


# Comparative Genome and Evolution Analyses of an Endangered Stony Coral Species *Dendrophyllia cribrosa* Near Dokdo Islands in the East Sea

Jungeun Kim<sup>1</sup>, Jae-Pil Choi<sup>1</sup>, Min Sun Kim<sup>1</sup>, Yejin Jo<sup>2</sup>, Won Gi Min<sup>3</sup>, Seonock Woo<sup>4</sup>, Seungshic Yum<sup>2,5,\*</sup>, and Jong Bhak <sup>1,6,7,8,\*</sup>

<sup>1</sup>Personal Genomics Institute (PGI), Genome Research Foundation (GRF), Cheongju, Republic of Korea

<sup>2</sup>Ecological Risk Research Division, Korea Institute of Ocean Science and Technology (KIOST), Geoje, Republic of Korea

<sup>3</sup>Ulleungdo-Dockdo Ocean Science Station, KIOST, Ulleung, Gyeongbuk, Republic of Korea

<sup>4</sup>Marine Biotechnology Research Center, KIOST, Busan, Republic of Korea

<sup>5</sup>The KIOST School, University of Science and Technology (UST), Geoje, Republic of Korea

<sup>6</sup>Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea

<sup>7</sup>Department of Biomedical Engineering, School of Life Sciences, UNIST, Ulsan, Republic of Korea

<sup>8</sup>Clinomics, Inc., Ulsan, Republic of Korea

\*Corresponding authors: E-mails: syum@kiost.ac.kr; jongbhak@genomics.org.

Accepted: 22 August 2022

## Abstract

Stony corals often harbor intracellular photosynthetic dinoflagellate algae that receive dissolved inorganic nutrients. However, *Dendrophyllia cribrosa* is a nonsymbiotic stony coral distributed in the western Pacific. We assembled a chromosome-level *D. cribrosa* genome using PacBio and Hi-C technologies. The final assembly was 625 Mb, distributed on 14 chromosomes, and contained 30,493 protein-coding genes. The Benchmarking Universal Single-Copy Orthologs analysis revealed a percentage of 96.8 of the metazoan genome. A comparative phylogenetic analysis revealed that *D. cribrosa*, which lacks symbionts, evolved to acquire cellular energy by expanding genes related to acyl-CoA metabolism and carbohydrate transporters. This species also has expanded immune-related genes involved in the receptor protein tyrosine kinase signaling pathway. In addition, we observed a specific expansion of calcification genes, such as *coral acid-rich proteins* and *carbonic anhydrase*, in *D. cribrosa*. This high-quality reference genome and comparative analysis provides insights into the ecology and evolution of nonsymbiotic stony corals.

**Key words:** *Dendrophyllia cribrosa*, comparative genome, comparative evolution, stony coral, chromosome-level assembly.

## Significance

*Dendrophyllia cribrosa* is a nonsymbiotic stony coral. We first provide a chromosome-level genome assembly of *D. cribrosa*, which has a size of 625 Mb forming 14 chromosomes. Our comparative analysis reveals a larger proportion of genes associated with acyl-CoA metabolism and carbohydrate transporters. We also find an expansion of the calcification-related genes. These results provide new insights into the metabolism of these stony corals which lack any symbiont.

## Introduction

*Dendrophyllia cribrosa*, belonging to the scleractinian coral family, is a rare subtropical–temperate coral species that is distributed in the western Pacific. *Dendrophyllia cribrosa* is a stony coral without symbiotic microalgae. In 2016, the Ministry of Oceans and Fisheries of Korea reported a single habitat of a *D. cribrosa* coral community with a width of 5 m and a height of 3 m, at depths of 18–20 m, near the Dokdo Islands in the East Sea. The morphological features of *D. cribrosa* resemble trees with irregular thick branches (fig. 1a), and their coloration ranges from deep yellow to orange. This species was designated as endangered in the “Endangered and Protected Wild Species List in Korea” in 1998 by the Korean Government.

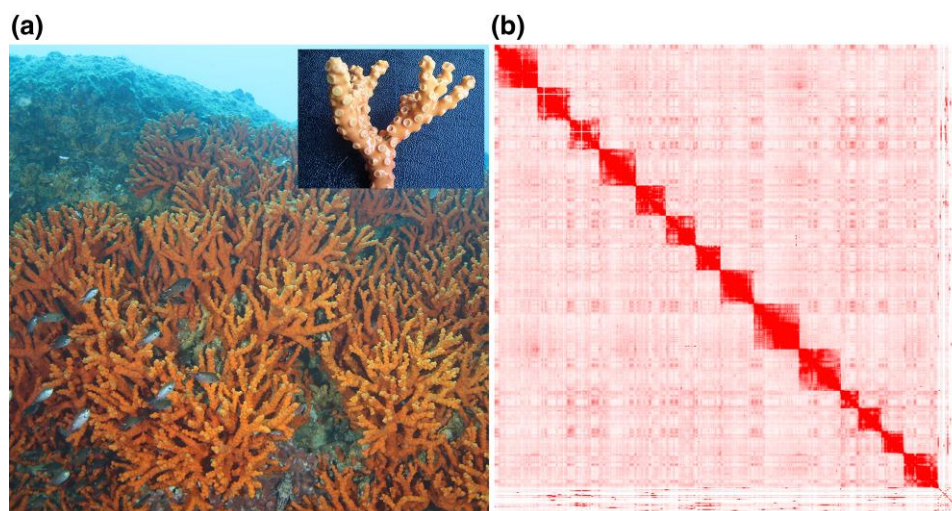
Here, we describe a chromosome-level assembly of *D. cribrosa* from the Dokdo Islands in Korea. The *Dendrophyllia cribrosa* genome provides a comparative study of coral genomes that exhibit evolutionary expansions related to coral calcification, metabolism and immune responses.

## Results and Discussion

### Genome Assembly of *Dendrophyllia cribrosa*

We produced 41 Gb next-generation sequencing (NGS) reads of *D. cribrosa* (supplementary table S1, Supplementary Material online). To estimate the genome size of *D. cribrosa*, we used Jellyfish, programmed with a K-mer range of 17–25. Jellyfish estimated the genome size of *D. cribrosa* to be 610 Mb, at  $K=25$ , with the lowest PCR error rate (0.19) and PCR duplicates (0.92), which is similar to the genome size of other closely related complex corals (*Montipora* spp. 615–

653). We added the GenomeScope result in supplementary figure S1, Supplementary Material online. At  $K=25$ , we estimated the *D. cribrosa* genome size to be 610 Mb with 0.30% heterozygosity in 25 bp of K-mer (supplementary fig. S1, Supplementary Material online). This estimation is similar to the genome size of closely related complex corals (Genus *Montipora*, 615–653 Mb) (Helmkamp et al. 2019). We also produced 120 Gb-long reads (~246-fold coverage of the genome) using a PacBio Sequel2 platform (DNA Link Inc., Seoul, Republic of Korea) with an N50 of 27 kb (supplementary table S2, Supplementary Material online). The FALCON\_unzip assembler constructed 1,174 contigs, with an assembly length of 765 Mb (table 1). After implementing purge haplotigs and error correction, we obtained a 680M assembly from 591 contigs. The Benchmarking Universal Single-Copy Orthologs (BUSCO) assessments showed that the number of “complete and duplicated BUSCO genes” was slightly decreased from 23 (9.0%) to 10 (3.9%) without any change in the total number of complete genes. After polishing the haplotigs, we could not find any changes in BUSCO values (table 1). The N50 of our contigs was 2.1 Mb and L50 was 104 Mb. Using 105 Gb Hi-C reads (~172-fold coverage), we obtained 22 scaffolds with a 627 Mb *D. cribrosa* genome (fig. 1b). Our Hi-C scaffolding resulted in relatively clear chromosomal compartments as shown in supplementary figure S2, Supplementary Material online. However, as denoted in the blue boxes (supplementary fig. S2a, Supplementary Material online), the results included two erroneously generated pseudo-scaffolds. The pseudo-scaffolds comprised seven scaffolds grouped into one gigantic chromosome-scale scaffold and two scaffolds grouped into a comparatively small chromosome-scale scaffold (supplementary fig. S2a, Supplementary Material online).



**Fig. 1.**—*Dendrophyllia cribrosa* close-up image and its chromosome contact map. (a) *Dendrophyllia cribrosa* inhabiting the sea near the Dokdo Islands and its close-up image. (b) Chromosome contact map of *D. cribrosa*.

**Table 1**Statistics of *Dendrophyllia cribrrosa* Assembly

	FALCON	Purge Haplotig	Error Correction (Pilon)	Hi-C Assembly
<b>No. of contigs</b>	1,174	591	591	14
<b>Assembly length</b>	764,923,991	680,856,317	680,577,677	627,238,274
<b>Longest contigs</b>	6,877,747	6,877,747	6,877,747	62,021,193
<b>N50</b>	1,750,786	2,113,998	2,112,715	48,602,881
<b>L50</b>	126	104	104	6
<b>BUSCO<sup>a</sup></b>	C: 98.8% [S: 89.8%, D: 9.0%], F: 0.0%, M: 1.2%, n: 255	C: 98.4% [S: 94.5%, D: 3.9%], F: 0.4%, M: 1.2%, n: 255	C: 98.4% [S: 94.5%, D: 3.9%], F: 0.4%, M: 1.2%, n: 255	C: 93.7% [S: 92.5%, D: 1.2%], F: 0.8%, M: 5.5%, n: 255

<sup>a</sup>BUSCO version: eukaryota\_odb10 (10 September 2020); C, complete BUSCOs; S, complete and single-copy BUSCOs; D, complete and duplicated BUSCOs; F, fragmented BUSCOs; M, missing BUSCOs.

To resolve this, we manually split these pseudo-scaffolds (supplementary fig. S2b, Supplementary Material online) and removed seven contigs to generate the separated scaffolds. Additionally, we did not use any contigs that were <1 kb in the BUSCO assessment, which resulted in a lower score (93.7%). The N50 of the *D. cribrrosa* assembly was 19 Mb, and the maximum assembly length was 62 Mb. Based on the BUSCO assessment score, we measured 93.7% completeness of genes, including 92.5% completeness of 236 single-copy genes and 1.2% completeness of three duplicated BUSCO genes. We found 14 (0.8%) missing genes in 14 pseudo-chromosomes. During the scaffolding, several genes were not integrated in the scaffolds.

Approximately 364 Mb (58.10%) of repeats were found in the *D. cribrrosa* genome (supplementary table S3, Supplementary Material online). This proportion is similar to that of other coral genomes, such as *Trachythela* (57.88%) (Zhou et al. 2021). We predicted 30,493 protein-coding genes in the *D. cribrrosa* genome from these data. They showed a slightly higher number of protein-coding genes compared with other coral genes (supplementary table S4, Supplementary Material online). We conducted BUSCO assessment in the protein-coding genes. It resulted in 231 (90.6%) complete eukaryote genes, with 226 (88.6%) single copy and 5 (2.0%) duplicated in the BUSCO gene set. Among them, 8 (3.1%) were fragmented and 16 (6.3%) were missed. The BUSCO assessment showed a higher number of complete genes in the *D. cribrrosa* genome (supplementary table S4, Supplementary Material online).

## Materials and Methods

### Sample Collections and Genome Sequencing

*Dendrophyllia cribrrosa* colonies were collected at 37° 14.6498' N and 131° 51.6516' E at a depth of 18–20 m using SCUBA diving equipment. The colonies were snap-frozen in liquid nitrogen and stored at –75 °C. Total DNA

was extracted from a colony of *D. cribrrosa* and processed according to a previously described method optimized for marine invertebrates at the Korea Institute of Ocean Science and Technology (KIOST, Geoje, Republic of Korea) (Kim et al. 2019b).

DNA libraries were constructed using a TruSeq Nano HT Sample Preparation Kit (Illumina, San Diego, CA, USA), and paired-end reads were generated on a NovaSeq 6000 (Illumina) according to the manufacturer's instructions (supplementary table S2, Supplementary Material online). We then removed adaptors and low-quality reads ( $Q < 20$ ) using Trimmomatic (ver. 0.64; RRID: SCR\_011848) (Bolger et al. 2014).

A long-read sequence library was constructed using the SMRTbell Express Template Preparation Kit (101-357-000) and sequenced using the PacBio Sequel2 platform. An Arima-Hi-C kit (Arima Genomics Inc., San Diego, CA, USA) was used according to the manufacturer's instructions. The Hi-C library was sequenced using the NovaSeq 6000 platform (Novogene Co. Ltd, CA, USA).

An Illumina RNA library from *D. cribrrosa* was constructed using the Illumina TruSeq Stranded mRNA LT Sample Prep Kit (Illumina, San Diego, CA, USA) and sequenced using the NovaSeq 6000 platform (DNA Link Inc., Seoul, Republic of Korea). Adaptor and low-quality reads ( $Q < 20$ ) were removed using Trimmomatic (ver. 0.39; RRID: SCR\_011848) (Bolger et al. 2014).

### Genome Assembly

Using cleaned Illumina reads, we estimated the genome size of *D. cribrrosa* using Jellyfish (ver. 2.2.4; RRID: SCR\_005491), a tool for fast, memory-efficient counting of K-mers in DNA (Marcais and Kingsford 2011) and GenomeScope (ver. 2; RRID: SCR\_017014) (Ranallo-Benavidez et al. 2020) (supplementary fig. S1, Supplementary Material online). We used the Jellyfish program set with a K-mer range of 17–25 at  $K = 25$ , with the lowest PCR error rate (0.19) and

PCR duplicates (0.92). We have added the GenomeScope result in [supplementary figure S1, Supplementary Material](#) online. We assembled the *D. cribrosa* genome using the FALCON\_unzip assembler (ver. 1.22; RRID: SCR\_016089) with default options and raw reads. We also constructed a deduplicated haploid assembly using the Purge Haplotigs (ver. 1.1.2; RRID: SCR\_017616) (Roach et al. 2018). To polish our assembly, short reads (61x coverage) were aligned to the *D. cribrosa* haplotigs using the Burrows-Wheeler Alignment tool (BWA) (ver. 0.7.17; RRID: SCR\_010910) (Li and Durbin 2009) and possible errors were corrected using Pilon (ver. 1.23; RRID: SCR\_014731) (Walker et al. 2014). Using 172x Hi-C reads, scaffolding was conducted with Juicer (ver. 1.6; RRID: SCR\_017226) and the 3D-DNA pipeline. A total of 14 pseudo-chromosomes were constructed after a manual curation of the assembly using Juicebox Assembly Tools (ver. 1.13.01; RRID: SCR\_021172).

To estimate the number of repetitive sequences in the *D. cribrosa* genome, we built a custom-repeat library using RepeatModeler2 (RRID: SCR\_015027) (Flynn et al. 2020) and predicted repeats using RepeatMasker (ver. 4.1.0; RRID: SCR\_012954) ([supplementary table S3, Supplementary Material](#) online). To estimate the number of protein-coding genes in *D. cribrosa*, we assembled the RNA-seq data using Trinity (ver. 2.10.0; RRID: SCR\_013048) (Haas et al. 2013) and aligned the transcript assembly with GMAP (downloaded on February 22, 2021; RRID: SCR\_008992) (Wu and Watanabe 2005). We also aligned the RNA-seq data to repeat the masked assembly using HISAT2 (ver. 2.2.1) (Kim et al. 2019a). We applied the Gaius-Augustus/BRAKER pipeline (ver. 2.1.5; RRID: SCR\_018964) (Lomsadze et al. 2014; Hoff et al. 2016; Bruna et al. 2021) and assembled the transcripts from RNA-seq data using Trinity (ver. 2.10.0; RRID: SCR\_013048) (Haas et al. 2013). We used *Acropora millepora* (Ying et al. 2019) and *Acropora acuminata* (Shinzato et al. 2020) protein sequences for alignment by using exonerate (RRID: SCR\_016088) (Slater and Birney 2005). These two genomes showed higher BUSCO values of 96.0% and 93.8%, respectively. For de novo gene prediction, we used AUGUSTUS (ver. 3.4.0; RRID: SCR\_008417), which was trained with RNA-seq data with default options. Finally, we predicted protein-coding genes by integrating evidence sequences with the EVIDENCEModeler (ver. 1.1.1; RRID: SCR\_014659) (Haas et al. 2008).

### Evolutionary Study of the Coral Genomes

We collected nine coral genomes from public databases, and of these, three were soft corals and six were hard corals, and used a sponge genome (*Amphimedon queenslandica*) as the outgroup. Orthologous relationships were defined using OrthoMCL (Ver. 2.0.9; RRID: SCR\_007839) (Li et al. 2003). We aligned one-to-one orthologs using MUSCLE (ver. 3.8.31; RRID: SCR\_011812) (Edgar 2004)

and eliminated ambiguously aligned regions using Gblocks (ver.0.9.1; RRID: SCR\_015945) (Talavera and Castresana 2007). A phylogenetic tree was constructed using RAxML software (ver. 8.2.12; RRID: SCR\_006086) (Stamatakis 2014) and employing the PROTGMMMAUTO model with an outgroup of the sponge genome. We estimated the divergence times using the MCMCtree (ver. 4.9) based on fossil calibration times. We used the café algorithm (ver. 4.2.1; RRID: SCR\_018924) (Han et al. 2013) to estimate gene expansion and contraction throughout the coral evolution.

### Supplementary Material

[Supplementary data](#) are available at *Genome Biology and Evolution* online.

### Acknowledgments

This research was supported by the Collaborative Genome Program (No. 2018043012) and a sustainable research and development of Dokdo, both of which were funded by the Ministry of Oceans and Fisheries.

### Author Contributions

S.Y. and J.B. designed and supervised this project, and Y.J., W.G.M., and S.Y. provided samples. J.K., J.-P.C., and M.S.K. conducted the bioinformatics data processing and analyses. J.K., S.Y., and J.B. wrote and revised the manuscript. All authors read and approved the final manuscript.

### Data Availability

All sequences generated in this study, including PacBio long reads and Illumina short reads, were deposited in the NCBI under BioProject PRJNA782406. The genome assembly and annotation files are available from the Marine Genome Information Center (<http://www.magic.re.kr/assembly/MA00395>) and the NCBI assembly accession GCA\_024195265.1 ([https://www.ncbi.nlm.nih.gov/assembly/GCA\\_024195265.1](https://www.ncbi.nlm.nih.gov/assembly/GCA_024195265.1)). The gene annotation files, including the functional gene and repeat gene, are available in the FigShare repository (<https://doi.org/10.6084/m9.figshare.20400987.v1>).

### Literature Cited

- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.
- Bruna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform.* 3:lqaa108.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.



- Flynn JM, et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*. 117: 9451–9457.
- Haas BJ, et al. 2008. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol*. 9:R7.
- Haas BJ, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 8:1494–1512.
- Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol*. 30:1987–1997.
- Helmkamp M, Bellinger MR, Geib SM, Sim SB, Takabayashi M. 2019. Draft genome of the rice coral *Montipora capitata* obtained from linked-read sequencing. *Genome Biol Evol*. 11:2045–2054.
- Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 32:767–769.
- Kim HM, et al. 2019b. The genome of the giant Nomura's jellyfish sheds light on the early evolution of active predation. *BMC Biol*. 17:28.
- Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019a. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 37:907–915.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
- Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13:2178–2189.
- Lomsadze A, Burns PD, Borodovsky M. 2014. Integration of mapped RNA-seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res*. 42:e119.
- Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27:764–770.
- Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 11:1432.
- Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* 19:460.
- Shinzato C, et al. 2020. Eighteen coral genomes reveal the evolutionary origin of *Acropora* strategies to accommodate environmental changes. *Mol Biol Evol*. 38:16–30.
- Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
- Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*. 56:564–577.
- Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963.
- Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21:1859–1875.
- Ying H, et al. 2019. The whole-genome sequence of the coral *Acropora millepora*. *Genome Biol Evol*. 11:1374–1379.
- Zhou Y, et al. 2021. The first draft genome of a cold-water coral *Trachythela* sp. (Alcyonacea: Stoloniifera: Clavulariidae). *Genome Biol Evol*. 13:evaa265.

Associate editor: Christopher Wheat