PLOS ONE

RESEARCH ARTICLE

Complete chloroplast genome sequence and comparative analysis of loblolly pine (*Pinus taeda* L.) with related species

Sajjad Asaf[1], Abdul Latif Khan[1], Muhammad Aaqil Khan[2], Raheem Shahzad[2], Lubna[3], Sang Mo Kang[2], Ahmed Al-Harrasi[1], Ahmed Al-Rawahi[1], In-Jung Lee[2,4]*

1 Chair of Oman's Medicinal Plants & Marine Natural Products, University of Nizwa, Nizwa, Oman, **2** School of Applied Biosciences, Kyungpook National University, Daegu, Republic of Korea, **3** Department of Botany, Garden Campus, Abdul Wali Khan University Mardan, Mardan, Pakistan, **4** Research Institute for Dok-do and Ulleung-do Island, Kyungpook National University, Daegu, Republic of Korea

* ijlee@knu.ac.kr

Citation: Asaf S, Khan AL, Khan MA, Shahzad R, Lubna , Kang SM, et al. (2018) Complete chloroplast genome sequence and comparative analysis of loblolly pine (*Pinus taeda* L.) with related species. PLoS ONE 13(3): e0192966. https://doi.org/10.1371/journal.pone.0192966

Editor: Hikmet Budak, Montana State University Bozeman, UNITED STATES

Received: May 30, 2017

Accepted: February 1, 2018

Published: March 29, 2018

Data Availability Statement: Data are available from GenBank with accession number KY964286.

Funding: This research was supported by a Basic Science Research Program grant to Prof. In-Jung Lee through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A1B04035601).

Competing interests: The authors have declared that no competing interests exist.

Abstract

Pinaceae, the largest family of conifers, has a diversified organization of chloroplast (cp) genomes with two typical highly reduced inverted repeats (IRs). In the current study, we determined the complete sequence of the cp genome of an economically and ecologically important conifer tree, the loblolly pine (*Pinus taeda* L.), using Illumina paired-end sequencing and compared the sequence with those of other pine species. The results revealed a genome size of 121,531 base pairs (bp) containing a pair of 830-bp IR regions, distinguished by a small single copy (42,258 bp) and large single copy (77,614 bp) region. The chloroplast genome of *P. taeda* encodes 120 genes, comprising 81 protein-coding genes, four ribosomal RNA genes, and 35 tRNA genes, with 151 randomly distributed microsatellites. Approximately 6 palindromic, 34 forward, and 22 tandem repeats were found in the *P. taeda* cp genome. Whole cp genome comparison with those of other *Pinus* species exhibited an overall high degree of sequence similarity, with some divergence in intergenic spacers. Higher and lower numbers of indels and single-nucleotide polymorphism substitutions were observed relative to *P. contorta* and *P. monophylla*, respectively. Phylogenomic analyses based on the complete genome sequence revealed that 60 shared genes generated trees with the same topologies, and *P. taeda* was closely related to *P. contorta* in the subgenus *Pinus*. Thus, the complete *P. taeda* genome provided valuable resources for population and evolutionary studies of gymnosperms and can be used to identify related species.

Introduction

Gymnosperms are represented by a diverse and magnificent group of coniferous species distributed across eight families, consisting of 70 genera containing more than 630 species [1]. They are thought to have arisen from seed plants approximately 300 million years ago and are one of the ancient main plant clades. Gymnosperms possess larger genomes than flowering plants [2–5]. Recently, rapid progress has been made in angiosperm genome sequencing and

analysis, but because of the complexity and order of magnitude increase in genome sizes, similar progress has not been attained for gymnosperms. Furthermore, comparative studies revealed that transposable elements, repetitive sequences, and gene duplication are common in gymnosperm genomes [4, 6–8]. Conifers are the main representatives of the gymnosperms, predominant in various ecosystems and representing 82% of terrestrial biomass [9].

Pinus taeda (loblolly pine) is a model species for the largest genus in the division Coniferae. It is an economically important and relatively fast-growing representative of conifers native to the southeastern United States. Previously, the loblolly pine was famous for providing pulp, lumber, and paper to commercial markets, but recently became a main bioenergy feedstock in lignocellulosic ethanol production [10]. Moreover, loblolly pine is considered an important species for comparative genomic studies between angiosperms and gymnosperms [8]. For example, microsatellites and single-nucleotide polymorphisms (SNPs) have been studied to determine population genetic parameters and the associations of phenotypes [11–13], create genetic maps [14–16], and develop genomic selection prediction models [17]. However, the number of available genetic markers remains small, particularly considering the large size of the pine genome. According to recent evaluations [18], the loblolly pine nuclear genome size is 21–24 Gbp. This is approximately four-fold larger than that of the angiosperm with the largest genome, *Hordeum vulgare* (barley), for which a reference genome is available, and approximately 7–8-fold larger than the human genome [19].

Chloroplasts are known to be derived from cyanobacterium through endosymbiosis and co-evaluation over time [20]. The gymnosperm chloroplast (cp) genome, particularly in conifers, has distinguishing characteristics among angiosperms. These features such as the high levels of variation (intra-specific) [21–24], paternal inheritance [25–28], and a different RNA editing pattern [29] were observed in studies. Generally, in angiosperms, cp genomes range from 130,000 to 160,000 base pairs (bp), with two duplicate inverted repeats (IRs) containing large single copy (LSC) and small single copy (SSC) regions. However, the comparative sizes of IRs, SSC, and LSC, are nearly unchanged, while the gene order and content are significantly conserved [30]. In contrast, the IR sizes of species form gymnosperms highly fluctuate among taxa [31–33]. Similarly, previous reports showed that the IR size for *Cycas taitungensis* is 23 kbp [34] and *Ginkgo biloba* is 17 kbp [35]. In contrast, *P. thunbergii* has a very small IR of 495 bp [36, 37]. Furthermore, in synergism with *P. thunbergii*, various conifer species have been found to lack the comparatively large IRs typically found in gymnosperms [31, 33, 38, 39]. This decrease in IR size is thought to cause extensive rearrangement in conifer cp genomes [33]. Based on the IRs, the cp genomes can be classified into three categories: (i) with two IRs, (ii) with one IRs, and (iii) with additional tandem repeats [30]. The cp genomes are essential and extremely valuable for understanding the phylogenetic relationships and designing specific molecular markers because of their firm mode of inheritance. Using a total evidence approach [40], the cp genomes or various concatenated sequences were studied to elucidate the phylogeny among various species [41–43]. Similarly, Steane [44] showed that the organization of the *P. thunbergii* cp genome differs from that of other related angiosperms.

The advent of high-throughput next-generation sequencing technologies from Illumina, Pacific Biosciences, Life Technologies, and Roche, among others, have rapidly improved genomic studies [45, 46]. In addition to draft or whole genomes of microbes and animals, genomic studies were performed to determine the chromosomal structures and molecular organization of wheat [47, 48] and maize [49]. In addition, these technologies have been extensively used to evaluate organelles, particularly chloroplast. Although the first complete nucleotide sequence of *Nicotiana tabacum* was generated by clone sequencing of plasmid and cosmid libraries over a long time [50], more than 800 cp genomes (including 300 from crops and trees) have now

been sequenced and deposited in the NCBI Organelle Genome Resources database [51]. The evolution of cp genomes in terrestrial plants can now be studied using these database resources [51]. To date, a total of 16 complete chloroplast genomes in the genus *Pinus* have been sequenced and submitted to NCBI. In the current study, the complete cp genome of *P. taeda* (GenBank accession number: KY964286) was sequenced using next-generation sequencing tools. The goal of this study was to determine the cp genome organization of *P. taeda* and its global pattern of structural and comparative variation in the cp genome of *P. taeda* with 14 *Pinus* species (*P. koraiensis*, *P. sibirica*, *P. armandii*, *P. lambertiana*, *P. krempfii*, *P. bungeana*, *P. gerardiana*, *P. monophylla*, *P. nelsonii*, *P. contorta*, *P. massoniana*, *P. tabuliformis*, *P. taiwanensis*, *P. strobus*, and *P. thunbergii*).

Materials and methods

Chloroplast genome sequencing and assembly

Plastid DNA was extracted from the fresh needle leaf parts of *P. taeda* using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany), and the resulting cpDNA was sequenced using an Illumina HiSeq-2000 platform (San Diego, CA, USA) at Macrogen (Seoul, Korea). The *P. taeda* cp genome was then assembled *de novo* using a bioinformatics pipeline (http://www.phyzen.com). Specifically, a 400-bp paired-end library was produced according to the Illumina standard method, which generated 28,110,596 bp of sequence data with a 100-bp average read length. Raw reads with Phred scores of $\leq$20 were removed from the total PE reads using the CLC-quality trim tool, and *de novo* assembly of trimmed reads was accomplished using CLC Genomics Workbench v7.0 (CLC Bio, Aarhus, Denmark) with a minimum overlap of 200–600 bp. The resulting contigs were compared against the *P. thunbergii* and *P. contorta* plastomes using BLASTN with an E-value cutoff of 1e-5, and five contigs were identified and temporarily arranged based on their mapping positions on the reference genome. After initial assembly, primers were designed (S1 Table) based on the terminal sequences of adjacent contigs, and PCR amplification and subsequent DNA sequencing were conducted to fill in the gaps. PCR amplification was performed in 20-μL reactions containing 1× reaction buffer, 0.4 μL dNTPs (10 mM), 0.1 μL Taq (Solg h-Taq DNA Polymerase), 1 μL (10 pm/μL) primers, and 1 μL (10 ng/μL) DNA, using the following conditions: initial denaturation at 95˚C for 5 min; 32 cycles of 95˚C for 30 s, 60˚C for 20 s, and 72˚C for 30 s; and a final extension step of 72˚C for 5 min. After incorporating the additional sequencing results, the complete cp genome was used as a reference to map the remaining unmapped short reads to improve the sequence coverage of the assembled genome.

Analysis of gene content and sequence architecture

The *P. taeda* cp genome was annotated using DOGMA [52], checked manually, and the codon positions were adjusted by comparison with homologs in the cp genome of *P. taeda* and *P. contorta*. Transfer RNA sequences of the *P. taeda* cp genome were verified using tRNAscan-SE version 1.21 [53] with default settings, and the structural features were illustrated using OGDRAW [54]. To examine deviations in synonymous codon usage by avoiding the influence of amino acid composition, the relative synonymous codon usage was determined using MEGA 6 software [55], and finally the divergence of the *P. taeda* cp genome from six other *Pinus* species (five from subgenus *Pinus* and one from subgenus *Strobus*) cp genomes was assessed using mVISTA [56] in Shuffle-LAGAN mode and using the *P. taeda* genome as a reference.

Elucidation of repeat sequences and simple sequence repeat (SSRs)

Repeat sequences, including direct, reverse, and palindromic repeats, were identified within the cp genome using REPuter [57] with the following settings: Hamming distance of 3, ≥90% sequence identity, and minimum repeat size of 30 bp. Furthermore, SSRs were detected using Phobos version 3.3.12 [58] with the search parameters set to ≥10 repeat units for mononucleotide repeats, ≥8 repeat units for dinucleotide repeats, ≥4 repeat units for trinucleotide and tetranucleotide repeats, and ≥3 repeat units for pentanucleotide and hexanucleotide repeats. Tandem repeats were identified using Tandem Repeats Finder version 4.07 b [59] with default settings.

Sequence divergence and phylogenetic analyses

The average pairwise sequence divergence of 60 shared genes and complete plastomes of 15 *Pinus* species was analyzed, using data from *P. taeda*, *P. koraiensis*, *P. sibirica*, *P. armandii*, *P. lambertiana*, *P. krempfii*, *P. bungeana*, *P. gerardiana*, *P. monophylla*, *P. nelsonii*, *P. contorta*, *P. massoniana*, *P. tabuliformis*, *P. taiwanensis*, *P. strobus*, and *P. thunbergii*. In cases of missed and unclear genes, annotation was confirmed by comparison with the reference sequence after assembling a multiple sequence alignment tool. The complete genome data set was aligned using MAFFT version 7.222 [60] with default parameters. For pairwise sequence divergence, a Kimura's model was used [61]. Indel polymorphisms among the complete genomes were identified using DnaSP 5.10.01 [62], and a custom Python script (https://www.biostars.org/p/119214/) was used to identify SNPs. To resolve the phylogenetic position of *P. taeda* within the genus *Pinus*, 14 published *Pinus* species plastomes were downloaded from the NCBI database for phylogenetic analysis. Multiple alignments of the complete plastomes were constructed based on the conserved structure and gene order of the plastid genomes [63], and four methods were employed to construct phylogenetic trees, including Bayesian inference (BI), which was implemented using MrBayes 3.1.2 [64], maximum parsimony (MP), which was implemented using PAUP 4.0 [65], and maximum likelihood (ML) and neighbor-joining (NJ), which were implemented using MEGA 6 [55] using previously described settings [66, 67]. In a second phylogenetic analysis, 60 shared cp genes from 15 *Pinus* species, including *P. taeda*, and one outgroup species (*Juniperus bermudiana*) were aligned using ClustalX with default settings, followed by manual adjustment to preserve the reading frames. Finally, the same four phylogenetic inference methods were used to infer trees from the 60 concatenated genes using the same settings [66, 67].

Results and discussion

The *P. taeda* cp genome was assembled by mapping all Illumina sequence reads into a draft cp genome. Approximately 2,5131,617 reads with 100-bp average lengths were retrieved to obtain 1619.4X coverage of the cp genome. The complete cp genome of *P. taeda* was 121,131 bp, with 38.5% GC content and only one bp less than the previously sequenced *P. taeda* cp genome (Table 1). The cp genome size of *P. taeda* was within the expected range (116–121 Kb) of other sequenced cp genomes of Pinaceae members [41, 68, 69]. The *P. taeda* cp genome was circular and contained two short-inverted repeats (IRa and IRb) of 830 bp, divided into SSC (42,258 bp) and LSC (77,614 bp) (Fig 1). The *P. taeda* cp genome encodes 120 genes, including 81 protein-coding genes, four ribosomal RNA (rRNA) genes, and 35 tRNA genes (Table 2). Of these genes, 11 genes (*atpF*, *petB*, *petD*, *rpoC1*, *rpl2*, *rpl16*, *trnI-GAU*, *trnG-UCC*, *trnA-UGC*, *trnV-UAC*, and *trnL-UAA*) contained one intron and two genes (*rps12* and *ycf3*) harbored two introns (Table 3). Furthermore, *trnK-UUU* was identified as the gene containing the longest intron (3,307 bp), which included *matK* (Table 3); similarly, *rps12* was recognized as a trans-

Table 1. Summary of complete chloroplast genomes for 15 *Pinus* species.

	P. tae	*P.tae**	*P.arm*	*P. bung*	*P. cont*	*P. gerar*	*P. kor*	*P. krem*	*P. lamb*	*P. mass*	*P. mono*	*P. nel*	*P. sib*	*P. tab*	*P. taiw*	*P. stro*	*P. thu*
Size (bp)	121,531	121,530	117,265	117,861	120,438	117,618	117,190	116,989	117,239	119,739	116,479	116,834	116,635	119,646	119,741	115,576	119,707
Overall GC contents	38.5	38.5	38.8	38.1	38.4	38.7	38.8	38.7	38.7	38.5	38.6	-	38.7	38.5	38.5	38.8	38.5
LSC size in bp	77,614	77,615	64,548	65,373	59,591	-	64,523	-	64,750	51,458	74,357	-	64,080	75,628	65,670	74,634	65,696
SSC size in bp	42,258	42,532	51,767	51,538	60,131	-	51,717	-	51,715	43,197	41,691	-	51,782	42,329	53,080	40,310	53,020
IR size in bp	830	693	475	475	358	-	475	-	387	378	431	-	387	845	409	467	495
Protein coding regions size in bp	61,691	60,765	61,227	60,702	58,469	60,364	60,496	59,753	60,847	60,519	60,015	69,598	62,988	60,549	65,133	53,919	70,395
tRNA size in bp	2,661	2,587	2,778	2,725	2,582	2,583	2,778	2,428	2,511	2,725	2,577	2,575	2,131	2,725	2,785	2,657	2,652
rRNA size in bp	4,517	4,517	4,555	4,515	4,517	4,515	4,555	4,514	4,515	4,515	4,515	4,515	4,555	4,518	4,518	4,516	4,518
Number of genes	122	111	115	113	110	110	110	108	110	109	111	111	113	116	137	111	171
Number of protein coding genes	83	71	74	71	70	70	70	69	71	73	70	70	81	74	92	70	123
Number of rRNA	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Number of tRNA	35	34	36	36	34	34	36	32	33	36	34	34	28	36	36	35	35
Genes duplicated in IR	3	2	2	2	1		4		1	1	1		1	3	1	1	2
Genes with introns	13	13	13	14	13	13	15	13	13	15	13	13	13	14	13	13	15

P. tae = *P. taeda*; **P. tae*** = *P. taeda* (old); **P.arm** = *P. armandii*; **P. bung** = *P. bungeana*; **P. cont** = *P. contorta*; **P. gerar** = *P. gerardiana*; **P. kor** = *P. koraiensis*; **P. krem** = *P. krempfii*; **P. lamb** = *P. lambertiana*; **P. mass** = *P. massoniana*; **P. mono** = *P. monophylla*; **P. nel** = *P. nelsonii*; **P. sib** = *P. sibirica*; **P. tab** = *P. tabuliformis*; **P. taiw** = *P. taiwanensis*; **P. stro** = *P. strobus*; **P. thu** = *P. thunbergii*

https://doi.org/10.1371/journal.pone.0192966.t001

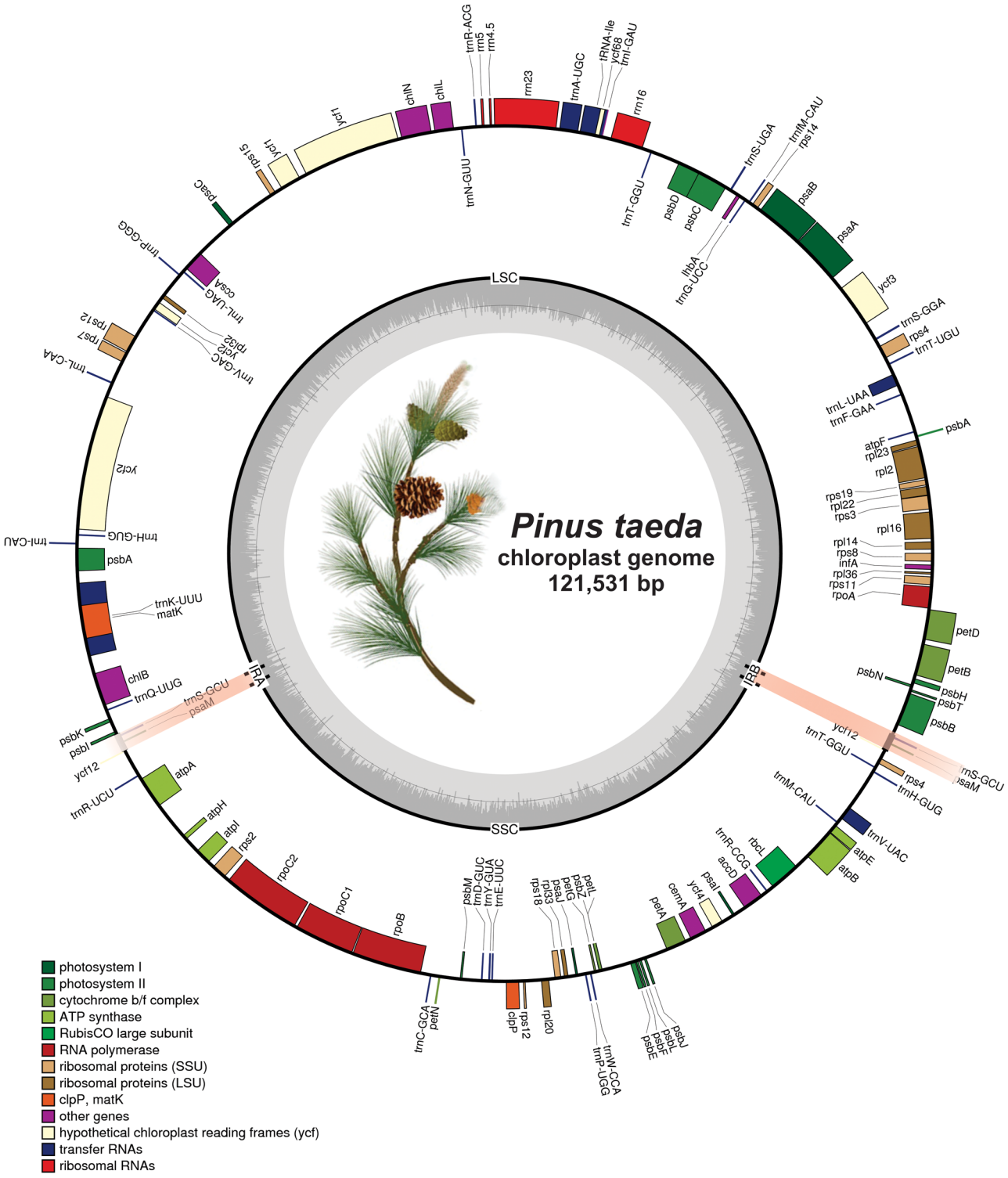

Fig 1. Gene map of the *Pinus taeda* plastid genome. Thick lines in the red area indicate the extent of the inverted repeat regions (IRa and IRb; 850 bp), which separate the genome into small (SSC; 42,258 bp) and large (LSC; 77,614 bp) single copy regions. Genes drawn inside the circle are transcribed clockwise, and those outside are transcribed counter clockwise. Genes belonging to different functional groups are color-coded. The dark grey in the inner circle corresponds to the GC content and the light grey corresponds to the AT content.

https://doi.org/10.1371/journal.pone.0192966.g001

Table 2. Genes in the sequenced *P. taeda* chloroplast genome.

Category	Group of genes	Name of genes
Self-replication	Large subunit of ribosomal proteins	*rpl2, 14, 16, 20, 22, 23, 32, 33, 36*
	Small subunit of ribosomal proteins	*rps2, 3, 4, 7, 8, 11, 12, 14, 15, 18, 19*
	DNA-dependent RNA polymerase	*rpoA, B, C1, C2*
	rRNA genes	*RNA*
	tRNA genes	*trnA-UGC, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnfM-CAU, trnG-UCC, trnH-GUG, trnI-CAU, trnI-GAU, trnK-UUU, trnL-CAA, trnL-UAA, trnL-UAG, trnM-CAU, trnN-GUU, trnP-GGG, trnP-UGG, trnQ-UUG, trnR-ACG, trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC, trnV-UAC, trnW-CCA, trnY-GUA*
Photosynthesis	Photosystem I	*psaA, B, C, I, J, M*
	Photosystem II	*psbA, B, C, D, E, F, H, I, J, K, L, M, N, T, Z*
	Cytochrome b6/f complex	*petA, B, D, G, L, N*
	ATP synthase	*atpA, B, E, F, H, I*
	Rubisco	*rbcL*
	Chlorophyll biosynthesis	*chlB, L, N*
Other genes	Maturase	*matK*
	Protease	*clpP*
	Envelop membrane protein	*cemA*
	Subunit acetyl-CoA-carboxylate	*accD*
	c-Type cytochrome synthesis gene	*ccsA*
Unknown	Conserved open reading frames	*ycf1, 2, 3, 4, 12, 68*

https://doi.org/10.1371/journal.pone.0192966.t002

Table 3. Genes with introns in the *Pinus taeda* chloroplast genome and length of exons and introns.

Gene	Location	Exon I (bp)	Intron 1 (bp)	Exon II (bp)	Intron II (bp)	Exon III (bp)
atpF	LSC	159	740	408		
petB	LSC	6	799	648		
petD	LSC	8	698	667		
rpl2	IR	402	668	429		
rpl16	LSC	9	835	396		
rpoC1	LSC	432	674	1665		
rps12		114	-	232	540	26
ycf3	LSC	124	726	230	709	156
trnA-UGC	IR	38	770	35		
trnI-GAU	IR	42	974	35		
trnL-UAA	LSC	50	488	35		
trnK-UUU	LSC	35	3307	37		
trnV-UAC	LSC	39	541	37		

https://doi.org/10.1371/journal.pone.0192966.t003

Table 4. Base compositions in the *Pinus taeda* chloroplast (cp) genome.

	T/U	C	A	G	Length (bp)
Genome	30.8	19.3	30.7	19.3	121,531
LSC	30.7	19.0	30.3	20.0	77,614
SSC	31.3	19.5	31.0	18.3	42,258
IR	31.1	20.2	31.1	17.6	830
tRNA	23.7	24.9	22.4	29.0	2661
rRNA	18.8	23.6	26.4	31.1	4517
Protein coding genes	30.5	18.1	30.5	20.9	61,691
1st position	20.4	16.03	30.26	28.3	20,563
2nd position	31.5	20.7	28.49	18.2	20,563
3rd position	38.18	13.94	31.79	16.07	20,563

https://doi.org/10.1371/journal.pone.0192966.t004

spliced gene, with the N-terminal exon-I located at 92 Kb from C-terminal exons-II and III as reported previously for various gymnosperms [70].

The protein coding regions containing 81 genes were 61,691 bp and accounted for 50.76% of the *P. taeda* cp genome. In the *P. taeda* cp genome, the gene proportion for tRNA was 2.18% and for rRNA it was 3.71%. A total of 43.35% of the non-coding region was composed of introns and intergenic spacers. The total protein-coding sequences encoded 20,563 codons (Table 4). The codon-usage frequency was calculated based on protein-coding and tRNA gene sequences (Table 5). Leucine was the most coded (2,067, 10.1%) and cysteine was the least coded (244, 1.2%) amino acid (Fig 2). Similar ratios for amino acids were found in previously reported cp genomes [71, 72]. The maximum GAA (835; 4.06%) and minimum TGC (65; 0.316%) codons used coded for glutamic acid and encoding cysteine, respectively. The A-T content was 50.6%, 59.99%, and 69.97% at the three consecutive codon positions (Table 4). The preference for the high A-T content at the 3rd codon position is similar to the A and T concentrations reported in various terrestrial plant cp genomes [72–74].

Difference in gene contents of *P. taeda*

We selected 16 cp genomes in the *Pinus* genus (*P. taeda* (old), *P. koraiensis*, *P. sibirica*, *P. armandii*, *P. lambertiana*, *P. krempfii*, *P. bungeana*, *P. gerardiana*, *P. monophylla*, *P. nelsonii*, *P. contorta*, *P. massoniana*, *P. tabuliformis*, *P. taiwanensis*, *P. strobus*, and *P. thunbergii*) for comparison with *P. taeda* (new) (121,531 bp). *Pinus taeda* had the largest genome. The differentiation can be ascribed to the variation in size of LSC (Table 1). Analysis of known genes functions revealed that *P. taeda* shared 60 different protein-coding genes with 15 other *Pinus* species. Furthermore, pairwise alignment between the cp genome of *P. taeda* and six related cp genomes showed the highest synteny. Annotation of the *P. taeda* cp genome was used for plotting the total sequence identity of the six cp genomes of *Pinus* species in mVISTA (Fig 3). The results revealed high sequence identity with five species from the subgenus *Pinus* (*P. contorta*, *P. massoniana*, *P. tabuliformis*, *P. taiwanensis*, and *P. thunbergii*) compared to *P. armandii* from the subgenus *Strobus*. However, for all species, relatively lower identity was observed in various comparable genomic regions, particularly the *trnK-UUU*, *matK*, *atpI*, *rpl16*, *petB*, *petD*, *ycf1*, and *ycf2* regions (Fig 3). Similarly, non-coding regions exhibited greater bifurcation than the coding-regions. Among the diverging regions, *psbA-chlB*, *psbM- clpP*, *ycf4-accD*, *ycf3-psaA*, *psaC-ccsA*, *ndhH- psaC*, *ycf3-psaA*, *trnG-UUU- chlL*, and *petL- psbF* were significant. The current findings agree with the results previously reported for these genes in angiosperm cp genomes [43, 72]. Our results confirmed similar variations among the coding-regions of the

Table 5. Codon–anticodon recognition pattern and codon usage for the *Pinus taeda* chloroplast genome.

Amino acid	Codon	No	RSCU	tRNA	Amino acid	Codon	No	RSCU	tRNA
Phe	UUU	1394	1.11		Tyr	UAC	562	0.66	*trnY-GUA*
Phe	UUC	1108	0.89	*trnF-GAA*	Tyr	UAU	1137	1.34	
Leu	UUA	841	1.23	*trnL-UAA*	Stop	UAA	776	1.05	
Leu	UUG	815	1.19	*trnL-CAA*	Stop	UGA	781	1.06	
Leu	CUU	818	1.2		Stop	UAG	662	0.89	
Leu	CUC	533	0.78		Cyc	UGC	378	0.9	*trnC-GCA*
Leu	CUA	642	0.94	*trnL-UAG*	Trp	UGG	677	1	*trnW-CCA*
Leu	CUG	444	0.65		His	CAU	839	1.43	
Ile	AUU	1233	1.09		His	CAC	337	0.57	*trnH-GUG*
Ile	AUC	963	0.85	*trnI-GAU*	Gln	CAA	842	1.27	*trnQ-UUG*
Ile	AUA	1194	1.06	*trnI-CAU*	Gln	CAG	481	0.73	
Met	AUG	807	1	*trn(f)M-CAU*	Asn	AAU	1318	1.34	
Val	GUU	652	1.29		Asn	AAC	644	0.66	*trnN-GUU*
Val	GUC	365	0.72	*trnV-GAC*	Lys	AAA	1444	1.3	*trnK-UUU*
Val	GUA	606	1.2	*trnV-UAC*	Lys	AAG	770	0.7	
Val	GUG	391	0.78		Asp	GAU	917	1.43	
Ser	UCC	752	1.22	*trnS-GGA*	Asp	GAC	368	0.57	*trnD-GUC*
Ser	UCA	767	1.25	*trnS-UGA*	Glu	GAA	1043	1.33	*trnE-UUC*
Ser	UCG	431	0.7		Glu	GAG	529	0.67	
Pro	CCU	516	1.11		Arg	CGU	278	0.67	*trnR-ACG*
Pro	CCC	400	0.86	*trnP-GGG*	Arg	CGC	163	0.39	
Pro	CCA	624	1.35	*trnP-UGG*	Arg	CGA	439	1.06	
Pro	CCG	313	0.68		Arg	CGG	284	0.68	
Thr	ACU	448	1.05		Ser	AGU	499	0.81	
Thr	ACC	497	1.17		Ser	AGC	387	0.63	*trnS-GCU*
Thr	ACA	441	1.03	*trnT-UGU*	Arg	AGA	821	1.97	*trnR-UCU*
Thr	ACG	320	0.75		Arg	AGG	511	1.23	
Ala	GCU	397	1.38		Gly	GGU	456	0.99	
Ala	GCC	233	0.81		Gly	GGC	214	0.46	*trnG-GCC*
Ala	GCA	347	1.21	*trnA-UGC*	Gly	GGA	728	1.57	*trnG-UCC*
Ala	GCG	172	0.6		Gly	GGG	451	0.98	

https://doi.org/10.1371/journal.pone.0192966.t005

investigated species. This was also suggested by Kumar et al. [75]. Furthermore, comparison of the *P. taeda* whole cp genome with those of related species revealed lower SNP and indel substitutions for the subgenus *Pinus* cp genomes, which ranged from 809 in *P. taeda* (old) to 2,636 in *P. thunbergii*. However, the results revealed higher SNP and indel substitutions within the subgenus *Strobus* cp genomes, which ranged from 9,211 in *P. gerardiana* to 19,196 in *P. monophylla* (S2 Table). These results indicate the presence of interspecific mutations in the highly conservative cp genome that may be useful for analyzing genetic diversity and evolution. Similarly, we evaluated pairwise-sequence differentiation among the 16 pine species (S3 Table). The results showed that the *P. taeda* genome had 0.0274 average sequence divergences, high divergence was detected for *P. nelsonii* (0.0402), and *P. taeda* (old) had the lowest average sequence divergence (0.00321) followed by *P. contorta* (0.00807).

The gene organization and gene contents of the cp genomes are generally conserved compared with those in the mitochondrial and nuclear genomes [76]. The cp genome organization and structure are extremely conserved in angiosperms, i.e. there is a distinctive quadripartite structure containing an SSC region and LSC region separated by a pair of inverted repeats

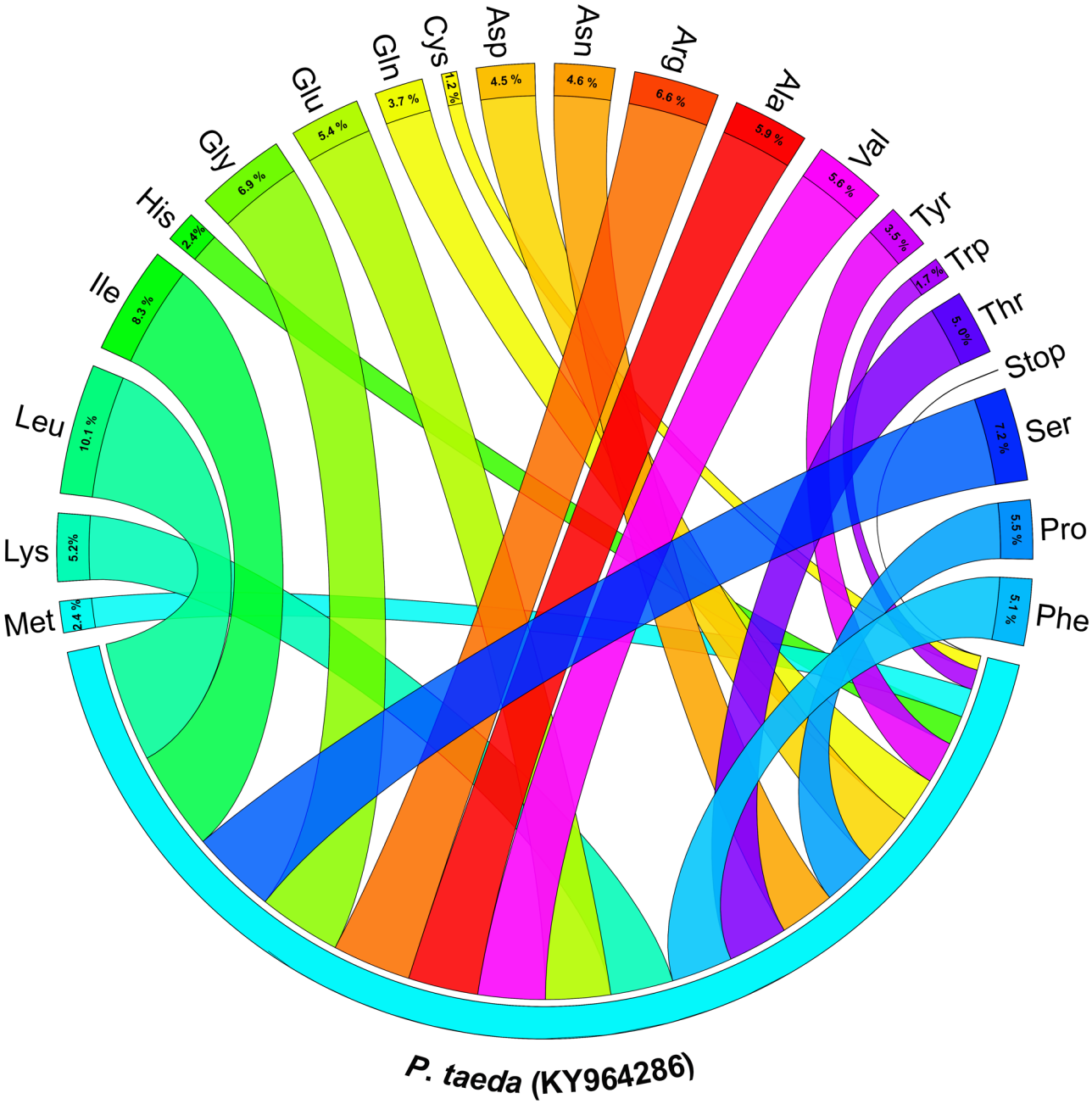

Fig 2. Amino acid frequencies of the *Pinus taeda* chloroplast (cp) protein coding sequences. The frequencies of amino acids were calculated for all 81 protein-coding genes from the start to the stop codon.

https://doi.org/10.1371/journal.pone.0192966.g002

[77]. In contrast, various genome rearrangements have been detected in various gymnosperms cp genomes [78, 79]. While the *P. taeda* cp genome shared some similar characteristics with other plants, we detected noticeable differentiation in numerous genes among gymnosperms. For example, significant divergence was noted in the gene content between *P. taeda* and other gymnosperms. For instance, in *Cryptomeria japonica*, eleven intact NADH dehydrogenase genes were identified, which were correlated to 5 other plant species [37], but were not present

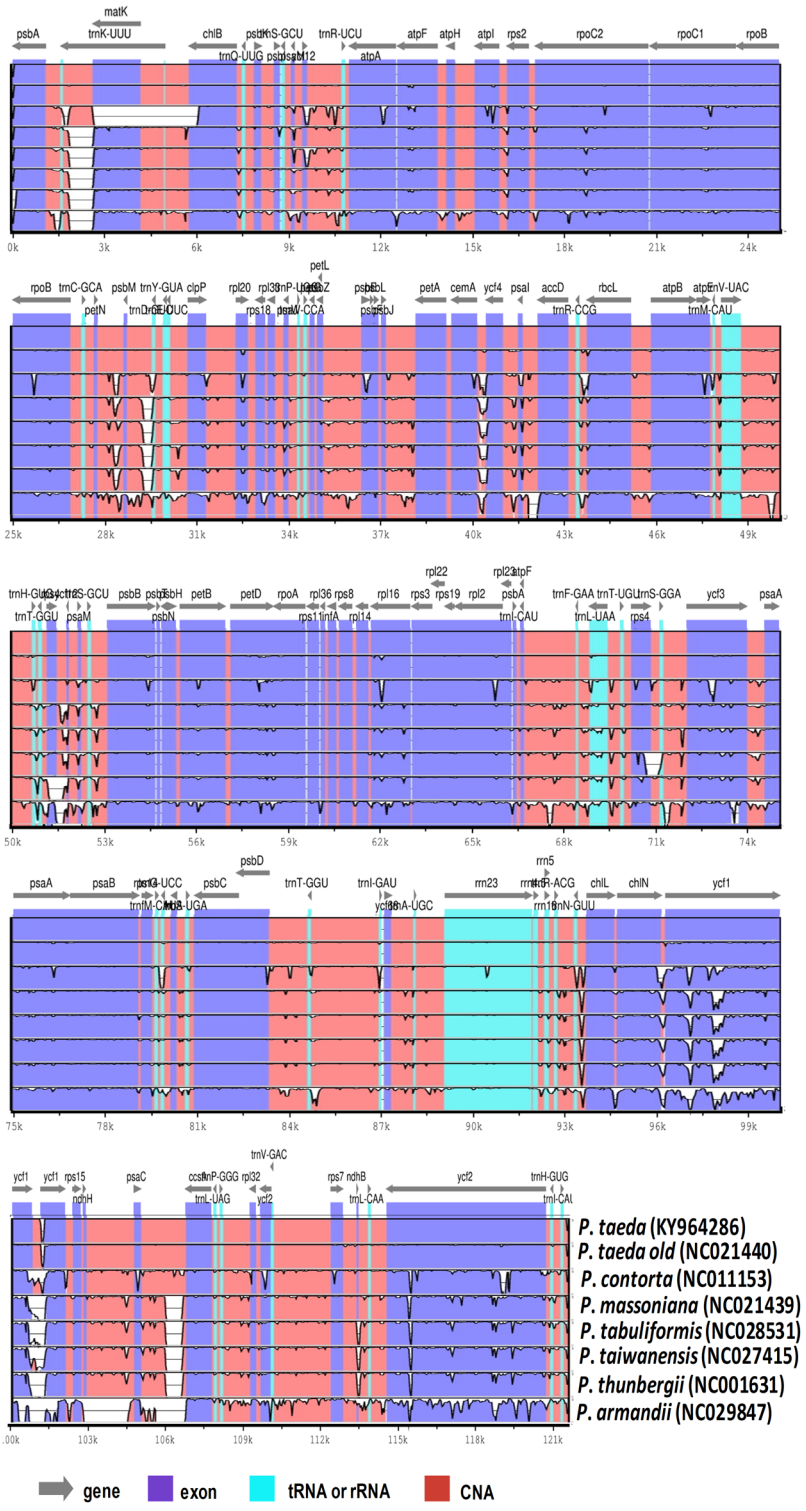

Fig 3. Visual alignment of plastid genomes from *Pinus taeda* and six other *Pinus* species (five from the subgenus *Pinus* and one from the subgenus *Strobus*). VISTA-based identity plot showing sequence identity among seven species, using *P. taeda* as a reference.

https://doi.org/10.1371/journal.pone.0192966.g003

in the *P. taeda* and *P. thunbergii* cp genomes [37]. Previously, it was reported that the loss of NADH dehydrogenases was caused by specific mutations in the cp genome of *Pinus* [79].

In contrast, an essential gene, *rps16*, was completely absent from the *P. taeda* cp genome. Similar results were reported for the *P. thunbergii* and *Marchantia polymorpha* [36, 80] cp genomes, in addition to various terrestrial plants species, including *Eucommia*, *Epifagus*, *Fugus*, *Malpighia*, *Krameria*, *Passiflora*, *Connarus*, *Linum*, *Turnera*, *Securidaca*, *Medicago*, *Selaginella*, *Viola*, and *Adonis* [81–86]. In contrast, *rps16* is present in the angiosperms *Oryza sativa* and *E. globulus*, in the fern *Adiantum capillus*, and in the gymnosperms *C. japonica* and *C. taitungensis*. However, the position of *rps16* is different in gymnosperms from that in angiosperm cp genomes. The position is intermediate between *chlB* and *trnK-UUU* in the gymnosperm cp genomes and halfway between *trnQ-UUG* and *trnK-UUU* and between *chlB* and *matK* in angiosperms and ferns, respectively. Doyle *et al.* [83] suggested the functional transfer of *rps16* to the nucleus from chloroplasts and the absence of this gene from various terrestrial plants. Furthermore, it was reported that the loss of *rps16* and its functional transfer to the nucleus may have occurred autonomously in gymnosperms, particularly in coniferous species.

trnR-CCG and *trnP-GGG* are also found in *P. taeda* cp genomes. These genes are reported as pseudo genes and are likely relics of cp genome evolution in mosses and gymnosperms [29, 87, 88]. *trnP-GGG* was previously reported in two gymnosperms, *C. taitungensis* and *P. thunbergii*, as well as in *C. japonica*, in the fern *A. capillus* and liverwort *M. polymorpha*, and but was absent from the cp genomes of angiosperms. This gene was also identified in *Ginkgo* and *Gnetum* [34], revealing that the gene is common in numerous gymnosperm species. Similarly, *trnR-CCG* in *P. taeda* was previously reported in *C. taitungensis*, *A. capillus*, *P. thunbergii*, and *M. polymorpha*. However, the absence of this gene in *C. japonica* and various cp genomes of angiosperms suggests that *trnR-CCG* is not well-maintained in the cp genomes of all gymnosperms and may have been lost in various taxa during plant evolution [79].

Furthermore, *clpP*, which encodes a proteolytic subunit of the ATP-dependent *clpP* protease, contains no intron in the *P. taeda* cp genome. Similar results were previously reported for *P. thunbergii*, *P. mugo*, *P. dabeshanensis*, and *P. taiwanensis* [37, 41, 68, 89]. In contrast, *clpP* is found in the cp genome of other land plants, such as *A. capillus*, *E. globulus*, *M. polymorpha*, and *C. taitungensis* with two or three exons [29]. However, in the *P. taeda* cp genome, only the *clpP* second exon remained, and as such, it occurs as a pseudogene. Similarly, the *rpl20 and clpP* order is conserved in the *P. taeda* cp genome and *clpP* is co-transcribed with the 5'-end of *rps12* and *rpl20*, as reported previously for the cp genomes of various gymnosperms [90, 91] [92]. *accD* encodes acetyl-CoA-carboxylase and has been found in the *P. taeda* cp genome. The reading frame length of *accD* was similar to that of the cp genomes of other Pinaceae members and has 321 codons, which is fewer than that in *C. japonica* (700 codons) and more than the 309 codons of *A. capillus* and 316 codons of *M. polymorpha*. Furthermore, in angiosperms, particularly monocots, the reading-frame size of *accD* has been reduced from 106 codons in *Oryza sativa* to none in *Zea mays*. This has also been suggested as reason for the loss of *accD* in monocot plant species [93]. In contrast, the *accD* reading-frame in gymnosperms, particularly in coniferous species and *C. japonica*, may have diverted in the ascending direction.

Loss of large IR region within the *P. taeda* cp genome

The large inverted repeat regions, which have been reported in various land plant cp genomes, were reduced to two very short inverted repeat (IRa and IRb) regions of 830 bp in *P. taeda*, and were separated by a SSC region of 42,258 bp and LSC region of 77,614 bp (Fig 1). However, in the previously sequenced *P. taeda* cp genome submitted to NCBI, the short inverted repeat regions were 693 bp (Table 1). Similar results were observed in other Pinaceae

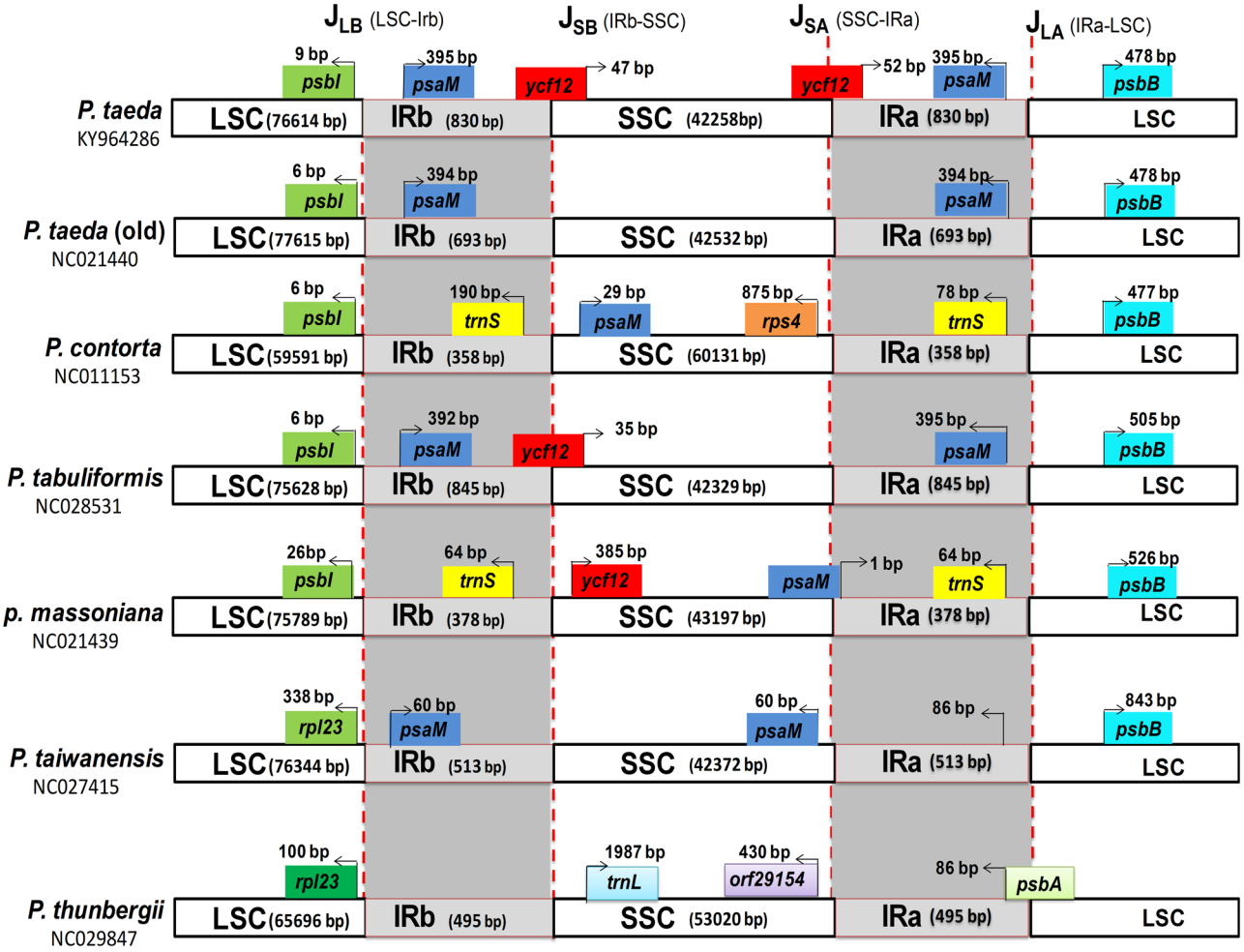

Fig 4. Distance between adjacent genes and junctions of the small single-copy (SSC), large single-copy (LSC), and two inverted repeat (IR) regions among plastid genomes from six *Pinus* species. Boxes above and below the main line indicate the adjacent border genes. The figure is not to scale regarding sequence length, and only shows relative changes at or near the IR/SC borders.

https://doi.org/10.1371/journal.pone.0192966.g004

members, such as *P. taiwanensis*, *P. armandii*, and *P. dabeshanensis*, where the inverted repeat sizes were reduced to 513, 475, and 473 bp, respectively [68, 69, 89]. The IR of *P. taeda* contained duplicated *psaM* and *trnS-GCU* and partial *ycf12*, apparently caused by incomplete loss of the large IR, as reported previously for various gymnosperms [36, 37]. Detailed comparison of four junctions (J_{LA}, J_{LB}, J_{SA}, and J_{SB}) between the two IRs (IRa and IRb) and two single-copy regions (LSC and SSC) was performed between *Pinus* species (*P. contorta*, *P. tabuliformis*, *P. massoniana*, *P. taiwanensis*, and *P. thunbergii*) and *P. taeda* by carefully analyzing the exact IR border positions and adjacent genes (Fig 4). Some IR expansion and contraction were observed in the *P. taeda* cp genome compared to that of the other five *Pinus* species, which ranged from 358 bp (*P. contorta*) to 845 bp (*P. tabuliformis*) (Fig 4). The genes marking the beginning and end of the IRs were only partially duplicated. *psbI* in *P. taeda* was located 9 bp from J_{LB} in the LSC region. In *P. contorta*, *P. tabuliformis*, and *P. taeda* (old), this distance was 6 bp, whereas in *P. massoniana* and *P. taiwanensis* the distances were 26 and 338 bp, respectively. However, variation was found in *P. thunbergii*, and *rpl23* was 100 bp away from J_{LB} in the LSC region. Similarly, hypothetical chloroplast *ycf12* was partially duplicated by 47 bp (*P. taeda*) and 35 bp in *P. tabuliformis*. However, in *P. massoniana*, *ycf12* was located in the SSC

region, 385 bp away from J_{SB}. In *P. taeda* and *P. tabuliformis*, J_{LA} was located between *psaM* and *psbB* and the difference in distance between *psaM* and J_{LA} was 395 bp. However, in *P. contorta* and *P. taiwanensis*, *psaM* was located in the SSC region, whereas in *P. massoniana*, it was located at the J_{SA} border (Fig 4). Similarly, in *P. taeda*, *P. contorta*, *P. tabuliformis*, *P. massoniana*, and *P. taiwanensis*, *psbB* was located in the LSC region at 478, 477, 505, 526, and 843 bp away from the J_{LA} border, respectively.

Large IRs play a significant role in stabilizing and maintaining the conserved structure of the cp genomes [94]. Various studies have reported that during the evolutionary process of angiosperms, a copy of an IR was lost, particularly in the subfamily Papilionoideae [95–97], and rearrangement in the chloroplast genome was observed because of IR loss in these genomes as compared to cp genomes with normal IRs [94]. Similarly, in gymnosperms, complete IRs were lost in conifers, particularly in cupressophytes and Pinaceae cp genomes, and greater rearrangement was observed in these genomes compared to in higher plants [33]. The remaining IR parts in various Pinaceae member and cupressophyte cp genomes were shown to differ, suggesting that these two conifer clades lost their large IRs independently during evolution from a common ancestor [78, 98]. Previously, it was reported that specific repeats in Pinaceae replaced the reduced IRs [99]. Compared to other conifers, a greater number of rearrangements occurred in *Pseudotsuga menziesii* and *P. radiate* cp genomes because of the lack of a large IR in these cp genomes [33]. Therefore, variation in the genome structure between *P. taeda* and related terrestrial plants, such as *C. japonica*, suggest that an IR is essential for structural stability of the cp genome.

Repeat analysis

Repeat analysis of the *P. taeda* cp genome revealed six palindromic repeats, 34 forward repeats, and 22 tandem repeats (S1 Fig and Table 6). Among these, three forward repeats were 45–59 bp in length, with 14 tandem repeats of 15–29 bp in length (S1 Fig). Additionally, two palindromic repeats were 75–89 bp and four repeats were >90 bp (S1 Fig). Overall, 62 repeats were found in the *P. taeda* cp genome. Among tandem repeats, 12 repeats were in coding regions, eight repeats in intergenic regions, one repeat extending from an intergenic region into a coding region, and one repeat in the *petB* intron region (Table 7). The length of tandem repeats in these regions varied between eight and 14, and up to 10 repeat units were present. Various numbers of repeats have been identified in conifer cp genomes [100, 101] and the mechanisms implicit in the origin of these tandem repeats remain unclear. Nevertheless, they are known to be associated with chloroplast DNA rearrangement [102], gene expansion [100, 101], and gene duplication [103]. Previous reports suggested that repeat sequences, which play a role in genome rearrangement, are very helpful in phylogenetic studies [74, 104]. Furthermore, analyses of different cp genomes revealed that repeat sequences are important causes of indels and substitutions [101]. Sequence variation and cp genome re-arrangement occurs because of the slipped strand mis-pairing and improper recombination of repeat sequences [104–106]. The presence of such repeats shows that the locus is an important hotspot for cp genome re-configuration [74, 107]. In addition, such repeats contain crucial information for developing genetic markers for phylogenetic and population studies [74].

SSR analysis

SSRs are repeating sequences of typically 1–6 bp that are distributed throughout the genome. SSRs generally have a high mutation rate compared to neutral DNA regions because of slipped-strand mispairing. Because these short repeats are uniparentally inherited and haploid, they can be used as molecular markers in genetic studies analyzing population structures [108,

Table 6. Repeat sequences in the *Pinus taeda* chloroplast genome.

Repeat type	Repeat size	Repeat Position 1	Repeat location 1	Repeat Position 2	Repeat location 2
P	830	8692	*psbl-psbM-ycf12*	51,779	*ycf12-psbM*
P	399	66,445	*psbA-atpF*	121,132	IGS
P	304	50,503	IGS	120,845	IGS
P	277	50,530	IGS	120,845	IGS
P	86	0	*psbA*	66,359	*psbA*
P	79	9017	IGS	52,205	*psbM*-IGS
F	800	175	*psbA*	1815	IGS
F	376	109,649	*ycf2*	120,134	*ycf2*
F	288	50,861	IGS	84,618	IGS
F	284	50,843	IGS	84,600	IGS
F	275	50,825	IGS	84,582	IGS
F	247	51,131	*rps4*	70,403	*rps4*
F	185	50,964	IGS	84,721	IGS
F	171	51,207	*rps4*	70,479	*rps4*
F	165	100,638	*ycf1*	100,659	*ycf1*
F	124	101,059	IGS-*ycf1*	101,068	IGS-*ycf1*
F	97	9677	IGS	30,444	IGS
F	97	101,059	IGS-*ycf1*	101,113	IGS-*ycf1*
F	85	9737	IGS	30,504	IGS
F	70	100,733	*ycf1*	100,754	*ycf1*
F	79	9017	IGS	52,205	psbM
F	73	9701	IGS	30,468	IGS
F	71	100,638	*ycf1*	100,701	*ycf1*
F	70	100,712	*ycf1*	100,754	IGS
F	70	101,059	IGS-*ycf1*	101,122	*ycf1*
F	70	101,086	*ycf1*	101,140	*ycf1*
F	62	93,524	IGS	93,579	IGS
F	69	115,329	*ycf2*	115,395	ycf2
F	71	9777	*ycf1*	30,544	IGS
F	71	101,086	*ycf1*	101,149	*ycf1*
F	70	101,077	*ycf1*	101,140	*ycf1*
F	69	9714	IGS	30,481	IGS
F	58	71,811	IGS	71,831	IGS
F	67	101,149	*ycf1*	101,167	*ycf1*
F	61	101,059	*ycf1*	101,131	*ycf1*
F	64	101,057	*ycf1*	101,138	*ycf1*
F	63	101,057	*ycf1*	101,147	*ycf1*
F	59	101,043	*ycf1*	101,133	*ycf1*
F	55	100,895	ycf1 intron	100,976	*ycf1* intron
F	61	101,068	*ycf1*	101,149	*ycf1*

https://doi.org/10.1371/journal.pone.0192966.t006

109]. In this study, we detected perfect SSRs in the *P. taeda* cp genome (Fig 5). Specific attributes were set for the analysis because SSRs (10 bp or longer) are exposed to slipped strand mis-pairing, the main mechanism of SSR polymorphisms [110–112]. A total of 151 perfect microsatellites were found in the *P. taeda* cp genome (Fig 5). Most (71) SSRs in this cp genome possessed a mononucleotide repeat motif. Dinucleotide SSRs were the second most common repeat motif (Fig 5B). Using our search criterion, four tetranucleotide SSRs and one

Table 7. Tandem repeat sequences in the *Pinus taeda* chloroplast genome.

Serial No	Indices	Repeat Length	Size of repeat unit × Copy number	A	C	G	T	Location
1	9274–9310	36	2 × 18	16	16	16	50	*PsaM/ycf12* (IGS)
2	15,199–15,235	36	2 × 18	44	8	23	23	*atpI* (CDS)
3	20,648–20,678	30	2 × 15	50	10	20	20	*rpoC2* (CDS)
4	28,466–28,534	68	2 × 34	30	24	12	33	*petN/psbM* (IGS)
5	31,275–31,313	38	2 × 19	23	13	36	26	*clpP*/IGS
6	33,103–33,166	63	3 × 21	29	16	19	33	*rps18* (CDS)
7	43,597–43,625	28	2 × 14	46	0	10	43	*accD/rbcL* (IGS)
8	43,615–43,659	44	2 × 22	40	12	8	38	*accD/rbcL* (IGS)
9	45,578–45,620	42	2 × 21	31	2	24	41	*rbcL/atpB* (IGS)
10	51,993–52,029	36	2 × 18	50	16	16	16	*ycf12/psbM* (IGS)
11	56,031–56,069	38	2 × 19	18	12	12	57	*petB* (intron)
12	93,544–93,631	87	3 × 29	37	16	10	35	*ycf68/chlL* (IGS)
13	93,525–93,635	110	2 × 55	35	15	11	36	*ycf68/chlL* (IGS)
14	97,002–97,056	54	2 × 27	28	20	24	26	*ycf1*(CDS)
15	100,583–100,631	48	2 × 24	54	9	18	16	*ycf1*(CDS)
16	100,639–100,828	189	9 × 21	45	9	28	16	*ycf1*(CDS)
17	100,827–101,025	198	6 × 33	31	1	43	23	*ycf1*(CDS)
18	100,866–101,016	150	10 × 15	30	1	44	23	*ycf1*(CDS)
19	100,827–101,953	126	2 × 63	31	1	43	23	*ycf1*(CDS)
20	100,823–101,985	162	2 × 81	32	2	42	22	*ycf1*(CDS)
21	100,939–101,047	108	2 × 54	34	4	38	22	*ycf1*(CDS)
22	115,330–115,452	122	2 × 66	21	22	11	45	*ycf2* (CDS)

https://doi.org/10.1371/journal.pone.0192966.t007

hexanucleotide SSR were detected in the *P. taeda* cp genome (Fig 5A). In *P. taeda*, most mononucleotide SSRs were A (92.5%) and C (8.45%) motifs, with most dinucleotide SSRs being A/T (47.3%) and A/G (52.63%) motifs (Fig 5B and Table 8). Approximately 59.60% of SSRs were in non-coding regions, approximately 2.64% were present in rRNA sequences, and 1.98% were in tRNA genes (Fig 5A). These results are similar to those of previous reports showing that SSRs were unevenly distributed in cp genomes, and these findings may provide more information for selecting effective molecular markers for detecting intra- and interspecific polymorphisms [113–116]. Furthermore, analysis of various gymnosperm cp genomes revealed that most mononucleotides and dinucleotides are composed of A and T, which may contribute to bias in base composition, which is consistent with other cp genomes [117–119]. For SSR identification, although different criteria and algorithms were used, their distribution and characteristics were similar to the cp genomes of conifers [71, 119], 30 asterid [72], and 14 monocot [112]. Our findings were comparable to those of previous reports in which SSRs in cp genomes were found to be largely composed of polythymine (polyT) or polyadenine (polyA) repeats, and infrequently contained tandem cytosine (C) and guanine (G) repeats [118, 120].

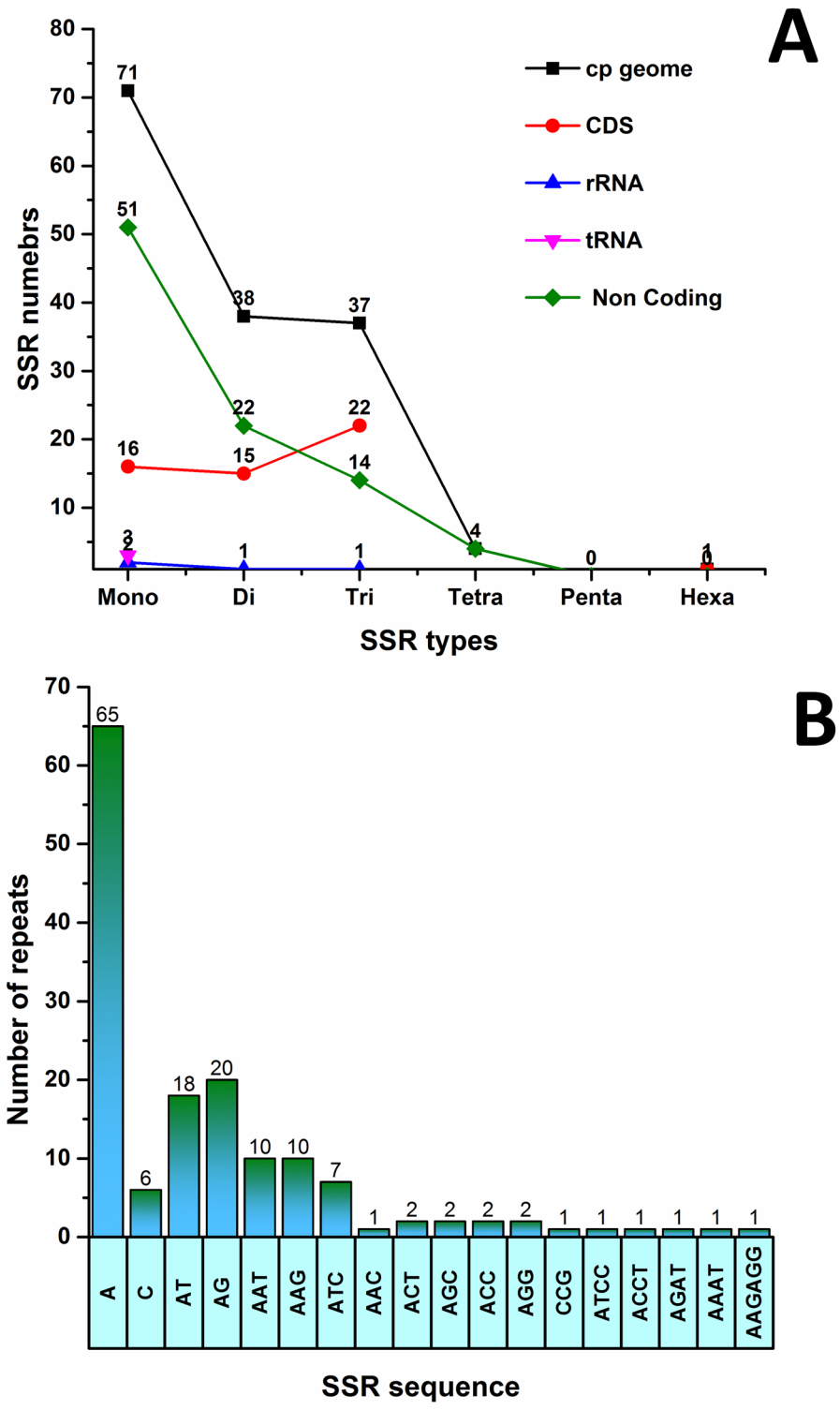

Fig 5. Analysis of simple sequence repeat (SSR) in the *Pinus taeda* plastid genome. A, Number of SSR types in complete genome, coding, and non-coding regions; **B**, Frequency of identified SSR motifs in different repeat class types.

https://doi.org/10.1371/journal.pone.0192966.g005

Table 8. Simple sequence repeats (SSRs) in the *Pinus taeda* chloroplast genome.

Unit	Length	No	SSR start
A	15	2	1375, 28,440
	14	3	68,741, 72,734, 106,240
	12	2	10,316, 110,251
	11	4	10,755, 26,980, 109,368, 11,873
	10	8	16,119, 22,252, 48,967, 83,427, 86,798, 88,062, 102,308, 111,412
	9	15	40,699, 41,827, 45,769, 70,952, 80,498, 80,744, 95,259, 102,053, 108,265, 110,985, 112,374, 113,688, 117,432, 119,716, 120,740
	8	31	4819, 10,738, 10,950, 16,110, 17,113, 30,189, 30,427, 30,701, 31,373, 33,345, 38,678, 41,893, 50,753, 51,485, 52622, 55,355, 56,042, 63,021, 64,394, 64,437, 92,458, 94,554, 95,822, 97,307, 103,868, 108,971, 114,282, 117065, 118885, 119,819, 120,893
C	9	4	16,101, 22,497, 71,353, 105,552
	8	2	31,381, 120,721
AT	13	1	41,344
	10	4	26,392, 96,162, 104,388, 113,787
	9	6	19,814, 24,397, 34,072, 42,422, 48,777, 74,253
	8	7	19,352, 19,904, 80,532, 83,639, 99,803, 105,218, 110,933
AG	9	10	8774, 22,311, 26,631, 47,568, 51,573, 52,520, 65,195,79,220, 80,699, 106,488,
	8	10	14,675, 22,384, 30,793, 42,926, 51,556, 69,139, 75,721, 83,721, 90,777, 91,093
AAT	11	1	78,353
	10	1	42,354
	9	8	13,934, 49,935, 65,369, 66,308, 71,749, 94,150, 98,727, 109,563
AAG	10	5	3167, 22,135, 106,110, 108,709, 120,693
	9	5	28,380, 79,051, 79,226, 81,004, 100,527
ATC	10	1	77,667
	9	6	2957, 16,215, 21,127, 75,445, 77,964, 111,780
AAC	9	1	32,982
ACT	9	2	43,692, 94,864
AGC	9	2	43,798, 89,223
ACC	9	2	54,293, 94,538
AGG	9	2	60,538, 80,037
CCG	9	1	
ATCC	17	1	48,863
ACCT	14	1	90,739
AGAT	13	1	51,753
AAAT	12	1	42,147
AAGAGG	23	1	117,038

https://doi.org/10.1371/journal.pone.0192966.t008

Therefore, these SSRs contributed to the A-T richness of the *P. taeda* cp genome, which was also previously observed in the cp genomes of plant species [43, 71, 120]. The SSRs identified in the cp genome of *P. taeda* can be evaluated for polymorphisms at the intra-specific levels and used as markers for evaluating the genetic diversity of wild populations of plants from the Pinaceae family.

Phylogenetic analysis

In plants, the cp genome is a valuable resource for exploring intra- and interspecific evolutionary histories [121–127]. Compared to nuclear genomes in chloroplasts, the uniparental inheritance (for exceptions, see [122, 128]) is systematically striking because a single, independent

genealogical history can be readily obtained for developing hypotheses [129–131]. Moreover, in some land plants (a few flowering plant lineages and conifers), the chloroplast is paternally inherited and independent of the nuclear and mitochondrial genome [132].

Recently, cp genomes have shown significant power in phylogenetic, evolution, and molecular systematics studies. During the last decade, various analyses have revealed the phylogenetic relationships at deep nodes based on comparisons of multiple protein coding genes, intergenic spacers [133, 134], and complete genome sequences in chloroplast genomes [135] that have enhanced our understanding of the evolutionary relationships among angiosperms and gymnosperms. According to the most recent classification, the genus *Pinus* is comprised of approximately 110 species and is shared by two subgenera, *Strobus* and *Pinus*, which are divided into further sections [136]. Furthermore, some evolutionary hypotheses suggest that the subgenera *Strobus* and *Pinus* originated from the Eocene [137, 138], whereas others indicated these subgenera were already present during the Cretaceous [138–140]. The *Pinus* subgenus has undergone significant distributional as well as environmental changes during their evolution, such as moving multiple times between America and Eurasia [140]. Chloroplast DNA polymorphisms in *P. taeda* have been used in numerous studies to assess paternal inheritance lineage and cytoplasmic diversity [141–146]. Continued efforts have expanded our ability to differentiate and understand the genomic structure and phylogenetic relationships of *Pinus* species [147]. The phylogeny and taxonomy of *Pinus* species have largely relied on chloroplast markers [140, 148, 149]. However, compared to nuclear genes, these markers are linked and offer independent information on species phylogeny. Previously, the phylogenetic study of pine based on multiple nuclear genes was reported by Syring et al. [150], where four low-copy nuclear loci were analyzed in 12 pine species and combined with internal transcribed spacers and chloroplast data. Various studies revealed that the addition of more genes increased the chance for improving the phylogenetic tree [151–153]. However, this does not resolve all phylogenetic problems [154, 155].

Complete genome sequencing provides detailed insight into an organism [43, 66, 156]. In this study, the phylogenetic position of *P. taeda* within the *Pinus* genus was established by employing the complete cp genome and 60 shared genes of 16 species. Phylogenetic analyses using Bayesian inference, maximum parsimony, maximum likelihood, and neighbor-joining methods were performed. The phylogenetic analysis revealed that the complete dataset and 60 shared genes of *P. taeda* contained the same phylogenetic signals. In the datasets for the genome and 60 shared genes, *P. taeda* formed a single clade with *P. contorta* with high Bayesian interference and bootstrap support using the four different methods (Fig 6 and S2 Fig). Moreover, tree topology confirmed the relationship inferred from the phylogenetic work previously conducted based on cp genomes [89, 141, 157], in which *P. taeda* was genetically similar to *P. contorta*. These results revealed good agreement with classical taxonomy, where similar concordance was observed in the cp genome and mitochondrial genome-based reconstructions of *Pinus* phylogeny [136, 140]. Furthermore, these results are in broad agreement with previous results reported by Niu et al., where *P. taeda* formed a single clade with *P. contorta* based on pairwise non-synonymous substitution rates of orthologous transcripts [158]. Additionally, the results suggest that there is no conflict between the entire genome dataset and 60 shared genes in these cp genomes.

Conclusion

The current study determined the complete genome sequence of the chloroplast from *P. taeda* (121,531 bp). The gene order and genome structure of *P. taeda* was similar to that of cp genomes of other *Pinus* species. Furthermore, the distribution and location of repeat sequences

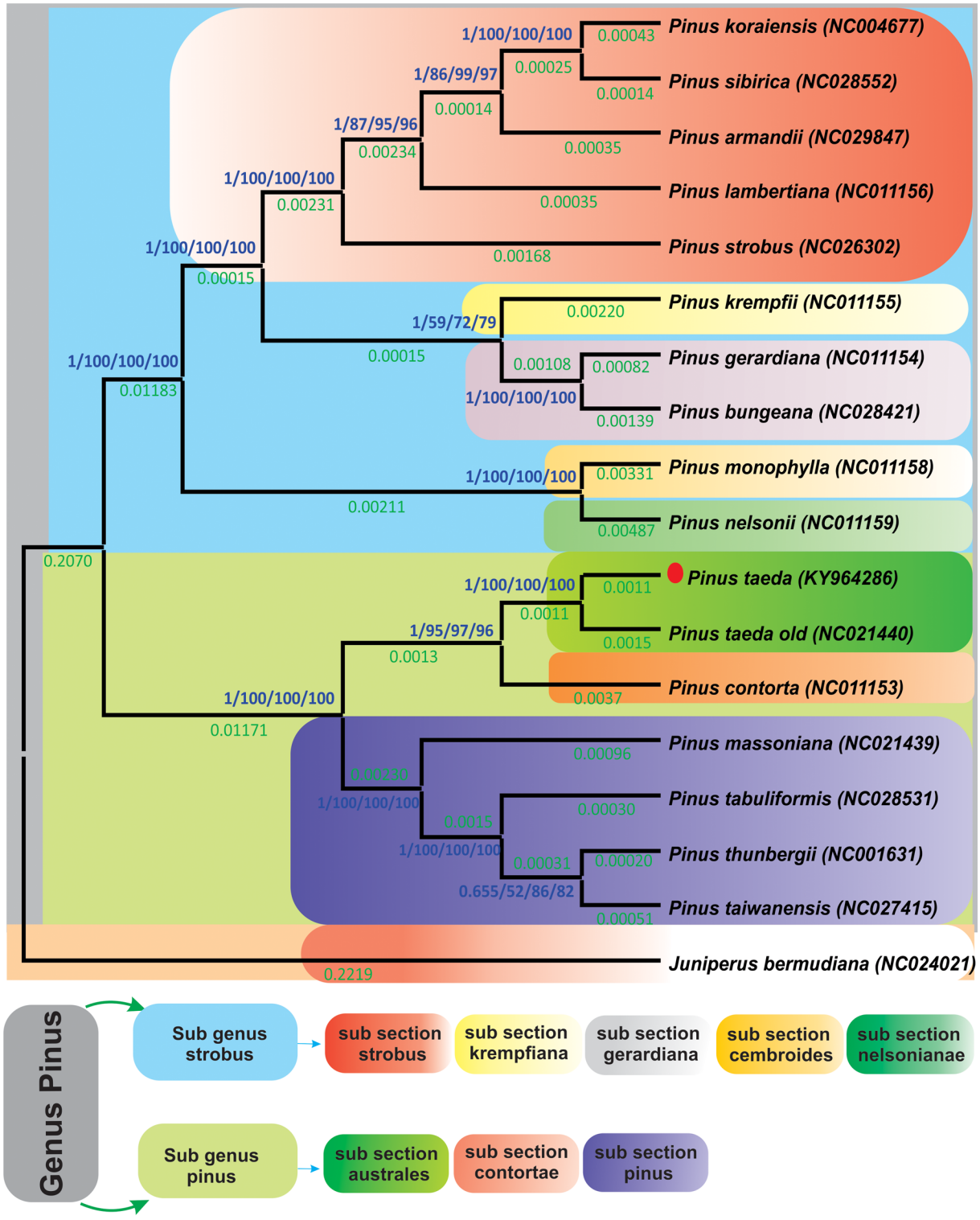

Fig 6. Phylogenetic trees of 15 *Pinus* species. The entire genome dataset was analyzed using four different methods: Bayesian inference (BI), maximum parsimony (MP), maximum likelihood (ML), and neighbor-joining (NJ). Numbers above the branches represent bootstrap values in the MP, ML, and NJ trees and posterior probabilities in the BI trees, whereas the number below the branches represents branch length. The red dot represents the position of *P. taeda* (KY964286).

https://doi.org/10.1371/journal.pone.0192966.g006

were determined, and average pairwise sequence divergences among cp genomes of related species were identified. SSR, SNP, and phylogenetic analyses were performed on 16 *Pinus* species cp genomes. No major structural rearrangement of *Pinus* species cp genomes was observed. Phylogenetic analyses revealed that the dataset based on 60 shared genes and that of the entire genome generated trees with the same topologies regarding the placement of *P. taeda*. Such investigations are an essential source of important information on the complete cp genome of *P. taeda* and related species, which can be used to facilitate biological study, identify species, and clarify taxonomic questions.

Supporting information

S1 Table. Primers used for gap closing and sequencing verification in *Pinus taeda*.
(DOCX)

S2 Table. Indel and SNP analysis of plastid genomes from *Pinus taeda* and 15 other *Pinus* species.
(XLSX)

S3 Table. Average pairwise distance of plastid sequences from *Pinus taeda* and 15 other *Pinus* species.
(XLS)

S1 Fig. Analysis of repeated sequences in *Pinus taeda* plastid genome. Total forward, tandem, and palindromic repeat sequences in the genome and their length distributions.
(TIF)

S2 Fig. Phylogenetic trees were constructed for 15 species in the genus *Pinus* using different methods and the Bayesian tree is shown for the entire genome sequence. Data for 60 shared genes were used with four different methods: Bayesian inference (BI), maximum parsimony (MP), maximum likelihood (ML), and neighbor-joining (NJ). Numbers above the branches represent bootstrap values in the MP, ML, and NJ trees and posterior probabilities in the BI trees. The red dot represents the position of *P. taeda* (KY964286).
(TIF)

Author Contributions

Conceptualization: Sajjad Asaf, Abdul Latif Khan, Raheem Shahzad, Ahmed Al-Rawahi.

Data curation: Sajjad Asaf, Lubna.

Formal analysis: Sang Mo Kang.

Methodology: Abdul Latif Khan, Sang Mo Kang.

Resources: Ahmed Al-Harrasi.

Software: Ahmed Al-Harrasi.

Supervision: In-Jung Lee.

Validation: Muhammad Aaqil Khan.

Writing – original draft: Muhammad Aaqil Khan, Raheem Shahzad.

Writing – review & editing: Ahmed Al-Rawahi.

References

1. Farjon A. World checklist and bibliography of conifers: Royal Botanic Gardens; 2001.
2. Bowe LM, Coat G. Phylogeny of seed plants based on all three genomic compartments: Extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Proceedings of the National Academy of Sciences. 2000; 97(8):4092–7.
3. Peterson DG, Schulze SR, Sciara EB, Lee SA, Bowers JE, Nagel A, et al. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res. 2002; 12(5):795–807. https://doi.org/10.1101/gr.226102 PMID: 11997346
4. Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, Garcia SA, et al. Evolution of genome size and complexity in Pinus. Plos One. 2009; 4(2):e4332. https://doi.org/10.1371/journal.pone.0004332 PMID: 19194510
5. Zonneveld B. Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. Nordic Journal of Botany. 2012; 30(4):490–502.
6. Ahuja MR, Neale DB. Evolution of genome size in conifers. Silvae genetica. 2005; 54(3):126–37.
7. Kovach A, Wegrzyn JL, Parra G, Holt C, Bruening GE, Loopstra CA, et al. The Pinus taeda genome is characterized by diverse and highly diverged repetitive sequences. Bmc Genomics. 2010; 11(1):420.
8. Mackay J, Dean JF, Plomion C, Peterson DG, Cánovas FM, Pavy N, et al. Towards decoding the conifer giga-genome. Plant Mol Biol. 2012; 80(6):555–69. https://doi.org/10.1007/s11103-012-9961-7 PMID: 22960864
9. Neale DB, Kremer A. Forest tree genomics: growing resources and applications. Nat Rev Genet. 2011; 12(2):111–22. https://doi.org/10.1038/nrg2931 PMID: 21245829
10. Frederick W, Lien S, Courchene C, DeMartini N, Ragauskas A, Iisa K. Production of ethanol from carbohydrates from loblolly pine: A technical and economic assessment. Bioresource Technol. 2008; 99(11):5051–7.
11. González-Martínez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB. Association genetics in Pinus taeda LI Wood property traits. Genetics. 2007; 175(1):399–409. https://doi.org/10.1534/genetics.106.061127 PMID: 17110498
12. Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, González-Martínez SC, et al. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (Pinus taeda L., Pinaceae). Genetics. 2010; 185(3):969–82. https://doi.org/10.1534/genetics.110.115543 PMID: 20439779
13. Stewart JF, Tauer CG, Nelson C. Bidirectional introgression between loblolly pine (Pinus taeda L.) and shortleaf pine (P. echinata Mill.) has increased since the 1950s. Tree Genet Genomes. 2012; 8(4):725–35.
14. Elsik C, Williams C. Low-copy microsatellite recovery from a conifer genome. Theor Appl Genet. 2001; 103(8):1189–95.
15. Eckert AJ, Pande B, Ersoz ES, Wright MH, Rashbrook VK, Nicolet CM, et al. High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (Pinus taeda L.). Tree Genet Genomes. 2009; 5(1):225–34.
16. Echt CS, Saha S, Krutovsky KV, Wimalanathan K, Erpelding JE, Liang C, et al. An annotated genetic map of loblolly pine based on microsatellite and cDNA markers. Bmc Genet. 2011; 12(1):17.
17. Resende M, Munoz P, Acosta J, Peter G, Davis J, Grattapaglia D, et al. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytol. 2012; 193(3):617–24. https://doi.org/10.1111/j.1469-8137.2011.03895.x PMID: 21973055
18. Wegrzyn JL, Lin BY, Zieve JJ, Dougherty WM, Martínez-García PJ, Koriabine M, et al. Insights into the loblolly pine genome: characterization of BAC and fosmid sequences. Plos One. 2013; 8(9):e72439. https://doi.org/10.1371/journal.pone.0072439 PMID: 24023741
19. Consortium IBGS. A physical, genetic and functional sequence assembly of the barley genome. Nature. 2012; 491(7426):711–6. https://doi.org/10.1038/nature11543 PMID: 23075845
20. Dagan T, Roettger M, Stucken K, Landan G, Koch R, Major P, et al. Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. Genome Biol Evol. 2013; 5(1):31–44. https://doi.org/10.1093/gbe/evs117 PMID: 23221676
21. Wagner DB, Furnier GR, Saghai-Maroof MA, Williams SM, Dancik BP, Allard RW. Chloroplast DNA polymorphisms in lodgepole and jack pines and their hybrids. Proc Natl Acad Sci U S A. 1987; 84(7):2097–100. Epub 1987/04/01. PMID: 3470779
22. Hong YP, Hipkins VD, Strauss SH. Chloroplast DNA diversity among trees, populations and species in the California closed-cone pines (Pinus radiata, Pinus muricata and Pinus attenuata). Genetics. 1993; 135(4):1187–96. Epub 1993/12/01. PMID: 7905846

23. Dong J, Wagner DB. Paternally inherited chloroplast polymorphism in Pinus: estimation of diversity and population subdivision, and tests of disequilibrium with a maternally inherited mitochondrial polymorphism. Genetics. 1994; 136(3):1187–94. Epub 1994/03/01. PMID: 8005423

24. Tsumura Y, Suyama Y, Taguchi H, Ohba K. Geographical cline of chloroplast DNA variation in Abies mariesii. Theor Appl Genet. 1994; 89(7–8):922–6. Epub 1994/12/01. https://doi.org/10.1007/BF00224518 PMID: 24178104.

25. Neale DB, Sederoff RR. Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in loblolly pine. Theor Appl Genet. 1989; 77(2):212–6. Epub 1989/02/01. https://doi.org/10.1007/BF00266189 PMID: 24232531.

26. Szmidt AE, Alden T, Hallgren JE. Paternal inheritance of chloroplast DNA in Larix. Plant Mol Biol. 1987; 9(1):59–64. Epub 1987/01/01. https://doi.org/10.1007/BF00017987 PMID: 24276798.

27. Neale DB, Marshall KA, Sederoff RR. Chloroplast and mitochondrial DNA are paternally inherited in Sequoia sempervirens D. Don Endl. Proc Natl Acad Sci U S A. 1989; 86(23):9347–9. Epub 1989/12/01. PMID: 16594091

28. Kondo T, Tsumura Y, Kawahara T, Okamura M. Paternal inheritance of chloroplast and mitochondrial DNA in interspecific hybrids of Chamaecyparis spp. Japanese Journal of Breeding. 1998; 48(2):177–9.

29. Wakasugi T, Hirose T, Horihata M, Tsudzuki T, Kossel H, Sugiura M. Creation of a novel protein-coding region at the RNA level in black pine chloroplasts: The pattern of RNA editing in the gymnosperm chloroplast is different from that in angiosperms. P Natl Acad Sci USA. 1996; 93(16):8766–70.

30. Sugiura M. The chloroplast genome. Plant Mol Biol. 1992; 19(1):149–68. Epub 1992/05/01. PMID: 1600166.

31. Lidholm J, Szmidt AE, Hallgren JE, Gustafsson P. The chloroplast genomes of conifers lack one of the rRNA-encoding inverted repeats. Mol Gen Genet. 1988; 212(1):6–10. Epub 1988/04/01. PMID: 24649523.

32. Tsumura Y, Ogihara Y, Sasakuma T, Ohba K. Physical map of chloroplast DNA in sugi, Cryptomeria japonica. Theor Appl Genet. 1993; 86(2–3):166–72. Epub 1993/04/01. https://doi.org/10.1007/BF00222075 PMID: 24193456.

33. Strauss SH, Palmer JD, Howe GT, Doerksen AH. Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. Proc Natl Acad Sci U S A. 1988; 85(11):3898–902. Epub 1988/06/01. PMID: 2836862

34. Wu CS, Wang YN, Liu SM, Chaw SM. Chloroplast genome (cpDNA) of Cycas taitungensis and 56 cp protein-coding genes of Gnetum parvifolium: Insights into cpDNA evolution and phylogeny of extant seed plants. Mol Biol Evol. 2007; 24(6):1366–79. https://doi.org/10.1093/molbev/msm059 PMID: 17383970

35. Palmer JD, Stein DB. Conservation of chloroplast genome structure among vascular plants. Curr Genet. 1986; 10(11):823–33.

36. Tsudzuki J, Nakashima K, Tsudzuki T, Hiratsuka J, Shibata M, Wakasugi T, et al. Chloroplast DNA of black pine retains a residual inverted repeat lacking rRNA genes: nucleotide sequences of trnQ, trnK, psbA, trnI and trnH and the absence of rps16. Mol Gen Genet. 1992; 232(2):206–14. Epub 1992/03/01. PMID: 1557027.

37. Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M. Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine Pinus thunbergii. Proc Natl Acad Sci U S A. 1994; 91.

38. White EE. Chloroplast DNA in Pinus monticola: 1. Physical map. Theor Appl Genet. 1990; 79(1):119–24. Epub 1990/01/01. https://doi.org/10.1007/BF00223797 PMID: 24226130.

39. Lidholm J, Gustafsson P. The chloroplast genome of the gymnosperm Pinus contorta: a physical map and a complete collection of overlapping clones. Curr Genet. 1991; 20(1–2):161–6. Epub 1991/07/01. PMID: 1682061.

40. Kluge AG. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Syst Biol. 1989; 38(1):7–25.

41. Duan R-Y, Yang L-M, Lv T, Wu G-L, Huang M-Y. The complete chloroplast genome sequence of Pinus dabeshanensis. Conservation Genetics Resources. 2016; 8(4):395–7.

42. Zheng W, Chen J, Hao Z, Shi J. Comparative Analysis of the Chloroplast Genomic Information of Cunninghamia lanceolata (Lamb.) Hook with Sibling Species from the Genera Cryptomeria D. Don, Taiwania Hayata, and Calocedrus Kurz. Int J Mol Sci. 2016; 17(7):1084.

43. Asaf S, Waqas M, Khan AL, Khan MA, Kang S-M, Imran QM, et al. The Complete Chloroplast Genome of Wild Rice (Oryza minuta) and Its Comparison to Related Species. Front Plant Sci. 2017; 8(304). https://doi.org/10.3389/fpls.2017.00304 PMID: 28326093

44. Steane DA. Complete nucleotide sequence of the chloroplast genome from the Tasmanian blue gum, Eucalyptus globulus (Myrtaceae). DNA Res. 2005; 12(3):215–20. Epub 2005/11/24. https://doi.org/10.1093/dnares/dsi006 PMID: 16303753.

45. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016; 107(1):1–8. Epub 2015/11/12. https://doi.org/10.1016/j.ygeno.2015.11.003 PMID: 26554401

46. Lucas SJ, Akpınar BA, Šimková H, Kubaláková M, Doležel J, Budak H. Next-generation sequencing of flow-sorted wheat chromosome 5D reveals lineage-specific translocations and widespread gene duplications. BMC genomics. 2014; 15(1):1080.

47. Akpinar BA, Yuce M, Lucas S, Vrána J, Burešová V, Doležel J, et al. Molecular organization and comparative analysis of chromosome 5B of the wild wheat ancestor Triticum dicoccoides. Scientific reports. 2015; 5.

48. Akpinar BA, Lucas SJ, Vrána J, Doležel J, Budak H. Sequencing chromosome 5D of Aegilops tauschii and comparison with its allopolyploid descendant bread wheat (Triticum aestivum). Plant biotechnology journal. 2015; 13(6):740–52. https://doi.org/10.1111/pbi.12302 PMID: 25516153

49. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017.

50. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 1986; 5.

51. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 2016; 17. Artn 134 https://doi.org/10.1186/s13059-016-1004-2 PMID: 27339192

52. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004; 20(17):3252–5. https://doi.org/10.1093/bioinformatics/bth352 PMID: 15180927

53. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005; 33:W686–W9. https://doi.org/10.1093/nar/gki366 PMID: 15980563

54. Lohse M, Drechsel O, Bock R. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet. 2007; 52(5–6):267–74. https://doi.org/10.1007/s00294-007-0161-y PMID: 17957369

55. Kumar S, Nei M, Dudley J, Tamura K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform. 2008; 9(4):299–306. Epub 2008/04/18. https://doi.org/10.1093/bib/bbn017 PMID: 18417537.

56. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004; 32:W273–W9. https://doi.org/10.1093/nar/gkh458 PMID: 15215394

57. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001; 29(22):4633–42. https://doi.org/10.1093/nar/29.22.4633 PMID: 11713313

58. Kraemer L, Beszteri B, Gäbler-Schwarz S, Held C, Leese F, Mayer C, et al. STAMP: Extensions to the STADEN sequence analysis package for high throughput interactive microsatellite marker design. Bmc Bioinformatics. 2009; 10(1):41. https://doi.org/10.1186/1471-2105-10-41 PMID: 19183437

59. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27(2):573–80. Epub 1998/12/24. PMID: 9862982.

60. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013; 30(4):772–80. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

61. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980; 16(2):111–20. Epub 1980/12/01. PMID: 7463489.

62. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25(11):1451–2. https://doi.org/10.1093/bioinformatics/btp187 PMID: 19346325

63. Wicke S, Schneeweiss GM, dePamphilis CW, Muller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 2011; 76(3–5):273–97. https://doi.org/10.1007/s11103-011-9762-4 PMID: 21424877

64. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003; 19(12):1572–4. Epub 2003/08/13. PMID: 12912839.

65. Swofford DL. Paup—a Computer-Program for Phylogenetic Inference Using Maximum Parsimony. J Gen Physiol. 1993; 102(6):A9–A.

66. Wu Z, Tembrock LR, Ge S. Are Differences in Genomic Data Sets due to True Biological Variants or Errors in Genome Assembly: An Example from Two Chloroplast Genomes. Plos One. 2015; 10(2): e0118019. https://doi.org/10.1371/journal.pone.0118019 PMID: 25658309

67. Asaf S, Khan AL, Khan AR, Waqas M, Kang S-M, Khan MA, et al. Complete chloroplast genome of Nicotiana otophora and its comparison with related species. Front Plant Sci. 2016; 7. https://doi.org/10.3389/fpls.2016.00843 PMID: 27379132

68. Celiński K, Kijak H, Barylski J, Grabsztunowicz M, Wojnicka-Półtorak A, Chudzińska E. Characterization of the complete chloroplast genome of Pinus uliginosa (Neumann) from the Pinus mugo complex. Conservation Genetics Resources. 2016:1–4.

69. Li ZH, Qian ZQ, Liu ZL, Deng TT, Zu YM, Zhao P, et al. The complete chloroplast genome of Armand pine Pinus armandii, an endemic conifer tree species to China. Mitochondr DNA. 2016; 27(4):2635–6. https://doi.org/10.3109/19401736.2015.1041130 PMID: 26024147

70. Hildebrand M, Hallick RB, Passavant CW, Bourque DP. Trans-splicing in chloroplasts: the rps 12 loci of Nicotiana tabacum. Proceedings of the National Academy of Sciences. 1988; 85(2):372–6.

71. Chen J, Hao Z, Xu H, Yang L, Liu G, Sheng Y, et al. The complete chloroplast genome sequence of the relict woody plant Metasequoia glyptostroboides Hu et Cheng. Front Plant Sci. 2015; 6:447. Epub 2015/07/03. https://doi.org/10.3389/fpls.2015.00447 PMID: 26136762.

72. Qian J, Song J, Gao H, Zhu Y, Xu J, Pang X. The complete chloroplast genome sequence of the medicinal plant Salvia miltiorrhiza. Plos One. 2013; 8. https://doi.org/10.1371/journal.pone.0057607 PMID: 23460883

73. Morton BR. Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J Mol Evol. 1998; 46(4):449–59. https://doi.org/10.1007/PI00006325 PMID: 9541540

74. Nie XJ, Lv SZ, Zhang YX, Du XH, Wang L, Biradar SS, et al. Complete Chloroplast Genome Sequence of a Major Invasive Species, Crofton Weed (Ageratina adenophora). Plos One. 2012; 7(5). ARTN e36869 https://doi.org/10.1371/journal.pone.0036869 PMID: 22606302

75. Kumar S, Hahn FM, McMahan CM, Cornish K, Whalen MC. Comparative analysis of the complete sequence of the plastid genome of Parthenium argentatum and identification of DNA barcodes to differentiate Parthenium species and lines. Bmc Plant Biol. 2009; 9. Artn 131 https://doi.org/10.1186/1471-2229-9-131 PMID: 19917140

76. Raubeson LA, Jansen RK. Chloroplast genomes of plants. Plant diversity and evolution: genotypic and phenotypic variation in higher plants. 2005; 45.

77. Palmer JD. Plastid chromosomes: structure and evolution. The molecular biology of plastids. 1991; 7:5–53.

78. Wu CS, Chaw SM. Highly rearranged and size-variable chloroplast genomes in conifers II clade (cupressophytes): evolution towards shorter intergenic spacers. Plant Biotechnol J. 2014; 12(3):344–53. https://doi.org/10.1111/pbi.12141 PMID: 24283260

79. Hirao T, Watanabe A, Kurita M, Kondo T, Takata K. Complete nucleotide sequence of the Cryptomeria japonica D. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. Bmc Plant Biol. 2008; 8. https://doi.org/10.1186/1471-2229-8-70 PMID: 18570682

80. Shimda H, Sugiuro M. Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucleic Acids Res. 1991; 19(5):983–95. PMID: 1708498

81. Umesono K, Inokuchi H, Shiki Y, Takeuchi M, Chang Z, Fukuzawa H, et al. Structure and organization of Marchantia polymorpha chloroplast genome: II. Gene organization of the large single copy region from rps' 12 to atpB. J Mol Biol. 1988; 203(2):299–331. PMID: 2974085

82. Downie SR, Palmer JD. Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. Molecular systematics of plants: Springer; 1992. p. 14–35.

83. Doyle JJ, Doyle JL, Palmer JD. Multiple independent losses of two genes and one intron from legume chloroplast genomes. Syst Bot. 1995:272–94.

84. Johansson JT. There large inversions in the chloroplast genomes and one loss of the chloroplast generps16 suggest an early evolutionary split in the genusAdonis (Ranunculaceae). Plant Syst Evol. 1999; 218(1):133–43.

85. Saski C, Lee S-B, Daniell H, Wood TC, Tomkins J, Kim H-G, et al. Complete chloroplast genome sequence of Glycine max and comparative analyses with other legume genomes. Plant Mol Biol. 2005; 59(2):309–22. https://doi.org/10.1007/s11103-005-8882-0 PMID: 16247559

86. Tsuji S, Ueda K, Nishiyama T, Hasebe M, Yoshikawa S, Konagaya A, et al. The chloroplast genome from a lycophyte (microphyllophyte), Selaginella uncinata, has a unique inversion, transpositions and many gene losses. J Plant Res. 2007; 120(2):281–90. https://doi.org/10.1007/s10265-006-0055-y PMID: 17297557

87. Kugita M, Kaneko A, Yamamoto Y, Takeya Y, Matsumoto T, Yoshinaga K. The complete nucleotide sequence of the hornwort (Anthoceros formosae) chloroplast genome: insight into the earliest land plants. Nucleic Acids Res. 2003; 31(2):716–21. PMID: 12527781

88. Sugiura C, Sugita M. Plastid transformation reveals that moss tRNAArg-CCG is not essential for plastid function. The Plant Journal. 2004; 40(2):314–21. https://doi.org/10.1111/j.1365-313X.2004.02202.x PMID: 15447656

89. Fang MF, Wang YJ, Zu YM, Dong WL, Wang RN, Deng TT, et al. The complete chloroplast genome of the Taiwan red pine Pinus taiwanensis (Pinaceae). Mitochondr DNA. 2016; 27.

90. Kanno A, Hirai A. A transcription map of the chloroplast genome from rice (Oryza sativa). Curr Genet. 1993; 23(2):166–74. PMID: 8381719

91. Clarke AK, Gustafsson P, Lidholm JÅ. Identification and expression of the chloroplast clpP gene in the conifer Pinus contorta. Plant Mol Biol. 1994; 26(3):851–62. PMID: 7999999

92. Kohch T, Ogural Y, Umesono K, Yamada Y, Komano T, Ozeki H, et al. Ordered processing and splicing in a polycistronic transcript in liverwort chloroplasts. Curr Genet. 1988; 14(2):147–54. PMID: 2846189

93. Maier RM, Neckermann K, Igloi GL, Kössel H. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J Mol Biol. 1995; 251(5):614–28. https://doi.org/10.1006/jmbi.1995.0460 PMID: 7666415

94. Palmer JD, Thompson WF. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. Cell. 1982; 29(2):537–50. PMID: 6288261

95. Palmer JD, Thompson WF. Rearrangements in the chloroplast genomes of mung bean and pea. Proceedings of the National Academy of Sciences. 1981; 78(9):5533–7.

96. Lavin M, Doyle JJ, Palmer JD. Evolutionary significance of the loss of the chloroplast-DNA inverted repeat in the Leguminosae subfamily Papilionoideae. Evolution. 1990:390–402. https://doi.org/10.1111/j.1558-5646.1990.tb05207.x PMID: 28564377

97. Liston A. Use of the polymerase chain reaction to survey for the loss of the inverted repeat in the legume chloroplast genome. 1995.

98. Wu C-S, Wang Y-N, Hsu C-Y, Lin C-P, Chaw S-M. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. Genome Biol Evol. 2011; 3:1284–95. https://doi.org/10.1093/gbe/evr095 PMID: 21933779

99. Wu C-S, Lin C-P, Hsu C-Y, Wang R-J, Chaw S-M. Comparative chloroplast genomes of Pinaceae: insights into the mechanism of diversified genomic organizations. Genome Biol Evol. 2011; 3:309–19. https://doi.org/10.1093/gbe/evr026 PMID: 21402866

100. do Nascimento Vieira L, Faoro H, Rogalski M, de Freitas Fraga HP, Cardoso RLA, de Souza EM, et al. The complete chloroplast genome sequence of Podocarpus lambertii: genome structure, evolutionary aspects, gene content and SSR detection. Plos One. 2014; 9(3):e90618. https://doi.org/10.1371/journal.pone.0090618 PMID: 24594889

101. Yi X, Gao L, Wang B, Su YJ, Wang T. The Complete Chloroplast Genome Sequence of Cephalotaxus oliveri (Cephalotaxaceae): Evolutionary Comparison of Cephalotaxus Chloroplast DNAs and Insights into the Loss of Inverted Repeat Copies in Gymnosperms. Genome Biol Evol. 2013; 5(4):688–98. https://doi.org/10.1093/gbe/evt042 PMID: 23538991

102. Cosner ME, Jansen RK, Palmer JD, Downie SR. The highly rearranged chloroplast genome of Trachelium caeruleum (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. Curr Genet. 1997; 31(5):419–29. PMID: 9162114

103. Do HDK, Kim JS, Kim J-H. A trnI_CAU triplication event in the complete chloroplast genome of Paris verticillata M. Bieb. (Melanthiaceae, Liliales). Genome Biol Evol. 2014; 6(7):1699–706. https://doi.org/10.1093/gbe/evu138 PMID: 24951560

104. Cavalier-Smith T. Chloroplast evolution: Secondary symbiogenesis and multiple losses. Curr Biol. 2002; 12(2):R62–R4. https://doi.org/10.1016/S0960-9822(01)00675-3 PMID: 11818081

105. Asano T, Tsudzuki T, Takahashi S, Shimada H, Kadowaki K. Complete nucleotide sequence of the sugarcane (Saccharum officinarum) chloroplast genome: A comparative analysis of four monocot chloroplast genomes. DNA Res. 2004; 11(2):93–9. https://doi.org/10.1093/dnares/11.2.93 PMID: 15449542

106. Timme RE, Kuehl JV, Boore JL, Jansen RK. A comparative analysis of the Lactuca and Helianthus (Asteraceae) plastid genomes: Identification of divergent regions and categorization of shared repeats. Am J Bot. 2007; 94(3):302–12. https://doi.org/10.3732/ajb.94.3.302 PMID: 21636403

107. Gao L, Yi X, Yang YX, Su YJ, Wang T. Complete chloroplast genome sequence of a tree fern Alsophila spinulosa: insights into evolutionary changes in fern chloroplast genomes. Bmc Evol Biol. 2009; 9. Artn 130 https://doi.org/10.1186/1471-2148-9-130 PMID: 19519899

108. Echt CS, DeVerno L, Anzidei M, Vendramin G. Chloroplast microsatellites reveal population genetic diversity in red pine, Pinus resinosa Ait. Mol Ecol. 1998; 7(3):307–16.

109. Leclercq S, Rivals E, Jarne P. Detecting microsatellites within genomes: significant variation among algorithms. Bmc Bioinformatics. 2007; 8(1):125.

110. Rose O, Falush D. A threshold size for microsatellite expansion. Mol Biol Evol. 1998; 15(5):613–5. https://doi.org/10.1093/oxfordjournals.molbev.a025964 PMID: 9580993

111. Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, et al. Comparative chloroplast genomics: analyses including new sequences from the angiosperms Nuphar advena and Ranunculus macranthus. Bmc Genomics. 2007; 8. Artn 174 https://doi.org/10.1186/1471-2164-8-174 PMID: 17573971

112. Huotari T, Korpelainen H. Complete chloroplast genome sequence of Elodea canadensis and comparative analyses with other monocot plastid genomes. Gene. 2012; 508(1):96–105. https://doi.org/10.1016/j.gene.2012.07.020 PMID: 22841789

113. Powell W, Morgante M, Mcdevitt R, Vendramin GG, Rafalski JA. Polymorphic Simple Sequence Repeat Regions in Chloroplast Genomes—Applications to the Population-Genetics of Pines. P Natl Acad Sci USA. 1995; 92(17):7759–63.

114. Provan J, Corbett G, McNicol JW, Powell W. Chloroplast DNA variability in wild and cultivated rice (Oryza spp.) revealed by polymorphic chloroplast simple sequence repeats. Genome. 1997; 40 (1):104–10. https://doi.org/10.1139/G97-014 PMID: 9061917

115. Pauwels M, Vekemans X, Gode C, Frerot H, Castric V, Saumitou-Laprade P. Nuclear and chloroplast DNA phylogeography reveals vicariance among European populations of the model species for the study of metal tolerance, Arabidopsis halleri (Brassicaceae). New Phytol. 2012; 193(4):916–28. https://doi.org/10.1111/j.1469-8137.2011.04003.x PMID: 22225532

116. Powell W, Morgante M, Andre C, Mcnicol JW, Machray GC, Doyle JJ, et al. Hypervariable Microsatellites Provide a General Source of Polymorphic DNA Markers for the Chloroplast Genome. Curr Biol. 1995; 5(9):1023–9. https://doi.org/10.1016/S0960-9822(95)00206-5 PMID: 8542278

117. Li XW, Gao HH, Wang YT, Song JY, Henry R, Wu HZ, et al. Complete chloroplast genome sequence of Magnolia grandiflora and comparative analysis with related species. Sci China Life Sci. 2013; 56 (2):189–98. https://doi.org/10.1007/s11427-012-4430-8 PMID: 23329156

118. Hao Z, Cheng T, Zheng R, Xu H, Zhou Y, Li M, et al. The Complete Chloroplast Genome Sequence of a Relict Conifer Glyptostrobus pensilis: Comparative Analysis and Insights into Dynamics of Chloroplast Genome Rearrangement in Cupressophytes and Pinaceae. Plos One. 2016; 11(8):e0161809. https://doi.org/10.1371/journal.pone.0161809 PMID: 27560965

119. Yap JY, Rohner T, Greenfield A, Merwe M, McPherson H, Glenn W. Complete chloroplast genome of the wollemi pine (Wollemia nobilis): structure and evolution. Plos One. 2015; 10. https://doi.org/10.1371/journal.pone.0128126 PMID: 26061691

120. Kuang DY, Wu H, Wang YL, Gao LM, Zhang SZ, Lu L. Complete chloroplast genome sequence of Magnolia kwangsiensis (Magnoliaceae): implication for DNA barcoding and population genetics. Genome. 2011; 54(8):663–73. https://doi.org/10.1139/G11-026 PMID: 21793699

121. Birky CW Jr. Transmission genetics of mitochondria and chloroplasts. Annu Rev Genet. 1978; 12(1):471–512.

122. Birky CW Jr. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Annu Rev Genet. 2001; 35(1):125–48.

123. Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, et al. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. Annals of the Missouri Botanical Garden. 1993:528–80.

124. McCauley DE. The use of chloroplast DNA polymorphism in studies of gene flow in plants. Trends in ecology & evolution. 1995; 10(5):198–202.

125. Newton A, Allnutt T, Gillies A, Lowe A, Ennos R. Molecular phylogeography, intraspecific variation and the conservation of tree species. Trends in ecology & evolution. 1999; 14(4):140–5.

126. Provan J, Powell W, Hollingsworth PM. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends in ecology & evolution. 2001; 16(3):142–7.

127. Petit RJ, Aguinagalde I, de Beaulieu J-L, Bittkau C, Brewer S, Cheddadi R, et al. Glacial refugia: hotspots but not melting pots of genetic diversity. Science. 2003; 300(5625):1563–5. https://doi.org/10.1126/science.1083264 PMID: 12791991

128. Mogensen HL. The hows and whys of cytoplasmic inheritance in seed plants. Am J Bot. 1996:383–404.

129. Ennos R. Estimating the relative rates of pollen and seed migration among plant populations. Heredity. 1994; 72(3):250–9.

130. Hu X-S, Ennos R. On estimation of the ratio of pollen to seed flow among plant populations. Heredity. 1997; 79(5):541–52.

131. Petit RJ, Duminil J, Fineschi S, Hampe A, Salvini D, Vendramin GG. Invited review: comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. Mol Ecol. 2005; 14(3):689–701.

132. Neale D, Sederoff R. Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in loblolly pine. TAG Theoretical and Applied Genetics. 1989; 77(2):212–6. https://doi.org/10.1007/BF00266189 PMID: 24232531

133. Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proceedings of the National Academy of Sciences. 2010; 107(10):4623–8. https://doi.org/10.1073/pnas.0907801107 PMID: 20176954

134. Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH. The chloroplast genome of Nymphaea alba: Whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol. 2004; 21(7):1445–54. https://doi.org/10.1093/molbev/msh147 PMID: 15084683

135. Hohmann N, Schmickl R, Chiang T-Y, Lučanová M, Kolář F, Marhold K, et al. Taming the wild: resolving the gene pools of non-model Arabidopsislineages. Bmc Evol Biol. 2014; 14(1):1–21. https://doi.org/10.1186/s12862-014-0224-x PMID: 25344686

136. Gernandt DS, López GG, García SO, Liston A. Phylogeny and classification of Pinus. Taxon. 2005; 54 (1):29–42.

137. Miller CN Jr. Silicified cones and vegetative remains of Pinus from Eocene of British Columbia. 1973.

138. Willyard A, Syring J, Gernandt DS, Liston A, Cronn R. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for Pinus. Mol Biol Evol. 2007; 24(1):90–101. https://doi.org/10.1093/molbev/msl131 PMID: 16997907

139. Millar C. Early evolution of pines. Ecology and biogeography of Pinus. 1998:69–91.

140. Eckert AJ, Hall BD. Phylogeny, historical biogeography, and patterns of diversification for Pinus (Pinaceae): phylogenetic tests of fossil-based hypotheses. Mol Phylogenet Evol. 2006; 40(1):166–82. https://doi.org/10.1016/j.ympev.2006.03.009 PMID: 16621612

141. Shoemaker R, Hatfield P, Palmer R, Atherly A. Chloroplast DNA variation in the genus Glycine subgenus Soja. J Hered. 1986; 77(1):26–30.

142. Close P, Shoemaker R, Keim P. Distribution of restriction site polymorphism within the chloroplast genome of the genus Glycine, subgenus Soja. Theor Appl Genet. 1989; 77(6):768–76. https://doi.org/10.1007/BF00268325 PMID: 24232890

143. Hirata T, Abe J, Shimamoto Y. RFLPs of chloroplast and mitochondrial genomes in summer and autumn maturing cultivar groups of soybean in Kyushu district of Japan. Soybean genetics newsletter (USA). 1996.

144. Lee D, Caha C, Specht J, Graef G. Chloroplast DNA evidence for non-random selection of females in an outcrossed population of soybeans [Glycine max (L.)]. Theor Appl Genet. 1992; 85(2–3):261–8. https://doi.org/10.1007/BF00222868 PMID: 24197313

145. Shimamoto Y, Hasegawa A, Abe J, Ohara M, Mikami T. Glycine soja germplasm in Japan: isozyme and chloroplast DNA variation. Soybean genetics newsletter-US Department of Agriculture, Agricultural Research Service (USA). 1992.

146. Abe J, Hasegawa A, Fukushi H, Mikami T, Ohara M, Shimamoto Y. Introgression between wild and cultivated soybeans of Japan revealed by RFLP analysis for chloroplast DNAs. Economic Botany. 1999; 53(3):285–91.

147. Khush GS. Origin, dispersal, cultivation and variation of rice. Plant Mol Biol. 1997; 35(1–2):25–34. https://doi.org/10.1023/A:1005810616885 PMID: 9291957

148. Wang X-R, Tsumura Y, Yoshimaru H, Nagasaka K, Szmidt AE. Phylogenetic relationships of Eurasian pines (Pinus, Pinaceae) based on chloroplast rbcL, matK, rpl20-rps18 spacer, and trnV intron sequences. Am J Bot. 1999; 86(12):1742–53. PMID: 10602767

149. Geada López G, Kamiya K, Harada K. Phylogenetic relationships of Diploxylon pines (subgenus Pinus) based on plastid sequence data. Int J Plant Sci. 2002; 163(5):737–47.

150. Syring J, Willyard A, Cronn R, Liston A. Evolutionary relationships among Pinus (Pinaceae) subsections inferred from multiple low-copy nuclear loci. Am J Bot. 2005; 92(12):2086–100. https://doi.org/10.3732/ajb.92.12.2086 PMID: 21646125

151. Parkinson CL, Adams KL, Palmer JD. Multigene analyses identify the three earliest lineages of extant flowering plants. Curr Biol. 1999; 9(24):1485–91. PMID: 10607592

152. Soltis PS, Soltis DE, Chase MW. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature. 1999; 402(6760):402–4. https://doi.org/10.1038/46528 PMID: 10586878

153. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 2003; 425(6960):798–804. https://doi.org/10.1038/nature02053 PMID: 14574403

154. Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 2005; 6(5):361–75. https://doi.org/10.1038/nrg1603 PMID: 15861208

155. Jeffroy O, Brinkmann H, Delsuc F, Philippe H. Phylogenomics: the beginning of incongruence? Trends Genet. 2006; 22(4):225–31. https://doi.org/10.1016/j.tig.2006.02.003 PMID: 16490279

156. Wambugu P, Brozynska M, Furtado A, Waters D, Henry R. Relationships of wild and domesticated rices (Oryza AA genome species) based upon whole chloroplast genome sequences. Sci Rep. 2015; 5. https://doi.org/10.1038/srep13957 PMID: 26355750

157. Gao C-W, Gao L-Z. The complete chloroplast genome sequence of wild soybean, Glycine soja. Conservation Genetics Resources. 2016:1–3.

158. Niu S-H, Li Z-X, Yuan H-W, Chen X-Y, Li Y, Li W. Transcriptome characterisation of Pinus tabuliformis and evolution of genes in the Pinus phylogeny. Bmc Genomics. 2013; 14(1):263.