

1 **Chromosomal evolution and speciation inferred from chromosome-scale genome**

2 **assemblies in three Cupressaceae species**

3

4 Kenta Shirasawa^{1*}, Yuta Aoyagi Blue¹, Kentaro Mishima², Hideki Hirakawa³, Tomonori Hirao⁴

5

6 ¹Kazusa DNA Research Institute, Chiba 292-0818, Japan

7 ²Kyushu Regional Breeding Office, Forest Tree Breeding Center, Forestry and Forest Products

8 Research Institute, Forest Research and Management Organization, Kumamoto 861-1102, Japan

9 ³Faculty of Agriculture, Kyushu University, Fukuoka 819-0395, Japan

10 ⁴Forest Tree Breeding Center, Forestry and Forest Products Research Institute, Forest Research

11 and Management Organization, Ibaraki 319-1301, Japan

12

13 *To whom correspondence should be addressed:

14 Kenta Shirasawa

15 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan

16 Tel.: +81-438-52-3935

17 Fax: +81-438-52-3934

18 E-mail: shirasaw@kazusa.or.jp

19

20 **Abstract**

21 Chromosome-scale genome assemblies in gymnosperms have lagged behind those of
 22 angiosperms, likely due to their large genomes. Coniferous tree species, which belong to the
 23 gymnosperms, are important resources for wood production in the forestry industry. To
 24 elucidate the evolution and speciation of these species and establish genome resources for
 25 breeding, we integrated draft assemblies with optical and genetic mapping to construct
 26 chromosome-scale genomes for Japanese cypress (*Chamaecyparis obtusa*, 8.7 Gb), Japanese
 27 cedar (*Cryptomeria japonica*, 9.6 Gb), and Chinese fir (*Cunninghamia lanceolata*, 13.4 Gb).
 28 Additionally, we assembled and annotated their chloroplast and mitochondrial genomes.
 29 Comparative analysis of the nuclear genomes revealed that while synteny is largely conserved,
 30 distinct translocations and inversions occurred in chromosomes 2, 6, and 9. Notably, the
 31 significantly larger genome of *C. lanceolata* was associated with frequent tandem gene
 32 duplications rather than transposon expansion. These findings suggest that chromosomal
 33 rearrangements and segmental duplications played key roles in the divergence of these species.
 34 The genomic resources presented here including chromosome-scale sequences, gene annotations,
 35 and genetic maps will facilitate advanced conifer genetics and accelerate forest tree breeding
 36 programs.

37

38 **Keywords:** Chromosome-scale genome assembly; Comparative genomics; Cupressaceae;

39 Forestry; Genetic mapping

40

41 **Introduction**

42 The Spermatophyta, or seed plants, are broadly categorized into angiosperms and gymnosperms.
 43 Whereas angiosperms include more than 60 orders (Chase *et al.* 2016), gymnosperms currently
 44 consist of only eight orders (Christenhusz *et al.* 2011): Araucariales, Cupressales, Cycadales,
 45 Ephedrales, Ginkgoales, Gnetales, Pinales, and Welwitschiales. Although the classification of
 46 both angiosperms and gymnosperms has been debated, molecular techniques have played a
 47 significant role in resolving these arguments. Gene sequences of organelle genomes, which are
 48 conserved across species yet slightly divergent, are recognized as DNA barcodes for identifying
 49 plant species (Hebert *et al.* 2003). More recently, with great advancements in genome
 50 sequencing technologies, nuclear genome sequences at the chromosome level have been applied
 51 to the classification of angiosperms (Zhao *et al.* 2021). Even though the genome sizes of
 52 angiosperms range from 100 Mb to 100 Gb (Pellicer and Leitch 2020), sequenced plant species
 53 have been biased toward those with small genome sizes (<1 Gb) (Shirasawa *et al.* 2021). In
 54 gymnosperms, on the other hand, most species possess massive genomes, approximately 10 Gb
 55 or more (Pellicer and Leitch 2020). Therefore, the assembly of chromosome-scale genome
 56 sequences in gymnosperms has lagged behind that of angiosperms.

57 Among gymnosperms, the orders Pinales, Cupressales, and Araucariales are
 58 collectively known as conifers (Coniferophyta). Coniferous trees are vital resources for the
 59 forestry industry and for housing materials. In Japan, approximately 60% of the land is forested;
 60 of this, 40% comprises plantation (artificial) forests dominated by Japanese cedar (*Cryptomeria*
 61 *japonica*), followed by Japanese cypress (*Chamaecyparis obtusa*) (Forestry Agency 2022,
 62 Tsumura *et al.* 2007) . Recently, Chinese fir (*Cunninghamia lanceolata*) has been recognized as
 63 a promising new woody resource due to its fast growth and desirable wood properties (Fujisawa
 64 2017). However, breeding programs for coniferous trees are generally time-consuming because

their generation times and harvest rotations often exceed 50 years. Genomic information would help accelerate these breeding programs by identifying genes that control traits of importance to the forestry industry. In our previous study (Shirasawa *et al.* 2024), we released draft genome sequences for *C. obtusa*, *C. japonica*, and *C. lanceolata*; however, these sequences were highly fragmented and lacked gene annotations.

Comparative analysis of genome structures using chromosome-scale assemblies can elucidate the evolution and speciation processes of species (Aoyagi Blue *et al.* 2025). Moreover, gene annotations, combined with chromosome-scale sequences, are valuable resources for identifying genes for use in breeding programs (Shirasawa *et al.* 2021). In this study, we extended the contiguity of the draft sequences for *C. obtusa*, *C. japonica*, and *C. lanceolata* using optical mapping and anchored the sequences to chromosomes using a genetic mapping strategy. Comparative genomics using the resulting chromosome-scale sequences provided insights into the evolution and speciation of these three coniferous tree species.

Materials and methods

Plant materials

We used the same individual trees of *C. obtusa* (tree ID: GFB00119), *C. japonica* (GFA01029), and *C. lanceolata* (GFHN00090) that were employed for the draft genome assemblies in our previous study (Shirasawa *et al.* 2024). For genetic mapping, we utilized four mapping populations for *C. obtusa*, five for *C. japonica*, and two for *C. lanceolata*.

Optical mapping

Genomic DNA was extracted from young leaves of the three species using the Plant DNA Isolation Kit (Bionano Genomics, San Diego, CA, USA), following the Bionano Prep Plant

89 Tissue DNA Isolation Base Protocol. The isolated genomic DNA was treated with the DLE-1
90 enzyme and labeled with a fluorescent dye supplied in the DLS DNA Labeling Kit (Bionano
91 Genomics). The labeled DNA was analyzed on the Saphyr Optical Genome Mapping
92 Instrument (Bionano Genomics). The resulting molecules were assembled and merged with the
93 draft assembly to generate hybrid scaffold sequences using Bionano Solve (Bionano Genomics)
94 with default parameters.

95

96 *Genetic mapping and pseudomolecule sequence construction*

97 DNA was extracted from all individuals of the mapping populations and their parental lines
98 using the sbeadex DNA extraction kit on the oKtopure system (LGC, United Kingdom). The
99 extracted DNA was digested with PstI and MspI restriction endonucleases and subjected to
100 ddRAD-Seq library preparation (Shirasawa *et al.* 2016). The resulting libraries were sequenced
101 on a DNBSEQ G400 instrument (MGI Tech) to generate 100-bp paired-end reads. Alternatively,
102 specific populations were genotyped using GRAS-Di (Enoki and Takeuchi 2018), and the
103 sequence reads were processed as described below. After removing adapter sequences
104 (AGATCGGAAGAGC for ddRAD-Seq or CTGTCTCTTATACACATCT for GRAS-Di) using
105 fastx_clipper in the FASTX-Toolkit (version 0.0.14) and trimming low-quality reads (quality
106 score < 10) using PRINSEQ (version 0.20.4) (Schmieder and Edwards 2011), high-quality reads
107 were aligned to the hybrid scaffold sequences using Bowtie2 (version 2.3.5.1) (Langmead and
108 Salzberg 2012). High-confidence biallelic SNPs were identified using the mpileup and call
109 options of BCFtools (version 1.9) (Li 2011) and filtered using VCFtools (version 0.1.16)
110 (Danecek *et al.* 2011) based on the following criteria: read depth ≥ 5 ; SNP quality = 999; minor
111 allele frequency ≥ 0.2 ; and proportion of missing data < 20%. Linkage analysis of the SNPs was
112 performed using Lep-Map3 (version 0.2) (Rastas 2017) to construct a genetic map. Finally, the

113 hybrid scaffold sequences were anchored to the genetic map to establish chromosome-level
 114 pseudomolecule sequences using ALLMAPS (version 0.7.3) (Tang *et al.* 2015).
 115
 116 *Transcriptome analysis with long- and short-read sequencing*
 117 Total RNA was extracted from four tissues (male and female strobili, stems, and leaves) of *C.*
 118 obtusa and three tissues (shoots, stems, and leaves) of *C. lanceolata* using the FavorPrep Plant
 119 Total RNA Extraction Kit (Favorgen Biotech Corp, Ping Tung, Taiwan). Full-length RNA
 120 libraries for long-read cDNA sequencing were prepared using the Iso-Seq Express 2.0 Kit or the
 121 Kinnex Full-Length RNA Kit, and sequenced on the Sequel II or Revio systems (PacBio, Menlo
 122 Park, CA, USA). The resulting full-length cDNA sequences were processed using the Iso-Seq
 123 pipeline to generate non-redundant transcript sequences.
 124 RNA libraries for short-read cDNA sequencing were prepared using the TruSeq
 125 Stranded mRNA Sample Preparation Kit (Illumina) and sequenced on a DNBSEQ G400
 126 instrument (MGI Tech, Shenzhen, China). Adapter sequences and low-quality bases were
 127 trimmed as described above.
 128
 129 *Gene and repetitive sequence annotation*
 130 Protein-coding genes were predicted in three steps: 1) Transcript-based prediction was
 131 performed using GeneMarkS-T (Besemer and Borodovsky 2005) with full-length cDNA
 132 sequences generated from Iso-Seq or Kinnex; 2) *Ab initio* prediction was conducted using
 133 Helixer (Stiehler *et al.* 2021); and 3) BRAKER2 (Brůna *et al.* 2021) was used in conjunction
 134 with short-read RNA-Seq data to identify genes missed by the first two methods. Gene
 135 prediction completeness was evaluated using Benchmarking Universal Single-Copy Orthologs

(BUSCO) with default parameters (Simão *et al.* 2015) and the gymnosperms_odb10 lineage dataset (Wu *et al.* 2023).

Repetitive sequences in the genome assemblies were identified using RepeatMasker (<https://www.repeatmasker.org>) with reference repeat sequences from Repbase and *de novo* repeat libraries constructed using RepeatModeler (<https://www.repeatmasker.org>).

Comparative analysis of the genome structure

Pairwise collinear blocks were identified using MCScanX (Wang *et al.* 2012) with a match score threshold (-k) of 120 and an E-value cutoff of 10^{-50} . The results were visualized using the synteny browser SynVisio (Bandi and Gutwin 2020).

Organelle genome assembly and annotation

Chloroplast and mitochondrial genomes were assembled using HiFi reads obtained in our previous study (Shirasawa *et al.* 2024) via Oatk (version 1.0), with the syncmer coverage threshold (-c) set to 30 (Zhou *et al.* 2025). While Oatk successfully generated both chloroplast and mitochondrial sequences for *C. obtusa* and *C. lanceolata*, it did not automatically assemble the mitochondrial genome of *C. japonica*. Consequently, we extracted mitochondrial unitigs from the Oatk assembly graph—identified via HMM search—and manually constructed a contig using Bandage (version 0.9.0) (Wick *et al.* 2015).

Chloroplast genes (protein-coding, tRNA, and rRNA) were annotated using LiftOn (version 1.0.5) (Chao *et al.* 2025) based on alignments to published reference genomes: *C. obtusa* (GenBank accession PV469661.1), *C. japonica* (NC_010548.1), and *C. lanceolata* (NC_021437.1). Mitochondrial genes were similarly annotated using LiftOn, utilizing references from *Platycladus orientalis* (OL703044.1–OL703045.1), *Thuja sutchuenensis*

(ON603305.1–ON603308.1), *Taxus chinensis* (NC_069591.1), *Taxus wallichiana* (NC_072607.1), and *Taxus cuspidata* (MN593023.1). Additionally, mitochondrial genes were predicted using PMGA (Li *et al.* 2025), a pipeline for plant mitochondrial genome annotation. All gene structures were manually curated and corrected, and final annotations were visualized using OGDRAW (version 1.3.1) (Greiner *et al.* 2019).

165

166 Results

167 Genome assembly and gene annotation for *C. obtusa*

168 In a previous study (Shirasawa *et al.* 2024), we established a draft genome assembly for *C.*
169 *obtusa* containing primary contigs spanning 8.5 Gb (N50 = 1.2 Mb) and alternate contigs
170 spanning 7.8 Gb (N50 = 496.6 kb). To extend sequence contiguity, the combined 16.3 Gb of
171 primary and alternate sequences were scaffolded using 2,058.1 Gb of optical mapping data
172 (7,538,216 molecules; N50 length = 289.7 kb). The resulting hybrid scaffold assembly
173 comprised 848 sequences with a total length of 9.6 Gb and an N50 of 54.2 Mb.

174 Subsequently, four genetic maps of *C. obtusa* were constructed via linkage analysis of
175 SNPs across four mapping populations. Each genetic map consisted of 11 linkage groups
176 (Supplementary Table S1), corresponding to the basic chromosome number of *C. obtusa*. Based
177 on these genetic maps, 378 hybrid scaffold sequences were anchored to the 11 linkage groups,
178 establishing chromosome-scale pseudomolecule sequences spanning a total of 8.7 Gb (Table 1).
179 The remaining 470 sequences (totaling 899.6 Mb) were unplaced.

180 For gene prediction, transcriptome sequences were generated using both long- and
181 short-read sequencing. Full-length cDNA sequences were clustered into 456,139 sequences
182 containing 283,181 isoforms. From these, 27,384 non-redundant sequences were selected to
183 predict 19,938 gene models, achieving a complete BUSCO score of 82.6%. In parallel, 1,728

184 million RNA-Seq reads from 96 samples were used to predict 41,266 gene models (complete
185 BUSCO score: 61.1%). Furthermore, the *ab initio* method predicted 50,494 gene models
186 (complete BUSCO score: 60.6%). Finally, we merged these three datasets to select a total of
187 54,034 non-redundant gene models (Table 2), achieving a final complete BUSCO score of
188 90.1% (Table 3).

189

190 *Genome assembly and gene annotation for C. japonica*

191 The draft genome assembly for *C. japonica* comprised primary contigs spanning 9.2 Gb (N50 =
192 8.3 Mb) and alternate contigs spanning 8.7 Gb (N50 = 2.1 Mb) (Shirasawa *et al.* 2024). A total
193 of 17.9 Gb of sequences was scaffolded using 2,151.5 Gb of optical mapping data (7,318,858
194 molecules; N50 length = 288.9 kb). The resulting hybrid scaffold assembly consisted of 805
195 sequences with a total length of 11.6 Gb and an N50 of 37.9 Mb.

196 Five genetic maps were constructed for *C. japonica* using five mapping populations.
197 Each genetic map consisted of 11 linkage groups (Supplementary Table S2), corresponding to
198 the basic chromosome number of *C. japonica*. Based on these maps, 294 hybrid scaffold
199 sequences were anchored to the 11 linkage groups to establish chromosome-scale
200 pseudomolecules spanning 9.6 Gb (Table 1). The remaining 511 sequences (2.0 Gb total) were
201 unplaced.

202 For gene prediction, we used 75,348 isoforms derived from full-length cDNA
203 sequences to predict 14,449 gene models (BUSCO completeness: 83.4%). Next, we utilized
204 34,731 unigene sequences from our previous study (Mishima *et al.* 2018) to predict 25,856 gene
205 models (BUSCO completeness: 59.9%). Additionally, the *ab initio* method predicted 58,119
206 gene models (BUSCO completeness: 61.4%). These three datasets were merged to select a final

207 set of 49,899 non-redundant gene models (Table 2), achieving a complete BUSCO score of
208 90.5% (Table 3).

209

210 *Genome assembly and gene annotation for C. lanceolata*

211 The draft genome assembly for *C. lanceolata* contained primary contigs spanning 11.5 Gb (N50
212 = 11.7 Mb) and alternate contigs spanning 11.0 Gb (N50 = 2.9 Mb) (Shirasawa *et al.* 2024). A
213 total of 22.5 Gb of sequence was scaffolded using 2,191.6 Gb of optical mapping data
214 (5,368,583 molecules; N50 = 397.9 kb). The resulting hybrid scaffold assembly comprised 822
215 sequences with a total length of 16.9 Gb and an N50 of 31.8 Mb.

216 Two genetic maps for *C. lanceolata* were constructed using two distinct mapping
217 populations. One genetic map consisted of 11 linkage groups (Supplementary Table S3),
218 corresponding to the basic chromosome number of *C. lanceolata*. The second map consisted of
219 16 linkage groups (Supplementary Table S3); in six cases, pairs of linkage groups were assigned
220 to a single chromosome sequence, while one chromosome was not represented by any linkage
221 group. Ultimately, 298 hybrid scaffold sequences were anchored to the 11 linkage groups to
222 establish chromosome-scale pseudomolecule sequences spanning a total of 13.4 Gb (Table 1).
223 The remaining 524 sequences (3.5 Gb total) were unplaced.

224 For gene prediction, we utilized 1,088,667 non-redundant full-length cDNA sequences
225 to predict 35,735 gene models (complete BUSCO score: 90.5%). Next, 113.2 million RNA-Seq
226 reads from stem, root, and leaf samples (NCBI SRA accession numbers
227 SRR5282557–SRR5282559) were used to predict 47,922 gene models (complete BUSCO
228 score: 62.0%). Additionally, the *ab initio* method predicted 45,934 gene models (complete
229 BUSCO score: 40.6%). These three datasets were merged to select a total of 72,169

230 non-redundant gene models (Table 2), achieving a final complete BUSCO score of 93.8%

231 (Table 3).

232

233 *Repetitive sequence analysis*

234 Repetitive sequences account for 77.5%, 80.5%, and 80.9% of the genomes of *C. obtusa*, *C.*

235 *japonica*, and *C. lanceolata*, respectively. LTR retrotransposons were the most abundant class

236 of repetitive sequences (48.5–53.6%), followed by unclassified repeats (14.3–21.6%), DNA

237 transposons (5.2–6.8%), and LINEs (3.3–4.9%) (Table 4).

238

239 *Comparative analysis of the three coniferous genomes*

240 The chromosome-scale genome sequences of the three species were compared pairwise (Figure

241 1A). This comparison revealed that while the structures of eight chromosomes (Chr 1, 3, 4, 5, 7,

242 8, 10, and 11) were well conserved across all three species, three chromosomes (Chr 2, 6, and 9)

243 underwent significant rearrangements (Figure 1B).

244 In *C. obtusa* chromosome 2, the top region (COBch02t) was conserved across all three

245 species. However, the bottom region (COBch02b) was inverted and translocated to the top of

246 chromosome 6 in both *C. japonica* (CJAch06t) and *C. lanceolata* (CLAch06t). Concurrently,

247 regarding *C. obtusa* chromosome 6, the bottom region was conserved across species, while the

248 top region (COBch06t) was translocated to the top of chromosome 9 in *C. japonica* (CJAch09t)

249 and *C. lanceolata* (CLAch09t).

250 The structure of *C. obtusa* chromosome 9 was more complex. The top region

251 (COBch09t) was inverted and translocated to the bottom of *C. japonica* chromosome 2

252 (CJAch02b) and the bottom of *C. lanceolata* chromosome 9 (CLAch09b). Conversely, the

253 bottom region (COBch09b) was translocated to the bottom of *C. japonica* chromosome 9

(CJAch09b) and the bottom of *C. lanceolata* chromosome 2 (CLAch02b). These patterns indicate a reciprocal translocation event involving the bottom regions of chromosomes 2 and 9 between *C. japonica* and *C. lanceolata*.

Synteny analysis (Figure 2A) identified 216 pairwise collinear blocks (17,449 gene pairs) between *C. obtusa* and *C. japonica*. Similarly, 271 blocks (16,918 gene pairs) were identified between *C. japonica* and *C. lanceolata*, and 254 blocks (16,022 gene pairs) were identified between *C. lanceolata* and *C. obtusa*. Finally, the analysis detected tandem gene duplications (Figure 2B), revealing 23 blocks (525 gene pairs) in *C. obtusa*, 37 blocks (924 gene pairs) in *C. japonica*, and a notably higher frequency of 89 blocks (3,149 gene pairs) in *C. lanceolata*.

264

265 *Organelle genome assemblies and gene annotations*

The chloroplast genomes of *C. obtusa*, *C. japonica*, and *C. lanceolata* were assembled as single circular sequences with lengths of 127,726 bp, 131,741 bp, and 135,331 bp, respectively (Figure 3A). The *C. obtusa* chloroplast genome contained 120 genes (83 protein-coding, 33 tRNA, and 4 rRNA). In *C. japonica*, 118 genes were annotated (82 protein-coding, 32 tRNA, and 4 rRNA), while in *C. lanceolata*, 120 genes were annotated (81 protein-coding, 35 tRNA, and 4 rRNA).

The mitochondrial genome of *C. obtusa* was assembled as a single linear sequence of 1,908,789 bp, and that of *C. japonica* as a single linear sequence of 3,743,212 bp (Figure 3B). In contrast, the mitochondrial genome of *C. lanceolata* consisted of six sequences with a total length of 1,993,458 bp, comprising one circular sequence and five linear sequences (Figure 3B). Gene annotation identified 70 genes in the *C. obtusa* mitochondrial genome (33 protein-coding, 26 tRNA, 11 rRNA); 71 genes in *C. japonica* (36 protein-coding, 22 tRNA, 13 rRNA); and 65 genes in *C. lanceolata* (36 protein-coding, 23 tRNA, 6 rRNA).

278

279 Discussion

280 We present chromosome-scale genome assemblies for three species of the Cupressaceae: *C.*
 281 *obtusa* (Japanese cypress), *C. japonica* (Japanese cedar), and *C. lanceolata* (Chinese fir) (Table
 282 1). Consistent with previous genome size estimates (Shirasawa *et al.* 2024), the resulting
 283 assembly length of *C. lanceolata* (13.4 Gb) was significantly larger than those of *C. obtusa* (8.7
 284 Gb) and *C. japonica* (9.6 Gb). Since the proportion of repetitive sequences remained
 285 approximately 80% across all three genomes (Table 4), the larger genome size of *C. lanceolata*
 286 appears to be driven by the multiplication of genomic segments containing both repetitive and
 287 non-repetitive sequences, rather than solely by a burst of transposon activity. This hypothesis is
 288 supported by the higher number of predicted genes in *C. lanceolata* (72,169) compared to *C.*
 289 *obtusa* (54,034) and *C. japonica* (49,899) (Table 2), as well as its elevated duplicated BUSCO
 290 score (34.2%) relative to *C. obtusa* (12.6%) and *C. japonica* (14.5%) (Table 3). These results
 291 suggest that frequent tandem gene duplications (Figure 2B) have contributed to the expansion of
 292 both genome size and gene content in *C. lanceolata*.

293 Comparative genome analysis revealed structural rearrangements in three
 294 chromosomes (2, 6, and 9) between *C. obtusa* and *C. japonica*—a pattern consistent with a
 295 previous study (Dogan *et al.* 2024)—as well as between *C. obtusa* and *C. lanceolata*. In contrast,
 296 rearrangements were observed in only two chromosomes (2 and 9) between *C. japonica* and *C.*
 297 *lanceolata* (Figure 1). Based solely on the number of chromosomal rearrangements, one might
 298 infer a closer relationship between *C. japonica* and *C. lanceolata*. However, phylogenetic
 299 studies (Liu *et al.* 2022) indicate that *Cunninghamia* diverged earlier from the common ancestor
 300 of *Chamaecyparis* and *Cryptomeria*. The chromosomal rearrangements and tandem gene

301 duplications identified in this study (Figures 1 and 2) provide new insights into the complex
302 processes of evolution and speciation in these coniferous trees.

303 In conclusion, this study provides essential genomic resources, including
304 chromosome-scale assemblies, gene annotations, and genetic maps, for *C. obtusa*, *C. japonica*,
305 and *C. lanceolata*. Although one chromosome-scale assembly for *C. japonica* was recently
306 published (Fujino *et al.* 2024), few other genomic resources have been publicly available for
307 these species. The availability of multiple genome assemblies, facilitating pan-genome analysis,
308 is crucial for understanding the genetic mechanisms underlying species diversity (Aoyagi Blue
309 *et al.* 2025). The resources presented here, combined with existing data, will facilitate further
310 research into the evolution and speciation of these conifers and accelerate the identification of
311 key genes for breeding programs.

312

313 **Acknowledgements**

314 We thank Y. Kishida, C. Minami, K. Ozawa, H. Tsuruoka, and A. Watanabe (Kazusa DNA
315 Research Institute) for technical assistance.

316

317 **Funding**

318 This study was supported in part by the MAFF commissioned project study on “Development of
319 efficient breeding technique aiming at forestry trees with superior carbon storage capacity”
320 (Grant Number JPJ009841), JSPS KAKENHI (22H05172 and 22H05181), and the Kazusa
321 DNA Research Institute Foundation.

322

323 **Data availability**

324 The genome assembly files and gene annotation files for the three species are available in

325 Kazusa Genome Atlas (<https://genome.kazusa.or.jp>).

326

327 **Conflicts of interest**

328 All authors declare no competing interests.

329

330 **References**

- 331 Aoyagi Blue, Y., H. Iimura, M.P. Sato and K. Shirasawa (2025) The impact of
332 telomere-to-telomere genome assembly in the plant pan-genomics era. *Breed Sci* 75:
333 3–12.
- 334 Bandi, V. and C. Gutwin (2020) Interactive Exploration of Genomic Conservation. *Proc Graph*
335 *Interface* 74–83.
- 336 Besemer, J. and M. Borodovsky (2005) GeneMark: web software for gene finding in
337 prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* 33: W451–4.
- 338 Brůna, T., K.J. Hoff, A. Lomsadze, M. Stanke and M. Borodovsky (2021) BRAKER2:
339 automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS
340 supported by a protein database. *NAR Genom Bioinform* 3: lqaa108.
- 341 Chao, K.-H., J.M. Heinz, C. Hoh, A. Mao, A. Shumate, M. Pertea and S.L. Salzberg (2025)
342 Combining DNA and protein alignments to improve genome annotation with LiftOn.
343 *Genome Res* 35: 311–325.
- 344 Chase, M.W., M.J.M. Christenhusz, M.F. Fay, J.W. Byng, W.S. Judd, D.E. Soltis, D.J.
345 Mabberley, A.N. Sennikov, P.S. Soltis and P.F. Stevens (2016) An update of the
346 Angiosperm Phylogeny Group classification for the orders and families of flowering
347 plants: APG IV. *Bot J Linn Soc* 181: 1–20.
- 348 Christenhusz, M.J.M., J.L. Reveal, A. Farjon, M.F. Gardner, R.R. Mill and M.W. Chase (2011)
349 A new classification and linear sequence of extant gymnosperms. *Phytotaxa* 19: 55–70.

350 Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G.
351 Lunter, G.T. Marth, S.T. Sherry *et al.* (2011) The variant call format and VCFtools.
352 Bioinformatics 27: 2156–2158.

353 Dogan, G., S. Ueno, Y. Hasegawa, A. Matsumoto, R. Saito, H. Saito, S. Hirayama and Y.
354 Moriguchi (2024) New linkage maps for the Cupressaceae species *Chamaecyparis*
355 *obtusa* and *Thujopsis dolabrata* var. *hondae* and comparison with the linkage map of
356 *Cryptomeria japonica*. J Forest Res 1–9.

357 Enoki, H. and Y. Takeuchi (2018) New Genotyping Technology, GRAS-Di, Using Next
358 Generation Sequencer. Plant & Animal Genome XXVI Poster Abstracts P0153.

359 Forestry Agency (2022) Annual Report on Forest and Forestry in Japan. .

360 Fujino, T., K. Yamaguchi, T.T. Yokoyama, T. Hamanaka, Y. Harazono, H. Kamada, W.
361 Kobayashi, T. Ujino-Ihara, K. Uchiyama, A. Matsumoto *et al.* (2024) A
362 chromosome-level genome assembly of a model conifer plant, the Japanese cedar,
363 *Cryptomeria japonica* D. Don. BMC Genomics 25: 1039.

364 Fujisawa, Y. (2017) The future of forestry and Chinese fir. Forest Genetics and Tree Breeding
365 6: 132–136.

366 Greiner, S., P. Lehwark and R. Bock (2019) OrganellarGenomeDRAW (OGDRAW) version
367 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic
368 Acids Res 47: W59–W64.

369 Hebert, P.D.N., A. Cywinska, S.L. Ball and J.R. deWaard (2003) Biological identifications
370 through DNA barcodes. Proc Biol Sci 270: 313–321.

371 Langmead, B. and S.L. Salzberg (2012) Fast gapped-read alignment with Bowtie 2. Nat
372 Methods 9: 357–359.

373 Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping
374 and population genetical parameter estimation from sequencing data. Bioinformatics 27:
375 2987–2993.

376 Li, J., Y. Ni, Q. Lu, H. Chen and C. Liu (2025) PMGA: A plant mitochondrial genome
377 annotator. Plant Commun 6: 101191.

378 Liu, X.-Q., X.-M. Xia, L. Chen and X.-Q. Wang (2022) Phylogeny and evolution of
379 Cupressaceae: Updates on intergeneric relationships and new insights on ancient
380 intergeneric hybridization. Mol Phylogenet Evol 177: 107606.

381 Mishima, K., T. Hirao, M. Tsubomura, M. Tamura, M. Kurita, M. Nose, S. Hanaoka, M.
382 Takahashi and A. Watanabe (2018) Identification of novel putative causative genes and
383 genetic marker for male sterility in Japanese cedar (*Cryptomeria japonica* D.Don).
384 BMC Genomics 19: 277.

385 Pellicer, J. and I.J. Leitch (2020) The Plant DNA C-values database (release 7.1): an updated
386 online repository of plant genome size data for comparative studies. New Phytol 226:
387 301–305.

388 Rastas, P. (2017) Lep-MAP3: robust linkage mapping even for low-coverage whole genome
389 sequencing data. Bioinformatics 33: 3726–3732.

390 Schmieder, R. and R. Edwards (2011) Quality control and preprocessing of metagenomic
391 datasets. Bioinformatics 27: 863–864.

- 392 Shirasawa, K., D. Harada, H. Hirakawa, S. Isobe and C. Kole (2021) Chromosome-level de
393 novo genome assemblies of over 100 plant species. *Breed Sci* 71: 117–124.
- 394 Shirasawa, K., H. Hirakawa and S. Isobe (2016) Analytical workflow of double-digest
395 restriction site-associated DNA sequencing based on empirical and in silico
396 optimization in tomato. *DNA Res* 23: 145–153.
- 397 Shirasawa, K., K. Mishima, H. Hirakawa, T. Hirao, M. Tsubomura, S. Nagano, T. Iki, S. Isobe
398 and M. Takahashi (2024) Haplotype-resolved *de novo* genome assemblies of four
399 coniferous tree species. *J Forest Res* 29: 151–157.
- 400 Simão, F.A., R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva and E.M. Zdobnov (2015)
401 BUSCO: assessing genome assembly and annotation completeness with single-copy
402 orthologs. *Bioinformatics* 31: 3210–3212.
- 403 Stiehler, F., M. Steinborn, S. Scholz, D. Dey, A.P.M. Weber and A.K. Denton (2021) Helixer:
404 cross-species gene annotation of large eukaryotic genomes using deep learning.
405 *Bioinformatics* 36: 5291–5298.
- 406 Tang, H., X. Zhang, C. Miao, J. Zhang, R. Ming, J.C. Schnable, P.S. Schnable, E. Lyons and J.
407 Lu (2015) ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol*
408 16: 3.
- 409 Tsumura, Y., A. Matsumoto, N. Tani, T. Ujino-Ihara, T. Kado, H. Iwata and K. Uchida (2007)
410 Genetic diversity and the genetic structure of natural populations of *Chamaecyparis*
411 *obtusa*: implications for management and conservation. *Heredity* 99: 161–172.

412 Wang, Y., H. Tang, J.D. Debarry, X. Tan, J. Li, X. Wang, T.-H. Lee, H. Jin, B. Marler, H. Guo
413 *et al.* (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene
414 synteny and collinearity. *Nucleic Acids Res* 40: e49.

415 Wick, R.R., M.B. Schultz, J. Zobel and K.E. Holt (2015) Bandage: interactive visualization of
416 de novo genome assemblies. *Bioinformatics* 31: 3350–3352.

417 Wu, J.-J., Y.-W. Han, C.-F. Lin, J. Cai and Y.-P. Zhao (2023) Benchmarking gene set of
418 gymnosperms for assessing genome and annotation completeness in BUSCO. *Hortic*
419 *Res* 10: uhad165.

420 Zhao, T., A. Zwaenepoel, J.-Y. Xue, S.-M. Kao, Z. Li, M.E. Schranz and Y. Van de Peer (2021)
421 Whole-genome microsynteny-based phylogeny of angiosperms. *Nat Commun* 12: 3498.

422 Zhou, C., M. Brown, M. Blaxter, Darwin Tree of Life Project Consortium, S.A. McCarthy and
423 R. Durbin (2025) Oatk: a de novo assembly tool for complex plant organelle genomes.
424 *Genome Biol* 26: 235.

425

426 **Figure legends**

427 **Figure 1.** Comparative analysis of genome sequences and structures of *C. obtusa*, *C. japonica*,
428 and *C. lanceolata*. Chromosome numbers are indicated on the top (x-axis) and left (y-axis).
429 Genome sizes (Mb) are shown on the bottom (x-axis) and right (y-axis).

430
431 **Figure 2.** Synteny analysis of *C. obtusa*, *C. japonica*, and *C. lanceolata*. (A) Lines connect
432 pairwise collinear blocks between the genomes of *C. obtusa*, *C. japonica*, and *C. lanceolata*. (B)
433 Lines connect collinear blocks within each genome to visualize tandem duplications.

434
435 **Figure 3.** Organelle genomes of *C. obtusa*, *C. japonica*, and *C. lanceolata*. Chloroplast and
436 mitochondrial genome structures are represented by solid lines. Boxes on the genome tracks
437 indicate genes. A key to the gene functions is provided in the bottom-right corner.

438

439 **Table 1.** Statistics of the chromosome-scale genome assemblies for *C. obtusa*, *C. japonica*, and *C. lanceolata*

Chromosome	<i>C. obtusa</i>		<i>C. japonica</i>		<i>C. lanceolata</i>	
	Length (bp)	% of length	Length (bp)	% of length	Length (bp)	% of length
Chr. 1	753,361,174	7.8	808,184,979	7.0	1,071,765,491	6.3
Chr. 2	643,942,463	6.7	787,791,099	6.8	1,098,708,507	6.5
Chr. 3	1,044,836,393	10.8	1,101,135,464	9.5	1,959,509,874	11.6
Chr. 4	815,834,336	8.5	803,702,909	6.9	1,350,266,392	8
Chr. 5	926,357,518	9.6	1,043,635,872	9.0	1,594,571,851	9.4
Chr. 6	607,243,501	6.3	760,652,447	6.5	906,465,489	5.4
Chr. 7	938,369,984	9.7	921,502,449	7.9	1,534,292,120	9.1
Chr. 8	754,287,409	7.8	787,691,810	6.8	922,755,377	5.5
Chr. 9	651,761,957	6.8	828,523,594	7.1	734,130,669	4.3
Chr. 10	888,336,793	9.2	950,971,039	8.2	1,497,158,838	8.8
Chr. 11	722,417,066	7.5	797,504,455	6.9	753,462,524	4.5
Subtotal	8,746,748,594	90.7	9,591,296,117	82.5	13,423,087,132	79.3
Unplaced contigs	899,622,983	9.3	2,035,395,917	17.5	3,497,085,311	20.7
Total	9,646,371,577	100.0	11,626,692,034	100.0	16,920,172,443	100

440

441 **Table 2.** Summary of predicted genes in the genomes of *C. obtusa*, *C. japonica*, and *C. lanceolata*

Chromosome	<i>C. obtusa</i>		<i>C. japonica</i>		<i>C. lanceolata</i>	
	#Genes	% of genes	#Genes	% of genes	#Genes	% of genes
Chr. 1	4,303	8.0	3,325	6.7	4,651	6.4
Chr. 2	4,042	7.5	3,323	6.7	4,808	6.7
Chr. 3	5,928	11.0	5,037	10.1	7,902	10.9
Chr. 4	4,605	8.5	3,503	7.0	5,688	7.9
Chr. 5	5,047	9.3	4,421	8.9	6,207	8.6
Chr. 6	3,337	6.2	2,958	5.9	4,320	6
Chr. 7	5,661	10.5	4,192	8.4	6,612	9.2
Chr. 8	3,829	7.1	2,992	6.0	3,963	5.5
Chr. 9	3,922	7.3	3,241	6.5	3,152	4.4
Chr. 10	4,957	9.2	3,943	7.9	6,135	8.5
Chr. 11	3,760	7.0	3,117	6.2	3,485	4.8
Subtotal	49,391	91.4	40,052	80.3	56,923	78.9
Unplaced contigs	4,643	8.6	9,847	19.7	15,246	21.1
Total	54,034	100.0	49,899	100.0	72,169	100

442

443 **Table 3.** Completeness assessment of gene predictions for *C. obtusa*, *C. japonica*, and *C. lanceolata*

	<i>C. obtusa</i>	<i>C. japonica</i>	<i>C. lanceolata</i>
Complete	90.1%	90.5%	93.8%
Single-copy complete	77.5%	76.0%	59.6%
Duplicated complete	12.6%	14.5%	34.2%
Fragmented	0.7%	0.9%	0.6%
Missing	9.2%	8.6%	5.6%

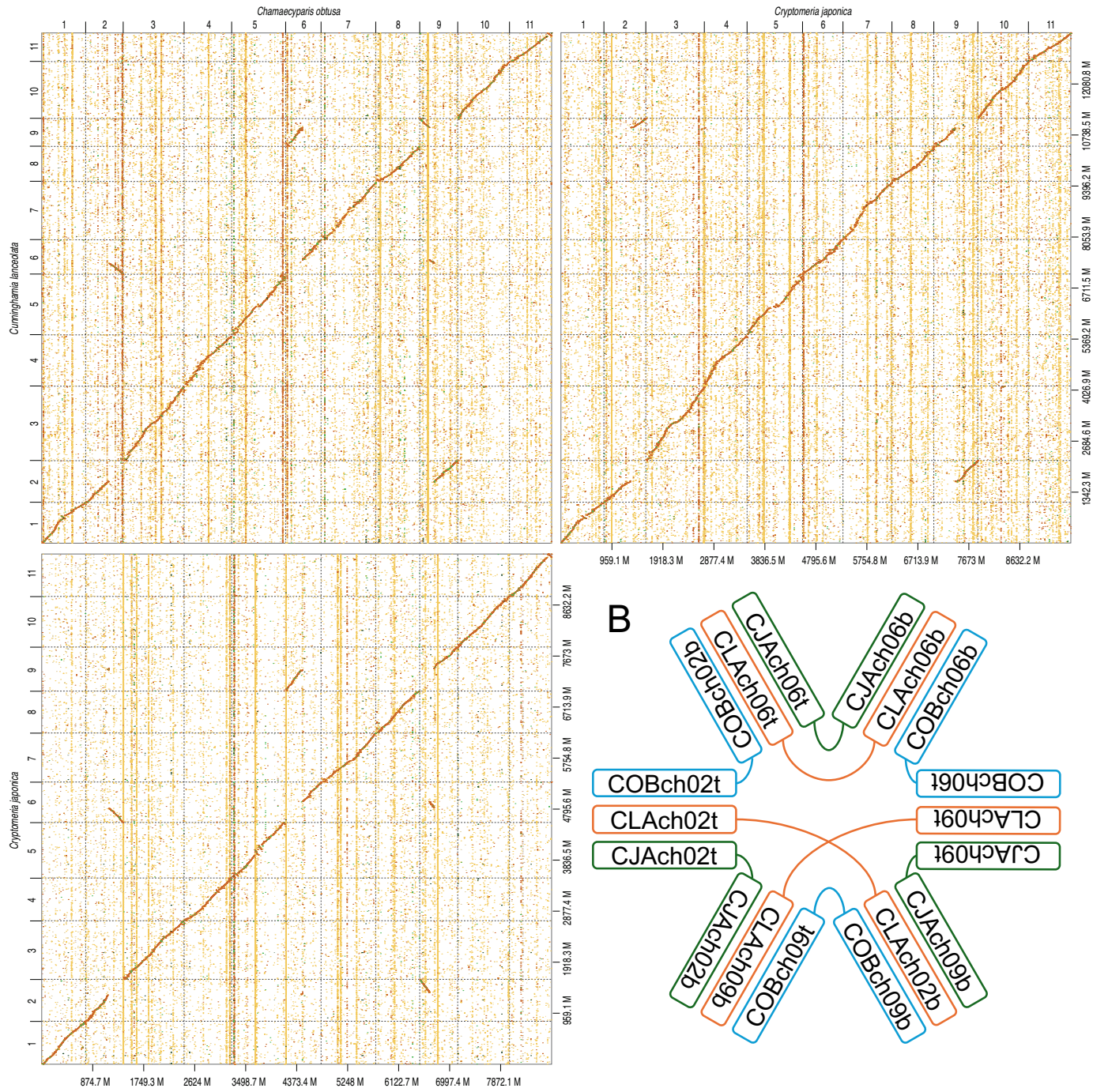
444

445 **Table 4.** Classification and abundance of repetitive sequences in the genomes of *C. obtusa*, *C. japonica*, and *C. lanceolata*

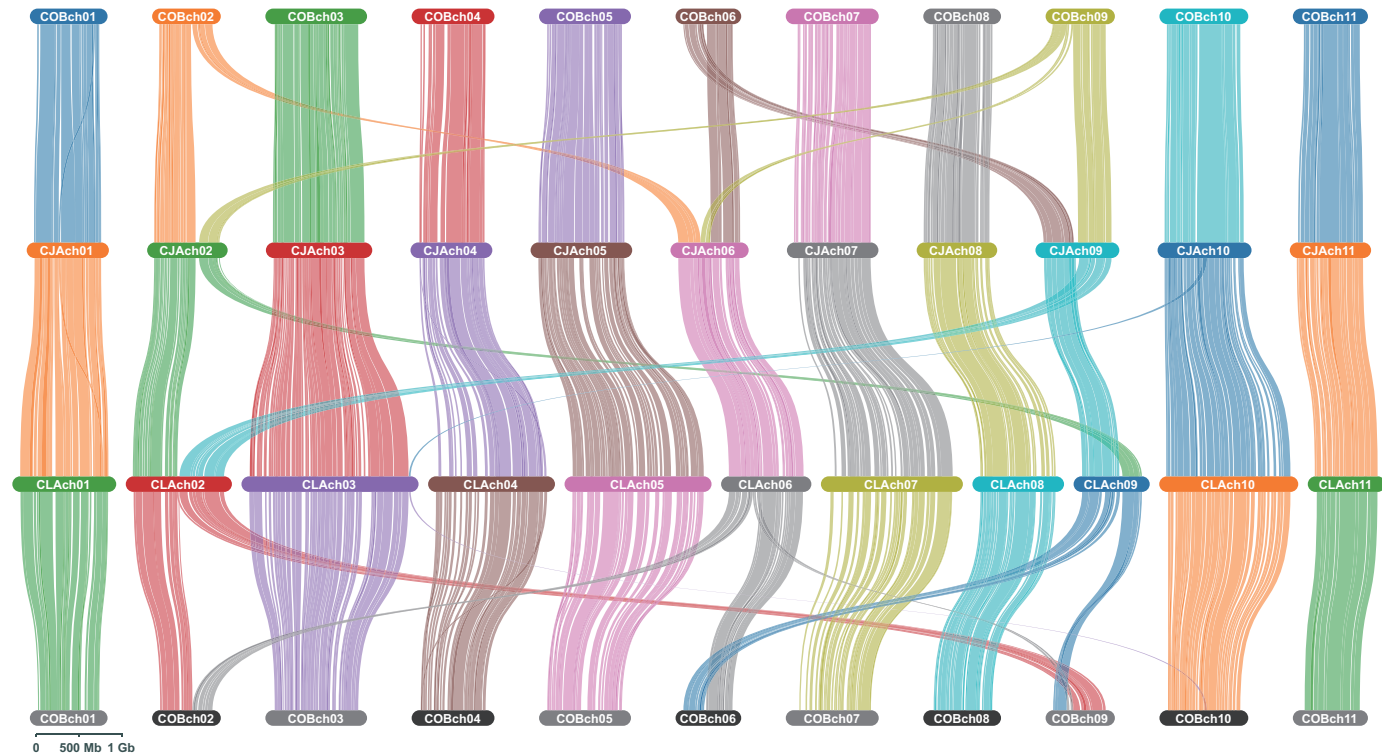
Repeat type	<i>C. obtusa</i>			<i>C. japonica</i>			<i>C. lanceolata</i>		
	#Elements	Length (bp)	%	#Elements	Length (bp)	%	#Elements	Length (bp)	%
SINEs	11,430	1,313,246	0.0	12,828	2,413,789	0.0	50,707	15,072,777	0.1
LINEs	536,344	356,067,530	3.7	647,369	568,704,286	4.9	793,194	558,315,070	3.3
LTR elements	2,900,360	4,680,905,729	48.5	3,606,721	6,235,301,883	53.6	6,005,684	8,319,335,602	49.2
DNA transposons	738,498	505,285,736	5.2	1,137,352	785,356,255	6.8	1,277,378	903,211,596	5.3
Small RNA	9,934	4,679,246	0.1	23,970	9,465,882	0.1	68,538	20,640,162	0.1
Satellites	29,335	26,779,510	0.3	19,979	10,042,049	0.1	75,861	45,335,388	0.3
Simple repeats	764,656	55,374,637	0.6	902,078	50,675,895	0.4	980,278	58,458,100	0.4
Low complexity	125,181	8,185,083	0.1	144,624	8,419,661	0.1	166,010	10,190,191	0.1
Unclassified	3,210,805	1,807,553,850	18.7	3,289,375	1,666,957,226	14.3	8,058,531	3,655,997,225	21.6

446

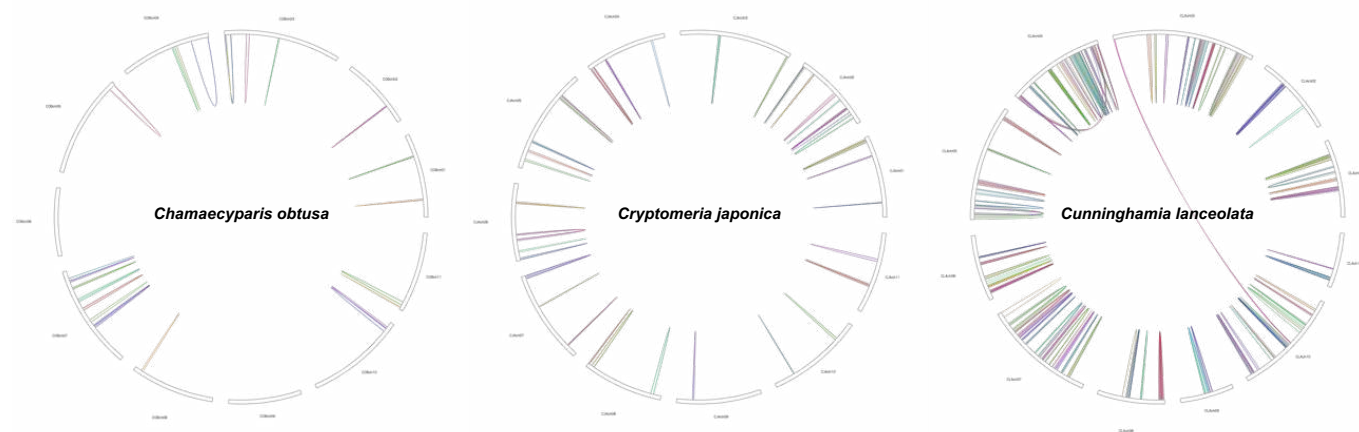
A



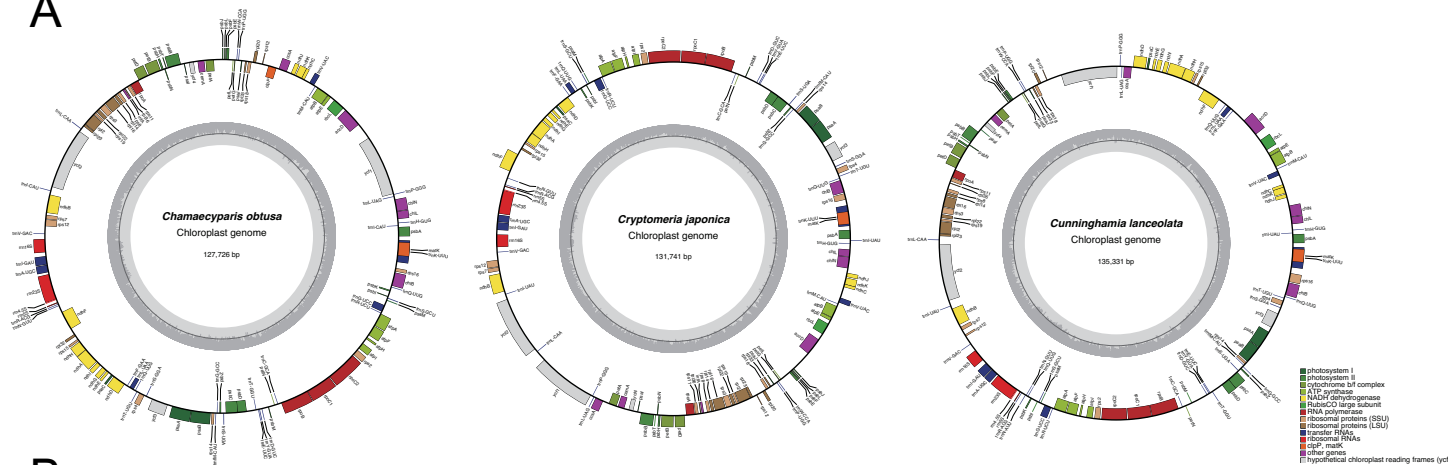
A



B



A



B

