forests

MDPI

Article

Complete Chloroplast Genome of *Fokienia hodginsii* (Dunn) Henry et Thomas: Insights into Repeat Regions Variation and Phylogenetic Relationships in Cupressophyta

Mingyue Zang [1], Qian Su [1], Yuhao Weng [1], Lu Lu [1], Xueyan Zheng [2], Daiquan Ye [2], Renhua Zheng [3], Tielong Cheng [1,4], Jisen Shi [1] and Jinhui Chen [1,4,*]

[1] Key Laboratory of Forest Genetics & Biotechnology of Ministry of Education of China, Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing 210037, China; sanskritm@163.com (M.Z.); suqian@njfu.edu.cn (Q.S.); gianl13851756619@163.com (Y.W.); luluzhubifu@hotmail.com (L.L.); ctielong@126.com (T.C.); jshi@njfu.edu.cn (J.S.)

[2] National Germplasm Bank of Chinese fir at Fujian Yangkou Forest Farm, Shunchang 353211, China; zxy0553@163.com (X.Z.); yklcydq@163.com (D.Y.)

[3] The Key Laboratory of Timber Forest Breeding and Cultivation for Mountainous Areas in Southern China, State Forestry Administration Engineering Research Center of Chinese Fir, Fujian Academy of Forestry, Fuzhou 350012, China; zrh08@126.com

[4] College of Biology and the Environment, Nanjing Forestry University, Nanjing 210037, China

* Correspondence: chenjh@njfu.edu.cn; Tel./Fax: +86-25-85428948

Received: 24 May 2019; Accepted: 21 June 2019; Published: 26 June 2019

Abstract: *Fokienia hodginsii* (Dunn) Henry et Thomas is a relic gymnosperm with broad application value. It is a fit candidate when choosing species for the construction of artificial forests. We determined the complete chloroplast genome sequence of *F. hodginsii*, which is 129,534 bp in length and encodes 83 protein genes, 33 transfer RNA (tRNA) genes, as well as four ribosomal RNA genes. The GC content of the complete sequence and protein coding regions is 34.8% and 36.2%, respectively. We identified 11 tandem repeats, 11 forward repeats, and three palindromic repeats and classified them by size. Following our microsatellite analysis, a total number of 73 simple sequence repeats were detected, preferentially within the intergenic space. Being a member of Cupressophyta, *F. hodginsii* owns several common characters; the *trnR-CCG* gene has been deleted, while the *trnI-CAU* and *trnQ-UUG* genes have been duplicated. Moreover, the *accD* gene, which encodes acetyl-CoA carboxylase, contains 771 codons in *F. hodginsii*, similar to *Cryptomeria japonica* (L. F.) D. Don, further supporting the diversity of *accD* and its size expansion in Cupressophyta. Concerning the loss of inverted repeat (IR) regions, the 86-bp sequence with the duplicated *trnI-CAU* gene is inferred to be the footprint of IR contraction. Phylogenetically, *F. hodginsii* is placed as a sister taxon to *Chamaecyparis hodginsii* (Dunn) Rushforth. This work offers meaningful guidance as well as reference value to the breeding research and improvement of *F. hodginsii*. Moreover, it gives us a better understanding of the genomic structure and evolutionary history of gymnosperms, especially coniferales.

Keywords: chloroplast genome; *Fokienia hodginsii*; inverted repeats; Cupressophyta

1. Introduction

Fokienia hodginsii (Dunn) Henry et Thomas, a relic gymnosperm from the tertiary period, is the only species of *Fokienia* Henry et Thomas. This threatened species was listed as vulnerable by the International Union for Conservation of Nature (IUCN) and was also under the second class state protection recorded by China Species Red List. *F. hodginsii* grows well in warm and humid subtropical zones, especially at altitudes from 500 to 1800 m above sea level; thus, this species is naturally distributed in southern Chinese provinces like Fujian, Jiangxi, and Guangdong [1,2]. Its high quality timber along with its extensive exploitation make *F. hodginsii* an important afforestation species worth promoting [2].

Gymnosperms, with typical exposed ovules, are divided into five groups: the cycads, *Ginkgo*, Pinaceae, gnetophytes, and Cupressophyta. The Cupressophyta are aggregations of conifers, excluding Pinaceae. This phylum consists of six families: Araucariaceae, Cephalotaxaceae, Cupressaceae, Podocarpaceae, Sciadopityaceae, and Taxaceae. While the complete chloroplast (cp) genome sequences of tobacco [3] and liverwort [4] were initially reported in 1986, the first cp genome sequence from members of Cupressophyta was not released until 2008, which was represented by *Cryptomeria japonica* (L. F.) D. Don [5].

Typically, cp genomes have a circular molecular structure ranging from 120 to 160 kbp [6], including a pair of IRs separated by a large single-copy region (LSC) and a small single-copy region (SSC) [7]. Across most plant species, the LSC and SSC sequences have always been conserved, but the inverted repeat (IR) regions, especially in Gymnospermae, vary in length and directionality [8–10]. For example, previous reports have suggested that *Ginkgo biloba* L. possess the structure of both IR_A and IR_B, but have lost several ancestral parts of the complete sequence, leaving only the duplication of some genes, such as one of the largest cp genome-encoded open reading frames, *ycf*2, whose gene products are essential for cell survival [11–13]. Furthermore, the Pinaceae and Cupressophyta, represented by *Cedrus deodara* (Roxb.) G. Don and *Cephalotaxus wilsoniana* Hay, have retained different IR copies [14]. The diversity in IR regions caused by its contraction or expansion ultimately led to a large variety in cp genomic size across the plant kingdom [15]. Therefore, they can provide valuable information for us to study the evolutionary process. As an endemic species in *Fokienia* Henry et Thomas, the genome sequence of *F. hodginsii* is still unpublished. The objectives of our work were to (1) acquire the complete cp genome sequence of *F. hodginsii*; (2) perform sequence statistics and structure comparing analyses to explore the typical characteristics of the *F. hodginsii* cp genome; and (3) together with related studies, establish a more abundant systematic relationship within the major lineages of conifers, especially in Cupressophyta, to understand better the coniferous evolutionary process.

2. Materials and Methods

2.1. DNA Extraction and Genome Assembly

Using the high salt concentration method [16], genomic DNA was isolated from young leaves of *F. hodginsii* grown in Fujian Yangkou Forest Farm, Shunchang, China. A 500 bp paired-end library was constructed with 5 µg of the isolated cp DNA. About 2 Gb of sequence with an average read length of 301 bp was obtained using the Illumina MiSeq platform. As to the assembly of the whole cp genome sequence, MiSeq raw reads were trimmed to 200 bp in length using a script developed in-house ('fasta_length_trimmer') to remove the potential low-quality bases. Then, initial contigs were assembled using Velvet Assembler version 1.2.07 [17]. Contigs with high similarity (E-value $<1^{e-10}$) compared with the reference cp genome (KX832622.1) were chosen to be spliced using SSPACE Premium version 2.2 [18], followed by a manual check. Finally, we acquired one single circular cp genome sequence (129,534) without ambiguous bases.

2.2. Genome Annotation and Sequence Analyses

The online program Dual Organellar GenoMe Annotator [19] and the Basic Local Alignment Search Tool (https://blast.ncbi.nlm.nih.gov/Blast.cgi) were used to annotate the *F. hodginsii* cp genome. Compared with homologous genes from other cp genomes, we manually checked the initial annotation and determined the putative start and stop codons, as well as intron positions. All transfer RNA genes were further verified by using tRNAscan-SE [20]. The map of the circular cp genome was drawn using OrganellarGenomeDRAW [21]. Codon usage and GC content were calculated by MEGA7 [22].

2.3. Repeat Sequence Statistics and IR Identification

Tandem repeats were identified using Tandem Repeats Finder [23] with parameters set to 2 for match, 7 for mismatch, and 7 for indel. The minimum alignment score and the maximum period length were 50 bp and 500 bp, respectively. Statistics of forward, reverse, complement, and palindromic repeats were computed by REPuter [24]. The minimal size was set to 30 bp and >90% identity with a hamming distance of 3. Upon the detection of the inverted repeat sequence, we conducted a comparison of IR regions among *F. hodginsii* and the other four conifers (*Podocarpus lambertii* Klotzch ex Endl, *Cephalotaxus oliveri* Mast, *Taiwania cryptomerioides* Hayata, and *C. japonica*). Additionally, in order to explore the conserved region as well as the repeated region of *accD*, we performed an amino acid sequence alignment using online MAFFT version7 [25] and visualized the results with jalview [26].

For the analysis of simple sequence repeats (SSRs), the Perl script MISA (https://pgrc.ipk-gatersleben.de/misa/) was used. The threshold of minimal repeat units was 10 for mononucleotides, five for dinucleotides, and four for tri-, tetra-, penta-, and hexanucleotides. All the outputs from the programs above were manually adjusted and redundant data were removed.

2.4. Phylogenetic Research

In order to construct a phylogenetic tree, which shows the evolutionary relationships between candidates in a clear tree-like map, we selected 21 species (Table S1) and downloaded the sequences of their cp protein-coding genes from the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/). We excluded all *ndh* (NADH dehydrogenase) genes due to their absence across Pinaceae [27]. After manual adjustment, 63 eligible genes were ultimately aligned using Clustal W [28] and concatenated by SequenceMatrix [29]. Phylogenetic trees were estimated by MEGA7 [16] using an array with 48967 bp nucleotide sites by maximum likelihood (ML) and neighbor-joining (NJ) methods. In the ML tree, the model General Time Reversible + Proportion Invariant + Gamma (GTR + I + R) was preferred for nucleotide substitution. We selected the Kimura two-parameter model [30] to estimate evolutionary rates of base substitutions and nearest-neighbor-interchange (NNI) on random trees, while initial trees were constructed automatically. For NJ analyses, we selected the the maximum composite likelihood method. One thousand bootstrap replicates were applied to evaluate clade supports in both methods.

3. Results

3.1. Genome Features

We determined the full length of the *F. hodginsii* cp genome (MK890147) to be 129,534 bp (Figure 1). The complete cp DNA does not share the common quadripartite structure, which contains a pair of IRs separated by the LSC and SSC regions. However, the *trnI-CAU* and *trnQ-UUG* genes were found to be duplicated within the short inverted repeats. We identified 120 genes (Table 1) in total, including 83 protein genes, 33 transfer RNA genes, and four ribosomal RNA genes. Among the 16 intron-containing genes (Table 2), eight protein-coding genes and six tRNA genes contain just one intron while another two genes, ribosomal protein S12 (*rps12*), and hypothetical chloroplast reading frame 3 (*ycf3*) contain two introns. It is worth mentioning that *rps12* is trans-spliced [31], meaning that exonI is about 35 kilobase pairs downstream of the nearest copy of the other two exons. Moreover, the largest intron within *trnK-UUU* (2457 bp) includes the entire *matK* gene.

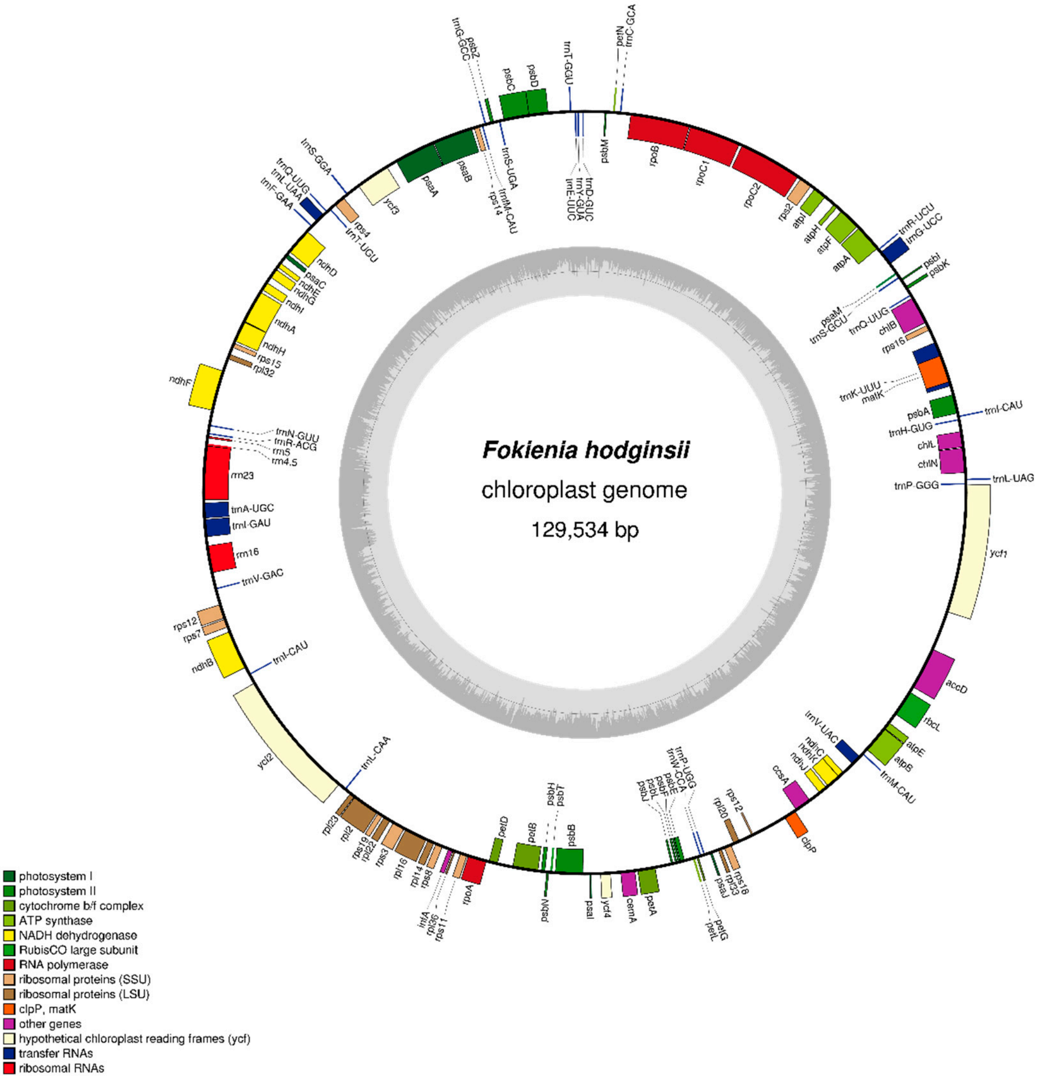

Figure 1. Gene map of the *F. hodginsii* cp genome. Genes drawn outside the circle are transcribed clockwise, while those shown inside are transcribed counter-clockwise. Genes were coded with diversified colors to distinguish their different functional groups. The darker and lighter gray circles represent GC and AT contents, respectively.

The GC contents of the complete cp genome sequence and protein-coding regions in *F. hodginsii* are 34.8% and 36.2%, respectively. Eighty-three genes constitute the protein-coding regions, being 74,715 bp in length. The first codon position shows a higher GC content than the second one; by contrast, the lowest GC content at the third codon position indicates a codon usage bias in A and T (Table 3), which contribute to the enrichment of AT content in the cp genome [32]. A total of 24,905 codons were identified, with the most and least frequently used codons being AAA (1234) and UGC (77), encoding lysine and cysteine, respectively. Leucine (10.95%) and cysteine (1.14%) respectively were the most and least often encoded amino acids (Figure 2).

Table 1. List of genes found in the *F. hodginsii* cp genome.

Functional Category	Group of Genes	Gene Names
Self-replication	Ribosomal RNA genes	rrn4.5, rrn5, rrn16, rrn23
	Transfer RNA genes	trnP-GGG, trnL-UAG, trnH-GUG, trn-I CAU(×2), trnK-UUU*, trnQ-UUG(×2), trnS-GCU, trnG-GCC, trnR-UCU, trnC-GCA, trnD-GUC, trnY-GUA, trnE-UUC, trnT-GGU, trnS-UGA, trnG-UCC*, trnfM-CAU, trnS-GGA, trnT-UGU, trnL-UAA*, trnF-GAA, trnN-GUU, trnR-ACG, trnA-UGC*, trnI-GAU*, trnV-GAC, trnL-CAA, trnW-CCA, trnP-UGG, trnV-UAC*, trnM-CAU
	Small subunit of ribosome	rps2, rps3, rps4, rps7, rps8, rps11, rps12**, rps14, rps15, rps16*, rps18, rps19
	Large subunit of ribosome	rpl2*, rpl14, rpl16*, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36
	DNA-dependent RNA polymerase	rpoA, rpoB, rpoC1*, rpoC2
	Translational initiation factor	infA
Genes for photosynthesis	Subunits of photosystem I	psaA, psaB, psaC, psaI, psaJ, psaM, ycf3**, ycf4
	Subunits of photosystem II	psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ
	Subunits of cytochrome	petA, petB*, petD, petG, petL, petN
	Subunits of ATP synthase	atpA, atpB, atpE, atpF*, atpH, atpI
	Chlorophyll biosynthesis	chlB, chlL, chlN
	Large subunit of Rubisco	rbcL
	Subunits of NADH dehydrogenase	ndhA*, ndhB*, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK
Other genes	Maturase	matK
	Envelope membrane protein	cemA
	Subunit of acetyl-CoA	accD
	C-type cytochrome synthesis gene	ccsA
	Protease	clpP
	Component of TIC complex	ycf1
Genes of unknown function	Conserved open reading frames	ycf2

One or two asterisks after a gene indicate that the gene contains one or two introns.

Table 2. Genes containing introns in the *F. hodginsii* cp genome.

Gene	Exon I (bp)	Intron I (bp)	Exon II (bp)	Intron II (bp)	Exon III (bp)
atpF	145	660	410		
ndhA	558	745	540		
ndhB	711	659	756		
rpl2	398	643	433		
rpl16	9	824	411		
rps12	114	/	232	526	26
rps16	40	235	41		
rpoC1	442	726	1673		
petB	6	770	642		
ycf3	126	688	226	690	158
trnA-UGC	38	776	35		
trnG-UCC	23	739	48		
trnI-GAU	42	892	35		
trnK-UUU	37	2457	35		
trnL-UAA	35	438	50		
trnV-UAC	39	518	37		

Table 3. Nucleotide composition of the *F. hodginsii* cp genome.

	T(U)%	C%	A%	G%	Length (bp)	Number	GC (%)
Total	31.9	17.6	33.2	17.2	129,534	-	34.8
CDS	31.8	16.6	32.0	19.6	74,715	-	36.2
First codon position	24	18.1	30.8	27.5	-	24,905	45.6
Second codon position	33	19.5	30.8	16.6	-	24,905	36.1
Third codon position	39	12.1	34.3	14.8	-	24,905	26.9

CDS: protein-coding genes.

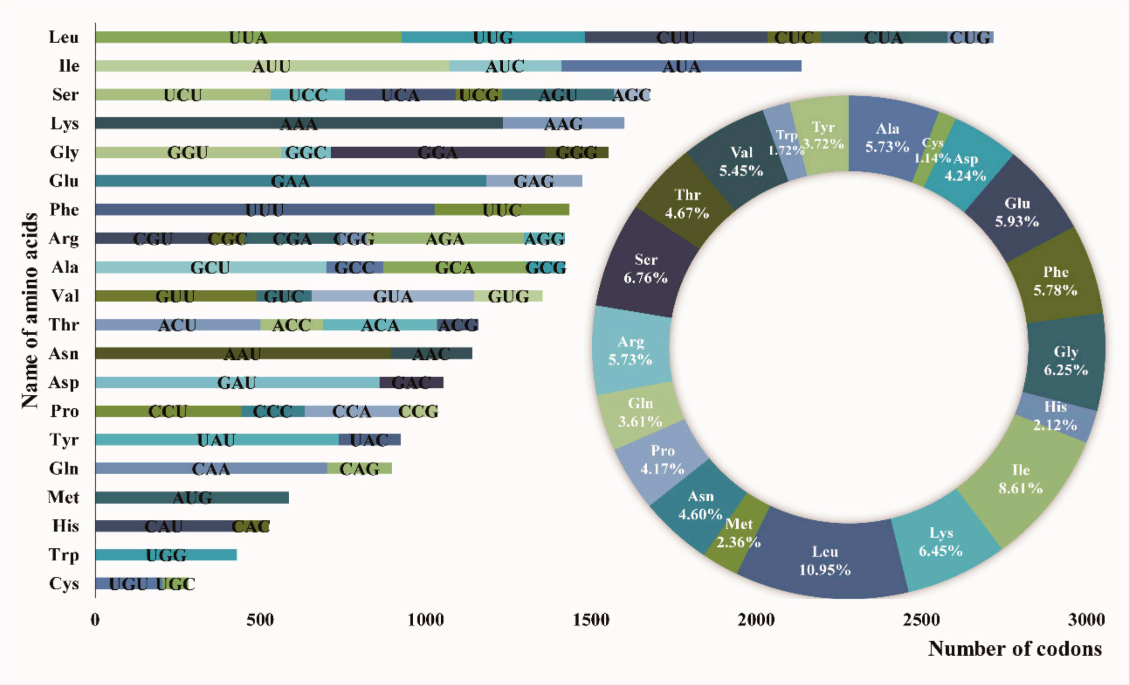

Figure 2. Codon usage and amino acid frequencies in the *F. hodginsii* cp genes. The composition of amino acids and the quantities of codons are listed in the bar chart. The frequencies of amino acids were calculated for all of the 83 protein-coding genes from the *F. hodginsii* cp genome shown in the pie chart.

3.2. Repeat Sequence and SSR Analysis

Repeat sequences play a crucial role in the occurrence of genomic rearrangements and the invention of evolutionary novelties. Occurring frequently in genomic sequences, they are important laboratory and analytic tools [23]. Our repeat sequence analyses are shown in Figure 3. We classified tandem, forward, and palindromic repeats, of which we identified 11, 11, and three, respectively (Figure 3A). Then we subsequently categorized these by size as shown in Figure 3B. A tandem repeat in DNA is two or more adjacent, approximate copies of a pattern of nucleotides. In *F. hodginsii*, the majority of tandem repeats (81.81%) are below 60 bp in length. As to forward repeats, the direct sequences that have the same nucleotide bases and directions in common, they are mainly between 60 and 74 bp long. Palindromic repeats, which frequently occur as essential elements in regulatory regions [33], vary in length from 52 to 233 bp (Table S2). Concerning the distribution of all these repeats, 59% of them are located within intergenic regions, whereas the remaining 41% occur within protein-coding genes (Figure 3C).

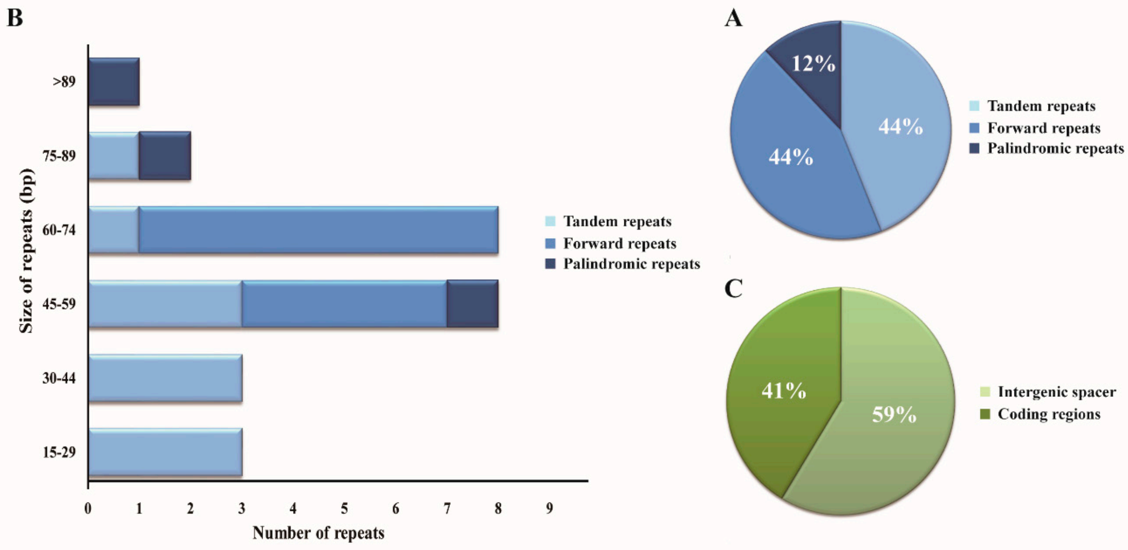

Figure 3. Analyses of repeat sequences in *F. hodginsii* cp genome. (**A**) Occupation of the three kinds of sequences. (**B**) Frequency distribution of tandem repeats, forward repeats, and palindromic repeats. The cutoff value was 15, 30, and 30 bp, respectively. (**C**) The proportion of regions containing repeat sequences.

Simple sequence repeats (SSRs), also called microsatellites, have a high level of polymorphisms and are considered potential genetic markers in ecological and evolutionary studies of plants [34]. From our SSR analysis, we detected 73 microsatellites in total (Table S3), with the minimum number of 10, five, four, and six for mono-, di-, tri, and hexa-nucleotides, respectively (Figure 4A). However, we did not detect any tetra- or penta-nucleotide motifs. The most common motif was mononucleotide A, followed by mononucleotide T, which together make up 53.4% of all SSRs. Among dinucleotide and trinucleotide repeats, AT and AAG occur the most frequently. The only hexanucleotide repeat found in the *F. hodginsii* cp genome was AGGAAC. Most (68.49%) of the SSRs are located in the intergenic space, followed by protein-coding genes (20.55%) and introns (10.96%) (Figure 4C). The exact numbers of SSRs found in these three regions were 50, 15, and 8, respectively (Figure 4B). This distribution preference supports previous findings that the SSR frequency varies between different regions of the genome [35]. Among all SSRs, a majority of homopolymers is entirely composed of A/T sequences (76.7%), indicating an enrichment in A/T content and a bias in base composition.

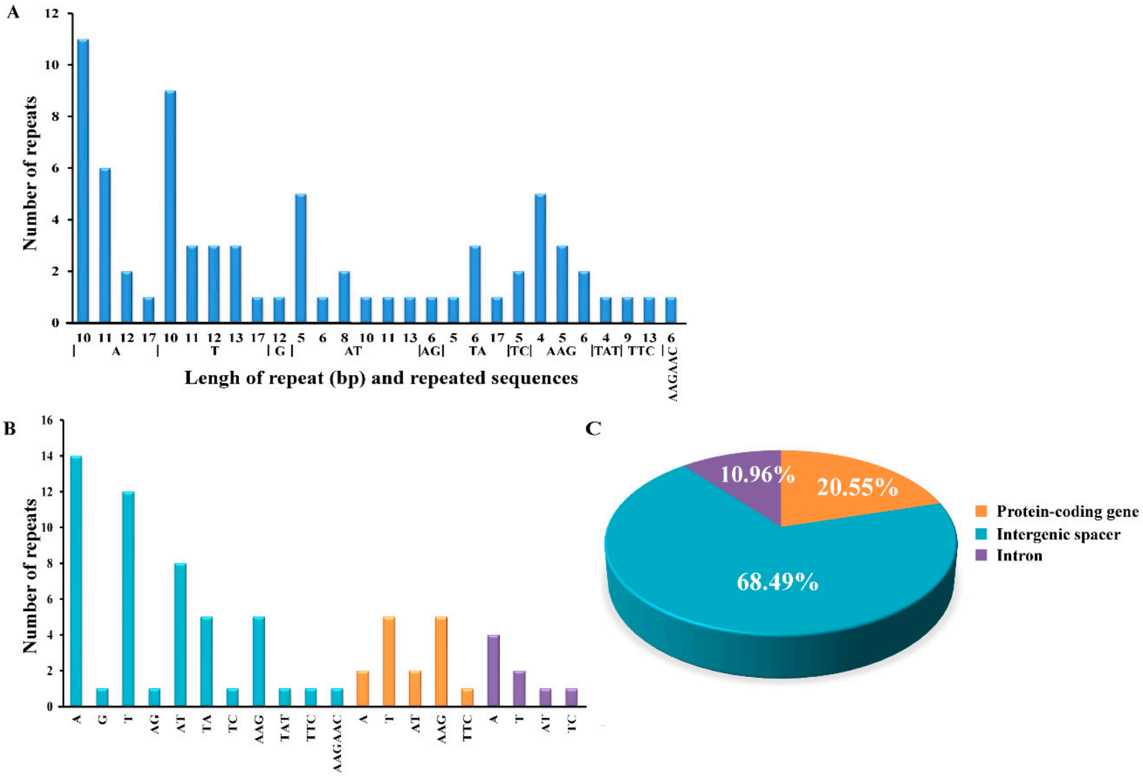

Figure 4. Simple sequence repeats in the cp genome of *F. hodginsii*. (**A**) Statistics of simple sequence repeats (SSRs) classified by sequence length. (**B**) Quantity statistics of SSRs according to their locations. (**C**) The proportion of regions containing SSRs.

When examining repeat location, we noticed that the protein-coding gene *accD* and the intergenic space in its close proximity contain a high number of repeated sequences of various types. This gene encodes acetyl-CoA carboxylase (ACCase), which is one of the key enzymes regulating the rate of fatty acid biosynthesis in plants [36,37]. We also performed an alignment of the *accD* gene from five Cupressophyta members including *F. hodginsii* (Figure 5). The accD reading frame of *F. hodginsii* contains 771 codons. Concerning the repetitive sequence, which most frequently occurs in medial portion, in *F. hodginsii* three repeats of EEEEQ were detected.

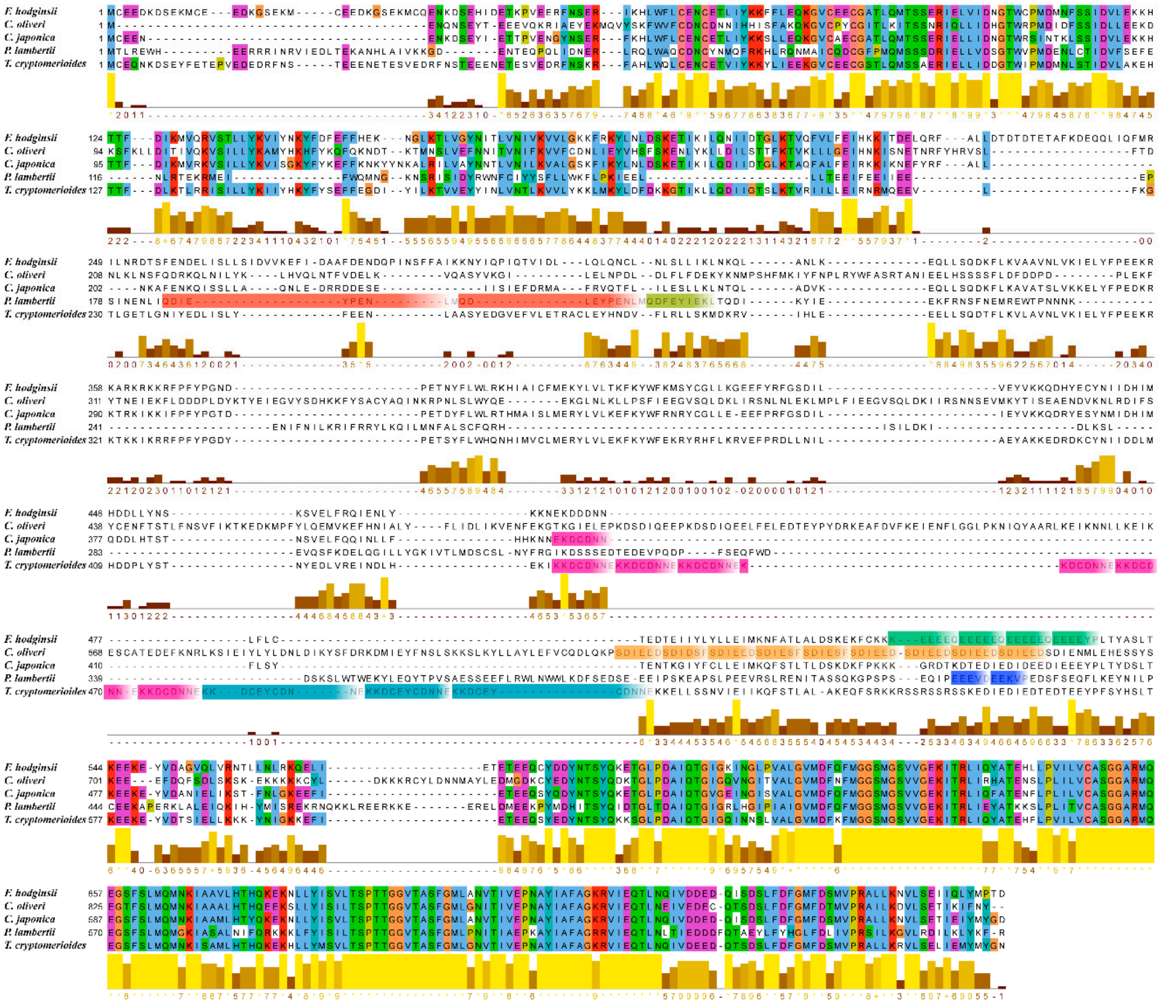

Figure 5. Comparison of *accD* sequences from five Cupressophyta members. Yellow histograms below each line indicate the conservation of aligned sequences. Different repetitive elements are marked with different gradient colored boxes.

3.3. Residual Inverted Repeat (IR) Regions

Several previous studies support that IRs in gymnosperms display a high amount of variation in sequence length and composition [5,11,38]. In order to explore the loss of IR regions in *F. hodginsii*, we compared it with four other conifers, the results of which are shown below (Figure 6). All five conifers conserved one copy of the IR$_B$, and left footprints of the IR$_A$ region with duplicated genes that can be detected in the sequence. We identified two main palindromic repeats in the *F. hodginsii* cp DNA sequence, each containing a duplicated gene: *trnQ-UUG* and *trnI-CAU*. Based on our comparison and preceding analysis, the 86 bp sequence that contains *trnI-CAU* and extends into the intergenic space between the *trnI-CAU* and *psbA* genes is inferred as the residual IR in *F. hodginsii*, which is the same length as the residual IR in *Metasequoia glyptostroboides* Hu et Cheng [39].

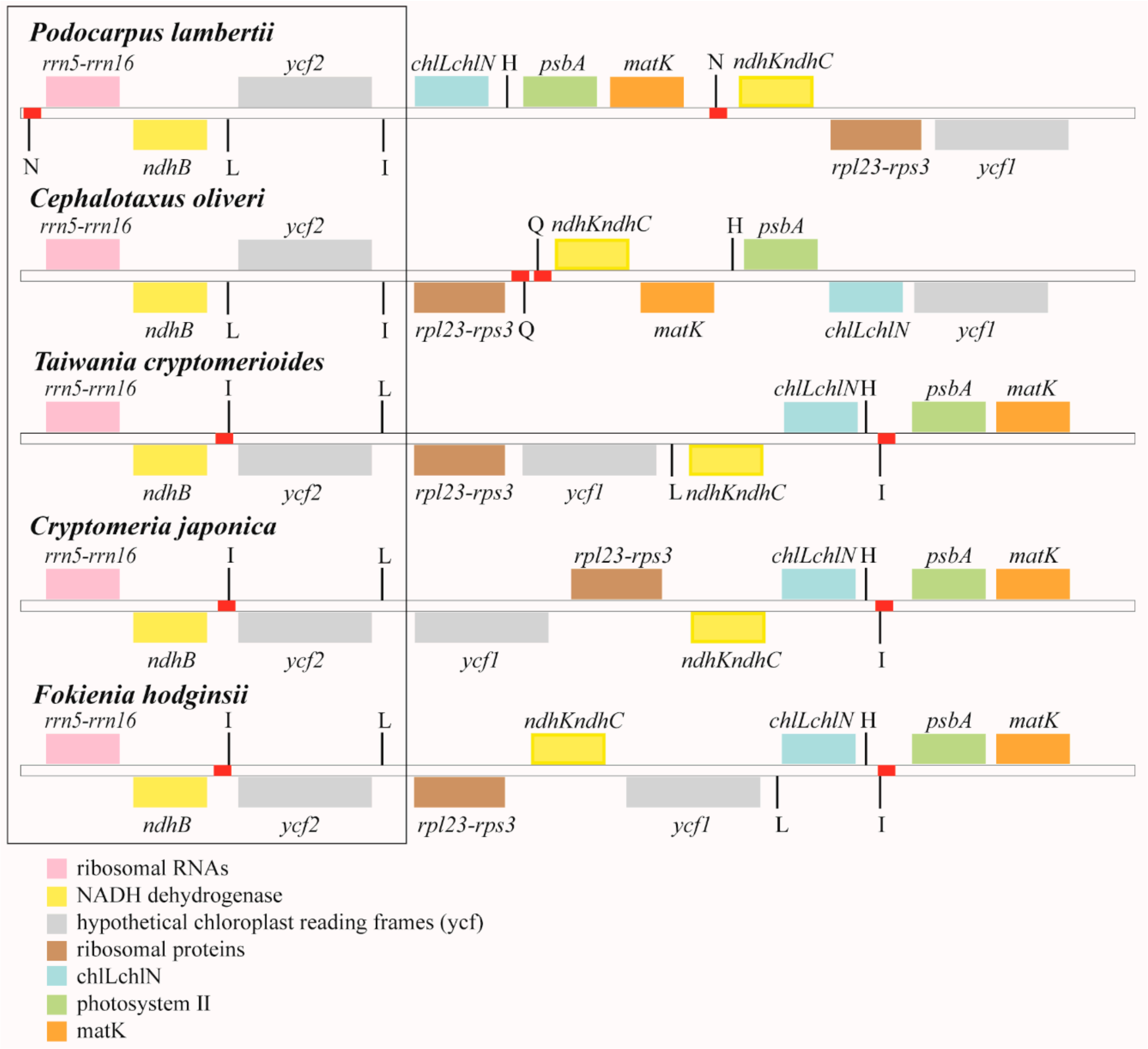

Figure 6. Comparisons of inverted repeat (IR) regions among five species from Cupressophyta. Genes located above the white strip were transcribed counter-clockwise, others below the strip were transcribed in the inverse direction. Different colors are used to distinguish gene functions. Single letters indicate the abbreviation of the transfer RNA. N for *trnN-GUU*, L for *trnL-CAA*, I for *trnI-CAU*, H for *trnH-GUG*, and Q for *trnQ-UUG*. The black frames represent the retained IR regions and the little red boxes show the track of short IR regions.

3.4. Phylogenetic Research

Ginkgo biloba was set as an outgroup for the construction of our phylogenetic tree with 21 plant species selected for this study. In the ML tree, 17 out of 18 nodes show bootstrap values of 100%, accounting for 94.44% (Figure 7), which is identical to our constructed NJ tree. Remarkably, in both topology structures, *F. hodginsii* belongs to the Cupressaceae family with a node bootstrap value of 100% and is placed as a sister clade to *Chamaecyparis hodginsii* (Dunn) Rushforth, indicating that *F. hodginsii* and *C. hodginsii* are very closely related. The 18 Cupressaceae gymnosperms included in our analysis form six sub-families of Cupressophyta (Figure 7).

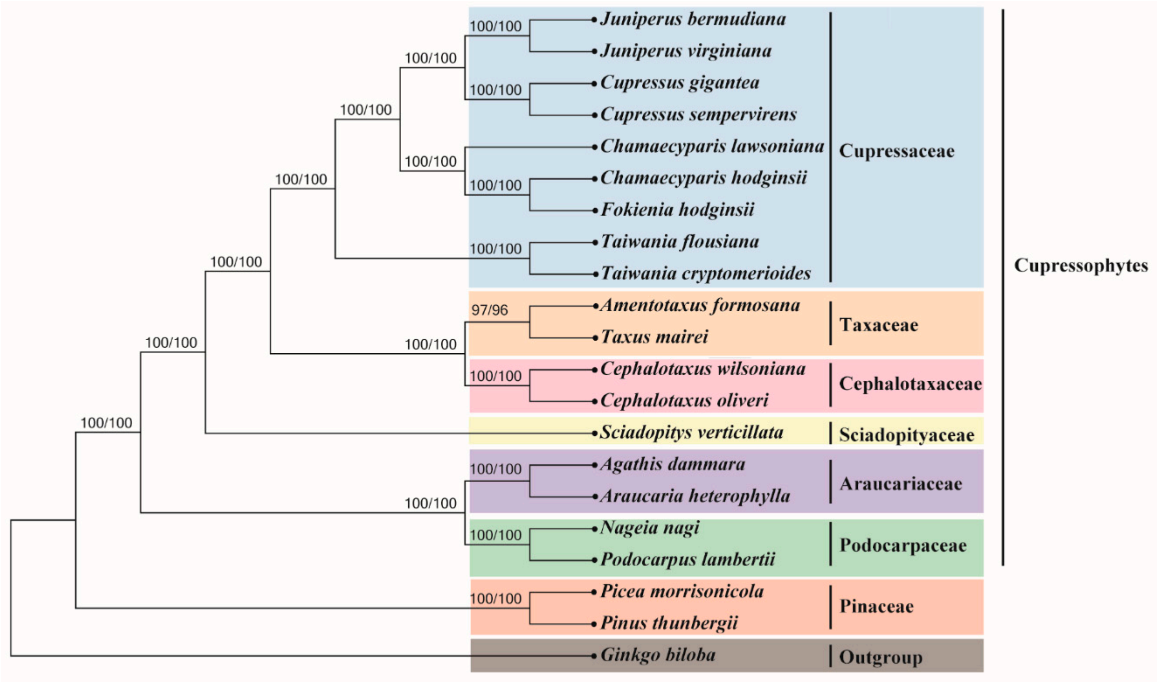

Figure 7. The maximum likelihood (ML) and neighbor-joining (NJ) phylogenetic tree based on 63 eligible genes from 21 species. Numbers above the nodes are bootstrap values. *G. biloba* was set as the outgroup.

4. Discussion

In recent years, with the rapid development of large-scale sequencing technologies, more and more research studies concerning the evolutionary relationships of plants have been carried out. Chloroplast is an organelle specialized for carrying out photosynthesis in plants. Cp genomes have a great deal of merits. The conservatism of the genome content and the relatively small size of the sequence make cp genomes easy to be sequenced. Using single-copy genes located outside the inverted repeat regions, one can effectively avoid the disturbance of paralogs when conducting phylogenetic analyses. Moreover, the variation of the evolutionary rate between coding and noncoding regions can be applied in distinguishing different categories [40]. Based on these benefits, phylogeny has undergone rapid progress with the help of cp genome sequencing.

From the respect of cp genomic studies in Cupressophyta, previous experiments paid enough attention to controversial issues about the taxonomy among Taxaceae, Cephalotaxaceae, and Podocarpaceae as well as the phylogenetic relationships among the genera of these families [41–43]. However, the taxonomic status of three genera from Araucariaceae remains to be solved [44,45]. Chloroplast genes *rbcL* and *matK* were the most commonly used gene segments because their sequences variation contained historical evidence appropriate for evolutionary analysis [46]. As the complete cp genome of more Cupressophyta species were published, new markers and methods can be invented in related fields.

In our research, the complete cp genome of *F. hodginsii* was found to be 129,534 bp in length, which falls in between the 128,290 bp of *Taxus mairei* [47] and the 131,810 bp of *C. japonica* [5]. Compared with Pinaceae cp genomes, which are mainly around 119 kbp [48–50], the Cupressophyta plastomes are of a bigger size. When annotating genes, the absence of *trnR-CCG*, which has been reported in Cupressophyta before and also was found to occur in *F. hodginsii*, supports the hypothesis based on phytochrome phylogenetic trees that the *trnR-CCG* gene was lost during the second split separating Araucariaceae and Podocarpaceae in the evolutionary progress of plants [51].

The comparative alignment of the *accD* gene revealed that the *accD* reading frame of *F. hodginsii* contains 771 codons, slightly more than *C. japonica* (700 codons) [5] and similar to other Cupressophyta species such as *T. cryptomerioides* (800 codons), *P. lambertii* (864 codons), and *C. oliveri* (936 codons). Concerning the repetitive sequence that most frequently occurs in medial portion, in *F. hodginsii*, three repeats of EEEEQ were detected, almost in the same region where nine repeats of SDIEED in *C. oliveri* occur [52]. In *P. lambertii* as well as *T. cryptomerioides*, more than one type of repeat sequence was found. The insertion of repetitive elements is thought to lead to the variation of the whole sequence, in an attempt to accelerate the substitution rate [37]. As indicated by the yellow histograms below each line (Figure 5), the amino acid sequence of both terminals remained more conserved than the middle part. Moreover, the C-terminal appeared as more conserved than the N-terminal. Since the most significant function of this region is considered as encoding the carboxyl transferase, this specific utilization may explain why the C-terminal remains more conserved in some degree. These results support that the *accD* reading frame displays a tendency towards enlarging size in Cupressophyta [6,37,51], which is mainly caused by the great number of insertions that consist of tandemly repeated motifs.

Rearrangements of cp genomes occurred more frequently in conifers and the loss of typical IR sequences may be an outcome of the reduction of gene content. For example, it is believed that the *Pinus thunbergii* Parl IR sequence only contains a region encompassing *trnI* and a portion of 3′*psbA* as a remnant of the whole IR [53]. The residual IR of *C. japonica* is 114 bp in length [5]. Short remaining sequences of 544 bp and 326 bp can be detected in *C. oliveri* [52] and *P. lambertii* [51], respectively. Wu et al. (2011) concluded that in conifers, the IR_B region was retained in Pinaceae whereas the IR_A region was retained in Cupressophyta. The regions encompassing the whole *ycf2* gene and the adjoined *psbA* or *rpl23* genes are considered as ancestral IRs [14]. This conclusion is supported by the 86 bp IR contraction footprint in *F. hodginsii* in our study. Notably, repeat regions still have function after expansion [41]. However, the complex mechanism of the insertion and loss of genes, as well as the induction of shrinkage and expansion in specific regions of cp genomes, remains unclear.

Our phylogenetic topology supports that the gymnosperm representative *G. biloba* forms a monophyletic branch. Gymnosperms are considered to contain three separate clades: Cycadales-Ginkgoales, Gnetales, and Coniferales [54,55]. A previous phylogenetic study using 43 Coniferales species concluded that they were divided into two main sub-clades: Pinaceae and Cupressophyta. The Pinaceae are represented by *Picea morrisonicola* Hayata and *P. thunbergii* in our analysis and are considered basal to the remaining conifer families [55]. Within the Cupressophyta, which is another branch of Coniferales, the relationship between *Sciadopitys verticillate* (Thunb.) Sieb. et Zucc (representing Sciadopityaceae) and the other species stood out. Its position supports that Sciadopityaceae form their own separate family.

5. Conclusions

In this work, we sequenced the complete cp genome of *F. hodginsii* (129,534 bp) by using Illumina high-throughput sequencing technology. Repeat motifs found by our statistical analyses can be further used for the development of molecular markers, which can find broad applications in genetic studies as well as phylogenetic research. We will also be able to exploit new strategies to protect this endangered species based on developed markers. As the only endemic species of *Fokienia* Henry et Thomas, the phylogenetic position of *F. hodginsii* we constructed is of great value in terms of understanding the evolutionary history of this genus and enriching our comprehension of the systematic status of Cupressophyta. However, there are still some problems that remain to be explored, such as the elaborate explanation of different cpDNA forms in Cupressophyta, the principle of the insertion and loss of genes, as well as the mechanism of IR shrinkage and expansion in cp genomes. We believe that the complete cp genome of *F. hodginsii* is an ideal example when studying these issues.

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4907/10/7/528/s1, Table S1: Accession numbers of cpDNA sequences used in the phylogenetic analysis, Table S2: Repeated sequences found in the *F. hodginsii* cp genome., Table S3: Distribution of SSRs in the *F. hodginsii* cp genome.

Author Contributions: Conceptualization, M.Z. and J.C.; methodology, M.Z., Y.W., and L.L.; software, M.Z. and Y.W.; formal analysis, M.Z.; investigation, X.Z. and D.Y.; resources, X.Z. and D.Y.; data curation, M.Z. and L.L.; writing—original draft preparation, M.Z.; writing—review and editing, M.Z. and J.C.; supervision, J.S. and J.C.; project administration, Q.S., J.S., and J.C.; funding acquisition, R.Z., T.C., and J.C.

Funding: This work was supported by the Key Research and Development Plan of Jiangsu Province (BE2017376), the Foundation of Jiangsu Forestry Bureau (LYKJ[2017]42), the Nature Science Foundation of China (31770715), the Program for Fujian Province Science and Technology Research funding on the fourth Tree Breeding Cycle Program of Chinese Fir (Min Lin Ke 2016-35), the Qinglan Project of Jiangsu Province, and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

Acknowledgments: We would like to thank Zhaodong Hao at the Key Laboratory of Forest Genetics and Biotechnology of Ministry of Education of China, Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University (NJFU), for his efforts in offering valuable suggestions about phylogenetic analysis.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gao, Z.W. A precious timber species—*Fokienia hodginsii*. *J. Fujian For. Sci. Tech.* **1994**, *21*, 62–66.
2. Hou, B.X.; YU, G.F.; Lin, F.; Cheng, Z.H. Study on *Fokienia hodginsii* natural wild wood community. *Hunan For. Sci. Tech.* **2004**, 31.
3. Shinozaki, K.; Ohme, M.; Tanaka, M.; Wakasugi, T.; Hayashida, N.; Matsubayashi, T.; Zaita, N.; Chunwongse, J.; Obokata, J.; Shinozaki, K.Y.; et al. The complete nucleotide sequence of the tobacco chloroplast genome: Its gene organization and expression. *Plant Mol. Biol. Rep.* **1986**, *5*, 2043–2049. [CrossRef]
4. Ohyama, K.; Fukuzawa, H.; Kohchi, T.; Shirai, H.; Sano, T.; Sano, S.; Umesono, K.; Shiki, Y.; Takeuchi, M.; Chang, Z.; et al. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature* **1986**, *322*, 572–574. [CrossRef]
5. 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*. [CrossRef] [PubMed]
6. Palmer, J.D. Comparative Organization of Chloroplast Genomes. *Annu. Rev. Genet.* **1985**, *19*, 325–354. [CrossRef]
7. Sugiura, M. The chloroplast chromosomes in land plants. *Annu. Rev. Cell Biol.* **1989**, *5*, 51–70. [CrossRef]
8. Jansen, R.K.; Ruhlman, T.A. Plastid Genomes of Seed Plants. In *Genomics of Chloroplasts and Mitochondria*; Springer: Dordrecht, The Netherlands, 2012; Volume 35, pp. 103–126.
9. Strauss, S.H.; Palmer, J.D.; Howe, G.T.; Doerksen, A.H. Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 3898–3902. [CrossRef]
10. Palmer, J.D.; Thompson, W.F. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell* **1982**, *29*, 537–550. [CrossRef]
11. Palmer, J.D.; Stein, D.B. Conservation of chloroplast genome structure among vascular plants. *Curr. Genet.* **1986**, *10*, 823–833. [CrossRef]
12. Lin, C.P.; Wu, C.S.; Huang, Y.Y.; Chaw, S.M. The Complete Chloroplast Genome of *Ginkgo biloba* Reveals the Mechanism of Inverted Repeat Contraction. *Genome Biol. Evol.* **2012**, *4*, 374–381. [CrossRef] [PubMed]
13. Drescher, A.; Ruf, S.; Calsa, J.T.; Carrer, H.; Bock, R. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *Plant J.* **2010**, *22*, 97–104. [CrossRef]
14. 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*, 1284–1295. [CrossRef] [PubMed]
15. Palmer, J.D. Chloroplast DNA Evolution and Biosystematic Uses of Chloroplast DNA Variation. *Am. Nat.* **1987**, *130*, 6–29. [CrossRef]
16. Sandbrink, J.; Vellekoop, P.; Van, H.R.; Van, B.J. A method for evolutionary studies on RFLP of chloroplast DNA, applicable to a range of plant species. *Biochem. Syst. Ecol.* **1989**, *17*, 45–49. [CrossRef]
17. Zerbino, D.R.; Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **2008**, *18*, 821–829. [CrossRef] [PubMed]

18. Boetzer, M.; Henkel, C.V.; Jansen, H.J.; Butler, D.; Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **2008**, *27*, 578–579. [CrossRef] [PubMed]
19. Wyman, S.K.; Jansen, R.K.; Boore, J.L. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **2004**, *20*, 3252–3255. [CrossRef] [PubMed]
20. Schattner, P.; Brooks, A.N.; Lowe, T.M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **2005**, *33*, 686–689. [CrossRef] [PubMed]
21. 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*, 267–274. [CrossRef] [PubMed]
22. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **2016**, *33*, 7. [CrossRef] [PubMed]
23. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **1999**, *27*, 573–580. [CrossRef] [PubMed]
24. Kurtz, S.; Choudhuri, J.V.; Ohlebusch, E.; Schleiermacher, C.; Stoye, J.; Giegerch, R. REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **2001**, *29*, 4633–4642. [CrossRef] [PubMed]
25. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef] [PubMed]
26. Waterhouse, A.M.; Procter, J.B.; Martin, D.M.A.; Clamp, M.; Barton, G.J. Jalview Version 2 - A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **2009**, *25*, 1189–1191. [CrossRef] [PubMed]
27. Braukmann, T.W.; Kuzmina, M.; Stefanović, S. Loss of all plastid ndh genes in Gnetales and conifers: Extent and evolutionary significance for the seed plant phylogeny. *Curr. Genet.* **2009**, *55*, 323–337. [CrossRef] [PubMed]
28. Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22*, 4673–4680. [CrossRef]
29. Vaidya, G.; Lohman, D.J.; Meier, R. SequenceMatrix: Concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics* **2011**, *27*, 171–180. [CrossRef]
30. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **1980**, *16*, 111–120. [CrossRef]
31. Hildebrand, M.; Hallick, R.B.; Passavant, C.W.; Bourque, D.P. Trans-splicing in chloroplasts: The rps 12 loci of Nicotiana tabacum. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 372–376. [CrossRef]
32. Shimada, H.; Sugiura, M. Fine structural features of the chloroplast genome: Comparison of the sequenced chloroplast genomes. *Nucleic Acids Res.* **1991**, *19*, 983–995. [CrossRef] [PubMed]
33. Eun, H.M. Enzymes and Nucleic Acids: General Principles. In *Enzymology Primer for Recombinant DNA Technology*; Elsevier: Amsterdam, The Netherlands, 1996; pp. 1–108.
34. Provan, J.; Powell, W.; Hollingsworth, P.M. Chloroplast microsatellites: New tools for studies in plant ecology and evolution. *Trends Ecol. Evol.* **2001**, *16*, 142–147. [CrossRef]
35. Cardle, L.; Ramsay, L.; Milbourne, D.; Macaulay, M.; Marshall, D.; Waugh, R. Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* **2000**, *156*, 847–854. [PubMed]
36. Madoka, Y.; Tomizawa, K.I.; Mizoi, J.; Nishida, I.; Nagano, Y.; Sasaki, Y. Chloroplast transformation with modified accD operon increases acetyl-CoA carboxylase and causes extension of leaf longevity and increase in seed yield in tobacco. *Plant Cell Physiol.* **2002**, *43*, 1518. [CrossRef] [PubMed]
37. Li, J.; Su, Y.J.; Wang, T. The Repeat Sequences and Elevated Substitution Rates of the Chloroplast accD Gene in Cupressophytes. *Frontiers Plant Sci.* **2018**, *9*, 533. [CrossRef]
38. Zhang, Y.; Ma, J.; Yang, B.; Li, R.; Zhu, W.; Sun, L.; Tian, J.; Zhang, L. The complete chloroplast genome sequence of *Taxus chinensis* var. mairei (Taxaceae): Loss of an inverted repeat region and comparative analysis with related species. *Gene* **2014**, *540*, 201–209. [CrossRef]
39. Chen, J.; Hao, Z.; Xu, H.; Yang, L.; Liu, G.; Sheng, Y.; Zheng, C.; Zheng, W.; Cheng, T.; Shi, J. The complete chloroplast genome sequence of the relict woody plant *Metasequoia glyptostroboides* Hu et Cheng. *Frontiers Plant Sci.* **2015**, *6*, 447. [CrossRef]

40. Zhang, Y.J.; Li, D.Z. Advances in Phylogenomics Based on Complete Chloroplast Genomes. *Plant Diver. Resour.* **2011**, *33*, 365–375.
41. Chaw, S.M.; Parkinson, C.L.; Cheng, Y.; Vincent, T.M.; Palmer, J.D. Seed plant phylogeny inferred from all three plant genomes: Monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 4086–4091. [CrossRef]
42. Conran, J.G.; Wood, G.M.; Martin, P.G.; Dowd, J.M.; Quinn, C.J.; Gadek, P.A.; Price, R.A. Generic relationships within and between the gymnosperm families Podocarpaceae and Phyllocladaceae based on an analysis of the chloroplast gene rbcL. *Aust. J. Bot.* **2000**, *48*, 715–724. [CrossRef]
43. Wang, X.Q.; Shu, Y.Q. Chloroplast matK gene phylogeny of Taxaceae and Cephalotaxaceae, with additional reference to the systematic position of Nageia. *Acta Phytotax. Sin.* **2000**, *38*, 201–210.
44. Setoguchi, H.; Osawa, T.A.; Pintaud, J.C.; Jaffre, T.; Veillon, J.M. Phylogenetic Relationships within Araucariaceae Based on rbcL Gene Sequences. *Am. J. Bot.* **1998**, *85*, 1507–1516. [CrossRef] [PubMed]
45. Kershaw, P.; Wagstaff, B. The Southern Conifer Family Araucariaceae: History, Status, and Value for Paleoenvironmental Reconstruction. *Annu. Rev. Ecol. Syst.* **2001**, *32*, 397–414. [CrossRef]
46. Chase, M.W.; Soltis, D.E.; Olmstead, R.G.; Morgan, D.; Les, D.H.; Mishler, B.D.; Duvall, M.R.; Price, R.A.; Hills, H.G.; Qiu, Y.L.; et al. Phylogenetics of Seed Plants: An Analysis of Nucleotide Sequences from the Plastid Gene rbcL. *Ann. Mo. Bot. Gard.* **1993**, *80*, 528–548. [CrossRef]
47. Hsu, C.Y.; Wu, C.S.; Chaw, S.M. Ancient nuclear plastid DNA in the yew family (taxaceae). *Genome Biol. Evol.* **2014**, *6*, 2111–2121. [CrossRef] [PubMed]
48. Ni, Z.X.; Ye, Y.J.; Bai, T.; Xu, M.; Xu, L.A. Complete Chloroplast Genome of *Pinus massoniana* (Pinaceae): Gene Rearrangements, Loss of ndh Genes, and Short Inverted Repeats Contraction, Expansion. *Molecules* **2017**, *22*, 1528. [CrossRef] [PubMed]
49. 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. *Conserv. Genet. Resour.* **2016**, *9*, 1–4. [CrossRef]
50. Asaf, S.; Khan, A.L.; Khan, M.A.; Shahzad, R.; Lubna; Kang, S.M.; Harrasi, A.A.; Rawahi, A.A.; Lee, I.J. Complete chloroplast genome sequence and comparative analysis of loblolly pine (*Pinus taeda* L.) with related species. *PLoS ONE* **2018**, *13*, e0192966. [CrossRef]
51. Vieira, L.N.; Faoro, H.; Rogalski, M.; Fraga, H.P.F.; Cardoso, R.L.A.; Souza, E.M.; Pedrosa, F.O.; Nodari, R.O.; Guerra, M.P. The complete chloroplast genome sequence of *Podocarpus lambertii*: Genome structure, evolutionary aspects, gene content and SSR detection. *PLoS ONE* **2014**, *9*, e90618. [CrossRef]
52. Yi, X.; Gao, L.; Wang, B.; Su, Y.J.; 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*, 688–698. [CrossRef]
53. Tsudzuki, J.; Nakashima, K.; Tsudzuki, T.; Hiratsuka, J.; Shibata, M.; Wakasugi, T.; Sugiura, M. 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*, 206–214. [PubMed]
54. Hart, J.A. A CLADISTIC ANALYSIS OF CONIFERS: PRELIMINARY RESULTS. *J Arnold. Arbor.* **1987**, *68*, 269–307.
55. Chaw, S.M.; Zharkikh, A.; Sung, H.M.; Lau, T.C.; Li, W.H. Molecular phylogeny of extant gymnosperms and seed plant evolution: Analysis of nuclear 18S rRNA sequences. *Mol. Biol. Evol.* **1997**, *14*, 56–68. [CrossRef] [PubMed]