Analyzing and characterization of the chloroplast genome of *Salix suchowensis*

Congrui Sun [1], **Jie Li** [1], **Xiaogang Dai** [1], **Yingnan Chen** Corresp. [1]

[1] College of Forestry, Nanjing Forestry University, Nanjing, China

Corresponding Author: Yingnan Chen
Email address: chenyingnan@njfu.edu.cn

By screening sequence reads from the chloroplast (cp) genome of *S. suchowensis* that generated by the next generation sequencing platforms, we built the complete circular pseudomolecule for its cp genome. This pseudomolecule is 155,508 bp in length, which has a typical quadripartite structure containing two single copy regions, a large single copy region (LSC 84,385 bp), and a small single copy region (SSC 16,209 bp) separated by inverted repeat regions (IRs 27,457 bp). Gene annotation revealed that the cp genome of *S. suchowensis* encoded 119 unique genes, including 4 ribosome RNA genes, 30 transfer RNA genes, 82 protein-coding genes and 3 pseudogenes. Analyzing the repetitive sequences detected 15 tandem repeats, 16 forward repeats and 5 palindromic repeats. In addition, a total of 188 perfect microsatellites were detected, which were characterized as A/T predominance in nucleotide compositions. Significant shifting of the IR/SSC boundaries was revealed by comparing this cp genome with that of other rosids plants. We also built phylogenetic trees to demonstrate the phylogenetic position of *S. suchowensis* in Rosidae, with 66 orthologous protein-coding genes presented in the cp genomes of 32 species. By sequencing 30 amplicons based on the pseudomolecule, experimental verification achieved accuracy up to 99.84% for the cp genome assembly of *S. suchowensis*. In conclusion, this study built a high quality pseudomolecule for the cp genome of *S. suchowensis*, which is a useful resource for facilitating the development of this shrub willow into a more productive bioenergy crop.

Analyzing and characterization of the chloroplast genome of *Salix suchowensis*

Congrui Sun, Jie Li, Xiaogang Dai, Yingnan Chen*

Co-Innovation Center for Sustainable Forestry in Southern China, College of Forestry, Nanjing Forestry University, Nanjing, China

Running title: The chloroplast genome of *Salix suchowensis*

*Author for Correspondence:

Yingnan Chen

College of Forestry, Nanjing Forestry University, Nanjing, China

Tel: 01186-25-85428165

Fax: 01186-25-85427165

Email: chenyingnan@njfu.edu.cn

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2388v1 | CC BY 4.0 Open Access | rec: 24 Aug 2016, publ: 24 Aug 2016

Abstract: By screening sequence reads from the chloroplast (cp) genome of *S. suchowensis* that generated by the next generation sequencing platforms, we built the complete circular pseudomolecule for its cp genome. This pseudomolecule is 155,508 bp in length, which has a typical quadripartite structure containing two single copy regions, a large single copy region (LSC 84,385 bp), and a small single copy region (SSC 16,209 bp) separated by inverted repeat regions (IRs 27,457 bp). Gene annotation revealed that the cp genome of *S. suchowensis* encoded 119 unique genes, including 4 ribosome RNA genes, 30 transfer RNA genes, 82 protein-coding genes and 3 pseudogenes. Analyzing the repetitive sequences detected 15 tandem repeats, 16 forward repeats and 5 palindromic repeats. In addition, a total of 188 perfect microsatellites were detected, which were characterized as A/T predominance in nucleotide compositions. Significant shifting of the IR/SSC boundaries was revealed by comparing this cp genome with that of other rosids plants. We also built phylogenetic trees to demonstrate the phylogenetic position of *S. suchowensis* in Rosidae, with 66 orthologous protein-coding genes presented in the cp genomes of 32 species. By sequencing 30 amplicons based on the pseudomolecule, experimental verification achieved accuracy up to 99.84% for the cp genome assembly of *S. suchowensis*. In conclusion, this study built a high quality pseudomolecule for the cp genome of *S. suchowensis*, which is a useful resource for facilitating the development of this shrub willow into a more productive bioenergy crop.

Key words: *Salix suchowensis*; chloroplast; genome structure; gene content; phylogenetic tree

Introduction

Chloroplasts (cps) are the plant plastid organelles responsible for photosynthesis (Shinozaki et al., 1986), whose genomes provide essential information for study of the biological processes in plant cells (Raubeson et al., 2005), such as biosynthesis of starch, fatty acids, pigments and amino acids (Neuhaus and Emes, 2000). It is generally accepted that cps have originated from endosymbiosis of cyanobacteria (Timmis et al., 2004). Cp genomes are typically inherited paternally or biparentally in gymnosperms (Reboud and Zeyl, 1994). By contrast, cp genomes are inherited maternally in most angiosperms (Palmer et al., 1988). The chloroplast genomes of angiosperms have a typical quadripartite structure containing a large single copy region (LSC) and a small single copy region (SSC) separated by two inverted repeats regions (IRs), and range from 120 to 160 kb in length with closed circular DNA (Sugiura, 1995). The cp genomes are more conserved in genome structure and organization than the nuclear and mitochondrial genomes (Raubeson et al., 2005). A study by Pyke (1999) revealed that there are approximately 400-1,600 copies of cp genomes in each cell, which led to a high expression level of the cp genes. In recent years, cp transformation has emerged as an environmentally friendly approach for plant genetic engineering (Daniell et al., 2002). Foreign genes in the transformed cps cannot be disseminated by pollen since this plastid is maternal inheritance in most flowering plants, thus posing significantly lower environmental risks. Cp transformation also possesses many other unique advantages over the nuclear transformation, such as permitting the introduction of

thousands of copies of foreign genes per plant cell, allowing the uniformly and extraordinarily high expression levels of foreign genes, and eliminating the gene silence and the 'position effect' (Qian et al., 2013; Daniell, 2007; Verma and Daniell, 2007). With the development of the next generation sequencing technologies, almost 1,078 cp genomes in Viridiplantae have been completely sequenced and deposited at the NCBI Organelle Genome Resources (http://www.ncbi.nlm.nih.gov/genome/organelle/) up to now.

Salix suchowensis is a small and early flowering shrub willow endemic in China, which belongs to subgenus *Vetrix* in genus of *Salix* (Wang, 1984). This willow species mainly distributes in Jiangsu, Shandong, Zhejiang and Henan provinces of China (Fang et al., 1999). Over thousands of years, it has been used as basket-weaving material. Nowadays, this shrub willow is developing into a promising source for bioenergy crops due to its high biomass yield (Smart and Cameron, 2008). The main function of cp is its role in photosynthesis, and biomass yield is highly correlated with the plant photosynthetic efficiency. Therefore, analyzing and characterization of the cp genome of this shrub willow will provide essential information for helping improve productivity and facilitating the establishment of plastid transformation system in this woody crop. In 2014, the whole genome of *S. suchowensis* was sequenced by using a whole-genome shotgun strategy incorporating Roche/454 and Illumina/HiSeq-2000 sequencing technologies, which produced 10.1 Gb 454 GS FLX reads and 230.2 Gb Illumina reads (Dai et al., 2014). Since the sequencing libraries were constructed with leaf tissue, the generated reads include a huge number of sequence reads from the willow cp genome, which provide sufficient sequence information to assemble the cp genome of this shrub willow. In this study, our

perspectives are to assemble and characterize the cp genome of *S. suchowensis* by screening the organelle reads from the willow genome sequencing project; and to experimentally assess the quality of the cp genome assembly derived from the proposed approach.

Materials and Methods

Sequence reads and Cp Genome Assembly

Sequence reads were selected from database generated by the genome sequencing project of *S. suchowensis* as described in Dai et al.'s study (2014). By mapping the raw reads to 660 cp genomes of terrestrial plants in the NCBI Organelle Genome Resources database (http://www.ncbi.nlm.nih.gov/genome/organelle/), we screened the willow cp sequence reads by using BLASTN with an E value of $1e^{-50}$ according to Ma et al.'s description (2016). The obtained reads were further assembled by using software Amos21.0 (De and Mccombie, 2007). Finally, a complete circular cp genome were established by using software Phrap (De and Mccombie, 2007) according to the reference cp genomes of *S. purpurea* (Wu, 2015), *Populus trichocarpa* (Tuskan et al., 2006) and *Arabidopsis thaliana* (Sato et al., 1999). The complete circular cp genome of *S. suchowensis* was deposited at GenBank with accession No. KU341117.

Gene Annotation

Annotation of the *S.suchowensis* cp genome was performed with the online program Dual Organellar GenoMe Annotator (DOGMA, Wyman et al., 2004). Then the start and stop codons and open reading frames (ORF) that may not have been annotated were manually identified by referring to the annotation of *Populus trichocarpa* (Tuskan et al., 2006). In addition, all tRNA genes were predicted with online program tRNAscan-SE v1.21 (Schattner et al., 2007). The

circular chloroplast genome map was generated using the OrganellarGenomeDRAW tool (OGDRAW) (http://ogdraw.mpimp-golm.mpg.de/). In this study, the genome content of Salicaceae species was referred to the previously published annotations on NCBI Organelle Genome Resources database (http://www.ncbi.nlm.nih.gov/genome/organelle/) by using blast-2.3.0 (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.3.0/).

Cp Genome Structure and Sequence Analysis

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

Phylogenetic Analysis

We selected 66 protein-coding gene sequences to explore the phylogenetic relation of *Salix* to 32 rosids lineages with complete cp genomes available. These lineages were from six families of rosids (Salicaceae, Rosaceae, Moraceae, Fagaceae, Chrysobalanaceae and Fabaceae), and *Ginkgo biloba* were used as the outgroup species. The sequences were aligned with ClustalW (Larkin et al., 2007), and a matrix consisting of 83,072 amino acids in default length was obtained. Optimal trees of maximum likelihood (ML) and neighbor joining (NJ) were

constructed by using MEGA6.0 (Tamura et al., 2013). The MP analyses were performed by the Nearest-Neighbor-Interchange (NNI) model under 1,000 bootstrap replicates. The NJ was assessed with the Poisson model under 1,000 bootstrap replicates.

Experimental Assess the Cp Genome Assembly

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

Results and Discussion

Cp Genome Assembly and the Genome Structure

By mapping the raw reads of the *S. suchowensis* genome sequencing project to the NCBI Organelle Genome Resources database (http://www.ncbi.nlm.nih.gov/genome/organelle/), a total of 1,171,821 reads (~533Mb) from willow cp genome were obtained. The sequence depth of the

cp genome would expecte to be more than 3,000×. *De novo* assembly by Amos21.0 (De and Mccombie, 2007) yielded 3,773 contigs. Referring to the cp genomes of *S. purpurea* (Wu, 2015), *Populus trichocarpa* (Tuskan et al., 2006) and *Arabidopsis thaliana* (Sato et al., 1999), these contigs were integrated into a complete circular pseudomolecule in length of 155,508 bp (GenBank accession KU341117). Physical map of the derived cp genome (Figure 1) showed that it possessed a typical quadripartite structure containing a pair of IRs (27,457 bp) separated by LSC (84,385 bp) and SSC (16,209 bp) regions. We compared the cp genomes across nine Salicaceae species. It revealed that, although cp genome were more conserved in genome structure and organization than the nuclear and mitochondrial genomes (Raubeson et al., 2005), the statistics of the cp genomes varied among these closely related species. In these species, the length of IRs, LSC and SSC regions were in range of 27167 bp to 28132 bp, 84377 bp to 85979 bp and 15945 bp to 16600 bp, respectively, with very high sequence similarities.

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

Gene Annotation

Annotation of the cp genome of *S.suchowensis* detected a total of 143 genes. According to gene functions (Yap et al., 2015), they were classified into four categories, including genes associated with self replication, genes associated with photosynthesis, genes associated with other functions, and genes of unknown function. Among these genes, 119 were unique genes, including 4 rRNA genes, 30 tRNA genes, 82 protein-coding genes and 3 pseudogenes. Besides, 4 rRNA genes, 7 tRNA genes and 13 protein-coding genes were found to duplicate in the IR regions. In the unique genes, most of them contain no intron, but one intron was found in six tRNA genes and eight protein-coding genes, and two introns were found in two protein-coding genes (Table 2).

The *ycf1* is one of the longest open reading frames in cp genome and is present in nearly all the plastid genomes sequenced to date (Raubeson et al., 2005). Drescher et al., (2000) predicted that *ycf1* was involved in some essential pathway in cellular metabolism or served some structural function for the plastid compartment. Vries et al. (2015) assumed that *ycf1* encoded the translocon on the inner envelope of chloroplasts (TIC). The function of *ycf1* gene has not been resolved clearly, nevertheless, it is deemed to be essential to plant survival (Drescher et al., 2000). In the sequenced cp genome, the *ycf1* gene usually spans the boundary of the IR and the SSC regions (Raubeson et al., 2005). In accordance with the common location of *ycf1* in the plastid genome, a copy of *ycf1* gene (5,424 bp) was found at the IRA/SSC border (Figure 1), and a truncate copy of *ycf1* pseudogene (1,878 bp) appeared at the IRB/SSC border (Figure 1) in the cp genome of *S. suchowensis*. The *ycf1* gene is highly variable, at approximately 5,500 bp in

plant plastid genome. Compare with chlorophyta species, the length of *S. suchowensis* ycf1 protein (1,807 aa) is much longer than that of *Nephroselmis olivacea* (956 aa; NC_0000927), and much shorter than that of *Schizomeris leibleinii* (3,212 aa; NC_015645).

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

Repeat sequence analysis

Previous studies have shown that gene duplication, gene expansion and chloroplast DNA rearrangement seemed to be associated with repetitive sequences (Cavalier-Smith, 2002). We identified 31 tandem repeats, 16 forward repeats and 5 palindromic repeats in the *S. suchowensis* cp genome (Table S2). The tandem repeat units are in lengths of 7-26 bp, and almost all the tandem repeats locate at the intergenic spacer regions (IGS) except one locating in the intron region. As for the forward repeats, the repeat units are in lengths of 30-76 bp. The majority of these repeats distribute in the IGS region, with some of them detected in the protein-coding regions and the tRNA genes regions. Whereas the palindromic repeats, the repeat units are in length of 30-42 bp. Four repeats were detected in IGS regions and one located at tRNA genes

regions.

Microsatellite or simple sequence repeat (SSR) are composed of short tandem repeats of 1-6 bp nucleotide motifs and they appear commonly in the plant cp genome. MISA output a total of 148 perfect SSRs. Among which, 126 are mononucleotide repeats, 10 are dinucleotide repeats, 11 are tetranucleotide repeats, and one is a pentanucleotide repeat (Table S3). Among the monomers, 121 consist of A/T repeats, and only 5 consist of G/C repeats. The A/T content of monomers is similar with that in the cp genome of *M. glyptostroboides* (96.03%) (Chen et al., 2015). All the dimmers in the cp genome of *S. suchowensis* are AT/TA repeats, and A/T contents in tetramers and pentamers are 86.36% and 80% respectively. Analyzed with the same parameters MISA, the average SSR repeat length and SSR density are found to be lower than those in the cp genome *W. nobilis* (Yap et al., 2015) and *M. glyptostroboides* (Chen et al., 2015).

IRs contraction and expansion

IR regions are prominent features of the cp genomes in most angiosperms. In gymnosperms, they always lack one copy of the IRs (Strauss et al., 1988). Previous studies proposed that the cp genome size might be influenced by IRs contraction and expansion during the evolutionary process of angiosperms (Goulding et al., 1996; Wang et al., 2008). Hereby, we compared the IR regions of four Rosid plants, including *S. suchowensis*, *S. integra*, *Prunus padus* and *Morus notabilis* (Figure 3). It showed that the borders of the IR regions contained the *rpl22* gene or the *rpl22* pseudogene in the cp genomes of *S. suchowensis* and *S. integra*, while the borders of the IR regions contained the *rps19* gene or the *rps19* pseudogene in the cp genomes of *P. padus* and *M. notabilis*. IR junctions between LSC and SSC showed remarkable changes, in detail, the IRb

region extended 52 bp into the *rpl22* gene in the *S. suchowensis* and *S. integra* cp genomes, which created a short *rpl22* pseudogene of 52 bp at the IRa/LSC border. The IRb region only extended 39 bp into the *rps19* gene in *P. padus*, which created a short *rps19* pseudogene of 39 bp at the IRa/LSC border. As for the *M. notabilis*, the IRb region was found to be immediately adjacent to the *rps19* gene. In addition, the IRb/SSC border extended into the *ycf1* genes to create truncated *ycf1* pseudogenes in *S. integra* and *M. notabilis* cp genomes. It was commonly found that contraction and expansion of IRs could create pseudogenes that cannot be transcripted (Wang et al., 2008). In the cp genomes of most land plants, there are always a *ycf1* pseudogene located in the LSC/IR border, e.g. in *P. trichocarpa* (Tuskan et al., 2006), *M. glyptostroboides* (Chen et al., 2015), and *S. miltiorrhiza* (Qian et al., 2013). We also detected a *ycf1* pseudogene in the LSC/IR border of the cp genome assembly for *S. suchowensis*.

Phylogenetic trees

To gain the phylogenetic position of *S. suchowensis* in Rosids, we selected 66 orthologous protein-coding genes presented in the cp genomes of 32 species (Table S4). The ML bootstrap analysis resolved into 29 nodes, of which 25 nodes had bootstrap values ≧90 % and 18 of these had bootstrap support of 100 % (Figure 4). With the NJ tree we obtained the sum of branch length of 0.61439509. The NJ bootstrap analysis was same to the ML tree that resolved into 29 nodes, of which 24 nodes had bootstrap values ≧90 % and 20 of these had bootstrap support of 100 % (Figure S1). Both the ML and NJ trees showed that these species were evident into three categories of Rosids Ⅰ (Salicaceae and Chrysobalanaceae), Rosids Ⅱ (Fagaceae, Moraceae and Rosaceae), and Rosids Ⅲ (Fabaceae). In the Rosids Ⅰ, *S. suchowensis* and *S. babylonica* were

the closest relatives. The topological orders in the derived ML tree is very similar with that of the established NJ tree, the only incongruence between them is the position of *P. fremontii* and *P. balsamifera* in relating to *P. trichocarpa*. It is noteworthy that the bootstrap supports for grouping *P. fremontii* or *P. balsamifera* with *P. trichocarpa* are relatively lower on both the ML and NJ trees. Relationship of these three species might not be resolved properly merely based on information at plastid level.

Assess the cp genome assembly of *S. suchowensis*

In this study, the raw reads were generated by the next generation sequencing platforms, and the screening of reads and the assembly of the cp genome were conducted merely based on bioinformatics tools. To assess the quality of the assembly, 30 sites were sequentially selected from the cp genome. All the synthesized primers succeed in PCR amplification. Sequenced by a Sanger sequencer, the 30 amplicons covers a total physical length of 18,639 bp. Align the amplicon sequences to the genome assembly, sequence errors were found in seven of the tested sites, while 100% match were revealed with amplicons at the other 23 sites in the cp genome assembly. The overall accuracy of the derived assembly was estimated to be 99.84%. Therefore, the cp genome obtained in this study is in high quality. As aforementioned, we detected raw reads that covered over 3000×sequence depth of the cp genome of *S. suchowensis*. The high sequence depth ensures the accuracy and integrity of the obtained pseudomolecule of the cp plastid. In conclusion, we derive a highly reliable pseudomolecule of the cp genome for *S. suchowensis* based on raw reads generated by the next generation sequencing platforms, which is highly desirable for facilitating the biological study of this promising biofuel plant.

Acknowledgments

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

References

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

Cavalier-Smith T. 2002. Chloroplast evolution: secondary symbiogenesis and multiple losses. *Current Biology Cb* **12(2):** 62-64.

Chen J, Hao Z, Xu H, Yang L, Liu G, Yu S, Chen Z, Weiwei Z, Tielong C, Jisen S. 2015. The complete chloroplast genome sequence of the relict woody plant *Metasequoia glyptostroboides* hu et cheng. *Frontiers in Plant Science* **6:** 447.

Dai X, Hu Q, Cai Q, Feng K, Ye N, Tuskan GA, Luo W. 2014. The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Research* **24(10):**1274-1277.

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

Daniell H. 2007. Transgene containment by maternal inheritance: effective or elusive?. *Proceedings of the National Academy of Sciences* **104(17):** 6879-6880.

De lBM, Mccombie WR. 2007. Assembling genomic dna sequences with phrap. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis* **Chapter 11:** Unit11.4-Unit11.4.

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

Fang Z, Zhao S, Skvortsov AK. 1999. Saliceae. In W. Zheng-yi and P.H. Raven (editors). *Flora of China*. St. Louis USA: Missouri Botanical Garden Press 139-274.

Goulding SE, Wolfe KH, Olmstead RG, Morden CW. 1996. Ebb and flow of the chloroplast inverted repeat. *Molecular and General Genetics MGG* **252(1-2):**195-206.

Huang CY, Ayliffe MA, Timmis JN. 2003. Direct measurement of the transfer rate of chloroplast dna into the nucleus. *Nature* **422(6927):** 72-6.

Lalitha S. 2000. Primer premier 5. *Biotech Software and Internet Report* **1(6):** 270-272.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal w and clustal x version 2.0. *Bioinformatics* **23(21):** 2947-2948.

Ma Q, Li S, Bi C, Hao Z, Sun C, Ning Y. 2016. Complete chloroplast genome sequence of a major economic species, *Ziziphus jujuba* (Rhamnaceae). *Current Genetics* 1-13.

Mcpherson H, Merwe MVD, Delaney SK, Edwards MA, Henry RJ, Mcintosh E, Rymer PD, Milner ML, Siow J, Rossetto M. 2013. Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. *BMC Ecology* **13(1):** 53-65.

Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Julian MH, Gray JC, Morden CW, Calle PJ. 2001. Many parallel losses of infa from chloroplast dna during angiosperm evolution with multiple independent transfers to the nucleus. *Plant Cell* **13(3):** 645-658.

Neuhaus HE, Emes MJ. 2000. Nonphotosynthetic metabolism in plastids. *Annual Review of Plant Biology* **51(4):**111-140.

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

Palmer JD, Jansen RK, Michaels HJ, Chase MW, Manhart JR. 1988. Chloroplast DNA variation and plant phylogeny. *Annals of the Missouri Botanical Garden* **75(4):**1180-1206.

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

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

Raubeson LA, Jansen RK, Henry RJ. 2005. Chloroplast genomes of plants. *Plant Diversity and Evolution* 45.

Reboud X and Zeyl C. 1994. Organelle inheritance in plants. *Heredity* **72(2):**132-140.

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

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

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

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayshida N, Matsubayasha T, Zaita N, Chunwongse J, Obakata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugirua M. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *The EMBO Journal* **5(9):** 2043-2049.

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

Stegemann S, Bock R. 2006. Experimental reconstruction of functional gene transfer from the tobacco plastid genome to the nucleus. *Plant Cell* **18(11):** 2869-78.

Strauss SH, Palmer JD, Howe GT, Doerksen AH. 1988. Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. *Proceedings of the National Academy of Sciences* **85(11):** 3898-3902.

Sugiura M. 1995. The chloroplast genome. *Essays in Biochemistry* **30(1):** 49-57.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. Mega6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* **30(12):** 2725-2729.

Timmis JN, Ayliffe, MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics* **5(2):**123-35.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Déjardin A, DePamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé JC, Locascio P, LouY, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, PeterG, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson R, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de PeerY, Rokhsar D. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). **Science 313(5793):**1596-1604.

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

Vries JD, Sousa FL, Bölter B, Soll J, Gould SB. 2015. Ycf1: a green tic?. *Plant Cell* **27(7):** 1827-1833.

Wang C, Fang CF, Zhao SD. 1984. Salicaceae. *Flora Reipublicae Popularis Sinicae* **20(2):** 79-

403.

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

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

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

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

Yap JY, Rohner T, Greenfield A, Van DMM, Mcpherson H, Glenn W, Kornfeld G, Marendy E, Annie YHP, Wilton A, Wilkins MR, Rossetto M, Delaney SK. 2015. Complete chloroplast genome of the wollemi pine (*Wollemia nobilis*): structure and evolution. *PLOS ONE* **10(6):** e0128126.

Tables

Table 1 Comparison the statistics of cp genomes across nine Salicaceae species

Species	IR (bp)	SSC (bp)	LSC (bp)	GC content (%)	Number of genes
Populus alba	27 660	16 567	84 618	36.74	109+1
P. balsamifera	27 836	16 499	84 921	36.65	109+3
P. fremontii	27 838	16 316	85 454	36.67	106+3
P. tremula	27 600	16 490	84 377	36.76	111
P. trichocarpa	27 652	16 600	85 129	36.68	119+1
Salix babylonica	27 646	16 273	85 255	36.65	109+1
S. interior	27 167	16 307	85 979	37.00	106+2
S. purpurea	27 459	16 220	84 455	36.69	110+1
S. suchowensis	27 457	16 209	84 385	36.73	116+3

Note: The digital following the "+" is the number of pseudogenes.

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

Category for genes	Group of genes	Name of genes
Self replication	Transfer RNA genes	30 tRNAs (6 contain an intron)
	Ribosomal RNA genes	*rrn4.5, rrn5, rrn16, rrn23*
	DNA dependent RNA polymerase	*rpoA, rpoB, rpoC1*, rpoC2*
	Small subunit of ribosome	*rps2, rps3, rps4, rps7, rps8, rps11, rps12**, rps14, rps15, rps18, rps19*
	Large subunit of ribosome	*rpl2*, rpl14, rpl16*, rpl20, rpl22, rpl23, rpl33, rpl36*
Large subunit of ribosome	Subunits of photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Subunits of photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
	Subunits of cytochrome	*petA, petB*, petD*, petG, petL, petN*
	Subunits of ATP synthase	*atpA, atpB, atpE, atpF*, atpH, atpI*
	Subunits of NADH dehydrogenase	*ndhA*, ndhB*, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Large subunit of Rubisco	*rbcL*

	ATP-dependent protease subunit p gene	*clpP***
Other genes	Subunit of acetyl-CoA-carboxylase	*accD*
	c-type cytochrome synthesis gene	*ccsA*
	Envelop membrane protein	*cemA*
	Maturase	*matK*
Genes of unknown function	pseudogene	*Pseudo-ycf68,Pseudo-ycf1， Pseudo-infA*
	Conserved open reading frames	*ycf1, ycf2, ycf3**, ycf4, ycf15,cp001,cp002,cp003,cp004,cp005*

Note: * : contains one intron; ** : contains two introns.

Figure Legends

Figure 1 Physical map of Salix suchowensis complete chloroplast genome

Genes shown outside the circle are transcribed counterclockwise and genes shown inside the circle are transcribed clockwise. Genes in the same color represent the same functional groups. Internal circle of darker gray and lighter gray indicate GC content and AT content, respectively.

Figure 2 Sequence alignment of *infA* from cp genome and that from nuclear genome

a: cp genome of *S. suchowensis*; b: nuclear genome of *S. suchowensis*.

Figure 3 Comparison of the borders of IR regions among the cp genomes of four Rosids plant

Three different colors were used to indicate the LSC, IR and SSC regions, respectively. The figure mainly indicates the shift of the genes located in the IR border. “ψ” means pseudogene, “overlap” means overlap of ψycf1 and ndhF gene.

Figure 4 The Maximum Likelihood (ML) phylogenetic tree

422

423 **Supplemental information files**

424 Table S1 Primer sequences and results of sequences alignment of the amplicons

425 Table S2 Statistics of tandem repeats, forward repeats and palindromic repeats in *S. suchowensis*

426 chloroplast genome

427 Table S3 Distribution of SSRs in the *S. suchowensis* chloroplast genome

428 Table S4 Species classification in phylogenetic tree and GenBank accession numbers of the cp

429 genome

430 Figure S1 The neighbor joining (NJ) phylogenetic tree

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2388v1 | CC BY 4.0 Open Access | rec: 24 Aug 2016, publ: 24 Aug 2016

Figure 1(on next page)

Figure 1

Figure 1 Physical map of Salix suchowensis complete chloroplast genome<?xml:namespace prefix = o ns = "urn:schemas-microsoft-com:office:office" />

Genes shown outside the circle are transcribed counterclockwise and genes shown inside the circle are transcribed clockwise. Genes in the same color represent the same functional groups. Internal circle of darker gray and lighter gray indicate GC content and AT content, respectively.

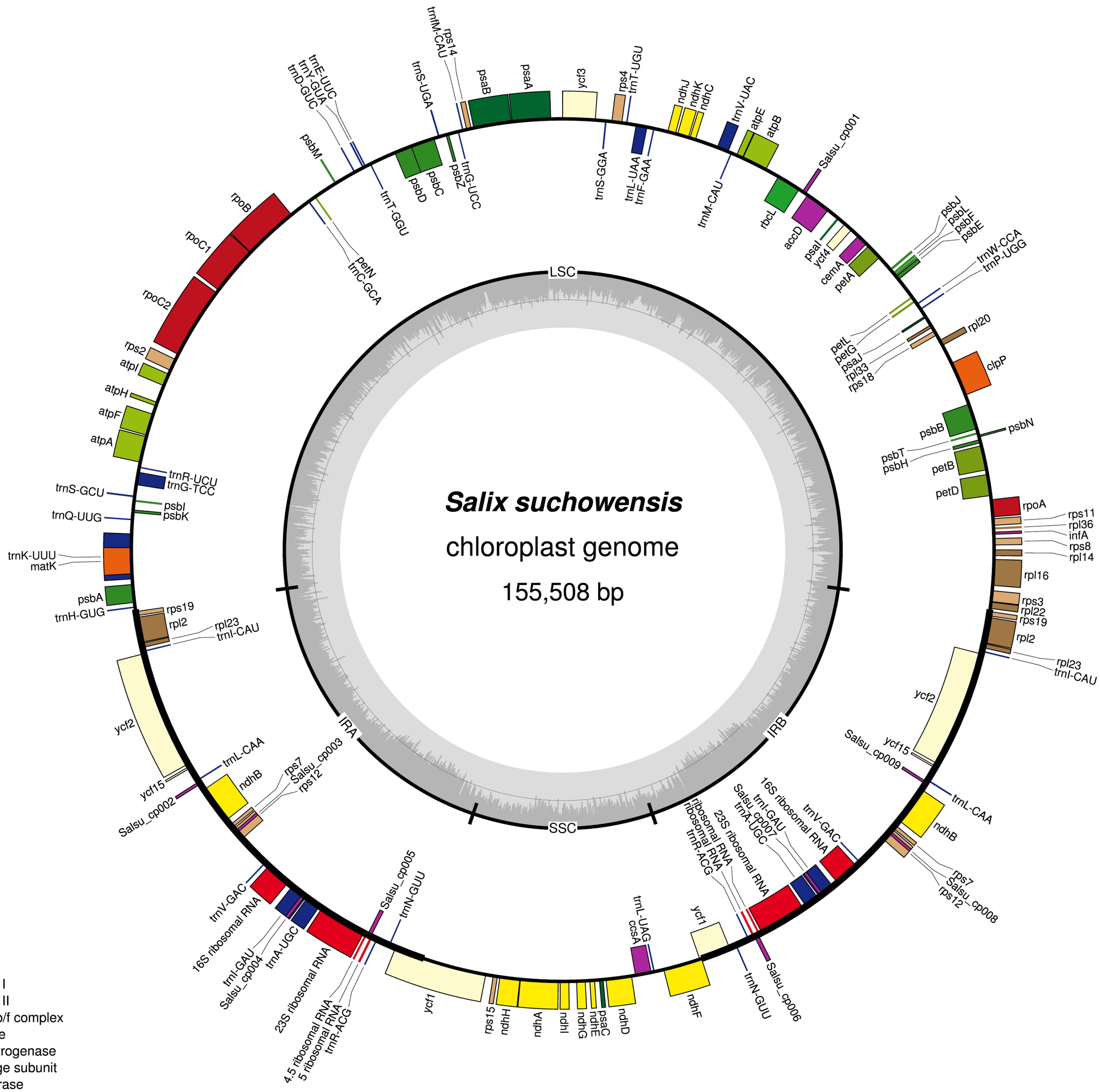
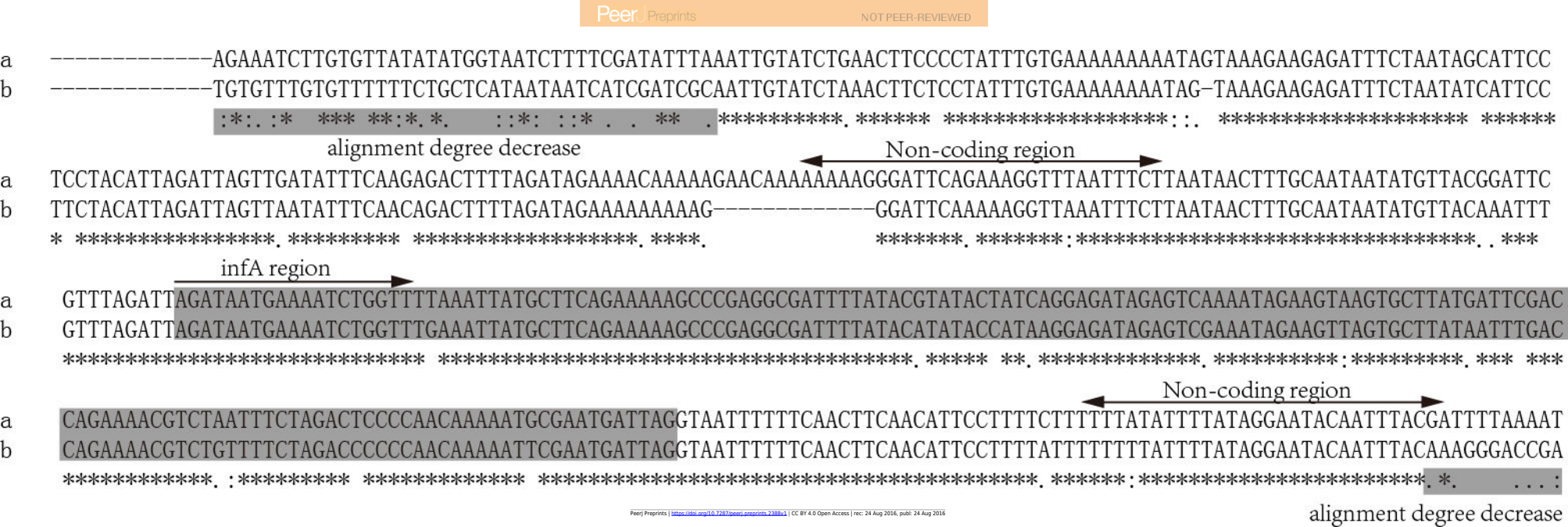


Figure 2(on next page)

Figure 2

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

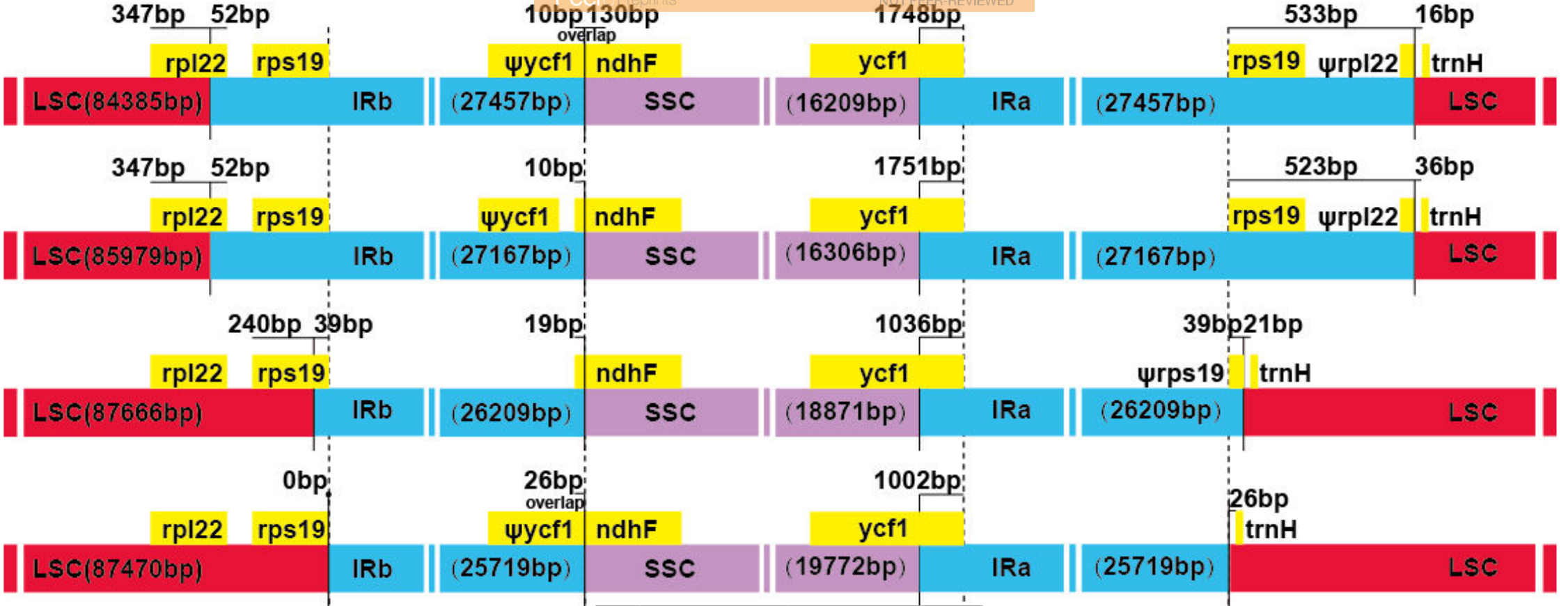


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2388v1 | CC BY 4.0 Open Access | rec: 24 Aug 2016, publ: 24 Aug 2016

Figure 3(on next page)

Figure 3

Figure 3 Comparison of the borders of IR regions among the cp genomes of four Rosids plant Three different colors were used to indicate the LSC, IR and SSC regions, respectively. The figure mainly indicates the shift of the genes located in the IR border. “ψ” means pseudogene, “overlap” means overlap of ψycf1 and ndhF gene.

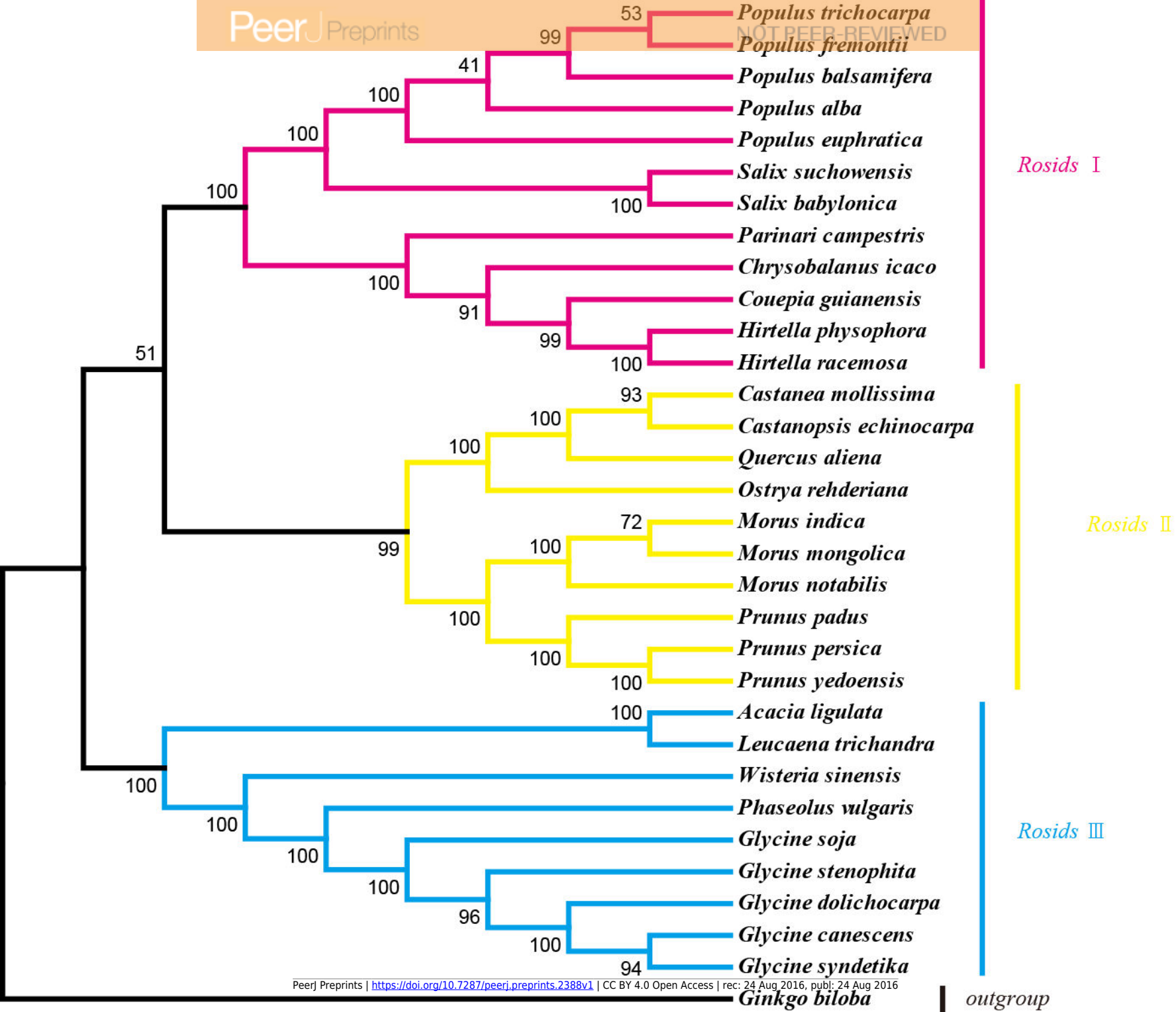


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2388v1 | CC BY 4.0 Open Access | rec: 24 Aug 2016, publ: 24 Aug 2016

Figure 4(on next page)

Figure 4

Figure 4 The Maximum Likelihood (ML) phylogenetic tree<?xml:namespace prefix = o ns = "urn:schemas-microsoft-com:office:office" />



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2388v1 | CC BY 4.0 Open Access | rec: 24 Aug 2016, publ: 24 Aug 2016