pdam-genome

# Materials and Methods

## Sequencing and assembly

The *Pocillopora damicornis* colony used for genome sequencing was collected at Isla de Saboga, Panama in March 2005, and cultured indoors at the University of Miami Coral Resource Facility until the time of sampling. Genomic DNA was extracted from two fragments of this genotype in XXX 2016 using XXX protocol and delivered to Dovetail Genomics (Santa Cruz, CA, USA), where Chicago libraries were prepared and sequenced on an Illumina XXX platform, and genome scaffolds were assembled *de novo* using the HiRise software pipeline [@Putnam:2016gk]. The Dovetail HiRise scaffolds were then filtered to remove those of potential non-coral origin using BLAST [@Altschul:1990dw] searches against three databases: 1) *Symbiodinium*, containing the genomes of *S. minutum* [@Shoguchi:2013bx] and *S. microadriaticum* [@Aranda:2016ez], 2) bacteria, containing 6954 complete bacterial genomes from NCBI, and 3) viruses, containing 2996 viral genomes from the phantome database (phantome.org; accessed 2017-03-01). Scaffolds with a BLAST hit to any of these databases with an e-value < 10-20 and a bitscore > 1000 were considered to be non-coral in origin and removed from the assembly [@Baumgarten:2015gu].

## Annotation

The filtered assembly was analyzed for completeness using BUSCO [@Simao:2015kk] to search for 978 universal metazoan single-copy orthologs. The --long option was passed to BUSCO to use these genes in training the *ab initio* gene prediction software Augustus [@Stanke:2004ih]. Subsequently, the Augustus gene prediction parameters were used in the MAKER pipeline [@Campbell:2014go] to annotate gene models, using as supporting 'evidence' two RNA-seq datasets from *P. damicornis* [Mayfield?; Bhattacharya/Traylor-Knowles?] and one from the closely-related species *Stylophora pistillata* (Voolstra et al., unpublished dataset), and protein sequences from 20 coral species [@Bhattacharya:2016bz]. Results from this initial MAKER run were used to train a second gene predictor (SNAP; <http://korflab.ucdavis.edu/software.html>) prior to an iterative MAKER run to refine gene models. Protein sequences were then extracted from the assembly and putative functional annotations were added by searching for homologous proteins in the Uniprot-Swissprot database [@TheUniProtConsortium:2017fy] using BLAST, and protein domains (and associated gene ontology (GO) terms) using InterProScan [@Finn:2017bz]. Annotation statistics were generated using the Genome Annotation Generator software [@GAGtheGenomeAnno:Bux0SZoV].

# Results

## Genome statistics

The final genome assembly of 234 Mb comprised 4,393 scaffolds with an N50 of 326 kb. We identified 26,077 gene models, with 21,389 (82%) of these being apparently complete with start and stop codons. The total gene length was 152 Mb (65.2% of the genome), and the average gene was 5,860 bp with 7 exons and 6 introns. Exons averaged 245 bp, while introns averaged 667 bp in length. Out of 978 metazoan single-copy orthologs searched using BUSCO, 865 (88.4%) were present and complete in the *P. damicornis* genome assembly (five of these were duplicated). Twenty-eight orthologs were present but fragmented, and 85 (8.7%) were missing.

|  |  |
| --- | --- |
| Parameter | Value |
| Assembly size | 234 Mb |
| Scaffold N50 | 326 kb |
| Number of gene models | 26,077 |
| Number of complete gene models | 21,389 |
| % of genome covered by genes | 65.2 |
| % of genome covered by CDS | 15.2 |
| Mean gene model length | 5,860 bp |
| Mean predicted protein length | 455 amino acids |
| % of metazoan BUSCOs present | 88.4 |