# Montastrea cavernosa draft genome

**Website**: https://www.bco-dmo.org/dataset/875253
**Data Type**: Other Field Results
**Version**: 1
**Version Date**: 2022-12-07

**Project**
» Barriers to cross-shelf coral connectivity in the Florida Keys (KeysCoralPopgen)

| Contributors | Affiliation | Role |
|---|---|---|
| Matz, Mikhail V. | University of Texas at Austin (UT Austin) | Principal Investigator, Contact |
| Soenen, Karen | Woods Hole Oceanographic Institution (WHOI BCO-DMO) | BCO-DMO Data Manager |

**Abstract**
Draft de novo genome assembly of M. cavernosa specimen collected from the West Bank of the Flower Garden Banks National Marine Sanctuary (27.88° N, 93.83° W) in 2014.

## Table of Contents



## Coverage

**Spatial Extent**: **Lat**:27.88 **Lon**:-93.83
**Temporal Extent**: 2014 - 2014

## Dataset Description

Data is published in Rippe et al., 2021.

Data can also be downloaded at https://matzlab.weebly.com/data--code.html. Direct link to archive: https://www.dropbox.com/s/yfqefzntt896xfz/Mcavernosa_genome.tgz.

Draft genome stands for not full-length chromosome assembly.

### Methods & Sampling

One specimen of *M. cavernosa* was collected for genome sequencing from the West Bank of the Flower Garden Banks National Marine Sanctuary (27.88° N, 93.83° W) in 2014. We constructed a *de novo* genome assembly of *M. cavernosa* using a combination of PacBio reads and Illumina paired-end reads with 10X Genomics Chromium barcodes. The PacBio sequencing (9 SMRT cells, yielding 1.9 million subreads with N50 = 8kb, 15.7 Gb total) was performed at the Duke Center for Genomic and Computational Biology. 10X barcode libraries were generated and paired-end reads sequenced on an Illumina HiSeqX platform at the New York Genome Center.

### Data Processing Description

Our assembly approach follows a similar method used to generate a reference genome for the coral *Acropora millepora* (Fuller et al., 2020). We first constructed an initial draft assembly using only PacBio long reads with canu v1.6 (Koren et al., 2017). We then performed two rounds of scaffolding on this PacBio-only assembly using the 10X paired-end reads with scaff10x v1.1 (https://github.com/wtsi-hpag/Scaff10X). The resulting assembly underwent additional scaffolding using the 10X barcodes and paired-end reads with the ARCS/LINKS pipeline (Warren et al., 2015; Yeo et al., 2018). The PacBio long reads were then mapped to the assembly and used to join contigs with SSPACE-longread v1.1 (Boetzer & Pirovano, 2014). A final two rounds of scaffolding were then performed using scaff10x v1.1 again. We filled in gaps with PBJelly v15.8.24 (English et al., 2012) and error corrected with PacBio reads using the Arrow v5.0.1.9585 algorithm (https://github.com/PacificBiosciences/GenomicConsensus). To create the final assembly, we used the paired-end reads with 10X barcodes removed to polish with pilon v1.22 (Walker et al., 2014). In total, this resulted in 10,835 scaffolds with an N50 of 248Kb and maximum length of 1.8Mb.

An initial all-vs-all alignment of the assembly revealed the presence of redundant contigs. To further refine the assembly, we used the HaploMerger2 (Huang et al., 2017) pipeline to rebuild both haploid sub-assemblies from the potential polymorphic diploid genome assembly. Afterwards, the final phased 448Mb assembly contained 5,161 sequences, with an N50 of 343Kb and average length of 86kb. The maximum scaffold length was unchanged.

Protein coding gene models of the *M. cavernosa* genome were annotated using MAKER-P v2.31.9 (Campbell et al., 2014, Cantarel et al. 2008; Holt & Yandell 2011). MAKER-P was implemented using the pipeline available at the CyVerse website (http://www.cyverse.org/), available as an Atmosphere image and MPI-enabled for parallel processing. To predict gene models, we used previously published transcriptome data (SRA accession: SRP063463) (Kitchen et al., 2015) and mapped these reads using a reference-based assembly strategy with the TopHat v.2.1.1 and Cufflinks v2.2.1 pipeline (Trapnell et al., 2012), resulting in 43,577 transcripts. In addition, we generated a *de novo* transcriptome from the raw RNA-seq data with Trinity v2.0.2 (Grabherr et al., 2011), producing 200,233 transcripts. Both sets of transcripts were used as evidence in the MAKER-P pipeline.

For protein-level evidence in the annotation pipeline, we used peptide sequences from 8 robust coral species: *Pseudodiploria strigosa*, *Seriatopora hystrix*, *Platygyra carnosus*, *Montastraea faveolata*, *Madracis auretenra*, *Fungia scutaria*, *Favia sp,* and *Montastraea cavernosa* itself (Bhattacharya et al., 2016). We excluded other relatively evolutionary distant coral species due to the significant divergence between genomes, as they are likely to provide only limited evidence for gene annotation.

Before running MAKER-P, we also performed *de novo* repetitive element identification for the genome assembly using RepeatScout v1.0.5 (Price et al., 2005). The assembly was then soft masked with this resulting repeat library using RepeatMasker (http://www.repeatmasker.org, RepeatMasker 4.0.7 and Library, Smit et al., 2013-2015). 44.7% of the genome was masked and identified as repetitive. To generate the final gene set, we performed three rounds of training and predicting gene models with MAKER-P, in combination with AUGUSTUS (Hoff & Stanke, 2019). In total, 30,390 non-redundant gene models were produced from the pipeline. We assessed the completeness and quality of the predicted final gene set using 978 near-universal single-copy orthologs from the Metazoan set available in BUSCO v3 (Simão et al., 2015; Waterhouse et al., 2018). Of these, 793 are completely present and 96 are present as fragments, for a total of 90% matches to BUSCO groups searched. Finally, functional annotation was performed for the gene set using the Trinotate pipeline (Bryant et al., 2017) and eggNOG-mapper (Huerta-Cepas et al., 2017).

[ table of contents | back to top ]

## Data Files

| File |
|---|
| **Genome assembly M. cavernosa**<br>filename: Mcav_genome.zip<br>(ZIP Archive (ZIP), 504.29 MB)<br>MD5:73ebd9ebb3de868d60056e68ae0f2989<br>de novo genome assembly of M. cavernosa specimen collected from the West Bank of the Flower Garden Banks National Marine Sanctuary (27.88° N, 93.83° W). Data can also be downloaded at https://matzlab.weebly.com/data--code.html. |

[ table of contents | back to top ]

## Related Publications


Boetzer, M., & Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics, 15(1). https://doi.org/10.1186/1471-2105-15-211
*Methods*

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., Lee, T. J., Leigh, N. D., Kuo, T.-H., Davis, F. G., Bateman, J., Bryant, S., Guzikowski, A. R., Tsai, S. L., Coyne, S., Ye, W. W., Freeman, R. M., Peshkin, L., Tabin, C. J., … Whited, J. L. (2017). A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Reports, 18(3), 762–776. https://doi.org/10.1016/j.celrep.2016.12.063
*Methods*

Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C. J., Ware, D., Shiu, S.-H., Childs, K. L., Sun, Y., Jiang, N., & Yandell, M. (2013). MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations . Plant Physiology, 164(2), 513–524. https://doi.org/10.1104/pp.113.230144
*Methods*

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., & Yandell, M. (2007). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. In Genome Research (Vol. 18, Issue 1, pp. 188–196). Cold Spring Harbor Laboratory. https://doi.org/10.1101/gr.6743907
*Software*

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., & Gibbs, R. A. (2012). Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE, 7(11), e47768. https://doi.org/10.1371/journal.pone.0047768
*Methods*

Fuller, Z. L., Mocellin, V. J. L., Morris, L. A., Cantin, N., Shepherd, J., Sarre, L., Peng, J., Liao, Y., Pickrell, J., Andolfatto, P., Matz, M., Bay, L. K., & Przeworski, M. (2020). Population genetics of the coral Acropora millepora : Toward genomic prediction of bleaching. Science, 369(6501). https://doi.org/10.1126/science.aba4674
*Methods*

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology, 29(7), 644–652. doi:10.1038/nbt.1883
*Software*

Hoff, K. J., & Stanke, M. (2018). Predicting Genes in Single Genomes with AUGUSTUS. Current Protocols in Bioinformatics, e57. Portico. https://doi.org/10.1002/cpbi.57
*Methods*

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics, 12(1). https://doi.org/10.1186/1471-2105-12-491
*Software*

Huang, S., Kang, M., & Xu, A. (2017). HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. Bioinformatics, 33(16), 2577–2579. https://doi.org/10.1093/bioinformatics/btx220
*Methods*

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Molecular Biology and Evolution, 34(8), 2115–2122. https://doi.org/10.1093/molbev/msx148
*Methods*

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research, 27(5), 722–736. https://doi.org/10.1101/gr.215087.116
*Methods*

Matz, M. V. (2022). *z0on/2bRAD_denovo: best RAD-seq method ever* (Version v1) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.7392161 https://doi.org/10.5281/zenodo.7392161
*Software*

Matz, M. V. (2022). *z0on/tag-based_RNAseq: Tag-Seq: low-cost alternative to RNAseq* (Version v1) [Computer

software]. Zenodo. https://doi.org/10.5281/ZENODO.7392165 https://doi.org/10.5281/zenodo.7392165
*Software*

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. Bioinformatics, 21(Suppl 1), i351–i358. https://doi.org/10.1093/bioinformatics/bti1018
*Methods*

Rippe, J. P., Dixon, G., Fuller, Z. L., Liao, Y., & Matz, M. (2021). Environmental specialization and cryptic genetic divergence in two massive coral species from the Florida Keys Reef Tract. Molecular Ecology, 30(14), 3468–3484. Portico. https://doi.org/10.1111/mec.15931
*Results*

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351
*Methods*

Smit AFA, Hubley R, and Green P. “RepeatMasker-Open 4.0.” 2013-2015.http://www.repeatmasker.org/
*Software*

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols, 7(3), 562–578. https://doi.org/10.1038/nprot.2012.016
*Methods*

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE, 9(11), e112963. https://doi.org/10.1371/journal.pone.0112963
*Methods*

Warren, R. L., Yang, C., Vandervalk, B. P., Behsaz, B., Lagman, A., Jones, S. J. M., & Birol, I. (2015). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. GigaScience, 4(1). https://doi.org/10.1186/s13742-015-0076-3
*Methods*

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2017). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Molecular Biology and Evolution, 35(3), 543–548. https://doi.org/10.1093/molbev/msx319
*Software*

Yeo, S., Coombe, L., Warren, R. L., Chu, J., & Birol, I. (2017). ARCS: scaffolding genome drafts with linked reads. Bioinformatics, 34(5), 725–731. https://doi.org/10.1093/bioinformatics/btx675
*Methods*

[ table of contents | back to top ]

## Parameters

*Parameters for this dataset have not yet been identified*

[ table of contents | back to top ]

## Project Information

**Barriers to cross-shelf coral connectivity in the Florida Keys (KeysCoralPopgen)**

NSF Award Abstract:

Coral reefs in the Florida Keys are in severe decline, which is the most prominent along the offshore reef tract while many nearshore reefs retain high coral cover. Why then coral larvae produced from surviving nearshore patches do not recolonize offshore reefs? The investigators hypothesize that such cross-shelf migrants do not survive in the new habitat due to genetic specialization for different environmental conditions, specific to their reef of origin. This project will analyze genetic diversity of coral populations to quantify the severity of this

barrier in three common coral species. This will be the first study to assess the strength of ecological barriers to coral dispersal across the seascape, which will fill an important knowledge gap that currently precludes informed assessment of threats to Florida reefs and development of accurate models to forecast their future. The project includes a variety of broader impact activities. Public outreach: This project is very well poised to raise public awareness of ongoing biodiversity loss. The investigators regularly give public lectures at diverse Austin venues, such as Science Under the Stars, Science in the Pub, and Nerd Nite. The progress of the project will be followed by press releases, materials on the University of Texas public outreach web page Know and in The Daily Texan. K-12 outreach: Two interns from Crockett High School and will be involved in the research. Graduate education: The project will be the main PhD theme for one full-time graduate student. Undergraduate education: The primary investigator regularly employs undergraduates. In this project they will participate in field experiments and sample processing, and later assigned individual sub-projects. Promotion of rapid data sharing: All sequencing data will be made available for unconditional use prior to publication. Specifically, the investigators will rapidly share new coral genome data, as well as data on genome-wide variation in coral populations.

The project consists of four parts, each of which is designed to demonstrate the action of divergent selection among nearshore and offshore reefs and to obtain quantitative estimate of its demographic impact. (1) To look for genomic signatures of ongoing selection between inshore and offshore habitats. The research team will perform genome-wide genotyping in three coral species representing alternative life histories and replicate the nearshore-offshore population comparison along the Florida Reef Tract. (2) To confirm continuous action of selection by comparing the extent of inshore-offshore divergence among juveniles and adults. Juveniles are presumed to have experienced local selection for shorter time and hence should show less cross-shelf divergence at the candidate loci. (3) To demonstrate association of genotypes at the candidate loci with local fitness by quantifying in situ growth and survival of genotyped juveniles. This part as well as part 2 is expected to provide quantitative estimates of the strength of selection against cross-shelf immigrants. (4) To verify the obtained estimates by simulating genome evolution under divergent selection and confirming that the proposed selection scenario is compatible with the observed genomic signatures.

[ table of contents | back to top ]

## Funding


| Funding Source | Award |
|---|---|
| NSF Division of Ocean Sciences (NSF OCE) | OCE-1737312 |


[ table of contents | back to top ]