivers in Russia (Moskalyuk et al. 2016). In China, the natural range of *C. arbutifolia* includes the Greater or Lesser Khingan Mountains, the Changbai Mountains, and the montane regions of eastern Liaoning Province (Wang & Fang 1984). *C. arbutifolia* is mainly severe in the northeast of China owing to its particular features, including frost tolerance, beautiful shape, and scarlet branches, and its usage as an important wood source for different industries, including construction, paper and furniture production (Tu 1982). However, as a consequence of inappropriate deforestation and utilization, the nature populations of *C. arbutifolia* in China have drastically decreased, which is why the plant has been classified as critically endangered and is currently listed in the National Key Protected Wild Plants.

The genus *Salix*, which comprises more than 500 species of shrubs and trees worldwide and has multiple natural interspecific hybrids, is known to be a taxonomically difficult genus. The position of *Chosenia* in Salicaceae phylogeny is highly controversial (Chen et al. 2010). As an ancient tree species, *Chosenia* was considered a transitional form between *Populus* and *Salix*, and it was segregated into a separate genus by the Japanese botanist Takenoshi Nakai in 1920 (Nakai 1920), which was also acknowledged by a Russian taxonomist A. Skvortsov (1999). This opinion was supported by morphological evidence, such as missing nectaries, different structures of stamens, pistils, and bracts, and pollination by wind (Kadis 2005). In contrast, according to Ohashi, *Chosenia* should be treated as a member of the genus *Salix* (Ohashi 2001) and this was confirmed based on molecular data and phylogenetic analysis in recent studies (Leskinen & Alström-Rapaport 1999, Azuma et al. 2000, Chen et al. 2010, Feng et al. 2019). Moreover, *C. arbutifolia* was defined as the synonymic species *S. arbutifolia* based on the Plant List (http://www.theplantlist.org/tpl/record/tro-28300001) and Flora of China (http://www.efloras.org/florataxon.aspx?flora_id=2&taxon_id=200005635).

The chloroplast (cp) is an essential plastid organelle in higher plant cells and plays an important role in photosynthesis and other cellular functions, including the synthesis of fatty acids, amino acids, pigments, and starch (Neuhaus & Emes 2000, Wicke et al. 2011). The cp genome of most land plants has a typical circular quadripartite structure consisting of one large single copy region (LSC), one small single copy region (SSC), and two inverted repeated regions (IRa and IRb) (Wicke et al. 2011). In angiosperms, the size of the cp genome commonly ranges from 120 kb to 180 kb, and the gene number varies from 100 to 120, with a highly conserved genomic order and composition (Palmer 1985). Nevertheless, minor rearrangements in the cp genome have been observed in various plant species (Tangphatsornruang et al. 2011, Jheng et al. 2012, Walker et al. 2014, Daniell et al. 2016), and many mutation regions with single nucleotide polymorphisms and indels have been identified (Ingvarsson et al. 2003, Eguiluz et al. 2017). Due to their highly conserved structure, low recombination, uniparental inheritance, and abundant genetic information, cp genomes are of great value for taxonomic and phylogenetic analysis of plant species and individuals, especially woody plants (Chen et al. 2015, Song et al. 2015, Asaf et al. 2018, Li et al. 2018, Mader et al. 2018, Zhao et al. 2018).

To date, a total of 41 accessions of *Salix*

cp genome have been deposited in the NCBI database, including two synonymic species *S. arbutifolia* (MG262340 and KX781246). In the present study, the cp genome of *C. arbutifolia* was sequenced using high-throughput sequencing technology and *de novo* assembly. To further understand the phylogenetic relationships of *C. arbutifolia*, cp genome sequences of 14 *Salix* species and 6 *Populus* species were also obtained for comparative analysis. The results of our study will provide molecular evidence for validation of the complex evolutionary relationships in the family Salicaceae and will be beneficial for the development of cp genetic markers for Salicaceae species in the future.

Materials and Methods

Plant materials and DNA extraction

Branches of *C. arbutifolia* were sampled in Hunchun, Jilin Province, China (43.00.519°N, 130.52.920°E), and brought back to the laboratory in Nanjing, Jiangsu Province, China for hydroponic cultivation. Young leaves were collected, immediately frozen in liquid nitrogen, and stored at -80°C for furture analyses. Total genomic DNA was isolated using an improved extraction method (McPherson et al. 2013).

Genome sequencing, assembly, and annotation

Short-insert libraries (insert size 430 bp) were prepared and sequenced by Shanghai BIOZERON Biotechnology Co., Ltd (Shanghai, China) using the Illumina Hiseq 4000 platform according to the manufacturer' s protocols. All low-quality reads and adaptor sequences were removed. The cp genome was reconstructed using a combination of *de novo* and reference-guided assemblies (Cronn et al. 2008). The online tool DOGMA with default parameters was used for the annotation of cp genes, including protein-coding genes, tRNA genes, and rRNA genes (Wyman et al. 2004). A whole cp genome Blast search was compared against diverse protein databases, including Nr (Non-redundant Protein Databases), COG (Clusters of Orthologous Groups), KEGG (Kyoto Encyclopedia of Genes and Genomes), Swiss-Port, and GO (Gene Ontology). The circular map of the completely annotated genome was drawn using OGDRAW v1.2 (Lohse et al. 2007).

Repeat sequence and microsatellites

Four types of repeats, including forward, reverse, palindromic, and complement repeats, were identified using REPuter (Kurtz et al. 2001). To screen for SSRs, the MISA tool (http://pgrc.ipk-gatersleben.de/misa/misa.html) was used with the following parameters: the motif size was set to mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides at a minimum of 8, 5, 4, 3, 3, and 3 repeats, respectively; compound SSRs were those with less than five interval spaces between repeats, with each space appearing at a maximum of 100 nucleotides.

Comparative genome analysis

To carry out a comparative analysis, five cp genome sequences of the Salicaceae model species were obtained from NCBI, consisting *P. trichocarpa* (EF489041), *P. euphratcia* (KJ624919), *S. suchowensis* (KM983390), *S. babylonica* (KT449800), and *S. tetrasperma* (MF189169). The mVISTA software (Mayor et al. 2000) was employed to determine the interspecific variation among the six cp genome sequences (i.e. the five mentioned above and the investigated species). Differences in the types and gene sizes of IR, LSC, and SSC border regions among these species were also analyzed.

Phylogenetic relationships

The cp genome sequences of an additional 15 species (11 *Salix* species and four *Populus* species), consisting *S. interior* (KJ742926), *S. oreinoma* (MF189168), *S. arbutifolia*

(MG262340), *S. taoensis* (MG262369), *S. chaenomeloides* (MG262362), *S. hypoleuca* (MG262363), *S. purpurea* (KP019639), *S. minjiangensis* (MG262365), *S. paraplesia* (MG262366), *S. rehderiana* (MG262367), *S. rorida* (MG262368), *P. alba* (AP008956), *P. balsamifera* (KJ664927), *P. adenopoda* (KX425622), and *P. lasiocarpa* (KX641589), were selected and obtained from NCBI. The MAFFT v7.149 program was used to align the cpDNA sequences under default parameters (Katoh et al. 2005), and the alignment was trimmed by Gblocks_0.91b to remove low-quality regions (Castresana 2000). The maximum-likelihood (ML) methods were employed for genome-wide phylogenetic analysis using PhyML v3.0 (http://www.atgc-montpellier.fr/phyml/). Nucleotide substitution model selection was carried out using jModelTest 2.1.10 (Darriba et al. 2012) and Smart Model Selection in PhyML 3.0. The GTR+I+G model was selected for ML analyses with 1000 bootstrap replicates to calculate the bootstrap values of the obtained topology.

Results

Genome features

In the study, a total of 4177 Mb of raw data was generated by cp genome sequencing. After removing low-quality sequences and adaptors, 3869 Mb of clean data with a GC content of 36.72% were obtained. The Q30 value was high and reached 94.6%. After assembly, the complete cp genome of *C. arbutifolia* was 155684 bp in size (Table 1 and Fig. 1), and the unknown base rate was zero. Four regions were detected in the *C. arbutifolia* cp genome, including a large single-copy region (LSC, 84551 bp), a small single-copy region (SSC, 16217 bp), and two inverted repeat regions (IRa and IRb, 27458 bp each) (Table 1). The overall GC content of *C. arbutifolia* cp genome was 38.68%, with the IR regions having higher GC content (41.89%) than that in the LSC (34.39%) and SSC regions (30.94%) (Table 1).

Table 1 Base composition of *Chosenia arbutifolia* cp genome.

Region	Length (bp)	T/U%	C%	A%	G%
Genome	155684	32.04	18.66	31.29	18.02
LSC	84551	33.56	17.62	32.05	16.77
SSC	16217	34.19	16.30	34.86	14.64
IRa	27458	28.97	20.09	29.14	21.80
IRb	27458	29.14	21.80	28.97	20.09
Protein-coding genes	78300	31.59	17.50	30.87	20.04
tRNA	2803	25.01	23.69	22.12	29.18
rRNA	9048	18.77	23.61	25.80	31.83

Table 2 Functional annotation statistics based on the public databases (DB).

DB name	Total unigenes	Annotated unigenes	Percent
Nr	85	83	0.9765
GO	85	82	0.9647
COG	85	80	0.9412
KEGG	85	74	0.8706
SWSS	85	85	1
In_all_DB	85	72	0.8471
AT_least_one_DB	85	85	1

In total, 130 genes were predicted across the whole cp genome, of which 85 genes were protein-coding genes, 37 were tRNA genes, and eight were rRNA genes. To investigate their putative functions, all 85 protein-coding genes were compared against five databases, including Nr, GO, COG, KEGG, and SWSS (Table 2). Overall, 83 genes (97.65%) had hits in the Nr database, 82 genes (96.47%) were assigned at least one GO term, and 80 genes (94.12%) were annotated with COG classifications. The metabolic pathway analysis revealed that 74 genes (87.06%) were related. Additionally, all genes were assigned to the SWSS database and significantly corresponded with sequences from at least one public database.

The genes involved in photosynthesis and self-replication formed two dominant gene families, and they contained six gene groups and five gene groups, respectively (Table 3). Five genes with unknown functions (*ycf1, ycf2, ycf3, ycf4,* and *ycf15*) were also identified and were considered conserved open reading frames.

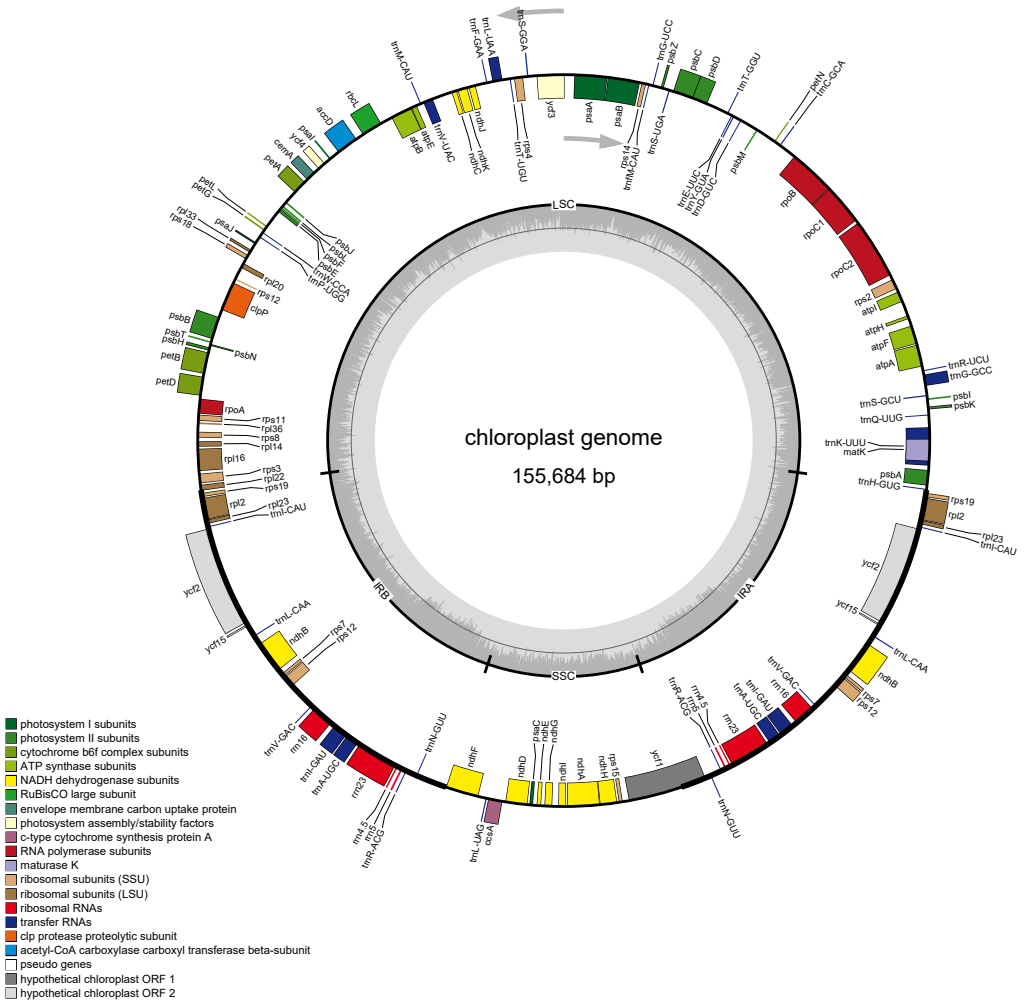

Figure 1 Circular gene map of *Chosenia arbutifolia*. Genes marked on the outside of the outer circle are transcripted counterclockwise, those on the inside are transcripted clockwise. Dashed area in the inner circle indicates the GC content of the chloroplast genome.

In the *C. arbutifolia* cp genome, 17 intron-containing genes were detected, three of which (*clpP, rps12*, and *ycf3*) contained two introns (Table 4). Of these 17 genes, 12 were located in the LSC region and four in the IR region, whereas only one was located in the SSC region. The exons varied in length, ranging from 6 bp to 1617 bp. The *trnK*-UUU gene had the largest intron, (2552 bp), followed by the *rpl16* gene (1122 bp) and *ndhA* gene (1113 bp). As a trans-spliced gene, the *rps12* gene had the smallest intron of 537 bp.

Repeat analysis

Four types of long repeat sequences were screened in the *C. arbutifolia* cp genome using the REPuter program, including 23 forward (F), 15 palindromic (P), one complement (C), and one reverse (R) repeat (Table 5). All repeat sizes ranged from 30 bp to 76 bp, except

for the R24 repeat, which reached a length of 27458 bp. Of the 40 repeats, eight (four forward and four palindromic) were associated with *ycf2* genes, and 11 were grouped with the other genes. A total of 21, 4, and 15 repeats were discovered in the LSC, SSC, and IR regions, respectively. Both complement (R5) and reverse (R21) repeats were found in the LSC region. Furthermore, more than half of the repeats (21; 52.5%) were located in the intergenic spacers (IGS), and 14 were distributed in the LSC region.

Using MISA, 221 putative simple sequence repeats (SSRs) were identified in the cp genome of *C. arbutifolia*. As shown in Table 6, 54 were of the compound type, and mono-nucleotide repeat motifs were the most abundant type (N = 195; 88.2%), followed by tetra- (N = 12; 5.4%) and di-nucleotides (N = 10; 4.5%), whereas tri- (N = 1; 0.045%) and penta-nucleotides (N = 1; 0.045%) were the least abundant types. Four types of mono-nucleotides and ten types of tetra-nucleotides were detected. The remaining four motif types were relatively scarce and contained two types of hexa-nucleotides and one type of di-, tri-, and penta-nucleotide. In the mono- and tetra-nucleotides, the A/T and ATTA repeat motifs were the most abundant, respectively.

Table 3 List of genes in the cp genome of *Chosenia arbutifolia*.

Gene functions	Gene groups	Gene names
Photosynthesis	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 NADH dehydrogenase	*ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Subunits of cytochrome	*petA, petB, petD, petG, petL, petN*
	Subunits of ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Large subunit of Rubisco	*rbcL*
Self-replication	Large subunits of ribosome	*rpl14, rpl16, rpl2, rpl20, rpl22, rpl23, rpl33, rpl36*
	Small subunits of ribosome	*rps11, rps12, rps14, rps15, rps18, rps19, rps2, rps3, rps4, rps7, rps8*
	DNA-dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	Ribosomal RNAs	*rrn16, rrn23, rrn4.5, rrn5*
	Transfer RNAs	*trnA*-UGC, *trnC*-GCA, *trnD*-GUC, *trnE*-UUC, *trnF*-GAA, *trnG*-GCC, *trnG*-UCC, *trnH*-GUG, *trnI*-CAU, *trnI*-GAU, *trnK*-UUU, *trnL*-CAA, *trnL*-UAA, *trnL*-UAG, *trnM*-CAU, *trnN*-GUU, *trnP*-UGG, *trnQ*-UUG, *trnR*-ACG, *trnR*-UCU, *trnS*-GCU, *trnS*-GGA, *trnS*-UGA, *trnT*-GGU, *trnT*-UGU, *trnV*-GAC, *trnV*-UAC, *trnW*-CCA, *trnY*-GUA, *trnfM*-CAU
Other genes	Maturase	*matK*
	Protease	*clpP*
	Envelope membrane protein	*cemA*
	Acetyl-CoA carboxylase	*accD*
	C-type cytochrome synthesis gene	*ccsA*
Unknown function	Conserved open reading frames	*ycf1, ycf15, ycf2, ycf3, ycf4*

Table 4 The genes within introns in the cp genome of *Chosenia arbutifolia*.

Gene	L	E1 (bp)	I1 (bp)	E2 (bp)	I2 (bp)	E3 (bp)
atpF	LSC	144	738			
clpP	LSC	71	840	292	594	228
ndhA	SSC	552	1113	546		
ndhB	IR	777	682	756		
petB	LSC	6	811	642		
petD	LSC	9	779	489		
rpl16	LSC	9	1122	399		
rpl2	IR	396	668	435		
rpoC1	LSC	453	775	1617		
*rps12**	LSC	114	-	29	537	232
ycf3	LSC	126	723	228	670	153
trnK-UUU	LSC	37	2552	29		
trnG-GCC	LSC	23	693	48		
trnL-UAA	LSC	37	587	50		
trnV-UAC	LSC	39	607	37		
trnI-GAU	IR	42	944	35		
trnA-UGC	IR	38	802	35		

Note: L:location; E: Exon ; I: Intron;The *rps12* gene is a trans-spliced gene with the 5' end located in the LSC region and the duplicated 3' end located in the IR region.

Comparative genome analysis

To evaluate genome divergence, the overall sequence identity analysis of the whole cp genomes of six typical Salicaceae species (*C. arbutifolia*, *P. trichocarpa*, *P. euphratica*, *S. suchowensis*, *S. babylonica*, and *S. tetrasperma*) was conducted with the mVISTA program using *C. arbutifolia* as a reference (Fig. 2). Overall, the comparison results showed that the LSC region was more divergent than the SSC and IR regions, and a higher divergence was observed in non-coding regions than in coding regions. The highest level of divergence was found in intergenic regions, such as *trnV-ndhC*. Slight divergences were also observed in some coding sequences, such as *ropC2*, *ndhF*, *ccsA*, and *ycf1*. However, the four rRNA genes, namely *rrn16*, *rrn23*, *rrn4.5,* and *rrn5*, were the most conserved among the six cp genomes analyzed.

IR contraction and expansion

The boundaries between the IRs and the two single-copy regions of the cp genome were compared between *C. arbutifolia* and five Salicaceae species (*P. trichocarpa, P. euphratica, S. suchowensis, S. babylonica*, and *S. tetrasperma*). As shown in Fig. 3, the *rpl22, ndhF, ycf1, rps19,* and *trnH*-GUG genes were detected at the junctions between IRs, LSC, and SSC. The *rpl22* gene was similarly located in the LSC region in all species, 49 - 52 bp away from the boundary of LSC and IRa. On the contrary, the *rps19* genes were present at the same location in the LSC region, 1 - 4 bp from the boundary of LSC and IRb. Moreover, the boundaries of SSC and IRb in all species were located in the *ycf1* gene, which extended into the SSC and IRb regions by variational lengths from 3675 bp to 3762 bp and from 1706 bp to 1748 bp, respectively. Interestingly, the duplications of the *ycf1* gene were not found at the borders of SSC and IRa in *C. arbutifolia* and *S. tetrasperma* cp genomes, where their location was replaced by the *ndhF* gene of the same length. Additionally, the replication of *ycf1* genes produced a pseudogene of divergent length at the SSC/IRa border in *P. trichocarpa, P. euphratica, S. suchowensis,* and *S. babylonica* cp genomes. The lengths of these pseudogenes ranged from 1408 bp to 1805 bp.

Phylogenetic analysis

To elucidate the position of *C. arbutifolia* within Salicaceae, a molecular phylogenetic tree was constructed using maximum likelihood (ML) and Bayesian inference (BI) methods (Fig. 4). A total of 20 cp genome sequences of Salicaceae species were obtained from GenBank, including six *Populus* species, 13 *Salix* species, and the same species as *C. arbutifolia* in that particular region (*S. arbutifolia*). Finally, 18 nodes were resolved by bootstrap analysis, ten of which had 100%

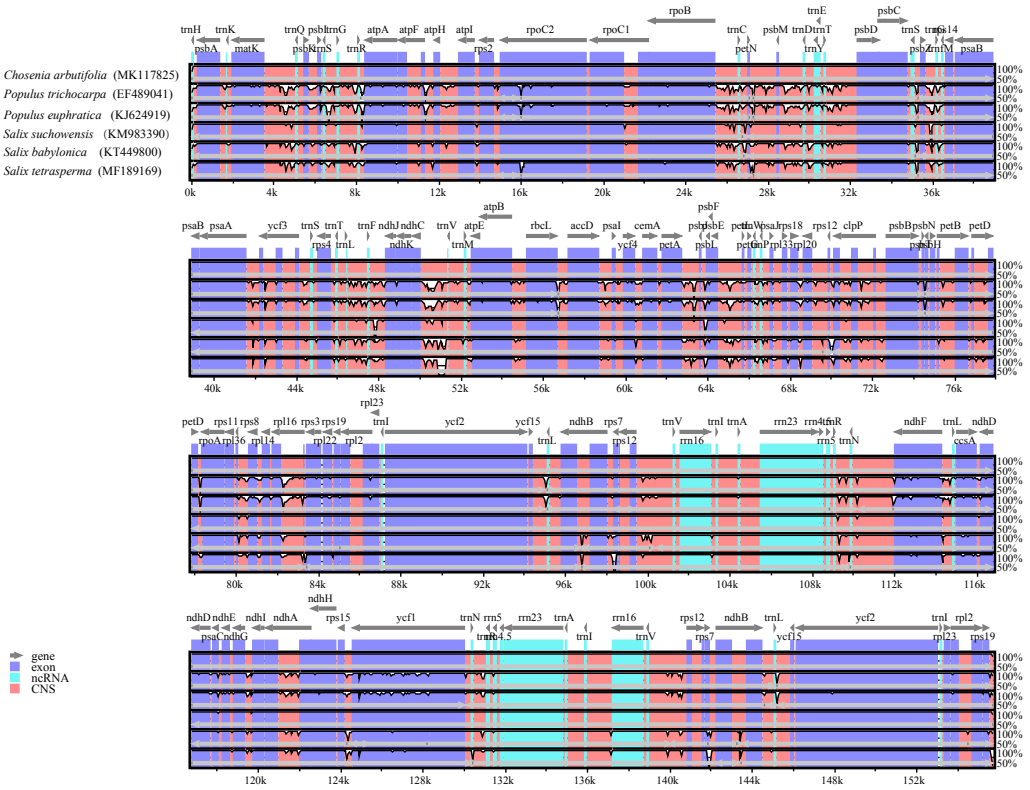

Figure 2 Comparison of six cp genomes using mVISTA. The x-axis represents the genome coordinate positions, and the y-axis represents the percentage of identity within 50-100%. Gray arrows above the alignment indicate the orientation of genes. Genome regions are color-coded as follows: protein-coding (exon, purple), tRNA- or rRNA-coding genes (blue), and conserved non-coding sequences (CNS, pink).

support. The remaining eight bootstrap values were > 90%. Two main groups with 100% bootstrap values were obtained, and they were consistent with the traditional morphological taxonomy in Salicaceae, namely the genera *Populus* and *Salix*. In the *Populus* group, *P. adenopoda* and *P. alba*, which belonged to section Leuce, were closely related. In addition, *P. balsamifera* and *P. trichocarpa*, which belonged to section Tacamahaca, were also clustered together. *P. lasiocarpa* and *P. euphratica*, which belonged to section Leucoides and Turanga, respectively, were in a single branch, separated from the other species. In the *Salix* group, all species belonged to different sections and form two clades with 100% bootstrap values. The smaller branch comprised five *Salix* species, in which *S. interior* was further away from the other four species and formed a single clade. The larger branch comprised *C. arbutifolia*, *S. arbutifolia*, and eight other *Salix* species. Thus, it could be seen that *C. arbutifolia* was closely related to *Salix* species.

Table 5 Long repeat sequences in the cp genome of *Chosenia arbutifolia*.

ID	L (bp)	Type	Position1	Position2	Gene	Region
R1	30	F	6306	65210	ISG	LSC; LSC
R2	30	F	6363	34980	*trnS*-GCU; *trnS*-UGA	LSC; LSC
R3	30	P	6366	44728	*trnS*-GCU; *trnS*-GGA	LSC; LSC
R4	30	F	7814	36187	*trnG*-GCC; *trnG*-UCC	LSC; LSC
R5	30	C	8223	50466	ISG	LSC; LSC
R6	30	P	14687	70589	ISG	LSC; LSC
R7	32	F	14825	14865	ISG	LSC; LSC
R8	43	P	27248	27248	ISG	LSC; LSC
R9	30	F	28811	28820	ISG	LSC; LSC
R10	37	P	30566	30566	ISG	LSC; LSC
R11	37	F	38391	40615	*psaB; psaA*	LSC; LSC
R12	55	F	38401	40625	*psaB; psaA*	LSC; LSC
R13	33	F	38432	40656	*psaB; psaA*	LSC; LSC
R14	30	F	38441	40665	*psaB; psaA*	LSC; LSC
R15	39	F	43307	99451	ISG	LSC; IRa
R16	39	P	43307	140747	ISG	LSC; IRb
R17	76	F	44541	57627	*accD*	LSC; LSC
R18	61	F	44556	57642	*accD*	LSC; LSC
R19	32	F	44585	57671	*accD*	LSC; LSC
R20	30	F	50233	50259	ISG	LSC; LSC
R21	30	R	63835	63872	ISG	LSC; LSC
R22	32	P	70591	83350	ISG	LSC; LSC
R23	30	F	72218	72245	ISG	LSC; LSC
R24	27458	P	84552	128227	*rpl22; ndhF*	IRa; IRb
R25	32	F	89841	89862	*ycf2*	IRa; IRa
R26	32	P	89841	150343	*ycf2*	IRa; IRb
R27	32	P	89862	150364	*ycf2*	IRa; IRb
R28	39	F	92255	92273	*ycf2*	IRa; IRa
R29	39	P	92255	147925	*ycf2*	IRa; IRb
R30	39	P	92273	147943	*ycf2*	IRa;IRb
R31	42	F	99449	120943	ISG	IRa;SSC
R32	32	F	99459	120953	ISG	IRa;SSC
R33	30	F	110151	110164	ISG	IRa; IRa
R34	30	P	110151	130043	ISG	IRa; IRb
R35	30	P	110164	130056	ISG	IRa; IRb
R36	42	P	120943	140746	ISG	SSC; IRb
R37	32	P	120953	140746	ISG	SSC; IRb
R38	30	F	130046	130059	ISG	IRb; IRb
R39	39	F	147925	147943	*ycf2*	IRb; IRb
R40	32	F	150343	150364	*ycf2*	IRb; IRb

Table 6 The composition of SSRs identified in the cp genome of *Chosenia arbutifolia*.

Motif types	Repeat motif	Number of repeats	Total
Mono-nucleotide	A/T	183	195
	G/C	12	
Di-nucleotide	AT/TA	10	10
Tri-nucleotide	AAT/ATT	1	1
Tetra-nucleotide	AAAC/GTTT	1	12
	AAAG/CTTT	2	
	AAAT/ATTT	2	
	AATG/ATTC	2	
	AATT/AATT	4	
	AGAT/ATCT	1	
Penta-nucleotide	AATAG/ATTCT	1	1
Hexa-nucleotide	AAAGTC/ACTTTG	1	2
	AATATC/ATATTG	1	
Compound		54	
Total			221

Discussion

In recent years, next generation sequencing technology accompanied by bioinformatics has developed considerably and has been extensively applied in genetic and genomic research, particularly of some woody plants (Wullschleger et al. 2013). Although 25 complete cp genome sequences, including those of 19 *Salix* species and one *Chosenia* species, have been deposited in NCBI to date, the information on the cp genome of *Salix* that comprised more than 500 species remains largely insufficient. Therefore, in the present study, the whole cp genome of *C. arbutifolia* was sequenced and compared with that of other Salicaceae species to elucidate the phylogenetic position of *C. arbutifolia* within Salicaceae.

After *de novo* assembly with an unknown base rate of zero, we found that the *C. arbutifolia* cp genome was 155684 bp in size. The obtained genome size was slightly greater than that of *S. arbutifolia* deposited in NCBI (KX781246), which is 155661 bp in length. However, the obtained genome size was over 600 bp longer than that of another *S. arbutifolia* individual (155055 bp, MG262340). Furthermore, of the 20 investigated species, the size of the *C. arbutifolia* cp genome is smaller than that of four species, namely *S. babylonica* (156819 bp of KT449800 and 155697 bp of MG262361), *S. interior* (156620 bp of KJ742926), *S. chaenomeloides* (156154 bp of MG262362), and *S. triandra* (155821 bp of MK722343). It is longer than the remaining 15 species. Some previous studies have proposed that the cp genome size of most land plants ranged from 107 kb to 218 kb in length and has a typical quadripartite structure (Daniell et al. 2016) that might be influenced by the length variation of

Figure 3 Comparison of cp genome boundaries between LSC, SSC, and IRs among six Salicaceae species.

IRs (Wang et al. 2008, Guisinger et al. 2011). Similar to these reports, length divergences in IRs were also observed in *C. arbutifolia* (Fig. 3) as well as in some other *Salix* species (Wu et al. 2015, Sun et al. 2018, Chen et al. 2019). These divergences are considered an important factor for the variation in cp genome size across species.

Generally, the cp genome includes 120 - 130 genes with photosynthesis functions, transcription, and translation (Wicke et al. 2011, Daniell et al. 2016). As expected, 130 genes were predicted in the present study. Excluding tRNA and rRNA genes, 85 genes were protein-coding genes, which is comparable to the number obtained for an *S. arbutifolia* individual (87 protein-coding genes, KX781246), but a little higher than the number obtained for another *S. arbutifolia* individual (80 protein-coding genes, MG262340). Furthermore, the two genes in *S. arbutifolia* (KX781246) confirmed as the *ycf68* gene were absent in the present study. However, the *ycf68* gene was also discovered in *S. wilsonii*, where it was considered to be a pseudogene located in an intron (Chen et al. 2019). As described in the report by Raubeson, the *ycf68* gene in rice, corn, and *Pinus* species seems to be a functional protein-coding gene, but in the majority of other species, it is likely to be a nonfunctional gene because of abundant frameshifts and premature stop codons (Raubeson et al. 2007). Nevertheless, a pseudogene *ycf1* was screened in the SSC region of *S. arbutifolia* (MG262340) but missed in the present study. A similar situation is also demonstrated in Fig. 3, in which

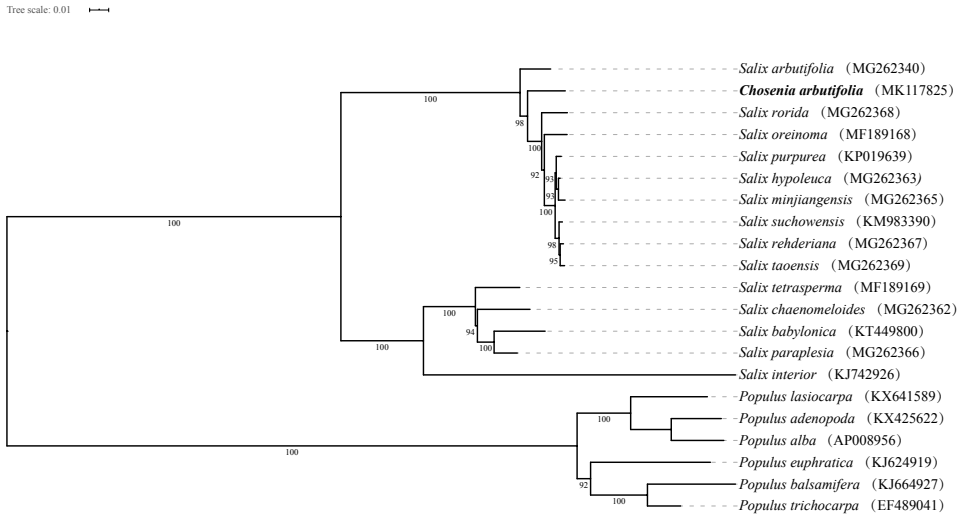

Figure 4 Phylogenetic tree of *Chosenia arbutifolia* and other Salicaceae species. Bootstrap values are shown on each node. The *Salix arbutifolia* is the synonym species as *Chosenia arbutifolia*.

the pseudogene *ycf1* existed in *S. suchowensis* and *S. babylonica*, but disappeared in *C. arbutifolia* and *S. tetrasperma*. As the second longest gene in the cp genome, the *ycf1* gene normally crosses the boundary of the IR and SSC regions and is considered a pseudogene in most plant cp genomes (De Las Rivas et al. 2002). Although the function of the *ycf1* gene has not been clarified to date, this gene might be indispensable and essential for plant survival (Drescher et al. 2000).

SSRs or microsatellites, are short tandem repetitive sequences consisting of one to six base pairs and are abundantly scattered within the nucleus (Tóth et al. 2000), chloroplast (Powell et al. 1999), and mitochondrial genomes (Soranzo et al. 1999). SSRs have been widely employed as ideal molecular markers for genetic and genomic studies (Provan et al. 2001, Varshney et al. 2005). In the present study, a total of 221 cp SSRs was identified. Such a frequency of occurrence is in accordance with the number of SSRs in *S. babylonica* (227, KT449800) and *S. tetrasperma* (229, MF189169), but much higher than that in *S. suchowensis* (148, Sun et al. 2018) and *S. wilsonii* (155, Chen et al. 2019) under a similar stringent criterion used for mining. Additionally, the frequencies of SSR occurrence in other woody plants are quite diverse, such as 415 in *A. miaotaiense* (Zhao et al. 2018), 151 in *P. taeda* (Asaf et al. 2018), 65 in *Q. acutissima* (Li et al. 2018), and 188 in *M. glyptostroboides* (Chen et al. 2015), which might be due to differences in genome size, mining tools, or search parameters. However, in all of the above mentioned species, the mono-nucleotide was the most abundant type of repeat, within which the A/T motif was predominant. As described by Powell (Powell et al. 1995), the number of A/T repeats is commonly lower than 15, but in the present study, 16 and 17 A-repeats were also observed, as well as 17 T-repeats. Otherwise, similar to *S. suchowensis* (Sun et al. 2018) and *S. wilsonii* (Chen et al. 2019), all of the ten recorded di-nucleotides in *C. arbutifolia* were composed of AT/TA, and the tri- and penta-nucleotide types were scarce. On the contrary, two hexa-nucleotide types were detected in *C. arbutifolia*, which was not detected in these two *Salix* species. Overall, a significant bias in the base composition of AT-rich repeat motifs was observed in the *C. arbutifolia* cp genome, which

is consistent with the results for *Salix* species (Sun et al. 2018, Chen et al. 2019) and other plants (Ebert et al. 2009). Compared to those EST-SSR markers developed in *Salix* (Tian et al. 2019), the cp SSRs in *C. arbutifolia* still have great potential to provide useful resources for species identification and evolution studies.

Due to the presence of many species, innumerable interspecific hybrids, insufficient morphological characteristics, and ineffective molecular description, the taxonomy and phylogenetic relationships of the genus *Salix* are exceedingly ambiguous. Furthermore, the phylogenetic position of *C. arbutifolia* within Salicaceae remains highly disputable. In the present study, a phylogenetic tree which includs *Populus*, *Salix*, and *Chosenia* species was constructed. Six *Populus* species in different sections were clustered as a main clade, which is in accordance with Lu's results (Lu et al. 2020). As a single branch, *S. interior* showed a distant relationship with other *Salix* species, which was also quite similar to previous reports (Chen et al. 2010, 2019, Lu et al. 2020). Although numerous different morphological characteristics were observed among *C. arbutifolia* and other *Salix* species, all evidence from earlier reports revealed by the *rbcL* gene sequences (Azuma et al. 2000), *matK* gene sequences (Hardig et al. 2010), whole cp genome sequences (Chen et al. 2010, Zhang et al. 2018, Chen et al. 2019, Lu et al. 2020), and ribosomal DNA sequences (Leskinenet al. 1999, Hardig et al. 2010) show that *C. arbutifolia* is closely related to *Salix* species. Our study provides additional support to the close relationship of *C. arbutifolia* to *Salix* species. This suggests that the placement of *C. arbutifolia* in a separate genus might be inappropriate and that it should rather be a member of the genus *Salix*.

Conclusions

In this study, the complete cp genome of *C. arbutifolia* was sequenced and characterized. Through a comparative analysis, the *C. arbutifolia* cp genome was found to be similar to other *Populus* and *Salix* species in structure and organization, but different in genome size, gene number, and gene order. The results of phylogenetic relationships analysis among 20 Salicaceae species demonstrate that *C. arbutifolia* is closely related with *Salix* species and should not be treated as a separate genus. Overall, our work provides comprehensive evidence of the phylogenetic taxonomy of *C. arbutifolia* within Salicaceae and has generated useful resources for future research regarding Salicaceae species.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant No. 31670662), the Independent Scientific Research Project of Jiangsu Academy of Forestry (Grant No. 2018022-1). The authors are grateful to Shanghai BIOZERON Biotechnology Co., Ltd. for assistance with the bio-information analysis.

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

Data archiving statement

The whole chloroplast genome sequences of *C. arbutifolia* has been deposited in the GenBank database under the accession number MK117825.

References

Asaf S., Khan A.L., Khan M.A., Shahzad R., Lubna, Kang S.M., Al-Harrasi A., Al-Rawahi A., Lee I.J. 2018. Complete chloroplast genome sequence and comparative analysis of loblolly pine (*Pinus taeda* L.) with related species. PLoS ONE 13(3): e0192966. https://doi.org/10.1371/journal.pone.0192966

Azuma T., Kajita T., Yokoyama J., Ohashi H. 2000. Phylogenetic relationships of *Salix* (Salicaceae) based on *rbcL* sequence data. Am. J. Bot. 87(1): 67-75. https://doi.org/10.2307/2656686

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17: 540-552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Chen J.H., Hao Z.D., Xu H.B., Yang L.M., Liu G.X.,

Sheng Y., Zheng C., Zheng W.W., Cheng T.L., Shi J.S. 2015. The complete chloroplast genome sequence of the relict woody plant *Metasequoia glyptostroboides* Hu et Cheng. Front. Plant Sci. 6: 447. https://doi.org/10.3389/fpls.2015.00447

Chen J.H., Sun H., Wen J., Yang Y.P. 2010. Molecular phylogeny of *Salix* L. (Salicaceae) inferred from three chloroplast datasets and its systematic implications. Taxon 59: 29-37. https://doi.org/10.1002/tax.591004

Chen Y.N., Hu N., Wu H.T. 2019. Analyzing and characterizing the chloroplast genome of *Salix wilsonii.* BioMed Res. Int. Article ID: 5190425. https://doi.org/10.1155/2019/5190425

Cronn R., Liston A., Parks M., Gernandt D.S., Shen R., Mockler T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res. 36: e122. https://doi.org/10.1093/nar/gkn502

Daniell H., Lin C.S., Yu M., Chang W.J. 2016. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. Genome Biol. 17: 134. https://doi.org/10.1186/s13059-016-1004-2

Darriba D., Taboada G.L., Doallo R., Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9: 772. https://doi.org/10.1038/nmeth.2109

De Las Rivas J., Lozano J.J., Ortiz A.R. 2002. Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny, and deduced evolutionary patterns. Genome Res. 12(4): 567-583. https://doi.org/10.1101/gr.209402

Drescher A., Ruf S., Calsa Jr 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. https://doi.org/10.1046/j.1365-313X.2000.00722.x

Ebert D., Peakall R. 2009. Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. Mol. Ecol. Resour. 9(3): 673-690. https://doi.org/10.1111/j.1755-0998.2008.02319.x

Eguiluz M., Rodrigues N.F., Guzman F., Yuyama P., Margis R. 2017. The chloroplast genome sequence from E*ugenia uniflora*, a Myrtaceae from Neotropics. Plant Syst. Evol. 303: 1199-1212. https://doi.org/10.1007/s00606-017-1431-x

Feng C. H., He C. Y., Wang Y., Zeng Y. F., Zhang J. G. 2019. Phylogenetic position of *Chosenia arbutifolia* in the Salicaceae inferred from whole chloroplast genome. Forest Res. 32(2): 73-77. https://doi.org/10.13275/j.cnk.lykxyj.2019.02.011

Guisinger M.M., Kuehl J.V., Boore J.L., Jansen R.K. 2011. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol. Biol. Evol. 28: 583-600. https://doi.org/10.1093/molbev/msq229

Hardig T.M., Anttila C.K., Brunsfeld S.J. 2010. A phylogenetic analysis of *Salix* (Salicaceae) based on *matK* and ribosomal DNA sequence data. J. Bot. Article ID: 197696. https://doi.org/10.1155/2010/197696

Ingvarsson P.K., Ribstein S., Taylor D.R. 2003. Molecular revolution of insertions and deletion in the chloroplast genome of Silene. Mol. Biol. Evol. 20: 1737-1740. https://doi.org/10.1093/molbev/msg163

Jheng C.F., Chen T.C., Lin J.Y., Chen T.C., Wu W.L., Chang C.C. 2012. The comparative chloroplast genomic analysis of photosynthetic orchids and developing DNA markers to distinguish Phalaenopsis orchids. Plant Sci. 190:62-73. https://doi.org/10.1016/j.plantsci.2012.04.001

Kadis I. 2005. Chosenia: an amazing tree of Northeast Asia. Arnoldia 63: 8-17.

Katoh K., Kuma K.I., Toh H., Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33: 511-518. https://doi.org/10.1093/nar/gki198

Kurtz S., Choudhuri J.V., Ohlebusch E., Schleiermacher C., Stoye J., Giegerich R. 2001. REPuter: The manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 29: 4633-4642. https://doi.org/10.1093/nar/29.22.4633

Leskinen E., Alström-Rapaport C. 1999. Molecular phylogeny of Salicaceae and closely related Flacourtiaceae: evidence from 5.8 S, ITS 1 and ITS 2 of the rDNA. Plant Syst. Evol. 215: 209-227. https://doi.org/10.1007/BF00984656

Li X., Li Y.F., Zang M.Y., Li M.Z., Fang Y.M. 2018. Complete chloroplast genome sequence and phylogenetic analysis of *Quercus acutissima*. Int. J. Mol. Sci. 19: 2443. https://doi.org/10.3390/ijms19082443

Lohse M., Drechsel O., Bock R. 2007. OrganellarGenomeDRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr. Genet. 52: 267-274. https://doi.org/10.1007/s00294-007-0161-y

Lu D.Y., Zhang L., Zhang G.S., Hao L. 2020. Chloroplast genome structure and variation of Salicaceae plants. J. Northwest A&F Univ. (Nat. Sci. Ed.), 48(2): 87-94. https://doi.org/10.13207/j.cnki.jnwafu.2020.02.011

Mader M., Pakull B., Blanc-Jolivet C., Paulini-Drewes M., Bouda Z.H.N., Degen B., Small I., Kersten B. 2018. Complete chloroplast genome sequences of four Meliaceae species and comparative analyses. Int. J. Mol. Sci. 19: 701. https://doi.org/10.3390/ijms19030701

Mayor C., Brudno M., Schwartz J.R., Poliakov A., Rubin E.M., Frazer K.A., Pachter L.S., Dubchak I. 2000. Vista: Visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16: 1046-1047. https://doi.org/10.1093/bioinformatics/16.11.1046

McPherson H., van der Merwe M., Delaney S.K., Edwards M.A., Henry R.J., McIntosh E., Rymer P.D., Milner M.L., Siow J., Rossetto M. 2013. Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. BMC Ecol. 13: 8. https://doi.org/10.1186/1472-6785-13-8

Moskalyuk T.A. 2016. *Chosenia arbutifolia* (Salicaceae): life strategies and introduction perspectives. Sibirskij Lesnoj Zurnal (Sib J For Sci) 3: 34-45 (in English with Russian abstract).

Nakai T. 1920. *Chosenia,* a new genus of Salicaceae. Bot. Mag. (Tokyo) 34: 66-69.

Neuhaus H.E., Emes M.J. 2000. Nonphotosynthetic metabolism in plastids. Annu. Rev. Plant Biol. 51: 111-140. https://doi.org/10.1146/annurev.arplant.51.1.111

Ohashi H. 2001. Salicaceae of Japan. Sci. Rep. Tôhoku Imp. Univ. Ser. 4 40: 269-396.

Palmer J.D. 1985. Comparative organization of chloroplast genomes. Annu. Rev. Genet. 19: 325-354. https://doi.org/10.1146/annurev.ge.19.120185.001545

Powell W., Morgante M., Andre C., McNicol J.W., Machray G.C., Doyle J.J., Tingey S.V., Rafalski J.A. 1995. Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome. Curr. Biol. 5: 1023-1029. https://doi.org/10.1016/S0960-9822(95)00206-5

Provan J., Powell W., Hollingsworth P.M. 2001. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends Ecol. Evol.

16(3): 142-147. https://doi.org/10.1016/S0169-5347(00)02097-8
Raubeson L.A., Peery R., Chumley T.W., Dziubek C., Fourcade H.M., Boore J.L., Jansen R.K. 2007. Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. BMC Genomics 8(1): 174. https://doi.org/10.1186/1471-2164-8-174
Skvortsov A.K. 1999. Willows of Russia and adjacent countries. Taxonomical and geographical revision (English translation with additions). Univ. Joensuu. Fac. Math. Nat. Sci. Rep. Ser. 39: 1-307.
Song Y., Dong W.P., Liu B., Xu C., Yao X., Gao J., Corlett R.T. 2015. Comparative analysis of complete chloroplast genome sequences of two tropical trees *Machilus yunnanensis* and *Machilus balansae* in the family Lauraceae. Front. Plant Sci. 6: 662. https://doi.org/10.3389/fpls.2015.00662
Soranzo N., Provan J., Powell W. 1999. An example of microsatellite length variation in the mitochondrial genome of conifers. Genome 42: 158-161. https://doi.org/10.1139/g98-111
Sun C.R., Li J., Dai X.J., Chen Y.N. 2018. Analysis and characterization of the *Salix suchowensis* chloroplast genome. J. For. Res. 29(4): 1003-1011. https://doi.org/10.1007/s11676-017-0531-3
Tangphatsornruang S., Uthaipaisanwong P., Sangsrakru D., Chanprasert J., Yoocha T., Jomchai N., Tragoonrung S. 2011. Characterization of the complete chloroplast genome of *Hevea brasiliensis* reveals genome rearrangement, RNA editing sites and phylogenetic relationships. Gene 475: 104-112. https://doi.org/10.1016/j.gene.2011.01.002
Tian X.Y., Zheng J.W., Jiao Z.Y., Zhou J., He K.Y., Wang B.S., He X.D. 2019. Transcriptome sequencing and EST-SSR marker development in *Salix babylonica* and *S. suchowensis*. Tree Genet. Genomes 15: 9. https://doi.org/10.1007/s11295-018-1315-4
Tóth G., Gáspári Z., Jurka J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10: 967-98. https://doi.org/10.1101/gr.10.7.967
Tu Z.Y. 1982. Breeding and cultivation of *Salix*. Jiangsu Science and Technology Press, Nanjing, p. 154-196.
Varshney R.K., Graner A., Sorrells M.E. 2005. Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 23(1): 48-55. https://doi.org/10.1016/j.tibtech.2004.11.005
Walker J.F., Zanis M.J., Emery N.C. 2014. Comparative analysis of complete chloroplast genome sequence and inversion variation in *Lasthenia burkei* (Madieae, Asteraceae). Am. J. Bot. 101: 722-729. https://doi.org/10.3732/ajb.1400049
Wang R.J., Cheng C.L., Chang C.C., Wu C.L., Su T.M., Chaw S.M. 2008. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. BMC Evol. Biol. 8: 36. https://doi.org/10.1186/1471-2148-8-36
Wang Z., Fang C.F. 1984. Salicaceae. In: Flora Reipublicae Popularis Sinicae. Science Press, Beijing, pp 79-81.
Wicke S., Schneeweiss G.M., DePamphilis C.W., Müller K.F., Quandt D. 2011. The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. Plant Mol. Biol. 76: 273-297. https://doi.org/10.1007/s11103-011-9762-4
Wu Z.Q. 2015. The new completed genome of purple willow (*Salix purpurea*) and conserved chloroplast genome structure of Salicaceae. JNSCI 1(3): e49.
Wullschleger S.D., Weston D.J., DiFazio S.P., Tuskan G.A. 2013. Revisiting the sequencing of the first tree genome: *Populus trichocarpa.* Tree Physiol. 33: 357-364. https://doi.org/10.1093/treephys/tps081
Wyman S.K., Jansen R.K., Boore J.L. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252-3255. https://doi.org/10.1093/bioinformatics/bth352
Zhang L., Xi Z.X., Wang M.C., Guo X.Y., Ma T. 2018. Plastome phylogeny and lineage diversification of Salicaceae with focus on poplars and willows. Ecol. Evol. 8: 7817-7823. https://doi.org/10.1002/ece3.4261
Zhao J.T., Xu Y., Xi L.J., Yang J.W., Chen H.W., Zhang J. 2018. Characterization of the chloroplast genome sequence of *Acer miaotaiense*: comparative and phylogenetic analyses. Molecules 23: 1740. https://doi.org/10.3390/molecules23071740