Characterization of the complete chloroplast genomes of five *Populus* species from the western Sichuan plateau, southwest China: comparative and phylogenetic analyses (#32464)

First submission

Editor guidance

Please submit by **26 Nov 2018** for the benefit of the authors (and your $200 publishing discount).

✓ **Structure and Criteria**
Please read the 'Structure and Criteria' page for general guidance.

! **Custom checks**
Make sure you include the custom checks shown below, in your review.

Raw data check
Review the raw data. Download from the materials page.

Image check
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the materials page.

10 Figure file(s)
12 Table file(s)
1 Raw data file(s)
1 Other file(s)

! Custom checks

DNA data checks

! Have you checked the authors data deposition statement?

! Can you access the deposited data?

! Has the data been deposited correctly?

! Is the deposition information noted in the manuscript?

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

🗋 You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- ✓ Clear, unambiguous, professional English language used throughout.
- ✓ Intro & background to show context. Literature well referenced & relevant.
- ✓ Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- ✓ Figures are relevant, high quality, well labelled & described.
- ✓ Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- ✓ Original primary research within Scope of the journal.
- ✓ Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- ✓ Rigorous investigation performed to a high technical & ethical standard.
- ✓ Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- ℹ Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- ✓ Data is robust, statistically sound, & controlled.
- ✓ Speculation is welcome, but should be identified as such.
- ✓ Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Characterization of the complete chloroplast genomes of five *Populus* species from the western Sichuan plateau, southwest China: comparative and phylogenetic analyses

Dan Zong [1, 2] , **Anpei Zhou** [1, 2] , **Yao Zhang** [1, 2] , **Xinlian Zou** [1, 2] , **Dan Li** [3] , **Anan Duan** [1, 2, 4] , **Chengzhong He** Corresp. [1, 2, 4]

[1] Key Laboratory for Forest Genetic and Tree Improvement &Propagation in Universities of Yunnan Province, Southwest Forestry University, Kunming, Yunnan, China
[2] Key Laboratory of State Forestry Administration on Biodiversity Conservation in Southwest China, Southwest Forestry University, Kunming, Yunnan, China
[3] Yunnan Academy of Biodiversity, Southwest Forestry University, Kunming, Yunnan, China
[4] Key Laboratory for Forest Resources Conservation and Use in the Southwest Mountains of China, Ministry of Education, Southwest Forestry University, Kunming, Yunnan, China

Corresponding Author: Chengzhong He
Email address: hcz70@163.com

Species of the genus *Populus*, which is widely distributed in the northern hemisphere from subtropical to boreal forests, are one of the most commercially exploited groups of forest trees. In this study, the complete chloroplast genomes of five *Populus* species (*Populus cathayana*, *P. kangdingensis*, *P. pseudoglauca*, *P. schneideri* and *P. xiangchengensis*) were compared. The chloroplast genomes of the five *Populus* species are very similar. The total chloroplast genome sequence lengths for the five plastomes were 156,789 bp, 156,523 bp, 156,512 bp, 156,513 bp and 156,465 bp, respectively. A total of 130 genes were identified in each genome, including 85 protein-coding genes, 37 tRNA genes and eight rRNA genes. Seven genes were duplicated in the protein-coding genes, whereas 11 genes were duplicated in the RNA genes. The GC content was 36.7% for all plastomes. We analyzed nucleotide substitutions, small inversions, SSRs and long repeats in the chloroplast genomes and found nine divergence hotspots (*ccsA-ndhD, ndhC-trnV*, *psbZ-trnfM*, *trnG-atpA*, *trnL-ndhJ*, *trnR-trnN*, *ycf4-cemA*, *ycf1-ndhF*, and *trnR-trnN*), which could be useful molecular genetic markers for future population genetic and phylogenetic studies. We also observed that two genes (*rpoC2* and *rbcL*) were subject to positive selection. The phylogenetic analysis based on whole cp genomes showed that *P. schneideri* had a close relationship with *P. kangdingensis* and *P. pseudoglauca*, while *P. xiangchengensis* was a sister to *P. cathayana.*

Characterization of the complete chloroplast genomes of five *Populus* species from the western Sichuan plateau, southwest China: comparative and phylogenetic analyses

Dan Zong[1,2], Anpei Zhou[1,2], Yao Zhang[1,2], Xinlian Zou[1,2], Dan Li[3], Anan Duan[1,2,4] and Chengzhong He[1,2,4]

[1] Key Laboratory for Forest Genetic and Tree Improvement &Propagation in Universities of Yunnan Province, Southwest Forestry University, Kunming, Yunnan, China

[2] Key laboratory of State Forestry Administration on Biodiversity Conservation in Southwest China, Southwest Forestry University, Kunming, Yunnan, China

[3] Yunnan Academy of Biodiversity, Southwest Forestry University, Kunming, Yunnan, China;

[4] Key Laboratory for Forest Resources Conservation and Use in the Southwest Mountains of China, Ministry of Education, Southwest Forestry University, Kunming, Yunnan, China

Corresponding Author:
Chengzhong He,

Bailong Road, Kunming, Yunnan, 650224, China

Email address: Chengzhong He hecz@swfu.edu.cn

Characterization of the complete chloroplast genomes of five *Populus* species from the western Sichuan plateau, southwest China: comparative and phylogenetic analyses

Dan Zong[1,2], Anpei Zhou[1,2], Yao Zhang[1,2], Xinlian Zou[1,2], Dan Li[3], Anan Duan[1,2,4] and Chengzhong He[1,2,4]

[1] Key Laboratory for Forest Genetic and Tree Improvement &Propagation in Universities of Yunnan Province, Southwest Forestry University, Kunming, China

[2] Key Laboratory of State Forestry Administration on Biodiversity Conservation in Southwest China, Southwest Forestry University, Kunming, China

[3] Yunnan Academy of Biodiversity, Southwest Forestry University, Kunming, China;

[4] Key Laboratory for Forest Resources Conservation and Use in the Southwest Mountains of China, Ministry of Education, Southwest Forestry University, Kunming, China

Corresponding Author:
Chengzhong He,
Bailong Road, Kunming, Yunnan, 650224, China
Email address: Chengzhong He hecz@swfu.edu.cn

Abstract

Species of the genus *Populus*, which is widely distributed in the northern hemisphere from subtropical to boreal forests, are one of the most commercially exploited groups of forest trees. In this study, the complete chloroplast genomes of five *Populus* species (*Populus cathayana*, *P. kangdingensis*, *P. pseudoglauca*, *P. schneideri* and *P. xiangchengensis*) were compared. The chloroplast genomes of the five *Populus* species are very similar. The total chloroplast genome sequence lengths for the five plastomes were 156,789 bp, 156,523 bp, 156,512 bp, 156,513 bp and 156,465 bp, respectively. A total of 130 genes were identified in each genome, including 85 protein-coding genes, 37 tRNA genes and eight rRNA genes. Seven genes were duplicated in the protein-coding genes, whereas 11 genes were duplicated in the RNA genes. The GC content was 36.7% for all plastomes. We analyzed nucleotide substitutions, small inversions, SSRs and long repeats in the chloroplast genomes and found nine divergence hotspots (*ccsA-ndhD, ndhC-trnV*, *psbZ-trnfM*, *trnG-atpA*, *trnL-ndhJ*, *trnR-trnN*, *ycf4-cemA*, *ycf1-ndhF*, and *trnR-trnN*), which could be useful molecular genetic markers for future population genetic and phylogenetic studies. We also observed that two genes (*rpoC2* and *rbcL*) were subject to positive selection. The phylogenetic analysis based on whole cp genomes showed that *P. schneideri* had a close

relationship with *P. kangdingensis* and *P. pseudoglauca*, while *P. xiangchengensis* was a sister to *P. cathayana.*

Subjects Evolutionary Studies, Genomics, Plant science
Keywords *Populus*; western Sichuan Plateau; chloroplast genome; phylogenetic relationship

Introduction

The species of the genus *Populus*, collectively known as poplar, are widely distributed in the northern hemisphere from subtropical to boreal forests and one of the most commercially exploited groups of forest trees (*Hamzeh & Dayanandan, 2004*). Because of their small genome size, fast growth rates, profuse vegetative propagation, adaptability to a variety of ecological sites, and their wood's numerous uses, *Populus* species have become one of the most economically important groups of forest trees and a model organism for the study of tree biology (*Braatne, et al., 1992*; *Stettler, et al., 1996*). According to a recent classification, the genus *Populus* is classified into six sections (*Fang, et al., 1999*; *Zsuffa, 1975*; *Eckenwalder, 1996*). To date, 100 and more *Populus* species have been reported worldwide, of which approximately 53 are endemic to China.

As a concentrated area of *Populu*s resources in southwest China, the western Sichuan Plateau is dominated by mountainous and plateau geomorphology, and the mountains play a critical role in isolating plant distribution (*He, et al., 2015*). Meanwhile, the complex and unique natural and geographical conditions of this area provide not only diversified refuges where plants retreat in response to climatic changes but also great opportunities to develop new hybrid species (*Lu, et al., 2014*). However, the extensive interspecific hybridization and the high levels of morphological variation in *Populus* have posed great difficulties in species delimitation for systematic and comparative evolutionary studies (*Hamzeh & Dayanandan, 2004*; *Eckenwalder, 1996*; *Cronk, 2005*).

Populus kangdingensis, *P. pseudoglauca, P. schneideri* and *P. xiangchengensis* are native to the western Sichuan Plateau, and they are distributed at altitudes above 3000 m and even above 4000 m, whereas *P. cathayana* widely occurs in China, at altitudes ranging from 800 m to 3000 m. All five species overlap in the western Sichuan Plateau. Previous research has focused on their phylogenetic relationships. Liu & Fu (*2004*) considered *P. xiangchengensis* a hybridization of *P. schneideri* and *P. pseudoglauca* based on morphological characteristics, while another study suggested that *P. xiangchengensis* was a likely hybrid species of *P. kangdingensis* and *P. pseudoglauca* (*Wan et al., 2009*). *P. schneideri* was classified into section *Tacamahaca*. Meanwhile, it was also considered a natural hybrid formed by *P. kangdingensis* and *P. cathayana* (Chen, 2007; Wang, 2012). *P. pseudoglauca* was originally classified in section *Leucoides*, while it was suggested to be assigned to the section *Tacamahaca* (*Zhao, 1994*), and this assignment was supported by inter-simple sequence repeat (ISSR) markers and nuclear internal transcribed spacer (ITS) sequence (*Wang, 2012*). All these findings suggested that the phylogenetic relationship of the five *Populus* species is rather complex and unclear.

Organellar DNA, as the uniparental inheritance, is well conserved and allows for the development of informative universal markers (*Howe, et al., 2003*; *Wicke, et al., 2011*). The chloroplast (cp) genome, because of its relatively conserved size, gene content, structure and slow rate of nucleotide substitution within protein-coding genes, has been an ideal source of data on the phylogenetic relationships of plant taxa and their evolution and has been used to make significant contributions concerning evolutionary mechanisms for species and phylogenetic reconstruction (*Khan et al., 2012*; *Asheesh & Vinav, 2012*; *Liu et al., 2017*).

With the development of sequencing technology in recent years, in addition to nuclear genome sequences, cp genes, gene spacer regions, and cp genome information have been widely used to study plant molecular systematics. Whole cp genomes of several species from the genus *Populus* have been sequenced and deposited in the GenBank database. Here, we compare the complete cp genome of *P. cathayana*, *P. kangdingensis*, *P. pseudoglauca*, *P. schneideri* and *P. xiangchengensis* (MH910611) (*Zong et al., in press*). The codon usage bias, sequence divergences, mutation events, pattern of single nucleotide polymorphisms (SNP) and distribution of SSRs are compared, and phylogenetic tree is reconstructed based on 27 complete cp genome sequences from Salicaceae. Our study provides cp genomic information for further phylogenetic reconstruction, molecular evolution research, selective breeding and crossbreeding of the genus *Populus*.

Materials & Methods

Plant materials and DNA extraction

The fresh leaves of *P. cathayana* were collected in Kangding (101°56'26"E, 29°59'36"N, Sichuan, China; Altitude: 3109 m), while the samples of *P. kangdingensis*, *P. pseudoglauca*, *P. schneideri* and *P. xiangchengensis* were collected in Kangding (101°36'43"E, 30°05'20"N, Sichuan, China; Altitude: 3554 m), Yajiang (100°54'06"E, 29°59'14"N, Sichuan, China; Altitude: 3598 m), Litang (101°36'43"E, 30°05'20"N, Sichuan, China; Altitude: 4018 m) and Xiangcheng (99º40'33"E, 28º55'47"N, Sichuan, China; Altitude: 3530 m), respectively. The voucher specimens of the five species were deposited at the herbarium of Southwest Forestry University, Kunming, China. Total genome DNA was extracted with the Ezup plant genomic DNA prep kit (Sangon Biotech, Shanghai, China), and DNA samples were properly stored at the Key Laboratory of State Forestry Administration on Biodiversity Conservation in Southwest China, Southwest Forestry University, Kunming, China.

Genome sequencing, assembly and annotation

Total DNA was used to generate libraries with an average insert size of 400 bp, which were sequenced using the Illumina HiSeq X platform. Approximately 15.0 GB of raw data were generated with 150 bp paired-end read lengths. Then, the raw data were used to assemble the complete cp genome using GetOrganelle software (*Jin et al., 2018*) with *P. trichocarpa* as the reference. Genome annotation was performed with the program Geneious R8 (Biomatters Ltd, Auckland, New Zealand) by comparing the sequences with the cp genome of *P. trichocarpa.* The tRNA genes were further confirmed through online tRNAscan-SE web servers (*Schattner et al., 2005*). A gene map of the annotated *Populus* cp genome was drawn by OGdraw online (*Lohse, et al., 2013*).

Indices of codon usage

The amino acid composition and relative synonymous codon usage (RSCU) value of the five *Populus* cp genomes were calculated using the CodonW program, and the latter metric is an important indicator of codon usage bias (*Sharp, et al., 1986*). Because short CDSs generally resulted in large estimation errors for codon usage, CDSs shorter than 300 bp in length were excluded in codon usage calculations to avoid sampling bias (*Rosenberg, et al., 2003*). Finally, 58 CDSs for the five cp genomes were analyzed in this study.

Genome comparison

To investigate divergence in cp genomes, the identity across the whole cp genomes was visualized using the mVISTA viewer in the Shuffle-LAGAN mode among the five species, with the *P. xiangchengensis* genome as the reference. MAFFT version 7 software (*Katoh, et al., 2005*) was used to align the five plastome sequences. After manual adjustment with BioEdit software, we then performed sliding window analysis to assess the pairwise variability (Pi) over the plastomes in DnaSP version 5 software (*Librado & Rozas, 2009*). The window length was set to 600 bp, and the step size was set to 200 bp. The SNP variation was detected using the “find variation” function in Geneious R8.

Identification of simple sequence repeats (SSRs) and long sequence repeats

SSRs in five *Populus* cp genomes were detected using MISA (*Thiel et al., 2003*) with the minimal repeat number set to 12, 6, 5, 5, 5 and 5 for mono-, di-, tri-, tetra-, penta-, and hexa nucleotide sequences, respectively. We used the online REPuter software to identify and locate forward (F), reverse (R), complemented (C) and palindromic (P) repeats. The following settings for repeat identification were used: (1) Hamming distance equal to 3; (2) minimal repeat size was set to 30 bp; (3) maximum computed repeats was set to 90 bp (*Kurtz et al., 2001*).

Gene selective pressure analysis of five *Populus* plastomes

To examine variation in the evolutionary rates of cp genes, we calculated the nonsynonymous substitution rates (Ka), synonymous substitution rates (Ks), and their ratio (Ka/Ks) using model averaging in the Ka_Ks Calculator program according to the LWL85 method (*Yang & Bielawski, 2000*; *Zhang et al., 2006*).

Phylogenetic analysis

To explore the genetic relationships of the five species among the *Populus* genus, a total of 17 complete cp genomes of *Populus* and five plastomes of *Salix* were obtained from GenBank, and *Itoa orientalis* and *Idesia polycarpa* were used as the outgroups (Table S7). The cp genomes were aligned using MAFFT under default settings. A maximum likelihood method for phylogenetic analysis was performed based on the GTR + I + G model in RAxML version 8 (*Stamatakis, 2014*).

Results

Features of the five *Populus* plastomes

The complete cp genomes of the five *Populus* species were a double-stranded molecule ranging from 156,465 bp (*P. xiangchengensis*) to 156,789 bp (*P. cathayana*) in length. The plastome size of *P. schneideri* was only one bp larger than that of *P. pseudoglauca*. The plastome size of *P. kangdingensis* was 11 bp larger than that of *P. pseudoglauca* and 166 bp smaller than that of *P. cathayana*. The five cp genomes included a pair of inverted repeats (IRs) of 27,620 bp in the three species *P. kangdingensis*, *P. pseudoglauca* and *P. schneideri* and a pair of 27,672 bp in *P. cathayana*, 27,570 bp in *P. xiangchengensis.* The GC contents were consistent in *P. kangdingensis*, *P. pseudoglauca* and *P. schneideri,* with 34.5%, 30.5% and 42.0% in the large single copy (LSC), short single copy (SSC) and IR regions, respectively (Tables 1 and 2). The high GC content in the IR regions was possibly due to the presence of four ribosomal RNA sequences in these regions.

Each of the *P. cathayana, P. kangdingensis*, *P. pseudoglauca*, *P. schneideri* and *P. xiangchengensis* cp genomes encoded 130 functional genes; 112 of these were unique genes, including 78 protein-coding genes, 30 tRNA genes and 4 rRNA genes. Most of these genes occurred as a single copy, while 18 genes were double copies: seven protein-coding genes, seven tRNA genes and four rRNA genes. The LSC region contained 59 protein-coding genes and 22 tRNA genes, whereas the SSC region contained 10 protein-coding genes and one tRNA gene. Among the functional genes, 12 genes had one intron, and three genes had two introns (Table S1). The *trnK-UUU* gene had the largest intron, where another gene, *matK*, was nested within it. For the *rps12* gene, the 5' end was located in the LSC region, and the 3' end was located in the IR regions (Fig. 1).

Codon usage

Most protein-coding genes had the standard AUG sequence as the start codon, but *ndhD* started with GUG, and *rpl16* started with ATC. GUG start codons have been reported in tobacco, but they are very rare in eukaryotic genomes (*Kuroda et al., 2007*). When GUG was the start codon of a protein, it was still translated as Met because of the separate tRNA used for initiation. Furthermore, the codon usage patterns of the 58 distinct protein-coding genes in the five plastomes were examined, and the plastomes of *P. kangdingensis*, *P. pseudoglauca*, and *P. schneideri* were consistent, with a length of 75,990 bp and encoding 25,330 codons, while that of *P. cathayana* and *P. xiangchengensis* were 75,864 bp and 75,840 bp in size and encoded 25,288 and 25280 codons, respectively, which are presented in Table S2.

As an important indicator of codon usage bias, the RSCU value is the frequency observed for a codon divided by its expected frequency (*Sharp & Li, 1987*). Coding ending with A and

T/U had RSCU values > 1 for the five *Populus* cp genomes, indicating that they were used more frequently than synonymous codons and may play major roles in the A+T bias of entire cp genomes. There was a general excess of A- and U-ending codons. All three stop codons were present, with UAA being the most frequently used among the five plastomes (Table S2). Interestingly, leucine (Leu, 10.67%, 10.65%, 10.65%, 10.65%, 10.65% and 10.65%) and cysteine (Cys, 1.14%) were the most and least commonly coded amino acids, respectively, among the five plastomes (Table S2 and Fig. 2).

Comparative analysis of the five *Populus* plastomes

IRs is the most conserved regions of the cp genome. However, the construction and expansion of IR borders are common evolutionary events and the major reason for size differences between cp genomes (*Shen et al., 2017*). In this study, the cp genomes of the five *Populus* species were well conserved, and no rearrangement occurred in gene organization when *P. xiangchengensis* was used as a reference (Fig. 3 and Fig. 4).

The variations in the length of angiosperm cp genomes mainly result from the contraction and expansion of boundary regions between the IR regions with single copy (SC) regions (*Wu et at., 2018*). LSC, SSC and IR sections of the three *Populus* species of *P. kangdingensis*, *P. pseudoglauca*, and *P. schneideri* were highly conserved and smaller than those of *P. cathayana*, while the IR regions were larger than *P. xiangchengensis*. Detailed comparisons of the IR-SSC and IR-LSC boundaries among the cp genomes of the five species are presented in Fig. 5. Two complete or fragmented copies of *rpl22* and *ycf1* were located at the boundaries between the LSC or SSC regions and IR regions among the five *Populus* plastomes. The *rpl22* gene crossed the IR-LSC with only one bp variation in sequence length among the five plastomes. The gene *ycf1*, in the IRb region, extended from 15 bp (*P. cathayana*) to 158 bp (*P. kangdingensis*, *P. pseudoglauca*, and *P. schneideri*), whereas the gene *ycf1* in the IRa region extended from 1689 bp (*P. xiangchengensis*) to 1707 bp (*P. cathayana*). A 61 bp overlap between *ycf1* and *ndhF* was found in *P. kangdingensis*, *P. pseudoglauca*, and *P. schneideri*.

To elucidate the level of sequence variation, the Pi values in five cp genomes were calculated with DnaSP 5.0 software. The Pi values within 600 bp in the five plastomes varied from 0.00001 to 0.00335, with a mean of 0.00210 (Table 3). The results indicate high sequence similarity across the five plastomes, suggesting that the cp genomes of these five *Populus* species are highly conserved. Using sliding window analysis, we identified the nine most divergent regions, *trnG-atpA*, *psbZ-trnfM*, *trnL-ndhJ*, *ndhC-trnV*, *ycf4-cemA*, *trnN-trnR*, *ycf1-ndhF*, *ccsA-ndhD* and *trnR-trnN* (Fig. 6), which had a Pi>0.01, indicating that these variations are mainly present in the intergenic space. The nine divergent regions could be utilized as potential molecular markers for population genetic and phylogenetic studies in *Populus*.

Number and forms of mutations

We investigated SNPs, as the most abundant type of mutation, in the five plastomes, with *P. xiangchengensis* as the reference. In gene-coding regions, we detected 70 SNPs in the plastome of *P. cathayana*, including 33 Ts and 37 Tv SNPs, and 160 (97 Ts and 63 Tv), 166 (101 Ts and 65 Tv) and 164 (99 Ts and 65 Tv) SNPs in the plastomes of *P. kangdingensis*, *P. pseudoglauca* and *P. schneideri* (Table 4). Furthermore, 106 (38 Ts and 68 Tv), 323 (130 Ts and 193 Tv), 316 (130 Ts and 186 Tv) and 314 (131 Ts and 183 Tv) SNPs were detected in noncoding regions among the plastomes of *P. cathayana*, *P. kangdingensis*, *P. pseudoglauca* and *P. schneideri*, respectively (Table S3).

It has been reported that small inversions is commonly associated with a hairpin secondary structure in the cp genomes (*Kim & Lee, 2005*; *Catalano et al., 2009*). In this study, a total of six small inversions (*petA-psbJ*, *ndhC-trnV*, *trnN-trnR*, *ccsA-ndhD*, *ndhD-psaC* and *ndhF-trnL*) were uncovered based on the sequence alignment of the five complete chloroplast genomes (Fig. 7A-F). The small inversions from *ndhC-trnV* and *ndhD-psaC* occurred in only *P. xiangchengensis*, those from *ndhF-trnL* occurred in *P. pseudoglauca* and *P. schneideri*, those from *trnN-trnR* occurred in *P. kangdingensis*, *P. pseudoglauca* and *P. schneideri*, and those from *ccsA-ndhD* occurred in the four species other than *P. cathayana*, while the inversion from *petA-psbJ* occurred in the four species other than *P. xiangchengensis*.

Synonymous (Ks) and nonsynonymous (Ka) substitution rate analysis

The synonymous (Ks) and nonsynonymous (Ka) nucleotide substitution patterns are very important markers in gene evolution studies (*Kimura, 1979*). The nonsynonymous to synonymous ratio (Ka/Ks) is indicative of changes in selective pressures. Ka/Ks values >1, =1, and <1 indicate positive selection, natural evolution and purifying selection affecting the coding portions, respectively (*Sharp & Li, 1987*; *Yang & Bielawski, 2000*; *Lawrie et al., 2013*). However, the ratio of Ka/Ks was less than one in most protein-coding regions (*Makalowski & Boguski, 1998*). In this study, when compared the plastomes of four *Populus* with that of *P. xiangchengensis*, only 19 protein-coding genes had the values from 85 comparison numerations (Fig. 8 and Table S4). The Ka/Ks value of the remaining protein-coding genes could not be calculated because Ka or Ks was equal to 0, indicating that these sequences were conserved without nonsynonymous or synonymous nucleotide substitution. The Ka/Ks ratio of all genes except *rpoC2* in *P. pseudoglauca* (1.00903) and the *rbcL* gene in *P. kangdingensis* (2.26407), *P. pseudoglauca* (2.26407) and *P. schneideri* (2.26407) was less than 1 (Fig. 8), indicating that the two genes suffered from positive selection and that at least some of the mutations concerned must be advantageous.

SSR and long repeat analysis

Cp simple sequence repeats (cpSSRs) are effective molecular markers. They have not only the advantages of abundance, codominant inheritance and high repeatability but also the characteristics of simple genomic structure, relatively conserved sequences and maternal inheritance, which makes them widely used in species identification, phylogenetic analysis, breeding analysis, population genetics and ecological studies at the individual and population levels (*Cavaliersmith, 2002*; *Kaundun & Matsumoto, 2002*; *Jiao et al., 2012*).

With MISA, a total of 170 SSR loci were detected, of which mononucleotide repeats occurred with high frequency constituted 148 (87.06%) of all the SSRs and all of the mononucleotides composed of poly A (polycytosine) and poly T (polythymine) repeats (Table 5). Within the five plastomes, SSR loci were primarily located in the LSC region, followed by the SSC region. A total of 15 SSR loci were detected in the protein-coding genes *rpoB*, *rpoC2* and *rps8*, with all others situated in gene spacers and introns (Table S5). A total of 28, 39, 39, 39 and 25 SSR loci were detected in *P. cathayana*, *P. kangdingensis*, *P. pseudoglauca*, *P. schneideri* and *P. xiangchengensis* cp genomes, respectively (Table 5). Among these, there were 26 and 23 mononucleotide repeats in the *P. cathayana* and *P. xiangchengensis* cp genomes, respectively, while they both had one dinucleotide repeat and one compound nucleotide repeat. The corresponding numbers of these repeats in *P. kangdingensis*, *P. pseudoglauca* and *P. schneideri* matched each other and were 33 mononucleotide, two dinucleotide and four compound repeats. Comparison among the five plastomes revealed that two loci were in only *P. xiangchengensis*, three loci were in only *P. cathayana*, and 36 loci were detected in the plastomes of *P. kangdingensis*, *P. pseudoglauca* and *P. schneideri* (Table S5).

In addition to SSRs, dispersed repeats are thought to play an important role in genome recombination and rearrangement through illegitimate recombination and slipped-strand mispairing (*Saski et al., 2007*; *Huang et al., 2014*). In the plastomes of the five *Populus* species, we found 58 repeats (27 forward repeats, 22 palindromic repeats, seven reverse repeats and two complement repeats) in *P. cathayana*; these numbers of repeats were higher than those found in the other four species (Fig. 9A). *P. pseudoglauca* and *P. schneideri* shared the same number and types. The majority of repeats (84.86%) varied from 30 to 39 bp in length (Fig. 9B). Variation in the number of repeat sequences has been observed between species belonging to different regions (Table S6). The dispersed repeats identified in the five *Populus* species provide a basis for the development of markers for phylogenetic and population genetic studies.

Phylogenetic analysis based on the cp genome

To futher elucidate the genetic relationship between the five *Populus* species, an improved resolution of phylogenetic relationships was achieved by using these complete plastomes of 17 public *Populus* species and five public *Salix* species, and plastomes of *Idesia polycarpa* and *Itoa orientalis* were used as the outgroups (Table S7). All of the *Populus* were divided into four main highly supported clades (Fig. 10). Three species of section *Turanga* were clade I members. Clade II included seven species (*P. adenopoda*, *P. alba*, *P. davidiana*, *P. qiongdaoensis*, *P. rotundifolia*, *P. tremula*, and *P. tremula* × *alba*) in section *Populus* and one species in section *Aigeiros* (*P. nigra*). Clade III consisted of three species in section *Tacamahaca* (*P. kangdingensis*, *P. schneideri* and *P. yunnanensis*) and two species in section *Leucoides* (*P. lasiocarpa* and *P. pseudoglauca*). Clade IV included the four species in section *Tacamahaca* (*P. balsamifera*, *P. cathayana*, *P. trichocarpa* and *P. xiangchengensis*), one species in section *Aigeiros* (*P. fremontii*) and one species in section *Leucoides* (*P. wilsonii*). Our results showed that *P. kangdingensis*, *P.*

pseudoglauca, and *P. schneideri* were in clade III, while *P. xiangchengensis* formed a sister relationship with a 100% bootstrap support to *P. cathayana* in clade IV.

Discussion

In the present study, we compared the five *Populus* plastomes, all of these assembled into single circule, double-stranded DNA sequences, presenting a typical quadripartite structure, with a length of 156,465 bp (*Zong et al., in press*) to 156,789 bp, which was similar to most *Populus* cp genomes (*Wang, 2016*; *Zhang & Gao, 2016*; *Zheng, 2016*; *Han et al., 2017*). LSC, SSC and IR sections of the three *Populus* species of *P. kangdingensis*, *P. pseudoglauca*, and *P. schneideri* were highly conserved and smaller than those of *P. cathayana*, while the IR regions were larger than *P. xiangchengensis*. The variations in the length of angiosperm cp genomes mainly result from the contraction and expansion of boundary regions between the IR regions with single copy (SC) regions (*Wu et at., 2018*). To elucidate the level of sequence variation, the Pi values in five cp genomes were calculated with DnaSP 5.0 software. We identified nine most divergent regions, *trnG-atpA*, *psbZ-trnfM*, *trnL-ndhJ*, *ndhC-trnV*, *ycf4-cemA*, *trnN-trnR*, *ycf1-ndhF*, *ccsA-ndhD* and *trnR-trnN* (Fig. 6), which could be utilized as potential molecular markers for population genetic and phylogenetic studies in *Populus*.

Understanding nucleotide substitution rates is of fundamental importance in molecular evolution (*Muse & Gaut, 1994*). During the process of searching for SNPs, we found that the cp genome sequences of *P. kangdingensis*, *P. schneideri* and *P. pseudoglauca* had similar mutation number, while *P. cathayana* had smaller mutation when compared with *P. xianchengensis*. Futhermore, we also found that the number and type of SSRs and long repeats of the three species *P. kangdingensis*, *P. pseudoglauca*, and *P. schneideri* were basically identical. Therefore, the phylogenetic relationships of these five species may be affected by the different mutation modes.

Small inversions in the cp genome of angiosperms are ubiquitous and commonly associated with a hairpin secondary structure in the cp genomes (*Kim & Lee, 2005*; *Catalano et al., 2009*). A distinctive feature of these inversions is that they are flanked by IRs such that the IRs form the stem and that the segment between them forms the loop (*Catalano et al., 2009*). These small inversions are generally recognized by pairwise comparisons between sequences. In this study, a total of six small inversions were uncovered based on the sequence alignment of the five complete cp genomes. Among these small inversions, *ndhC-trnV* and *ndhD-psaC* occurred in only *P. xiangchengensis*, *ndhF-trnL* occurred in *P. pseudoglauca* and *P. schneideri*, *trnN-trnR* occurred in *P. kangdingensis*, *P. pseudoglauca* and *P. schneideri*, and *ccsA-ndhD* occurred in the four species other than *P. cathayana*, while the inversion from *petA-psbJ* occurred in the four species other than *P. xiangchengensis*. Small inversions in the *ccsA-ndhD* and *petA-psbJ* intergenic regions have been reported in other studies (*Song et al., 2015*; *2016*; *Dong et al., 2017*). However, small inversions of noncoding sequences may influence sequence alignment and character interpretation in phylogeny reconstructions, so caution is necessary when using cp noncoding sequences for phylogenetic analysis.

The cp genome is widely employed in the study of evolution through phylogenetics, and it has been suggested to be useful for phylogenetic reconstruction at low taxonomic levels (*Zhang et al., 2011*; *Ma et al., 2014*; *Yang et al., 2014*; *Zhang et al., 2016*). It has also been postulated to be a potential ultrabarcode or organelle-scale barcode for taxonomically complex groups (*Kane et al., 2012*). The key interest in the current study is to resolve the previous phylogenetic controversies in the *Populus* (*Zhao, 1994*; *Liu & Fu, 2004*; Chen et al., 2007; *Wan et al., 2009*; *Wang, 2012*) by using the complete cp genome sequences. All of the *Populus* were divided into four main highly supported clades (Fig. 10). Three species of section *Turanga* were clade I members. Clade II included seven species in section *Populus* and one species in section *Aigeiros* (*P. nigra*), which is supported by previous studies (*Rajora & Dancik, 1995*; *Hamzeh & Dayanandan, 2004*). Both studies found that *P. nigra* showed higher similarities to *P. alba* than to other species. Clade III consisted of three species in section *Tacamahaca* (*P. kangdingensis*, *P. schneideri* and *P. yunnanensis*) and two species in section *Leucoides* (*P. lasiocarpa* and *P. pseudoglauca*). Clade IV included the four species in section *Tacamahaca* (*P. balsamifera*, *P. cathayana*, *P. trichocarpa* and *P. xiangchengensis*), one species in section *Aigeiros* (*P. fremontii*) and one species in section *Leucoides* (*P. wilsonii*). Our results showed that *P. kangdingensis*, *P. pseudoglauca*, and *P. schneideri* were in clade III, while *P. xiangchengensis* formed a sister relationship with a 100% bootstrap support to *P. cathayana* in clade IV.

The position of *P. pseudoglauca* confirms the previously published phylogeny described by Chao & Liu (*1991*), in which *P. pseudoglauca* was classified into section *Tacamahaca* according to fossil evidence, paleogeography, paleoclimate, and modern distribution. The species *P. schneideri*, which is distributed in the western Sichuan Plateau at altitudes of 3000 to 4000 m, has remained a topic of debate among scientists. According to the morphology, it is similar to *P. cathayana* (*Fang et al., 1999*). Wan *et al*. (*2013*) suggested that *P. schneideri* is generally closer to *P. cathayana* than *P. kangdingensis*, and it is a natural hybrid between the ancestors of *P. cathayana* and *P. kangdingensis* based on cpDNA and nuclear DNA sequence data as well as amplified restriction fragment polymorphism (AFLP) analyses. Other studies considered *P. schneideri* to be a variety of *P. kangdingensis* based on morphological traits (*Chao & Liu, 1991*; *Yu et al., 2003*; *Liu & Fu, 2004*). Chen *et al*. (*2007*) suggest that *P. schneideri* is generally more highly related to *P. kangdingensis* than to *P. cathayana* based on the cpSSR analysis. Our data also reveal that *P. schneideri* had a close relationship with *P. kangdingensis*. *P. schneideri* and *P. kangdingensis* are both unique to the western Sichuan Plateau, and they share similar altitude and habitat requirements (*Yu et al., 2003*). However, *P. xiangchengensis* was a sister to *P. cathayana*, as revealed by cp genome sequence analysis, which did not support the viewpoint that it was a natural hybrid species of either *P. schneideri* and *P. pseudoglauca* or *P. kangdingensis* and *P. pseudoglauca*. It is our hope that the five plastomes will provide useful resources for better understanding the phylogeny and the relationships of the genus *Populus*.

Conclusions

This study reports the comparative analysis of five *Populus* cp genome sequences with detailed gene annotation. Comparing the five plastomes showed that the plastomes are similar in structure and have a high degree of synteny. Nine divergent regions (*trnG-atpA*, *psbZ-trnfM*, *trnL-ndhJ*, *ndhC-trnV*, *ycf4-cemA*, *trnR-trnN*, *ycf1-ndhF*, *ccsA-ndhD* and *trnR-trnN*) were identified, which could be utilized as potential molecular markers for population genetic and phylogenetic studies in *Populus*. Furthermore, among the five cp genomes, *P. kangdingensis*, *P. pseudoglauca* and *P. schneideri* had little difference in their SNP loci and SSRs. The results of phylogenetic analyses showed that *P. schneideri* had the closest affinity to *P. kangdingensis* and was sister to *P. pseudoglauca*, while *P. cathayana* had a close relationship with *P. xiangchengensis*. The characterization of these five plastomes will provide useful resources for better understanding the phylogeny and the relationships of the genus *Populus*.

Acknowledgements

We would like to thank Prof. Wuyuan Yin for his help with sample collection and species identification. We are grateful to Dr. Hui Zhan for language review.

References

Asheesh S, Vinay S. 2012. Evolutionary analysis of plants using chloroplast. German: LAP Lambert Academic Publishing.

Braatne JH, Hinckly TM, Stettler, RF. 1992. Nuclear ribosomal DNA phylogeny water supply on the physiological and morpgological components of plant water balance in *Populus trichocarpa*, *Populus deltoids* and their F1 hybrids. *Tree physiology* **11:** 325-340.

Catalano SA, Saidman BO, Vilardi JC. 2009. Evolution of small inversions in chloroplast genome: a case study from a recurrent inversion in angiosperms. *Cladistics* **25(1):** 93-104.

Cavaliersmith T. 2002. Chloroplast evolution: secondary symbiogenesis and multiple losses. *Curr. Biol.* **12(2):** 62-64.

Chao N, Liu J. 1991. Taxonnmic studies on *Populus* L. in southwestern China (I). *J. Wuhan Bot. Res*. **9(3):** 229-238.

Chen K, Peng YH, Wang YH, Korpelainen H, Li CY. 2007. Genetic relationships among poplar species in section *Tacamahaca* (*Populus* L.) from western Sichuan, China. *Plant Science* **172(2):** 196-203.

Cronk Q. 2005. Plant eco-devo: the potential of poplar as a model organism. *New Phytologist* **166(1):** 39-48.

Dong WP, Xu C, Li WQ, Xie XM, Lu YZ, Liu YL, Jin XB, Suo ZL. 2017. Phylogenetic resolution in Juglans based on complete chloroplast genomes and nuclear DNA sequences. *Frontiers in Plant Science* **8.**

Eckenwalder JE. 1996. Systematics and evolution of *Populus*. In Stettler, R.F., Bradshaw, H.D., Heilman, P. E., Hinckley, T.M. Biolology of *Populus* and its implications for management and conservation. Canada: NRC research press, 7-32.

Fang ZF, Zhao SD, Skvortsov AK. 1999. Flora of China (English version). Beijing: Science press, **4:** 139-274.

Hamzeh M, Dayanandan S. 2004. Phylogeny of *Populus* (Salicaceae) based on nucleotide sequences of chloroplast *trnT-trnF* region and nuclear rDNA. *American Journal of Botany* **91(9):** 1398-1408.

Han XM, Wang YM, Liu YJ. 2017. The complete chloroplast genome sequence of *Populus wilsonii* and its phylogenetic analysis. *Mitochondrial DNA Part B*: resources **2(2):** 932-933.

He CZ, Li JM, Yun T, Zong D, Zhou AP, Ou GL, Yin WY. 2015. SRAP analysis on the effect of geographic isolation on population genetic structure of *Populus davidiana* in Tibetan-inhabited regions in Southwest China. *Forest Research* **28(2):** 152-157.

Howe CJ, Barbrook AC, Koumandou VL, Nisbet RER, Symington HA, Wightman TF. 2003. Evolution of the chloroplast genome. *Phil. Trans. R Soc. L and B* **358(1429):** 99-107.

Huang H, Shi C, Liu Y, Mao SY, Gao LZ. 2014. Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evol. Biol*. **14(1):151.**

Jin JJ, Yu WB, Yang JB, Song Y, Yi TS, Li DZ. 2018. GetOrganelle: a simple and fast pipeline for de novo assemble of a complete circular chloroplast genome using genome skimming data. *BioRxiv* Preprint.

Jiao Y, Jia HM, Li XW, Chai ML, Jia HJ, Chen Z, Wang GY, Chai CY. Van de WE, Gao ZS. 2012. Development of simple sequence repeat (SSR) markers from a genome survey of Chinese bayberry (*Myrica rubra*). *BMC Genomics* **13(1):** 201.

Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JM, Cronk Q. 2012. Ultra-barcoding in cacao (*Theobram* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *Am. J. Bot*. **99(2):** 320-329.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33(2):** 511-518.

Kaundun SS, Matsumoto S. 2002. Heterologous nuclear and chloroplast microsatellite amplification and variation in tea *Camellia sinensis*. *Genome* **45(6):** 1041-1048.

Khan A, Khan I, Heinze B, Azim MK. 2012. The chloroplast genome sequence of date palm (*Pheonix dactylifera* L. cv. 'Aseel'). *Plant Mol. Biol. Rep.* **30(3):** 666-678.

Kim KJ, Lee HL. 2005. Wide spread occurrence of small inversions in the chloroplast genomes of land plants. *Mol. Cells* **19(1):** 104-113.

Kimura M. 1979. The neutral theory of molecular evolution. *Sci. Am*. **241:** 98.

Kuroda H, Suzuki H, Kusumegi T, Hirose T, Yukawa Y, Sugiura M. 2007. Translation of *psbC* mRNAs starts from the downstream GUG, not the upstream AUG, and requires the extended shine-dalgarno sequence in tobacco chloroplasts. *Plant Cell Physiol*. **48(9):** 1374-1378.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoyem J, Giegerich R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **29(22):** 4633-4642.

Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genet.* **9(5):** 261-270.

Librado P, Rozas J. 2009. Dnasp v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25(11):** 1451-1452.

Liu LX, Li R, Worth JP, Li X, Li P, Cameron KM, Fu CX. 2017. The complete chloroplast genome of Chinese Bayberry (*Morella rubra*, Myricaceae): Implications for understangding the evolution of Fagales. *Frontiers in Plant Science* **8:** 968.

Liu YQ, Fu DQ. 2004. Development and utilization of sect. Ⅲ *Tacamahaca* gene resources on the plateau of western Sichuan. *Journal of Central South Forestry University* **24(5):** 129-132.

Lohse M, Drechsel O, Kahlau S, Bock R. 2013. OrganellarGenomeDRAW-a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res* **41:** 575-581.

Lu ZQ, Tian B, Liu BB, Yang C, Liu JQ. 2014. Origin of *Ostryopsis intermedia* (Betulaceae) in the southeast Qinghai-Tibet Plateau through hybrid speciation. *Journal of Systematics and Evolution* **52(3):** 250-259.

Ma PF, Zhang YX, Zeng CX. Guo ZH, Li DZ. 2014. Chloroplast phylogenomic analyses resolve deep-level relationships of an intractable bamboo tribe Arundinarieae (Poaceae). *Syst. Biol.* **63(6):** 933-950.

Makalowski W, Boguski MS. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Sci. U.S.A.* **95(16):** 9407-9412.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11(5):** 715-724.

Rajora OP, Dancik BP. 1995. Chloroplast DNA variation in *Populus* II interspecific restriction fragment polymorphisms and genetic relationships among *Populus* deltoids, *P. nigra*, *P. maximowiczii*, and *P.* ∅ *canadensis*. *Theor. Appl. Genet.* **90(3-4):** 324-330.

Rosenberg MS, Subramanian S, Kumar S. 2003. Patterns of transitional mutation biases within and among mammalian genomes. *Mol. Biol. Evol.* **20(6):** 988-993

Saski C, Lee SB, Fjellheim S, Guda C, Jansen RK, Luo H, Tomkins J, Rognli OA, Daniell H, Clarke JL. 2007. Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. *Theoretical and Applied Genetics* **115(4):** 571-590.

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:** 686-689.

Sharp PM, Li WH. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.* **4(3):** 222-230.

Sharp PM, Tuohy TMF, Mosurski KR. 1986. Codon usage in yeast cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* **14(13):** 1281-1295.

Shen XR, Wu ML, Liao BS, Liu ZX, Bai R, Xiao SM, Li XW, Zhang BL, Xu J, Chen SL. 2017. Complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant *Artemisia annua*. *Molecules* **22(8):** 1330.

Song Y, Dong W, Liu B, Xu C, Yao X, Gao J, Corlet RT. 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.

Song Y, Yao X, Tan YH, Gan Y, Corlet RT. 2016. Complete chloroplast genome sequence of the avocado: gene organization, comparative analysis, and phylogenetic relationships with other *Lauraceae*. *Can. J. For. Res.* **46 (11):** 1293-1301.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30(9):** 1312-1313.

Stettler RF, Bradshaw HD, Heilman PE, Hinckley TM. 1996. The role of hybridization in genetic manipulation of *Populu*s. In: Biology of *Populus* and its implications for management and conservation. Canada : NRC Research Press.

Thiel T, Michalek W, Varshney RK. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics* **106(3):** 411-422.

Wan XQ, Zhang F, Zhong Y, Ding YH, Wang LW, Hu TX. 2013. Study of genetic relationships and phylogeny of the native *Populus* in southwest China based on nucleotide sequences of chloroplast *trnT-trnF* and nuclear DNA. *Plant Syst*. Evol. **299(1):** 57-65.

Wan XQ, Zhang F, Zhong Y, Wang CL, Ding YH, Hu TX, Zhai MP, Qian ZL. 2009. Conservation and application of the genetic resources of native poplars in southwest China. *Scientia Slivae Sinicae* **45(4):** 139-144.

Wang MF. 2012. Analysis the genetic relationship of the native *Populus* in Sichuan by ISSR and ITS sequences. *Sichuan Agricultrual University*.

Wang TJ, Fan LQ, Guo XL, Wang K. 2016. Characterization of the complete chloroplast genome of *Populus qiongdaoensis* T. Hong et P. Luo. *Conservation Genet Resour* **8(4):** 1-3.

Wicke S, Schneeweiss GM, Müller KF, Quandt D. 2011. The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. *Plant Mol. Biol.* **76(3-5):** 273-297.

Wu ML, Qing L, Xu J, Li XW. 2018. Complete chloroplast genome of the medicinal plant *Amomum compactum*: gene organization, comparative analysis, and phylogenetic relationships within Zingiberales. *Chinese Medicine* **13(1):** 10.

Yang JB, Li DZ, Li HT. 2014. Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. *Molecular ecology resources* **14(5):** 1024-1031.

Yang ZH, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol.* **15(12):** 496-503.

Yu SQ, Liu J, Fu DR, Liu DJ, Liu YQ. 2003. Characteristics of *Tacamachaca* genes in the western Sichuan Plateau. *Journal of Zhejiang Forestry College* **20:** 27-31.

Zhang QJ, Gao LZ. 2016. The complete chloroplast genome sequence of desert poplar (*Populus euphratica*). *Mitochondrial DNA* **27(1):** 721.

Zhang YJ, Du LW, Liu A, Chen JJ, Wu L, Hu WM, Zhang W, Kim K, Lee SC, Yang TJ, Wang Y. 2016. The complete chloroplast genome sequences of five *Epimedium* species: lights into phylogenetic and taxonomic analysis. *Front. Plant Sci*. **7:** 306.

Zhang YJ, Ma PF, Li DZ. 2011. High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). *PloS ONE* **6(5):** e20596.

Zhang Z, Li J, Zhao XQ, Wang J, Wong KSG, Yu J. 2006. KaKs_calculator: Calculating Ka and Ks through model selection and model averaging. *Genomics, Proteomics & Bioinformatics*, **4(4):** 259-263.

Zhao N. 1994. Taxonomic study on Salicaceae in Sichuan and its adjacent regions (Third). S*ichuan Forestry Science and Technology* **15: 1-11.**

Zheng HL, Fan LQ, Wang TJ, Zhang L, Ma T, Mao KS. 2016. The complete chloroplast genome of *Populus rotundifolia* (Salicaceae). *Conservation Genet. Resour.* **8(4):** 1-3.

Zong D, Zhou, AP, Li, D, He, CZ. 2018. The complete chloroplast genome of *Populus xiangchengensis*, an endemic species in southwest China. *Mitochondrial DNA part B* In press.

Zsuffa L. 1975. A summary review of interspecific breeding in the genus *Populus*. In proceedings of the 14th annual meeting of the Canadian tree improvement association, part 2. Canadian Forest Service, Ottawa, Ontario, Canada 107-123.

Additional Information and Declarations

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 31460205, 31860219, 31360184) and the Forestry Public Benefit Research Program (No. 201104076).

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

Dan Zong conceived and performed the experiments, analyzed the data, wrote the paper, and prepared figures and tables.

An-Pei Zhou performed the experiments, prepared figures and/or tables, and reviewed drafts of the paper.

Yao Zhang, Xin-Lian Zou and Dan Li performed the experiments and reviewed drafts of the paper.

Anan Duan and Chengzhong He conceived and designed the experiments, contributed the materials, authored or reviewed drafts of the paper, and approved the final draft.

Data Archiving Statement

Our raw data (Complete chloroplast genome sequences for the four *Populus* species) will be submitted to Genebank of NCBI through the revision process. The accession numbers from Genebank will be supplied before the final acceptance of the manuscript.

Figure captions

Figure 1 Gene map of the five *Populus* species cp genomes

Figure 2 Amino acid frequencies of the protein-coding sequences of the five plastomes

Figure 3 Comparison of the cp genome sequences of five *Populus* plastomes

Figure 4 Mauve alignment of the chloroplast genomes of five *Populus* species

Figure 5 Comparison of LSC, SSC and IR region borders among chloroplast genomes of five *Populus* species

Figure 6 Sliding window analysis of the whole plastomes for five *Populus* species

Figure 7A-F Predicted hairpin loops of inversions in the five plastomes of *Populus.* The structures of hairpin loops in the regions of the (A) *ccsA-ndhD*, (B) *ndhC-trnV*, (C) *ndhD-psaC*, (D) *ndhF-trnL*, (E) *petA-psbJ* and (F) *trnN-trnR* were drawn with RNAstructure. The arrows in the figure indicated the break points in inversion events.

Figure 8 Ka/Ks values of 19 protein-coding genes of the four species

Figure 9 Comparison of long repeat among five *Populus* plastomes. (A) Number of each repeat type; (B) Frequency of each repeat type by length

Figure 10 Molecular phylogenetic tree of 27 species in the family *Salicaceae* based on the complete plastome sequence

Supporting Information

Supplementary file1

The raw data of the four *Populus* species *P. cathayana*, *P. kangdingensis*, *P. pseudoglauca* and *P. schneideri*.

Table S1 List of genes in the chloroplast genomes

Table S2 Codon usage in the five *Populus* plastomes

Table S3 Transitions (Ts) and transversions (Tv) in the four plastomes compared with the plastome of *P. xiangchengensis*

Table S4 Synonymous (Ks) and nonsynonymous (Ka) analysis of the five species, with *P. xiangchengensis* as the reference

Table S5 SSRs in the five chloroplast genomes

Table S6 Repeat sequences in the five chloroplast genomes

Table S7 GenBank accession numbers of the *Populus* and *Salix* and outgroups with chloroplast genome sequences used for phylogenetic analyses

Figure 1

Gene map of the five *Populus* species cp genomes

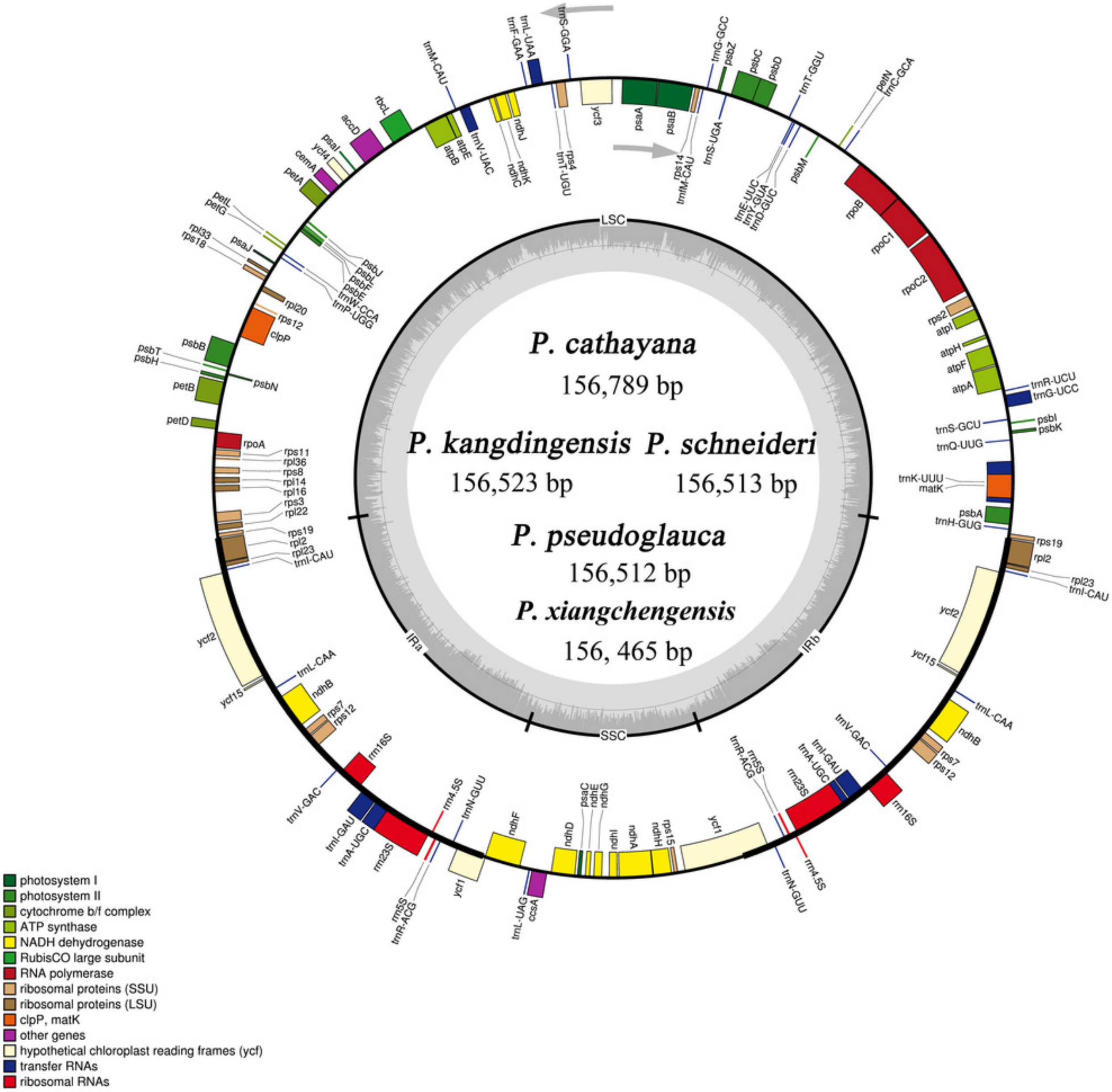

Figure 2

Amino acid frequencies of the protein-coding sequences of the five plastomes

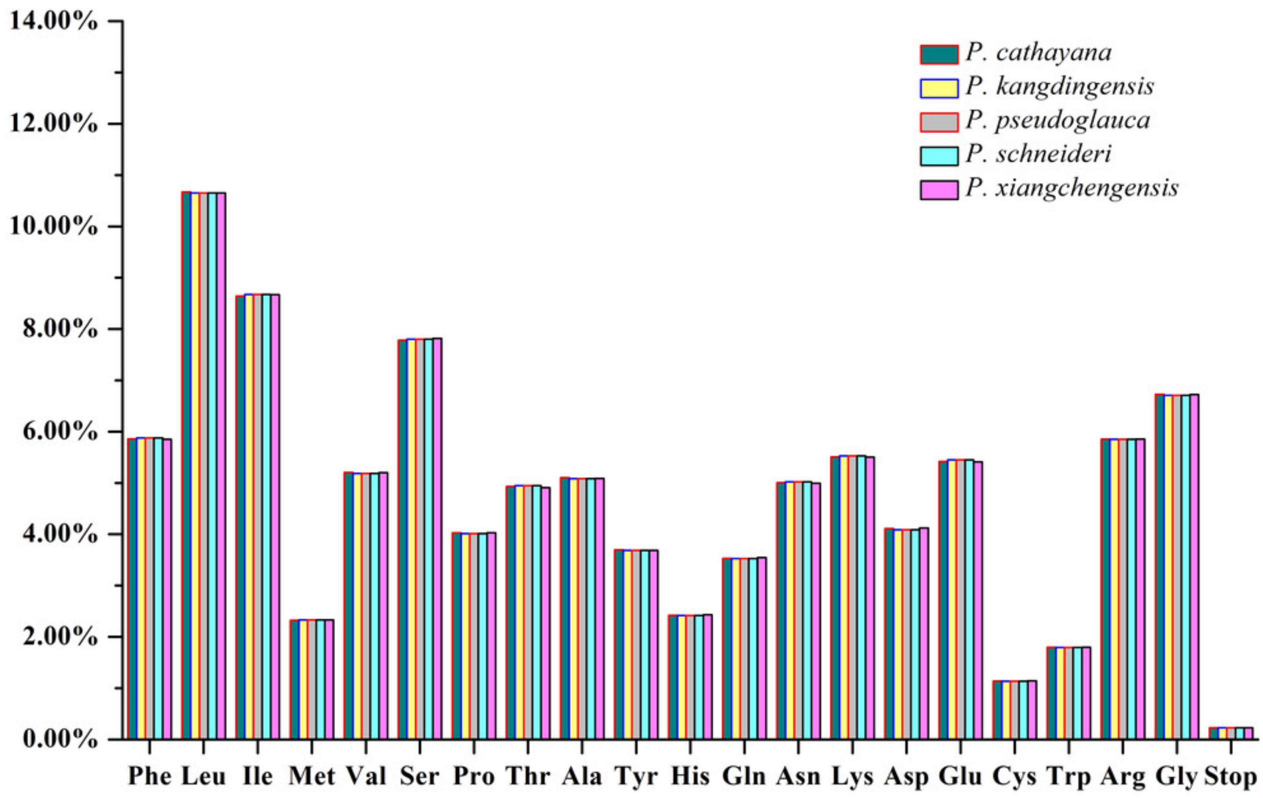

Figure 3

Comparison of the cp genome sequences of five *Populus* plastomes

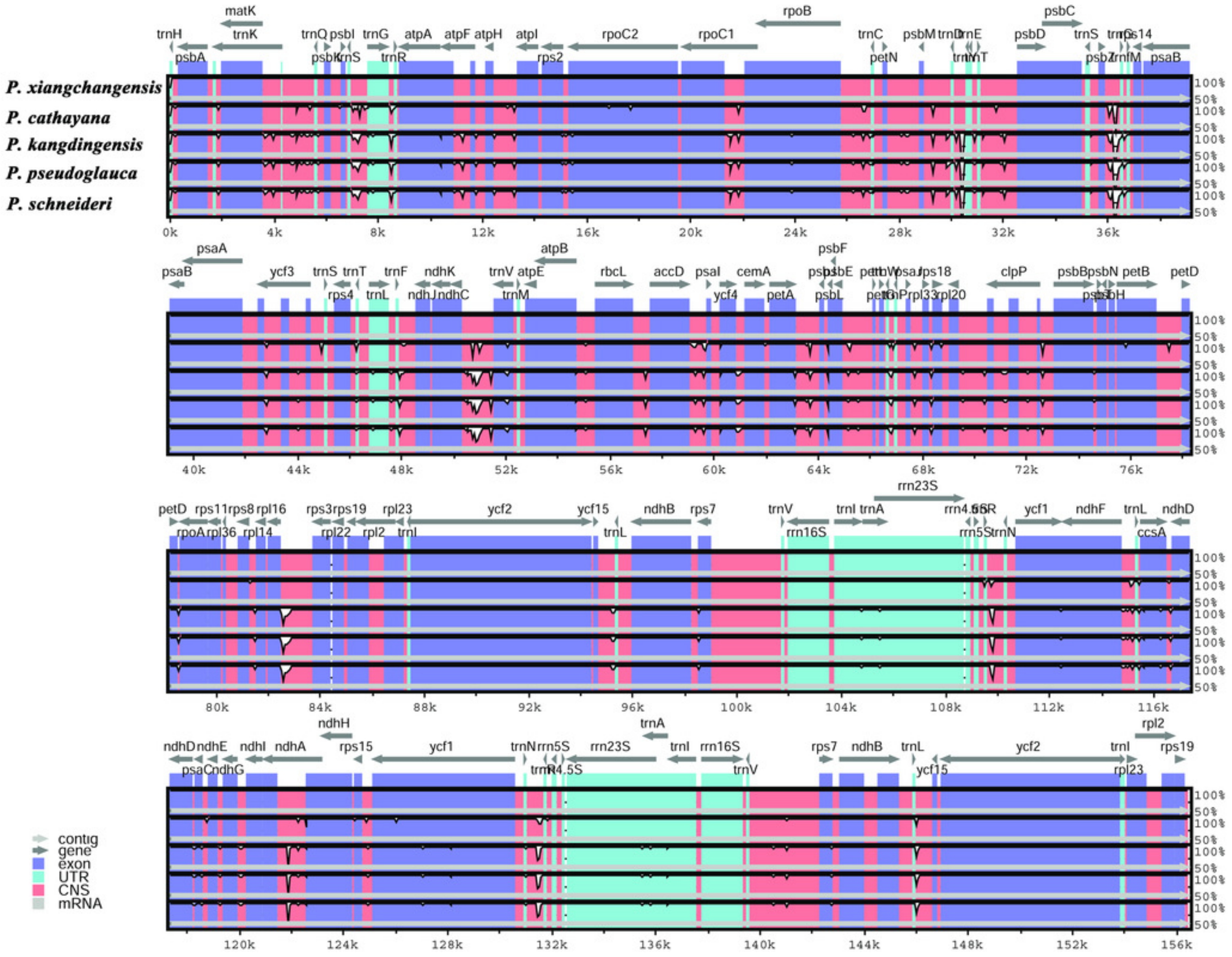

Figure 4

Mauve alignment of the chloroplast genomes of five *Populus* species

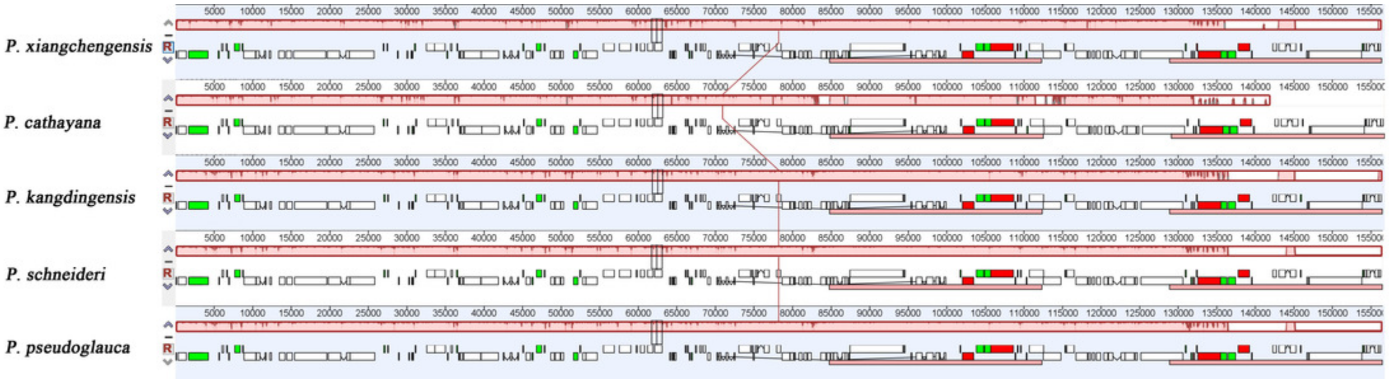

Figure 5

Comparison of LSC, SSC and IR region borders among chloroplast genomes of five *Populus* species

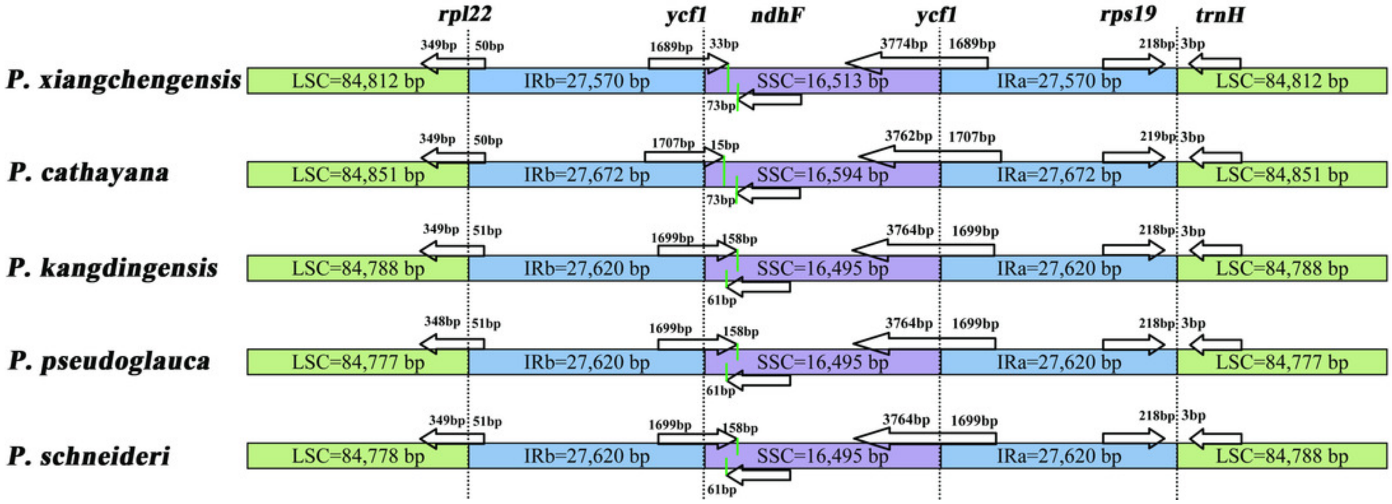

Figure 6

Sliding window analysis of the whole plastomes for five *Populus* species

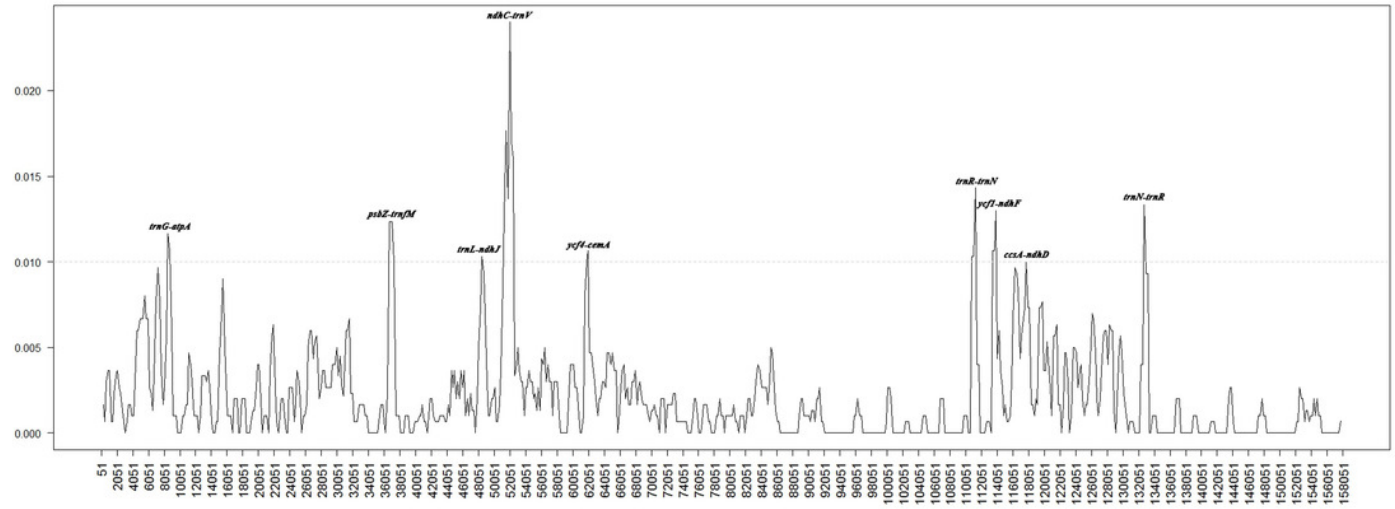

Figure 7

Predicted hairpin loops of inversions in the five plastomes of *Populus*

The structures of hairpin loops in the regions of the (A) *ccsA-ndhD*, (B) *ndhC-trnV*, (C) *ndhD-psaC*, (D) *ndhF-trnL*, (E) *petA-psbJ* and (F) *trnN-trnR* were drawn with RNAstructure. The arrows in the figure indicated the break points in inversion events.

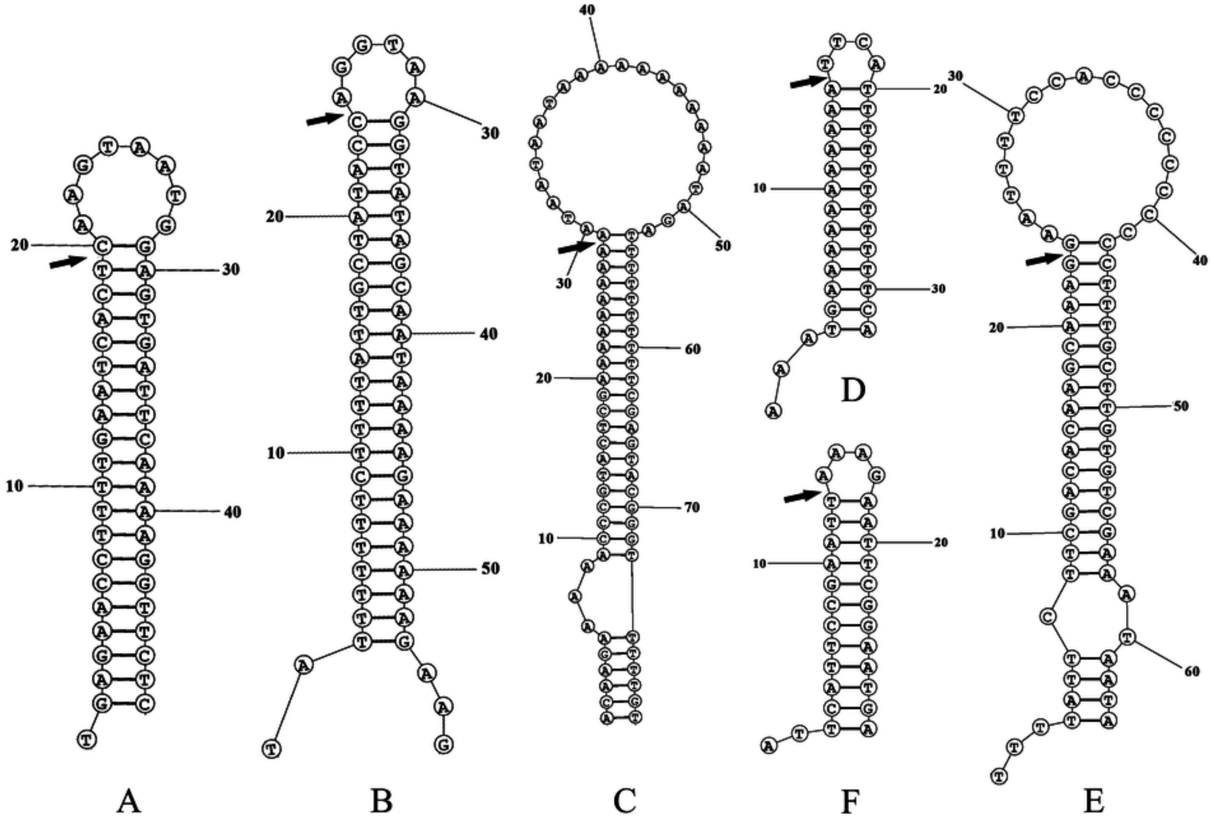

Figure 8

Ka/Ks values of 19 protein-coding genes of the four species

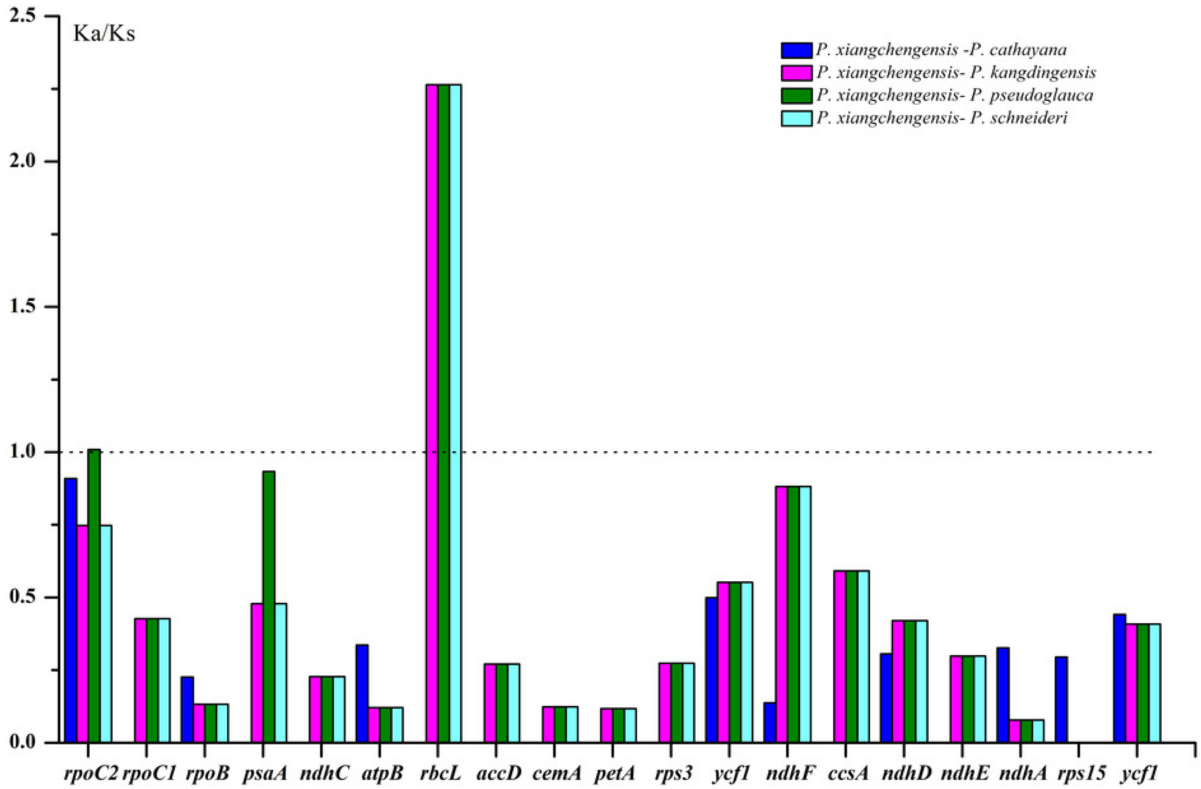

Figure 9

Comparison of long repeat among five *Populus* plastomes

(A) Number of each repeat type; (B) Frequency of each repeat type by length

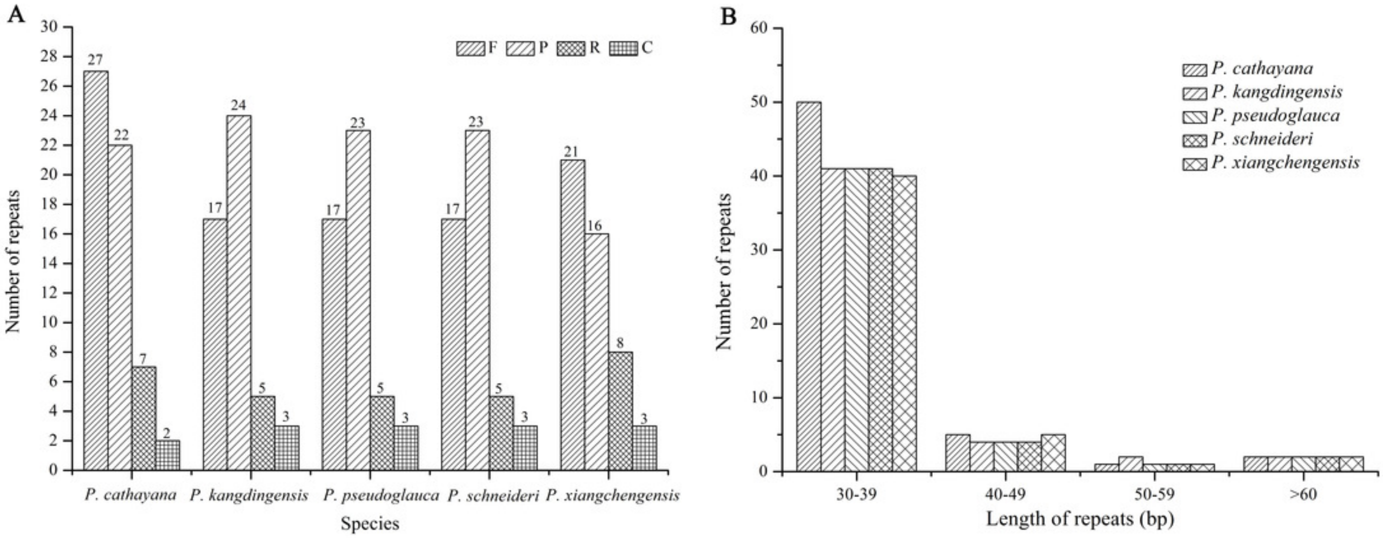

Figure 10

Molecular phylogenetic tree of 27 species in the family *Salicaceae* based on the complete plastome sequence

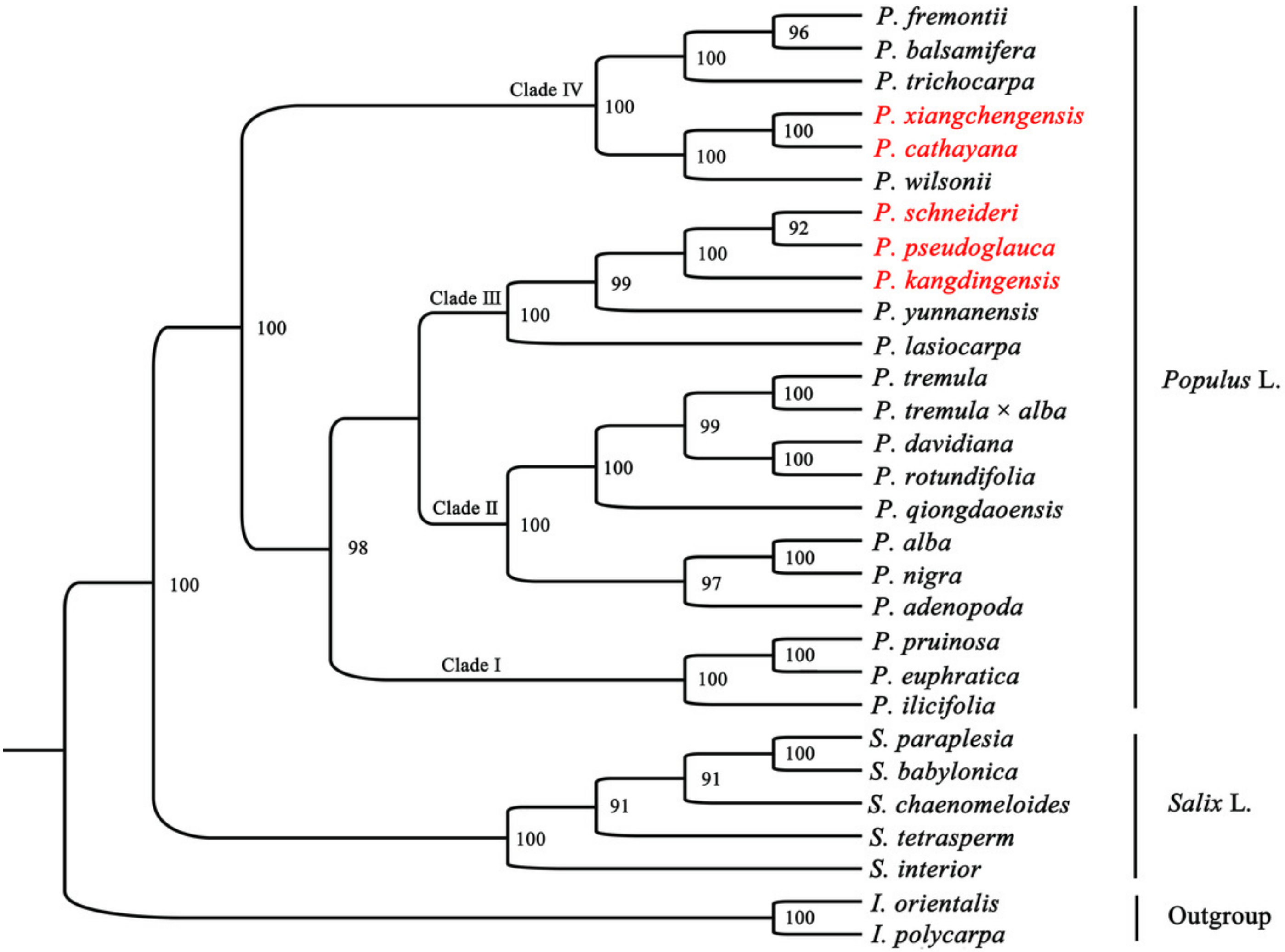

Table 1(on next page)

The features of five *Populus* plastomes

Table 1. The features of five *Populus* plastomes

Species	**Size (bp)**	**LSC (bp)**	**SSC(bp)**	**IR(bp)**	**Number of proteincoding genes**	**Number of tRNA genes**	**Number of rRNA genes**	**GC content （%）**
P. cathayana	156,789	84,851	16,594	27,672	85(7)	37(7)	8(4)	36.7
P. kangdingensis	156,523	84,788	16,495	27,620	85(7)	37(7)	8(4)	36.7
P. pseudoglauca	156,512	84,777	16,495	27,620	85(7)	37(7)	8(4)	36.7
P. schneideri	156,513	84,778	16,495	27,620	85(7)	37(7)	8(4)	36.7
P. xiangchengensis	156,465	84,812	16,513	27,570	85(7)	37(7)	8(4)	36.7

Table 2(on next page)

Base composition of the five *Populus* plastomes

Table 2. Base composition of the five *Populus* plastomes

Region		***P. cathayana***	***P. kangdingensis***	***P. pseudoglauca***	***P. schneideri***	***P. xiangchengensis***
LSC(%)	A	32.0	32.1	32.1	32.1	32.1
	T	33.4	33.4	33.4	33.4	33.4
	C	17.7	17.7	17.7	17.7	17.7
	G	16.8	16.8	16.8	16.8	16.8
	GC	34.6	34.5	34.5	34.5	34.5
SSC(%)	A	34.9	34.9	34.9	34.9	34.9
	T	34.5	34.6	34.6	34.6	34.3
	C	16.1	16.1	16.1	16.1	16.1
	G	14.6	14.4	14.4	14.4	14.6
	GC	30.6	30.5	30.5	30.5	30.7
IR(%)	A	28.9	29.0	29.0	29.0	29.0
	T	29.1	29.0	29.0	29.0	29.1
	C	21.8	21.8	21.8	21.8	21.8
	G	21.8	20.1	20.1	20.1	20.2
	GC	41.9	42.0	42.0	42.0	42.0
Overall length (%)	A	31.3	31.3	31.3	31.3	31.3
	T	32.0	32.0	32.0	32.0	32.0
	C	18.7	18.7	18.7	18.7	18.7
	G	18.0	18.0	18.0	18.0	18.1
	GC	36.7	36.7	36.7	36.7	36.7

Table 3(on next page)

Pairwise nucleotide divergences of the five *Populus* plastomes

Table 3. Pairwise nucleotide divergences of the five *Populus* plastomes

Species	*P. cathayana*	*P. kangdingensis*	*P. pseudoglauca*	*P. schneideri*	*P. xiangchengensis*
P. cathayana	-				
P. kangdingensis	0.00124	-			
P. pseudoglauca	0.00333	0.00326	-		
P. schneideri	0.00335	0.00327	0.00003	-	
P. xiangchengensis	0.00333	0.00325	0.00001	0.00002	-

Table 4(on next page)

Transitions (Ts) and transversions (Tv) in the protein-coding regions of the four plastomes compared with the plastome of *P. xiangchengensis*

Table 4. Transitions (Ts) and transversions (Tv) in the protein-coding regions of the four plastomes compared with the plastome of *P. xiangchengensis*

Species	**Ts**		**Tv**				Total
	A-G	C-T	A-T	A-C	T-G	G-C	
P. cathayana	14	19	13	7	9	8	70
P. kangdingensis	42	55	12	14	24	13	160
P. pseudoglauca	44	57	12	15	24	14	166
P. schneideri	43	56	12	15	24	14	164

Table 5(on next page)

Statistics of chloroplast SSRs detected in five *Populus* plastomes

Table 5. Statistics of chloroplast SSRs detected in five *Populus* plastomes

SSR type		***P. cathayana***	***P. kangdingensis***	***P. pseudoglauca***	***P. schneideri***	***P. xiangchengensis***
P1	(A)12	3	9	9	9	3
	(A)13	5	5	5	5	4
	(A)14	2	2	2	2	3
	(A)15	2	0	0	0	1
	(A)16	1	2	2	2	0
	(A)17	0	1	1	1	0
	(T)12	4	5	5	5	3
	(T)13	3	3	3	3	3
	(T)14	3	1	1	1	1
	(T)15	1	1	1	1	3
	(T)16	2	4	4	4	1
	(T)17	0	0	0	0	1
	ALL	26	33	33	33	23
P2	TA/AT	1	2	2	2	1
C		1	4	4	4	1
Total		28	39	39	39	25