DATA NOTE

The scaffold-level genome sequence of an encrusting sponge, *Halisarca caerulea* Vacelet & Donadey, 1987, and its associated microbial metagenome sequences

[version 1; peer review: 2 approved]

Jasper M. de Goeij[1-3], Benjamin Mueller[1-3], Michelle Achlatis[1,2], Sara Campana[1], Meggie Hudspith[1], Niklas A. Kornder[1], Ute Hentschel[4], Graeme Oatley[5], Elizabeth Sinclair[5], Eerik Aunin[5], Noah Gettle[5], Camilla Santos[5], Michael Paulini[5], Haoyu Niu[5], Victoria McKenna[5], Rebecca O'Brien[5], Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory Team, Wellcome Sanger Institute Scientific Operations: Sequencing Operations, Wellcome Sanger Institute Tree of Life Core Informatics Team, EBI Aquatic Symbiosis Genomics Data Portal Team, Aquatic Symbiosis Genomics Project Leadership

[1]Department of Freshwater and Marine Ecology, University of Amsterdam, Amsterdam, The Netherlands
[2]CARMABI Foundation, Willemstad, Curaçao, The Netherlands
[3]Department for Marine Ecology, University of Bremen, Bremen, Germany
[4]GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany
[5]Tree of Life, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 09 Jul 2025, **10**:344
https://doi.org/10.12688/wellcomeopenres.24281.1
Latest published: 09 Jul 2025, **10**:344
https://doi.org/10.12688/wellcomeopenres.24281.1

Abstract

We present a scaffold-level genome assembly from a *Halisarca caerulea* specimen (encrusting sponge; Porifera; Demospongiae; Chondrillida; Halisarcidae). The genome sequence is 195.70 megabases in span. The mitochondrial genome has also been assembled and is 19.15 kilobases in length. Gene annotation of this assembly on Ensembl identified 26,722 protein-coding genes. The metagenome of the specimen was also assembled and four binned bacterial genomes related to the relevant sponge symbiont clades Alphaproteobacteria bacterium GM7ARS4 and Gammaproteobacteria bacterium *AqS2* ((Tethybacterales) were identified.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 09 Jul 2025	✓ view	✓ view

1. **Matsapume Detcharoen**, Prince of Songkla University, Hat Yai, Thailand
2. **Rodrigo Costa**, Universidade de Lisboa, Lisbon, Portugal

Any reports and responses or comments on the article can be found at the end of the article.

Keywords
Halisarca caerulea, encrusting sponge, genome sequence, chromosomal, Chondrillida; metagenome assembly

This article is included in the Tree of Life gateway.

Corresponding author: Aquatic Symbiosis Genomics Project Leadership (Mark.Blaxter@sanger.ac.uk)

Author roles: de Goeij JM: Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Mueller B**: Resources, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Achlatis M**: Resources, Writing – Review & Editing; **Campana S**: Resources, Writing – Review & Editing; **Hudspith M**: Resources, Writing – Review & Editing; **Kornder NA**: Resources, Writing – Review & Editing; **Hentschel U**: Project Administration, Supervision, Writing – Review & Editing; **Oatley G**: Investigation; **Sinclair E**: Investigation; **Aunin E**: Formal Analysis; **Gettle N**: Data Curation, Formal Analysis, Methodology; **Santos C**: Formal Analysis; **Paulini M**: Formal Analysis; **Niu H**: Project Administration; **McKenna V**: Project Administration; **O'Brien R**: Project Administration;

Competing interests: No competing interests were disclosed.

Grant information: This work was funded by the Gordon and Betty Moore Foundation through a grant (GBMF8897) to the Wellcome Sanger Institute to support the Aquatic Symbiosis Genomics Project, and by Wellcome through core funding to the Wellcome Sanger Institute (220540). Fieldwork to collect the samples was funded by the European Research Council under the European Union's Horizon 2020 research and innovation programme (Starting Grant agreement number 715513 to JMdG).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: de Goeij JM, Mueller B, Achlatis M *et al.* **The scaffold-level genome sequence of an encrusting sponge, *Halisarca caerulea* Vacelet & Donadey, 1987, and its associated microbial metagenome sequences [version 1; peer review: 2 approved]** Wellcome Open Research 2025, **10**:344 https://doi.org/10.12688/wellcomeopenres.24281.1

First published: 09 Jul 2025, **10**:344 https://doi.org/10.12688/wellcomeopenres.24281.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Porifera; Demospongiae; Verongimorpha; Chondrillida; Halisarcidae; *Halisarca*; *Halisarca caerulea* Vacelet & Donadey, 1987 (NCBI:txid858189).

Background

The marine sponge genus *Halisarca* (Dujardin, 1838) has a worldwide distribution and currently includes 22 accepted species (de Voogd *et al.*, 2024). Species identification in this genus is particularly difficult due to the lack of a mineral (i.e., spicules) or proteinaceous (i.e., spongin fibres) skeleton (Ereskovsky, 2007; Ereskovsky *et al.*, 2011). Histological, ultrastructural and embryological analyses are needed to identify the species (Bergquist & Kelly, 2004; Vacelet & Donadey, 1987).

The encrusting 1–3 mm thick "star" sponge *Halisarca caerulea* Vacelet & Donadey 1987 is found on coral reefs (de Goeij *et al.*, 2008a; Diaz & Rützler, 2009) and in mangroves (Diaz & Rützler, 2009; Rützler *et al.*, 2000) throughout the wider Caribbean, extending to the east coast of Brazil (de Moraes, 2011). Individuals can be as small as a few mm^2 up to several hundred cm^2 in size and range in colour from violet to blue (Figure 1A). They contain distinctive star-shaped canal systems surrounding their osculum (i.e. outflow opening) (Figure 1B), distributed evenly every 1–3 cm across its surface. *Halisarca caerulea* grows on dead coral substrate or mangrove roots from approximately 2 to at least 50 m water depth (Lesser *et al.*, 2020). In shallow reef areas it is found in low-light, concealed, "cryptic" reef spaces, such as under coral overhangs and in crevices or cavities (de Goeij *et al.*, 2008a; de Goeij *et al.*, 2013). Below a water depth of around 30 m, *H. caerulea* can also be found in large patches on exposed surfaces of the reef (Figure 1D).

This sponge is known for its important role in cycling of dissolved organic matter (DOM; the largest available organic resource in the oceans) on coral reefs through the sponge loop (de Goeij *et al.*, 2013). As one of the four model species described in the original study, *H. caerulea* was found to consume >90 % of its daily carbon diet as DOM (de Goeij *et al.*, 2008a), of which approximately half is turned over as cellular debris through very rapid cell division and shedding of its filter cells (choanocytes) (Alexander *et al.*, 2014; De Goeij *et al.*, 2009).

Previously, the classification as either low or high microbial abundance sponge species has been controversial (LMA/HMA) (Alexander *et al.*, 2014; Maldonado *et al.*, 2012). The thick, collagenous tissue structure resembles that of HMA sponges. However, its physiological performance in terms of uptake and processing rates of particulate and dissolved food sources as well as the excretion of inorganic nutrients resembles that of other LMA species (Campana *et al.*, 2021; Campana *et al.*, 2024). In combination with the relatively low abundance (10^5–10^6 prokaryotes per mL), composition, and small cell size of its endosymbiotic prokaryote community, that are more similar to seawater planktonic prokaryotes than to HMA prokaryotic endosymbionts, *H. caerulea* is now considered a LMA species. The prokaryote community of *H. caerulea* is dominated by the phyla Pseudomonadota, Cyanobacteriota, Bacteroidota, and Spirochaetota (Campana *et al.*, 2024). Within the sponge holobiont (i.e., functional unit consisting of sponge host and associated microbiota), ammonium-oxidising Archaea appear to play a relevant role in internal nitrogen cycling (Cardoso *et al.*, 2013). Furthermore, carbon fixation and chitin degradation were suggested to support the holobiont metabolism on the shallow and deep reef, respectively (Lesser *et al.*, 2020).

Halisarca caerulea is also a model species in animal-microbe symbiosis studies, for example to show translocation of isotope-tracer labelled food between sponge host and microbial symbionts (Campana *et al.*, 2021; Campana *et al.*, 2022; de Goeij *et al.*, 2008b; Hudspith *et al.*, 2021). In addition, the species is used as a metazoan model for wound healing and tissue regeneration (Alexander *et al.*, 2015; Kenny *et al.*, 2018). It is also used as a model for metazoan uptake systems, as its cell kinetics, choanocyte shedding and structure of the choanosome show striking similarities with the human gastrointestinal tract (de Goeij *et al.* 2009).

The genome of the sponge, *H. caerulea*, was sequenced as part of the Aquatic Symbiosis Genomics (ASG) project, a collaborative effort to sequence 50 sponge species, both marine and freshwater, around the world. *Halisarca caerulea* is the first encrusting and cryptic sponge in this effort. The absence of a skeleton in this species is a unique feature within the Porifera and therefore its genome information adds to the morphological and taxonomic diversity of sponge genomics. Moreover, the family Halisarcidae are proposed to be either highly derived or primitive group of sponges to study the origin of multicellularity (Maldonado, 2004). Comparative genomics on the recently assembled genome of the sister species *H. dujardinii* will help provide new insights into this family (Borisenko *et al.*, 2024).

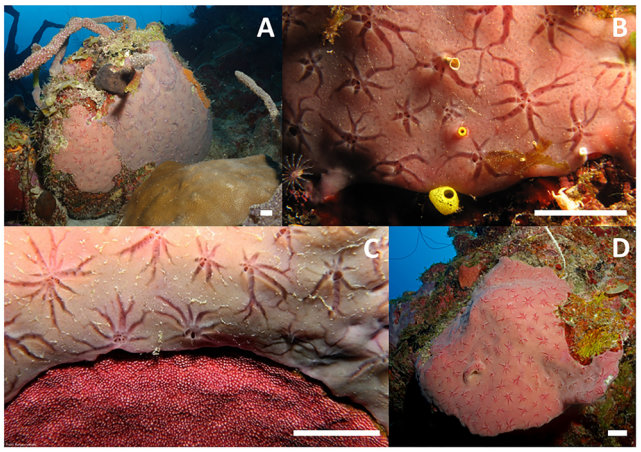

Figure 1. Morphological variations of *Halisarca caerulea* on the reefs of Curaçao. (**A**) Violet and purple colour morph next to each other. (**B**) Close up of star-shape canal system surrounding oscula. (**C**) Shallow water (~10 m) specimen concealed under a coral overhang. (**D**) Deep water specimen growing openly on the reef at 35 m depth. White scale bar in the lower right corner of each panel represents 2 cm.

Genome sequence report

The genome was sequenced from an adult *Halisarca caerulea* (Figure 1) collected from the southern Caribbean island of Curaçao. A total of 106-fold coverage in Pacific Biosciences single-molecule HiFi long reads was generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 19 missing joins or mis-joins and removed 7 haplotypic duplications, reducing the assembly length by 0.51%.

The final assembly has a total length of 195.70 Mb in 488 sequence scaffolds with a scaffold N50 of 2.8 Mb (Table 1). The snail plot in Figure 2 provides a summary of the assembly statistics, while the distribution of assembly scaffolds on GC proportion and coverage is shown in Figure 3. The cumulative assembly plot in Figure 4 shows curves for subsets of scaffolds assigned to different phyla. We did not manage to get the assembly to chromosome level, so it is submitted as curated scaffolds. Based on terminal telomers found in the scaffolds, the

Table 1. Genome data for *Halisarca caerulea*, odHalCaeu1.1.

Project accession data			
Assembly identifier	odHalCaeu1.1		
Species	*Halisarca caerulea*		
Specimen	odHalCaeu1		
NCBI taxonomy ID	858189		
BioProject	PRJEB62893		
BioSample ID	Genome sequencing: SAMEA8576948 Hi-C scaffolding: SAMEA8576954 RNA sequencing: SAMEA8576955		
Isolate information	odHalCaeu1: whole organism (genome sequence) odHalCaeu7: whole organism (Hi-C sequencing) odHalCaeu8: whole organism (RNA sequencing)		
Assembly metrics			
Consensus quality (QV)	50.8		
BUSCO*	C:69.7%[S:67.7%,D:2.0%],F:11.8%,M:18.5%,n:954		
Percentage of assembly mapped to chromosomes	0%		
Sex chromosomes	None		
Organelles	Mitochondrial genome: 19.15 kb		
Sequencing information			
Platform	**Run accession**	**Read count**	**Base count (Gb)**
Hi-C Illumina NovaSeq 6000	ERR11547915	5.54e+08	83.66
Hi-C Illumina NovaSeq 6000	ERR11547916	4.58e+08	69.18
Hi-C Illumina NovaSeq 6000	ERR11641141	1.07e+09	161.65
PacBio Sequel IIe	ERR11531687	1.70e+06	9.72
PacBio Sequel IIe	ERR11531686	1.81e+06	11.98
RNA Illumina NovaSeq X	ERR13093635	1.97e+08	29.81
Genome assembly			
Assembly accession	GCA_963170055.1		
Accession of alternate haplotype	GCA_963170115.1		
Span (Mb)	195.70		

Genome assembly	
Number of contigs	901
Contig N50 length (Mb)	1.7
Number of scaffolds	488
Scaffold N50 length (Mb)	2.8
Longest scaffold (Mb)	14.2
Genome annotation	
Number of protein-coding genes	26,722
Number of non-coding genes	733
Number of gene transcripts	39,196

* BUSCO scores based on the metazoa_odb10 BUSCO set using version 5.4.3. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/Halisarca%20caerulea/dataset/CAUJGJ01/busco.

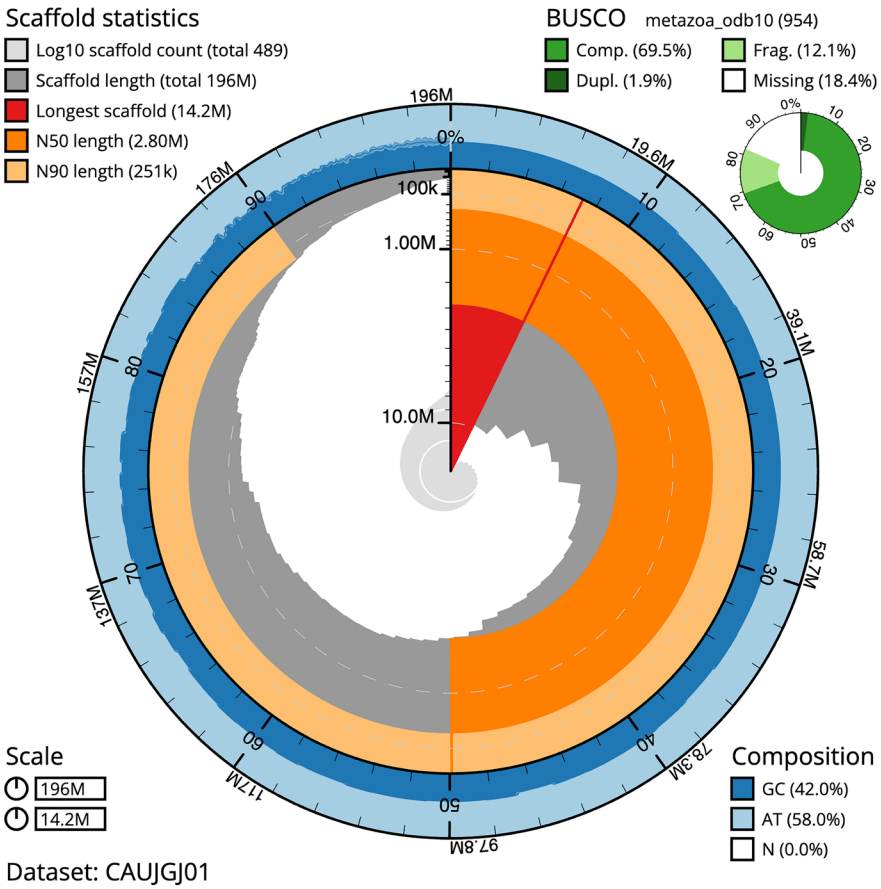

Figure 2. Genome assembly of *Halisarca caerulea*, odHalCaeu1.1: metrics. The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 195,692,738 bp assembly. The distribution of sequence lengths is shown in dark grey with the plot radius scaled to the longest sequence present in the assembly (14,164,103 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 sequence lengths (2,795,200 and 251,000 bp), respectively. The pale grey spiral shows the cumulative sequence count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the metazoa_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/Halisarca%20caerulea/dataset/CAUJGJ01/snail.

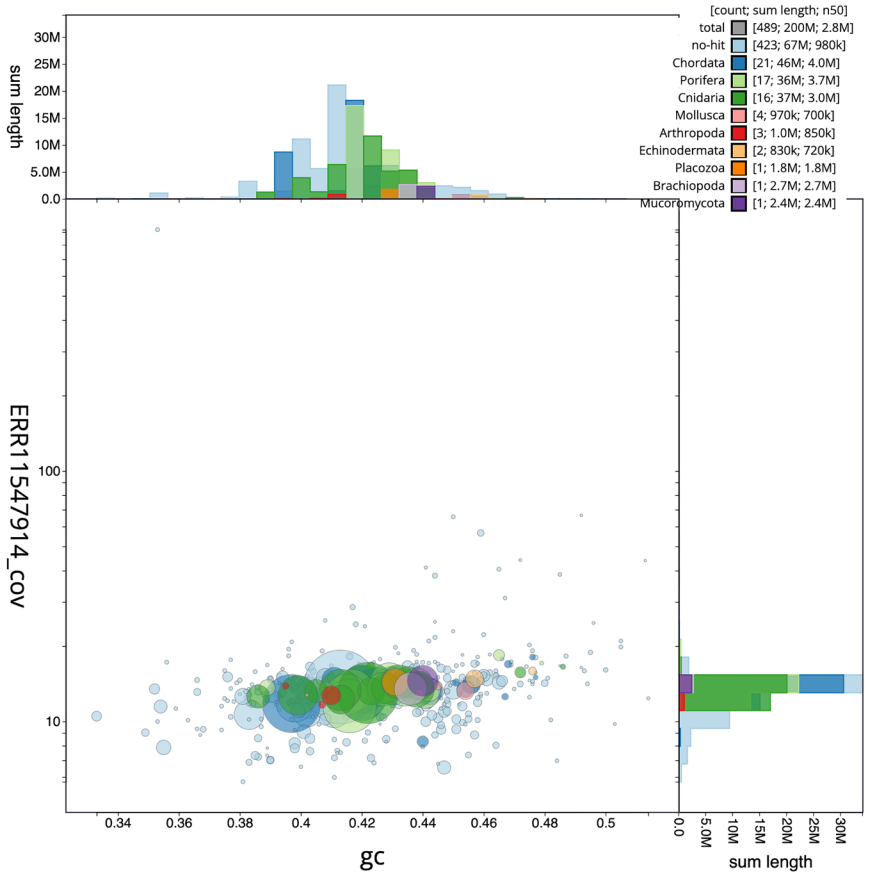

Figure 3. Genome assembly of *Halisarca caerulea*, odHalCaeu1.1: BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/Halisarca%20caerulea/dataset/CAUJGJ01/blob.

estimated chromosome number is between 9 and 17. While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial genome was also assembled and can be found as a contig within the multifasta file of the genome submission with GenBank accession OY720623.1.

The estimated Quality Value (QV) of the final assembly is 50.8 with *k*-mer completeness of 99.98%, and the assembly has a BUSCO v5.4.3 completeness of 69.7% (single = 67.7%, duplicated = 2.0%), using the metazoa_odb10 reference set (n = 954).

Metagenome report

Four binned genomes were generated from the metagenome assembly (Figure 5) of which one was classified as a high-quality metagenome assembled genome (MAG) (see methods). The completeness values for these assemblies range from approximately 71% to 88% with contamination below 2%. For details on binned genomes see Table 2. Interestingly, two binned genomes each fell into the alphaproteobacterial order GM7ARS4 and the gammaproteobacterial family AqS2, MAGs of which have previously been isolated from sponges.

Genome annotation report

The *Halisarca caerulea* genome assembly (GCA_963170055.1) was annotated at the European Bioinformatics Institute (EBI) on Ensembl Rapid Release. The resulting annotation includes 39,196 transcribed mRNAs from 26,722 protein-coding and 733 non-coding genes (Table 2; https://rapid.ensembl.org/Halisarca_caerulea_GCA_963170055.1/Info/Index). The average transcript length is 5,833.36. There are 1.43 coding transcripts per gene and 5.29 exons per transcript. An alternative annotation is available here: https://github.com/Aquatic-Symbiosis-Genomics-Project/sponge_annotations/tree/main/results/odHalCaeu1.

Methods

Sample acquisition

The genome was sequenced from an *H. caerulea* individual (specimen ID GHC0000080, ToLID odHalCaeu1) (Figure 1) collected on 2021-01-19 by SCUBA diving from reef station "Buoy 1" (latitude 12.12, longitude –68.97) near the Caribbean

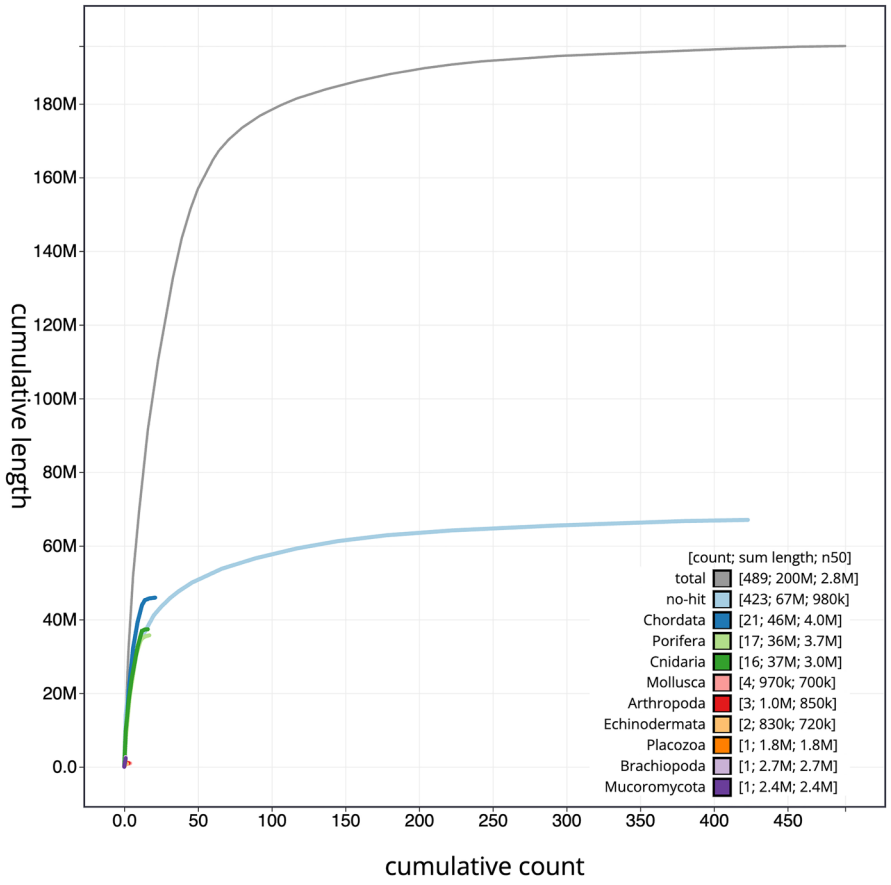

Figure 4. Genome assembly of *Halisarca caerulea*, odHalCaeu1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/Halisarca%20caerulea/dataset/CAUJGJ01/cumulative.

Research and Management of Biodiversity (CARMABI) marine research station on the southern Caribbean island of Curaçao. A second specimen (specimen ID GHC0000086, ToLID odHalCaeu7) was used for Hi-C sequencing and another for RNA sequencing (specimen ID GHC0000087, ToLID odHalCaeu8). The specimens were collected and identified by Jasper Merijn De Goeij and Benjamin Mueller (University of Amsterdam), and preserved by snap freezing.

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of core procedures: sample preparation; sample homogenisation, DNA extraction, fragmentation and clean-up. Protocols developed by the WSI Tree of Life laboratory are publicly available on protocols.io (Howard *et al.*, 2025). In sample preparation, the odHalCaeu1 sample was weighed and dissected on dry ice (Jay *et al.*, 2023). Tissue was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023). HMW DNA was extracted using the Automated MagAttract v1 protocol (Sheerin *et al.*, 2023). DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Todorovic *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation (Strickland *et al.*, 2023), using AMPure PB beads to eliminate shorter fragments and concentrate the DNA. The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

RNA was extracted from whole organism tissue of odHalCaeu8 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol (do Amaral *et al.*, 2023). The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

Sequencing

Pacific Biosciences HiFi circular consensus DNA sequencing libraries were constructed according to the manufacturers' instructions. Poly(A) RNA-Seq libraries were constructed using

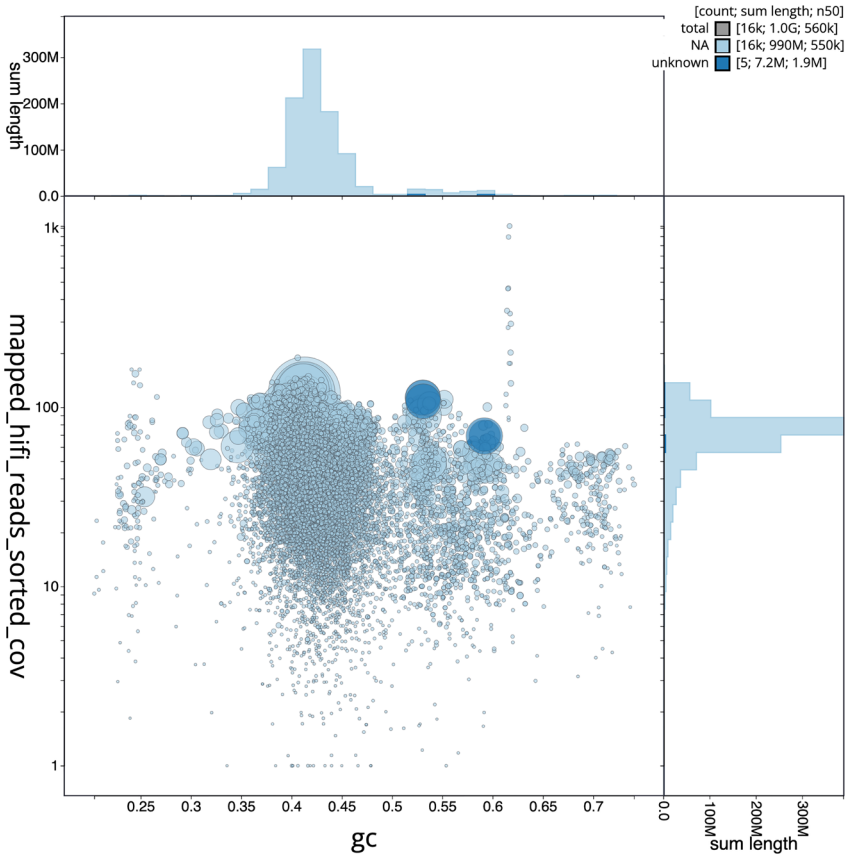

Figure 5. Blob plot of base coverage in mapped against GC proportion for sequences in the *Halisarca caerulea* metagenome. Binned contigs are coloured by family. Circles are sized in proportion to sequence length on a square root scale, ranging from 2,140 to 8,133,556. Histograms show the distribution of sequence length sum along each axis.

Table 2. Quality metrics and taxonomic assignments of the binned metagenomes.

NCBI taxon	Taxid	GTDB taxonomy	Quality	Size (bp)	Contigs	Circular	Mean coverage	Completeness (%)	Contamination (%)
Alphaproteobacteria bacterium	91750	o__GM7ARS4	High	1,869,963	1	Yes	27	87.67	0
Gammaproteobacteria bacterium	86473	f__AqS2	Medium	2,047,616	2	Partial	16.14	88.15	1.22
Gammaproteobacteria bacterium	86473	f__AqS2	Medium	1,470,391	1	No	17.19	71.27	1.22
Alphaproteobacteria bacterium	91750	o__GM7ARS4	High	1,859,046	1	Yes	23.42	87.75	0

the NEB Ultra II RNA Library Prep kit. DNA and RNA sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences Sequel IIe (HiFi) and Illumina NovaSeq X (RNA-Seq) instruments. Hi-C data were also generated from whole organism tissue of odHalCaeu7 using the Arima2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

Genome assembly and curation

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021) and haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2020). The assembly was then scaffolded with Hi-C data (Rao *et al.*, 2014) using YaHS (Zhou *et al.*, 2023). Manual curation was performed using JBrowse2 (Diesh *et al.*, 2023), HiGlass (Kerpedjiev *et al.*, 2018) and Pretext

(Harry, 2022). The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

Metagenome assembly

The metagenome assembly was generated using Hifiasm (version 0.16.1) and binned using MetaBAT2 (Kang *et al.*, 2019), MaxBin (Wu *et al.*, 2014), bin3C (DeMaere & Darling, 2019), and MetaTOR. The resulting bin sets of each binning algorithm were optimised and refined using MAGScoT (Rühlemann *et al.*, 2022). PROKKA (Seemann, 2014) was used to identify tRNAs and rRNAs in each bin, CheckM (Parks *et al.*, 2015) (checkM_DB release 2015-01-16) was used to assess bin completeness/contamination, and GTDB-TK (Chaumeil *et al.*, 2022) (GTDB release 214) was used to taxonomically classify bins. Taxonomic replicate bins were identified using dRep (Olm *et al.*, 2017) with default settings (95% ANI threshold). The final bin set was filtered for bacteria and archaea. All bins were assessed for quality and categorised as metagenome-assembled genomes (MAGs) if they met the following criteria: contamination ≤ 5%, presence of 5S, 16S, and 23S rRNA genes, at least 18 unique tRNAs, and either ≥ 90% completeness or ≥ 50% completeness with fully circularised chromosomes. Bins that did not meet these thresholds, or were identified as taxonomic replicates of MAGs, were retained as 'binned metagenomes' provided they had ≥ 50% completeness and ≤ 10% contamination. Software tool versions and sources are given in Table 3.

Evaluation of the final assembly

The MerquryFK tool (Rhie *et al.*, 2020), run within a Singularity container (Kurtzer *et al.*, 2017), was used to evaluate *k*-mer completeness and assembly quality for the primary and alternate haplotypes, using the *k*-mer databases (k = 31) computed before genome assembly. The analysis outputs included the assembly QV consensus quality value. A Hi-C map for the final host and eukaryote cobiont assemblies were produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). The genome was also analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021) were calculated.

Table 3 contains a list of relevant software tool versions and sources.

Genome annotation

The Ensembl Genebuild annotation system (Aken *et al.*, 2016) was used to generate annotation for the *Halisarca caerulea* assembly (GCA_963170055.1) in Ensembl Rapid Release at the EBI. Annotation was created primarily through alignment of transcriptomic data to the genome, with gap filling via

Table 3. Software tools: versions and sources.

Software tool	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
Blast	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.7	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.4.3 and 5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	427104ea91c78c3b8b8b49f1a7d6bbeaa869ba1c	https://github.com/thegenemyers/FASTK
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.15.3-r339	https://github.com/chhylp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84 aa44357826c0b6753eb28de	https://github.com/higlass/higlass
MerquryFK	d00d98157618f4e8d1a9190026b19b471055b22e	https://github.com/thegenemyers/MERQURY.FK
MitoHiFi	2	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14, 1.17, and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.04.0-5857	https://github.com/nextflow-io/nextflow
PretextView	0.2.5	https://github.com/wtsi-hpag/PretextView

Software tool	Version	Source
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
samtools	1.16.1, 1.17, and 1.18	https://github.com/samtools/samtools
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.0.0	https://github.com/sanger-tol/treeval
YaHS	1.1a.2	https://github.com/c-zhou/yahs

protein-to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 2019).

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Tree of Life collaborator. The Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible. The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international)

Each transfer of samples is undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Tree of Life collaborator, Genome Research Limited (operating as the Wellcome Sanger Institute) and in some circumstances other Tree of Life collaborators.

Data availability

European Nucleotide Archive: Halisarca caerulea. Accession number PRJEB62893; https://identifiers.org/ena.embl/PRJEB62893. The genome sequence is released openly for reuse. The *Halisarca caerulea* genome sequencing initiative is part of the Aquatics Symbiosis Genomics (ASG) project. All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

Author information

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory Team are listed here: https://doi.org/10.5281/zenodo.10066175.

Members of the Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: https://doi.org/10.5281/zenodo.10043364.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: https://doi.org/10.5281/zenodo.10066637.

Members of the European Bioinformatics Institute ASG Data Portal team are listed here: https://doi.org//10.5281/zenodo.10076466.

Members of the Wellcome Sanger Institute/Aquatic Symbiosis Genomics Project Leadership are listed here: https://doi.org/10.5281/zenodo.10184833.

References

Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316.
PubMed Abstract | Publisher Full Text | Free Full Text

Aken BL, Ayling S, Barrell D, *et al.*: **The Ensembl gene annotation system**. *Database (Oxford).* 2016; **2016**: baw093.
PubMed Abstract | Publisher Full Text | Free Full Text

Alexander BE, Achlatis M, Osinga R, *et al.*: **Cell kinetics during regeneration in the sponge *Halisarca caerulea*: how local is the response to tissue damage?** *PeerJ.* 2015; **3**: e820.
PubMed Abstract | Publisher Full Text | Free Full Text

Alexander BE, Liebrand K, Osinga R, *et al.*: **Cell turnover and detritus production in marine sponges from tropical and temperate benthic ecosystems.** *PLoS One.* 2014; **9**(10): e109486.
PubMed Abstract | Publisher Full Text | Free Full Text

Allio R, Schomaker-Bastos A, Romiguier J, *et al.*: **MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics.** *Mol Ecol Resour.* 2020; **20**(4): 892–905.
PubMed Abstract | Publisher Full Text | Free Full Text

Bergquist PR, Kelly M: **Taxonomy of some halisarcida and homosclerophorida (Porifera: Demospongiae) from the Indo-Pacific.**

N Z J Mar Freshwater Res. 2004; **38**(1): 51–66.
Publisher Full Text

Borisenko I, Predeus AV, Lavrov A, *et al.*: **First draft genome assembly and characterization of sponge *Halisarca dujardinii* reveals key components of basement membrane and broad repertoire of aggregation factors**. *bioRxiv.* 2024.
Publisher Full Text

Campana S, Arts MGI, Díez-Vives C, *et al.*: **Sponges on shifting reefs: holobionts show similar molecular and physiological responses to coral versus macroalgal food.** *Front Mar Sci.* 2024; **11**.
Publisher Full Text

Campana S, Hudspith M, Lankes D, *et al.*: **Processing of naturally sourced macroalgal- and coral-Dissolved Organic Matter (DOM) by high and low microbial abundance encrusting sponges.** *Front Mar Sci.* 2021; **8**.
Publisher Full Text

Campana S, Riesgo A, Jongepier E, *et al.*: **Meta-transcriptomic comparison of two sponge holobionts feeding on coral- and macroalgal-Dissolved Organic Matter.** *BMC Genomics.* 2022; **23**(1): 674.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Cardoso J, van Bleijswijk J, Witte H, *et al.*: **Diversity and abundance of ammonia-oxidizing *Archaea* and *Bacteria* in tropical and cold-water coral reef sponges.** *Aquat Microb Ecol.* 2013; **68**(3): 215–230.
Publisher Full Text

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Chaumeil PA, Mussig AJ, Hugenholtz P, *et al.*: **GTDB-Tk v2: memory friendly classification with the Genome Taxonomy Database.** *Bioinformatics.* 2022; **38**(23): 5315–5316.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

De Goeij JM, De Kluijver A, Van Duyl FC, *et al.*: **Cell kinetics of the marine sponge *Halisarca caerulea* reveal rapid cell turnover and shedding.** *J Exp Biol.* 2009; **212**(Pt 23): 3892–3900.
PubMed Abstract | **Publisher Full Text**

de Goeij JM, van Oevelen D, Vermeij MJ, *et al.*: **Surviving in a Marine Desert: the sponge loop retains resources within coral reefs.** *Science.* 2013; **342**(6154): 108–110.
PubMed Abstract | **Publisher Full Text**

de Goeij J, van den Berg H, van Oostveen M, *et al.*: **Major bulk Dissolved Organic Carbon (DOC) removal by encrusting coral reef cavity sponges.** *Mar Ecol Prog Ser.* 2008a; **357**: 139–151.
Publisher Full Text

de Goeij J, van den Berg H, van Oostveen M, *et al.*: **Major bulk Dissolved Organic Carbon (DOC) removal by encrusting coral reef cavity sponges.** *Mar Ecol Prog Ser.* 2008b; **357**: 139–151.
Publisher Full Text

de Moraes FC: **Esponjas das ilhas oceânicas Brasileiras**. Museu Nacional, 2011.

de Voogd NJ, Alvarez B, Boury-Esnault N, *et al.*: **World Porifera Database**. 2024; [Accessed 7 June 2024].
Reference Source

DeMaere MZ, Darling AE: **bin3C: exploiting Hi-C sequencing data to accurately resolve Metagenome-Assembled Genomes.** *Genome Biol.* 2019; **20**(1): 46.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Denton A, Oatley G, Cornwell C, *et al.*: **Sanger Tree of Life sample homogenisation: PowerMash.** *protocols.io.* 2023.
Publisher Full Text

Diaz MC, Rützler K: **Biodiversity and abundance of sponges in Caribbean Mangrove: indicators of environmental quality.** *Smithson Contrib Mar Sci.* 2009; (38): 151–172.
Publisher Full Text

Diesh C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of synteny and structural variation.** *Genome Biol.* 2023; **24**(1): 74.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

do Amaral RJV, Bates A, Denton A, *et al.*: **Sanger Tree of Life RNA extraction: automated MagMax™ mirVana.** *protocols.io.* 2023.
Publisher Full Text

Ereskovsky AV: **A new species of *Halisarca* (Demospongiae: Halisarcida) from the Sea of Okhotsk, North Pacific.** *Zootaxa.* 2007; **1432**: 57–66.
Publisher Full Text

Ereskovsky AV, Lavrov DV, Boury-Esnault N, *et al.*: **Molecular and morphological description of a new species of *Halisarca* (Demospongiae: Halisarcida) from Mediterranean Sea and a redescription of the type species *Halisarca dujardini*.** *Zootaxa.* 2011; **2768**(1): 5–31.
Publisher Full Text

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Harry E: **PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps**. 2022.
Reference Source

Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025.
Publisher Full Text

Hudspith M, Rix L, Achlatis M, *et al.*: **Subcellular view of host-microbiome nutrient exchange in sponges: insights into the ecological success of an early metazoan-microbe symbiosis.** *Microbiome.* 2021; **9**(1): 44.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life sample preparation: triage and dissection.** *protocols.io.* 2023.
Publisher Full Text

Kang DD, Li F, Kirton E, *et al.*: **MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies.** *PeerJ.* 2019; **7**: e7359.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Kenny NJ, de Goeij JM, de Bakker DM, *et al.*: **Towards the identification of ancestrally shared regenerative mechanisms across the Metazoa: a transcriptomic case study in the demosponge *Halisarca caerulea***. *Mar Genomics.* 2018; **37**: 135–147.
PubMed Abstract | **Publisher Full Text**

Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HiGlass: web-based visual exploration and analysis of genome interaction maps.** *Genome Biol.* 2018; **19**(1): 125.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Lesser MP, Mueller B, Pankey MS, *et al.*: **Depth-dependent detritus production in the sponge, *Halisarca caerulea*.** *Limnol Oceanogr.* 2020; **65**(6): 1200–1216.
Publisher Full Text

Maldonado M: **Choanoflagellates, choanocytes, and animal multicellularity.** *Invertebrate Biology.* 2004; **123**(1): 1–22.
Publisher Full Text

Maldonado M, Ribes M, van Duyl FC: **Nutrient fluxes through sponges.** *Adv Mar Biol.* 2012; **62**: 113–182.
PubMed Abstract | **Publisher Full Text**

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Olm MR, Brown CT, Brooks B, *et al.*: **dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication.** *ISME J.* 2017; **11**(12): 2864–2868.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Parks DH, Imelfort M, Skennerton CT, *et al.*: **CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.** *Genome Res.* 2015; **25**(7): 1043–55.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rühlemann MC, Wacker EM, Ellinghaus D, *et al.*: **MAGScoT: a fast, lightweight and accurate bin-refinement tool.** *Bioinformatics.* 2022; **38**(24): 5430–5433.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Rützler K, Díaz MC, van Soest RWM, *et al.*: **Diversity of sponge fauna in mangrove ponds, Pelican Cays, Belize.** *Atoll Res Bull.* 2000; **467**: 229–248.
Publisher Full Text

Seemann T: **Prokka: rapid prokaryotic genome annotation.** *Bioinformatics.* 2014; **30**(14): 2068–2069.
PubMed Abstract | **Publisher Full Text**

Sheerin E, Sampaio F, Oatley G, *et al.*: **Sanger Tree of Life HMW DNA Extraction: Automated MagAttract v.1.** *protocols.io.* 2023.
Publisher Full Text

Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *protocols.io.* 2023.
Publisher Full Text

Todorovic M, Sampaio F, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor®3 for PacBio HiFi.** *protocols.io.* 2023.
Publisher Full Text

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python**

pipeline for mitochondrial genome assembly from PacBio high fidelity reads. *BMC Bioinformatics.* 2023; **24**(1): 288.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–D515.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Vacelet J, Donadey C: **A new species of *Halisarca* (Porifera, Demospongiae) from the Caribbean, with remarks on the cytology and affinities of the genus.** In: Jones, W. C.(ed.) *European contribution to the taxonomy of sponges.* Publications of the Sherkin Island Marine Station, 1987; 5–12.
Reference Source

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.
Publisher Full Text

Wu YW, Tang YH, Tringe SG, *et al.*: **MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm.** *Microbiome.* 2014; **2**(1): 26.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 27 August 2025

https://doi.org/10.21956/wellcomeopenres.26779.r127598

✔ **Rodrigo Costa**
Universidade de Lisboa, Lisbon, Lisbon, Portugal

De Goeij et al report on the genome sequence and associated microbial metagenome of the marine sponge *Halisarca caerulea*. This species is a valuable resource within the Porifera owing to its unusual absence of mineral and proteinaceous skeletal structures, typical of other marine sponges. The genome could not be assembled at the chromosomal level and just four bacterial MAGs are reported, suggesting limitations in the retrieval of contig/scaffolds with high coverage. Nevertheless, making these datasets available, including RNA-seq data, will benefit the scientific community interested in sponge physiology, evolution, and symbiosis at large.

General comments:
Because of the stringency involved in the selection of MAGs to be reported, requiring MAGs to present rRNA genes, only four MAGs passed the quality thresholds used and, in some instances of the manuscript (particularly, in the abstract), this gives the impression that the *H. caerulea* microbiome is rather poor and constituted by a few bacterial species. In the minor/specific comments below, I suggest a few interventions to fix this issue, including modifications to Figure 5 that are needed to provide a more accurate perspective of microbiome diversity in *H. caerulea*. Then authors should ideally discuss, in the text (one-two sentences is enough), the extent to which microbial diversity observed in this study (including MAGs and metagenome contigs) compares with previous reports of microbial diversity in *H. caerulea* (e.g. Campana et al. 2024).

- Please revise the ms. throughout to highlight (e.g., using italics) scientific names with standing in nomenclature. This is valid for the sponge genus Halisarca, for the sponge species *H. caerulea*, and for all prokaryotic taxon names (from phylum to species) mentioned in the text.

Specific comments:
ABSTRACT:
“four binned bacterial genomes... were identified”.
In the context of the abstract, this gives the impression that the *H. caerulea* microbiome is composed of two species only. I suggest broadening this perspective here and adding a note of the family-level taxonomy (or even class or phylum level) of the *H. caerulea* microbiome, according

to the data provided in Figure 5 (more on this below).

BACKGROUND
2nd paragraph:
There is no mention of Figure 1C. Please correct this so that all items A through D are cited in the text in sequential order. Alternatively, simply refer to "Figure 1" in all instances where this figure is mentioned throughout the manuscript.
4th paragraph, first sentence:
Replace “classification” with “classification of *H. caerulea*” for better readability.

4th paragraph:
“…by the phyla Pseudomonadota… and Spirochaeota (Campana et al., 2024)”.
Considering that only two MAGs/phylotypes are reported, one could disclose a bit more precisely, here, what is the species-level (ASVs, single-species genome bins/MAGs etc), and family-level diversity known for this sponge. Then, in the “metagenome report” section, a note could be added on how microbiome diversity reported in this study compares with that reported by Campana et al and/or other studies.

6th paragraph
“primitive”: most likely “basal” and early-diverging” are better word choices.

GENOME SEQUENCE REPORT
1st paragraph, 1st line:
Replace “adult” with “adult specimen of” for better readability.

METAGENOME REPORT
1st paragraph, “(Figure 5)”: From Figure 5, as currently presented, it is not possible to identify the four binned genomes listed in Table 2. Please adjust the Figure accordingly or rephrase the sentence so it fits the display item(s) you are referring to.

Figure 5 (continued): There is no legend provided as an inlet within the figure to allow the visualization of different families (based on colours).

Where are the four assembled MAGs? Can they be located on this landscape (see comment above on this)?
As mentioned in my comment to the abstract, this figure can help discuss microbial diversity in *H. caerulea* in a broader context, beyond the two phylotypes represented by the MAGs, and in line with Campana et al (2024) regarding what is known in terms of microbial diversity in this sponge. Are your results similar to those reported by Campana et al? One or two sentences on the matter will be enough, in my opinion.

Legend to Figure 5: Add unit after “8,133,556”.

GENOME ANNOTATION REPORT
Add unit (bp) after “5,833.36”

METHODS – Sample acquisition

"A second specimen…was used…for RNA sequencing".
Perhaps this should be better contextualized in the genome report section, for improved readability. Only here the reader learns that RNA seq was employed. As RNA seq is apparently less frequent among genome notes, mentioning that dedicated RNA seq was performed while reporting the data in the sections above can be helpful.

Caption to Table 2. Replace "metagenomes" with "MAGs" or "genomes". You are listing individual genomes here.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Metagenomics, molecular microbial ecology, symbioses, microbiomes, blue biotechnology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 27 August 2025

https://doi.org/10.21956/wellcomeopenres.26779.r127599

Matsapume Detcharoen
Prince of Songkla University, Hat Yai, Thailand

This study reports a scaffold-level genome assembly of the sponge *Halisarca caerulea*. The genome is 195.70 Mb in length and distributed over 488 scaffolds. Genome annotation identified 26,722 protein-coding genes and 733 non-coding genes. Metagenome analysis classified four binned bacterial genomes related to typical sponge symbionts. Two of these belonged to the alphaproteobacterial order GM7ARS4, and two of the gammaproteobacterial family AqS2. These microbial clades are considered typical sponge symbionts and have previously been identified in

other sponge species. The authors acknowledge the absence of a chromosome-level assembly, achieving one would significantly improve genome completeness and contiguity. My comments would be consider a more in-depth analysis of the metagenome which can provide a fuller understanding of the diverse microbial community associated with *H. caerulea* (e.g., functional annotation).

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular ecology and symbiosis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.