

# 1 **Chromosome-scale genome assembly of *Cycas revoluta* provides insights into cycad** 2 **diversification and species diversity**

3 Running title: The genome assembly of *Cycas revoluta*

4

5 Mitsuhiro P. Sato<sup>1</sup>, Yuta B. Aoyagi<sup>1</sup>, Kazutoshi Yoshitake<sup>2</sup>, Yukiho Toyama<sup>3,†</sup>, Hideaki  
6 Iimura<sup>1</sup>, Fumika Matsuoka<sup>4</sup>, Atsushi Toyoda<sup>5</sup>, Takashi Akagi<sup>6</sup>, Satohiro Okuda<sup>3</sup>, Yasuko Ito-  
7 Inaba<sup>4,7</sup>, and Kenta Shirasawa<sup>1\*</sup>

8

9 1. Department of Frontier Research and Development, Kazusa DNA Research Institute,  
10 Chiba 292-0818, Japan

11 2. School of Marine Bioscience, Kitasato University, Kanagawa 252-0373, Japan

12 3. Department of Biological Science, Graduate School of Science, The University of Tokyo,  
13 Tokyo 113-0033, Japan

14 4. Faculty of Agriculture, University of Miyazaki, Miyazaki 889-2192, Japan

15 5. Advanced Genomics Center, National Institute of Genetics, Shizuoka 411-8540, Japan

16 6. Graduate School of Environmental and Life Science, Okayama University, Okayama,  
17 Japan

18 7. Graduate School of Life Science, Tohoku University, Miyagi 980-8577, Japan

19 †. Present address: Department of Frontier Research, National Institute of Genetics, Shizuoka  
20 411-8540, Japan

21

22 \*To whom correspondence should be addressed: Tel. +81 438 52 3935; Email  
23 shirasaw@kazusa.or.jp

24

## 25    **Abstract**

26    Cycads are an ancient group of seed plants. Despite their ancient origin, many extant cycad  
 27    genera exhibit high species diversity. The specialized reproductive traits in cycads, dioecy  
 28    governed by the XY sex-determination system and the elaborate co-evolutionary synergy  
 29    with insect pollinators, may facilitate lineage diversification. Here, we present a  
 30    chromosome-scale genome sequence of *Cycas revoluta*, the species in which plant  
 31    spermatozoids were discovered in 1896. The genome sequence spanned 11.6 Gb, 98.4% of  
 32    which were anchored onto the 11 cycad chromosomes. Repetitive sequences occupied 9.8 Gb,  
 33    and 31,481 genes were predicted. Based on this genome assembly, the X- and Y-associated  
 34    genomic regions were characterized, and candidate genes for sex determination were  
 35    identified. In addition, the genomic positions of genes for sex-related traits were determined.  
 36    Subsequently, we analyzed transcriptomes for thermogenesis responsible for attracting insect  
 37    pollinators. Through these comprehensive analyses, we provide new insights into the  
 38    genomic basis of cycad diversification and species diversity.

39    **Keywords:** Chromosome-scale assembly, Gymnosperm, Living fossil, Male-specific region,  
 40    Thermogenesis

41

## 42     **Introduction**

43     Cycad (order Cycadales) is an ancient group of seed plants with a lineage dating back to the  
 44     Carboniferous period, approximately 330 million years ago, on the Laurasian landmass and  
 45     colonized Gondwana in the Jurassic period<sup>1</sup>. Along with cycad, ginkgo (Ginkgoales, a sister  
 46     order of Cycadales) is also one of the most ancient lineages of extant seed plants. Both  
 47     lineages diverged from a common ancestor hundreds of millions of years ago. Due to their  
 48     remarkable morphological resemblance to fossil ancestors, cycad and ginkgo are often  
 49     termed 'living fossils.' Despite their ancient origin, while Ginkgoales comprises only a single  
 50     extant species, *Ginkgo biloba*, many extant cycad genera exhibit high species diversity  
 51     resulting from rapid radiations during the Neogene (~10–20 Ma)<sup>1–4</sup>. The order Cycadales  
 52     possesses 380 currently known living species, including 125 species in the genus *Cycas*<sup>5</sup>.  
 53     This raises a central paradox; the coexistence of the long-term genomic stability of an ancient  
 54     lineage and a species diversification system driven by genomic plasticity in cycad.

55             A key to understanding this paradox lies in their unique reproductive strategies. The  
 56     Cycadales and Ginkgoales lineages are characterized by a fertilization system possessing  
 57     primitive motile sperm like basal plants and derived pollen tubes like seed plants<sup>6</sup>, which  
 58     intermediate state between zooidogamy and siphonogamy in cycad is supported by molecular  
 59     evidence through tissue-specific transcriptome analysis<sup>7</sup>. Furthermore, both lineages exhibit  
 60     strict dioecy governed by stable XY sex-determination systems<sup>8</sup>. Recently, a candidate sex  
 61     determination gene was identified in the male-specific region (MSR) of the Zamiaceae and  
 62     *Ginkgo* genomes<sup>2</sup>, with homologous genes found in *Cycas panzhihuaensis*. Elucidating the  
 63     genomic basis of these traits is crucial for revealing the evolutionary history of reproduction  
 64     in land plants.

65             In contrast to the wind-pollinated and species-poor ginkgo, the rapid radiation of  
 66     cycads was likely driven by an elaborate co-evolutionary synergy with insect pollinators.  
 67     Cone thermogenesis, a widespread phenomenon in cycads, is closely associated with  
 68     reproductive function, such as attracting insect pollinators through the release of volatile  
 69     emissions and/or infrared radiation<sup>9,10</sup>. This physiological trait suggests that while cycads are  
 70     morphologically conservative, they possess highly specialized physiological adaptations to  
 71     synchronize with specific biotic partners. Understanding how these specialized reproductive  
 72     traits facilitated lineage diversification is vital to explaining why Cycadales, unlike their

sister order, successfully exploited modern ecological niches through intricate biotic interactions.

To address these questions, a chromosome-scale genome assembly serves as a foundation to reveal chromosome structures and transposon evolution, determine sex-associated regions, and accurately identify gene models and their expression profiles. Owing to the advancement of sequencing technologies and sophisticated bioinformatics tools, high-quality chromosome-scale genome assemblies are now available for many plant species<sup>11</sup>, including gymnosperms<sup>2,12,13</sup>. However, the large genome size and repeat-rich sequences of gymnosperm genomes are still challenging in general. Hi-C scaffolding is a reliable method for building chromosome-scale sequences<sup>14</sup>. Alternatively, genetic mapping is also a promising method to anchor sequences to genetic maps, establishing chromosome-scale sequences even in gymnosperms<sup>13</sup>. For genetic mapping, mapping populations consisting of hundreds of samples should be generated by crossing parental lines, which requires time and space and is not realistic in organisms with large body sizes and/or long life cycles. In gymnosperms, megagametophytes have been used for linkage analysis<sup>15</sup>. In our previous study, we used pollen, each possessing a recombinant haploid chromosome derived from the heterozygous diploid genomes of the parent, as a mapping population to demonstrate that single-pollen genotyping technology is useful to establish genetic maps in plants<sup>16</sup>. These techniques would also be helpful to establish a chromosome-scale genome assembly in cycad.

In this study, we focused on *Cycas revoluta*, a species representing the northernmost limit of cycad distribution, native to southern Japan. Beyond its ecological significance, *C. revoluta* has a profound ethnoecological and symbolic history within the culture of the Ryukyu archipelago<sup>17</sup>. Here, we report a chromosome-scale genome assembly of *C. revoluta*, spanning over 11 Gb in length. By integrating this high-quality genomic resource with tissue-specific transcriptomic data, we uncovered the genetic architecture that facilitates both ancestral preservation and modern adaptive innovation. This study provides a foundational tool for resolving the central paradox of cycad evolution and provides new insights into the genomic basis of their unique reproductive and sex-determination systems.

## Materials and methods

### Plant Materials

Six *Cycas revoluta* individuals were used, including one male and one female from each of University of Miyazaki (Miyazaki, Japan), Kazusa DNA Research Institute (Kisarazu, Japan), and the Koishikawa Botanical Garden of the University of Tokyo (Tokyo, Japan). The male individual from Miyazaki was used for *de novo* genome assembly and the other five lines were used for genetic diversity analysis.

#### *Whole genome sequencing and de novo assembly*

Leaves of the male individual from Miyazaki were subjected to high-molecular-weight genomic DNA extraction using a Genomic-tip 500/G (Qiagen, Hilden, Germany). The high-molecular-weight genomic DNA was sheared to 30 kb using Megaruptor2, and long-read sequencing libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 (PacBio, California, United States). The libraries were separated using BluePippin (Sage Science, Massachusetts, United States) to remove short DNA fragments (< 15kb). Sequence data were obtained using the Sequel II system (PacBio) and converted into high-fidelity (HiFi) reads with the SMRT Link pipeline (PacBio). The HiFi reads were assembled into contig sequences using Hifiasm version 0.19.6<sup>18</sup>, with the parameter that removed 20 bases from both ends of the reads. In parallel, genome DNA was extracted from leaves of the other five individuals and sequenced using the HiFi method on the Revio system (PacBio).

An Omni-C library, a type of high-throughput chromosomal conformation capture (Hi-C)<sup>14</sup>, was prepared using an Omni-C kit (Dovetail Genomics, California, United States) with a standard procedure for scaffolding the contigs. The library was sequenced using DNBSEQ-G400 (MGI Tech, Shenzhen, China). The adapter sequences of Omni-C were trimmed with Cutadapt<sup>19</sup> and PRINSEQ<sup>20</sup>. Low-quality bases were trimmed with fastx\_clipper in the FASTX-Toolkit ([http://hannonlab.cshl.edu/fastx\\_toolkit](http://hannonlab.cshl.edu/fastx_toolkit)). The cleaned reads were mapped to the contigs using Bowtie2<sup>21</sup>, and the contigs were scaffolded with YaHS<sup>22</sup> using default parameters. The contact map was constructed with Juicer<sup>23</sup> and subsequently visualized and manually curated with Juicebox<sup>24</sup>.

Potential contaminated sequences from organelles and foreign organisms were identified and removed using Tiara v1.0.3<sup>25</sup> and the gxdb\_20240308 dataset of FCS-GX v0.5.5<sup>26</sup>. The completeness of the assembly was evaluated using BUSCO v5.5.0 with embryophyta\_odb10<sup>27</sup> and the gene set of gymnosperms<sup>28</sup>, LTR Assembly Index (LAI)<sup>29</sup>, and telomere\_finder<sup>30</sup>. The assembly was aligned to *C. panzhihuaensis* using minimap2 v2.28<sup>31</sup> and visualized with D-genies<sup>32</sup>.

# *Chromosome-scale sequence construction based on SNP analysis*

For linkage disequilibrium (LD) analysis, SNPs from pollen were obtained from a single-pollen genotyping technology<sup>16</sup>. Pollen collected from the Miyazaki male individual was subjected to protoplast preparation using the previously reported protocol<sup>33</sup>. Briefly, the pollen of *C. revoluta* was washed with pre-buffer (1/2 MS, 1M sorbitol, 0.5M glucose, 5mM 2-morpholinoethanesulphonic acid, pH 5.8, through a 0.22 µm filter) and filtered through a cell strainer with a 70 µm pore size (pluriSelect Life Science, Saxony, Germany) twice. The washed pollen was soaked in the pre-buffer containing 1.0% Macerozyme R1-0 (Cosmo Bio, Tokyo, Japan), 2.0% Cellulose Onozuka RS (Cosmo Bio), 0.1% BSA as enzyme solution and then shaken at 28°C and 20 rpm for 6 days. The pollen protoplasts were collected by centrifugation at 100g for 5 min at room temperature and washed with the pre-buffer twice. Single-cell cDNA libraries were prepared using DNBelab C4 (MGI Tech) and sequenced on DNBSEQ-G400 to generate 150 bp paired-end reads. The obtained reads were mapped to the scaffold sequences as a reference through the standard pipeline of DNBelab C Series Single-Cell RNA Analysis Software ([https://github.com/MGI-tech-bioinformatics/DNBELAB\\_C\\_Series\\_scRNA-analysis-software](https://github.com/MGI-tech-bioinformatics/DNBELAB_C_Series_scRNA-analysis-software)) and SNPs were detected with GATK HaplotypeCaller<sup>34</sup>. LD analysis was performed with SELDLA<sup>35</sup>, in which pairwise correlations between loci were calculated using Fisher's exact test.

For genetic mapping analysis, we used megagametophytes, which were collected from the Kisarazu female individual, as a mapping population. Genomic DNA was extracted from the megagametophytes using the hexadecyltrimethylammonium bromide (CTAB) method<sup>36</sup> and subjected to ddRAD-Seq library construction using the PstI and MspI enzymes<sup>37</sup>. ddRAD-Seq reads were obtained using DNBSEQ-G400, and their low-quality bases and adaptor sequences were trimmed with PRINSEQ and fastx\_clipper in the FASTX-Toolkit, respectively. The cleaned reads were mapped to the genome assembly using Bowtie2<sup>21</sup>, and sequence variants were called using BCFtools<sup>38</sup>. High-confidence SNPs were selected using VCFtools<sup>39</sup> (parameters for minimum depth of 5; minimum SNP quality of 20; and minor allele frequency of 0.5) and subjected to linkage analysis using Lep-Map3<sup>40</sup>. The resulting genetic map was assigned to the genome assembly using ALLMAPS<sup>41</sup> to validate the accuracy of the genome assembly.

## *Repetitive sequence analysis*

Repetitive sequences were detected with RepeatMasker v4.1.5 (<https://www.repeatmasker.org>) using repeat sequences obtained from the genome assembly using RepeatModeler v2.0.5 (<https://github.com/Dfam-consortium/RepeatModeler>) and from a dataset registered in Repbase<sup>42</sup>.

#### *Protein-coding gene prediction and annotation*

Total RNA was isolated from a young leaflet, a mature leaflet, a rachis, a central axis, microsporangia, an upper microsporophyll, and a lower microsporophyll of the Miyazaki male individual with RNA extraction kit (Qiagen). A library for short-read sequencing was prepared with a TruSeq RNA Library Prep Kit v2 (Illumina, California, United States) and sequenced using HiSeq 4000 (Illumina). Potential protein-coding genes were predicted using BRAKER3 version 3.0.7<sup>43</sup> with the RNA-Seq data, protein sequences of Viridiplantae from orthoDB v11<sup>44</sup>, and a gene set of gymnosperms<sup>28</sup>.

In parallel, a long-read sequencing library was prepared using the KINEX full-length RNA kit (PacBio) from RNA of a leaflet, a rachis, three stages of microsporangia, and three stages of microsporophyll of the Miyazaki male individual. The library was sequenced on the Revio system (PacBio). The obtained full-length cDNA reads were treated with the Iso-Seq pipeline (<https://isoseq.how>, PacBio), in which the reads were aligned to the genome assembly with an intron size of 500 kb, to obtain transcript sequences. Since multiple isoforms were aligned to the same genomic region, we selected a single representative transcript per locus based on the following criteria: (1) support by more than three reads; (2) the presence of an open reading frame; and (3) a splicing pattern matching either the gene models of *C. panzhihuaensis* or those predicted by the RNA-Seq analysis described above. For genomic regions where no gene models were assigned, the transcript with the highest number of reads was selected.

The genes predicted with short- and long-read sequencing were compared using GffCompare<sup>45</sup> and single representative genes per locus were selected from the long-read data with highest priority followed by the short-read data. The predicted genes were assigned functional annotations with eggNOG-mapper v2.1.8<sup>46</sup> and clustered with those of *C. panzhihuaensis*<sup>2</sup> using OrthoFinder<sup>47</sup>.

#### *Organelle genome assembly and gene prediction*

For the six individuals, chloroplast and mitochondrial genome sequences were assembled with oak v1.0<sup>48</sup> and the number of repetitive motifs was manually curated by alignment of HiFi reads using GraphAligner<sup>49</sup>. Using Lifton v1.0.5<sup>50</sup>, protein-coding genes and ribosomal genes in the chloroplast assembly were transferred from *C. revoluta* (NC\_020319.1) and those in the mitochondrial assembly were transferred from *C. taitungensis* (NC\_010303.1), *Ginkgo biloba* (NC\_027976.1), *Welwitschia mirabilis* (NC\_029130.1), *Pinus Taeda* (NC\_039746.1), *Taxus chinensis* (NC\_069591.1), *T. wallichiana* (NC\_072607.1), and *Abies koreana* (NC\_071216.1). Transfer RNA genes were predicted using tRNAscan-SE2.0<sup>51</sup> and ARAGORN<sup>52</sup> in both chloroplast and mitochondrial genomes.

#### *Sequence and structure variations within the Cycas species and across its relatives*

HiFi reads from the six individuals were mapped to the *C. revoluta* genome assembly with pbmm2 v1.17.0. Single-nucleotide variations (SNVs) and structural variants (SVs) were called with BCFtools v1.21<sup>38</sup> and pbsv v2.11.0 (<https://github.com/PacificBiosciences/pbsv>), respectively. The effects of the variants on gene function were annotated using SnpEff v4.3<sup>53</sup>. The sequence depth for each sample across the genome was calculated using goleft v0.2.6 (<https://github.com/brentp/goleft>). Homology and synteny analyses between the genomes of *Cycas* and *Ginkgo* were performed using D-genies<sup>32</sup> and MCScanX<sup>54</sup>, respectively.

#### *Phylogenetic analysis of MADS-box genes*

To identify MADS-box genes from the gene models, we aligned SRF-TF (PF00319) and K-box (PF01486) from the Pfam database (<https://www.ebi.ac.uk/interpro/>) against the gene models of *C. revoluta* (this study), *C. panzhihuaensis*<sup>2</sup>, and *G. biloba*<sup>12</sup> with hmmsearch in HMMER v3.3.2 (<http://hmmer.org/>). Protein sequences of the resultant selected genes in addition to two Y chromosome-specific MADS-box genes, GbMADS18 (MN548764) and CpMADSY (CYCAS\_034085) were aligned with MAFFT v7.505<sup>55</sup>. A phylogenetic analysis was conducted using IQ-TREE v3.0.1<sup>56</sup> with 10,000 ultrafast bootstrap replicates.

#### *Comparison of candidate genes involved in sex determination and sexual dimorphisms*

Toyama et al.<sup>7</sup> explored gene expression in pollen tubes and capacitated motile sperm of *C. revoluta* in a tissue-specific transcriptome analysis to reveal the intermediate evolutionary state of motile sperm and pollen tubes. The transcriptome analysis was performed without genome sequence data, which were unavailable at the time of analysis.

Now, in this study, since we present the *C. revoluta* genome, we aligned a total of 77,449 transcript sequences for nine tissues to the genome sequence as a reference using BLASTX searches to determine their genome positions.

To identify putative genes involved in the biosynthesis and regulation of flagella, 58 flagellar-related genes of six categories (radial spoke protein, central pair, dynein, inner dynein, outer dynein, and intraflagellar transport protein) were collected from a previous study in algae<sup>57</sup>. Similarity searches of the flagellar gene sequences were performed using DIAMOND<sup>58</sup> against the predicted genes in the genomes of *C. revoluta* (from this study) in addition to 13 land plant species (*Azolla filiculoides*, *Salvinia cucullata*, *G. biloba*, *C. panzhihuaensis*, *Gnetum montanum*, *Sequoiadendron giganteum*, *Abies alba*, *Picea abies*, *Pinus lambertiana*, *P. taeda*, *Pseudotsuga menziesii*, *Oryza sativa*, and *Arabidopsis thaliana*). We employed the reciprocal best hit method with an E value cut-off of  $1 \times 10^{-6}$  to select the top hit genes.

#### *Transcriptome analysis for thermogenesis*

Microsporophylls (ML), thermogenic cone tissue, and microsporangia (MG), non-thermogenic cone tissue, were sampled at early thermogenic, middle thermogenic, and late non-thermogenic stages during the flowering period in 2024 from three flowers of the individual used for genome sequencing. The thermogenic tissues and timing in these samples were identified using FLIR thermal image cameras. Total RNA was extracted with an RNA extraction kit (Qiagen), prepared for the library using TruSeq Stranded mRNA HT Sample Prep Kit, and sequenced with DNBSEQ-G400. The resulting reads were mapped to the genome assembly and quantified using hisat2 v2.2.1<sup>59</sup>, stringtie v2.2.1<sup>60</sup>, and edgeR v4.6.3<sup>61</sup>. Mitochondrial transit peptides were predicted from aligned transcripts using TargetP<sup>62</sup>.

## **Results**

### *Chromosome-scale genome assembly*

The genome sequence of *C. revoluta* was *de novo* assembled using 17 million HiFi reads totaling 397 Gb (~29.6x coverage of the estimated genome size of 13.4 Gb<sup>63</sup>), which were obtained from 14 SMRT cells on the Sequel II platform. The reads were assembled into 2,483 contigs spanning 11.6 Gb, with a contig N50 of 14.1 Mb (Table 1). Then, we generated 2.8 billion Omni-C short reads (426 Gb; 31.8x coverage) to scaffold the contigs. Following the

mapping of 74.0% of these reads, contigs were scaffolded into 821 sequences. Then, 408 potential contaminated sequences from organelle and foreign organisms were removed. As a result, nine large scaffold sequences exceeding 100 Mb were obtained. Since *C. revoluta* has 11 basic chromosomes, the scaffold sequences might include over-scaffolding or misassemblies.

To address these potential errors, we employed a LD approach. Since pollen grains possess haploid genomes containing recombinant parental chromosomes, haplotype blocks could be predicted using SNPs derived from pollen. We obtained transcriptome sequence data from 210 individual pollen grains and mapped the reads to the scaffold assembly to detect 283 SNPs (Supplementary Figure S1). LD analysis of these SNPs indicated distinct misassembly points in the two largest sequences, which were subsequently broken at these high-confidence junctions.

The accuracy of the resulting sequences was further validated using a genetic mapping strategy. We genotyped 880 SNPs across 224 megagametophytes to construct a genetic map. The resulting map comprised 12 linkage groups with a total length of 1,457.6 cM (Supplementary Figure S2). Ten of the linkage groups were each assigned to a single scaffold sequence, while the remaining two linkage groups were assigned to a single large scaffold, confirming its continuity.

This integrated approach successfully established 11 pseudomolecule sequences at the chromosome-scale (named CREV\_r1.0.pmol: Figure 1) together with 323 unplaced contigs, representing the haploid genome of *C. revoluta* (CREV\_r1.0.genome: Table 1). The complete BUSCO score was 97.1%, with single-copy and duplicated scores of 91.3% and 5.8%, respectively. The whole-genome LAI score was 15.15. Telomeric repeats (TTTAGGG) were identified at the 15 ends of the 11 chromosome-scale sequences, out of which 5 motifs at one end of 5 chromosome sequences included both the repeats of TTTAGGG and its variant type, TTCAGGG (Supplementary Figure S3).

The structure of the *C. revoluta* genome was consistent with that of *C. panzhihuaensis*<sup>2</sup> (Figure 2A). The orientations of the chromosomes are described based on the genomic coordinates, with the 5' end corresponding to the start of the assembled sequence and the 3' end to its terminus.

### *Repetitive sequence analysis*

Repetitive sequences occupied 9.82 Gb (84.62%) of the 11.60 Gb assembly (Supplementary Table S1). The dominant repeat types were long terminal repeat (LTR) retroelements, which accounted for 51.85% of the total genome, followed by unclassified repeats (20.42%). In contrast, DNA transposons represented only a small fraction of the genome at 3.86% (447.9 Mb).

### *Gene prediction*

Full-length cDNA sequences were obtained from eight tissues using KINNEX technology. Subsequent genome-guided analysis, which involved mapping these reads and collapsing redundant sequences, yielded 66,769 transcripts. In parallel, RNA-Seq data obtained from seven samples were mapped to the genome sequence. Gene prediction was performed by integrating the non-redundant full-length cDNA sequences with 21,221 genes predicted from RNA-Seq and protein evidence. A total of 31,481 representative protein-coding sequences were identified and functionally annotated with a complete BUSCO score of 87.5%.

Gene clustering analysis between *C. revoluta* and *C. panzhihuaensis* yielded 17,722 orthogroups, of which 16,063 were shared between both species, 864 were unique to *C. revoluta*, and 795 were unique to *C. panzhihuaensis* (Figure 2B). The number of genes between the two species was concordant in 12,340 orthogroups (69.63%), while species-specific gene family expansions or contractions were observed in 137 for *C. revoluta* and 140 for *C. panzhihuaensis*.

### *Organelle genome assemblies and polymorphism analysis*

The organelle genomes of the six individuals were assembled as single circular sequences in both chloroplasts and mitochondria. The chloroplast genome of the Miyazaki male individual was 162,174 bp in length and encoded 87 protein-coding sequences, 39 tRNA, and 8 rRNA genes. Among the six individuals, a single base insertion/deletion was found in homopolymer stretches, resulting in a total length difference of 1 bp. The mitochondrial genome of the Miyazaki male individual was 417,531 bp in length and encoded 41 protein-coding sequences, 33 tRNA, and 3 rRNA genes. Among the six individuals, nine indels were found, of which six were single base indels whereas the others were variants in the number of tandem repeats in the intergenic region, ranging from 26 to 40 repeats of a 63-bp unit.

### *Genetic variations within the species*

Genome sequence and structural variations were assessed using the HiFi reads from three male and three female individuals. We identified 10.57 million SNPs and 145,400 structural variants (SVs) over the genome. Both SNPs and SVs were significantly enriched in intergenic regions compared to genic regions (chi-square test,  $P$ -value  $< 1.0 \times 10^{-16}$ ). Even in genic regions, insertions and deletions of 5–20 kb in size were predominantly associated with LTR elements (Figure 3A), suggesting that recent retrotransposon activity continues to shape the genetic landscape of *C. revoluta*. The allele frequency distribution indicated a high prevalence of minor alleles at many loci, reflecting a diverse genetic reservoir within the species (Figure 3B).

### *Comparative analysis of genome structures across the relatives*

Comparative genomic analysis based on gene synteny revealed a highly conserved chromosome structure between *C. revoluta* and *C. panzhihuaensis* (Figure 4). This macrosynteny indicates that the large-scale genomic architecture is fundamentally maintained within the genus *Cycas*. Based on this structural conservation, the sex-determining region in *C. revoluta* was predicted to be located at the 5' end of chromosome 8, corresponding to the locus previously identified in *C. panzhihuaensis*. Furthermore, despite their deep phylogenetic divergence, large-scale synteny blocks were also observed between *C. revoluta* and *G. biloba* (Figure 4).

### *Identification of sex-associated genomic regions and genes*

Although we assembled the genome of a male individual (XY), the assembled sequences would represent either the X- or Y-associated region only. To identify and characterize the sex-specific region on the *C. revoluta* chromosome 8, we performed a depth-of-coverage analysis using the HiFi reads from the male and female individuals (Figure 5). In the three male samples, the depth-of-coverage on the 5' end of chromosome 8 was half compared to other regions which represent pseudoautosomal regions (PARs) or autosomes (Supplementary Figure S4). On the other hand, in the female samples, the 5' end of chromosome 8 was rarely mapped with reads while pseudoautosomal regions (PARs) or autosomes were uniformly covered (Supplementary Figure S4). This result suggested that the sequence of the 5' end of chromosome 8 is the Y-associated region, also known as the male-specific region (MSR) and that sequences for the X-associated region might be missing from the pseudomolecule sequences.

In parallel, genome sequences for the X-linked region were recovered from alternate contigs that were derived from alternative haplotypes of the *C. revoluta* diploid genome. First, the alternate contigs were aligned to the genome sequence of *C. panzhihuaensis*, which includes a 124-Mb X-associated region at the end of chromosome 8. As result, two contigs showed high homology to the *C. panzhihuaensis* X- associated region but low homology to the *C. revoluta* Y-linked region (Supplementary Figure S5). Then, the depth-of-coverage analysis was applied to these contigs. As expected, the depth-of-coverage on these contigs was half in the three male samples compared to other regions whereas it was even across these contigs and other regions in the three female samples (Figure 5, Supplementary Figure S4).

Subsequently, we searched for characterized genes within the sex-associated regions. We identified a gene encoding a MADS-box transcription factor (Cre1.0ch08.g000010.1) as a prime candidate for sex determination. This gene shares high sequence similarity with known potential sex-determining genes in basal gymnosperms, specifically CYCAS\_034085 (MADS-Y) in *C. panzhihuaensis* and GbMADS4 in *G. biloba*. A genome-wide survey identified 32, 28, and 47 MADS-box genes in *C. revoluta*, *C. panzhihuaensis*, and *G. biloba*, respectively (Supplementary Figure S6). Additionally, a *HOTHEAD*-like gene (Cre1.0ch08.g000080.1) was located within the sex-determining region; its orthologs in *Arabidopsis thaliana* are well-known to cause abnormal floral organ development when mutated<sup>64</sup>. Although the expression of neither the MADS-box nor the *HOTHEAD*-like gene was captured in the above male and female organ transcriptome datasets, their conserved functions and evolutionary positioning strongly support them as key candidates governing sex determination in *C. revoluta*.

#### *Genes for sex-related traits*

To identify genes associated with sex-related traits, we mapped 19,295 unique publicly available transcriptome sequences derived from six tissues<sup>7</sup> onto the *C. revoluta* genome assembly. Among these, the genomic positions of 16,668 transcripts were successfully determined. Notably, five transcripts were located within the Y-associated male-specific region (MSR), while only one was mapped to the X-associated sequences. Within the Y-associated MSR, three transcripts (Cre978\_c2\_g1\_i1.p1, Cre103\_c0\_g3\_i2.p1, and Cre39109\_c0\_g1\_i2.p1) were exclusively transcribed in male organs, making them strong tissue-specific candidates for sex determination. The remaining two Y-associated transcripts,

Cre6872\_c0\_g1\_i1.p1 and Cre94321\_c1\_g1\_i1.p1, were expressed in female organs and megasporophylls, respectively. Conversely, the single X-associated transcript (Cre160992\_c0\_g1\_i1.p1) was expressed in male organs. Among the three male-specific Y-associated transcripts, Cre103\_c0\_g3\_i2.p1 overlapped with a predicted gene model (Cre1.0ch08.g000220.1) encoding a small nuclear ribonucleoprotein, whereas the other two lacked any overlapping gene models or functional annotations.

In parallel, we searched for genes involved in flagellar formation across 14 vascular plant species including 2 angiosperms, 10 gymnosperms, and 2 ferns. The genes for flagellar biosynthesis and regulation, such as radial spoke, central pair, intraflagellar transport, and inner and outer dynein, were absent from the Y-associated region of chromosome 8. Over the genome, the copy numbers of the genes were higher in *Cycas*, *Ginkgo*, and ferns, which form sperm flagella, than other species having vascular systems but lacking flagella during evolution (Supplementary Table S2).

#### *Genes for thermogenesis*

To elucidate the molecular basis of cone thermogenesis, we performed transcriptome analysis during the diurnal thermogenic period. We identified 89 upregulated and 36 downregulated transcripts in thermogenic tissues (Supplementary Figure S7). Central to this heat-production mechanism was the identification of Alternative Oxidase 1 (AOX1). Three of the upregulated candidates, including AOX1, were predicted to possess mitochondrial transit peptides, supporting their role in diverting redox energy from the respiratory chain to heat production. These findings provide a direct link between genomic plasticity and the specialized physiological adaptations of cycads.

## **Discussion**

We present the chromosome-scale assembly of the *C. revoluta* genome spanning 11.6 Gb in length, of which 11.4 Gb was anchored to chromosomes as pseudomolecules (Figure 1, Table 1). This assembly represents the second chromosome-scale genome for the order Cycadales, enabling a direct comparison with the closely related *C. panzhihuaensis*<sup>2</sup> (Figure 2A). The *C. revoluta* chromosome-scale assembly (11.4 Gb) is 1.2 Gb longer than that of *C. panzhihuaensis* (10.2 Gb), a difference primarily attributable to repeat expansion at the distal ends of chromosomes, as no large-scale structural rearrangements were detected (Figure 2A).

While the two species share a high percentage of orthologous gene clusters (90.6%), each possesses a distinct set of unique gene clusters (4.9% in *C. revoluta* and 4.5% in *C. panzihuaensis*) that likely arose after their diversification approximately 19-20 million years ago (Figure 2B). This highlights a central paradox in cycad evolution: although they are an ancient group of 'living fossils', many cycad genera exhibit high species diversity resulting from more recent divergences. Further genomic research on cycads is needed to reveal the molecular mechanisms that enable both the conservation of a long-lived lineage and a species diversification system comparable to that of angiosperms.

Another key feature of this genome assembly was its extraordinary repeat content; repetitive sequences occupy 9.8 Gb (84.2%) of the assembly, with LTR retrotransposons alone constituting half of these repeats (Supplementary Table S1). The intraspecific genetic variation in *C. revoluta* revealed that both SNPs and SVs were biased toward intergenic regions (Figure 3A). This pattern likely reflects the prevalence of neutral mutations in intergenic regions and the impact of purifying selection against deleterious mutations in genic regions. In contrast, large indel polymorphisms (5,000 to 20,000 bp) associated with LTR retrotransposons showed a relative bias toward genic regions, including untranslated regions (UTRs) and introns. The predominance of low-frequency transposon insertion polymorphisms (Figure 3B) indicates that a recent increase in the activity of these insertions and deletions occurred during the distribution expansion, rather than being maintained at intermediate frequencies through long-term neutral evolution.

The genomic localization of TEs in plants has been documented in several pangenome studies of model organisms, such as *A. thaliana* and rice, and two non-mutually exclusive possibilities have been discussed<sup>65–67</sup>. First, the initial TE insertions may occur randomly, with their subsequent retention being favored by natural selection if they contribute to gene expression regulation and individual survival. Second, TE targeting could be directed by specific chromatin signatures, such as histone modifications that characterize transcriptionally active and accessible chromatin. These epigenetic landmarks serve as recruitment signals for new TE insertions toward genic regions and TE deletions by homologous recombination. Transposon insertion polymorphisms in UTRs and introns, which have been associated with local adaptation in angiosperms, may also occur in gymnosperms, despite their substantially larger genome sizes and higher proportions of intergenic regions compared to angiosperms. These results imply that these mechanisms may

be involved in adaptive evolution, although further functional studies are needed to confirm their role in speciation.

We constructed a continuous assembly of the sex chromosome, including the male-specific region (MSR) at the end of the chromosome 8. The sex-determining region in *C. revoluta* exhibited complex repetitive patterns expected to result in recombination suppression (Supplementary Figure S5). This structure suggests that the mechanism of sexual dimorphism maintenance is the XY system. Cycads and ginkgos are dioecious gymnosperms that possess a unique male trait intermediate between basal and other seed plants, specifically the production of large, multi-flagellated motile sperm within pollen tubes. Because the genes for sperm flagella and those expressed in male organs in *C. revoluta* are mainly located on autosomes, the sex-determination genes in the MSR could trigger the production of sperm flagella and male-specific traits.

Moreover, the phylogenetic analysis revealed species-specific MADS-box gene clusters in cycads and ginkgos. Tandem duplication and faster birth-and-death evolution of MADS-box genes are well known in model land plants<sup>68,69</sup>. These species-specific gene clusters in cycads and ginkgos might have emerged after rapid radiations ranging from 11 to 20 million years ago and might contribute to the development and structural diversification of reproductive organs. Our findings will contribute to understanding the unique sexual dimorphic traits and the evolutionary history of sex chromosomes from seedless plants to angiosperms.

We detected alternative oxidase (AOX) as a candidate associated with thermogenesis, consistent with a recent study in *Zamia furfuracea*, a cycad belonging to a different family<sup>70</sup>. The AOX detected in this study had different sequences and genomic positions from those of CrAOX1 and CrAOX2 reported in the previous study<sup>9</sup>, whereas its sequence was more similar to CrAOX1 than to CrAOX2. AOX1 is a transmembrane protein localized to the inner mitochondrial membrane, has mitochondrial transit peptides, and bypasses the proton-translocating multi-protein complexes<sup>71,72</sup>. Of the candidates, three genes with mitochondrial transit peptide, including those of unknown function, might also be associated with thermogenesis. This suggests that thermogenesis was comprehensively driven by multiple genes associated with the mitochondrial electron transport chain and cellular respiration<sup>73</sup>. The previous study detected circadian clock genes, but our study did not. The different result

is reasonable because *C. revoluta* does not clearly exhibit a circadian rhythm in thermogenic activity, unlike the genus *Zamia*<sup>9,70</sup>.

In conclusion, this high-quality genome resource establishes *C. revoluta* as a key model organism, directly resolving the genomic architecture underlying cycad diversification and species diversity. Crucially, the genomic data also offer vital genetic tools to mitigate the devastating cycad aulacaspis scale (*Aulacaspis yasumatsui*) damage currently threatening cycads in Okinawa<sup>74</sup>, while providing strategies to safely manage or reduce cycasin<sup>75</sup>, the toxin responsible for severe historical crises known as "cycad hell." Ultimately, protecting this lineage through genomic insights ensures the survival of both an ancient evolutionary wonder and a deeply rooted ethnoecological heritage.

## Acknowledgments

We thank the Koishikawa Botanical Garden (The University of Tokyo) for providing the plant materials. We also thank K. Ozawa, C. Minami, H. Tsuruoka, Y. Kishida, and A. Watanabe (Kazusa DNA Research Institute) for their technical assistance.

## Supplementary Data

**Supplementary Table S1.** Repetitive sequences in the *C. revoluta* genome assembly.

**Supplementary Table S2.** Counts of candidate genes of seven categories involved in sperm flagellum biosynthesis in 14 species.

**Supplementary Figure S1.** Relationship between markers obtained by pollen RNA-seq on Omni-C scaffolds.

The lower triangle of the matrix shows correlation coefficients between markers, while the upper triangle shows distances between marker positions.

**Supplementary Figure S2.** A genetic map of *C. revoluta*.

Left: SNP loci on the genetic map (vertical lines) and physical map (bars) are connected with horizontal lines. Right: Positions of SNP loci are indicated with dots on the genetic map (y-axis, cM) and physical map (x-axis, Mb). The vertical yellow lines indicate borders of contigs.

**Supplementary Figure S3.** Distribution of telomere repeats on each pseudomolecule with 1-Mb windows.

TTTAGGG and TTCAGGG are counted as plus, and CCCTAAA and CCCTGAA are counted as minus.

**Supplementary Figure S4.** Sequencing depth on whole genomic regions of male and female individuals from Miyazaki, Kisarazu, and Tokyo.

Red lines indicate the mean depth of the whole genome.

**Supplementary Figure S5.** Dot plot of sex-determining regions. Line colors represent sequence identity.

(A) Comparison of sex chromosomes and sex-determining regions between *C. panzhihuaensis* and *C. revoluta*. (B) Self-alignment of the X contigs in *C. revoluta*. Identical sequences are ignored. (C) Self-alignment of the initial 60 Mb of the sex chromosome in *C. revoluta*. Identical sequences are ignored.

**Supplementary Figure S6.** Phylogenetic relationship of MADS-box transcription factor genes in *C. revoluta*, *C. panzhihuaensis*, and *G. biloba*.

**Supplementary Figure S7.** Transcriptome analysis of thermogenesis-associated genes.

(A) A Venn diagram of the number of differentially expressed genes. (B) A heatmap of differentially expressed genes associated with thermogenesis. The x-axis indicates microsporophylls (ML) and microsporangia (MG) collected at early thermogenic (Early), middle thermogenic (Middle), and late non-thermogenic (Late) stages.

## Funding

This work was supported by JSPS KAKENHI (JP16H06279 (PAGS), 17K07762, 20H02917, 22H05172, 22H05173, 22H05178, 22H05181, 23KJ0741, and 25K01953) and the Kazusa DNA Research Institute Foundation.

## Data availability

533 Raw sequence reads and assembled sequences were deposited in the DNA Data Bank of  
534 Japan (DDBJ) BioProject database under the accession number PRJDB42334. The genome  
535 assembly files and gene annotation files are available in Kazusa Genome Atlas  
536 (<https://genome.kazusa.or.jp>).

537

## References

1. Coiro, M., Allio, R., Mazet, N., Seyfullah, L. J., and Condamine, F. L. 2023, Reconciling fossils with phylogenies reveals the origin and macroevolutionary processes explaining the global cycad biodiversity. *New Phytol.*, **240**, 1616–35.
2. Liu, Y., Wang, S., Li, L., et al. 2022, The Cycas genome and the early evolution of seed plants. *Nat Plants*, **8**, 389–401.
3. Gutiérrez-Ortega, J. S., Pérez-Farrera, M. A., Matsuo, A., et al. 2024, The phylogenetic reconstruction of the Neotropical cycad genus *Ceratozamia* (Zamiaceae) reveals disparate patterns of niche evolution. *Mol. Phylogenet. Evol.*, **190**, 107960.
4. Gutiérrez-Ortega, J. S., Pérez-Farrera, M. A., Sato, M. P., et al. 2024, Evolutionary and ecological trends in the Neotropical cycad genus *Dioon* (Zamiaceae): An example of success of evolutionary stasis. *Ecol. Res.*
5. Calonje, M., Stevenson, D. W., and Osborne, R. 2025, The World List of Cycads, online edition.
6. Ikeno, S., and Hirase, S. 1897, SPERMATIZOIDS IN GYMNOSPERMS. *Ann. Bot.*, **11**, 344–5.
7. Toyama, Y., Okuda, S., Suzuki, T., and Higashiyama, T. 2026, Intermediate evolutionary state of motile sperm and pollen tubes in the extant gymnosperm *Cycas revoluta*. *Proc. Natl. Acad. Sci. U. S. A.*, **123**, e2506320123.
8. Abraham, A., and Mathew, P. M. 1962, Cytological Studies in the Cycads: Sex Chromosomes in *Cycas*. *Ann. Bot.*, **26**, 261–6.
9. Ito-Inaba, Y., Sato, M., Sato, M. P., et al. 2019, Alternative Oxidase Capacity of Mitochondria in Microsporophylls May Function in Cycad Thermogenesis. *Plant Physiol.*, **180**, 743–56.
10. Peris, D., Postigo-Mijarra, J. M., Peñalver, E., et al. 2024, The impact of thermogenesis on the origin of insect pollination. *Nat. Plants*, **10**, 1297–303.
11. Aoyagi Blue, Y., Iimura, H., Sato, M. P., and Shirasawa, K. 2025, The impact of telomere-to-telomere genome assembly in the plant pan-genomics era. *Breed. Sci.*, **75**, 3–12.
12. Liu, H., Wang, X., Wang, G., et al. 2021, The nearly complete genome of *Ginkgo biloba* illuminates gymnosperm evolution. *Nat. Plants*, **7**, 748–56.

13. Shirasawa, K., Blue, Y. A., Mishima, K., Hirakawa, H., and Hirao, T. 2026, January 7, Chromosomal evolution and speciation inferred from chromosome-scale genome assemblies in three Cupressaceae species. *bioRxiv*. bioRxiv.
14. Dudchenko, O., Batra, S. S., Omer, A. D., et al. 2017, De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, **356**, 92–5.
15. Guries, R. P., Friedmanf, S. T., and Ledig, F. T. 1978, A megagametophyte analysis of genetic linkage in pitch pine (*Pinus rigida* mill.). *Heredity (Edinb.)*, **40**, 309–14.
16. Shirasawa, K., Yoshitake, K., Kondo, H., Kikuchi, S., Koiwai, K., and Fujii, S. 2026, Chromosome-scale genome assembly of *Malcolmia littorea* using long-read sequencing and single-pollen genotyping technologies. *DNA Res.*, **33**.
17. Englehardt, J. D., Carrasco, M. D., Gutiérrez-Ortega, J. S., Deloso, B. E., and Matsubara, N. 2025, Threats to cycad biocultural heritage in the Amami Islands, Japan. *Plant Species Biol.*, **40**, 292–310.
18. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. 2021, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods*, **18**, 170–5.
19. Martin, M. 2011, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, **17**, 10.
20. Schmieder, R., and Edwards, R. 2011, Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–4.
21. Langmead, B., and Salzberg, S. L. 2012, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–9.
22. Zhou, C., McCarthy, S. A., and Durbin, R. 2023, YaHS: yet another Hi-C scaffolding tool. *Bioinformatics*, **39**.
23. Durand, N. C., Shamim, M. S., Machol, I., et al. 2016, Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.*, **3**, 95–8.
24. Durand, N. C., Robinson, J. T., Shamim, M. S., et al. 2016, Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.*, **3**, 99–101.
25. Karlicki, M., Antonowicz, S., and Karnkowska, A. 2022, Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics*, **38**, 344–50.
26. Astashyn, A., Tvedte, E. S., Sweeney, D., et al. 2024, Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol.*, **25**, 60.
27. Manni, M., Berkeley, M. R., Seppely, M., Simão, F. A., and Zdobnov, E. M. 2021, BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper

- Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.*, **38**, 4647–54.
28. Wu, J.-J., Han, Y.-W., Lin, C.-F., Cai, J., and Zhao, Y.-P. 2023, Benchmarking gene set of gymnosperms for assessing genome and annotation completeness in BUSCO. *Hortic. Res.*, **10**, uhad165.
29. Ou, S., Chen, J., and Jiang, N. 2018, Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.*, **46**, e126.
30. Sato, M. P., Iwakami, S., Fukunishi, K., et al. 2023, Telomere-to-telomere genome assembly of an allotetraploid pernicious weed, *Echinochloa phyllopogon*. *DNA Res.*, **30**.
31. Li, H. 2018, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–100.
32. Cabanettes, F., and Klopp, C. 2018, D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, **6**, e4958.
33. Fan, L., Wang, Y.-F., Wang, H., and Wu, W.-H. 2001, In vitro Arabidopsis pollen germination and characterization of the inward potassium currents in Arabidopsis pollen grain protoplasts. *J. Exp. Bot.*, **52**, 1603–14.
34. Poplin, R., Ruano-Rubio, V., DePristo, M. A., et al. 2017, November 14, Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. bioRxiv.
35. Yoshitake, K., Igarashi, Y., Mizukoshi, M., et al. 2018, Artificially designed hybrids facilitate efficient generation of high-resolution linkage maps. *Sci. Rep.*, **8**, 16104.
36. Murray, M. G., and Thompson, W. F. 1980, Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.*, **8**, 4321–6.
37. Shirasawa, K., Hirakawa, H., and Isobe, S. 2016, Analytical workflow of double-digest restriction site-associated DNA sequencing based on empirical and in silico optimization in tomato. *DNA Res.*, **23**, 145–53.
38. Danecek, P., Bonfield, J. K., Liddle, J., et al. 2021, Twelve years of SAMtools and BCFtools. *Gigascience*, **10**.
39. Danecek, P., Auton, A., Abecasis, G., et al. 2011, The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–8.
40. Rastas, P. 2017, Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics*, **33**, 3726–32.
41. Tang, H., Zhang, X., Miao, C., et al. 2015, ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.*, **16**, 3.

42. Bao, W., Kojima, K. K., and Kohany, O. 2015, Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, **6**, 11.
43. Gabriel, L., Brûna, T., Hoff, K. J., et al. 2024, BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.*, **34**, 769–77.
44. Kuznetsov, D., Tegenfeldt, F., Manni, M., et al. 2023, OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.*, **51**, D445–51.
45. Pertea, G., and Pertea, M. 2020, GFF utilities: GffRead and GffCompare. *F1000Res.*, **9**, 304.
46. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. 2021, EggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.*, **38**, 5825–9.
47. Emms, D. M., and Kelly, S. 2019, OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.*, **20**, 238.
48. Zhou, C., Brown, M., Blaxter, M., Darwin Tree of Life Project Consortium, McCarthy, S. A., and Durbin, R. 2025, Oatk: a de novo assembly tool for complex plant organelle genomes. *Genome Biol.*, **26**, 235.
49. Rautiainen, M., and Marschall, T. 2020, GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.*, **21**, 253.
50. Chao, K.-H., Heinz, J. M., Hoh, C., et al. 2025, Combining DNA and protein alignments to improve genome annotation with LiftOn. *Genome Res.*, **35**, 311–25.
51. Chan, P. P., Lin, B. Y., Mak, A. J., and Lowe, T. M. 2021, tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.*, **49**, 9077–96.
52. Laslett, D., and Canback, B. 2004, ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.*, **32**, 11–6.
53. Cingolani, P., Platts, A., Wang, L. L., et al. 2012, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, **6**, 80–92.
54. Wang, Y., Tang, H., Debarry, J. D., et al. 2012, MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.*, **40**, e49.
55. Katoh, K., and Standley, D. M. 2013, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–80.

56. Wong, T., Ly-Trong, N., Ren, H., et al. 2025, April 7, IQ-TREE 3: Phylogenomic inference software using complex evolutionary models. *EcoEvoRxiv*.
57. Li, L., Wang, S., Wang, H., et al. 2020, The genome of *Prasinoderma coloniale* unveils the existence of a third phylum within green plants. *Nat. Ecol. Evol.*, **4**, 1220–31.
58. Buchfink, B., Xie, C., and Huson, D. H. 2015, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
59. Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. 2019, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–15.
60. Shumate, A., Wong, B., Pertea, G., and Pertea, M. 2022, Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLoS Comput. Biol.*, **18**, e1009730.
61. Chen, Y., Chen, L., Lun, A. T. L., Baldoni, P. L., and Smyth, G. K. 2025, edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Res.*, **53**.
62. Almagro Armenteros, J. J., Salvatore, M., Emanuelsson, O., et al. 2019, Detecting sequence signals in targeting peptides using deep learning. *Life Sci Alliance*, **2**.
63. Zonneveld, B. J. M. 2012, Genome sizes for all genera of Cycadales. *Plant Biol. (Stuttg.)*, **14**, 253–6.
64. Krolkowski, K. A., Victor, J. L., Wagler, T. N., Lolle, S. J., and Pruitt, R. E. 2003, Isolation and characterization of the *Arabidopsis* organ fusion gene HOTHEAD. *Plant J.*, **35**, 501–11.
65. Naito, K., Cho, E., Yang, G., et al. 2006, Dramatic amplification of a rice transposable element during recent domestication. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 17620–5.
66. Quadrana, L., Etcheverry, M., Gilly, A., et al. 2019, Transposition favors the generation of large effect mutations that may facilitate rapid adaption. *Nat. Commun.*, **10**, 3421.
67. Kang, M., Wu, H., Liu, H., et al. 2023, The pan-genome and local adaptation of *Arabidopsis thaliana*. *Nat. Commun.*, **14**, 6259.
68. Nam, J., Kim, J., Lee, S., An, G., Ma, H., and Nei, M. 2004, Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 1910–5.
69. Zhao, T., Holmer, R., de Bruijn, S., Angenent, G. C., van den Burg, H. A., and Schranz, M. E. 2017, Phylogenomic synteny network analysis of MADS-box transcription factor

- genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation. *Plant Cell*, **29**, 1278–92.
70. Valencia-Montoya, W. A., Liénard, M. A., Rosser, N., et al. 2025, Infrared radiation is an ancient pollination signal. *Science*, **390**, 1164–70.
71. Rasmusson, A. G., Geisler, D. A., and Møller, I. M. 2008, The multiplicity of dehydrogenases in the electron transport chain of plant mitochondria. *Mitochondrion*, **8**, 47–60.
72. Møller, I. M., Rasmusson, A. G., and Van Aken, O. 2021, Plant mitochondria - past, present and future. *Plant J.*, **108**, 912–59.
73. Otsubo, M., Sato, M. P., Suzuki-Masuko, H., et al. 2026, Symplocarpus renifolius SWEET4 and SWEET6 may contribute to stamen development via hexose transport during floral thermogenesis. *Biosci. Biotechnol. Biochem.*, **90**, 185–94.
74. Deloso, B. E., Gutiérrez-Ortega, J. S., Chang, J.-T., et al. 2025, Biological invasion by the cycad - specific scale pest *Aulacaspis yasumatsui* ( Diaspididae ) into *Cycas revoluta* ( Cycadaceae ) populations on Amami - Oshima and Okinawa - jima, Japan. *Plant Species Biol.*, **40**, 280–91.
75. Nishida, K., Kobayashi, A., and Nagahama, T. 1955, 12. Studies on cycasin, a new toxic glycoside, of *Cycas revoluta* Thunb. *Bull. Agric. Chem. Soc. Jpn.*, **19**, 77–84.

## Figure legends

**Figure 1.** A chromosome-scale genome assembly of *C. revoluta*.

(A) A Circos plot of the pseudomolecule assemblies with 1 Mb bins. The tracks from the outside to the inside are (a) pseudo-chromosome numbers, (b) GC content, (c) the number of genes, and (d) repeat density. (B) Omni-C contact map. The color indicates the intensity of interaction, ranging from low (white) to high (red). Blue boxes indicate pseudomolecule assemblies.

**Figure 2.** Comparison between *C. revoluta* and *C. panzhihuaensis*.

(A) A dot plot between the pseudomolecules of the two species. Line colors represent sequence identity. (B) A Venn diagram of gene clusters.

**Figure 3.** Structural variants within *C. revoluta*.

(A) Distribution of INDEL length. (B) Insertion allele frequency of large INDELs (5,000 - 20,000 bp).

**Figure 4.** Synteny and collinearity of genes in *C. revoluta*, *C. panzhihuaensis*, and *G. biloba*.

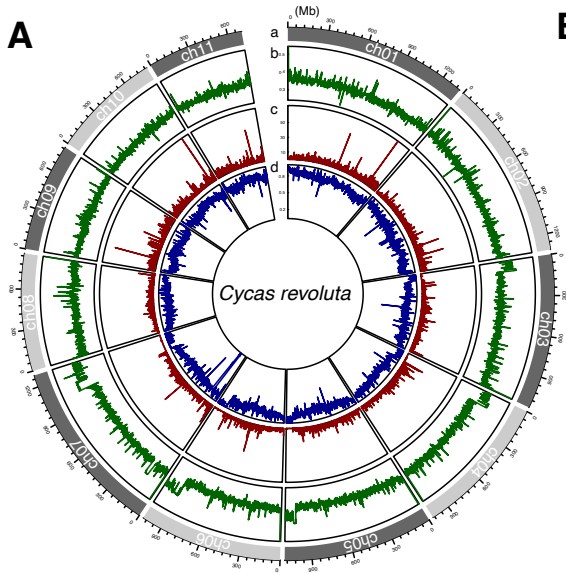
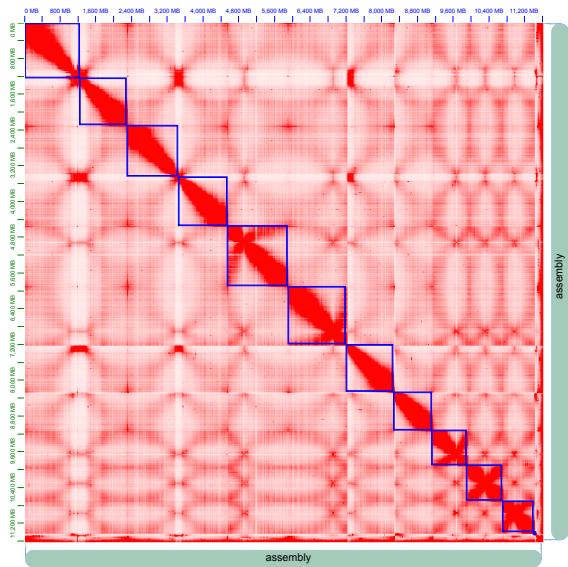
**Figure 5.** Sequencing depth on sex-determining regions of male and female individuals from Miyazaki.

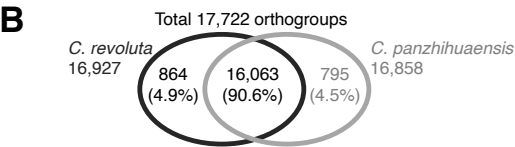
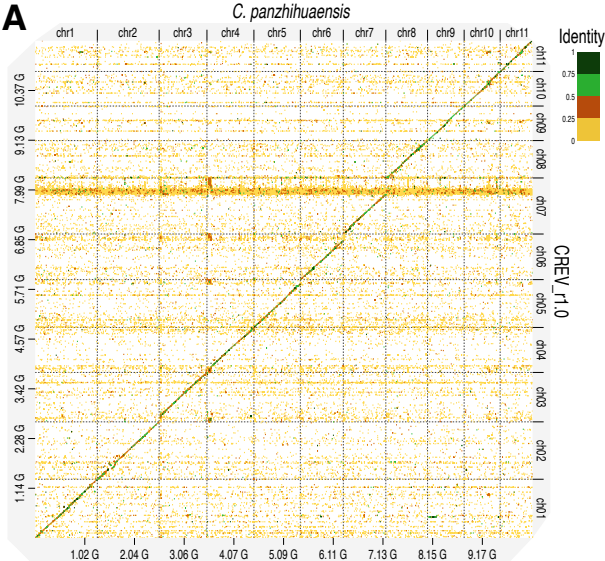
Red lines indicate the mean depth of the whole genome.

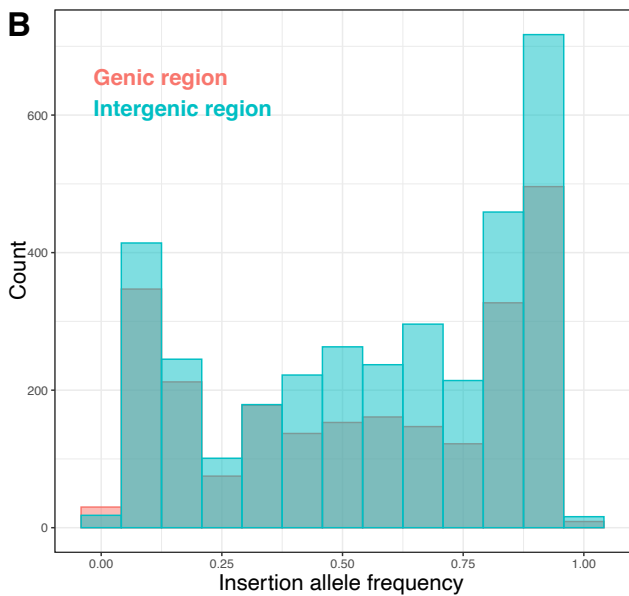
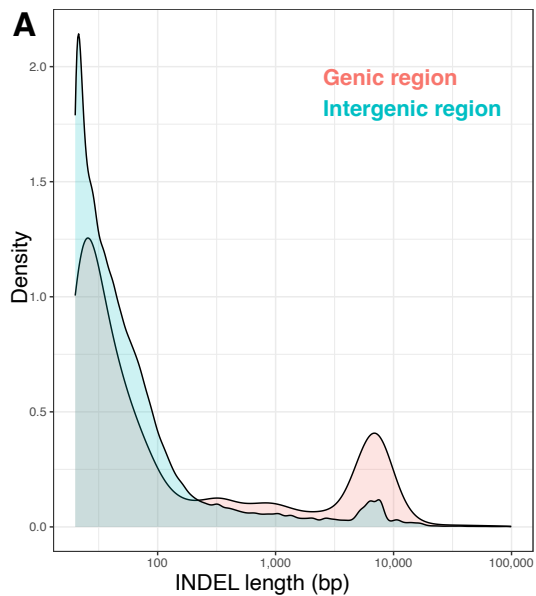
739 **Table 1.** Statistics of the *C. revoluta* genome assembly.

|                               | Contig sequence | Genome sequence<br>(CREV_r1.0.genome) | Chromosome-scale<br>sequence<br>(CREV_r1.0.pmol) |
|-------------------------------|-----------------|---------------------------------------|--------------------------------------------------|
| Total size (bp)               | 11,598,888,748  | 11,582,440,925                        | 11,413,479,678                                   |
| No. of sequences              | 2,483           | 334                                   | 11                                               |
| N50 length (bp)               | 14,138,399      | 1,098,390,435                         | 1,098,390,435                                    |
| Longest sequence size<br>(bp) | 176,690,324     | 1,348,290,124                         | 1,348,290,124                                    |
| No. of gap                    | 0               | 1,756                                 | 1,691                                            |
| Gap length (bp)               | 0               | 175,600                               | 169,100                                          |
| No. of predicted genes        | Not analyzed    | 31,481                                | 31,084                                           |

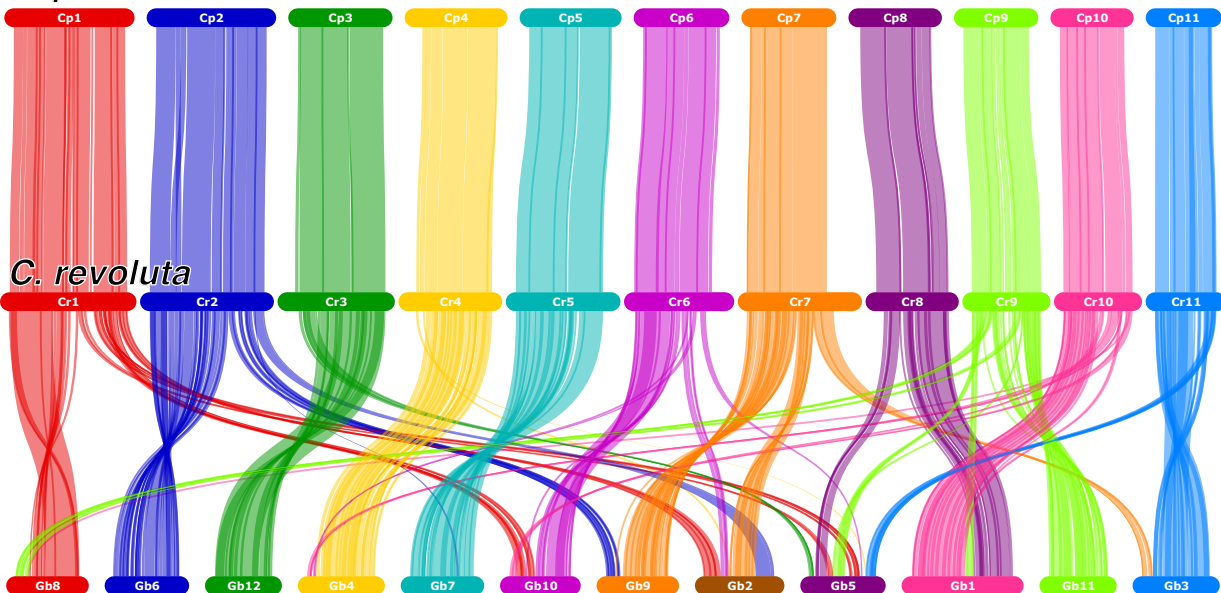
740

**A****B**





*C. panzhihuaensis*



*C. revoluta*

*G. biloba*

