



# cpSNP discovery and genotyping for a *Pinus taeda* breeding population with targeted comparison to related conifers

Ling Wang<sup>1,2</sup>, Jipeng Mao<sup>1,2</sup>, Kaibin Jiang<sup>1,2</sup>, Zhengyu Wu<sup>1,2</sup>, Chunxin Liu<sup>1,2</sup>, Huagong Ning<sup>3</sup>, Haibiao Zhou<sup>4</sup>, Jiehu Chen<sup>4</sup>, Shaowei Huang<sup>1,2</sup> and Tianyi Liu<sup>1,2</sup>

<sup>1</sup> College of Forestry and Landscape Architecture, South China Agriculture University, Guangzhou, China

<sup>2</sup> Guangdong Key Laboratory for Innovative Development and Utilization of Forest Plant Germplasm, Guangzhou, China

<sup>3</sup> Yingde Research Institute of Forestry, Yingde, China

<sup>4</sup> Science Corporation of Gene, Guangzhou, China

## ABSTRACT

*Pinus taeda* (Loblolly pine) is the most important commercial tree species in the southern United States and a significant non-native plantation species in China. Its genetic improvement program has been implemented in South China for 30 years. In this study, the chloroplast (cp) genome of *P. taeda* was sequenced, assembled, and compared with other available chloroplast genomes of Pinaceae using BLAST. Codon usage among 33 species of Pinaceae was analyzed using the relative synonymous codon usage (RSCU) value. The results were then visualized using the pheatmap v1.0.10 in R. The rates of nonsynonymous (Ka) and synonymous (Ks) substitutions in the chloroplast genomes among five species of *Pinus* were estimated using the seqinr package in R. Additionally, selected single nucleotide polymorphisms (SNPs) were used to genotype 33 individuals from the *P. taeda* breeding population. The *P. taeda* cp genome is 121,530 bp, with certain regions (e.g., *ycf1* and *ycf2*) showing lower sequence conservation compared to other Pinaceae species. Codon usage analysis revealed that codons ending in G or C were not prevalently used, with significant differences in natural selection pressure on chloroplast genes between three species (*P. taiwanensis*, *P. thunbergii*, and *P. koraiensis*) and the other 30 species. Phylogenomic analysis using 36 cp genomes (representing 25 species) resolved *Pinus* into two subgenera, with *P. taeda* clustered with *P. rigida* within the diploxylon pines. Notably, *ycf1*-based phylogenetic analysis recovered a topology highly congruent (normalized RF = 0.15) with whole-plastome phylogenies. This study validates the single-copy gene *ycf1* as a robust and low-cost phylogenetic marker for conifer genus-level reconstruction. The identified SNPs provide valuable molecular markers for genotyping individuals in *P. taeda* breeding programs, supporting germplasm characterization and management.

Submitted 29 May 2024

Accepted 27 August 2025

Published 6 October 2025

Corresponding author

Tianyi Liu, tianyiliu@scau.edu.cn

Academic editor

Douglas Domingues

Additional Information and  
Declarations can be found on  
page 14

DOI 10.7717/peerj.20092

© Copyright  
2025 Wang et al.

Distributed under  
Creative Commons CC-BY 4.0

OPEN ACCESS

**Subjects** Bioinformatics, Genomics, Plant Science

**Keywords** *Pinus taeda*, Chloroplast genome, DNA sequence, SNP analysis, Breeding population

## INTRODUCTION

The study of the chloroplast (cp) genome has become a valuable and widely used approach for evolutionary, taxonomic, and phylogenetic analyses of plants. This is due to its simple structure, highly conserved sequence, and single parental (maternal or paternal) inheritance characters (Wolfe, Li & Sharp, 1987; Daniell et al., 2016; Jakobsson et al., 2007; Nock et al., 2011; Wu & Chaw, 2014; Mao et al., 2023). Unlike most angiosperms, cpDNA is paternally inherited in all conifers (Neale, Marshall & Sederoff, 1989; Szmidt, Aldén & Hällgren, 1987; Wagner et al., 1987; Neale, Wheeler & Allard, 1986). Conifers hold immense ecological and economic value. Understanding the organization of their chloroplast genomes is therefore significant for phylogenetic studies and resolving evolutionary relationships. This knowledge also supports efforts to enhance the commercial use of conifers.

Loblolly pine is the most important commercial tree species in the southern United States due to its fast growth, desirable wood properties, and broad adaptability (McKeand et al., 2003). In China, it is a major non-native plantation species, especially in southern regions (Gwaze, Byram & Lowe, 2001). Genetic improvement programs for *P. taeda* in China began in the early 1980s. In the 1990s, comprehensive progeny and provenance tests were initiated, leading to advances in selective breeding (Zhong, Chen & Huang, 1995). Since 2002, China has implemented a breeding strategy (El-Kassaby & Lstibůrek, 2009). The paternal inheritance of the conifer cp genome enables paternity verification when the maternal parent is known, thereby facilitating precision breeding programs. Although cp genomes are generally conserved, recent studies have identified variable regions. For example, nine mutational loci were detected in *ycf1* of *P. taeda* (Li, 2012). Recent studies have demonstrated that *ycf1* exhibits high variability (Jiang et al., 2017; Li et al., 2020; Chen et al., 2020; Zhang et al., 2023). Such variability, while contrasting with overall cpDNA conservation, offers valuable markers for fine-scale genetic analyses.

While the complete chloroplast genome of *P. taeda* has been assembled and its broad structure and placement resolved (Asaf et al., 2018), three knowledge gaps remain. First, there is a need for population-level chloroplast single nucleotide polymorphism (cpSNP) discovery within breeding material. Second, a systematic comparison of codon-usage bias across Pinaceae is lacking. Third, the single-copy *ycf1* gene has not been empirically assessed as a cost-effective phylogenetic marker. To address these gaps, we generated a high-quality *P. taeda* cp genome and compared codon-usage patterns across 33 Pinaceae species. We quantified Ka/Ks in five *Pinus* species, identified 71 novel cpSNPs in 33 breeding individuals, and performed phylogenomic analyses using both whole plastomes and *ycf1* across 36 samples (25 species). The resulting cpSNP panel differentiates 72% of breeding individuals, and *ycf1* recapitulates the whole-genome tree (normalized RF = 0.15), demonstrating its utility for low-cost conifer phylogenetics and immediate application in germplasm management.



# MATERIALS & METHODS

## DNA extraction and sequencing

Fresh needles of *P. taeda* were collected from the Yingde Research Institute of Forestry in Guangdong Province, P.R. China (24°15'N, 113°25'E). Total cpDNA was extracted from approximately 100 g needles using a high-ionic-strength medium ([Bookjans, Stummann & Henningsen, 1984](#)). The integrity and purity of DNA were assessed by 1% agarose gel electrophoresis. The purity of the samples was further evaluated by measuring the OD260/280 ratio with a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific). The DNA concentration was quantified by an Invitrogen Qubit fluorometer. Libraries were prepared in accordance with the manufacturer's instructions with an average insert size of 350 bp. Purified cpDNA was sequenced on an Illumina HiSeq 2500 platform (Science Corporation of Gene, Guangzhou, China) under a paired-end 100 bp mode. Raw read quality was then assessed using fastp (v0.23.2) with the following parameters: a cut window size of 8 nucleotides and a qualified quality Phred score threshold of 20.

## De novo CP genome assembly

The chloroplast assembly software SPAdes (v3.15.5; [Bankevich et al., 2012](#)) was used to assemble the chloroplast genome of loblolly pine, with K-mer length set to 79 and 97. The assembly's integrity was verified by mapping reads back to the consensus sequence using Burrows-Wheeler Aligner (BWA; v0.7.17-r1188) and SAMtools (v1.14). The *P. taeda* cp genome was compared with other available cp genomes of Pinaceae by using the CGView Comparison Tool (CCT; [Grant & Stothard, 2008](#)). Gene annotations were performed using the Clusters of Orthologous Groups (COGs) database. The Basic Local Alignment Search Tool (BLAST) was employed to align the other genomes to *P. taeda*, facilitating the identification of homologous sequences. The complete annotated genome was shown as a circular map using OrganellarGenomeDRAW (OGDRAW; [Greiner, Lehwark & Bock, 2019](#)). The adenine (A) and thymine (T) distributions were measured based on the results of the AT-skew equation, as follows:  $AT-skew = (A - T) / (A + T)$ .

## Codon usage

Codon usage, quantified as relative synonymous codon usage (RSCU) values ([Sharp & Li, 1987](#)), was determined for all protein-coding genes across the 33 Pinaceae species. Statistical analyses of RSCU distributions were performed, and results were visualized as a heatmap using the pheatmap package (v1.0.10) in R ([R Core Team, 2022](#)).

## Ka/Ks analysis

The rates of nonsynonymous (Ka) and synonymous (Ks) substitutions within the chloroplast genomes of five *Pinus* species were estimated using seqinR (v3.4-5; [Gouy et al., 1984](#); [R Core Team, 2022](#)). Histograms of these substitution rates were then created using ggplot2 (v3.0.0) to visualize the distribution of Ka and Ks values.

## Single nucleotide polymorphisms analysis

There were 32 *P. taeda* cpDNA samples, selected from the Yingde Research Institute of Forestry in Guangdong Province, P.R. China, mixed and sequenced by the Illumina HiSeq

**Table 1** Primers for Eighteen SNPs in *Pinus taeda*.

| No. | Name of Primer | Location | Length (bp) | Sequence (5'-3')      | Length of Primer (bp) | GC%    | PCR Annealing Temperature | SNP                                                                                                                                                        |
|-----|----------------|----------|-------------|-----------------------|-----------------------|--------|---------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1   | F1             | 218      | 540         | TTCCCACAACTTTCATACCA  | 20                    | 40.00% | 51                        | ycf1_98314                                                                                                                                                 |
|     | R1             | 757      |             | CTTTAGGATAAGCGGGTATT  | 20                    | 40.00% |                           |                                                                                                                                                            |
| 2   | F2             | 277      | 423         | GTCTGATTGGACCATTTGTA  | 20                    | 40.00% | 51                        | ycf2_119153                                                                                                                                                |
|     | R2             | 699      |             | TCAATCTTTATGGGTCCTAC  | 20                    | 40.00% |                           |                                                                                                                                                            |
| 3   | F3             | 274      | 577         | TGGTGGAGATGGTGAAGATG  | 20                    | 50.00% | 53                        | ycf1_101076, ycf1_101085, ycf1_101094, ycf1_101099, ycf1_101101, ycf1_101103, ycf1_101112, ycf1_101121, ycf1_101130, ycf1_101139, ycf1_101148, ycf1_101157 |
|     | R3             | 850      |             | AAAGGCCATTAGACTCAGGT  | 20                    | 45.00% |                           |                                                                                                                                                            |
| 4   | F4             | 306      | 576         | GAATATCTAAACCCTGGACT  | 20                    | 40.00% | 51                        | rpoC1_23167                                                                                                                                                |
|     | R4             | 881      |             | ATTAGCTTCTCCCGAACAGA  | 20                    | 45.00% |                           |                                                                                                                                                            |
| 5   | F5             | 251      | 520         | CTTCTCATTTCCAATCCCTG  | 20                    | 45.00% | 53                        | IGS_23593                                                                                                                                                  |
|     | R5             | 770      |             | GAAATGGAAGTTTGGGCTCT  | 20                    | 45.00% |                           |                                                                                                                                                            |
| 6   | F6             | 243      | 705         | GACCCACATAAGAACAAACG  | 20                    | 45.00% | 53                        | IGS_50842, IGS_50999                                                                                                                                       |
|     | R6             | 947      |             | CGGATATGTCCATGATCACTA | 22                    | 40.91% |                           |                                                                                                                                                            |

2500 platform with a paired-end 100 bp sequencing strategy. The resulting sequences were aligned against the known reference sequence of *P. taeda* (GenBank accession no. [NC\\_021440.1](https://www.ncbi.nlm.nih.gov/nuccore/NC_021440.1)). Subsequently, 18 single nucleotide polymorphisms (SNPs) were selected for the detection of 33 individuals in the breeding population of *P. taeda*. Primer-BLAST (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/>) was used for the primer design (Table 1), and the PCR products were sequenced using the Sanger method.

## Phylogenetic analysis

To resolve the phylogenetic position of *P. taeda*, phylogenetic analyses were conducted using both complete cp genomes and the *ycf1* gene sequences from 36 samples. Sequence data were retrieved from NCBI (<https://www.ncbi.nlm.nih.gov/>), comprising 24 *Pinus* species (*P. rigida*, *P. jaliscana*, *P. oocarpa*, *P. elliottii*, *P. caribaea*, *P. taiwanensis*, *P. thunbergii*, *P. densata*, *P. wallichiana*, *P. strobiformis*, *P. bungeana*, *P. monophylla*, *P. nelsonii*, plus *P. taeda*, *P. greggii*, *P. contorta*, *P. sylvestris*, *P. tabuliformis*, *P. krempfii*, *P. strobus*, *P. massoniana*, *P. lambertiana*, *P. sibirica*, and *P. koraiensis* represented by two samples each) and the outgroup *Cedrus deodara*.

Both datasets were processed identically: multiple sequence alignment used MAFFT v7 (Katoh, Rozewicki & Yamada, 2019), and Maximum Likelihood trees were inferred in MEGA X (Kumar et al., 2018) with 1,000 bootstrap replicates.

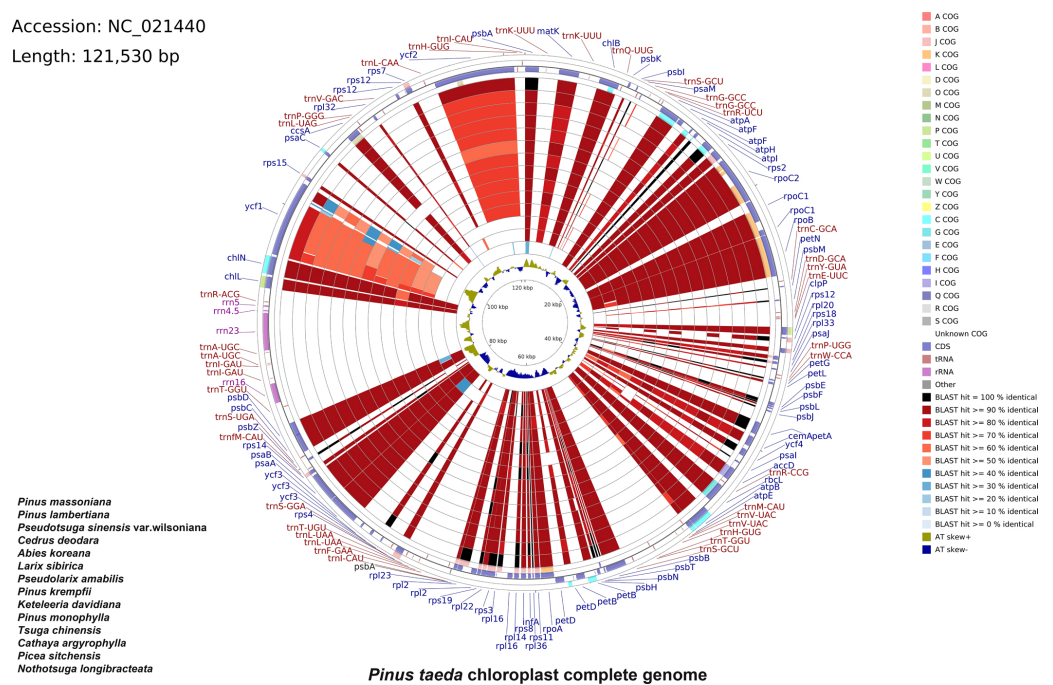
Topological congruence between whole-genome and *ycf1*-only trees was assessed by calculating the normalized Robinson-Foulds (RF) distance using ape (v5.8.1; Paradis & Schliep, 2019) in R v4.5.1 (R Core Team, 2022).

## RESULTS

### Genome comparison analysis

A total of 33,022,566 base pairs (bp) of raw reads with 38.41% GC content were generated with a paired-end 100 bp read length. The raw sequences of the *P. taeda* cp genome

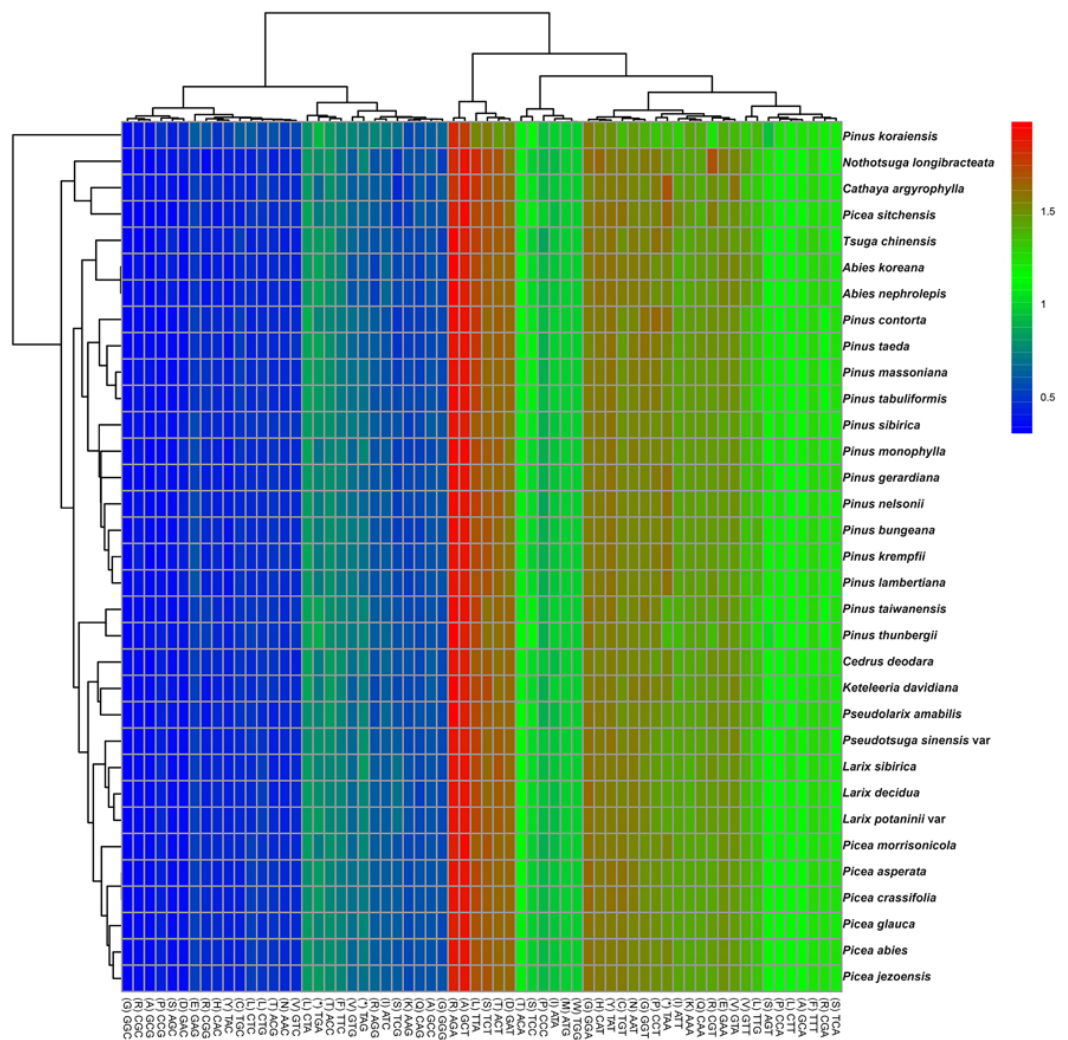
Accession: NC\_021440  
Length: 121,530 bp



**Figure 1** Comparative chloroplast genome analysis of fourteen Pinaceae species against *Pinus taeda*. From the outer to the inner color ring: *Pinus massoniana*, *Pinus lambertiana*, *Pseudotsuga sinensis* var. *wilsoniana*, *Cedrus deodara*, *Abies koreana*, *Larix sibirica*, *Pseudolarix amabilis*, *Pinus krempfii*, *Keteleeria davidiana*, *Pinus monophylla*, *Tsuga chinensis*, *Cathaya argyrophylla*, *Picea sitchensis*, and *Nothotsuga longibracteata*. Alignment method: BLASTn alignment of all genomes to *Pinus taeda*. Similarity scores: black (100%), red (50–99%), blue (<50%). Four outer narrow rings: Protein-coding gene positions (*Pinus taeda* chloroplast genome reference). Color code: Clusters of Orthologous Groups (COG). Innermost ring: AT skew of *P. taeda* (positive [+] = A>T; negative [–] = A).  
Full-size [DOI: 10.7717/peerj.20092/fig-1](https://doi.org/10.7717/peerj.20092/fig-1)

were deposited in GenBank with accession number [PRJNA1159385](https://doi.org/10.7717/peerj.20092), and the assembled sequences with accession number [NC\\_021440.1](https://doi.org/10.7717/peerj.20092). The size of the genome was 121,530 bp, with a GC content of 38.5% (Fig. 1). The positions of all the identified genes in the cp genome and their functional categorization are presented in Fig. 1.

The cp genomes of 14 species of Pinaceae (*P. massoniana*, *P. lambertiana*, *Pseudotsuga sinensis* var. *wilsoniana*, *Cedrus deodara*, *Abies koreana*, *Larix sibirica*, *Pseudolarix amabilis*, *P. krempfii*, *Keteleeria davidiana*, *P. monophylla*, *Tsuga chinensis*, *Cathaya argyrophylla*, *Picea sitchensis*, and *Nothotsuga longibracteata*) were selected for comparison with *P. taeda* by using CCT (Fig. 1). The sequence identity between *P. taeda* and other Pinaceae representatives was analyzed, revealing that certain regions were less conserved. For instance, *ycf1* showed sequence identity lower than 70%, while *ycf2* exhibited identities below 80%. Previous studies have confirmed that *ycf1* has an important role in the evolution and classification of *Pinus* (Daniell et al., 2016; Handy et al., 2011; Georgolopoulos, Parducci & Drouzas, 2016).



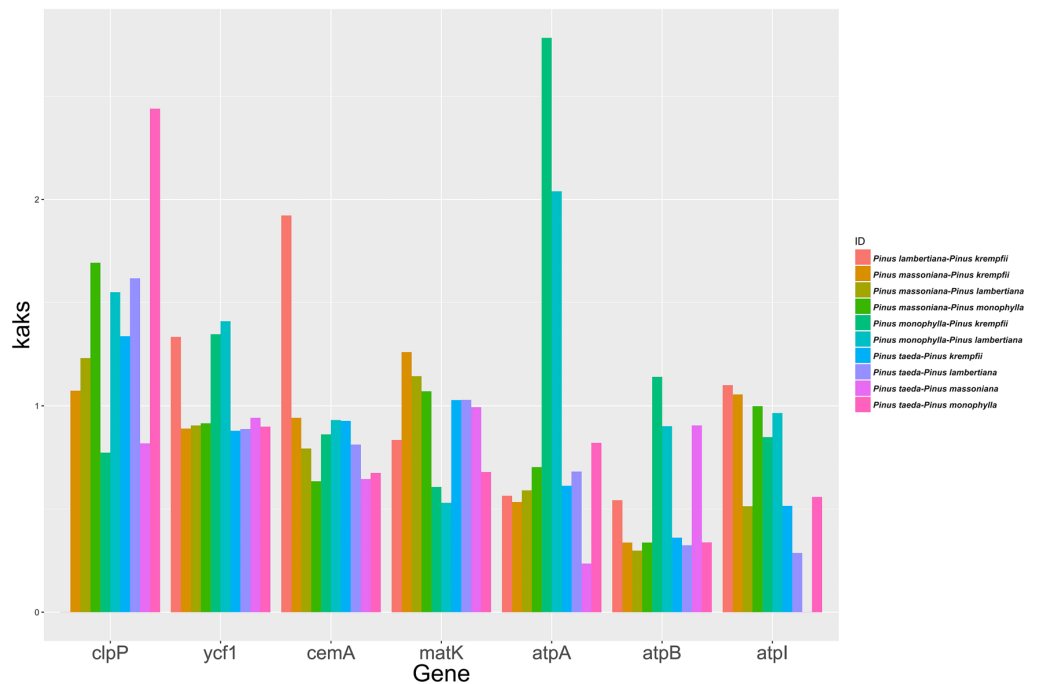
**Figure 2** Relative synonymous codon usage (RSCU) heatmap of protein-coding genes across 33 species. Color scale: RSCU values from low (blue, 0.5) to high (red, 2.0); x-x-axis: Codon patterns clustered by UPGMA (Lance & Williams, 1967; Euclidean distance).

Full-size DOI: 10.7717/peerj.20092/fig-2

## Codon usage

The codon usage patterns among the 71 distinct cp protein-coding genes in *P. taeda* were examined (Fig. 2). The 71 protein-coding genes consisted of 20,255 codons, with the majority ending in A or T and the remainder ending in C or G. Codons terminating in A or T exhibited significantly higher RSCU values compared to those ending in C or G, as detailed in Table 1. Statistical analysis of codon usage distributions across 33 Pinaceae species revealed infrequent use of codons ending with G or C, as shown in the heatmaps (Fig. 2). Similar patterns have been reported in other cp genomes (Delannoy et al., 2011; Raubeson et al., 2007; Ying et al., 2016).

Among the 33 analyzed Pinaceae species, codon usage patterns were largely conserved, except for *P. taiwanensis*, *P. thunbergii*, and *P. koraiensis*. Phylogenetic clustering revealed



**Figure 3** Ka/Ks ratios for the seven Chloroplast protein-coding genes among five *Pinus* species. [Full-size DOI: 10.7717/peerj.20092/fig-3](https://doi.org/10.7717/peerj.20092/fig-3)

that *P. taiwanensis* and *P. thunbergii* formed a distinct clade, divergent from other *Pinus* species. Additionally, *Picea* and *Abies* occupied phylogenetically distant positions both from *Pinus* and from each other. The clustering results were inconsistent with traditional classification results ([Gernandt et al., 2005](#)).

### Ka/Ks analysis

The study analyzed the molecular evolutionary characteristics of seven chloroplast genes (*clpP*, *ycf1*, *cemA*, *matK*, *atpA*, *atpB*, and *atpI*) from five species belonging to *Pinus*. The Ka/Ks ratios of the 10 species pairs of these genes are shown in [Fig. 3](#). Significant differences in the Ka/Ks ratios were found between the seven chloroplast genes. Ka/Ks >1 was detected in seven species pairs for *clpP* and in five species pairs for *matK*. For *ycf1* and *cemA*, Ka/Ks >1 was detected in three species pairs each, while the remaining species pairs had Ka/Ks ratios close to 1. For *atpA*, *atpB*, and *atpI*, Ka/Ks >1 was detected in only one or two species pairs, with the remaining species pairs having Ka/Ks ratios close to 1.

### SNP analysis of the chloroplast genome

Compared with the known reference sequence, 71 SNPs and four indels were identified from the 33 individuals in the breeding population of *P. taeda* ([Table 2](#)). Of these, 35 SNPs were located in the gene-coding regions of 11 genes, while the remaining SNPs were found in an intergenic region. The study found that 50 SNPs had mutation frequencies below 15%, which constituted 70.42% of the total number of SNPs. The lower mutation rate indicates that the cp genome of *P. taeda* was conservative. Notably, three SNP loci had



mutation rates between 30% and 50%, with two located in gene-coding regions (*ycf1* and *ycf2*). Specifically, *ycf1* contained 17 mutations (23.94% of the total mutation sites), and *ycf2* had three SNPs, including one with the highest mutation rate of 48%. *R poB* had three SNPs, two of which had the same mutation rate. *R poC2* had two SNPs with mutation rates at 24% and 3%. *R poC1* had two SNPs with a similar mutation rate, as did *psaA*. The intergenic regions between *matK* and *chlB*, *psbJ* and *petA*, and *rps15* and *psaC* contained 14 SNPs, eight of which had mutation rates higher than 10%.

### Individual distinction of the breeding population

A total of 18 SNPs were selected to detect 33 individuals from the breeding population of *P. taeda*. Of these, 13 SNPs were mapped to *ycf1*. Analysis revealed that 32 individuals exhibited SNPs across at 15 different loci, and 23 individuals could be distinguished based on these SNPs. Only one individual's genome matched the reference genome, as shown in Table 3. Among the 23 distinguished individuals, one additional SNP was identified, and seven SNPs were found to be the most prevalent.

Furthermore, three additional SNPs were found in 14 individuals, accounting for 60.87% of the distinguishable SNPs. A total of 13 individuals had two SNPs, which represented 39.40% of the total number of individuals (33). Compared to previous studies, the cpDNA markers had a comparatively high level of variability (Walter & Epperson, 2010). The chloroplast simple sequence repeats (cpSSRs) of *P. sylvestris* revealed a high level of intra-population polymorphism across the entire range (Semerikov et al., 2014). Additionally, the variation among populations accounted for 99% of the total variance (Pazouki et al., 2016).

Additionally, the remaining 10 individuals that could not be distinguished, each of which had only one or two SNPs, could be classified into four groups. Introducing more SNPs could potentially help distinguish these individuals. Specifically, individuals P043, P054, and P064 exhibited identical SNP profiles. Given their common origin, they may descend from the same paternal lineage. Similarly, individuals G09, G10, and 243 also had the same SNPs. G09 and G10 had the same origin, while 243 was selected from a different stand. The loblolly pine in China, which were all introduced from the southern United States, might be related to each other.

### Phylogenetic insights from cp Genomes and *ycf1*

Phylogenetic reconstruction using complete cp genomes resolved the 24 *Pinus* species and the outgroup *C. deodara* into two well-supported clades (Clade A and Clade B; Fig. 4A). The topology contained 33 nodes, with 26 nodes achieving maximal bootstrap support (100%), while the remaining nodes attaining values ranging from 64% to 99%. This bifurcation is consistent with the canonical classification proposed by Gernandt et al. (2005). Notably within Clade A, *P. taeda* and *P. rigida* converge into the same branch. *P. taeda* emerged as a focal lineage closely affiliated with *P. jaliscana*, *P. oocarpa*, *P. greggii*, *P. elliotii*, and *P. caribaea*, corroborating clade relationships reported by Duan et al. (2016) and Zeb et al. (2019); Zeb et al. (2022).

Analysis of *ycf1* gene sequences from the same chloroplast genomic resources recovered an overall congruent phylogenetic architecture (Fig. 4B), despite minor topological



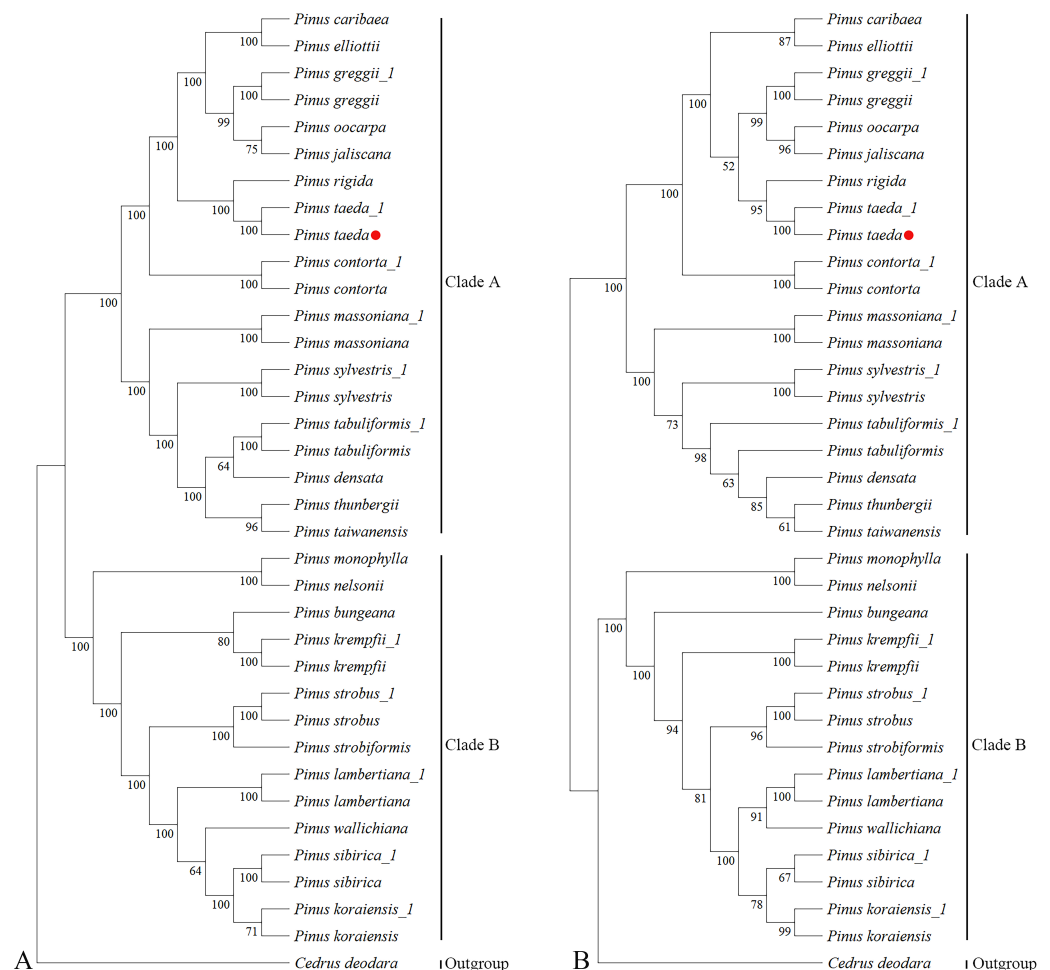
Table 2 Mutation sites of the *P. taeda* chloroplast genome.

| Coordinate | Ref | SNP/ Indels | Rate | Gene                 | Coordinate | Ref | SNP/ Indels | Rate | Gene                     | Coordinate | Ref | SNP/ Indels | Rate      | Gene                  |
|------------|-----|-------------|------|----------------------|------------|-----|-------------|------|--------------------------|------------|-----|-------------|-----------|-----------------------|
| 4409       | A   | G           | 0.1  | <i>matK~chlB</i>     | 39801      | A   | G           | 0.18 | <i>petA</i>              | 100910     | C   | -GATGGTGAA  | 0         | <i>ycf1</i>           |
| 4473       | C   | T           | 0.03 | <i>matK~chlB</i>     | 41670      | C   | T           | 0.1  | <i>psaI~accD</i>         | 101076     | T   | G           | 0.01      | <i>ycf1</i>           |
| 4609       | C   | T           | 0.04 | <i>matK~chlB</i>     | 41678      | C   | G           | 0.06 | <i>psaI~accD</i>         | 101085     | T   | G           | 0.04      | <i>ycf1</i>           |
| 4910       | G   | T           | 0.09 | <i>matK~chlB</i>     | 50842      | C   | T           | 0.06 | <i>trnH-GUG~trnT-GGU</i> | 101103     | T   | G           | 0.11      | <i>ycf1</i>           |
| 5514       | G   | A           | 0.27 | <i>matK~chlB</i>     | 58344      | C   | A           | 0.04 | <i>petD</i>              | 101112     | T   | G           | 0.07      | <i>ycf1</i>           |
| 5690       | G   | A           | 0.12 | <i>matK~chlB</i>     | 64071      | T   | C           | 0.04 | <i>rpl22</i>             | 101121     | T   | G           | 0.1       | <i>ycf1</i>           |
| 9984       | A   | C           | 0.23 | <i>trnG-GCC</i>      | 65304      | C   | A           | 0.15 | <i>rpl2</i>              | 101130     | T   | G           | 0.06      | <i>ycf1</i>           |
| 10349      | A   | C           | 0.06 | <i>trnG-GCC</i>      | 70068      | T   | G           | 0.04 | <i>trnT-UGU~rps4</i>     | 101139     | G   | T           | 0.4       | <i>ycf1</i>           |
| 11437      | G   | A           | 0.1  | <i>atpA</i>          | 71282      | A   | C           | 0.04 | <i>trnS-GGA~ycf3</i>     | 101157     | T   | G           | 0.09      | <i>ycf1</i>           |
| 19229      | G   | A           | 0.24 | <i>rpoC2</i>         | 71422      | C   | G           | 0.12 | <i>trnS-GGA~ycf3</i>     | 101166     | T   | G           | 0.02      | <i>ycf1</i>           |
| 20388      | C   | T           | 0.03 | <i>rpoC2</i>         | 74477      | T   | G           | 0.23 | <i>ycf3~psaA</i>         | 102292     | A   | C           | 0.36      | <i>ycf1~rps15</i>     |
| 22162      | A   | T           | 0.15 | <i>rpoC1</i>         | 74535      | A   | C           | 0.05 | <i>ycf3~psaA</i>         | 102658     | C   | T           | 0.06      | <i>rps15~psaC</i>     |
| 23394      | T   | G           | 0.1  | <i>rpoC1</i>         | 75726      | C   | T           | 0.04 | <i>psaA</i>              | 103514     | C   | T           | 0.06      | <i>rps15~psaC</i>     |
| 24094      | C   | T           | 0.04 | <i>rpoB</i>          | 76044      | A   | G           | 0.04 | <i>psaA</i>              | 103620     | G   | T           | 0.18      | <i>rps15~psaC</i>     |
| 24666      | G   | A           | 0.04 | <i>rpoB</i>          | 81006      | T   | G           | 0.04 | <i>trnfM-CAU~psbZ</i>    | 103785     | G   | A           | 0.07      | <i>rps15~psaC</i>     |
| 26193      | C   | A           | 0.12 | <i>rpoB</i>          | 83527      | G   | T           | 0.09 | <i>psbD~trnT-GGU</i>     | 104282     | G   | A           | 0.1       | <i>rps15~psaC</i>     |
| 29448      | T   | G           | 0.11 | <i>psbM~trnD-GCA</i> | 84756      | A   | G           | 0.05 | <i>tRNA-Thr~rrn16</i>    | 104944     | G   | A           | 0.2       | <i>psaC~ccsA</i>      |
| 29449      | T   | G           | 0.09 | <i>psbM~trnD-GCA</i> | 97585      | T   | G           | 0.08 | <i>ycf1</i>              | 105020     | T   | C           | 0.13      | <i>psaC~ccsA</i>      |
| 30550      | G   | T           | 0.07 | <i>trnE-UUC~clpP</i> | 98169      | T   | A           | 0.05 | <i>ycf1</i>              | 105446     | A   | C/G         | 0.11/0.01 | <i>psaC~ccsA</i>      |
| 34091      | T   | G           | 0.1  | <i>psaJ~trnP-UGG</i> | 100294     | A   | C           | 0.26 | <i>ycf1</i>              | 109795     | A   | C           | 0.28      | <i>rpl32~trnV-GAC</i> |
| 37238      | T   | A           | 0.24 | <i>psbJ~petA</i>     | 100387     | T   | G           | 0.35 | <i>ycf1</i>              | 113045     | T   | G           | 0.22      | <i>rps7~trnL-CAA</i>  |
| 37239      | T   | A           | 0.24 | <i>psbJ~petA</i>     | 100523     | G   | T           | 0.02 | <i>ycf1</i>              | 114839     | T   | C           | 0.09      | <i>ycf2</i>           |
| 37240      | T   | A           | 0.24 | <i>psbJ~petA</i>     | 100598     | A   | C           | 0.13 | <i>ycf1</i>              | 114997     | A   | T           | 0.04      | <i>ycf2</i>           |
| 37263      | A   | -ATCT       | 0    | <i>psbJ~petA</i>     | 100730     | A   | G           | 0.02 | <i>ycf1</i>              | 117038     | T   | +TCITCC     | 0         | <i>ycf2</i>           |
| 38094      | C   | -GAAG       | 0    | <i>psbJ~petA</i>     | 100732     | G   | T           | 0.02 | <i>ycf1</i>              | 119153     | G   | T           | 0.48      | <i>ycf2</i>           |

**Table 3** Fifteen single nucleotide polymorphisms (SNPs) for individual identification in the *Pinus taeda* breeding population.

| Gene  | <i>rpoC1</i> | <i>IGS</i> | <i>IGS</i> | <i>IGS</i> | <i>ycf1</i> | <i>ycf1</i> | <i>ycf1</i> | <i>ycf1</i> | <i>ycf1</i> | <i>ycf1</i> | <i>ycf1</i> | <i>ycf1</i> | <i>ycf1</i> | <i>ycf1</i> | <i>ycf2</i> |
|-------|--------------|------------|------------|------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Pos   | 23167        | 23593      | 50842      | 50999      | 98314       | 101076      | 101085      | 101103      | 101112      | 101121      | 101130      | 101139      | 101148      | 101157      | 119153      |
| Ref   | G            | C          | C          | G          | G           | T           | T           | T           | T           | T           | T           | G           | G           | T           | G           |
| 013   | –            | –          | –          | –          | T           | –           | –           | –           | –           | –           | G           | –           | T           | –           | T           |
| 023   | –            | A          | T          | A          | –           | –           | –           | G           | –           | –           | –           | T           | –           | –           | –           |
| 024   | A            | A          | T          | A          | –           | –           | –           | –           | –           | –           | –           | T           | –           | –           | –           |
| 026   | –            | A          | T          | –          | –           | –           | –           | G           | –           | –           | –           | T           | –           | –           | –           |
| 222aA | A            | A          | T          | –          | –           | –           | –           | –           | –           | G           | –           | –           | T           | –           | –           |
| 243   | –            | –          | –          | –          | –           | –           | –           | G           | –           | –           | –           | –           | –           | –           | T           |
| 248   | –            | –          | –          | –          | –           | –           | –           | –           | –           | –           | –           | –           | T           | –           | –           |
| 252   | –            | –          | –          | –          | T           | –           | –           | –           | –           | –           | –           | –           | –           | –           | T           |
| 257   | –            | –          | –          | –          | N           | –           | –           | –           | –           | –           | –           | T           | –           | –           | T           |
| 259   | –            | A          | T          | A          | T           | –           | –           | G           | –           | –           | –           | T           | –           | –           | –           |
| 262   | –            | –          | –          | –          | T           | –           | –           | –           | –           | –           | –           | T           | –           | G           | T           |
| N3    | –            | –          | –          | –          | –           | –           | –           | –           | –           | –           | –           | T           | –           | –           | –           |
| G01   | –            | N          | T          | A          | –           | –           | –           | G           | –           | –           | –           | T           | –           | –           | –           |
| G10   | –            | –          | –          | –          | –           | –           | –           | G           | –           | –           | –           | –           | –           | –           | T           |
| G9    | –            | –          | –          | –          | –           | –           | –           | G           | –           | –           | –           | –           | –           | –           | T           |
| P012  | –            | –          | –          | –          | T           | G           | G           | G           | G           | –           | G           | –           | –           | –           | T           |
| P034  | –            | –          | –          | –          | –           | –           | –           | –           | –           | G           | –           | T           | –           | –           | –           |
| P043  | –            | –          | –          | –          | –           | –           | –           | –           | –           | –           | –           | –           | –           | G           | T           |
| P052  | A            | A          | T          | A          | –           | –           | –           | –           | –           | G           | –           | T           | –           | –           | –           |
| P054  | –            | –          | –          | –          | –           | –           | –           | –           | –           | –           | –           | –           | –           | G           | T           |
| P063  | –            | –          | –          | –          | –           | –           | –           | G           | –           | –           | G           | –           | –           | –           | –           |
| P064  | –            | –          | –          | –          | –           | –           | –           | –           | –           | –           | –           | –           | –           | G           | T           |
| Q16   | –            | –          | –          | –          | –           | –           | –           | –           | –           | G           | –           | –           | –           | –           | T           |
| Q26   | –            | –          | –          | –          | –           | –           | –           | –           | –           | –           | G           | –           | –           | –           | T           |
| Q6    | –            | –          | –          | –          | T           | –           | –           | –           | –           | –           | –           | –           | –           | –           | T           |
| S11   | A            | A          | T          | A          | –           | –           | –           | G           | –           | G           | –           | T           | –           | –           | –           |
| S3    | –            | –          | –          | –          | –           | –           | –           | –           | –           | –           | G           | –           | –           | –           | –           |
| S6    | –            | –          | –          | –          | –           | –           | –           | –           | –           | –           | N           | –           | –           | –           | –           |
| W05   | –            | –          | –          | –          | T           | –           | –           | –           | –           | –           | –           | T           | T           | –           | T           |
| W11   | –            | –          | –          | –          | –           | –           | –           | –           | –           | –           | –           | T           | –           | –           | –           |
| W13   | –            | –          | –          | –          | –           | –           | –           | –           | –           | –           | –           | –           | –           | –           | –           |
| W16   | A            | A          | T          | A          | –           | –           | –           | –           | –           | –           | –           | –           | –           | –           | –           |
| W28   | –            | –          | –          | –          | –           | –           | –           | –           | –           | –           | –           | –           | –           | –           | T           |

variations. The resultant tree likewise recovered Clades A and B across 33 nodes, with 16 nodes exhibiting 100% bootstrap support and the remaining nodes having bootstrap values ranging from 52% to 99%. While *P. jaliscana*, *P. oocarpa*, and *P. greggii* retained their close clustering with *P. taeda* within Clade A, *P. elliottii* and *P. caribaea* failed to group with this lineage. Instead, they occupied distinct positions with bootstrap supports of 92% at their respective nodes.



**Figure 4** Phylogenetic trees of Pinaceae from 36 plastomes and *ycf1*. Phylogenetic relationships inferred by maximum likelihood (ML) from 36 complete chloroplast genome sequences representing 25 species of Pinaceae. (A) Whole-chloroplast genome phylogeny. (B) *ycf1* gene phylogeny. Major clades are indicated: Clade A, subgenus *Pinus* (two-needle pines); Clade B, subgenus *Strobus* (single-needle pines). The red dot represents the position of *P. taeda* (NC\_021440.1).

Full-size [DOI: 10.7717/peerj.20092/fig-4](https://doi.org/10.7717/peerj.20092/fig-4)

Quantitative assessment of topological concordance revealed minimal divergence between the whole-plastome and *ycf1*-based phylogenies, evidenced by a normalized RF distance of 0.15. This strong congruence highlights the efficacy of *ycf1* as a single genomic marker for resolving genus-level evolutionary relationships in *Pinus*.

## DISCUSSION

The complete cp sequence of *P. taeda*, a dominant non-native plantation species in China, has been assembled, annotated, and analyzed. This species is a valuable resource for investigating the intra- and interspecific evolutionary history of plants (Birky Jr, 1978; Birky Jr, 2001; Chase et al., 1993; McCauley, 1995; Newton et al., 1999; Provan, Powell &

*Hollingsworth, 2001; Petit et al., 2003*). The cp genome consisted of 121,530 bp and lacked inverted repeats, which are commonly found in most angiosperms.

Comparative genomic analysis between *P. taeda* and 14 species of Pinaceae showed that certain regions, such as *ycf1*, were less conserved. This finding was confirmed in a previous study (*Jiang et al., 2017; Li et al., 2020; Chen et al., 2020; Zhang et al., 2023*). The highest sequence diversity was identified in two regions of *ycf1*. These regions can be targeted in all pine subsections using three primer combinations (*Parks, Liston & Cronn, 2011; Handy et al., 2011*). Recently, *ycf1* has been identified as an essential component of the protein translocon at the chloroplast's inner envelope membrane (*Kikuchi et al., 2013*). Seven variable sites and nine haplotypes were found in an 840 bp fragment of the DNA-coding region of *ycf1*. These variable sites and haplotypes have been evaluated in the *Pinus* subsection *Australes* and have unquestionable value for studying evolution in the group (*Ortiz-Martínez & Gernandt, 2016*). The present study corroborates that *ycf1*, with 17 mutations identified, plays a significant role in individual distinction in *P. taeda*.

It is probable that the chloroplast genes of *P. taeda* exhibit large differences in the natural selection pressure during the evolutionary process. This is in comparison to the chloroplast genes of other species belonging to the same genus (*Kober & Pogson, 2013; Nasrullah et al., 2015; Kwon et al., 2016*). However, most species of the same genus have similar codon biases and are basically clustered together. These findings suggest that the codon usage bias of chloroplast genes is closely related to the genetic similarity among species, reflects the evolutionary relationships between species, and could provide additional insights when used in conjunction with phylogenetic analysis to investigate the evolutionary relationships and molecular evolutionary mechanisms of species.

The current study surveyed the evolutionary characteristics of seven chloroplast genes across five species of *Pinus*. The findings indicated that for *clpP* and *matK*, most species pairs exhibited Ka/Ks ratios >1. In contrast, *ycf1* and *cemA* showed ratios approximately half of that value, while the remaining species pairs for these genes had ratios near 1. For *atpA*, *atpB*, and *atpI*, Ka/Ks >1 was observed in only one or two species pairs, while the ratios for the majority of pairs were well below 1. These results suggest that *clpP* and *matK* experienced predominantly positive selection, *ycf1* and *cemA* experienced either positive or neutral selection, and *atpA*, *atpB*, and *atpI* were under purifying selection during *Pinus* evolution (*Yang, 1998; Zhang, 2005*). This variation may be attributed to the encoded proteins playing different roles in growth and metabolism processes (*Hahn, 2005; Ransay, Rieseberg & Ritland, 2009; Cork & Purugganan, 2004*). cpSNP markers helped distinguish individuals among populations of loblolly pine. Distinguishing individuals is crucial for identifying the male parent of open- or mixed-pollination offspring. Given the paternal inheritance of cpDNA in *Pinus*, the male parent can be identified by using the molecular marker developed from SNPs in the cp genome. Fifteen SNPs successfully distinguished 23 individuals within the *P. taeda* breeding population. The inability to distinguish the remaining individuals suggests they may be closely related. *P. taeda* was introduced to China nearly 120 years ago, and the individuals in the breeding population of the current study were selected from plantations or had just been introduced from the southern United States. The genetic relationships among individuals in the *P. taeda* breeding population

were previously undefined. The study of the SNPs in the cp genome may provide a new method for paternal identification in the breeding program of *P. taeda*.

Chloroplast genomes are valuable tools for plant phylogenetic analysis, with important insights often coming from the examination of protein-coding genes (Eckert & Hall, 2006; Gernandt, Liston & Piñero, 2003; Moore et al., 2010; Parks, Cronn & Liston, 2012; Zhu et al., 2016a; Zhu et al., 2016b). These earlier studies provided the foundation for the current investigation into the loblolly pine chloroplast genome. This study performed a phylogenetic analysis using the maximum likelihood (ML) method. The dataset comprised 36 chloroplast genomes and the *ycf1* gene from 24 *Pinus* species, with *C. deodara* as the outgroup. Phylogenetic trees based on both the complete chloroplast genomes and the single *ycf1* gene consistently resolved the 25 species (which were represented by 36 samples) into two major clades: the subgenus *Strobilus* (single needle section, clade B) and subgenus *Pinus* (double-needle section, clade A). Particularly significant is the placement of *P. taeda* within Clade A, where it forms a sister relationship with *P. rigida* while maintaining close affinity to *P. jaliscana*, *P. oocarpa*, and *P. greggii*. This configuration aligns with and extends earlier findings by Zeb et al. (2019); Zeb et al. (2022), resolving previous ambiguities regarding loblolly pine's phylogenetic position.

Notably, analysis of the *ycf1* gene region alone recovered a highly congruent phylogenetic architecture (Fig. 4B), despite minor topological variations affecting *P. elliottii*, *P. caribaea* and *C. deodara* placements. The high degree of concordance between whole-plastome and single-gene topologies-quantified by a normalized RF distance of 0.15-validates *ycf1* as a phylogenetically informative marker for genus-level reconstructions in conifers. This finding substantiates earlier suggestions (Zhu et al., 2016a; Zhu et al., 2016b) regarding the utility of protein-coding genes, while demonstrating that targeted sequencing of *ycf1* has comparable resolution to resource-intensive complete plastome analyses in resolving major clades (e.g., subgenera) in *Pinus*.

## CONCLUSIONS

We sequenced and deposited the 121,530-bp chloroplast genome of *Pinus taeda* (GenBank: NC\_021440.1). Regions such as *ycf1* and *ycf2* showed markedly lower identity (<70% and <80%, respectively) than 14 other *Pinus* species. Phylogenomic analyses placed *P. taeda* within the diploxylon clade, sister to *P. rigida*. Comparative codon-usage patterns across 33 Pinaceae species revealed signatures of natural selection. In the breeding population, 71 cpSNPs distinguished 72% of individuals; the single-copy gene *ycf1* recapitulated whole-plastome relationships (normalized RF = 0.15), validating it as an efficient barcode. These cpSNPs and the mutation map provide readily applicable markers for paternal identification in *P. taeda* breeding programs, while the hypervariable *ycf1* gene serves as a highly informative marker for future genus-wide phylogenetic and germplasm studies.

## ADDITIONAL INFORMATION AND DECLARATIONS

### Funding

This research was supported by the National Natural Science Foundation of China (No. 31570654), the Projects of the National Science & Technology Program (2012BAD01B0203), and Science and Technology Program of Guangdong, China (2011B020302004). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Grant Disclosures

The following grant information was disclosed by the authors:

National Natural Science Foundation of China: 31570654.

Projects of the National Science & Technology Program: 2012BAD01B0203.

Science and Technology Program of Guangdong, China: 2011B020302004.

### Competing Interests

The authors declare there are no competing interests.

### Author Contributions

- Ling Wang conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Jipeng Mao analyzed the data, prepared figures and/or tables, and approved the final draft.
- Kaibin Jiang analyzed the data, prepared figures and/or tables, and approved the final draft.
- Zhengyu Wu analyzed the data, prepared figures and/or tables, and approved the final draft.
- Chunxin Liu analyzed the data, prepared figures and/or tables, and approved the final draft.
- Huagong Ning performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Haibiao Zhou analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Jiehu Chen analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Shaowei Huang analyzed the data, prepared figures and/or tables, and approved the final draft.
- Tianyi Liu conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

### Data Availability

The following information was supplied regarding data availability:

The raw data is available at CNGBdb: [CNP0006272](https://cn.gdb.cn/CNP0006272).



The raw data is available at GenBank: [PRJNA1159385](#) and [NC\\_021440.1](#).

## REFERENCES

- Asaf S, Khan AL, Khan MA, Shahzad R, Lubna, Kang SM, Ahmed A-R, In-Jung L. 2018. Complete chloroplast genome sequence and comparative analysis of loblolly pine (*Pinus taeda* L.) with related species. *PLOS ONE* 13(3):e0192966 DOI [10.1371/journal.pone.0192966](#).
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin Valery M, Nikolenko Sergey I, Pham S, Prjibelski AD. 2012. SPAdes: a new genome assembly algorithm and its applications to single cell sequencing. *Journal of Computational Biology* 19(5):455–477 DOI [10.1089/cmb.2012.0021](#).
- Birky Jr CW. 1978. Transmission genetics of mitochondria and chloroplasts. *Annual Review of Genetics* 12(1):471–512 DOI [10.1146/annurev.ge.12.120178.002351](#).
- Birky Jr CW. 2001. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annual Review of Genetics* 35(1):125–148 DOI [10.1146/annurev.genet.35.102401.090231](#).
- Bookjans G, Stummam BM, Henningsen KW. 1984. Preparation of chloroplast DNA from pea plastids isolated in a medium of high ionic strength. *Analytical Biochemistry* 141(1):244–247 DOI [10.1016/0003-2697\(84\)90452-4](#).
- Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu Y-L, Kron KA, Rettig JH, Conti E, Palmer JD, Manhart JR, Sytsma KJ, Michaels HJ, John Kress W, Karol KG, Dennis Clark W, Hedren M, Gaut BS, Jansen RK, Kim K-J, Wimpee CF, Smith JF, Furnier GR, Strauss SH, Xiang QY, Plunkett GM, Soltis PS, Swensen SM, Williams SE, Gadek PA, Quinn CJ, Eguiarte LE, Golenberg E, Learn Jr GH, Graham SW, Barrett SCH, Dayanandan S, Albert VA. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden* 80(3):528–548+550–580 DOI [10.2307/2399846](#).
- Chen SF, Ishizuka W, Hara T, Goto S. 2020. Complete chloroplast genome of Japanese Larch (*Larix kaempferi*): insights into intraspecific variation with an isolated northern limit population. *Forests* 11(8):884 DOI [10.3390/f11080884](#).
- Cork JM, Purugganan MD. 2004. The evolution of molecular genetic pathways and networks. *Bioessays* 26(5):479–484 DOI [10.1002/bies.20026](#).
- Daniell H, Lin CS, Yu M, Chang WJ. 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology* 17(1):134 DOI [10.1186/s13059-016-1004-2](#).
- Delannoy E, Fujii S, Colasdes Francs-Small C, Brundrett M, Small I. 2011. Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlights evolutionary constraints on plastid genomes. *Molecular Biology and Evolution* 28(7):2077–2086 DOI [10.1093/molbev/msr028](#).

- Duan R-Y, Yang L-M, Lv T, Wu G-L, Huang M-Y. 2016. The complete chloroplast genome sequence of *Pinus dabeshanensis*. *Conservation Genetics Resources* 8(4):395–397 DOI 10.1007/s12686-016-0567-2.
- Eckert AJ, Hall BD. 2006. Phylogeny, historical biogeography, and patterns of diversification for *Pinus* (Pinaceae): phylogenetic tests of fossil-based hypotheses. *Molecular Phylogenetics and Evolution* 40(1):166–182 DOI 10.1016/j.ympev.2006.03.009.
- El-Kassaby YA, Lstibůrek M. 2009. Breeding without breeding. *Genetics Research* 91(2):111–120 DOI 10.1017/S001667230900007X.
- Georgolopoulos G, Parducci L, Drouzas AD. 2016. A short phylogenetically informative cpDNA fragment for the identification of *Pinus* species. *Biochemical Systematics and Ecology* 66:166–172 DOI 10.1016/j.bse.2016.03.001.
- Gernandt DS, Liston A, Piñero D. 2003. Phylogenetics of *Pinus* subsections Cembroides and Nelsoniae inferred from cpDNA sequences. *Systematic Botany* 28(4):657–673 DOI 10.1043/02-61.1.
- Gernandt DS, López GG, García SO, Liston A. 2005. Phylogeny and classification of pinus. *Taxon* 54(1):29 DOI 10.2307/25065300.
- Gouy M, Milleret F, Mugnier C, Jacobzone M, Gautier C. 1984. ACNUC: a nucleic acid sequence data base and analysis system. *Nucleic Acids Research* 12(1):121–127 DOI 10.1093/nar/12.1Part1.121.
- Grant JR, Stothard P. 2008. The CGView server: a comparative genomics tool for circular genomes. *Nucleic Acids Research* 36(Web Server issue):W181–W184 DOI 10.1093/nar/gkn179.
- Greiner S, Lehwark P, Bock R. 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Research* 47:W59–W64 DOI 10.1093/nar/gkz238.
- Gwaze DP, Byram TD, Lowe WJ. 2001. Genetic parameter estimates for growth and wood density in loblolly pine (*Pinus taeda* L.). *Forest Genetics* 8(1):47–54.
- Hahn MW. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein interaction networks. *Molecular Biology and Evolution* 22(4):803–806 DOI 10.1093/molbev/msi072.
- Handy SM, Parks MB, Deeds JR, Liston A, Diachenko GW. 2011. Use of the chloroplast gene *ycf1* for the genetic differentiation of pine nuts obtained from consumers experiencing dysgeusia. *Journal of Agricultural and Food Chemistry* 59(20):10995–11002 DOI 10.1021/jf203215v.
- Jakobsson M, Säll T, Lind-Halldén C, Halldén C. 2007. The evolutionary history of the common chloroplast genome of *Arabidopsis thaliana* and *A. suecica*. *Journal of Evolutionary Biology* 20(1):104–121 DOI 10.1111/j.1420-9101.2006.01217.x.
- Jiang D, Zhao Z, Zhang T, Zhong W, Liu C, Yuan Q, Huang L. 2017. The chloroplast genome sequence of *Scutellaria baicalensis* provides insight into intraspecific and interspecific chloroplast genome diversity in *Scutellaria*. *Gene* 8(9):227 DOI 10.3390/genes8090227.

- Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* 20:1160–1166 DOI 10.1093/bib/bbx108.
- Kikuchi S, Bedard J, Hirano M, Hirabayashi Y, Oishi M, Imai M, Takase M, Ide T, Nakai M. 2013. Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science* 339(6119):571–574 DOI 10.1126/science.1229262.
- Kober KM, Pogson GH. 2013. Genome-wide patterns of codon bias are shaped by natural selection in the purple sea urchin, *Strongylocentrotus purpuratus*. *G3—Genes Genomes Genetics* 3(7):1069–1083 DOI 10.1534/g3.113.005769.
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution* 35:1547–1549 DOI 10.1093/molbev/msy096.
- Kwon KC, Chan HT, León IR, Williams-Carrier R, Barkan A, Daniell H. 2016. Codon-optimization to enhance expression yields insights into chloroplast translation. *Plant Physiology* 172(1):62–77 DOI 10.1104/pp.16.00981.
- Lance GN, Williams WT. 1967. A general theory of classificatory sorting strategies 1, hierarchical systems. *Computer Journal* 9:373–380 DOI 10.1093/comjnl/9.4.373.
- Li Q. 2012. *Analyzing cpDNA complete sequence of Pinus taeda and detecting relationship among parents in the breeding population*. Guangzhou: South China Agriculture University.
- Li JL, Jiang M, Chen HM, Yu J, Liu C. 2020. Intraspecific variations among the chloroplast genomes of *Artemisia scoparia* (asteraceae) from Pakistan and China. *Mitochondrial DNA Part B* 5(3):3200–3202 DOI 10.1080/23802359.2020.1810167.
- Mao L, Zou Q, Sun Z, Dong Q, Cao X. 2023. Insights into chloroplast genome structure, intraspecific variation, and phylogeny of *Cyclamen* species (Myrsinoideae). *Scientific Reports* 13(1):87 DOI 10.1038/s41598-022-27163-1.
- McCauley DE. 1995. The use of chloroplast DNA polymorphism in studies of gene flow in plants. *Trends in Ecology & Evolution* 10(5):198–202 DOI 10.1016/S0169-5347(00)89052-7.
- McKeand S, Mullin T, Byram T, White T. 2003. Deployment of genetically improved loblolly and slash pines in the south. *Journal of Forestry* 101(3):32–37 DOI 10.1093/jof/101.3.32.
- Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences of the United States of America* 107(10):4623–4628 DOI 10.1073/pnas.0907801107.
- Nasrullah I, Butt AM, Tahir S, Idrees M, Tong Y. 2015. Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution. *BMC Evolutionary Biology* 15(1):174 DOI 10.1186/s12862-015-0456-4.
- Neale DB, Marshall KA, Sederoff RR. 1989. Chloroplast and mitochondrial DNA are paternally inherited in *Sequoia sempervirens* D. Don Endl. *Proceedings of the*

- National Academy of Sciences of the United States of America* **86**(23):9347–9349 DOI 10.1073/pnas.86.23.9347.
- Neale DB, Wheeler NC, Allard RW. 1986. Paternal inheritance of chloroplast DNA in Douglas-fir. *Canadian Journal of Forest Research* **16**(5):1152–1154 DOI 10.1139/x86-205.
- Newton A, Allnutt T, Gillies A, Lowe A, Ennos R. 1999. Molecular phylogeography, intraspecific variation and the conservation of tree species. *Trends in Ecology & Evolution* **14**(4):140–145 DOI 10.1016/S0169-5347(98)01555-9.
- Nock CJ, Waters DLE, Edwards MA, Bowen SG, Rice N, Cordeiro GM, Henry RJ. 2011. Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnology Journal* **9**(3):328–333 DOI 10.1111/j.1467-7652.2010.00558.x.
- Ortiz-Martínez A, Gernandt DS. 2016. Species diversity and plastid DNA haplotype distributions of *Pinus Subsection Australes* (pinaceae) in Guerrero and Oaxaca. *TIP* **19**(2):92–101 DOI 10.1016/j.recqb.2016.06.002.
- Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**(3):526–528 DOI 10.1093/bioinformatics/bty633.
- Parks M, Cronn R, Liston A. 2012. Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from *Pinus* L. (Pinaceae). *BMC Evolutionary Biology* **12**(1):100 DOI 10.1186/1471-2148-12-100.
- Parks M, Liston A, Cronn R. 2011. Newly developed primers for complete *ycf1* amplification in *Pinus* (Pinaceae) chloroplasts with possible family-wide utility. *American Journal of Botany* **98**(7):e185–e188 DOI 10.3732/ajb.1100088.
- Pazouki L, Shanjani PSS, Fields PD, Martins K, Suhhorutšenko M, Viinalass H, Niinemets Ü. 2016. Large within-population genetic diversity of the widespread conifer *Pinus sylvestris* at its soil fertility limit characterized by nuclear and chloroplast microsatellite markers. *European Journal of Forest Research* **135**(1):161–177 DOI 10.1007/s10342-015-0928-5.
- Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C, Brewer S, Cheddadi R, Ennos R, Fineschi S, Grivet D, Lascoux M, Mohanty A, Müller-Starck G, Demesure-Musch B, Palmé A, Martin JP, Rendell S, Vendramin GG. 2003. Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* **300**(5625):1563–1565 DOI 10.1126/science.1083264.
- Provan J, Powell W, Hollingsworth PM. 2001. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology & Evolution* **16**(3):142–147 DOI 10.1016/S0169-5347(00)02097-8.
- R Core Team. 2022. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available at <https://www.R-project.org/>.
- Ransay H, Rieseberg LH, Ritland K. 2009. The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. *Molecular Biology and Evolution* **26**(5):1045–1053 DOI 10.1093/molbev/msp021.
- Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK. 2007. Comparative chloroplast genomics: analyses including new sequences from the

- angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genomics* **8**(1):174 DOI [10.1186/1471-2164-8-174](https://doi.org/10.1186/1471-2164-8-174).
- Semerikov VL, Semerikova SA, Dymshakova OS, Zatsepina KG, Tarakanov VV, Tikhonova IV, Ekart AK, Vidyakin AI, Jamiyansuren S, Rogovtsev RV. 2014. Microsatellite loci polymorphism of chloroplast dna of scots pine (*Pinus sylvestris* L.) in Asia and eastern Europe. *Russian Journal of Genetics* **50**(6):577–585 DOI [10.1134/S1022795414040127](https://doi.org/10.1134/S1022795414040127).
- Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* **15**(3):1281–1295 DOI [10.1093/nar/15.3.1281](https://doi.org/10.1093/nar/15.3.1281).
- Szmidt AE, Aldén T, Hällgren JE. 1987. Paternal inheritance of chloroplast DNA in *Larix*. *Plant Molecular Biology* **9**(1):59–64 DOI [10.1007/BF00017987](https://doi.org/10.1007/BF00017987).
- Wagner DB, Furnier GR, Saghai-Marroof MA, Williams SM, Dancik BP, Allard RW. 1987. Chloroplast DNA polymorphisms in lodgepole and jack pines and their hybrids. *Proceedings of the National Academy of Sciences of the United States of America* **84**(7):2097–2100 DOI [10.1073/pnas.84.7.2097](https://doi.org/10.1073/pnas.84.7.2097).
- Walter R, Epperson BK. 2010. Geographic pattern of genetic variation in *Pinus resinosa*: area of greatest diversity is not the origin of postglacial populations. *Molecular Ecology* **10**(1):103–111 DOI [10.1046/j.1365-294X.2001.01177.x](https://doi.org/10.1046/j.1365-294X.2001.01177.x).
- Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences of the United States of America* **84**:9054–9058 DOI [10.1073/pnas.84.24.9054](https://doi.org/10.1073/pnas.84.24.9054).
- Wu CS, Chaw SM. 2014. Highly rearranged and size-variable chloroplast genomes in conifers II clade (cupressophytes): evolution towards shorter intergenic spacers. *Plant Biotechnology Journal* **12**(3):344–353 DOI [10.1111/pbi.12141](https://doi.org/10.1111/pbi.12141).
- Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15**(5):568–573 DOI [10.1093/oxfordjournals.molbev.a025957](https://doi.org/10.1093/oxfordjournals.molbev.a025957).
- Ying W, Di-Feng Z, Xian J, Wen-Li M, Hao-Fu D, Xiong-Ting C, Shi-Qing P. 2016. Complete chloroplast genome sequence of *Aquilaria sinensis* (Lour.) gilg and evolution analysis within the malvales order. *Frontiers in Plant Science* **7**:280 DOI [10.3389/fpls.2016.00280](https://doi.org/10.3389/fpls.2016.00280).
- Zeb U, Dong W-L, Zhang T-T, Wang R-N, Shahzad K, Ma X-F, Li Z-H. 2019. Comparative plastid genomics of *Pinus* species: insights into sequence variations and phylogenetic relationships. *Journal of Systematics and Evolution* **58**(2):118–132 DOI [10.1111/jse.12492](https://doi.org/10.1111/jse.12492).
- Zeb U, Wang X, Aziz Ullah AU, Fiaz S, Khan H, Ullah S, Ali H, Shahzad K. 2022. Comparative genome sequence and phylogenetic analysis of chloroplast for evolutionary relationship among *Pinus* species. *Saudi Journal of Biological Sciences* **29**(3):1618–1627 DOI [10.1016/j.sjbs.2021.10.070](https://doi.org/10.1016/j.sjbs.2021.10.070).

- Zhang J. 2005.** Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution* **22**(12):2472–2479 DOI [10.1093/molbev/msi237](https://doi.org/10.1093/molbev/msi237).
- Zhang W, Zhang Z, Liu B, Chen J, Zhao Y, Huang Y. 2023.** Comparative analysis of 17 complete chloroplast genomes reveals intraspecific variation and relationships among *Pseudostellaria heterophylla* (Miq.) Pax populations. *Frontiers in Plant Science* **14**:1163325 DOI [10.3389/fpls.2023.1163325](https://doi.org/10.3389/fpls.2023.1163325).
- Zhong WH, Chen BQ, Huang ZH. 1995.** Roguing in seed orchard of loblolly pine. In: *Forest tree improvement in the Asia-Pacific region*. Beijing: China Forestry Publishing House, 208–216.
- Zhu A, Guo W, Gupta S, Fan W, Mower JP. 2016a.** Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytologist* **209**(4):1747–1756 DOI [10.1111/nph.13743](https://doi.org/10.1111/nph.13743).
- Zhu ZY, Xie JW, Zhang MC, Zhou Q, Liu FQ. 2016b.** Insight into the adsorption of PPCPs by porous adsorbents: effect of the properties of adsorbents and adsorbates. *Environmental Pollution* **214**:524–531 DOI [10.1016/j.envpol.2016.04.070](https://doi.org/10.1016/j.envpol.2016.04.070).