PART OF A HIGHLIGHT ON GENOMIC EVOLUTION

Concerted and birth-and-death evolution of 26S ribosomal DNA in *Camellia* L.

Min Zhang^{1,2}, Yi-Wei Tang², Ying Xu^{2,3}, Takahiro Yonezawa⁴, Yang Shao², Yu-Guo Wang², Zhi-Ping Song²,
Ji Yang² and Wen-Ju Zhang^{2,*}

¹Co-Innovation Center for Sustainable Forestry in Southern China, College of Biology and the Environment, Nanjing Forestry University, Nanjing 210037, China, ²Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, Institute of Biodiversity Science, School of Life Sciences, Fudan University, Shanghai 200438, China, ³Shanghai Railway Entry-Exit Inspection and Quarantine Bureau, Shanghai 200070, China and ⁴Faculty of Agriculture, Tokyo University of Agriculture, Funako 1737, Atsugi, Kanagawa 14 243-0034, Japan

*For correspondence. E-mail wjzhang@fudan.edu.cn

Returned for revision: 23 July 2020 Editorial decision: 9 September 2020 Accepted: 15 September 2020
Electronically published: 17 September 2020

- **Background and Aims** The ribosomal DNA (rDNA) gene family, encoding ribosomal RNA (rRNA), has long been regarded as an archetypal example illustrating the model of concerted evolution. However, controversy is arising, as rDNA in many eukaryotic species has been proved to be polymorphic. Here, a metagenomic strategy was applied to detect the intragenomic polymorphism as well as the evolutionary patterns of 26S rDNA across the genus *Camellia*.
- **Methods** Degenerate primer pairs were designed to amplify the 26S rDNA fragments from different *Camellia* species. The amplicons were then paired-end sequenced on the Illumina MiSeq platform.
- **Key Results** An extremely high level of rDNA polymorphism existed universally in *Camellia*. However, functional rDNA was still the major component of the family, and was relatively conserved among different *Camellia* species. Sequence variations mainly came from rRNA pseudogenes and favoured regions that are rich in GC. Specifically, some rRNA pseudogenes have existed in the genome for a long time, and have even experienced several expansion events, which has greatly enriched the abundance of rDNA polymorphism.
- **Conclusions** *Camellia* represents a group in which rDNA is subjected to a mixture of concerted and birth-and-death evolution. Some rRNA pseudogenes may still have potential functions. Conversely, when released from selection constraint, they can evolve in the direction of decreasing GC content and structural stability through a methylation-induced process, and finally be eliminated from the genome.

Key words: *Camellia*, ribosomal DNA, pseudogenes, concerted evolution, birth-and-death evolution.

INTRODUCTION

Ribosomal RNA (rRNA), accounting for ~80 % of cellular RNA and 20 % of total cell dry weight, is closely related to the growth and development of the organism (Gourse *et al.*, 1996). In eukaryotes there are four types of rRNA molecule (18S, 5.8S, 26S and 5S), which are encoded by nuclear ribosomal DNA (rDNA). Generally, the 18S, 5.8S and 26S rRNA genes are transcribed as a single operon (45S rDNA in plants) and tandemly repeat at one or several chromosomal locations. Though they have tens to several thousands of repeats, different rDNA operons within most eukaryotic species are uniform and sequence variations exist only among different species. Consequently rDNA fragments are widely used as barcodes for species identification and phylogenetic analysis (Soltis *et al.*, 1997; Li *et al.*, 2011).

The traditional view holds that the homogeneity of rDNA within species is driven by concerted evolution, which refers to the idea that individual members of the rRNA gene family do not evolve independently but together as a unit by means of gene conversion or unequal crossing-over (Smith, 1974, 1976; Arnheim, 1983). However, this opinion is becoming controversial as the rDNA sequences of more and more species are

being sequenced. Intragenomic polymorphism of rDNA has been reported in many plant species. In gymnosperms, *Cycas revoluta* and some gnetophytes were reported to have an extraordinarily high level of rDNA diversity, unlike *Ginkgo biloba*, *Gnetum gnemon* and other gymnospermous taxa, which show a high degree of rDNA homogeneity (Wang *et al.*, 2016, 2019). In angiosperms, intragenomic polymorphism of rDNA is also pervasive. Muir *et al.* (2001) found that rDNA polymorphism existed in *Quercus* species. Xu *et al.* (2017) even identified large-scale polymorphism of rDNA within individuals of *Lespedeza* (Fabaceae) species. These findings indicate that the evolution of rDNA is more complex than people predicted and worth further study.

Camellia, comprising more than 14 sections and over 100 species, is the biggest genus in Theaceae and is of great economic and phylogenetic value (Chang and Ren, 1998; Ming, 1999). Study of the molecular systematics of *Camellia* is challenging. Vijayan *et al.* (2009) tried to develop a molecular phylogenetic frame for *Camellia* based on internal transcribed spacer (ITS) sequences. However, interspecific relationships among many taxa could not be well resolved, partly because the ITS sequences from 70 % of the studied species harboured

intragenomic variations. Similarly, our previous study on *Camellia sinensis* also detected a high level of intragenomic rDNA polymorphism (Xu *et al.*, 2015). Specifically, by applying a metagenomic strategy, we reported the most abundant rDNA polymorphism of seed plants in three *Camellia* species (Shao *et al.*, 2018). So *Camellia* represents a group of angiosperms in which concerted evolution seems to be ineffective. However, we do not know how widespread this polymorphism is across *Camellia* or what is the evolutionary pattern of rDNA in this genus. In this study, by improving the metagenomic strategy established by our previous study (Shao *et al.*, 2018), we sequenced the 26S rDNA fragments of 13 *Camellia* species that represent the main sections in this genus. Through evolution and sequence variation analysis, we want to reveal the evolutionary pattern of rDNA in *Camellia* and provide new insight into the evolution of rDNA in angiosperms.

MATERIALS AND METHODS

Sample collection

As shown in Table 1, 13 *Camellia* species belonging to eight sections (Ming, 1999) were collected for MiSeq sequencing in this study. Most of this material was collected from the International *Camellia* Species Garden located in Jinhua, Zhejiang province, China, while others were collected from their natural habitats. Voucher specimens of these collections are preserved in the Institute of Biodiversity Science of Fudan University. Fresh leaves of the above species, after collection, were kept in sealable packages containing silica gel and finally transferred to 4 °C for storage.

DNA extraction

A DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA) was used to extract genomic DNA from silica-gel-dried leaves of the above taxa following the manufacturer's instructions. Concentration and purity of the DNA were then checked with a NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA).

Library construction and sequencing

Primer pairs (Table 2) targeting the 3'-end of the 26S rDNA (Fig. 1) were designed based on our previous studies (Xu *et al.*, 2015; Shao *et al.*, 2018). The length of the target fragment was ~370 bp. For each primer pair, a six-nucleotide species-specific barcode sequence selected from the Illumina TruSeq Universal Adapter Index was added at the 5'-end of the forward primer (Table 3), which enabled the sequencing of pooled DNA on the Illumina MiSeq instrument. Five degenerate sites across the forward primer were designed to amplify as many ribotypes as possible.

A high-fidelity thermostable Ex Taq DNA polymerase (TaKaRa, Dalian, China) was used for the PCR. PCR amplifications were performed in 30-μL reactions containing 10 × Ex Taq Buffer (Biocolors, Shanghai, China), 3 μL; 50 mM MgCl₂, 2.4 μL; 10 mM dNTP, 1.8 μL; 10 μM forward primer, 1.5 μL; 10 μM reverse primer, 1.5 μL; TaKaRa Ex Taq (5 U μL⁻¹), 1.0 μL; template genomic DNA (50 ng μL⁻¹), 2 μL; 3d H₂O, 16.8 μL. All reactions were carried out in an ABI thermal cycler (Applied Biosystems, Foster City, CA, USA) with 5 min of initial denaturation at 95 °C and 38 cycles of 45 s at 94 °C, 30 s at 58 °C and 1 min at 72 °C, followed by a final extension for 10 min at 72 °C. The PCR products were purified using a Gel Extraction System B Kit (BioDev-Tech, Beijing, China). Then, the purified amplicons from different taxa were pooled in equimolar concentrations for MiSeq DNA library construction. Finally, the library was paired-end (2 × 300 bp) sequenced on the Illumina MiSeq platform (Illumina, San Diego, CA, USA) by Beijing Genomics Institute (BGI, Shenzhen, China).

Construction of internal control

Differently from our previous study (Shao *et al.*, 2018), two internal controls were set up to evaluate the potential error rate introduced by PCR amplification and sequencing. Three 26S rDNA sequences were cloned from *C. sinensis* var. *assamica* (Supplementary Data Table S1). PCR products were separated by agarose gel electrophoresis. The target fragments (~370 bp) were purified using a Gel Purification Kit (Biotech, Beijing, China) and then cloned into pMD18-T vector (TaKaRa,

TABLE 1. Taxa used for MiSeq sequencing in this study

Sample	Taxon	Section (Ming, 1999)	Section (Chang and Ren, 1998)	Place of collection
Camp	<i>Camellia amplexicaulis</i>	<i>Archecamellia</i>	<i>Longipedicellata</i>	Guangdong, China
Cpet	<i>C. petelotii</i>	<i>Archecamellia</i>	<i>Chrysanthia</i>	Zhejiang, China
Caza	<i>C. azalea</i>	<i>Camellia</i>	<i>Camellia</i>	Guangdong, China
Cche	<i>C. chekiangoleosa</i>	<i>Camellia</i>	<i>Camellia</i>	Zhejiang, China
Csp	<i>C. sp.</i>	<i>Camellia</i>	<i>Camellia</i>	Zhejiang, China
Claw	<i>C. lawii</i>	<i>Eriandria</i>	<i>Eriandria</i>	Zhejiang, China
Cyun	<i>C. yunnanensis</i>	<i>Heterogenea</i>	<i>Stereocarpus</i>	Yunnan, China
Cfur	<i>C. furfuracea</i>	<i>Heterogenea</i>	<i>Furfuracea</i>	Zhejiang, China
Clut	<i>C. luteoflora</i>	<i>Stereocarpus</i>	<i>Luteoflora</i>	Zhejiang, China
Chan	<i>C. handelii</i>	<i>Theopsis</i>	<i>Theopsis</i>	Hunan, China
Ctub	<i>C. tuberculata</i>	<i>Tuberculata</i>	<i>Tuberculata</i>	Zhejiang, China
Cpyx	<i>C. pyxidiacea</i>	<i>Tuberculata</i>	<i>Tuberculata</i>	Zhejiang, China
Csin_sin	<i>C. sinensis</i> var. <i>sinensis</i>	<i>Thea</i>	<i>Thea</i>	Zhejiang, China
Csin_ass	<i>C. sinensis</i> var. <i>assamica</i>	<i>Thea</i>	<i>Thea</i>	Zhejiang, China
Inter_ref1	<i>C. sinensis</i> var. <i>assamica</i>	<i>Thea</i>	<i>Thea</i>	Zhejiang, China
Inter_ref2	<i>C. sinensis</i> var. <i>assamica</i>	<i>Thea</i>	<i>Thea</i>	Zhejiang, China

Dalian, China). The recombinant plasmids were transformed into *Escherichia coli* TOP 10 (Tiangen, Beijing, China) and selected on Luria–Bertani (LB) plates containing ampicillin ($50 \mu\text{g mL}^{-1}$). Positive colonies were selected into a separate test tube with LB broth and incubated overnight at 37°C with shaking. After 12 h the broth was sent to Shanghai Majorbio Bio-Pharm Technology (Shanghai, China) for sequencing. For inter_ref1, one sequence was used as template for PCR and library construction. For inter_ref2, the above three known sequences, pooled in equimolar concentrations, served as template DNA, which could be used to evaluate the potential recombination rate in PCR amplification.

Data filtering, splitting and assembly

To obtain more accurate results, raw data generated from the Illumina MiSeq platform were pre-processed. Specific steps were as follows: (1) truncation of sequence reads with an average quality <20 over a 25-bp sliding window based on the phred algorithm (trimmed reads having $<75\%$ of their original length were removed); (2) removal of reads with adapter contamination (for the 15 bases overlapped by reads and adapter, a maximum of three mismatched bases was allowed) and ambiguous base (N base), as well as reads with low complexity (reads with ten or more consecutive bases the same). Based on the species-specific barcode sequences, reads from different taxa were split. During this process, no mismatch was allowed between the barcode sequences and the generated reads. Then, FLASH version 1.2.11 (Tanja and Salzberg, 2011) software was used to assemble the paired-end reads into one tag based on the overlapping relationship (minimal overlapping length, 15 bp; mismatch ratio of overlapped region, ≤ 0.1). Paired-end reads without overlaps were to be removed. The forward and reverse amplification primers were then mapped to the two ends of the tags. If four consecutive bases at the 3'-end of the primers could completely match with the tags, and the mismatch of the remaining primer was no more than three bases, the tags were to be retained. After this, primer sequences were removed from the tags using an internal script to ensure the accuracy of the following analysis.

TABLE 2. Primer pairs used in this study

Primer name	Primer sequence (5'–3')	Source
Forward primer	*****YRAGAGGAAYYRTTGATT	Shao <i>et al.</i> , 2018
Reverse primer	GGGAGGGASGAATCGGAGCGACGC	Xu <i>et al.</i> , 2015

*Species-specific barcode.

Chimera filtering and operational taxonomic unit clustering

USEARCH (Edgar, 2010) was applied to identify the unique ribotype as well as to calculate its size (number of read pairs uniquely mapped to one ribotype). A ribotype with size 1 was a singleton and would be discarded in the following analysis. To eliminate potential errors introduced by PCR and the subsequent sequencing, ribotypes from different taxa were further classified into operational taxonomic units (OTUs). The clustering method itself was an UCLUST algorithm, implemented as the cluster_smallmem command in USEARCH. Then, the UNOISE algorithm (Edgar and Henrik, 2015), implemented in the unoise3 command, was further used to perform denoising (error correction) of the OTU sequences. A denoised sequence is called a ZOTU (zero-radius OTU). In previous studies, an identity of 97.0 % was always used as the threshold for OTU clustering. In our study, two internal controls were added to evaluate the exact error rate during data processing. The results showed that the similarities between most of the ribotypes and the known sequences in internal controls were $>99.4\%$, i.e. for most of the ribotypes (~ 340 bp in length) there was no more than a two-base mismatch. So an identity of 99.4 % was used as the threshold for OTU clustering. After OTU clustering, there was still some contamination introduced for unknown reasons, which could be eliminated by the following denoising process. The results showed that an identity of 99.4 % for OTU clustering combined with -minsize 8 for denoising had good performance in reflecting the real sequence information of each control (Table 3). We therefore used them as criteria throughout this study. During these processes, chimeras were removed simultaneously, following a *de novo* strategy.

Secondary structure and free energy analyses

According to our previous study, the functional rRNA genes among *Camellia* species are relatively conserved (Xu *et al.*, 2015). In the subsequent analysis, one functional consensus sequence (Supplementary Data Table S1) cloned by our laboratory from *C. sinensis* was used as a basal taxon. MEGA version 6.0 (Tamura *et al.*, 2013) was applied to calculate the pairwise genetic distances between ZOTU sequence representatives and the consensus sequence under the Kimura two-parameter model. The GC content was also calculated using MEGA version 6.0. Then, DAMBE (Xia, 2018) software was used to predict the RNA minimum folding energy (MFE) secondary structure as well as DNA Gibbs free energy (ΔG) based on the ZOTU sequence representatives with a folding temperature of 37°C . For DAMBE, free energy is predicted based on the model described by SantaLucia (1998), which has a negative value. The more negative the value, the more stable the secondary structure.

FIG. 1. Region of 26S rDNA targeted in this study. ETS, external transcribed spacer; IGS, intergenic spacer.

TABLE 3. Summary of data changes during assembly and filtering

Samples	Taxon	Barcode	Reads pairs	Tag number	Tags without primer	Tag utilization ratio (%)	Ribotype	Ribotype discarded singletons	OTU number	ZOTU number	Average length (bp)
Camp	<i>Camellia amplexicaulis</i>	GTCCGC	402 694	402 222	227 242	56.50	42943	9068	1278	64	330 ± 15
Cpet	<i>C. petelotii</i>	ATTCCG	299 859	299 763	244 457	81.55	87333	16 352	1718	128	338 ± 8
Caza	<i>C. azalea</i>	CTTCGA	291 265	291 087	156 867	53.89	32823	7007	501	87	336 ± 11
Cche	<i>C. chekiangoleosa</i>	CCGGTG	172 079	172 011	133 037	77.34	33252	5941	834	78	335 ± 9
Csp	<i>C. sp.</i>	ACATCG	204 688	204 608	141 961	69.38	41247	7263	727	141	336 ± 10
Claw	<i>C. lawii</i>	GTCGTC	212 535	212 270	119 155	56.13	38795	5728	658	214	336 ± 12
Cyun	<i>C. yunnanensis</i>	TCGGGA	273 963	273 850	155 102	56.64	33948	5417	532	74	334 ± 13
Cfur	<i>C. furfuracea</i>	AGTTCC	319 617	319 429	174 514	54.63	41014	8195	417	174	335 ± 14
Clut	<i>C. luteoflora</i>	GATCAG	202 459	202 391	103 802	51.29	26309	4933	476	109	336 ± 10
Chan	<i>C. handelii</i>	GTAGCC	291 686	291 568	195 628	67.10	45 509	8173	689	79	338 ± 9
Ctub	<i>C. tuberculata</i>	GCTACC	146 104	146 013	108 453	74.28	22 942	3603	891	136	337 ± 8
Cpyx	<i>C. pyxidacea</i>	GGCTAC	198 549	198 506	137 733	69.38	24 947	3775	502	68	337 ± 8
Csin_sin	<i>C. sinensis</i> var. <i>sinensis</i>	CCACTC	506 619	505 472	363 350	71.88	79 024	15 798	1818	107	337 ± 4
Csin_ass	<i>C. sinensis</i> var. <i>assamica</i>	AAGCTA	249 059	247 389	139 661	56.45	45 002	7204	584	217	337 ± 11
Inter_ref1	<i>C. sinensis</i> var. <i>assamica</i>	CAGATC	567 398	567 209	454 487	80.13	24 898	3465	33	1	340 ± 0
Inter_ref2	<i>C. sinensis</i> var. <i>assamica</i>	AGGAAT	229 696	229 606	182 589	79.52	16 818	3234	94	3	340 ± 0

Phylogenetic analysis of the ZOTUs

Phylogenetic relationships among sequences were estimated with maximum likelihood. In addition to the ZOTU sequence representatives, we also collected 195 rDNA sequences (SubmissionID SUB7189345, accession numbers MT249843–MT250041) previously cloned by our laboratory from five genera (*Camellia*, *Stewartia*, *Schima*, *Ternstroemia* and *Tutcheria*) of Theaceae (Supplementary Data Tables S1 and S2) and downloaded four rDNA sequences of Theaceae and Pentaphylacaceae from GenBank (AY727952–AY727954 and AY727977). In total, 1876 sequences involving 26 taxa were included in the phylogenetic analysis of this study. The sequences were aligned using MUSCLE v. 3.8.31 (Edgar, 2004). Then, phylogenetic inference was performed in RAxML version 8 (Stamatakis, 2014) under the HKY85 model with γ distribution, taking the sequences from Pentaphylacaceae as an outgroup. To assess the confidence of the internal nodes, a rapid bootstrap method (Stamatakis *et al.*, 2008) was applied with 1000 replications.

RESULTS

Data assembly and filtering

After pre-filtering, 4 568 270 clean paired reads were obtained. The data utilization rate was ~75 %. All sequencing data were deposited in the National Center for Biotechnology Information Short Read Archive under BioProject ID PRJNA613609 (accession number SRP253494). Based on the barcode sequences, 146 104–402 694 paired reads, with a mean of 285 517, were assigned to the individual samples (Table 3). Most of the paired reads (>99.3 %) could be assembled into a tag. When primer sequences were removed, 103 802–454 487 tags could be kept for the downstream analysis, and the tag utilization rate ranged from 51.29 to 81.55 %. Finally, 22 942–87 333 ribotypes were obtained, with a mean of 42 506 for each sample. To make the data more credible, singletons were further discarded. As shown in Table 3, >80 % of the ribotypes were singletons.

Intragenomic polymorphism of 26S rDNA in *Camellia*

Ribotypes from different samples were further classed into OTUs based on the criteria established by the internal controls. As shown in Table 3, 417–1818 OTUs, with a mean of 830, were identified in the studied taxa. Then, denoising was performed to eliminate the potential contamination. Finally, 64–217 ZOTUs, representing a high level of intragenomic polymorphism, were obtained from these taxa. The functional sequences of 26S rDNA in *Camellia* are relatively conserved. We therefore compared the genetic distance between ZOTU sequence representatives and a functional consensus sequence cloned in our laboratory (Supplementary Data Table S1). The results showed that most of the representative sequences had variation of more than ten bases (genetic distance >0.032) compared with the consensus sequence. Taking *C. sinensis* as an example, the values of genetic distance between the ZOTU representative sequences and the consensus sequence ranged from 0.000 to 0.196, and 85 % of them were >0.032. According to

our previous studies (Xu *et al.*, 2015; Shao *et al.*, 2018), most of these sequences were from pseudogenes. However, more Miseq read pairs were from the functional sequence than from other ZOTU representative sequences. The functional sequence was more likely to be detected in most of the studied species compared with rRNA pseudogenes (Fig. 2).

Distribution of single-nucleotide polymorphisms

The GC contents of the ZOTU representative sequences were calculated. The result showed that the GC content of these sequences varied considerably within a species. Taking *C. sinensis* as an example, GC ranged from 47.35 to 61.76 %. We detected changes in the GC content of the ZOTU sequence representatives with their genetic distance to the functional consensus sequence. As shown in Fig. 3A, significant correlation was found between GC content and genetic distance, which could be formalized as $y = -72.308x + 62.067$ ($R^2 = 0.924$). Of which, y represents GC content, while x represents genetic distance. Taking all the ZOTU representative sequences from different species for fitting, a similar formula was obtained (Fig. 3B), formalized as $y = -71.862x + 61.988$ ($R^2 = 0.934$). Thus, GC content decreased linearly as ZOTUs became more divergent from the consensus sequence. The distribution of single-nucleotide polymorphisms across the fragment was analysed. As shown in Fig. 3C, most of the mutations were located at sites that were C (cytosine) and G (guanine) in the consensus sequence, indicating that the mutation rates at these sites were significantly higher than those at A (adenine) and T (thymine). In addition, the vast majority of mutations were C→T and G→A transitions (Fig. 3D). Significantly, deletion mutations were more likely to happen at regions that are rich in GC. In the 26s rDNA fragment of this study, there were two regions with a high deletion mutation rate. One was located at 117–125 bp and another at 180–212 bp (Fig. 3C). Both of these regions had a high GC content. Interestingly, as shown in Fig. 4, ribotypes with deletion mutation tended to have a greater genetic distance to the functional sequence than those without deletion mutation,

indicating that the median genetic distance to the functional sequence for the former was significantly higher than that for the latter (Wilcoxon's rank-sum test, $P < 2.2e^{-16}$).

Changes in secondary structure and Gibbs free energy

DAMBE software was used to predict the RNA MFE secondary structure and rDNA stability measured by Gibbs free energy (ΔG). We analysed how MFE changed with divergence from the consensus sequence. The results showed that as ZOTU representative sequences became more divergent from the functional sequence, the absolute value of MFE declined linearly (Fig. 5A). A strong linear relationship was found ($R^2 = 0.626$) between MFE and the genetic distance. A similar trend was also observable for Gibbs free energy, with $R^2 = 0.368$. When fitting ZOTU representative sequences from all the studied species, the linear relationship between MFE, ΔG and genetic distance became more robust, with R^2 of 0.652 and 0.478, respectively (Fig. 5B). Compared with the consensus sequence, most of the rRNAs encoded by ZOTU sequence representatives exhibited less compact secondary structure (Fig. 5C, D), i.e. the stability of these rRNAs decreased gradually as they diverged from the consensus sequence.

Phylogeny and evolution of 26S rDNA in *Camellia*

The maximum likelihood tree of the ZOTU sequence representatives was inferred (Fig. 6A). In total, 1876 sequences from 26 taxa were included in the tree with sequences from Pentaphragmaceae as an outgroup. Subtrees were formed by dividing the ZOTU representative sequences into different clades. However, most of them were not highly supported by the bootstrap test. Phylogenetic resolution within *Camellia* in this tree was poor, and the relationships among most taxa were not resolved. High bootstrap values appeared only at some terminal branches. In these branches, *C. handelii* from section *Theopsis* and *C. lawii* from *Eriandria* in Ming's (1999) system could always be clustered together with a high support rate. Sequences from the five species (*C. azalea*, *C. chekiangoleosa*, *C. saluenensis*, *C. pitardii* and *C. sp.*) of section *Camellia* could always be classified into the same cluster. *Camellia tuberculata* and *C. pyxidiacea*, two closely related species, were highly supported to be clustered together. This supports the systematic delimitation established by both Ming (1999) and Chang and Ren (1998).

The functional 26S rDNA sequences among *Camellia* species were relatively conserved and distributed around the consensus sequence in the tree. Other sequences, mostly from pseudogenes, formed many branches and dispersed across the tree (Fig. 6A). These branches had divergent branch lengths, and the phylogenetic relationships between these sequences and the consensus sequences were unclear. However, there were still some highly supported groups (HSGs) at the terminal of the tree, as shown in Fig. 6B, C. HSG-I could further be clustered into two subclades. The two subclades both contained species from section *Camellia*, section *Heterogenea*, section *Eriandria*, section *Theopsis*, section *Thea* and section *Archecamellia* according to Ming's (1999) system, i.e. one duplication event happened before their divergence

FIG. 2. Change in detection rate of ZOTU sequence representatives with their genetic distance to the functional sequence. Coloured dots show ZOTU sequence representatives from different accessions used for MiSeq sequencing in this study.

FIG. 3. Mutation types and their frequencies across the studied rDNA region in *Camellia*. (A, B) Changes in GC content of rRNA pseudogenes with their genetic distance to the functional sequence for *C. sinensis* (A) and all studied *Camellia* species (B). (C) Distribution of mutation and miss rates at each site of the studied rDNA region. Green, red, blue and purple dots represent the four bases (adenine, thymine, guanine and cytosine, respectively) in the functional sequence. The yellow line shows the distribution of miss rate for each site across the studied region. (D) Frequency of different mutation types across the studied region. Statistical analysis was based on all the ribotypes detected in this study with the functional sequence as reference. ‘-’ represents a deletion mutation.

from each other. Similar patterns were also found in HSG-II, and at least one duplication event happened before the divergence of section *Eriandria*, section *Theopsis* and section *Camellia*. In addition, duplication also happened repeatedly within species. In HSG-I, *C. azalea* from section *Camellia*, *C. furfuracea* from section *Heterogenea*, *C. lawii* from section *Eriandria*, *C. handelii* from section *Theopsis*, as well as *C. petelotii* and *C. amplexicaulis* from section *Archecamellia*, experienced several duplication events (Fig. 6). In HSG-II, three species (*C. azalea*, *C. chekiangoleosa* and *C. sp.*) from section *Camellia*, *C. lawii* from section *Eriandria* and *C. amplexicaulis* from section *Archecamellia* experienced many duplication events (Fig. 6).

DISCUSSION

Widespread rDNA polymorphism in *Camellia*

In previous studies (Xu *et al.*, 2015; Shao *et al.*, 2018) we have revealed the intragenomic polymorphism of rDNA in some

Camellia species. A new finding here is that rDNA polymorphism is extremely high and is pervasive in the genus *Camellia*. In this study, an improved metagenomic method was applied to detect the polymorphism of rDNA. In previous studies, an identity of 97.0 % was always used as the threshold for OTU clustering. If using this parameter value (97.0 %), 29–154 ZOTUs were detected in the species of our study (Supplementary Data Table S3). Since mismatches are common, two internal controls were set up to evaluate the exact error rate during data processing. The results showed that the similarity between most of the ribotypes and the known sequences in internal controls was >99.4 %, i.e. for most of the ribotypes (~340 bp in length) there were no more than two bases mismatched. We therefore used an identity of 99.4 % for OTU clustering. Finally, 64–217 ZOTUs were detected from individuals of different species (Table 3). To our knowledge, this represents the most abundant rDNA polymorphism in angiosperms.

The question thus arises as to why *Camellia* harbours such an abundance of rDNA copies. For rDNA and other multigene

families, sequence diversity might be associated with genome size (Kelly *et al.*, 2015), since species with large genomes always tend to be rich in repeat sequences (Hancock, 2002). In this study, all the species used for Illumina MiSeq sequencing are diploid (Ming, 2000). However, the rDNA polymorphism detected in these species was divergent, and there seemed to be no significant linear relationship between genome size and rDNA diversity, as *C. luteoflora* ($2C = 2.54$ Gb) processed more ZOTUs than *C. chekiangoleosa* ($2C = 4.94$ Gb) (Huang *et al.*, 2013). In addition, the abundant polymorphism could also be related to the number of rDNA loci. In theory, species with multiple loci should have more diverged rDNA sequences (Schlötterer *et al.*, 1994). Considering that *Camellia* species have much more limited rDNA loci but a higher level of rDNA

polymorphism than some gymnosperms (Gu and Xiao, 2003; Islam-Faridi *et al.*, 2007), neither genome size nor number of loci accounts for their exceptionally high level of rDNA polymorphism, just as Wang *et al.* (2016) concluded from *Cycas revoluta*. However, within *Camellia* the detected rDNA polymorphism seemed to be positively associated with rDNA loci. Existing research shows that the number of rDNA loci among *Camellia* species is variable (Gu and Xiao, 2003; Vijayan *et al.*, 2012; Heitkam *et al.*, 2015). Our previous study using fluorescence *in situ* hybridization (FISH) revealed that 45S rDNA was mainly distributed in three loci (six chromosomes) in *C. sinensis* (Xu *et al.*, 2015), which was more than in *C. yunnanensis* (two loci and four chromosomes), as revealed by Gu and Xiao (2003). Meanwhile, the rDNA polymorphism detected was significantly higher in *C. sinensis* than in *C. yunnanensis* (Table 3). This means that *Camellia* species with more rDNA loci tend to have more diverged rDNA sequences. However, what is surprising is that individuals within one species could harbour different numbers of rDNA copies, just like *C. sinensis* (Table 3). We speculate that rDNA loci within species may also be variable in *Camellia*, as reported in *C. japonica* by different researchers (Gu and Xiao, 2003; Vijayan *et al.*, 2012; Heitkam *et al.*, 2015). This is worthy of further research.

In addition, the complex evolutionary history is an important reason for the extremely high level of rDNA polymorphism in *Camellia*. As described above, molecular phylogenies of *Camellia* reconstructed using nuclear (Vijayan *et al.*, 2009) and chloroplast (Yang *et al.*, 2013; Huang *et al.*, 2014) sequences were always in conflict with the results of morphological studies (Chang and Ren, 1998; Ming, 1999). Hybridization and polyploidization may occur frequently during the evolution of *Camellia* and make this genus extremely complex. So it is difficult to identify the orthologous rDNA sequences and reconstruct the molecular phylogeny of this genus, though there is still some evolutionary information hiding in the rDNA polymorphism (Fig. 6).

Another factor contributing to the rDNA polymorphism in *Camellia* is the presence of pseudogenes. According to previous criteria (Xu *et al.*, 2015), most of the ribotypes detected in this study were from rDNA pseudogenes. Significantly, some rRNA pseudogenes have experienced several instances

FIG. 4. Boxplot showing the distribution of genetic distance to the functional sequence of ZOTU sequence representatives with and without deletion mutation. AZOTU, all ZOTU sequence representatives; ZOTUW, ZOTU sequence representatives with deletion mutation; ZOTUWO, ZOTU sequence representatives without deletion mutation; ***Significant difference between medians of two data sets; Open circle, outlier in each data set.

FIG. 5. Comparison of free energy and secondary structure between 26S rRNA pseudogenes and the functional sequence in *Camellia*. (A, B) Changes in free energy of rRNA pseudogenes with genetic distance to the functional sequence for *C. sinensis* (A) and all the studied *Camellia* species (B). (C, D) RNA MFE secondary structure of the functional sequence (C) and that of a typical rRNA pseudogene (D).

FIG. 6. Phylogenetic relationship of the 26 rDNA sequences detected in this study. (A) Maximum likelihood tree of the representative ribotypes of ZOTUs detected in different *Camellia* species. In total, 1876 rDNA sequences of 26 species (19 from *Camellia*, eight from other genera) are included in this tree. The red dot shows the functional sequence; green and blue dots show two highly supported groups (HSGs). (B) and (C) Two highly-supported groups in (A).

of expansion (Fig. 6), some having happened recently, which has greatly enriched the abundance of rDNA in *Camellia* (Table 3). The expansions may involve whole-genome

duplication or be driven by unknown causes. A reasonable explanation for this phenomenon is that, like other protein-coding genes, some rRNA pseudogenes may still have

potential functions (Tam *et al.*, 2008; Saka *et al.*, 2013), so they could stay in the genome for a long time and continue their expansion. We know that rRNA is indispensable for protein synthesis. In addition, recent studies have documented that rRNA could also regulate hundreds of genes and was a valuable source of phenotypic variation (Li *et al.*, 2018). Different from the consensus functional gene, rRNA pseudogenes are functionally less constrained and have accumulated more mutations. If they have some functions in gene regulation, this would allow more rapid functional diversification (Yano *et al.*, 2004). In fact, our previous study did find that some rRNA pseudogenes were actively transcribed (Xu *et al.*, 2015). So we speculate that some rRNA pseudogenes may play an important role in accelerating phenotypic changes as well as genomic evolution, which needs further study.

A mixed model for the evolution of 26S rDNA in *Camellia*

For the evolution of multigene families there are two classic models, i.e. concerted evolution and birth-and-death evolution. In the former, all member genes of a family evolve as a unit in concert. In the latter, new genes are created by gene duplication, and some duplicate genes could stay in the genome for a long time, whereas others would be deleted or become non-functional during the following evolutionary processes (Nei *et al.*, 1997). The rRNA genes have long been regarded as a typical example of genes subjected to concerted evolution, since rDNA copies within most eukaryotes are conserved. However, *Camellia* represents a group in which concerted evolution seems to be somewhat ineffective. As shown in Table 3, sequence variations of rDNA are ubiquitous in *Camellia* species. Significantly, this study presents some new features. (1) The functional sequences were more likely to be detected (Fig. 2). (2) The functional sequences among most *Camellia* species were relatively conserved and even identical. (3) The high level of rDNA polymorphism was mainly due to the presence of pseudogenes. Nearly all the representative sequences of ZOTUs were from rDNA pseudogenes. (4) rDNA pseudogenes formed many highly supported branches. Many branches consisted of rDNA pseudogenes from different sections (Fig. 6B, C). These pseudogenes have existed in the genome since the divergence of these sections. However, they were not eliminated but continued their expansion. According to these results, we deduce that functional rRNA genes still make up the majority of rDNA copies in the genomes of *Camellia*. This is in line with expectation as rRNA is very much needed for the biosynthesis of proteins in the organism and is intolerant of sequence variation (Dahlberg, 1989). From this perspective, concerted evolution still plays an important role in the homogenizing of thousands of rDNA copies in *Camellia* through gene conversion or unequal crossing-over (Smith, 1974; Arnheim, 1983). However, concerted evolution cannot explain the high variability of rDNA in *Camellia*. In this case, the presence and expansion of rRNA pseudogenes indicates that the birth-and-death model provides an important complementary mechanism for the evolution of rDNA in *Camellia*. A potential process may be that, after duplication, some rDNA copies experience mutation and neofunctionalization, so they can stay in the genome and continue their expansion, whereas others without function would be deleted from the genome gradually during

the following evolutionary processes. Thus, the rRNA genes in *Camellia* follow a mixed model of concerted evolution and birth-and-death evolution.

The destiny of rRNA pseudogenes in *Camellia*

Pseudogenes have been considered as ‘molecular fossils’ (Perna and Kocher, 1996). In fact, pseudogenes are more like ‘remains’ buried in the genome, which would decay with time. Since the functional rRNA genes among most *Camellia* species were relatively conserved and even identical, we could examine the evolutionary history of rRNA pseudogenes by comparing their sequences with that of the functional one. In this study, we analysed the sequence and structural characters of rDNA pseudogenes and found that the GC content of the pseudogenes decreased linearly as they became more divergent from the functional ones (Fig. 3). A main factor contributing to this result is that transitions (mainly C→T and G→A) existed previously in the rRNA pseudogenes compared with the functional sequence (Fig. 3). Interestingly, transitions favour regions that are rich in GC. According to previous studies, these regions are always heavily methylated (Varriale and Bernardi, 2010). Thus, the formation of rRNA pseudogenes may be closely related to DNA methylation. Wang *et al.* (2016) proposed a model for methylation-induced base substitution. In this model, methylation promotes transitions from GC to AT, prevents gene conversion, and finally makes the rDNA sequences increasingly AT-rich and divergent. Our results support this hypothesis. In addition, we found that the miss rate was obviously higher in regions that are rich in GC. Significantly, as shown in Fig. 4, pseudogenes with deletion mutations tended to have a greater genetic distance to the functional sequence than those without deletion mutations. We speculate that deletion mutations represent the last evolutionary stage of a region, which are becoming AC-rich after methylation induced substitutions. Accordingly, the MFA and DNA Gibbs free energy of these

Fig. 7. A hypothetical evolutionary model of ribosomal DNA in *Camellia*. Triangles show rDNA clades with divergent branch length. Longitudinal span of triangle shows the relative copy number of rDNA in each clade. Some of the rRNA pseudogenes marked with an asterisk (*) may continue their expansion and have potential function.

sequences would become lower, and the sequences themselves would become unstable (Fig. 5). So when rRNA pseudogenes escaped from selection they would develop in the following direction: increased methylation-induced substitutions, low GC content, reduced secondary structure stability, and finally disappearance from the genome.

Conclusions

In this study, by applying a metagenomic strategy we detected intragenomic polymorphism of rDNA across the genus of *Camellia*. The results showed that an extremely high level of rDNA polymorphism existed previously in *Camellia*. However, the functional rRNA genes among different *Camellia* species were relatively conserved, and sequence variations mainly came from rRNA pseudogenes. *Camellia* represents a group in which rDNA is subjected to a mixture of concerted and birth-and-death evolution. Some rRNA pseudogenes may have potential functions. When they escape from selection constraint, pseudogenes would evolve in the direction of decreasing GC content, free energy and structural stability, and would finally be eliminated from the genome (Fig. 7).

SUPPLEMENTARY DATA

Supplementary data are available online at <https://academic.oup.com/aob> and consist of the following. Table S1: 26S rDNA sequences cloned in this study. Table S2: taxa used for Sanger sequencing in this study. Table S3: detected ZOTU numbers with an identity of 97.0 %.

FUNDING

This work was supported by the National Natural Science Foundation of China (grants 31670223 and 31270407) and the Key Research and Development Program of Guangdong Province (grant 2018B020200002). The authors have no conflicts of interest to declare.

LITERATURE CITED

- Arnheim N. 1983. Concerted evolution of multigene families. In: Nei M, Koehn RK, eds. *Evolution of genes and proteins*. Sunderland, MA: Sinauer Associates, 38–61.
- Chang HD, Ren SX. 1998. *Flora reipublicae popularis sinicae. Delectis florum reipublicae popularis sinicae, agenda academiae sinicae edita*. Beijing: Science Press, 101–133.
- Dahlberg AE. 1989. The functional role of ribosomal RNA in protein synthesis. *Cell* 57: 525–529.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.
- Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461.
- Edgar RC, Henrik F. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* 31: 3476–3482.
- Gourse RL, Gaal T, Bartlett MS, Appleman JA, Ross W. 1996. rRNA transcription and growth rate-dependent regulation of ribosome synthesis in *Escherichia coli*. *Annual Review of Microbiology* 50: 645–677.
- Gu Z, Xiao H. 2003. Physical mapping of the 18S–26S rDNA by fluorescent in situ hybridization (FISH) in *Camellia reticulata* polyploid complex (Theaceae). *Plant Science* 164: 279–285.
- Hancock JM. 2002. Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. *Genetica* 115: 93–103.
- Heitkam T, Petrasch S, Zakrzewski F, et al. 2015. Next-generation sequencing reveals differentially amplified tandem repeats as a major genome component of Northern Europe's oldest *Camellia japonica*. *Chromosome Research* 23: 791–806.
- Huang H, Shi C, Liu Y, Mao SY, Gao LZ. 2014. Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evolutionary Biology* 14: 151.
- Huang H, Tong Y, Zhang QJ, Gao LZ. 2013. Genome size variation among and within *Camellia* species by using flow cytometric analysis. *PLoS ONE* 8: e64981.
- Islam-Faridi MN, Nelson CD, Kubisiak TL. 2007. Reference karyotype and cytomolecular map for loblolly pine (*Pinus taeda* L.). *Genome* 50: 241–251.
- Kelly LJ, Renny-Beyfield S, Pellicer J, et al. 2015. Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytologist* 208: 596–607.
- Li B, Kremling KAG, Wu P, et al. 2018. Core regulation of ribosomal RNA with hundreds of genes contributes to phenotypic variation. *Genome Research* 28: 1555–1565.
- Li DZ, Gao LM, Li HT, et al. 2011. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences of the USA* 108: 19641–19646.
- Ming TL. 1999. A systematic synopsis of the genus *Camellia*. *Acta Botanica Yunnanica* 21: 149–159.
- Ming TL. 2000. *Monograph of the genus Camellia*. Kunming: Yunnan Science and Technology Press.
- Muir G, Fleming CC, Schlötterer C. 2001. Three divergent rDNA clusters predate the species divergence in *Quercus petraea* (Matt.) Liebl. and *Quercus robur* L. *Molecular Biology and Evolution* 18: 112–119.
- Nei M, Gu X, Sitnikova T. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences of the USA* 94: 7799–7806.
- Perna NT, Kocher TD. 1996. Mitochondrial DNA: molecular fossils in the nucleus. *Current Biology* 6: 128–129.
- Saka K, Ide S, Ganley AR, Kobayashi T. 2013. Cellular senescence in yeast is regulated by rDNA noncoding transcription. *Current Biology* 23: 1794–1798.
- Schlötterer C, Hauser MT, von Haeseler A, Tautz D. 1994. Comparative evolutionary analysis of rDNA ITS regions in *Drosophila*. *Molecular Biology and Evolution* 11: 513–522.
- Shao Y, Zhang M, Xu Y, et al. 2018. An improved metagenomic strategy reveals an unprecedentedly high level of intragenomic polymorphism of ribosomal DNA in three species of *Camellia*. *Journal of Systematics and Evolution* 56: 250–258.
- Smith GP. 1974. Unequal crossover and the evolution of multigene families. *Cold Spring Harbor Symposia on Quantitative Biology* 38: 507–513.
- Smith GP. 1976. Evolution of repeated DNA sequences by unequal crossover. *Science* 191: 528–535.
- Soltis DE, Soltis PS, Nickrent DL, et al. 1997. Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. *Annals of the Missouri Botanical Garden* 84: 1–49.
- SantaLucia J. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences of the USA* 95: 1460–1465.
- Stamatakis A. 2014. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.
- Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAXML web servers. *Systematic Biology* 57: 758–771.
- Tam OH, Aravin AA, Stein P, et al. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* 453: 534.
- Tamura K, Stecher G, Peterson D, Filipowski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* 30: 2725–2729.
- Tanja M, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27: 2957–2963.

- Varriale A, Bernardi G. 2010.** Distribution of DNA methylation, CpGs, and CpG islands in human isochores. *Genomics* **95**: 25–28.
- Vijayan K, Zhang WJ, Tsou CH. 2009.** Molecular taxonomy of *Camellia* (Theaceae) inferred from nrITS sequences. *American Journal of Botany* **96**: 1348–1360.
- Vijayan K, Chung M, Tsou C. 2012.** Dispersion of rDNA loci and its implications on intragenomic variability and phylogenetic studies in *Camellia*. *Scientia Horticulturae* **137**: 59–68.
- Wang W, Ma L, Becher H, et al. 2016.** Astonishing 35S rDNA diversity in the gymnosperm species *Cycas revoluta* Thunb. *Chromosoma* **125**: 683–699.
- Wang W, Wan T, Becher H, et al. 2019.** Remarkable variation of ribosomal DNA organization and copy number in gnetophytes, a distinct lineage of gymnosperms. *Annals of Botany* **123**: 767–781.
- Xia X. 2018.** DAMBE7: new and improved tools for data analysis in molecular biology and evolution. *Molecular Biology and Evolution* **35**: 1550–1552.
- Xu B, Zeng XM, Gao XF, Jin DP, Zhang LB. 2017.** ITS non-concerted evolution and rampant hybridization in the legume genus *Lespedeza* (Fabaceae). *Scientific Reports* **7**: 40057.
- Xu J, Xu Y, Yonezawa T, et al. 2015.** Polymorphism and evolution of ribosomal DNA in tea (*Camellia sinensis*, Theaceae). *Molecular Phylogenetics and Evolution* **89**: 63–72.
- Yang JB, Yang SX, Li HT, Yang J, Li DZ. 2013.** Comparative chloroplast genomes of *Camellia* species. *PLoS ONE* **8**: e73053.
- Yano Y, Saito R, Yoshida N, et al. 2004.** A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. *Journal of Molecular Medicine* **82**: 414–422.