Genome assembly of the deep-sea coral *Lophelia pertusa*

Santiago Herrera[1*], Erik E. Cordes[2]

[1]Department of Biological Sciences, Lehigh University, Bethlehem, PA USA

[2]Biology Department, Temple University, Philadelphia, PA USA

*santiago.herrera@lehigh.edu

Abstract

Like their shallow-water counterparts, cold-water corals create reefs that support highly diverse communities, and these structures are subject to numerous anthropogenic threats. Here, we present the genome assembly of *Lophelia pertusa* from the southeastern coast of the USA, the first one for a deep-sea scleractinian coral species. We generated PacBio CLR data for an initial assembly and proximity ligation data for scaffolding. The assembly was annotated using evidence from transcripts, proteins, and *ab initio* gene model predictions. This assembly is comparable to high-quality reference genomes from shallow-water scleractinian corals. The assembly comprises 2,858 scaffolds (N50 1.6 Mbp) and has a size of 556.9 Mbp. Approximately 57% of the genome comprises repetitive elements and 34% of coding DNA. We predicted 41,089 genes, including 91.1% of complete metazoan orthologs. This assembly will facilitate investigations into the ecology of this species and the evolution of deep-sea corals.

Keywords

Scleractinia, cold water, azooxanthellate, stony coral, PacBio

Data Description

Context

Stony corals (Order Scleractinia) are foundational species in marine seafloor ecosystems worldwide. Due to their ecological importance, more than 40 whole genome assemblies of shallow-water scleractinian corals have been published to date [1–3]. Although most commonly associated with warm, shallow, tropical reefs, scleractinian stony corals are at least as diverse in cold water, particularly below the sunlit surface ocean (i.e., deeper than 50 meters below sea level or mbsl) [4]. However, no genome assemblies for deep-sea or cold-water scleractinian corals have been available previously.

Cold-water coral reefs support highly diverse communities comprising faunal biomass orders of magnitude above the surrounding seafloor [5–7]. In addition to this tightly-associated community, cold-water corals may also serve as important breeding, nursery, and feeding areas for a multitude of fishes and invertebrates [8,9]). These communities rely on the transport of surface productivity to depth because of the lack of photosynthetic symbionts in the corals. Like their shallow-water counterparts, deep-sea corals are subject to ongoing anthropogenic threats, from ocean warming and acidification [10] to oil pollution [11]. Among deep-sea corals, *Lophelia pertusa* (Linnaeus, 1758), also known as *Desmophyllum pertusum* (NCBI:txid174260) [12], is one of the most ecologically important species. *Lophelia pertusa* is a scleractinian coral that builds reef structures (**Fig. 1**). This coral has a nearly-cosmopolitan distribution, spanning from approximately 80 mbsl off the coast of Norway to over 1000 mbsl on the Mid-Atlantic Ridge. Although *L. pertusa* is arguably the best-studied deep-sea coral species, a high-quality reference

genome assembly has yet to be available. This hinders our understanding of the biology of this coral species, its ecological functions, and capacity to survive anthropogenic threats.

Here, we present the genome assembly of *Lophelia pertusa,* the first one for a deep-sea scleractinian coral species. Only one other published genomic-level DNA sequence dataset exists for *D. pertusum*. Emblem and collaborators [13] produced 73 million SOLIiD ligation sequencing reads and 1.2 million 454 pyrosequencing reads with average lengths of 46 bp and 580 bp, respectively. The Emblem dataset was useful for detecting mitochondrial single nucleotide polymorphisms but needed higher coverage and more cohesive to produce a useful genome assembly. Our study used PacBio CLR data for the initial assembly, followed by proximity ligation data for scaffolding and RNA-seq data for annotation. Our approach yielded a genome assembly of comparable quality to those obtained from shallow-water scleractinian corals [14–17].

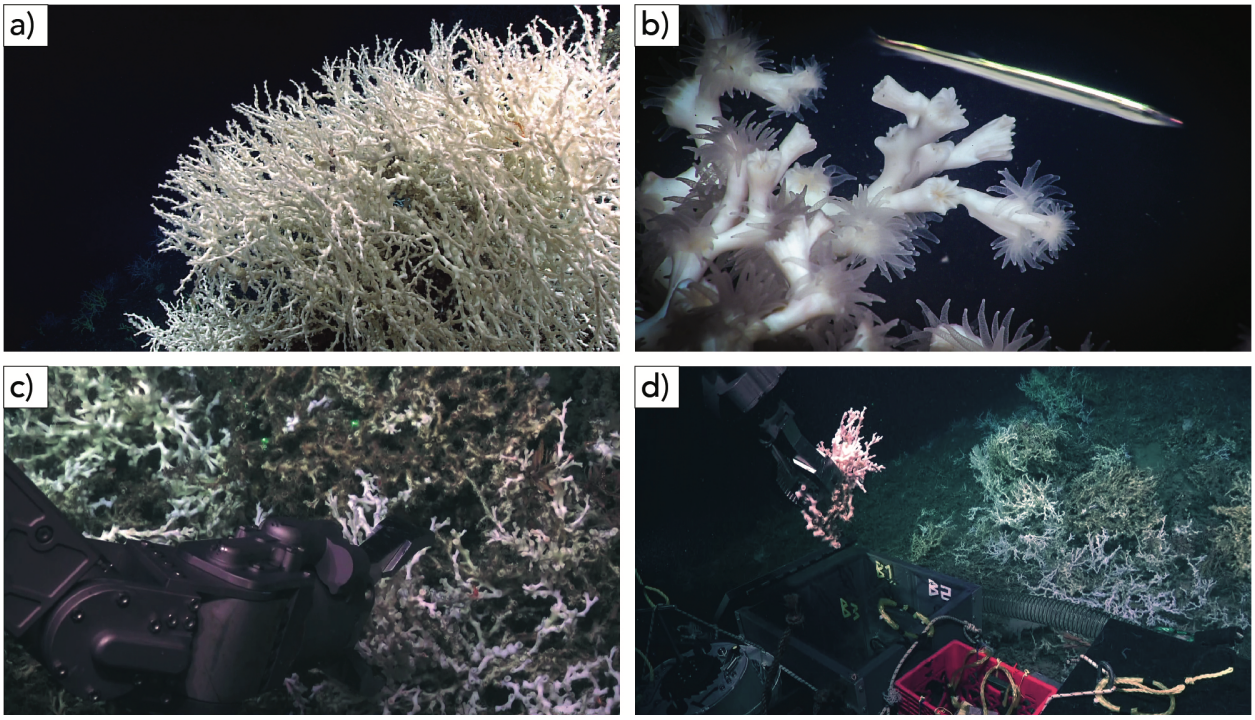

Figure 1. *In situ* images of the coral *Lophelia pertusa* in the Atlantic U.S. southeast shelf. (a) *Lophelia* reef. (b) Close-up of *Lophelia* polyps. (c) Collection of *Lophelia* sample sequenced in this study using the hydraulic arm of ROV Jason. (d) *Lophelia* sample being placed in ROV Jason's biobox. Images (a) and (b) courtesy of NOAA OER, Windows to the Deep 2019. Images (c) and (d) courtesy of the Deep SEARCH program and copyright Woods Hole Oceanographic Institution (CC BY).

Methods

Sample collection

Branches of *Lophelia pertusa* were obtained from the Savannah Banks site, off the southeastern coast of the continental USA, Atlantic Ocean (latitude 31.75420, longitude -79.19442, depth 515 mbsl), while aboard the NOAA Ship *Ronald Brown* (expedition RB1903) using ROV *Jason* (Dive 1130) on April 17, 2019 (BioSample accession SAMN31822850). The branches were collected using a hydraulic robotic arm and stored in an insulted bio-box until they reached the surface (**Figs. 1c-d**). Once onboard the ship, they were immersed in cold RNALater (Thermo Fisher), left to soak in the refrigerator (4°C) for

24 hours, and then frozen at -80°C. Samples remained at that temperature until DNA was purified back in the laboratory.

DNA purification

Polyp tissue was scraped from the skeleton and digested in 2% Cetyltrimethyl Ammonium Bromide (CTAB) buffer with 0.5% β-mercaptoethanol for 15 minutes at 68°C. The DNA was purified through two rounds of phenol: chloroform: isoamyl alcohol (25:24:1) and one round of chloroform: isoamyl alcohol (24:1) mixing and partitioning through centrifugation at 10,000 rpm for 10 minutes. The DNA was precipitated out of the solution with 100% isopropanol. The resulting pellet was washed with 70% ethanol, then air-dried and resuspended in Qiagen G2 buffer. DNA concentration was quantified using a Qubit fluorometer (Invitrogen). The DNA was further purified using the Blood & Cell Culture DNA Midi Kit (Qiagen kit #13343) following the manufacturer's protocol after one hour of protease digestion. The average DNA fragment size was determined using pulsed-field gel electrophoresis (PFGE).

DNA sequencing

A total of 19.3 Gbp contained in 2.07 million continuous long reads (CLR) were generated using a PacBio Sequel sequencer. For this, a 20 kb PacBio SMRTbell library was constructed using Blue Pippin Size selection. Long-insert chromosome conformation capture Chicago [18] and Hi-C [19] libraries (one each) were constructed and sequenced on an Illumina Hiseq X sequencer (PE 150bp), yielding 46.7 Gbp (156 million pairs) for the Chicago library and 72.6 Gbp (242 million pairs) for the Hi-C library.

De novo genome assembly

The analytical pipeline to generate the *de novo* assembly of *Lophelia pertusa* is depicted in Fig. 2. *De novo* genome assembly of PacBio data was performed using the assemblers flye v2.9 [20], wtdbg2 v2.5 [21], and FALCON [22], in combination with the polishing tools NextPolish v1.3.1 [23] and Arrow as implemented in the Pacific Biosciences GenomicConsensus package (https://github.com/PacificBiosciences/GenomicConsensus), and the haplotig and contig overlap removal program purge_dups v.1.2.3 [24]. First, we generated an assembly with flye using default parameters, followed by purging with purge_dups and polishing with NextPolish (assembly A). Using default parameters, we generated a second assembly with wtdbg2 and polished it with NextPolish (assembly B). A third assembly was generated using FALCON, followed by polishing with Arrow and purging with purge_dups (assembly C). Assemblies A and B were combined by aligning the flye assembly against the wtdbg2 assembly using MUMmer v4.0 [25] followed by merging with Quickmerge v0.3 [26] (-hco 5.0 -c 1.5 -l 248998 -ml 5000). The resulting assembly was polished with NextPolish (assembly D). Assembly D was aligned against assembly C using MUMmer and merged with Quickmerge. Finally, the resulting merged assembly between C and D was polished with NextPolish and purged with purge_dups (assembly E). Assemblies generated with other programs were not included because they had lower assembly contiguity or completeness (see Data validation and quality control section, Appendix 1).

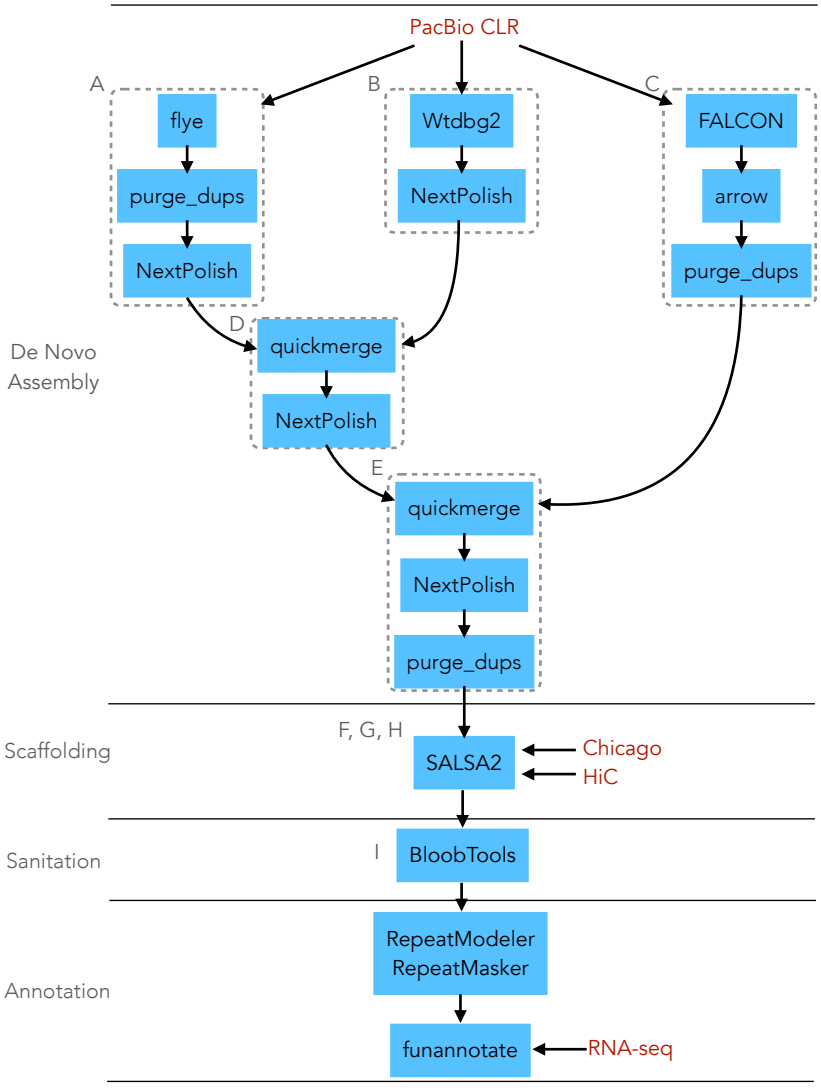

Figure 2. Flow chart depicting the assembly pipeline for the *Lophelia pertusa* genome. Dotted boxes indicate the different *de novo* assemblies. Letters indicate the designed nomenclature of each assembly as reflected in the text and Appendix 1. Data inputs are indicated in maroon font. Software packages are highlighted with blue background.

Scaffolding

Assembly E was scaffolded with long-insert Chicago, and Hi-C reads following the Arima Genomics mapping pipeline A160156 v02 (retrieved from https://github.com/ArimaGenomics/mapping_pipeline). First, the reads from the Chicago library were aligned to assembly E using the MEM algorithm of the program BWA v0.7.17 [27]. Chicago and Hi-C sequence data had mapping rates to the assembly of 96% and 98%, indicating high quality. Chimeric reads that mapped in the 3' direction were excluded using the filter_five_end.pl script. Reads were combined into pairs with the two_read_bam_combiner.pl script and sorted using Samtools v.1.10 [28]. The program Picard tools v2.26.6 (https://broadinstitute.github.io/picard/) was used to add read groups to the resulting bam file and remove PCR duplicates. The program SALSA2 v2.2 [29,30] (-e GATC -m yes)

was used for scaffolding assembly E with the mapped Chicago reads (assembly F). The Hi-C reads they were mapped to assembly H using the same procedure described above and re-scaffolded with SALSA2 (assembly H).

Sanitation

The program BloobToolsKit v2.2 [31] was used to identify non-target scaffolds from assembly H. First; scaffolds were queried against the nucleotide collection database (nt) from the National Center for Biotechnology Information (NCBI), retrieved on May 5, 2020, using NCBI BLAST+ blastn v2.10 [32]. Scaffolds were then queried against the UniProt protein sequence database [33], retrieved on May 5, 2020, using DIAMOND blastx vv0.9.14.115 [34]. Assembly coverage evenness was assessed by mapping the raw PacBio reads against assembly H using minimap2 v2.24-r1122 [35]. We excluded 6 scaffolds with significant matches to non-eukaryotic sequences (i.e., bacteria and viruses). We also excluded 4,531 scaffolds with significant deviations in coverage (<x0.01, >x65) or GC content (<26%, >52.5%) relative to the assembly-wide means (coverage = 3.27x, GC content = 39.81%) (assembly I). This Whole Genome Shotgun (WGS) project was deposited at DDBJ/ENA/GenBank under the accession JAPMOT000000000.

Annotation

Repetitive elements in the genome assembly I were identified *de novo* with the RepeatModeler v2.0.2 package, including the programs RECON v1.05 [36] and RepeatScout v1.06 [37]. Repetitive elements were classified using RepeatClassifier v 2.0.2 and soft-masked using RepeatMasker v4.1.2 [38]. This procedure resulted in 57.37% of the genome assembly being masked.

The masked genome assembly was used for functional annotation using the Funannotate v1.8.9 pipeline [39]. First, we performed a *de novo* genome-guided transcriptome assembly using the Funnannotate *train* script with the *Lophelia pertusa* RNA-seq data published by Glazier and colleagues [40]. In short, (1) The RNA-seq data reads were normalized with Trinity v2.8.5 [41] and mapped to the masked genome assembly using HISAT2 v2.2.1 [42]; (2) A transcriptome assembly was generated with these mapped reads using Trinity; (3) the PASA v2.4.1 [43] program was used to produce a likely set of protein-coding genes based on transcript alignments.

Second, we performed gene prediction using the Funnannotate *predict* script (--repeats2evm --max_intronlen 30000 --busco_db metazoa). With this script, we (1) Parsed transcripts alignments to the genome to use as transcript evidence; (2) Aligned the UniProtKB/SwissProt v2021_04 curated protein database [44] to the genomes and parsed alignments to use as protein evidence; (3) Generated *ab initio* gene model predictions from the masked assembly with GeneMark-ES/ET v4.68 [45,46], Augustus v3.3.3 [47], SNAP v2013_11_29 [48], and GlimmerHMM v3.0.4 [49], using PASA gene modes for training; (4) Computed a weighted consensus of gene models from transcript, protein, and *ab initio* evidence using EVidenceModeler v.1.1.1 [50] (evidence source/weight: transcript/1, protein/1, Augustus/1, Augustus HiQ/2, GeneMark/1, GlimmerHMM/1, PASA/6, Snap/1); (5) Filtered gene models to exclude transposable elements and lengths <50 aa; (6) Predictd tRNAs using tRNAscan-SE v2.0.9 [51]. In total, 37,945 coding genes and 3,144 tRNA genes were predicted in the genome assembly. The average gene length was 4,972 bp. This analysis indicates that approximately 34% of the *Lophelia pertusa* genome is coding DNA.

The protein products of the predicted coding gene models were functionally annotated using the Funnannotate *annotate* script. The following annotations were added: (1) Protein family domains from PFAM v35.0 using HMMer v3.3.2 to find sequence homologs [52]; (2) Gene and product names from UniProt DB v2021_04 using DIAMOND blastp v2.0.13 [53] alignments; (3) Orthologous groups, gene, and product names from EggNog v5.0 [54] using EggNOG-mapper v2.1.6 [55]; (4) Protease annotation

from MEROPS v12.0 [56] using Diamond blastp; (5) Metazonan single-copy orthologs from the OrthoDB v10 [57] using BUSCO v5 [58]; and (6) protein families and gene ontology(GO) terms from InterPro v87 using InterProScan v5.53 [59]. This procedure yielded 24,665 EggNog annotations, 24,471 InterPro annotations, 16,020 PFAM annotations, 16,646 GO terms, and 1,086 MEROPS annotations.

Quality control

The quality of each assembly was assessed using Quast v5.0.2 [60] and BUSCO v5 [58] (genome analysis with the metazoan lineage orthologs dataset OrthoDB v10 [57]). The steps described in the *de novo* assembly and scaffolding pipelines were implemented to maximize the contiguity, measured by the N50 statistic, and completeness, measured by the percentage of single-copy metazoan orthologs present, in the assembly. The final assembly, I, had an N50 of 1.61 Mbp, 5 to 10 times greater than the N50 of initial *de novo* assemblies without merging or scaffolding (assemblies A, B, and C). Similarly, assembly I had 89% complete single-copy metazoan orthologs of the 954 surveyed, which was between 7% and 18% more than initial *de novo* assemblies. Quality metrics for the final assembly (I) are shown in Fig. 3. Quality metrics for all intermediate assemblies (A-H) are shown in Appendix 1.

	Genome Size (Mbp)	Scaffolds	N50 (Mbp)	GC%	Genes	BUSCO% (n=954) S	D	F	M
Lophelia pertusa	556.9	2,858	1.6	39.4	41,089	89.0	2.1	2.1	6.8
Orbicella faveolata	485.5	1,933	1.2	28.6	30,178	84.7	0.4	8.7	6.2
Stylophora pistillata	400.1	5,688	0.5	34.5	28,912	91.6	0.9	4.6	2.9
Pocillopora damicornis	234.4	4,393	0.3	36.4	23,077	90.5	0.2	4.2	5.1
Porites lutea	552.0	2,975	0.7	35.7	31,126	93.5	1.8	2.1	2.6
Acropora millepora	475.4	854	19.8	39.1	42,775	86.2	5.7	3.2	4.9

Figure 3. Quality metrics for the final *Lophelia pertusa* genome assembly (I), compared to other reference genome assemblies of scleractinian corals. BUSCO percentages indicate the proportion of the 954 metazoan orthologs that are complete and single-copy (S), complete and duplicated (D), fragmented (F), and missing (M). The phylogeny shown on the left is the best-scoring maximum likelihood tree inferred from single-copy orthologs. All branches had 100% bootstrap confidence.

The quality of genome assembly I is comparable to those obtained from shallow-water scleractinian corals. For comparison, we retrieved available genome assemblies of scleractinian corals with RefSeq annotations from NCBI's Genome database. This genome set comprised assemblies for the species *Orbicella faveolata* [14], *Stylophora pistillata* [17], *Pocillopora damicornis* [15], and *Acropora millepora* [16]. We also retrieved the genome assembly of *Porites lutea* from reefgenomics.org. The quality of each of these assemblies was assessed using Quast and BUSCO as described above. The *Lophelia pertusa* assembly I has greater contiguity (N50) than most of the other scleractinian genomes in our comparison (0.3-1.2 Mbp), except for *A. millepora* (19.8 Mbp). The completeness of the *Lophelia pertusa* assembly I (91.1% complete metazoan orthologs, including single-copy and duplicated) is similar to the other scleractinian genomes (85.1-95.3%). The assembly size and the number of predicted genes of *Lophelia pertusa* (556.9 Mbp and 41,089 genes) are also similar, although larger than the other scleractinian genomes (234.4-552.0 Mbp and 20,267-31,834 genes). In our comparison, we used 242 single-copy orthologs present in all species to infer phylogenetic relationships among them. The amino-acid sequences of these orthologs

were aligned using MAFFT v7.453 [61] and concatenated for each species (the final concatenated alignment contained 16,619 amino-acid sites). A species phylogeny was inferred in RAxML v8.2.12 [62] using the GAMMA model of rate heterogeneity. Branch support values were estimated through 500 rapid bootstrap replicates. The resulting tree topology is congruent with the most recent phylogeny for the group [63].

Re-use potential

The assembly of the *Lophelia pertusa* genome will facilitate numerous investigations into the ecology and evolution of this important species. This reference resource will enable population-genomic studies of this species within the U.S. exclusive economic zone and comparative studies with populations throughout the Atlantic Ocean, Gulf of Mexico, Caribbean Sea, and Mediterranean Sea. This genome assembly will also be instrumental in resolving the taxonomic position of *Lophelia pertusa* as a monotypic genus instead of its proposed placement as a species, or set of species, within the genus *Desmophyllum*. This annotated genome assembly is the first one for a deep-sea scleractinian coral and thus will provide insights into the evolutionary history of deep-sea corals and the genomic adaptations to the deep-sea environment.

Data Availability

The sequence data and metadata supporting the results of this article are available at the U.S. National Library of Medicine, National Center for Biotechnology Information (NCBI) under BioProject accession PRJNA903949, BioSample accession SAMN31822850, WGS accession JAPMOT000000000, and SRA accessions SRR22387542 (Hi-C reads), SRR22387543 (Chicago reads), and SRR22387544 (PacBio reads). The RNA-seq data is available under BioProject accession PRJNA922177. A voucher of the *Lophelia pertusa* specimen sequenced in this study is available at the Smithsonian Institution National Museum of Natural History under accession number USNM 1676648.

List of abbreviations

Ethics approval and consent to participate

Not applicable

Consent for publication

Approved for publication by the Bureau of Ocean Energy Management.

Competing interests

The author(s) declare that they have no competing interests.

Funding

Sample collection was achieved through the Deep SEARCH project, funded by the Bureau of Ocean Energy Management (contract M17PC00009 to TDI Brooks International) and the NOAA Office of Ocean Exploration and Research (for ship time). Additional support came from the NOAA Deep-Sea Coral Research and Technology Program. Data generation was supported by an award from the Institute

for Genomics and Evolutionary Medicine (iGEM) of Temple University to EEC. The National Academies of Sciences, Engineering, and Medicine Gulf Research Program, Early-Career Fellowship 2000013668 to SH, supported analysis and writing time.

Authors' contributions

SH and EEC conceptualized the project. EEC acquired and managed the funding, collected the samples, and provided computational resources. SH and EEC generated the data. SH curated the data, performed analyses, generated visualizations, and wrote the original draft. SH and EEC reviewed and edited the manuscript.

Acknowledgments

Alexis Weinnig assisted with the collection of samples. Amanda Glazier assisted with laboratory logistics and the composition of the proposal to iGEM. We thank Andrea Quattrini for the helpful discussions. Thanks to the science parties, captains, and crews of the expedition RB1903 aboard the NOAA Ship Ronald H. Brown.

References

1. Voolstra CR, Quigley KM, Davies SW, Parkinson JE, Peixoto RS, Aranda M, et al.. Consensus guidelines for advancing coral holobiont genome and specimen voucher deposition. *Front Mar Sci*. Frontiers Media SA; 2021; doi: 10.3389/fmars.2021.701784.

2. Stephens TG, Lee J, Jeong Y, Yoon HS, Putnam HM, Majerová E, et al.. High-quality genome assembles from key Hawaiian coral species. *Gigascience*. 2022; doi: 10.1093/gigascience/giac098.

3. Liew YJ, Aranda M, Voolstra CR. Reefgenomics.org - a repository for marine genomics data. *Database* . 2016; doi: 10.1093/database/baw152.

4. Cairns SD. Deep-water corals: an overview with special reference to diversity and distribution of deep-water scleractinian corals. *Bull Mar Sci*. 81:311–222007;

5. Mortensen PB, Hovland M, Brattegard T, Farestveit R. Deep water bioherms of the scleractinian coral *Lophelia pertusa* (L.) at 64° n on the Norwegian shelf: Structure and associated megafauna. *Sarsia*. Taylor & Francis; 80:145–581995;

6. Cordes EE, McGinley MP, Podowski EL, Becker EL, Lessard-Pilon S, Viada ST, et al.. Coral communities of the deep Gulf of Mexico. *Deep Sea Res Part I*. 55:777–872008;

7. Henry L-A, Roberts JM. Biodiversity and ecological composition of macrobenthos on cold-water coral mounds and adjacent off-mound habitat in the bathyal Porcupine Seabight, NE Atlantic. *Deep Sea Res Part I*. 54:654–722007;

8. Fosså JH, Mortensen PB, Furevik DM. The deep-water coral *Lophelia pertusa* in Norwegian waters: distribution and fishery impacts. *Hydrobiologia*. 471:1–122002;

9. Ross SW, Quattrini AM. The fish fauna associated with deep coral banks off the southeastern United States. *Deep Sea Res Part I*. 54:975–10072007;

10. Sweetman AK, Thurber AR, Smith CR, Levin LA, Mora C, Wei C-L, et al.. Major impacts of climate

285 change on deep-sea benthic ecosystems. *Elementa: Science of the Anthropocene*. 5:42017;

286 11. White HK, Hsing P-Y, Cho W, Shank TM, Cordes EE, Quattrini AM, et al.. Impact of the Deepwater
287 Horizon oil spill on a deep-water coral community in the Gulf of Mexico. *Proc Natl Acad Sci U S A*.
288 109:20303–82012;

289 12. Addamo AM, Vertino A, Stolarski J, García-Jiménez R, Taviani M, Machordom A. Merging
290 scleractinian genera: the overwhelming genetic similarity between solitary *Desmophyllum* and colonial
291 *Lophelia*. *BMC Evol Biol*. 16:1082016;

292 13. Emblem A, Karlsen BO, Evertsen J, Miller DJ, Moum T, Johansen SD. Mitogenome polymorphism
293 in a single branch sample revealed by SOLiD deep sequencing of the *Lophelia pertusa* coral genome.
294 *Gene*. 506:344–92012;

295 14. Prada C, Hanna B, Budd AF, Woodley CM, Schmutz J, Grimwood J, et al.. Empty Niches after
296 Extinctions Increase Population Sizes of Modern Corals. *Curr Biol*. 26:3190–42016;

297 15. Cunning R, Bay RA, Gillette P, Baker AC, Traylor-Knowles N. Comparative analysis of the
298 *Pocillopora damicornis* genome highlights role of immune system in coral evolution. *Sci Rep*.
299 8:161342018;

300 16. Fuller ZL, Mocellin VJL, Morris LA, Cantin N, Shepherd J, Sarre L, et al.. Population genetics of the
301 coral *Acropora millepora*: Toward genomic prediction of bleaching. *Science*. 2020; doi:
302 10.1126/science.aba4674.

303 17. Voolstra CR, Li Y, Liew YJ, Baumgarten S, Zoccola D, Flot J-F, et al.. Comparative analysis of the
304 genomes of *Stylophora pistillata* and *Acropora digitifera* provides evidence for extensive differences
305 between species of corals. *Sci Rep*. 7:175832017;

306 18. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al.. Chromosome-scale
307 shotgun assembly using an in vitro method for long-range linkage. *Genome Res*. 26:342–502016;

308 19. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al..
309 Comprehensive mapping of long-range interactions reveals folding principles of the human genome.
310 *Science*. 326:289–932009;

311 20. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs.
312 *Nat Biotechnol*. 37:540–62019;

313 21. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods*. 17:155–82020;

314 22. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al.. Phased diploid
315 genome assembly with single-molecule real-time sequencing. *Nat Methods*. 13:1050–42016;

316 23. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read
317 assembly. *Bioinformatics*. 36:2253–52020;

318 24. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic
319 duplication in primary genome assemblies. *Bioinformatics*. 36:2896–82020;

320 25. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al.. Versatile and open
321 software for comparing large genomes. *Genome Biol*. Springer; 5:R122004;

26. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res*. 44:e1472016;

27. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. Oxford Academic; 25:1754–602009;

28. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al.. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021; doi: 10.1093/gigascience/giab008.

29. Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics*. 18:5272017;

30. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al.. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol*. 15:e10072732019;

31. Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit - Interactive Quality Assessment of Genome Assemblies. *G3* . 10:1361–742020;

32. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al.. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3389–4021997;

33. The UniProt Consortium, Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, et al.. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. Oxford Academic; 49:D480–92021;

34. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 12:59–602015;

35. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 34:3094–1002018;

36. Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*. 12:1269–762002;

37. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. *Bioinformatics*. 21 Suppl 1:i351-82005;

38. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*. Chapter 4:Unit 4.102004;

39. Palmer JM. Funannotate: a fungal genome annotation and comparative genomics pipeline. https://github.com/nextgenusfs/funannotate.

40. Glazier A, Herrera S, Weinnig A, Kurman M, Gómez CE, Cordes E. Regulation of ion transport and energy metabolism enables certain coral genotypes to maintain calcification under experimental ocean acidification. *Mol Ecol*. 29:1657–732020;

41. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al.. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 29:644–522011;

42. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 37:907–152019;

43. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, et al.. Improving the

Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*. 31:5654–662003;

44. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*. 2022; doi: 10.1093/nar/gkac1052.

45. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*. 33:6494–5062005;

46. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res*. 42:e1192014;

47. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*. 19 Suppl 2:ii215-252003;

48. Korf I. Gene finding in novel genomes. *BMC Bioinformatics*. 5:592004;

49. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*. 20:2878–92004;

50. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al.. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*. 9:R72008;

51. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res*. 49:9077–962021;

52. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res*. 41:e1212013;

53. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods*. 18:366–82021;

54. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al.. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*. 47:D309–142019;

55. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol*. 38:5825–92021;

56. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res*. 46:D624–322018;

57. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, et al.. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res*. 47:D807–112019;

58. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol*. 38:4647–542021;

59. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al.. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 30:1236–402014;

60. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 29:1072–52013;

61. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30:772–802013;

62. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30:1312–32014;

63. McFadden CS, Quattrini AM, Brugler MR, Cowman PF, Dueñas LF, Kitahara MV, et al.. Phylogenomics, Origin, and Diversification of Anthozoans (Phylum Cnidaria). *Syst Biol*. 70:635–472021;

405

Appendix 1. Statistics for *Lophelia pertusa* intermediate and final assemmblies.

Assembly ID	A			B		C		D		E			F	G	H	**I**
Input Data	PacBio CLR	PacBio CLR	PacBio CLR	PacBio CLR	PacBio CLR	PacBio CLR	PacBio CLR	PacBio CLR	PacBio CLR	PacBio CLR	PacBio CLR	PacBio CLR	PacBio CLR + Chicago	PacBio CLR + HiC	PacBio CLR + Chicago + HiC	H
Software	flye	flye + purge_dups	flye + purge_dups+ NextPolish	wtdbg2	wtdbg2 + NextPolish	FALCON + arrow	FALCON + arrow + purge_dups	quickmerge (A + B)	quickmerge (A + B) + Next polish	quickmerge (D + C)	quickmerge (D+C) + Next polish	quickmerge (D+C) + Next polish+ purge_dups	SALSA2 (E + Chicago)	SALSA2 (E + HiC)	SALSA2 (F + HiC)	BlobToolKit
Sanitation															Prokaryot.	GC, cov., no-hit, undef
# contigs	17,029	13,865	13,865	7,345	7,345	8,987	6,321	11,237	11,237	10,226	10,226	10,011	7,818	8,712	7,385	**2,858**
# contigs (>= 10 Kbp)	11,441	8,278	8,514	5,789	5,863	8,710	6,044	6,218	6,248	5,431	5,438	5,223	3,284	4,073	2,910	**2,019**
# contigs (>= 25 Kbp)	7,012	4,688	4,708	3,543	3,573	6,729	4,283	3,226	3,227	2,768	2,768	2,528	1,300	1,765	1,033	**924**
# contigs (>=50 Kbp)	4,449	3,271	3,274	2,119	2,121	4,292	2,835	2,142	2,145	1,843	1,844	1,630	897	1,148	676	**652**
Largest contig(Kbp)	1,284	1,284	1,278	2,222	2,198	1,100	1,100	3,039	3,036	3,134	3,136	3,136	5,013	6,202	10,677	**10,677**
Total length (Kbp)	781,392	615,714	618,620	546,887	548,050	685,805	487,642	620,046	619,797	635,659	635,608	588,242	589,370	588,927	589,296	**556,859**
Total length (>= 1 Kbp)	781,186	615,508	618,417	546,886	548,049	685,805	487,642	619,842	619,593	635,455	635,406	588,040	589,174	588,731	589,104	**556,857**
Total length (>= 5 Kbp)	774,389	608,711	611,986	545,841	547,109	685,800	487,637	613,412	613,222	629,083	629,053	581,686	583,014	582,517	583,001	**556,159**
Total length (>= 10 Kbp)	751,665	585,996	590,000	536,436	537,966	683,319	485,155	594,011	593,904	611,231	611,203	563,836	566,568	565,396	566,844	**551,248**
Total length (>= 25 Kbp)	680,889	529,945	530,627	498,949	499,636	648,954	455,537	547,800	547,280	570,231	570,093	522,303	537,194	530,454	539,228	**534,434**
Total length (>= 50 Kbp)	589,893	480,070	480,059	449,330	448,950	685,805	403,091	509,550	509,117	537,749	537,651	490,800	523,329	509,117	526,899	**525,028**
GC (%)	39.57	39.59	39.57	39.22	39.40	39.36	39.37	39.55	39.55	39.54	39.54	39.55	39.55	39.55	39.55	**39.53**
N50 (Kbp)	114	138	137	249	248	123	142	331	329	455	452	467	901	824	1,440	**1,614**
N75 (bp)	51	59	58	84	82	65	69	82	82	117	117	109	413	219	553	**689**
L50	1,817	1,229	1,243	586	590	1,582	977	455	457	366	366	329	186	155	94	**83**
L75	4,373	2,935	2,976	1,509	1,527	3,487	2,200	1,453	1,458	1,038	1,040	953	420	495	258	**219**
# N's / 100 kbp	2.23	3.01	0	0.00	0	0.00	0.51	0	0	0.05	0.0	0.5	191.90	116.74	238.68	**248.99**
Busco (metazoa, n=954)			87.4		90.7		75.1		87.9			88.6	89.0	88.9	89.0	**88.9**
			4.4		0.9		2.1		4.7			2.4	2.1	2.1	2.1	**2.2**
			3.4		2.8		4.2		2.1			2.1	2.1	2.1	2.1	**2.1**
			4.8		5.6		18.6		5.3			6.9	6.8	6.9	6.8	**6.8**

406