Phylogenomic Analysis and Dynamic Evolution of Chloroplast Genomes in Salicaceae

Yuan Huang1,2†, Jun Wang3†, Yongping Yang4*, Chuanzhu Fan3* and Jiahui Chen1,4*

1 Key Laboratory for Plant Diversity and Biogeography of East Asia, Chinese Academy of Sciences, Kunming, China, 2 School of Life Sciences, Yunnan Normal University, Kunming, China, 3 Department of Biological Sciences, Wayne State University, Detroit, MI, United States, 4 Institute of Tibetan Plateau Research at Kunming, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China

Chloroplast genomes of plants are highly conserved in both gene order and gene content. Analysis of the whole chloroplast genome is known to provide much more informative DNA sites and thus generates high resolution for plant phylogenies. Here, we report the complete chloroplast genomes of three *Salix* species in family Salicaceae. Phylogeny of Salicaceae inferred from complete chloroplast genomes is generally consistent with previous studies but resolved with higher statistical support. Incongruences of phylogeny, however, are observed in genus *Populus*, which most likely results from homoplasy. By comparing three *Salix* chloroplast genomes with the published chloroplast genomes of other Salicaceae species, we demonstrate that the synteny and length of chloroplast genomes in Salicaceae are highly conserved but experienced dynamic evolution among species. We identify seven positively selected chloroplast genes in Salicaceae, which might be related to the adaptive evolution of Salicaceae species. Comparative chloroplast genome analysis within the family also indicates that some chloroplast genes are lost or became pseudogenes, infer that the chloroplast genes horizontally transferred to the nucleus genome. Based on the complete nucleus genome sequences from two Salicaceae species, we remarkably identify that the entire chloroplast genome is indeed transferred and integrated to the nucleus genome. This observation, along with presence of the large nuclear plastid DNA (NUPTs) and NUPTs-containing multiple chloroplast genes in their original order in the chloroplast genome, favors the DNA-mediated hypothesis of organelle to nucleus DNA transfer. Overall, the phylogenomic analysis using chloroplast complete genomes clearly elucidates the phylogeny of Salicaceae. The identification of positively selected chloroplast genes and dynamic chloroplast-to-nucleus gene transfers in Salicaceae provide resources to better understand the successful adaptation of Salicaceae species.

Keywords: chloroplast genome, phylogenomics, phylogenetic incongruence, NUPT, evolution, organellar horizontal gene transfer, Salicaceae
INTRODUCTION

The chloroplast (cp) is the photosynthetic organelle that provides energy for plants and algae. It is believed that chloroplasts arose from endosymbiosis between a photosynthetic bacterium and a non-photosynthetic host (Dyall et al., 2004). The chloroplast has its own genome, which is generally non-recombinant and uniparentally inherited (Birky, 1995). Chloroplast genes are involved in major functions, which include sugar synthesis, starch storage, the production of several amino acids, lipids, vitamins, and pigments. They are also involved in key sulfur and nitrogen metabolic pathways. In angiosperms, most cp genomes are composed of circular DNA molecules ranging from 120 to 160 kb in length and have a quadripartite organization consisting of two copies of inverted repeats (IRs) of about 20–28 kb in size. These IRs divide the rest of cp genome into an 80–90 kb Large Single Copy (LSC) region and a 16–27 kb Small Single Copy (SSC) region (Jansen et al., 2005). The gene content and order of cp genomes of angiosperms are generally conserved, which encode four rRNAs, 30 tRNAs, and about 80 unique proteins (Chumley et al., 2006).

Chloroplast-derived DNA sequences have been widely used for phylogenetic studies, and complete cp genome sequences could provide valuable data sets for resolving complex evolutionary relationships (Jansen et al., 2007; Moore et al., 2010). However, acquiring large coverage of cp genomes has typically been limited by conventional DNA sequencing technology. As Next-Generation Sequencing technologies have revolutionized DNA sequencing (Shendure and Ji, 2008), it is now more convenient to obtain complete cp genome sequences with low cost, to extend gene-based phylogenetics to genome-based phylogenomics, and to examine phylogeny and evolutionary events of plant species using complete entire cp genome sequences.

Salicaceae s.str consists of two genera (Ohashi, 2001): Salix with about 450–580 species (Fang et al., 1999; Argus et al., 2010), and Populus with about 30 species (Argus et al., 2010). Species of Salicaceae are widely distributed in the world, except in Oceania and Antarctica. They are mostly found in the Northern Template Zone and are one of the main groups of trees and shrubs in those areas (Skvortsov, 1999; Argus et al., 2010). Plants of Salicaceae are often grown for ornament, shelterbelts, timber, pulp, and specialty wood products. Some shrub species of Salix are deemed as most suitable for bioenergy crops (Karp and Shield, 2008).

Because of dioecious reproduction, simple flowers, common natural hybridization, and large intraspecific phenotypic variation, both the resolution of taxonomy and the systematics of Salicaceae based on morphology, especially Salix, have been extremely difficult (Skvortsov, 1999; Argus et al., 2010). Molecular methods (e.g., molecular marker techniques, molecular phylogenetics and DNA barcoding) provide effective information for taxonomy, species identification, and phylogenetics of Salicaceae. However, previous molecular systematic analyses revealed that the phylogeny of Salicaceae, based on single or a few genetic markers, succeeds in resolving relationships in generic or sub-generic levels, but limits or has almost no resolution in infra-subgeneric level, specifically in subgenera Chameatia-Vetrix clade (Leskinen and Alstrom-Rapaport, 1999; Hamzeh and Dayanandan, 2004; Chen et al., 2008; Hardig et al., 2010). Therefore, DNA markers with higher resolution for phylogenetic analysis of unresolved lineages remain to be examined in Salicaceae.

Here, we report the complete cp genome sequences of three Salix species and further integrate the 11 available cp genomes of Salicaceae. Hence, all of the main lineages of Salicaceae have their representative species present in this study. The questions that we addressed in this study are: (1) What are potential DNA markers in cp genomes that can be used for phylogenetic analysis of Salicaceae? (2) What is the phylogenetic relation of Salicaceae based on phylogenomic analysis of complete Salicaceae cp genomes? (3) What are the structures and contents of cp genomes in Salicaceae? and (4) What are the evolution and dynamics patterns of cp genomes revealed by examining evolution of cp genes and DNA horizontally transferring events from cp to nucleus?

MATERIALS AND METHODS

Plant Materials

Three species, S. tetrasperma, S. babylonica, and S. oreinoma, representing two subgenera of the genus Salix, were sampled. We collected healthy, tender and fresh leaves from adult plants of target species. The voucher herbarium specimens for the three sampled Salix species are all deposited in Herbarium of Kunming Institute of Botany, Chinese Academy of Sciences (KUN) (Supplementary Table S1).

Chloroplast DNA Extraction, Sequencing, Genome, Assembly, and PCR-Based Validation

Total DNA enriched with cp DNA was extracted from 200 g of fresh leaves according to the methods of Jansen et al. (2005) and Zhang et al. (2011). Purified DNA (5 mg) was fragmented and used to construct short-insert libraries according to the manufacturer's manual (Illumina Inc., San Diego, CA, United States). DNA from the different individuals was indexed by tags and pooled together in one lane of Illumina's Genome Analyzer for sequencing.

We filtered out non-cp DNA reads from the raw sequences based on the known cp genome sequences. Next, the filtered reads were used to de novo assemble the cp genomes with SOAPdenovo software, which is specially designed to assemble Illumina short reads. SOAPdenovo pipeline (e.g., k = 31 bp and scaffolding contigs with a minimum size of 100 bp) can carry out accurate analyses of unexplored genomes, resolve repeat regions in contig assembly and improve gap closing, etc., in a cost effective way (Luo et al., 2012). Then, all contigs were mapped to the reference cp genome of P. trichocarpa using BLAST1 search from NCBI with default parameters. The orders of aligned contigs were determined according to the reference genome. Gaps between the

**Molecular Evolution Analysis**

We collected the coding DNA sequence (CDS) of each orthologous gene in eight *Populus* species and six *Salix* species, and aligned them with translation aligner using Geneious (Kearse et al., 2012), which could take into account frame shifts and premature stop codons, and generate codon-based alignments. We used the branch site model of PAML to compute the \( Ka/Ks \) ratio for orthologous genes in each external and internal branch of phylogeny trees that were generated based on protein coding sequence alignment with ML method. We tested two branch site models (with the parameters model = 2 and NSites = 2): the “model 1” with both the branch site specific \( Ka/Ks \) and background \( Ka/Ks \) varying freely, and the “model 2” with the branch site specific \( Ka/Ks \) fixed at 1 and background \( Ka/Ks \) varying freely (Yang, 2007). We then performed the Likelihood Ratio Test (LRT), which tests whether the likelihood of the “model 1” is significantly different from that of the “model 2” by comparing two times the log likelihood difference. We computed \( p \)-values using a chi-square distribution with one degree of freedom (Yang, 1998).

**Analysis of Chloroplast-Nuclear DNA Transfer**

We used the pipelines developed by UCSC genome browser to search for the homologous regions of the chloroplast genome in the nuclear genome for *P. trichocarpa* and *S. purpurea*, respectively, since the nuclear genome sequences are only available for these two species (Tuskan et al., 2006). We first aligned the chloroplast genome to the nuclear genome of the two reference genomes with LASTZ (Harris, 2007). We then transformed the “lav” output format of LASTZ to “axt” format using lavToAxt. Finally we chained the “axt” files using axtChain and generated chain format outputs (Kent et al., 2003; Schwartz et al., 2003).

Although we did not set the identity filter of LASTZ for blocks or HSPs (high-scoring segment pairs), we computed the identity of the final homologous sequence pairs generated by axtChain. The identity ranges from 60 to 99%. Based on the chain file, we generated the homologous regions between the chloroplast genome and nuclear genome, which could represent chloroplast-nuclear DNA transfer events.

**Monte Carlo Sampling Testing**

To test the hypotheses of whole chloroplast genome horizontally transferring to nucleus in *P. trichocarpa*, we conducted 100,000 Monte Carlo sampling test, which is implemented by the comparison of the observed data with random samples generated in accordance with the hypothesis being tested. The significance of the Monte Carlo sampling test is determined by the rank of the test criterion of the observed data relative to the test criteria of the random samples composing the reference set (Hope, 1968). Specifically, we assumed the whole chloroplast genome was split into eight fragments, which were inferred from the alignment between the chloroplast genome and nuclear genome in species *P. trichocarpa*. For each simulation, we randomly sampled the insertion locations of the eight fragments in the whole *P. trichocarpa* genome. We then counted the
number of simulations, in which the five fragments (i.e., the 1st, 3rd, 5th, 7th, and 8th fragments) were inserted into the same chromosome within the range of the length of chromosome 13 and were arranged according to their previous order in the chloroplast genome. Next, we divided this number by the number of simulations (i.e., 100,000). This mathematical derivative is treated as the P-value for inference of the testing hypothesis significance.

RESULTS

Genome Assembly and PCR-Based Validation

Using the Illumina Hiseq 2000 system, we obtain 1,392,310, 3,779,094, and 4,595,286 bp paired-end clean reads (average read length 91 bp) for S. babylonica, S. tetrasperma, and S. oreinoma, respectively. We mapped these sequence reads to the reference cp genome of S. purpurea and achieved at least 1600 × (1615 × for S. babylonica, 4384 × for S. tetrasperma, 5330 × for S. oreinoma) coverage for cp genome. Based on de novo and reference-guide assembly, we obtain the complete cp genome for S. tetrasperma. The assemblies of the other two cp genomes contain seven to eight gaps, and we filled the gaps using PCR-based sequencing.

Four junction regions of cp genomes were validated using PCR-based sequencing for each cp genome, respectively. Furthermore, in order to overcome the errors of heterogeneous indels from homopolymeric repeats (Moore et al., 2006; Yang et al., 2010), we conducted PCR-based validation to correct the errors. We designed 18 pairs of primers based on the variable regions of alignments to validate these sequences in each cp genome (Supplementary Table S2). In total, we amplified and sequenced ~152 kb from all the three Salix species. Then, we compared these sequences directly to the assembled genomes and observed no nucleotide mismatches or indels/insertions. This result confirmed the reliability of assembled chloroplast genome sequences. Finally, we obtain the complete cp genome sequences of S. babylonica, S. tetrasperma, and S. oreinoma.

Salicaceae Chloroplast Genome Structure and Content

The complete cp genomes of the three Salix species sequenced vary from 155,531 to 155,740 bp in size and exhibit a typical circular structure including a pair of IRs (range from 27384 to 27436 bp) and two single-copy regions (LSC, 84466–84580 bp; SSC, 15862–16323 bp). Each of the three cp genomes contains 111 unique genes (110 for S. babylonica) (Figure 1 and Table 1), including 77 unique CDSs (76 for S. babylonica because of the pseudogenization of ndhA), four unique rRNAs, and 30 unique tRNAs. Seventeen genes contain introns; 14 of them (atpF, ndhA, ndhB, petB, petD, rpl2, rpl16, rpoC1, trnA\textsuperscript{UGC}, trnG\textsuperscript{GCC}, trnI\textsuperscript{GAU}, trnK\textsuperscript{UUU}, trnL\textsuperscript{UAA}, and trnV\textsuperscript{UAC}) exhibit one intron and three of them (clpP, rpl12, and ycf3) contain two introns. All the CDSs have canonical ATG start codons except ndhD, which has GTG as the start codon.

Molecular Marker Identification

Based on cp genome alignment from 14 Salicaceae species, we identified 30 highly divergent non-coding regions for implementation of phylogenetic analysis in Salix, Populus and Salicaceae. Those molecular markers are mostly derived from intergenic or intronic non-coding regions, though four protein-coding genes including ccsA, rpl20, rps7 and ndhA/ndhE are also identified. Among the four genes, rps7 is highly diverged and lineage-specific within genus Salix. The length of rps7 is 468 bp for all Populus species except P. cathayana, whereas the three Salix species (i.e. S. oreinoma, S. suchowensis, and S. purpurea), which are closely related to subgenera Chamaetia and Vetrix, have identical rps7 sequences with a length of 270 bp. The other three Salix species (i.e. S. interior, S. babylonica, and S. tetrasperma) in subgenus Salix have identical rps7 sequences with a length of only 180 bp (Figures 2A, 3, and Supplementary Table S3).

The phylogenetic resolution of the genus Salix is proven to be extremely difficult (Skvortsov, 1999; Argus et al., 2010; Wu et al., 2015) and no agreement on Salix phylogeny is made thus far. Molecular phylogenetic analysis in Salix only succeeds in resolving relationships within the subgeneric level, but fails in the taxonomic level under subgenus and shows almost no resolution in the clade with about 73% of Salix species (i.e., subgenera Chamaetia and Vetrix) (Leskinen and Alstrom-Rapaport, 1999; Azuma et al., 2000; Chen et al., 2010; Hardig et al., 2010; Abdollahzadeh et al., 2011; Wu et al., 2015). Therefore, there is a need to identify highly divergent regions in Salix cp genomes as molecular markers.
TABLE 1 | Features of chloroplast complete genomes in Salicaceae.

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome size</th>
<th>Length of LSC (bp)</th>
<th>Length of SSC (bp)</th>
<th>Length of IRB (bp)</th>
<th>Length of IRB (bp)</th>
<th>GC content (%)</th>
<th>Unique protein coding gene</th>
<th>Unique non-coding RNA</th>
<th>Unique tRNA</th>
<th>Unique rRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salix tetrasperma</td>
<td>155246</td>
<td>84516</td>
<td>15830</td>
<td>27450</td>
<td>27450</td>
<td>58.28</td>
<td>79005</td>
<td>64763</td>
<td>112</td>
<td>76</td>
</tr>
<tr>
<td>S. babylonica</td>
<td>155531</td>
<td>84470</td>
<td>16213</td>
<td>27424</td>
<td>27424</td>
<td>59.17</td>
<td>80187</td>
<td>63505</td>
<td>113</td>
<td>77</td>
</tr>
<tr>
<td>S. oreinoma</td>
<td>156620</td>
<td>85980</td>
<td>16308</td>
<td>27166</td>
<td>27166</td>
<td>58.57</td>
<td>80079</td>
<td>64882</td>
<td>113</td>
<td>77</td>
</tr>
<tr>
<td>S. interior</td>
<td>155590</td>
<td>84452</td>
<td>16220</td>
<td>27459</td>
<td>27459</td>
<td>59.14</td>
<td>80175</td>
<td>63576</td>
<td>113</td>
<td>77</td>
</tr>
<tr>
<td>S. purpurea</td>
<td>155214</td>
<td>84077</td>
<td>16221</td>
<td>27458</td>
<td>27458</td>
<td>59.28</td>
<td>80175</td>
<td>63200</td>
<td>113</td>
<td>77</td>
</tr>
<tr>
<td>Populus alba</td>
<td>156766</td>
<td>84887</td>
<td>16589</td>
<td>27644</td>
<td>27646</td>
<td>58.95</td>
<td>80568</td>
<td>64360</td>
<td>113</td>
<td>77</td>
</tr>
<tr>
<td>P. fremontii</td>
<td>156848</td>
<td>84827</td>
<td>16589</td>
<td>27644</td>
<td>27646</td>
<td>58.95</td>
<td>80568</td>
<td>63868</td>
<td>113</td>
<td>77</td>
</tr>
<tr>
<td>P. trichocarpa</td>
<td>156067</td>
<td>84377</td>
<td>16490</td>
<td>27600</td>
<td>27600</td>
<td>59.25</td>
<td>80634</td>
<td>63594</td>
<td>113</td>
<td>77</td>
</tr>
<tr>
<td>P. yunnanensis</td>
<td>155449</td>
<td>83911</td>
<td>16490</td>
<td>27525</td>
<td>27525</td>
<td>59.46</td>
<td>80584</td>
<td>63200</td>
<td>113</td>
<td>77</td>
</tr>
</tbody>
</table>

Sequence lengths are measured by bp. cp genome sequenced in this study. *cp genome sequenced in this study.

Phylogenomic Analyses of Salicaceae

The genus *Populus* is comprised of ca. 30 species, which can be divided into six major clades (Eckenwalder, 1996; Argus et al., 2010). We selected eight species from four clades in our phylogenomic analyses. Our analysis indicates that *Populus* is monophyletic. Phylogeny, based on whole genome and concatenated non-coding datasets, shares the same topology (Figure 2B). However, phylogenetic relationship based on concatenated protein coding dataset is different in the topological positions of *P. euphratica*, *P. cathayana* and *P. balsamifera* (Figure 2A), as indicated by network analysis (Figure 2C). The three New World *Populus* species (*P. trichocarpa*, *P. balsamifera*, *P. fremontii*), Clade A in Figure 2A) form a robust monophyly. This clade is sister to the remaining *Populus* species in the protein coding gene tree (Figure 2A). However, *P. euphratica* is sister to *P. cathayana* from the tree based on whole cp genome and non-coding regions (Figure 2B). In the tree based on protein coding regions, *P. euphratica* and *P. cathayana* fall in the clade that contains all species from Old World with low support value and short branch length (clade B in Figure 2A). This topology is absent from previous studies with more *Populus* species sampled (Hamzeh and Dayanandan, 2004; Wang et al., 2014), indicating that insufficient sampling might have caused this topology.

Molecular Evolution of Salicaceae Chloroplast Genome

We identified at least four possible pseudogenization events in *Salicaceae*: (1) *ndhA* in *S. babonica*, which contains a 1039 bp deletion including a start codon in exon 1; (2) *accD* in *P. yunnanensis*, which loses 1056 bp (72%) of gene content; (3) *ycf1* in *P. yunnanensis*, which is deleted >5000 bp (94%) of DNA sequences; (4) *psbC* in *P. cathayana*, which misses ~1000 bp (~70%) of DNA sequences (Figure 2A).

We identified seven significantly positively selected genes (*atpE*, *rps7*, *ycf2*, *ccsA*, *petD*, *psbC*, and *psbJ*) that contain positively selected sites (Table 2). Among them, three positively selected genes (i.e., *rps7*, *petD*, and *psbC*) are identified in *P. cathayana*, *ycf2* is identified in *P. yunnanensis*, *ccsA* and *psbJ* are identified in *P. tremula*. An exception, *atpE*, shows significant positive selection on one site in the *Salix* subgenus *Salix* clade (i.e., *S. babonica*, *S. tetrasperma*, and *S. interior*) (Table 2).

Chloroplast-Nuclear DNA Transfer

We used LASTZ (Harris, 2007) to search for NUPT including DNA fragments less than 200 bp. We identified 571 and 713 NUPTs with total ungapped lengths of ca. 536 and 193 kb in *P. trichocarpa* and *S. purpurea*, respectively (Figure 4, Table 3, and Supplementary Table S7). This result is different from a previous study that used BLAST method for searching NUPT (Yoshida et al., 2013). The number of NUPTs in *S. purpurea* is larger than that in *P. trichocarpa*, but the total length of NUPTs in *S. purpurea* is much lower than that in *P. trichocarpa* (Table 3).

The fragmented assembly of *S. purpurea* genome (2015) might cause uncertainty for NUPT searching. Therefore, we subsequently discuss the plastid-to-nucleus DNA transfer in...
Salicaceae, based on the analysis from *P. trichocarpa*. The most abundant NUPTs in *P. trichocarpa* are around 70–199 bp (Figure 4). The number of NUPTs with 1000 bp longer is 77, and these large NUPTs are responsible for the majority of the total NUPT length (>75%, ca. 410 kb out of 536 kb). Furthermore, the average identity of those large NUPTs is 90.2%, and that of the six largest NUPTs with length greater than 10 kb is 99.4% (Supplementary Table S7), indicating that large NUPTs
TABLE 2 | Positive selection sites identified using Codeml under branch-site model.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Branch</th>
<th>Null</th>
<th>Alternative</th>
<th>p-value</th>
<th>Putative sites under positive selection, amino acid, and corresponding posterior probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>atpE</td>
<td>S. babylonica; S. tetrasperma; S. interior</td>
<td>−578.123</td>
<td>566.714</td>
<td>1.78E−06</td>
<td>130 S 0.970*</td>
</tr>
<tr>
<td>rps7</td>
<td>P. cathayana</td>
<td>−445.250</td>
<td>437.826</td>
<td>1.17E−04</td>
<td>E−06 130 S 0.970; 45 Y 0.998; 46 A 0.987; 47 M 0.961; 48 K 0.952; 52 Q 0.978; 54 T 0.965*</td>
</tr>
<tr>
<td>ycf2</td>
<td>P. yunnanensis</td>
<td>−9664.014</td>
<td>−9628.861</td>
<td>5.07E−17</td>
<td>2271 M 0.982; 2272 A 0.982; 2275 G 0.969*</td>
</tr>
<tr>
<td>ccsA</td>
<td>P. tremula</td>
<td>−1585.549</td>
<td>−1581.119</td>
<td>2.92E−03</td>
<td>319 I 0.959*</td>
</tr>
<tr>
<td>petD</td>
<td>P. cathayana</td>
<td>−499.511</td>
<td>491.763</td>
<td>8.27E−05</td>
<td>312 M 0.982; 313 A 0.982; 315 G 0.969*</td>
</tr>
<tr>
<td>psbC</td>
<td>P. cathayana</td>
<td>−872.644</td>
<td>829.745</td>
<td>1.99E−20</td>
<td>101 E 1.000; 102 V 1.000; 103 I 0.999; 104 D 0.997; 105 T 0.979; 106 F 0.998; 107 P 0.998; 108 Y 0.978; 109 F 0.974; 110 V 0.999; 111 S 1.000; 112 G 0.999; 113 V 1.000; 114 L 0.980; 115 H 1.000; 116 L 1.000; 117 I 0.974; 119 S 0.988; 120 A 1.000; 121 V 0.996; 122 L 0.995; 123 G 0.979; 124 F 0.999; 125 G 0.998; 127 I 0.978; 128 Y 0.998; 129 H 0.996; 130 A 1.000; 131 L 0.992; 132 L 0.996; 133 G 0.962; 134 P 0.980; 135 E 1.000; 136 T 0.987; 137 L 0.999; 138 E 0.998; 139 E 0.974*</td>
</tr>
<tr>
<td>psbJ</td>
<td>P. tremula</td>
<td>−196.515</td>
<td>−192.834</td>
<td>6.66E−03</td>
<td>20 P 0.981*</td>
</tr>
</tbody>
</table>

*Posterior probability > 95%, **Posterior probability > 99%. "Null" and "Alternative" columns list likelihood values obtained under the null model and alternative model.

FIGURE 4 | Illustration of NUPTs size distribution in Populus trichocarpa and Salix purpurea.

are relatively recently integrated into the nucleus (see Data Sheet S1 for detail). Additionally, all large NUPTs contain large gaps.

In our analyses, we identified five aligned regions (~4,000–16,000 bp in length) and three/four gaps (~7,000–15,000 bp in length) between the chloroplast and nuclear genomes in the same order as in the cp genome by LASTZ (Figure 5). We conducted MC sampling to demonstrate that it is more likely due to the insertion of the whole cp genome into the nuclear genome and generation of the gaps through insertion/translocation rather than the separate insertions of split cp genome fragments. Previously, it is also discovered that large cp fragments can be inserted into the nuclear genome based on the similarity between cp genome sequence and nuclear genome sequence using the sequence alignment tools (e.g., BLAST), and sequencing the insertion junction product (Yuan et al., 2002; Yoshida et al., 2013). Here, we used LASTZ as the alignment tool, which is a pairwise aligner for aligning DNA sequences and is originally designed to align the sequences in the size of human chromosomes and from different species.
Huang et al. Dynamic Chloroplast Genome Evolution in Salicaceae

We compared the three sequenced *Salix* cp genomes in this study with previous published complete cp genomes of 11 Salicaceae species, including eight *Populus* and three *Salix* species (Supplementary Table S2). We found that the structure and synteny of the cp genomes of the 14 Salicaceae species are highly conserved (Supplementary Figure S1). Also, the lengths of various segments, or parts (IRs, LSC, SSC, coding sequence and non-coding sequence), of the cp genomes are also quite conserved and vary in a small range (Supplementary Figure S2).

We observed that four genes (i.g. *infA*, *rps16*, *rpl32*, and *ycf68*) were lost from Salicaceae cp genomes compared with other angiosperms, among which, *infA* and *rps32* were reported to be transferred into the nucleus (Millen et al., 2001; Ueda et al., 2007). The inverted repeat of the Salicaceae cp genome results in the complete duplication of *rps19*, *rpl2*, *rpl23*, *ycf2*, *ycf15*, *ndhB* and *rps7*, as well as exons 1 and 2 of *rps12*, all four rRNA genes (4.5S, 5S, 16S, and 23S) and seven tRNA genes (*trnI*^G AU^, *trnL*^C AU^, *trnV*^G AC^, *trnG*^A UC^, *trnA*^U GC^, *trnR*^A CC^, and *trnN*^G UU^). Both IRs of the Salicaceae cp genome run 1,695–1,907 bp into *ycf1*, which differs largely in genera *Salix* and *Populus*: 1,747 bp in all *Salix* species and 1,695–1,907 bp in *Populus* species, which result in most proportions of fluctuations in IR length. Additionally, IRb runs 50–51 bp into *rpl22* (Supplementary Figure S3).

Similar to other angiosperms, the IR (pairwise identity 98.1% for both IRa and IRb) regions are more conserved in the nine species than the LSC (pairwise identity 94.0%) and SSC (pairwise identity 94.9%) regions. Differences are observed in the cp genome in the 14 Salicaceae species including genome (Harris, 2007). Therefore, it is more suitable to detect similarity in a large scale, such as between the cp genome and the nuclear genome.
size, gene losses, the pseudogenization of protein-coding genes, and IR expansion and contraction (Table 1 and Supplementary Figure S2).

**Phylogenomic Analyses of Salicaceae**

Phylogenetic analysis, based on the complete cp genomes and non-coding and protein-coding datasets, indicates that the phylogeny of Salicaceae s.str, *Populus* and *Salix* s.l. are all resolved as monophyletic, which is largely consistent with previous studies (Davis et al., 2005; Wurdack and Davis, 2009; Chen et al., 2010; Wang et al., 2014; Wu et al., 2015) (Figure 2 and Supplementary Table S4). However, we identified the incongruence of phylogenies in Salicaceae that are inferred from these three datasets (i.e., whole genome and concatenated coding sequence between non-coding sequences of the cp genome). This incongruence might be caused by homoplasy, which could result from convergence, parallelism or reversal. Because the cp genome is inherited maternally as a single unit, the observed phylogenetic tree incongruences in our result are most unlikely caused by lineage sorting or hybridization/introgression, which are generally used to explain the conflict signals among characters of plant taxa including Salicaceae (Wang et al., 2014; Wu et al., 2015).

Our phylogenetic tree, based on the whole cp genome, coding-sequence and non-coding sequences of cp genome, shows that all conflicting branches are short (Figures 2A, B), which could be caused by fast species radiation or short stem-lineages (Wagele and Mayer, 2007). In this case, apomorphies evolved in stem-lineage may be rare, and chance similarities (non-phylogenetic signal produced by nucleotide substitution processes) that evolved later can accumulate and dominate in the form of phylogenetic-signal-like (i.e., false and misleading phylogenetic signals) patterns. It is difficult to distinguish these kinds of homoplasies from apomorphies, and consequently, this might lead to the wrong phylogenetic tree (Wagele and Mayer, 2007). Moreover, multiple nucleotide substitutions along long branches may destroy synapomorphies, resulting in accumulation of homoplases along long branches and attracting distantly related clades to be clustered together in a topology (i.e., long-branch attraction) (Wagele and Mayer, 2007). Therefore, the conflict topologies in our phylogenetic trees most likely result from incomplete sampling or homoplasy. This is in line with previous cp DNA markers based on phylogenomic study of bamboo, which comes to the conclusion that homoplasy should be in account for conflicting phylogenetic signals between cp genome datasets (Ma et al., 2014). Moreover, our analyses show that the phylogenetic tree based on genomic data with large number of informative characters (e.g., cp genomes) should be used in caution when the examined taxa are quickly radiated or bear short branches.

Phylogenomic analysis indicates that the species of *Salix* were clustered in a robust monophyletic clade. Two robust subclades were resolved within *Salix*. One subclade includes the species of subgenera *Chametia* and Vetric. Two species (*S. suchowensis* and *S. purpurea*) from subgenus *Vetric* cluster as a monophyletic group, which is sister to a group including a species, *S. oceinoma* from subgenus *Chamaetia*. The other subclade is comprised of species from subgenus *Salix*, the New World species, *S. interior*, which is sister to a group containing two Old World species, *S. babylonica* and *S. tetrasperma* (Figure 2). The relationships resolved in this study are in line with previous phylogenetic studies of *Salix* (Azuma et al., 2000; Chen et al., 2010; Hardig et al., 2010; Wu et al., 2015), but we provide evidence from a phylogenomic perspective.

**Positively Selected Genes of Salicaceae Chloroplast Genomes**

In the evolutionary process of a certain lineage of an organism, changing environments (e.g., climate changing) impose selective pressures and result in adaptive evolution. Identification of genes involved in this process (i.e., positively selected genes) is central to understanding the evolutionary pattern of organisms (Yang, 1998), and pinpointing specific targets for adaptive studies.

Among seven significantly positively selected genes in Salicaceae cp genomes, gene *rps7* shows a frame shift mutation. A single nucleotide insertion at site 102 causes a frame shift mutation for gene *rps7* and results in a large variation of an amino acid sequence that is shown under positive selection (Figure 3). This frame shift mutation further introduces an early stop codon and shortens its amino acid sequence about 42% of the length compared to other *Populus* species. The *rps7* gene encodes the ribosome S7 protein, also known as ribosomal protein S7 (uS7), which is crucial for the assembly and stability of the ribosome. S7 protein is an important part of the translation process, and is universally present in the small subunit of prokaryotic and eukaryotic ribosomes, and might play either a general or a specific regulatory role in translation initiation in the chloroplast (Fargo et al., 2001). The *rps7* gene of *P. cathayana* is positively selected, although it contains a premature stop codon caused by a frame shift mutation compared with other *Populus* species surveyed in this study (Figure 3). Despite the premature stop codon, this truncated *rps7* gene in *P. cathayana* has a conserved domain for the uS7 superfamily, as revealed by NCBI-CDD blast results (Supplementary Figure S4). This reveals that this shortened *rps7* gene may still function normally. Alternatively, even if the truncated cp *rps7* cannot maintain its original function, it is possible that an intact copy of cp *rps7* has been transferred into the nucleus and can properly function in the nucleus genome. *P. cathayana* does not have complete genome sequences, so we searched the cp *rps7* gene in nucleus genome of *P. trichocarpa*, which has fine quality whole nuclear genome sequences. We observe that the Potri.013G138900.1 gene in chromosome 13 is identical to cp *rps7* in length and has 100% identity for both gene and coding sequences. This implies that *rps7* may have functionally transferred into the nucleus and the cp copy of *rps7* might be freely subject to natural selection.

Similarly, the protein coding sequences of *petD* and *psbC* genes in *P. cathayana* are shortened by about 29 and 70%, respectively. Compared to other Salicaceae species, both mutations are caused by a single nucleotide insertion. The *petD* gene encodes subunit IV of the cytochrome b6/f complex. It is required for photosynthetic electron transport and hence,
supports photosynthetic growth. The mutations in the 5′ UTR or initiation codon can affect its function (Chen et al., 1993; Sakamoto et al., 1994; Sturm et al., 1994). The \textit{psbC} gene encodes one of the components of the core complex of photosystem II. It binds chlorophyll and helps catalyze the primary light-induced photochemical processes of PSII (Rochaix et al., 1989; Cai et al., 2010). However, the effects of shortened coding sequences of \textit{petD} and \textit{psbC} on their function remains unknown, especially for \textit{psbC}, which is shortened about 70% of the length. Our analysis finds that \textit{petD} and \textit{psbC} have been transferred into the coding region of the nuclear genome for three and eight times in \textit{P. trichocarpa}, respectively (Supplementary Tables S5, S6).

Three of the last few amino acids of \textit{ycf2} in \textit{P. yunnanensis} are detected to be under significant positive selection. These amino acid changes are caused by a frame shift (a single nucleotide insertion) 6 bp ahead of them. Consequently, \textit{ycf2} in \textit{P. yunnanensis} is 21 bp shorter than in other Salicaceae species. The \textit{ycf2} gene is a putative ATPase with unknown function. This gene exists in many plants, including non-photosynthetic plants. Previous experiments in tobacco indicate that it plays an essential role in cell survival in the tobacco chloroplast (Drescher et al., 2000). Our analysis indicates that \textit{ycf2} is transferred into the nuclear genome 18 times and all targeted in protein-coding regions in \textit{P. trichocarpa} (Supplementary Tables S5, S6).

We identified two positively selected genes, \textit{ccsA} and \textit{psbf} in \textit{P. tremula}. Species of \textit{P. tremula} mainly distribute in cool temperate regions of Europe and Asia. The protein encoded by \textit{ccsA} is a component of the cytochrome \textit{c} synthase complex of the membrane-bound System II, and is required during the biogenesis of \textit{c}-type cytochromes at the step of heme attachment (Xie et al., 1998). The \textit{psbf} encodes one of the components of the core complex of photosystem II. Experiments in tobacco indicate that plants with a mutated \textit{psbf} gene are unable to grow photoautotrophically (Hager et al., 2002). Our analysis shows that \textit{ccsA} has been transferred to the coding region of the nuclear genome once, but \textit{psbf} is not transferred in \textit{P. trichocarpa} (Supplementary Tables S5, S6).

All the positively selected genes locate in terminal species, except \textit{atpE}, which shows significant positive selection on one site in the \textit{Salix} subgenus, \textit{Salix} clade (i.e., \textit{S. babylonica}, \textit{S. tetrasperma}, and \textit{S. interior}) (Table 2). The \textit{atpE} gene encodes the \textit{e} subunit of CF1 of the H\textsuperscript{+}-translocating ATP synthase, and functions in part to prevent wasteful ATP hydrolysis by the enzyme (Cruz et al., 1997). The common ancestors of this clade might be adapted to temperate conditions and diverge in about early Oligocene (Wu et al., 2015). The positively selected \textit{atpE} gene might be related to the adaptive evolution of the common ancestor group of \textit{Salix} subgenus \textit{Salix}.

As mentioned above, most of the positively selected protein-coding genes are transferred into the nucleus in \textit{P. trichocarpa}. Positive selection is generally regarded as evidence of adaptive evolution, and these positive selected genes might have driven the successful adaptation of the selected taxa and/or lineages. However, the specific functions or effects of these positively selected cp genes in the target species remain unknown, and structural biological studies are needed to clarify the implication of these findings.

### Chloroplast-Nuclear DNA Transfer

Most of the chloroplast genes are transferred to nucleus and then deleted from the plastome. However, a transferred cp gene is not readily expressed in the nucleus, nor would it be able to give rise to a product equipped with the capability of returning to the chloroplast and ousting its progenitor chloroplast gene. The NUPT events have happened repeatedly and still are active during endosymbiotic evolution (Timmis et al., 2004; Cullis et al., 2009; Yoshida et al., 2013). Although some previous studies investigated the patterns of genomic integration of NUPTs in plant species, the mechanisms of cp-to-nucleus gene transfer are not well understood (Timmis et al., 2004; Lane, 2011; Yoshida et al., 2013).

The number of NUPTs we identified in \textit{P. trichocarpa} is larger than that from a previous study (Yoshida et al., 2013), which used NCBI-BLASTN for NUPT identification instead of LASTZ, as in our study. However, the total length of NUPTs is similar. There is no clear explanation for the variation of the number and amount of NUPTs in plant species, but it may correlate with genome complexity, proportion of repetitive elements, and/or other factors (Yoshida et al., 2013).

Consistent with Yoshida et al. (2013), our analysis shows that large NUPTs are young. Therefore these large NUPTs may experience rapid recombination, insertion/deletion, and fragmentation; supporting previous results that the plant nuclear genome is in equilibrium between frequent integration and rapid elimination of plastid DNA (Matsuo et al., 2005).

There are two main hypotheses to explain the mechanism of DNA transferring from organelle-to-nucleus: the “bulk DNA” (DNA-mediated) view and “cDNA intermediate” (RNA-mediated) view (Timmis et al., 2004). Given the presence of large NUPTs containing both non-coding and coding regions, our results and most previous studies support “DNA-mediated” transfer hypothesis, (Matsuo et al., 2005; Michalovova et al., 2013; Yoshida et al., 2013). Despite the frequent occurrence of large NUPTs, e.g., nearly the entire chloroplast fragment in rice chromosome 10 (Yuan et al., 2002), whether the whole plastid genome can be transferred to the nucleus genome remains largely unknown or not specifically verified yet. We observe that the entire \textit{P. trichocarpa} chloroplast genome is aligned to a region on chromosome 13 of the \textit{P. trichocarpa} nuclear genome (Data Sheet S1). Together with MC sampling, our study clearly provide the evidence that the whole chloroplast genome horizontally transfers to the nuclear genome.

As discussed above, the whole chloroplast genome can be transferred to the nucleus, so the missing and pseudogenized genes in the Salicaceae chloroplast genome, e.g., \textit{infA}, \textit{ndhA}, \textit{rpl16}, \textit{rpl32}, could be transferred to the nucleus and may function properly. Similar cases have been reported previously, involving \textit{rpl32} in \textit{Populus} (Ueda et al., 2007) and \textit{infA} in some angiosperm species (Millen et al., 2001). Our analyses and other studies show that organellar DNA horizontal transfer in plants is frequent. These transfers are thought to play an important role in gene and genome evolution in plants, and functional transfer of the chloroplast genes may facilitate the regulation of gene expression.
However, functional gene horizontal transferring from organelle-to-nucleus is rare as the transferred coding sequences must acquire gene promoter and terminator sequences for proper transcription in the nucleus, and must also acquire transit peptides necessary for importing the protein back into organelle (Timmis et al., 2004; Stegemann and Bock, 2006). Our result provides additional evidence of the organelle-to-nucleus horizontal functional gene transfer in species of *P. trichocarpa*, where the majority of cp coding sequences are transferred to the nucleus genome, and most of them remain as coding regions in the nucleus genome (Supplementary Tables S5, S6). However, further investigation is needed to determine whether these plastid-coding sequences in nuclear genome could function properly. Because chloroplast-to-nucleus transferred functional genes must acquire promoter and terminator sequences (Eckardt, 2006), it is necessary to identify the regulatory motifs of these chloroplast-to-nucleus transferred genes. Our discovery that large cp genome segments are transferred into the nuclear genome, implies that the transfer of the coding genes might be accompanied by the transfer of their promoter and terminator sequences, which could provide materials for those cp genes to properly function in the nuclear genome. However, further experiments (e.g., western bolt and genetic screen) and analyses are needed to confirm the function of the genes transferred from the chloroplast to the nucleus. Furthermore, the identification of the whole cp genome transfer event in *P. trichocarpa* is based on single genome analysis. It remains unclear whether this event is usual or even fixed in *P. trichocarpa*. Therefore, sequencing the genomes of more individuals is needed to further verify the whole cp genome transfer event in *P. trichocarpa*.

### AUTHOR CONTRIBUTIONS

JC and YY designed research. YH and JC performed research. JC, JW, and YH analyzed data. JC, YH, JW, YY, and CF wrote the paper. All authors read and approved the final manuscript.

### FUNDING

This study is supported by grants from the National Natural Science Foundation of China (NSFC 31590820, 31590823 to Hang Sun, 31270271, 31670198 to JC, 31560062 to YH); the Science and Technology Research Program of Kunming Institute of Botany, the Chinese Academy of Sciences, Grant NO. KIB2016005; and the Youth Innovation Promotion Association, the Chinese Academy of Sciences. CF is supported by start-up fund from Wayne State University.

### SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.01050/full#supplementary-material

### REFERENCES


Huang et al. Dynamic Chloroplast Genome Evolution in Salicaceae...


**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Huang, Wang, Yang, Fan and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.