



DATA NOTE

The genome sequence of the sand star, *Astropecten irregularis* (Pennant, 1777) [version 1; peer review: 2 approved]

Patrick Adkins¹, Joanna Harley¹, John Bishop¹,
Marine Biological Association Genome Acquisition Lab,
Darwin Tree of Life Barcoding collective,
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,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

¹The Marine Biological Association, Plymouth, England, UK

V1 First published: 07 Aug 2024, 9:431
<https://doi.org/10.12688/wellcomeopenres.22821.1>
Latest published: 07 Aug 2024, 9:431
<https://doi.org/10.12688/wellcomeopenres.22821.1>

Abstract

We present a genome assembly from an individual *Astropecten irregularis* (the sand star; Echinodermata; Asteroidea; Paxillosida; Astropectinidae). The genome sequence spans 475.80 megabases. Most of the assembly is scaffolded into 22 chromosomal pseudomolecules. The mitochondrial genome has also been assembled and is 16.34 kilobases in length.

Keywords

Astropecten irregularis, sand star, genome sequence, chromosomal, Paxillosida



This article is included in the [Tree of Life](#) gateway.

Open Peer Review

Approval Status

	1	2
version 1 07 Aug 2024	 view	 view

1. **Hiroshi Wada**, University of Tsukuba, Tsukuba, Japan
2. **Ferdinand Marlétaz** , University College London, London, UK

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

Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: **Adkins P:** Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Harley J:** Investigation, Resources; **Bishop J:** Investigation, Resources;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute [206194, <https://doi.org/10.35802/206194>] and the Darwin Tree of Life Discretionary Award [218328, <https://doi.org/10.35802/218328>]. *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

Copyright: © 2024 Adkins P *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Adkins P, Harley J, Bishop J *et al.* **The genome sequence of the sand star, *Astropecten irregularis* (Pennant, 1777) [version 1; peer review: 2 approved]** Wellcome Open Research 2024, 9:431 <https://doi.org/10.12688/wellcomeopenres.22821.1>

First published: 07 Aug 2024, 9:431 <https://doi.org/10.12688/wellcomeopenres.22821.1>

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Echinodermata; Eleutherozoa; Asterozoa; Asteroidea; Valvatacea; Paxillosoida; Astropectinidae; *Astropecten*; *Astropectin irregularis* (Pennant, 1777) (NCBI:txid55651).

Background

Astropectin irregularis, commonly referred to as a Sand Star, is a starfish of the family Astropectinidae. It has five somewhat short inflexible arms fringed strongly by marginal plates and spines. In life, its colour is variable with individuals being brown, yellowish brown, orange or pink, with variations between these colour morphs.

A. irregularis is found from Norway to Morocco and throughout the Mediterranean at depths from 10–1000 m (Hayward & Ryland, 2017). Common sublittorally around the British Isles, it inhabits soft substrata, particularly sand, where it partially or completely buries itself within the sediment, emerging to hunt (Freeman *et al.*, 2001). It is a voracious predator of infaunal invertebrates, in particular molluscs and crustaceans (Morin *et al.*, 1985). Its presence strongly influences its prey, to the extent of near exclusion of some prey species, as in the case of *Spisula subtruncata* (Muus, 1966). *A. irregularis* is reproductively active during the summer months (June/July) where it migrates inshore to feed and reproduce before retreating to deeper waters in the winter (Freeman, 1999; Freeman *et al.*, 2001). As with other starfish, *A. irregularis* releases gametes into the water column where they then fertilise. To maximise the success of this strategy and produce high densities of gametes, fertile *A. irregularis* aggregate together and synchronize spawning events (Freeman *et al.*, 2001; Grant & Tyler, 1986).

This genome presented here is the first of its kind for this species. We hope this will provide a valuable tool for further study into this species and echinoderms in general.

Genome sequence report

The genome of an adult *Astropecten irregularis* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of 22.66 Gb (gigabases) from 2.30 million reads, providing approximately 49-fold coverage. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data, which produced 159.66 Gbp from 1,057.37 million reads, yielding an approximate coverage of 336-fold. Specimen and sequencing information is summarised in Table 1.

Manual assembly curation corrected 283 missing joins or mis-joins and 87 haplotypic duplications, reducing the assembly length by 4.72% and the scaffold number by 30.74%, and increasing the scaffold N50 by 13.87%. The final assembly has a total length of 475.80 Mb in 186 sequence scaffolds with a scaffold N50 of 21.5 Mb (Table 2). The total count of gaps in the scaffolds is 786. 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. Most (98.18%)



Figure 1. Photograph of the *Astropecten irregularis* (eaAstIrre1) specimen used for genome sequencing.

of the assembly sequence was assigned to 22 chromosomal-level scaffolds. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 3). Chromosome 19 contains a heterozygous inversion between approximately 1.1–7.4Mb. 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.

The estimated Quality Value (QV) of the final assembly is 56.8 with *k*-mer completeness of 99.99%, and the assembly has a BUSCO v5.4.3 completeness of 97.3% (single = 96.6%, duplicated = 0.7%), using the metazoa_odb10 reference set (*n* = 954).

Metadata for specimens, BOLD barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly statistics are given at <https://links.tol.sanger.ac.uk/species/55651>.

Methods

Sample acquisition

A single *Astropectin irregularis* individual was collected offshore from site L4 in the western English Channel (latitude 50.25, longitude -4.23) on 2021-05-27. The specimen was removed from its habitat of broken shell and muddy sand using an Agassiz trawl deployed from the RV Sepia. It was then kept in seawater in a flow tank and brought to the Marine Biological Association where it was identified by Patrick Adkins.

The initial identification was verified by an additional DNA barcoding process according to the framework developed by Twyford *et al.* (2024). A small sample was dissected from the specimen and stored in ethanol, while the remaining parts of the specimen were shipped on dry ice to the Wellcome Sanger Institute (WSI). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also

Table 1. Specimen and sequencing data for *Astropecten irregularis*.

Project information			
Study title	Astropecten irregularis (sand star)		
Umbrella BioProject	PRJEB64718		
Species	<i>Astropecten irregularis</i>		
BioSample	SAMEA110449760		
NCBI taxonomy ID	55651		
Specimen information			
Technology	ToIID	BioSample accession	Organism part
PacBio long read sequencing	eaAstIrre1	SAMEA110450618	bodywall
Hi-C sequencing	eaAstIrre1	SAMEA110450618	bodywall
RNA sequencing	eaAstIrre1	SAMEA110450617	bodywall
Sequencing information			
Platform	Run accession	Read count	Base count (Gb)
Hi-C Illumina NovaSeq 6000	ERR11814107	1.06e+09	159.66
PacBio Sequel IIe	ERR11809141	2.30e+06	22.66
RNA Illumina NovaSeq 6000	ERR11814108	6.52e+07	9.85

used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding have been deposited on protocols.io (Beasley *et al.*, 2023).

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. In sample preparation, the eaAstIrre1 sample was weighed and dissected on dry ice (Jay *et al.*, 2023). For sample homogenisation, bodywall tissue was cryogenically disrupted using the Covaris cryoPREP® Automated Dry Pulverizer (Narváez-Gómez *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 with speed setting 30 (Todorovic *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation (Strickland *et al.*, 2023): in brief, the method employs 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 using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

RNA was extracted from bodywall tissue of eaAstIrre1 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.

Protocols developed by the WSI Tree of Life laboratory are publicly available on protocols.io (Denton *et al.*, 2023).

Sequencing

Pacific Biosciences HiFi circular consensus DNA sequencing libraries were constructed according to the manufacturers' instructions. Poly(A) RNA-Seq libraries were constructed using 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 6000 (RNA-Seq) instruments. Hi-C data were also generated from bodywall tissue of eaAstIrre1 using the Arima-HiC v2 kit. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on the Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

Assembly

The original assembly of HiFi reads was performed using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed with purge_dups (Guan *et al.*, 2020). Hi-C reads were further mapped with bwa-mem2 (Vasimuddin *et al.*, 2019) to the primary

Table 2. Genome assembly data for *Astropecten irregularis*, eaAstIrre1.1.

Genome assembly		
Assembly name	eaAstIrre1.1	
Assembly accession	GCA_963971285.1	
Accession of alternate haplotype	GCA_963971295.1	
Span (Mb)	475.80	
Number of contigs	973	
Contig N50 length (Mb)	0.9	
Number of scaffolds	186	
Scaffold N50 length (Mb)	21.5	
Longest scaffold (Mb)	40.85	
Assembly metrics*		Benchmark
Consensus quality (QV)	56.8	≥ 50
k-mer completeness	99.99%	≥ 95%
BUSCO**	C:97.3%[S:96.6%,D:0.7%], F:1.2%,M:1.5%,n:954	C ≥ 95%
Percentage of assembly mapped to chromosomes	98.18%	≥ 95%
Sex chromosomes	None	localised homologous pairs
Organelles	Mitochondrial genome: 16.34 kb	complete single alleles

* Assembly metric benchmarks are adapted from column VGP-2020 of “Table 1: Proposed standards and metrics for defining genome assembly quality” from [Rhie et al. \(2021\)](#).

** 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/Astropecten_irregularis/dataset/GCA_963971285.1/busco.

contigs, which were further scaffolded using the provided Hi-C data ([Rao et al., 2014](#)) in YaHS ([Zhou et al., 2023](#)) using the --break option. Scaffolded assemblies were evaluated using Gfastats ([Formenti et al., 2022](#)), BUSCO ([Manni et al., 2021](#)) and MERQURY.FK ([Rhie et al., 2020](#)).

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.

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline (article in preparation). Flat files and maps used in curation were generated in TreeVal ([Pointon et al., 2023](#)). Manual curation was primarily conducted using PretextView ([Harry, 2022](#)), with additional insights provided by JBrowse2 ([Diesh et al., 2023](#)) and HiGlass

([Kerpedjiev et al., 2018](#)). Scaffolds were visually inspected and corrected as described by [Howe et al. \(2021\)](#). Any identified contamination, missed joins, and mis-joins were corrected, and duplicate sequences were tagged and removed. The entire process is documented at <https://gitlab.com/wtsi-grit/rapid-curation> (article in preparation).

Evaluation of the final assembly

The final assembly was post-processed and evaluated with the three Nextflow ([Di Tommaso et al., 2017](#)) DSL2 pipelines “sanger-tol/readmapping” ([Surana et al., 2023a](#)), “sanger-tol/genomenote” ([Surana et al., 2023b](#)), and “sanger-tol/blobtoolkit” ([Muffato et al., 2024](#)). The pipeline sanger-tol/readmapping aligns the Hi-C reads with bwa-mem2 ([Vasimuddin et al., 2019](#)) and combines the alignment files with SAMtools ([Danecek et al., 2021](#)). The sanger-tol/genomenote pipeline transforms the Hi-C alignments into a contact map with BEDTools ([Quinlan & Hall, 2010](#)) and the Cooler tool suite ([Abdennur & Mirny, 2020](#)), which is then visualised with HiGlass ([Kerpedjiev et al., 2018](#)). It also provides statistics about

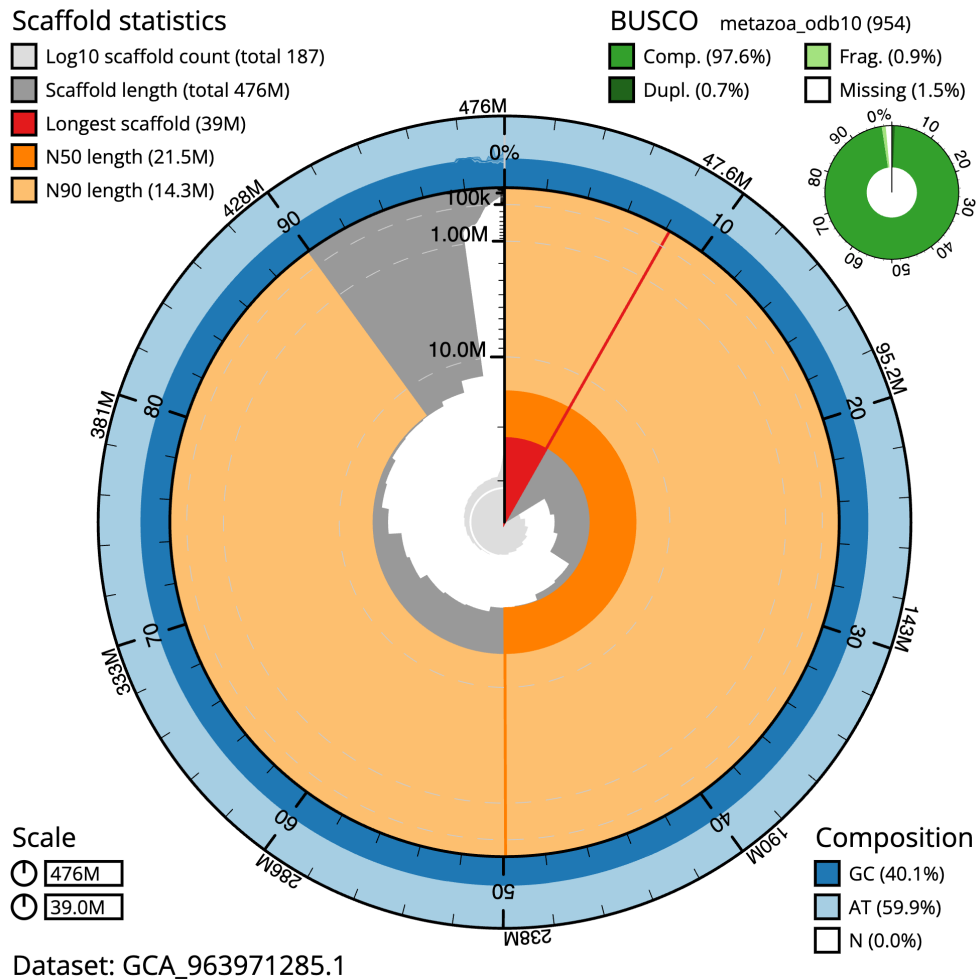


Figure 2. Genome assembly of *Astropecten irregularis*, eaAstIrre1.1: metrics. The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 475,837,774 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (39,043,871 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (21,462,565 and 14,331,079 bp), respectively. The pale grey spiral shows the cumulative scaffold 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/Astropecten_irregularis/dataset/GCA_963971285.1/snail.

the assembly with the NCBI datasets (Sayers *et al.*, 2024) report, computes k -mer completeness and QV consensus quality values with FastK and MERQURY.FK, and a completeness assessment with BUSCO (Manni *et al.*, 2021).

The sanger-tol/blobtoolkit pipeline is a Nextflow port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads with SAMtools and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoAT database (Challis *et al.*, 2023) to identify all matching BUSCO lineages

to run BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineage, the pipeline aligns the BUSCO genes to the Uniprot Reference Proteomes database (Bateman *et al.*, 2023) with DIAMOND (Buchfink *et al.*, 2021) blastp. The genome is also split into chunks according to the density of the BUSCO genes from the closest taxonomically lineage, and each chunk is aligned to the Uniprot Reference Proteomes database with DIAMOND blastx. Genome sequences that have no hit are then chunked with seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). All those outputs are combined with the blobtools suite into a blobdir for visualisation.

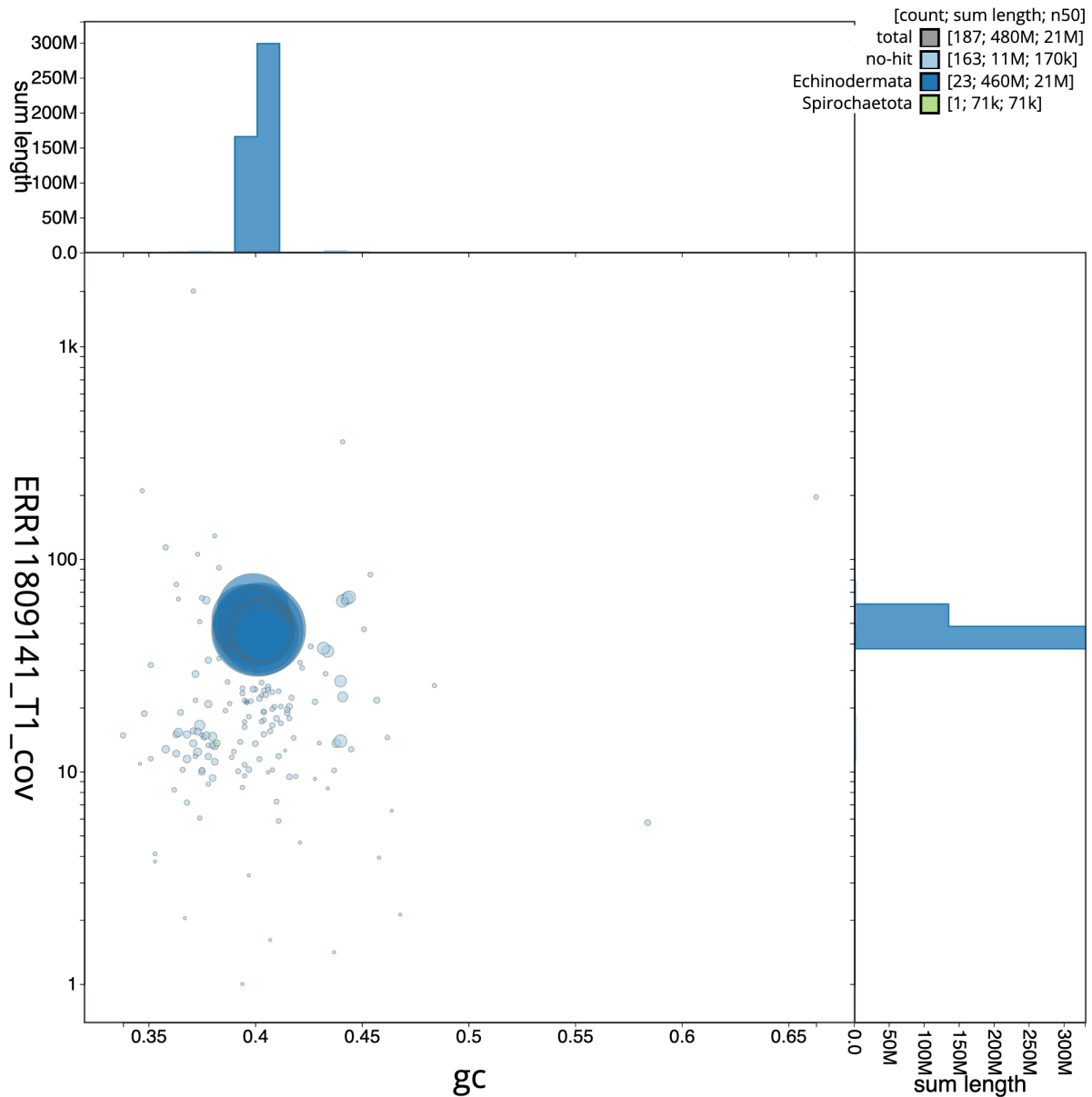


Figure 3. Genome assembly of *Astropecten irregularis*, eaAstIrre1.1. Blob plot of base coverage in ERR11809141 against GC proportion for sequences in assembly GCA_963971285.1. Sequences are coloured by phylum. Circles are sized in proportion to sequence length. Histograms show the distribution of sequence length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/Astropecten_irregularis/dataset/GCA_963971285.1/blob.

The genome assembly and evaluation pipelines were developed using the nf-core tooling (Ewels *et al.*, 2020), use MultiQC (Ewels *et al.*, 2016), and make extensive use of the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), and the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

Table 4 contains a list of relevant software tool versions and sources.

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the

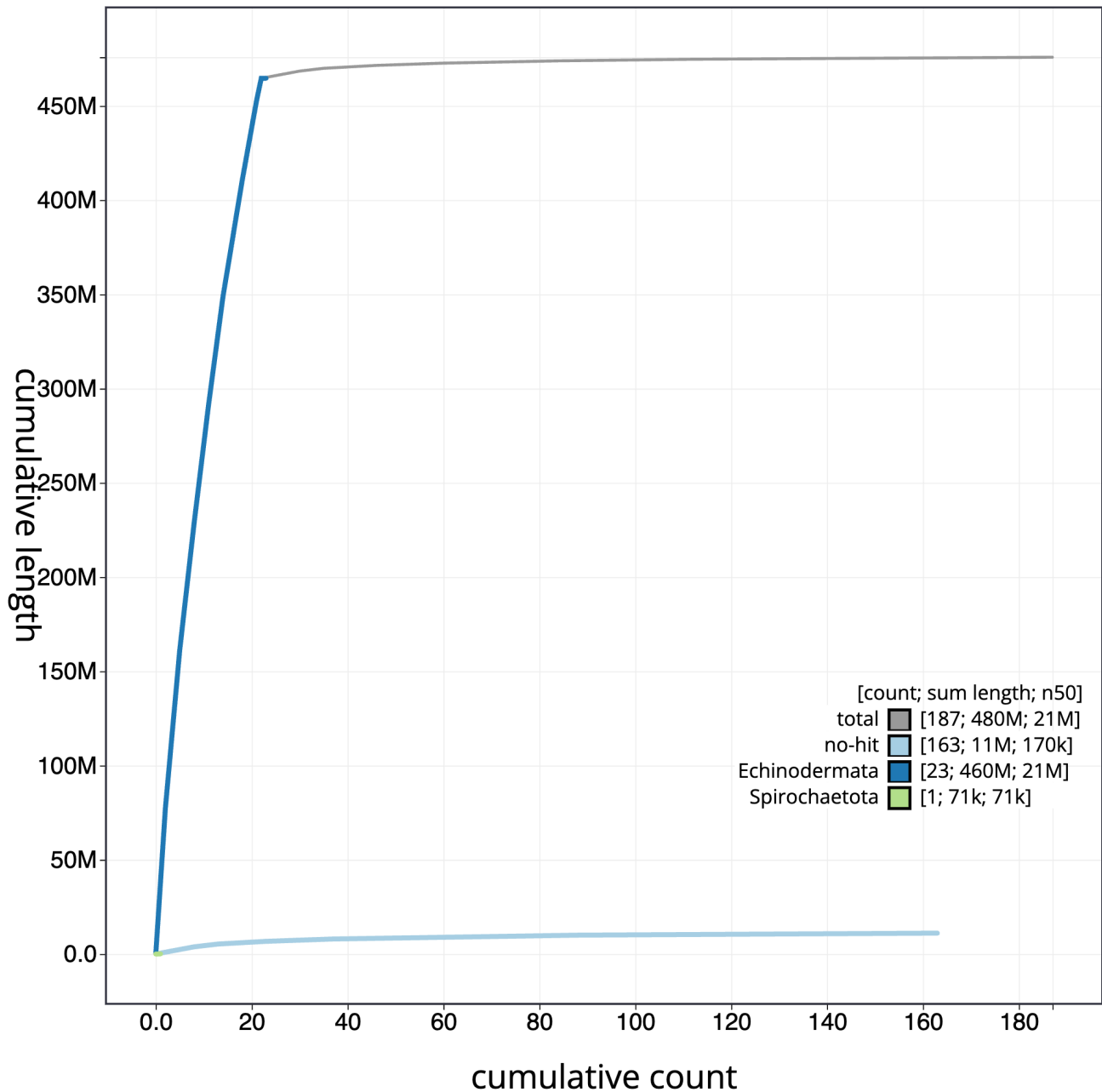


Figure 4. Genome assembly of *Astropecten irregularis* eaAstIrre1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all sequences. Coloured lines show cumulative lengths of sequences assigned to each phylum using the buscones taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/Astropecten_irregularis/dataset/GCA_963971285.1/cumulative.

‘Darwin Tree of Life Project Sampling Code of Practice’, which can be found in full on the Darwin Tree of Life website [here](#). By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within

this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project.

Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature

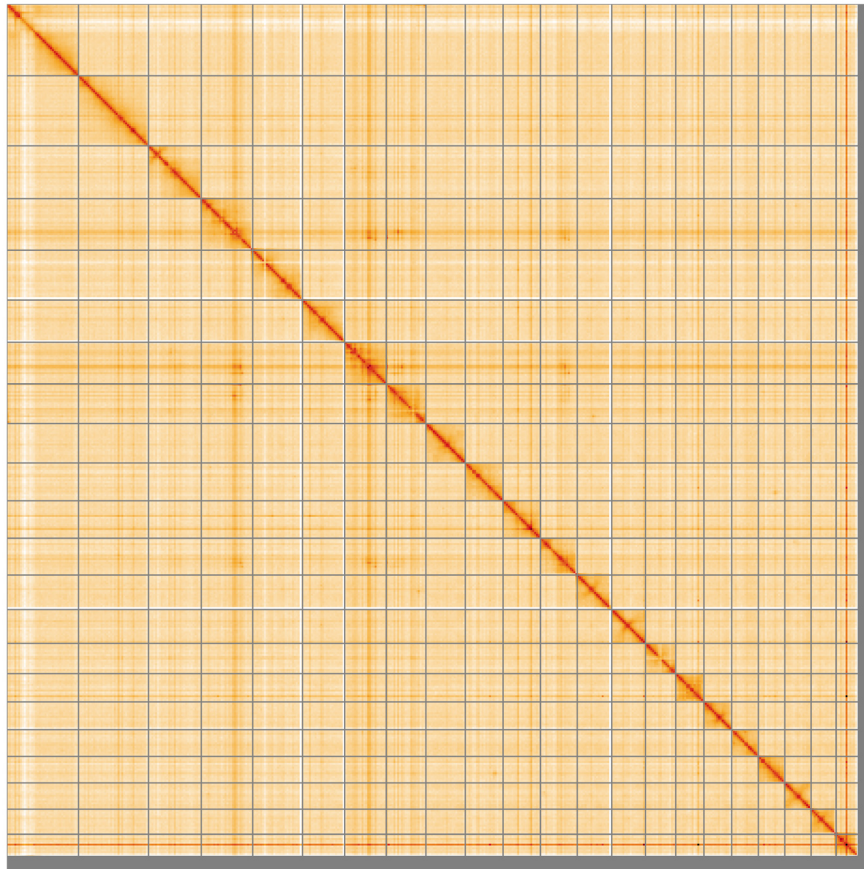


Figure 5. Genome assembly of *Astropecten irregularis* eaAstIrre1.1: Hi-C contact map of the eaAstIrre1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=TvB534skT0qAf4Af_6H0qg.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Astropecten irregularis*, eaAstIrre1.

INSDC accession	Name	Length (Mb)	GC%
OZ020244.1	1	39.04	40.0
OZ020245.1	2	38.23	40.0
OZ020246.1	3	28.81	40.0
OZ020247.1	4	28.06	40.0
OZ020248.1	5	27.23	40.0
OZ020249.1	6	22.69	40.0
OZ020250.1	7	23.0	40.0
OZ020251.1	8	21.66	39.5
OZ020252.1	9	21.46	40.0
OZ020253.1	10	20.61	40.0

INSDC accession	Name	Length (Mb)	GC%
OZ020254.1	11	20.41	40.0
OZ020255.1	12	20.07	40.0
OZ020256.1	13	18.85	40.0
OZ020257.1	14	18.44	40.0
OZ020258.1	15	16.46	40.5
OZ020259.1	16	15.41	40.0
OZ020260.1	17	15.15	40.5
OZ020261.1	18	14.44	40.5
OZ020262.1	19	14.56	40.5
OZ020263.1	20	14.33	40.0
OZ020264.1	21	13.68	40.5
OZ020265.1	22	12.08	40.5
OZ020266.1	MT	0.02	37.0

Table 4. 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
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.5-r587	https://github.com/chhylp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84aa44357826c0b6753eb28de	https://github.com/higlass/higlass
Mercury.FK	d00d98157618f4e8d1a9190026b19b471055b22e	https://github.com/thegenemyers/MERQUERY.FK
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14, 1.17, and 1.18	https://github.com/MultiQC/MultiQC
NCBI Datasets	15.12.0	https://github.com/ncbi/datasets
Nextflow	23.04.0-5857	https://github.com/nextflow-io/nextflow
PretextView	0.2	https://github.com/sanger-tol/PretextView
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
samtools	1.16.1, 1.17, and 1.18	https://github.com/samtools/samtools
sanger-tol/ascc	-	https://github.com/sanger-tol/ascc
sanger-tol/genomenote	1.1.1	https://github.com/sanger-tol/genomenote
sanger-tol/readmapping	1.2.1	https://github.com/sanger-tol/readmapping
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	1.2a.2	https://github.com/c-zhou/yahs

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

Data availability

European Nucleotide Archive: *Astropecten irregularis* (sand star). Accession number PRJEB64718; <https://identifiers.org/ena.embl/PRJEB64718> (Wellcome Sanger Institute, 2024). The genome sequence is released openly for reuse. The *Astropecten irregularis* genome sequencing initiative is part of the Darwin Tree of Life (DToL) 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 Marine Biological Association Genome Acquisition Lab are listed here: <https://doi.org/10.5281/zenodo.8382513>.

Members of the Darwin Tree of Life Barcoding collective are listed here: <https://doi.org/10.5281/zenodo.12158331>.

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: <https://doi.org/10.5281/zenodo.12162482>.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: <https://doi.org/10.5281/zenodo.12165051>.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: <https://doi.org/10.5281/zenodo.12160324>.

Members of the Tree of Life Core Informatics collective are listed here: <https://doi.org/10.5281/zenodo.12205391>.

Members of the Darwin Tree of Life Consortium are listed here: <https://doi.org/10.5281/zenodo.4783558>.

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](#)
- 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](#)
- Altschul SF, Gish W, Miller W, et al.: **Basic local alignment search tool.** *J Mol Biol.* 1990; **215**(3): 403–410. [PubMed Abstract](#) | [Publisher Full Text](#)
- Bateman A, Martin MJ, Orchard S, et al.: **UniProt: the universal protein knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Beasley J, Uhl R, Forrest LL, et al.: **DNA barcoding SOPs for the Darwin Tree of Life project.** *Protocols.io.* 2023. [Publisher Full Text](#)
- Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at tree-of-life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Challis R, Kumar S, Sotero-Caio C, et al.: **Genomes on a Tree (GoAT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free 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](#)
- 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](#)
- Crowley L, Allen H, Barnes I, et al.: **A sampling strategy for genome sequencing the British terrestrial arthropod fauna [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 123. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- da Veiga Leprevost F, Grüning BA, Alves Aflitos S, et al.: **BioContainers: an open-source and community-driven framework for software standardization.** *Bioinformatics.* 2017; **33**(16): 2580–2582. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Danecek P, Bonfield JK, Liddle J, et al.: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Denton A, Yatsenko H, Jay J, et al.: **Sanger Tree of Life wet laboratory protocol collection V.1.** *Protocols.io.* 2023. [Publisher Full Text](#)
- Di Tommaso P, Chatzou M, Floden EW, et al.: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319. [PubMed Abstract](#) | [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](#)

- Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Freeman SM: **The ecology of *Astropecten irregularis* and its potential role as a benthic predator in a soft-sediment community.** Bangor University (United Kingdom), 1999.
[Reference Source](#)
- Freeman SM, Richardson CA, Seed R: **Seasonal abundance, spatial distribution, spawning and growth of *Astropecten irregularis* (Echinodermata: Asteroidea).** *Estuar Coast Shelf Sci.* 2001; **53**(1): 39–49.
[PubMed Abstract](#) | [Free Full Text](#)
- Grant A, Tyler PA: **An analysis of the reproductive pattern in the sea star *Astropecten irregularis* (Pennant) from the Bristol Channel.** *Int J Invertebr Reprod Dev.* 1986; **9**(3): 345–361.
[PubMed Abstract](#) | [Free Full Text](#)
- Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free 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; [Accessed 19 October 2022].
[Reference Source](#)
- Hayward PJ, Ryland JS: **Handbook of the Marine Fauna of North-West Europe.** 2nd ed. Oxford: Oxford University Press, 2017.
[Reference Source](#)
- Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): giaa153.
[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.
[PubMed Abstract](#) | [Free 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](#)
- Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Manni M, Berkeley MR, Seppely 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](#)
- Merkel D: **Docker: lightweight linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2.
[Reference Source](#)
- Morin JG, Kastendiek JE, Harrington A, *et al.*: **Organization and patterns of interactions in a subtidal community on an exposed coast.** *Mar Ecol Prog Ser.* 1985; **27**: 163–185.
[Reference Source](#)
- Muffato M, Butt Z, Challis R, *et al.*: **Sanger-tol/blobtoolkit: v0.3.0 – poliwig.** 2024.
[PubMed Abstract](#) | [Free Full Text](#)
- Muus K: **A quantitative 3-year survey on the meiofauna of known macrofauna communities on the Øresund.** *Veroff Institute Meeresforschungen Bremerh Supplement.* 1966; **3**: 289–292.
[Reference Source](#)
- Narváez-Gómez JP, Mbye H, Oatley G, *et al.*: **Sanger Tree of Life sample homogenisation: covaris cryoPREP® automated Dry Pulverizer V.1.** *Protocols.io.* 2023.
[PubMed Abstract](#) | [Free Full Text](#)
- Pointon DL, Eagles W, Sims Y, *et al.*: **Sanger-tol/treeval v1.0.0 – ancient Atlantis.** 2023.
[PubMed Abstract](#) | [Free Full Text](#)
- Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–842.
[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, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rhie A, Walenz BP, Koren S, *et al.*: **Merquy: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sayers EW, Cavanaugh M, Clark K, *et al.*: **GenBank 2024 update.** *Nucleic Acids Res.* 2024; **52**(D1): D134–D137.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sheerin E, Sampaio F, Oatley G, *et al.*: **Sanger Tree of Life HMW DNA extraction: automated MagAttract v.1.** *Protocols.io.* 2023.
[PubMed Abstract](#) | [Free Full Text](#)
- Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *Protocols.io.* 2023.
[PubMed Abstract](#) | [Free Full Text](#)
- Surana P, Muffato M, Qi G: **Sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo.* 2023a.
[PubMed Abstract](#) | [Free Full Text](#)
- Surana P, Muffato M, Sadasivan Baby C: **Sanger-tol/genomenote (v1.0.dev).** *Zenodo.* 2023b.
[PubMed Abstract](#) | [Free Full Text](#)
- Todorovic M, Sampaio F, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode megaruptor® 3 for pacBio HiFi.** *protocols.io.* 2023.
[PubMed Abstract](#) | [Free Full Text](#)
- Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer review: awaiting peer review].** *Wellcome Open Res.* 2024; **9**: 339.
[PubMed Abstract](#) | [Free Full Text](#)
- Uliano-Silva M, Ferreira JGRN, Krashenninnikova 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](#)
- 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.
[PubMed Abstract](#) | [Free Full Text](#)
- Wellcome Sanger Institute: **The genome sequence of the sand star, *Astropecten irregularis* (Pennant, 1777).** European Nucleotide Archive. [dataset], accession number PRJEB64718, 2024.
- 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 15 October 2024

<https://doi.org/10.21956/wellcomeopenres.25127.r95804>

© 2024 Marlétaz F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Ferdinand Marlétaz 

Centre for Life's Origin & Evolution, Department of Ecology, Evolution & Environment, University College London, London, UK

This report described the sequencing and assembly of the sea star *Astropecten irregularis* at chromosome scale. It appears accurate and well-described. I will reiterate some usual comments I make on such genome notes: the HiC contact map appears difficult to read, and the HiC protocol is not sufficiently detailed. Also, the sex of the specimen used for assembly is not specified, it is a pity.

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?

Partly

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, phylogeny, Zoology

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 02 September 2024

<https://doi.org/10.21956/wellcomeopenres.25127.r93590>

© 2024 Wada H. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Hiroshi Wada

University of Tsukuba, Tsukuba, Japan

This is a valuable report of the genome of *A. irregularis*. Because I am not an expert of the genome analysis, I make comments from the aspect of echinoderm biology. Technical issues should be judged by other experts of genome analysis.

But actually, although the authors provided some biological information in the Background section, they describe no implication they found from the genome. I like to see some biological implication the authors draw from the genome. Why do they like to sequence the genome of this species?

Some information should be provided for the individual the DNA was extracted. Is the individual male or female? Which tissue DNA was extracted from?

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: Embryology and evolutionary biology

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.
