

SCIENTIFIC REPORTS



OPEN

Comparison of intraspecific, interspecific and intergeneric chloroplast diversity in Cycads

Guo-Feng Jiang*, Damien Daniel Hinsinger* & Joeri Sergej Strijk

Received: 11 May 2016

Accepted: 20 July 2016

Published: 25 August 2016

Cycads are among the most threatened plant species. Increasing the availability of genomic information by adding whole chloroplast data is a fundamental step in supporting phylogenetic studies and conservation efforts. Here, we assemble a dataset encompassing three taxonomic levels in cycads, including ten genera, three species in the genus *Cycas* and two individuals of *C. debaoensis*. Repeated sequences, SSRs and variations of the chloroplast were analyzed at the intraspecific, interspecific and intergeneric scale, and using our sequence data, we reconstruct a phylogenomic tree for cycads. The chloroplast was 162,094 bp in length, with 133 genes annotated, including 87 protein-coding, 37 tRNA and 8 rRNA genes. We found 7 repeated sequences and 39 SSRs. Seven loci showed promising levels of variations for application in DNA-barcoding. The chloroplast phylogeny confirmed the division of Cycadales in two suborders, each of them being monophyletic, revealing a contradiction with the current family circumscription and its evolution. Finally, 10 intraspecific SNPs were found. Our results showed that despite the extremely restricted distribution range of *C. debaoensis*, using complete chloroplast data is useful not only in intraspecific studies, but also to improve our understanding of cycad evolution and in defining conservation strategies for this emblematic group.

Cycads are iconic relict species, regarded as “living fossils” because of their recognizable intermediate morphological traits between angiosperms and gymnosperms¹. Cycads dominated the Mesozoic but their origin can be dated to the late Paleozoic (~265–290 Ma)^{2,3}. However, molecular dating studies indicate that living cycad species could be not much older than ~12 Ma, rejecting both the hypothesized role of dinosaurs in generating extant diversity and the use of “living fossils” to describe current cycad species⁴. Cycads are distributed in tropical and subtropical regions of Africa, Asia, Oceania and America⁵. Ten genera and 344 species are currently accepted⁶, with *Cycas* containing roughly 40% of the species in the Near Threatened and Vulnerable categories in the IUCN red list⁷.

The genus *Cycas* L., in the monotypic family Cycadaceae, is the oldest genus of cycads, holding about 113 species⁵. More than 20 species are found in China^{5,8}, with most of them endemic. *Cycas debaoensis* Y. C. Zhong & C. J. Chen, a critically endangered cycad species endemic to southwest China⁹, only occurs in 11 small populations near the border of Guangxi province and Yunnan province¹⁰. Previous studies have assessed genetic diversity in *C. debaoensis* using inter simple sequence repeat (ISSR) markers or nuclear microsatellites, showing limited gene flow among populations and low within-population diversity^{10–12}.

Chloroplasts (cps) are present in photosynthetically active green tissues and generally develop from proplastids in meristems or etioplasts after illumination of dark-grown tissues, and display a conserved structure of two inverted repeats (IR) separated by small (SSC) and large (LSC) single-copy regions¹³. Due to their natural abundance in plant cells (~3–5% of the cell DNA content comparing to nuclear DNA)¹⁴, cp sequences are a versatile tool for plant identification (DNA-barcoding) and evolutionary studies. They have been used at small and large temporal scales in plants^{15,16}. The use of cps is a very powerful tool to reconstruct plant phylogenies and infer historical biogeographic patterns of diversification^{17,18}. However, only a limited number of regions in the chloroplast genome have been used to address evolutionary, taxonomic and biodiversity questions in *Cycas*¹⁹. With the rapid development of Next Generation Sequencing (NGS), it is now feasible to obtain the entire sequence of

Plant Ecophysiology and Evolution Group, State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources and College of Forestry, Guangxi University, Nanning, Guangxi 530004, China. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to G.-F.J. (email: gfjiang@gxu.edu.cn)

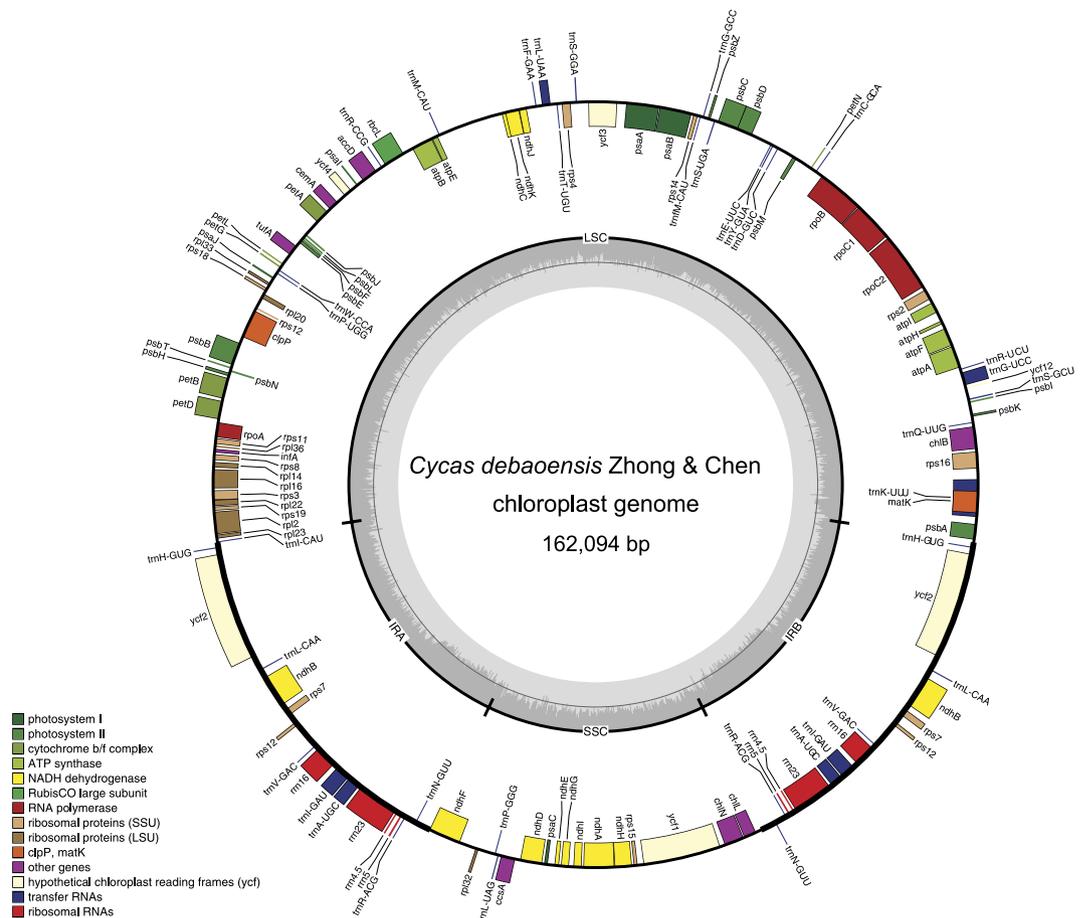


Figure 1. Circular gene map of the plastid genome of *Cycas debaoensis*. Genes drawn within the circle are transcribed clockwise, while those drawn outside are transcribed counter clockwise. Genes are colour-coded according to their functional groups. Inner circle: GC content.

Species	Length	LSC	SSC	IR	GC%				Coding	tRNA	rRNA
					Total	LSC	SSC	IR			
<i>C. debaoensis</i> KU743927	162,094	88,854	23,088	25,076	39.4	38.7	36.6	42.0	87	37	8
<i>C. debaoensis</i>	162,092	88,852	23,088	25,076	39.4	38.7	36.5	42.0	87	37	8
<i>C. revoluta</i>	162,489	88,977	23,376	25,068	39.4	38.6	38.9	42.0	109	39	8
<i>C. taitungensis</i>	163,403	90,216	23,039	25,074	39.5	38.8	36.5	42.0	122	38	8

Table 1. Characteristics of the complete chloroplasts used in the study. The Length, Large Single Copy (LSC) region; Small Single Copy (SSC) region; Inverted Repeats (IR), the GC content of each region, and the number of coding, tRNA and rRNA loci are shown. Data from GenBank for *C. debaoensis* (KM459003), *C. taitungensis*¹⁷ (NC_009618) and *C. revoluta* (NC_020319).

the chloroplast using a genome skimming approach, resulting in high resolution phylogenies and allowing for estimations of timing of historical diversification, biodiversity and extent of genomic divergence^{14,20,21}.

It is well known that genetic diversity can greatly vary between taxa, due to either different intrinsic characteristics (e.g. reproductive system, genome size and organization) or to extrinsic features (e.g. endemic vs. widespread species, young vs. old species)^{17,18,22}. In addition, the hypothesis of a linear accumulation of mutations in sequences across time (i.e. a molecular clock) has been refuted in many groups^{23–25}.

In this study, we analyse the *C. debaoensis* chloroplast as a reference together with molecular data of other *Cycas* species to identify potential DNA-barcode loci, and compared generic-level chloroplast features in cycads, to highlight the evolutionary history of this group. We also compared the chloroplast features of two individuals of *C. debaoensis* to provide new resources for marker development in this endangered species.

Results and Discussion

Genome size and features. Using genome skimming and reference-guided assembly, we reconstructed the 162,094 bp long chloroplast genome *C. debaoensis*_Jiang_DB-2015. The complete cp genome was submitted

	Group of genes	Gene names
1	Photosystem I	<i>psaA, psaB, psaC, psaI, psaJ</i>
2	Photosystem II	<i>psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ</i>
3	Cytochrome b/f complex	<i>petA, petB, petD, petG, petL, petN</i>
4	ATP synthase	<i>atpA, atpB, atpE, atpF^a, atpH, atpI</i>
5	NADH dehydrogenase	<i>ndhA^a, ndhB^a (×2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK</i>
6	RubisCO large subunit	<i>rbcl</i>
7	RNA polymerase	<i>rpoA, rpoB, rpoC1^a, rpoC2</i>
8	Ribosomal proteins (SSC)	<i>rps2, rps3, rps4, rps7 (×2), rps8, rps11, rps12 (×2), rps14, rps15, rps16, rps18, rps19</i>
9	Ribosomal proteins (LSC)	<i>rpl2^a, rpl14, rpl16, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36</i>
10	Other genes	<i>accD, ccsA, cemA, chlB, chlL, chlN, clpP^b, infA, matK, tufA</i>
11	Proteins of unknown function	<i>ycf1, ycf2 (×2), ycf3^b, ycf4, ycf12</i>
12	Ribosomal RNAs	<i>rrn4.5 (×2), rrn5 (×2), rrn16 (×2), rrn23 (×2)</i>
13	Transfer RNAs	<i>trnA-UGC^a (×2), trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnFM-CAU, trnG-GCC, trnG-UCC^a, trnH-GUG (×2), trnI-CAU, trnI-GAU^a (×2), trnK-UUU^a, trnL-CAA (×2), trnL-UAA^a, trnL-UAG, trnM-CAU, trnN-GUU (×2), trnP-GGG, trnP-UGG, trnQ-UUG, trnR-ACG (×2), trnR-CCG, trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-UGU, trnV-GAC (×2), trnW-CCA, trnY-GUA</i>

Table 2. Genes present in the *C.debaensis* (KU743927) chloroplast genome. (×2) Two gene copies in the IRs. ^aGene containing one intron. ^bGene containing two introns.

to GenBank under accession number KU743927. It was 2 bp longer than *C. debaoensis* (KM459003) due to two 1 bp indels in the LSC (Fig. 1). The two *C. debaoensis* individuals exhibited the typical composition of LSC, SSC regions and two IR copies of 88,854 bp, 23,088 bp, and 25,076 bp (Table 1). The overall GC content of *C. debaoensis* _Jiang_DB-2015 was 39.4%, and 38.7%, 36.6% and 42.0% in the LSC, SSC and IR regions, respectively. These values are similar to *C. debaoensis* KM459003. *C. debaoensis* (KU743927), *C. debaoensis* (KM459003) and *C. revoluta* showed similar GC content (39.4%), slightly lower than *C. taitungensis* (39.5%) (Table 1). In total, 133 genes were annotated, including 87 protein coding genes, 37 tRNA genes and 8 rRNA genes in *C. debaoensis* _Jiang_DB-2015, while 156 and 169 genes were annotated in *C. revoluta* and *C. taitungensis*, respectively. Twelve genes (*atpF, rpoC1, rpl2, clpP, ycf3, trnG-UCC, trnK-UUU, trnL-UAA* in LSC; *ndhA* locates in SSC; *ndhB, trnA-UGC, trnI-GAU* in the IRs regions) contain 1–2 introns, respectively; while fourteen genes (3 protein coding 7 tRNA and 4 rRNA) were duplicated in the IR regions (Table 2). The *tufA* gene, found in gymnosperms and hornworts and inherited from green algae is coding for a nonfunctional protein synthesis elongation factor (723-bp long in *C. taitungensis* and *Ginkgo*)¹⁷. Interestingly, this gene was 1 bp longer in *C. debaoensis* _Jiang_DB-2015 than in other *Cycas* (Table 2).

Repeat and SSR analysis. Using REPuter, seven repeats were found in the chloroplast of *C. debaoensis* _Jiang_DB-2015, which were three forward (F) and four palindrome (P), with no reverse and complement repeats discovered (Table 3). The repeats were mainly distributed in the intergenic spacers of transfer RNA genes, some of them being located in transfer RNA itself (Table 3). Interspecific comparison and analysis in broader *Cycadaceae* showed that *C. revoluta* had the highest number of repeats (24), while *C. debaoensis* contained the fewest (7), and *C. taitungensis* contained an intermediate number of repeats (16) (Supplementary Table S1, Fig. 2). In contrast, the comparison of simple sequence repeats (SSRs) revealed a relative conservatism in their numbers, with congeneric species showing similarities in both numbers and spatial patterns of SSRs occurrence (Fig. 3). This is of particular value in *Cycadaceae*, where species are usually scarcely distributed, and in which diagnostic morphological characters are often poorly defined or absent. These generic molecular biomarkers have the potential to provide useful diagnostic data in redefining complex paraphyletic and polyphyletic species groups in the family^{18,22,26}, aiding directly in conservation efforts.

In other groups, the presence and nature of repeats have been shown to be of great value in evolutionary and population analyses^{27,28}. Microsatellites (SSRs) are useful markers for population genetics, conservation of endangered species and species delineation^{22,29–31}, as previously highlighted for *C. debaoensis*^{10,12}. There were 39 SSRs in the chloroplast of *C. debaoensis*, 34 (87%) and 5 (13%) mono- and di-nucleotides SSRs, respectively (Table 4). These SSRs were mainly distributed in the IGS region (29; 74%), and the other 26% were distributed in CDS genes (Table 4). *C. revoluta* and *C. taitungensis* had similar SSRs numbers and locations than *C. debaoensis*, most of them being mononucleotides and distributed in the IGS (Fig. 3). Interestingly, *C. revoluta* lacked some SSR patterns (G and GA mono- and di- nucleotides, respectively) (Table S2). These diagnostic SSRs can be used in combination with nuclear SSRs developed in the genus for *Cycadaceae* conservation or reintroduction, species biodiversity assessments and phylogenetic studies in native or introduced areas^{12,30–33}.

Cycads phylogenetic reconstruction and comparison. In the maximum likelihood (ML) phylogenetic tree, all but two nodes were highly supported (bootstrap support ≥ 95), with the accessions of *C. debaoensis* closely related to the other *Cycas* species (Fig. 4A). *Cycas* spp. diverged first in the Cycadales, followed by *Dioon*, a clade containing *Zamia*, *Ceratozamia* and *Stangeria*. *Bowenia* diverged from the remaining Zamiaceae with relatively high support (bootstrap support 80%), *Macrozamia* being as sister to a clade containing both *Encephalartos* and *Lepidozamia* (Fig. 4A). This chloroplast phylogeny confirms the division of Cycadales into two suborders, each of them being monophyletic in our analyses (Fig. 4B), but contradict the current family delimitations³⁴, with the family Stangeriaceae being polyphyletic with high support, in agreement with the most recent phylogenetic

	Repeat1 start (location)	Repeat2 start (location)	Size (bp)	Type	Region
1	88,852 (<i>trnI</i> -CAU)	137,037 (IGS <i>chlL</i> - <i>trnN</i> -GUU)	20,257	P	LSC; IRb
2	28,396 (IGS <i>ropB</i> - <i>trnC</i> -GCA)	28,470 (IGS <i>ropB</i> - <i>trnC</i> -GCA)	39	F	LSC
3	55,293 (<i>trnM</i> -CAU)	55,340 (IGS <i>trnM</i> -CAU- <i>atpE</i>)	32	P	LSC
4	48,603 (IGS <i>trnF</i> -GAA- <i>ndhF</i>)	48,629 (IGS <i>trnF</i> -GAA- <i>ndhF</i>)	31	F	LSC
5	113,680 (IGS <i>trnN</i> -GUU- <i>ndhF</i>)	113,680 (IGS <i>trnN</i> -GUU- <i>ndhF</i>)	30	P	IRa
6	113,680 (IGS <i>trnN</i> -GUU- <i>ndhF</i>)	137,236 (IGS <i>chlL</i> - <i>trnN</i> -GUU)	30	F	IRa; IRb
7	137,236 (IGS <i>chlL</i> - <i>trnN</i> -GUU)	137,236 (IGS <i>chlL</i> - <i>trnN</i> -GUU)	30	P	IRb

Table 3. Repeat sequences and their distribution found by REPuter in the *C. debaensis* (KU743927) chloroplast genome. IGS: Intergenic spacer.

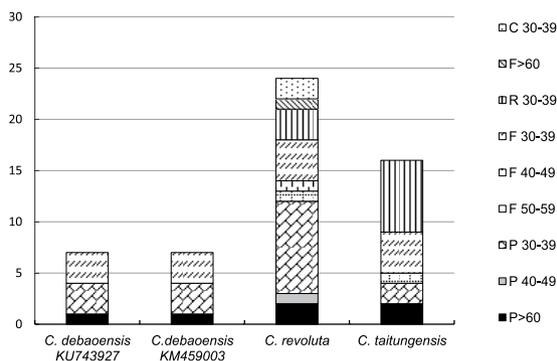


Figure 2. Repeat sequences in four chloroplast genomes of *Cycas*. REPuter was used to identify repeat sequences with length ≥ 30 bp and sequence identity $\geq 90\%$ in the chloroplast genomes. F and P indicate the repeat type F (forward) and P (palindrome), respectively. Repeats with different lengths are indicated in different patterns.

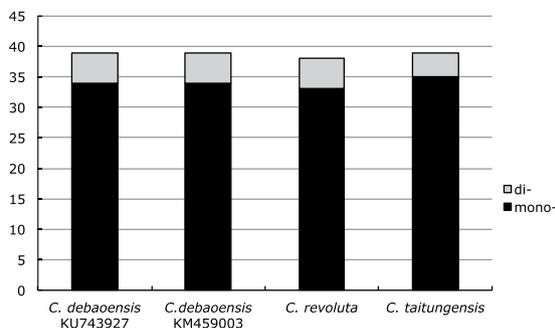


Figure 3. Number of simple sequence repeats in four chloroplast genomes of *Cycas*, classified by repeat type. mono-: mononucleotide SSRs; di-: dinucleotide SSRs.

work in cycads³⁵. However, we found each genus in Stangeriaceae grouping with one of the clade in Zamiaceae, contrary to Salas-Leiva *et al.*³⁵, in which *Bowenia* diverged in a basal position for all cycads but *Cycas* and *Dioon*, and *Stangeria* group with *Zamia* (*Microcycas* being absent from our dataset). *C. debaensis* diverged in a basal position of the genus, with *C. revoluta* and *C. taitungensis* being closely related. In addition, the branch leading to *Cycas* is longer than the branches leading to the other genera, in accordance with the hypothesis of a recent diversification during the Miocene¹⁹. These results demonstrate the suitability and efficiency of using complete chloroplast sequences in reconstructing the evolutionary history of cycads, as previously demonstrated for other groups^{36,37}.

Although the clear delineation of the genera in cycads are mostly due to the lengths of the sequences provided by complete chloroplast sequencing, the variability was unevenly distributed along the circular molecule (Fig. 5). Indeed, the ribosomal RNA genes as well as the region between *psbA* and *rpoC1* genes (0–25 kb) and between *ndhC* and *rpl20* (50–72 kb) of the cp sequences exhibited relatively few variations among cycads (Fig. 5). A cluster of four *ndh* genes (120–124.5 kb) appeared to be strikingly conserved. Overall, the level of variation increased with taxonomic distance, meaning that regions showing polymorphisms among *Cycas* species, exhibited higher variation among different cycad genera. However, some notable well-defined (<500 bp long) polymorphic

No.	SSR type	SSR	size	start	SSR-containing region
2	p1	A	20	8307	IGS (<i>trnQ-UUG-psbK</i>); LSC
	p1	A	20	73073	<i>clpP</i> ; LSC
1	p1	A	19	123912	IGS (<i>ndhG-ndhI</i>); SSC
1	p1	A	14	32963	IGS (<i>trnE-UUC-trnH-GUG</i>); LSC
1	p1	A	13	68694	IGS (<i>tufA-trnH-GUG</i>); LSC
2	p1	A	11	101911	IGS (<i>rps7-rps12</i>); IRa
	p1	A	11	148078	IGS (<i>trnV-GAC-rps12</i>); IRb
2	p1	A	10	1872	<i>trnK-UUU</i> ; LSC
	p1	A	10	11723	IGS (<i>trnG-UCC-trnG-UCC</i>); LSC
1	p1	T	19	15487	IGS (<i>atpH-atpI</i>); LSC
1	p1	T	18	10133	IGS (<i>trnS-GCU-ycf12</i>); LSC
2	p1	T	16	57648	IGS (<i>atpB-rbcL</i>); LSC
	p1	T	16	84489	<i>rpl16</i> ; LSC
3	p1	T	15	54643	IGS (<i>trnM-CAU-ndhC</i>); LSC
	p1	T	15	84795	<i>rpl16</i> ; LSC
	p1	T	15	134345	IGS (<i>ycf1-chlN</i>); SSC
1	p1	T	13	74064	<i>clpP</i> ; LSC
2	p1	T	12	5951	<i>rps16</i> ; LSC
	p1	T	12	118369	IGS (<i>rpl32-trnP-GGG</i>); SSC
4	p1	T	11	83227	<i>rps8</i> ; LSC
	p1	T	11	88618	IGS (<i>rpl23-trnI-CAU</i>); LSC
	p1	T	11	102859	IGS (<i>rps12-trnV-GAC</i>); IRa
	p1	T	11	149026	IGS (<i>rps7-rps12</i>); IRb
3	p1	T	10	63350	IGS (<i>ycf4-cemA</i>); LSC
	p1	T	10	69919	IGS (<i>trnP-UGG-psaI</i>); LSC
	p1	T	10	125905	<i>ndhA</i> ; SSC
1	p1	G	14	70451	IGS (<i>psaJ-trnH-GUG</i>); LSC
1	p1	G	11	52764	IGS (<i>ndhC-trnH-GUG</i>); LSC
2	p1	G	10	70921	IGS (<i>rpl33-trnH-GUG</i>); LSC
	p1	G	10	149995	IGS (<i>rps7-ndhB</i>); IRb
1	p1	C	14	44747	<i>ycf3</i> ; LSC
1	p1	C	11	5229	<i>rps16</i> ; LSC
2	p1	C	10	17681	IGS (<i>rps2-rpoC2</i>); LSC
	p1	C	10	100943	IGS (<i>ndhB-rps7</i>); IRa
3	p2	(TA)14	28	29832	IGS (<i>petN-psbM</i>); LSC
	p2	(TA)9	18	1453	IGS (<i>psbA-trnK-UUU</i>); LSC
	p2	(TA)6	12	15670	IGS (<i>atpH-atpI</i>); LSC
1	p2	(GA)6	12	68377	IGS (<i>psbE-petL</i>); LSC
1	p2	(AT)6	12	15808	IGS (<i>atpH-atpI</i>); LSC

Table 4. Simple sequence repeats in the *C. debaensis* (KU743927) chloroplast genome.

regions departed from this assertion in *ycf1* and *ycf2* IRa. Indeed, the polymorphisms in these regions were higher among *Cycas* than among cycads as a whole. Interestingly, three SNPs were located in other regions (*trnL-trnF*, *clpP* intron 2 and *ycf2* IRb) that exhibited a continuous increase in polymorphism levels across the family. *ycf1* has been recently proposed as a barcode locus³⁸, despite it not being present in all genera³⁹, and was identified with *clpP* among the most variable loci in *Parthenium* spp.⁴⁰. Finally, *trnL-trnF* was previously used in *Cycas* phylogenetic studies⁴¹, but also in species identification of trees and ferns^{42,43}. Here, we stress the need to further assess these loci as potential more informative substitutes to the official barcode loci⁴⁴.

Comparative interspecific chloroplast genomic analysis. Focusing on the three *Cycas* species available in GenBank, the mVISTA results showed that the four chloroplasts were highly conserved; however, the coding regions appeared to be globally more variable than the non-coding regions (Fig. 5 and Suppl. Fig. 1). Furthermore, the coding regions, e.g. *rpoB*, *psbC*, *clpP* (intron), *ycf1*, and *ycf2*; *psbA-trnH* and *trnL-trnF* intergenic spacers showed promising levels of variations for further development in applications such as DNA-barcoding or phylogenetic reconstruction.

Cycas have been considered as a difficult group for DNA-barcoding⁴⁵. Previously, it was reported that the *psbA-trnH* spacer was highly variable in Cycadales except *Cycas*⁴⁵, but *trnL-trnF* was used in *Cycas* for phylogenetic studies⁴¹. Although *rbcL* + *matK* were chosen as a two-locus DNA-barcoding for their universality and efficacy in land plant⁴⁴, it was not variable enough in *Cycas* (Fig. 5 and Suppl. Fig. 1). *rpoB*, *rpoC1*, and non-coding

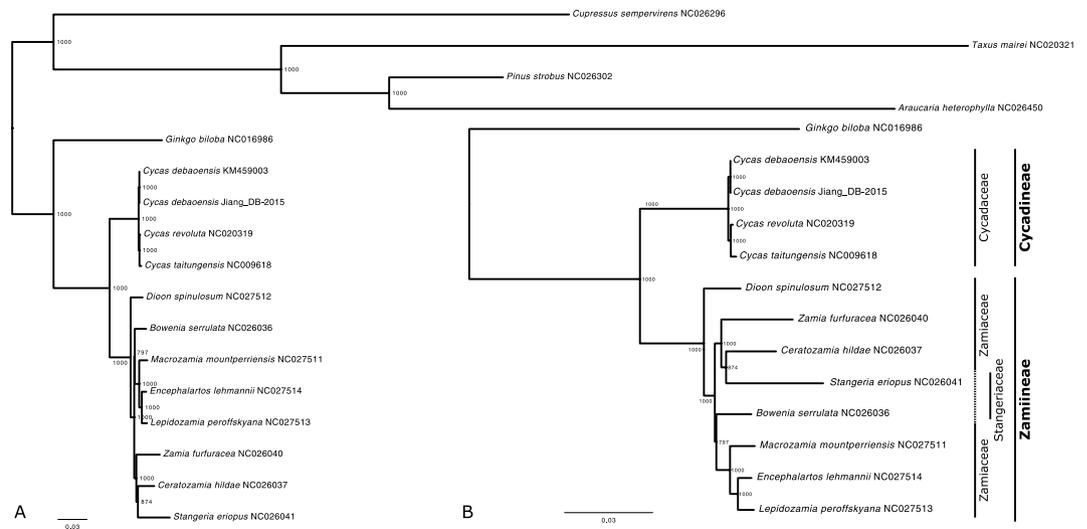


Figure 4. Maximum Likelihood phylogenetic tree of the available chloroplast sequences in GenBank for Cycadales, plus the chloroplast sequences of *Cupressus sempervirens*, *Taxus mairei*, *Pinus strobus*, *Araucaria heterophylla* and *Ginkgo biloba* as outgroups (A). For readability and better understanding of the branches lengths, a zoom on the Cycadales family is shown (B).

(e.g. *atpF-atpH*, *psbK-psbI*, and *rpl32-trnL*) regions have been shown to be variable enough at higher taxonomic levels⁴⁶, but also within cycads^{47,48}. In light of this, *psbC*, *clpP* (intron), *ycf1*, and *ycf2* should be considered as candidates for future phylogenetic studies in *Cycas*.

Intraspecific comparison. Comparing the two individuals of *C. debaoensis*, we found 10 SNPs and 1 indel, plus one “N” position in our data that didn’t allow us to confirm its status (Table 5). Their genomic positions are indicated in Figs 1 and 5. Six and four SNPs were located in IGS and coding regions, respectively, and one indel in the *clpC* intron. The genetic distance between the two individuals was 0.005553% (i.e. ≈ 1 SNPs/indels per 20 kb). This result is consistent with previous studies showing low within-population diversity in addition to limited gene flow among populations in *C. debaoensis*^{10,11}. Whittall *et al.*⁴⁹ found a comparable level of intraspecific divergence in pines, irrespective of the rarity of the considered species. However, in the pest species *Jacobaea vulgaris* (Asteraceae), the intraspecific divergence was four times higher⁵⁰, perhaps due to its short generation times as opposed to those prevailing in slow growing and long-lived trees. Further studies are still needed to determine whether or not intraspecific genetic diversity is linked to geographic ranges or the intrinsic characteristics of the taxonomic group.

Conclusions

Comparing genomic diversity at different taxonomical, but also spatial and temporal scales, we were able to reconstruct a robust phylogeny for cycads, and to identify regions showing promising levels of variation at three levels (familial, generic and intraspecific rank). These regions can provide useful and alternative loci for species identification and population-based studies for conservation, ecology and evolution. Despite their restricted geographic ranges, we showed that several, potentially diagnostic intraspecific variations can be found in the chloroplasts of different individuals *C. debaoensis*, including 10 SNPs and 1 indel in as of yet unstudied regions. Comparing results from the three scales, four regions appeared to be variable at the three considered taxonomic scales, namely *ycf3*, *clpP*, *psbD* and the *trnL-trnF* IGS. Therefore, we recommend future studies in cycads further evaluate these loci in details.

We expect that by providing and highlighting these new resources to the plant research community, it will allow for development of new diagnostic markers and innovative conservation strategies in this iconic, but highly threatened taxonomic group, especially in the case of *C. debaoensis*.

Materials and Methods

DNA sequencing and genome assembly. Total genomic DNA was extracted from 0.1 g of frozen fresh leaves, from an individual collected in Guangxi (23°69′40″N, 106°15′83″E) in 2015 (voucher deposited at our research group herbarium, Jiang_C2) according to the manufacturer instructions with the Plant Genomic DNA Kit (Tiangen Biotech Co., Ltd). A 350-bp paired-end library was then constructed using NEBNext Ultra II DNA Library Prep Kit (Ipswich, Massachusetts, USA) and sequenced by Novogene (Beijing, China). About 1 Gb of raw data were obtained on an Illumina HiSeq2500 platform (San Diego, California, USA), with a paired-end read length of 2×150 bp. The raw reads were submitted to the SRA under the accession number SRR3407155.

The raw data were imported in Geneious R9 (Biomatters Ltd, Auckland, New Zealand), and a cp genome was assembled according to Hinsinger and Strijk⁵¹. Raw reads were trimmed according to their 5′ and 3′-end quality, then a reference-guided assembly was performed, using the available cp of *Cycas debaoensis* (KM459003) as a

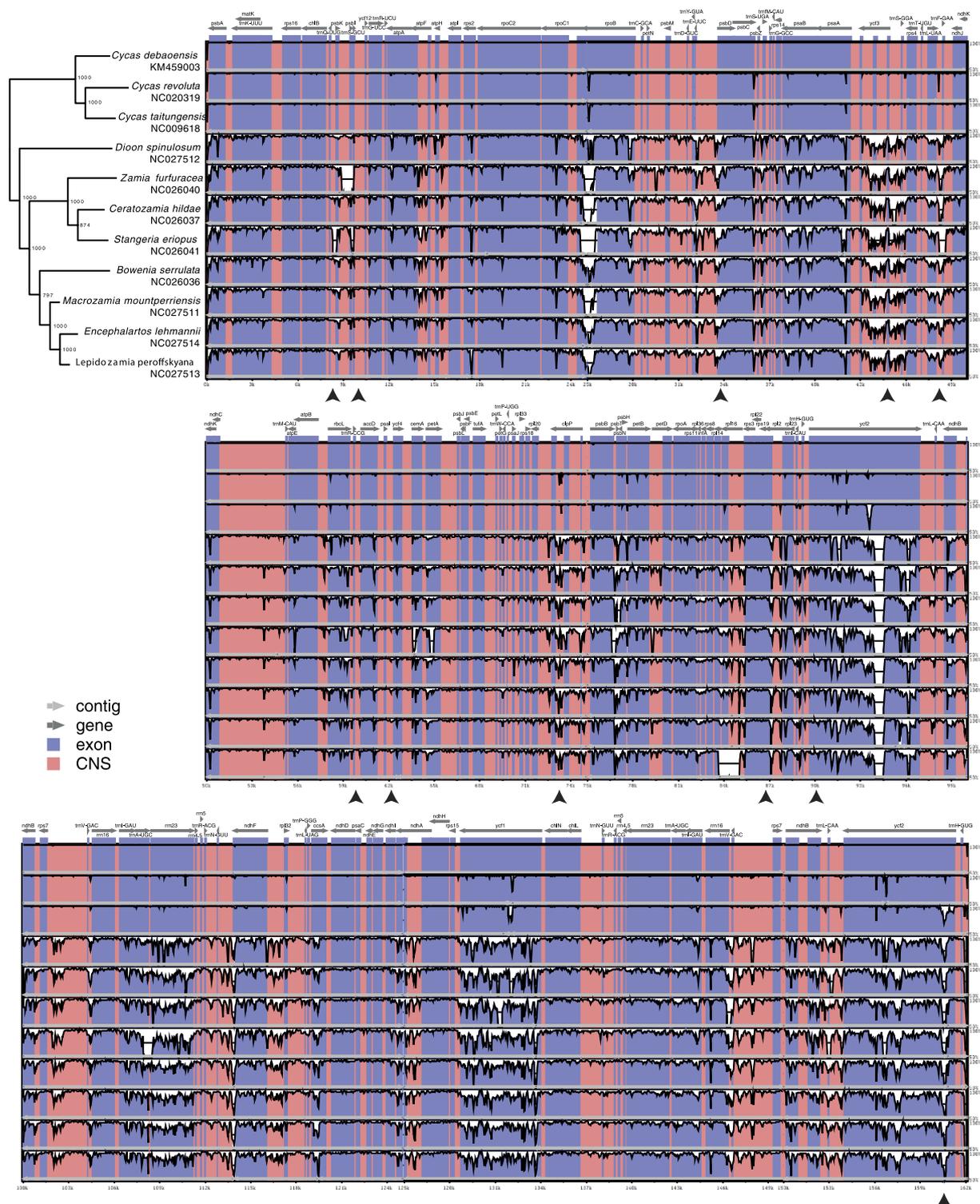


Figure 5. mVISTA percent identity plot of available cycad chloroplasts, using *C. debaensis* as a reference. Vertical scale indicates the percentage of identity ranging from 50% to 100%. Coding regions are in blue and noncoding regions are in pink. Cladogram redrawn from Fig. 5B; branch lengths are not representative of evolutionary changes; bootstrap support is indicated on the nodes. Black arrows indicate intraspecific variations in *C. debaensis*.

reference for the mapping step. The cp genome annotation was transferred from *C. debaensis* using the implemented function in Geneious R9 and their boundaries were manually checked. The circular cp genome map was generated using the Organellar Genome Draw program (OGDRAW).

Position	Type	Nucleotide	location	location type
8432		A/T	<i>trnQ-psbK</i>	IGS
9889		C/T	<i>trnS-ycf12</i>	IGS
33707		G/T	<i>psbD</i>	coding
44878		C/G	<i>ycf3</i>	exon 1
48140		A/C	<i>trnL-trnF</i>	IGS
60132		G/T	<i>trnR-accD</i>	IGS
62218		A/C	<i>psaI-ycf4</i>	IGS
73073		A/.	<i>clpP</i>	intron 2
86731		C/T	<i>rps19-rpl2</i>	IGS
90307		G/T	<i>ycf2</i>	coding (IRa)
160640		A/C	<i>ycf2</i>	coding (IRb)

Table 5. Intraspecific SNPs between two individuals of *C. debaoensis* (KU743927).

Intergeneric comparisons and phylogenetic reconstruction. Following the recommended best practices for complete organellar sequencing⁵², we performed a phylogenetic analysis to confirm the accuracy of our reconstructed plastid and sample identification. We retrieved all the Cycadales available in GenBank (accessed 2016/02/15), representing all the ten genera in the order except *Microcycas*⁶. To this Cycadales dataset, we added the sequences of *Ginkgo biloba* (NC016986), *Pinus strobus* (NC026302), *Araucaria heterophylla* (NC026450), *Taxus mairei* (NC020321) and *Cupressus sempervirens* (NC026296) as outgroups.

Sequences were aligned using MAFFT⁵³ with default options. We used the jModelTest⁵⁴ implementation in CIPRES⁵⁵, and set the substitution model accordingly. We built a maximum likelihood tree using PHYML⁵⁶, with a 012310 + G + F model using four gamma categories and 1000 bootstrap replicates. The ML tree was built using all positions. In addition, to identify regions with substantial variability, the complete cp genomes of nine cycad genera were compared using mVISTA⁵⁷, with *C. debaoensis* (KM459003) as a reference for the annotations.

Sequence divergences among cycads were estimated using the Kimura 2-parameter model⁵⁸, implemented in MEGA6⁵⁹. Codon positions included were 1st + 2nd + 3rd + Noncoding. All positions containing gaps and missing data were excluded prior to analyses.

Interspecific comparisons. The complete cp genomes of *C. debaoensis* and two other species in cycas (*C. revoluta* and *C. taitungensis*) were compared using mVISTA⁵⁷, as described above.

For each species, repeats (forward, palindrome, reverse and complement sequences) were identified using REPuter⁶⁰ with 30 bp and sequence identity greater than 90%. Simple sequence repeats (SSRs) of *C. debaoensis* and the two other species were detected using MISA⁶¹ by setting the minimum number of repeats to 10, 5, 4, 3, 3 and 3 for mono-, di-, tri-, tetra-, penta- and hexa nucleotides, respectively. Sequence divergences among the *Cycas* species were estimated as described above.

Intraspecific genome comparison. The two complete chloroplasts of *C. debaoensis* (KU743927, KM459003) were aligned in Geneious R9 (Biomatters Ltd, Auckland, New Zealand) using the MAFFT algorithm⁵³, and differences were identified using the “Find Variations/SNPs” function and checked individually. We recorded substitutions and indels separately, as well as their location in the chloroplast genome (e.g. SSRs/repeats, coding region/rRNA/tRNA/IGS). Sequence divergence extent between the two individuals was estimated as described above.

References

- Brenner, E. D., Stevenson, D. W. & Twigg, R. W. Cycads: evolutionary innovations and the role of plant-derived neurotoxins. *Trends Plant Sci.* **8**, 446–452 (2003).
- Mamay, S. H. Cycads: fossil evidence of late paleozoic origin. *Science*. **164**, 295–296 (1969).
- Gao, Z. & Thomas, B. A. A review of fossil cycad megasporophylls, with new evidence of Crossozamia Pomel and its associated leaves from the Lower Permian of Taiyuan, China. *Rev. Palaeobot. Palynol.* **60**, 205–223 (1989).
- Nagalingum, N. *et al.* Recent synchronous radiation of a living fossil. *Science*. **334**, 796–799 (2011).
- Hill, K. D., Stevenson, D. W. & Osborne, R. The world list of cycads. *Bot. Rev.* **70**, 274–298 (2004).
- Calonje, M., Stevenson, D. W. & Stanberg, L. The World List of Cycads. *Online edition [Internet]*, Available from: <http://www.cycadlist.org>. (2013–2016).
- Marler, P. N. & Marler, T. E. An assessment of Red List data for the Cycadales. *Trop. Conserv. Sci.* **8** (2015).
- Chen, C. J. & Liu, N. New discoveries of cycads and advancement of conservation of cycads in China. *Bot. Rev.* **70**, 93–100 (2004).
- Chen, C. J. & Zhong, Y. C. *Cycas debaoensis* Y. C. Zhong et C. J. Chen – a new cycad from China. *Acta Phytotaxon. Sin.* **35**, 571 (1997).
- Zhan, Q.-Q., Wang, J.-F., Gong, X. & Peng, H. Patterns of chloroplast DNA variation in *Cycas debaoensis* (Cycadaceae): conservation implications. *Conserv. Genet.* **12**, 959–970 (2011).
- Xie, J., Jian, S. & Liu, N. Genetic variation in the endemic plant *Cycas debaoensis* on the basis of ISSR analysis. *Aust. J. Bot.* **53**, 141–145 (2005).
- Yang, Y., Li, Y., Li, L.-F., Ge, X.-J. & Gong, X. Isolation and characterization of microsatellite markers for *Cycas debaoensis* YC Zhong et CJ Chen (Cycadaceae). *Mol. Ecol. Resour.* **8**, 913–915 (2008).
- Jansen, R. K. & Ruhlman, T. A. Plastid genomes of seed plants. In *Genomics of chloroplasts and mitochondria*. 103–126 (Springer, 2012).
- Staats, M. *et al.* Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLoS One.* **8**, e69189 (2013).
- Nock, C. J. *et al.* Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnol. J.* **9**, 328–333 (2011).

16. Coissac, E., Hollingsworth, P. M., Lavergne, S. & Taberlet, P. From barcodes to genomes: extending the concept of DNA barcoding. *Mol. Ecol.* (2016).
17. Wu, C.-S., Wang, Y.-N., Liu, S.-M. & Chaw, S.-M. Chloroplast genome (cpDNA) of *Cycas taitungensis* and 56 cp protein-coding genes of *Gnetum parvifolium*: insights into cpDNA evolution and phylogeny of extant seed plants. *Mol. Biol. Evol.* **24**, 1366–1379 (2007).
18. Jardón-Barbolla, L., Delgado-Valerio, P., Geada-López, G., Vázquez-Lobo, A. & Piñero, D. Phylogeography of *Pinus* subsection Australes in the Caribbean basin. *Ann. Bot.* mcq232 (2010).
19. Xiao, L.-Q. & Möller, M. Nuclear ribosomal ITS functional paralogs resolve the phylogenetic relationships of a late-Miocene radiation cycad *Cycas* (Cycadaceae). *PLoS One*. **10**, e0117971 (2015).
20. Straub, S. C. *et al.* Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *Am. J. Bot.* **99**, 349–364 (2012).
21. Hinsinger, D. D. & Strijk, J. S. Complete chloroplast genome sequence of *Castanopsis concinna* (Fagaceae), a threatened species from Hong Kong and South-Eastern China. *Mitochondrial DNA*. 1–2 (2015).
22. Wu, C.-S., Lin, C.-P., Hsu, C.-Y., Wang, R.-J. & Chaw, S.-M. Comparative chloroplast genomes of Pinaceae: insights into the mechanism of diversified genomic organizations. *Genome Biol. Evol.* **3**, 309–319 (2011).
23. Bell, C. D., Soltis, D. E. & Soltis, P. S. The age of the angiosperms: a molecular timescale without a clock. *Evolution*. **59**, 1245–1258 (2005).
24. Kay, K. M., Whittall, J. B. & Hodges, S. A. A survey of nuclear ribosomal internal transcribed spacer substitution rates across angiosperms: an approximate molecular clock with life history effects. *BMC Evol. Biol.* **6**, 36 (2006).
25. Pereira, S. L. & Baker, A. J. A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. *Mol. Biol. Evol.* **23**, 1731–1740 (2006).
26. 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.* **4**, 374–381 (2012).
27. Cavalier-Smith, T. Chloroplast evolution: secondary symbiogenesis and multiple losses. *Curr. Biol.* **12**, R62–R64 (2002).
28. Nie, X. *et al.* Complete chloroplast genome sequence of a major invasive species, crofton weed (*Ageratina adenophora*). *PLoS One*. **7**, e36869 (2012).
29. Powell, W., Morgante, M., McDevitt, R., Vendramin, G. & Rafalski, J. Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines. *Proc. Natl. Acad. Sci. USA* **92**, 7759–7763 (1995).
30. Cibrián-Jaramillo, A., Marler, T. E., DeSalle, R. & Brenner, E. D. Development of EST-microsatellites from the cycad *Cycas rumphii*, and their use in the recently endangered *Cycas micronesica*. *Conserv. Genet.* **9**, 1051–1054 (2008).
31. Calonje, M. *et al.* Cycad biodiversity in the Bahamas Archipelago and conservation genetics of the threatened *Zamia lucayana* (Zamiaceae). *Oryx*. **47**, 190–198 (2013).
32. Cibrián-Jaramillo, A., Daly, A., Brenner, E., Desalle, R. & Marler, T. When North and South don't mix: genetic connectivity of a recently endangered oceanic cycad, *Cycas micronesica*, in Guam using EST-microsatellites. *Mol. Ecol.* **19**, 2364–2379 (2010).
33. Ju, L.-P. *et al.* Microsatellite primers in the native perennial cycad *Cycas taitungensis* (Cycadaceae). *Am. J. Bot.* **98**, e84–e86 (2011).
34. Stevenson, D. W. A formal classification of the extant cycads. *Brittonia*. **44**, 220–223 (1992).
35. Salas-Leiva, D. E. *et al.* Phylogeny of the cycads based on multiple single-copy nuclear genes: congruence of concatenated parsimony, likelihood and species tree inference methods. *Ann. Bot.* **112**, 1263–1278 (2013).
36. Jansen, R. K. *et al.* Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evol. Biol.* **6**, 1 (2006).
37. Parks, M., Cronn, R. & Liston, A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol.* **7**, 1 (2009).
38. Dong, W. *et al.* *ycf1*, the most promising plastid DNA barcode of land plants. *Sci. Rep.* **5** (2015).
39. Dong, W., Liu, J., Yu, J., Wang, L. & Zhou, S. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS One*. **7**, e35071 (2012).
40. Kumar, S., Hahn, F. M., McMahan, C. M., Cornish, K. & Whalen, M. C. Comparative analysis of the complete sequence of the plastid genome of *Parthenium argentatum* and identification of DNA barcodes to differentiate *Parthenium* species and lines. *BMC Plant Biol.* **9**, 1 (2009).
41. Sangin, P., Lindstrom, A. J., Kokubugata, G., Chaiprasongsuk, M. & Mingmuang, M. Phylogenetic relationships within Cycadaceae inferred from non-coding regions of chloroplast DNA. *J. Nat. Sci.* **44**, 544–557 (2010).
42. Kojoma, M. *et al.* Genetic identification of cinnamon (*Cinnamomum* spp.) based on the *trnL-trnF* chloroplast DNA. *Planta Medica*. **68**, 94–96 (2002).
43. De Groot, G. A. *et al.* Use of *rbcl* and *trnL-F* as a two-locus DNA barcode for identification of NW-European ferns: an ecological perspective. *PLoS One*. **6**, e16371 (2011).
44. CBOL Plant Working Group. A DNA barcode for land plants. *Proc. Natl. Acad. Sci. USA* **106**, 12794–12797 (2009).
45. Sasser, C., Little, D. P., Stevenson, D. W. & Specht, C. D. DNA barcoding in the cycadales: testing the potential of proposed barcoding markers for species identification of cycads. *PLoS One*. **2**, e1154 (2007).
46. Li, X. *et al.* Plant DNA barcoding: from gene to genome. *Biol. Rev.* **90**, 157–166 (2015).
47. Nicolalde-Morejón, F., Vergara-Silva, E., González-Astorga, J. & Stevenson, D. W. Character-based, population-level DNA barcoding in Mexican species of *Zamia* L. (Zamiaceae: Cycadales). *Mitochondrial DNA*. **21**, 51–59 (2010).
48. Nicolalde-Morejón, F. *et al.* A character-based approach in the Mexican cycads supports diverse multigene combinations for DNA barcoding. *Cladistics*. **27**, 150–164 (2011).
49. Whittall, J. B. *et al.* Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Mol. Ecol.* **19**, 100–114 (2010).
50. Doorduyn, L. *et al.* The complete chloroplast genome of 17 individuals of pest species *Jacobaea vulgaris*: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. *DNA Res.* dsr002 (2011).
51. Hinsinger, D. D. & Strijk, J. S. Toward phylogenomics of Lauraceae: The complete chloroplast genome sequence of *Litsea glutinosa* (Lauraceae), an invasive tree species on Indian and Pacific Ocean islands. *Plant Gene*, submitted.
52. Botero-Castro, F., Delsuc, F. & Douzery, E. J. Thrice better than once: Quality control guidelines to validate new mitogenomes. *Mitochondrial DNA*. **27**, 449–454 (2016).
53. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
54. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods*. **9**, 772–772 (2012).
55. Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In *Gateway Computing Environments Workshop (GCE)*, 1–8 (IEEE) (2010).
56. Guindon, S., Delsuc, F., Dufayard, J.-F. & Gascuel, O. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.* **537**, 113–137 (2009).
57. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32**, W273–W279 (2004).

58. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).
59. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* mst197 (2013).
60. Kurtz, S. & Schleiermacher, C. REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics.* **15**, 426–427 (1999).
61. Thiel, T., Michalek, W., Varshney, R. & Graner, A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* **106**, 411–422 (2003).

Acknowledgements

We thank Guangxi Baise Forestry Bureau of Youjiang District and associate Prof. Zhuang Jia for assistance in sample collecting. This work was supported by the Natural Science Foundation of Guangxi Province of China (Grant No. 2014GXNSFBA118075 and 2015GXNSFAA139085), the Foundation of Guangxi Educational Committee of China (Grant No. YB2014006), the Postdoctoral Science Foundation of Guangxi Province of China, the Scientific Research Foundation of Guangxi University (Grant No. XJZ130365, SYJN20131630, 201410593198, and XDZ120929); the Guangxi University (Nanning) and the provincial government of Guangxi Province (“100 Talents” Program; recruitment of overseas talents for colleges and universities in Guangxi); the postdoctoral starting grant of the Guangxi University, the China Postdoctoral Science Foundation Grant (No. 2015M582481) and the Postdoctoral Project from GXU.

Author Contributions

G.-F.J. designed the study, collected the plant materials, extracted DNA, assembled and analyzed the data, and wrote the paper; D.D.H. designed the study, assembled and analyzed the data and wrote the paper; J.S.S. wrote the paper. All authors have read and approved the final manuscript.

Additional Information

Accession codes: Raw reads and assembled chloroplast of *C. debaoensis* are available under the accession numbers SRR3407155 and KU743927, respectively.

Supplementary information accompanies this paper at <http://www.nature.com/srep>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Jiang, G.-F. *et al.* Comparison of intraspecific, interspecific and intergeneric chloroplast diversity in Cycads. *Sci. Rep.* **6**, 31473; doi: 10.1038/srep31473 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016