DATA NOTE

The genome sequence of a mollusc, *Azorinus chamasolen* (da Costa, 1778)

[version 1; peer review: awaiting peer review]

Chris Fletcher[1], Crispin Little[2],
Natural History Museum 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

[1]Natural History Museum, London, England, UK
[2]University of Leeds, Leeds, England, UK

v1 **First published:** 15 Jan 2025, **10**:16
https://doi.org/10.12688/wellcomeopenres.23457.1
Latest published: 15 Jan 2025, **10**:16
https://doi.org/10.12688/wellcomeopenres.23457.1

Abstract

We present a genome assembly from a specimen of the mollusc, *Azorinus chamasolen* (Mollusca; Bivalvia; Cardiida; Solecurtidae). The genome sequence has a total length of 1,723.50 megabases. Most of the assembly (99.33%) is scaffolded into 19 chromosomal pseudomolecules. The mitochondrial genome has also been assembled and is 17.14 kilobases in length.

Keywords

Azorinus chamasolen, mollusc, genome sequence, chromosomal, Cardiida

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

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: Fletcher C: Investigation, Resources; **Little C**: 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.

How to cite this article: Fletcher C, Little C, Natural History Museum Genome Acquisition Lab *et al.* **The genome sequence of a mollusc, *Azorinus chamasolen* (da Costa, 1778) [version 1; peer review: awaiting peer review]** Wellcome Open Research 2025, **10**:16 https://doi.org/10.12688/wellcomeopenres.23457.1

First published: 15 Jan 2025, **10**:16 https://doi.org/10.12688/wellcomeopenres.23457.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Spiralia; Lophotrochozoa; Mollusca; Bivalvia; Autobranchia; Heteroconchia; Euheterodonta; Imparidentia; Neoheterodontei; Cardiida; Tellinoidea; Solecurtidae; *Azorinus*; *Azorinus chamasolen* (da Costa, 1778) (NCBI:txid2922058).

Background

The genome of *Azorinus chamasolen* was sequenced as part of the Darwin Tree of Life Project, a collaborative effort to sequence all named eukaryotic species in the Atlantic Archipelago of Britain and Ireland. Here we present a chromosomally complete genome sequence for *Azorinus chamasolen*, based on a specimen from Mayflower Marina, England, United Kingdom.

Genome sequence report

The genome of *Azorinus chamasolen* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of 62.75 Gb (gigabases) from 5.96 million reads, providing an estimated 35-fold coverage. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data, which produced 119.21 Gb from 789.48 million reads, yielding an approximate coverage of 69-fold. Specimen and sequencing details are summarised in Table 1.

Assembly errors were corrected by manual curation, including 94 missing joins or mis-joins and 63 haplotypic duplications. This reduced the assembly length by 2.32% and the scaffold number by 40.65%. The final assembly has a total length of 1,723.50 Mb in 145 sequence scaffolds, with 1,773 gaps, and a scaffold N50 of 97.7 Mb (Table 2).

The snail plot in Figure 2 provides a summary of the assembly statistics, indicating the distribution of scaffold lengths and other assembly metrics. Figure 3 shows the distribution of scaffolds by GC proportion and coverage. Figure 4 presents a cumulative assembly plot, with separate curves representing different scaffold subsets assigned to various phyla, illustrating the completeness of the assembly.

Most of the assembly sequence (99.33%) was assigned to 19 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by the Hi-C data, are named in order of size (Figure 5; Table 3). During manual curation it was noted that some telomeric repeat sequences could not be uniquely assigned to a chromosomal location.

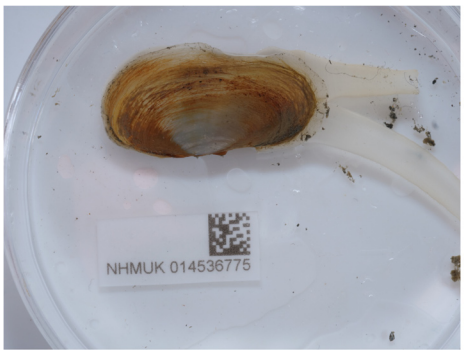

Figure 1. Photograph of the *Azorinus chamasolen* (xbAzoCham1) specimen used for genome sequencing.

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, and as a separate fasta file with accession OY755254.1.

The final assembly has a Quality Value (QV) of 60.5 and *k*-mer completeness of 98.56% for the combined assemblies. BUSCO (v5.4.3) analysis using the mollusca_odb10 reference set (n = 5,295) indicated a completeness score of 81.1% (single = 79.6%, duplicated = 1.5%).

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/2922058.

Methods

Sample acquisition and DNA barcoding

An adult specimen of *Azorinus chamasolen* (specimen ID NHMUK014536775, ToLID xbAzoCham1) was collected from Mayflower Marina, England, United Kingdom (latitude 50.36, longitude –4.17) on 2021-06-24. The specimen was collected by Chris Fletcher (Natural History Museum) and identified by Crispin Little (University of Leeds) and preserved in 80% ethanol.

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 specimens and stored in ethanol, while the remaining parts 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 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 and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io (Denton *et al.*, 2023b). The xbAzoCham1 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023), and tissue was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a).

Table 1. Specimen and sequencing data for *Azorinus chamasolen*.

Project information			
Study title	Azorinus chamasolen		
Umbrella BioProject	PRJEB62728		
Species	*Azorinus chamasolen*		
BioSample	SAMEA110043170		
NCBI taxonomy ID	2922058		
Specimen information			
Technology	**ToLID**	**BioSample accession**	**Organism part**
PacBio long read sequencing	xbAzoCham1	SAMEA14452984	Other somatic tissue
Hi-C sequencing	xbAzoCham1	SAMEA14452987	Mollusc foot
Sequencing information			
Platform	**Run accession**	**Read count**	**Base count (Gb)**
Hi-C Illumina NovaSeq 6000	ERR11526204	7.89e+08	119.21
PacBio Sequel IIe	ERR11512312	9.63e+05	11.34
PacBio Sequel IIe	ERR11512314	1.92e+06	23.68
PacBio Sequel IIe	ERR11512315	8.95e+05	9.09
PacBio Sequel IIe	ERR11512313	2.19e+06	18.64

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, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Strickland *et al.*, 2023). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. The fragment size distribution was evaluated by running the sample on the FemtoPulse system.

Hi-C preparation

Tissue from the foot of the xbAzoCham1 sample was processed at the WSI Scientific Operations core, using the Arima-HiC v2 kit. In brief, frozen tissue (stored at –80 °C) was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde. After crosslinking, the tissue was homogenised using the Diagnocine Power Masher-II and BioMasher-II tubes and pestles. Following the kit manufacturer's instructions, crosslinked DNA was digested using a restriction enzyme master mix. The 5'-overhangs were then filled in and labelled with biotinylated nucleotides and proximally ligated. An overnight incubation was carried out for enzymes to digest remaining proteins and for crosslinks to reverse. A clean up was performed with SPRIselect beads prior to library preparation.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Pacific Biosciences HiFi circular consensus DNA sequencing libraries were prepared using the PacBio Express Template Preparation Kit v2.0 (Pacific Biosciences, California, USA) as per the manufacturer's instructions. The kit includes the reagents required for removal of single-strand overhangs, DNA damage repair, end repair/A-tailing, adapter ligation, and nuclease treatment. Library preparation also included a library purification step using AMPure PB beads (Pacific Biosciences, California, USA) and size selection step to remove templates <3kb using AMPure PB modified SPRI. DNA concentration was quantified using the Qubit Fluorometer v2.0 and Qubit HS Assay Kit and the final library fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument and 165kb gDNA and 55kb BAC analysis kit. Samples were sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was between 40–135 pM. The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, as well as perform primary and secondary analysis of the data upon completion.

For Hi-C library preparation, DNA was fragmented to a size of 400 to 600 bp using a Covaris E220 sonicator. The DNA was then enriched, barcoded, and amplified using the NEBNext

Table 2. Genome assembly data for *Azorinus chamasolen*, xbAzoCham1.1.

Genome assembly		
Assembly name	xbAzoCham1.1	
Assembly accession	GCA_963576725.1	
Accession of alternate haplotype	*GCA_963576665.1*	
Span (Mb)	1,723.50	
Number of contigs	1,919	
Number of scaffolds	145	
Longest scaffold (Mb)	122.08	
Assembly metrics*		***Benchmark***
Contig N50 length (Mb)	1.7	*≥ 1 Mb*
Scaffold N50 length (Mb)	97.7	*= chromosome N50*
Consensus quality (QV)	60.5	*≥ 40*
k-mer completeness	primary: 71.25%; alternate: 70.48%; combined: 98.56%	*≥ 95%*
BUSCO v5.4.3 lineage: mollusca_odb10	C:81.1%[S:79.6%,D:1.5%], F:4.1%,M:14.8%,n:5,295	*S > 90%, D < 5%*
Percentage of assembly mapped to chromosomes	99.33%	*≥ 90%*
Organelles	Mitochondrial genome: 17.14 kb	*complete single alleles*

* Assembly metric benchmarks are adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024.

** BUSCO scores based on the mollusca_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/Azorinus_chamasolen/dataset/GCA_963576725.1/busco.

Ultra II DNA Library Prep Kit following manufacturers' instructions. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on an Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

Assembly

The HiFi reads were first assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). The Hi-C reads were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). The contigs were further scaffolded using the provided Hi-C data (Rao *et al.*, 2014) in YaHS (Zhou *et al.*, 2023) using the --break option. The 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 curation 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 using 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 readmapping pipeline aligns the Hi-C reads using bwa-mem2 (Vasimuddin *et al.*, 2019) and combines the alignment files with SAMtools (Danecek *et al.*, 2021). The genomenote pipeline converts the Hi-C

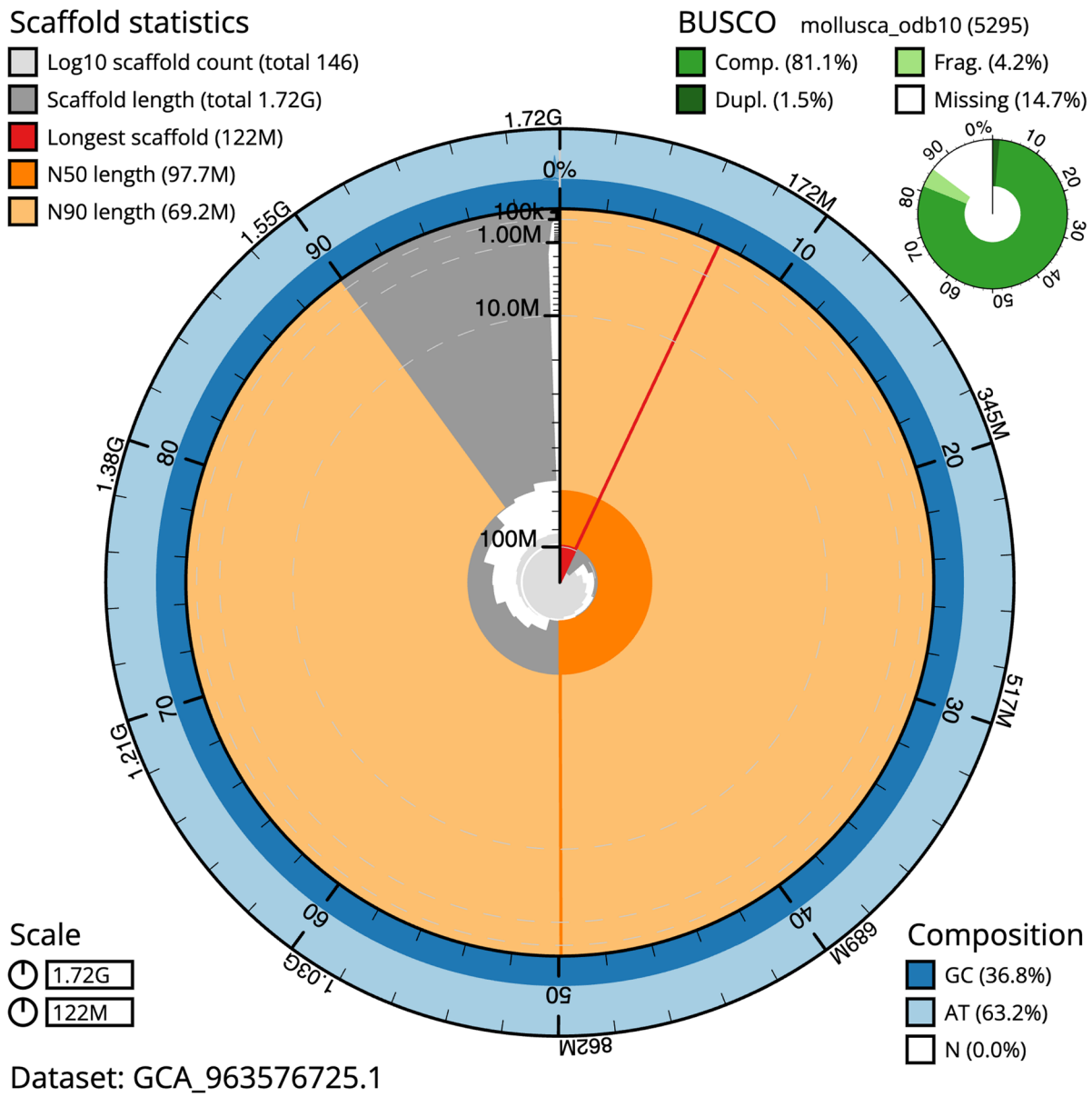

Figure 2. Genome assembly of *Azorinus chamasolen*, xbAzoCham1.1: metrics. The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the mollusca_odb10 set is shown at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963576725.1/dataset/GCA_963576725.1/snail.

alignments into a contact map using BEDTools (Quinlan & Hall, 2010) and the Cooler tool suite (Abdennur & Mirny, 2020). The contact map is visualised in HiGlass (Kerpedjiev *et al.*, 2018). This pipeline also computes *k*-mer completeness and QV consensus quality values with FastK and MERQURY.FK, and runs BUSCO (Manni *et al.*, 2021) to assess completeness.

The blobtoolkit pipeline is a Nextflow port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads in 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

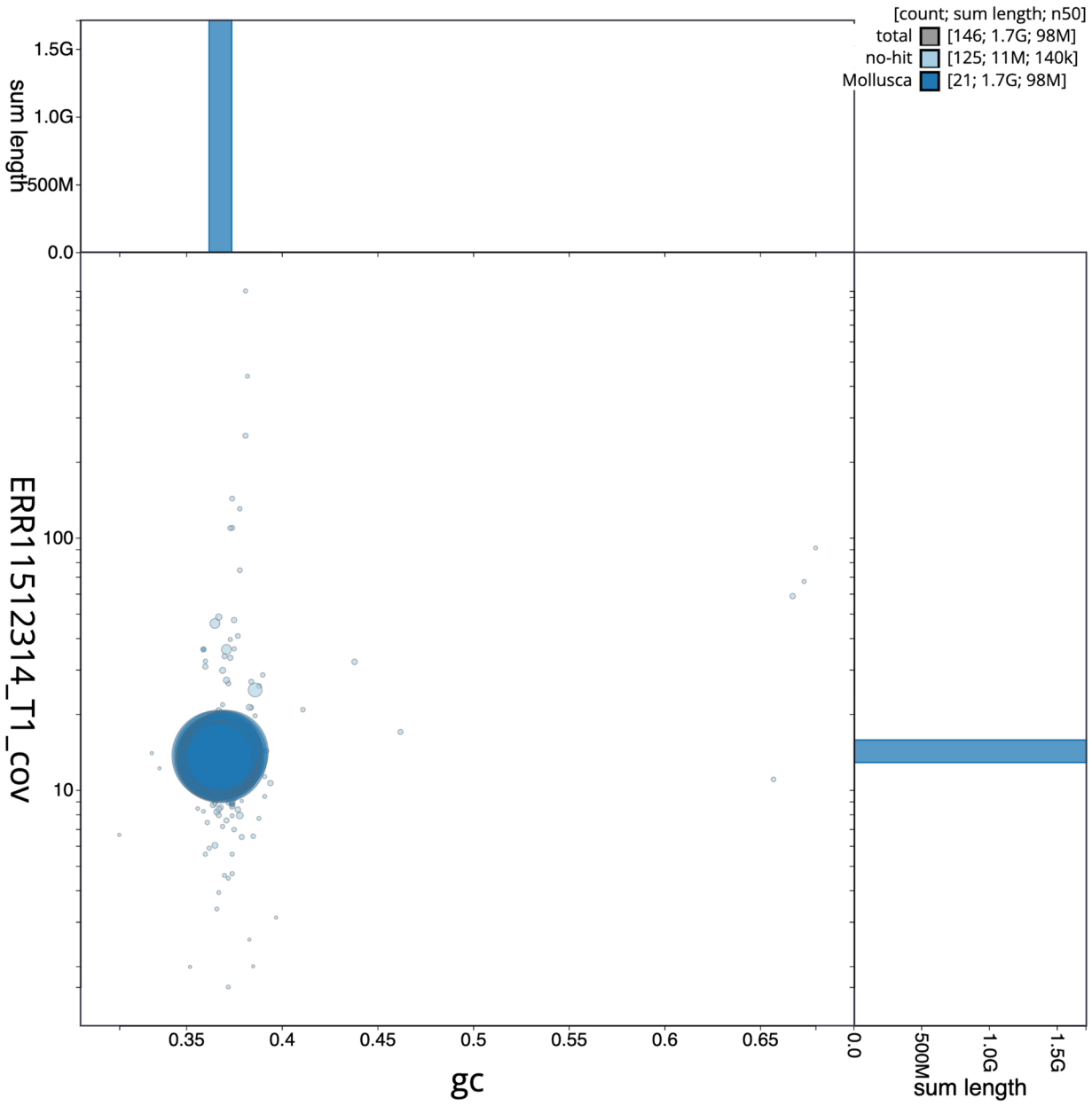

Figure 3. Genome assembly of *Azorinus chamasolen*: Blot plot of base coverage in the raw data against GC proportion for sequences in xbAzoCham1.1. Sequences are coloured by phylum. Circles are sized in proportion to sequence length. Histograms show the distribution of sequence length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963576725.1/dataset/GCA_963576725.1/blob.

(Manni *et al.*, 2021). For the three domain-level BUSCO lineages, 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 taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Genome sequences without a hit are chunked with seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The blobtools suite combines all these outputs into a blobdir for visualisation.

The genome assembly and evaluation pipelines were developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC

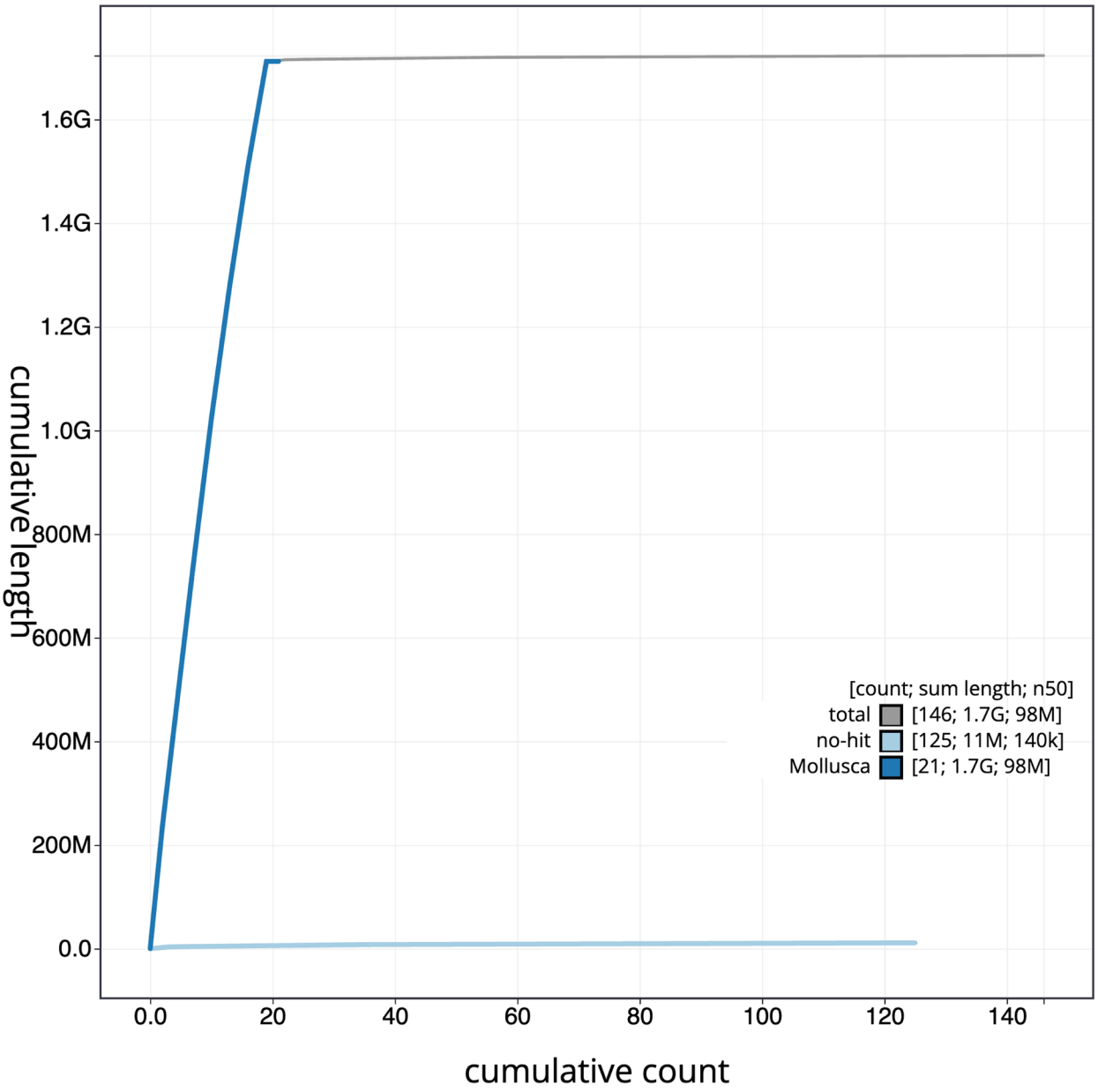

Figure 4. Genome assembly of *Azorinus chamasolen* xbAzoCham1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963576725.1/dataset/GCA_963576725.1/cumulative.

(Ewels *et al.*, 2016), relying on the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), as well as 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 **'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

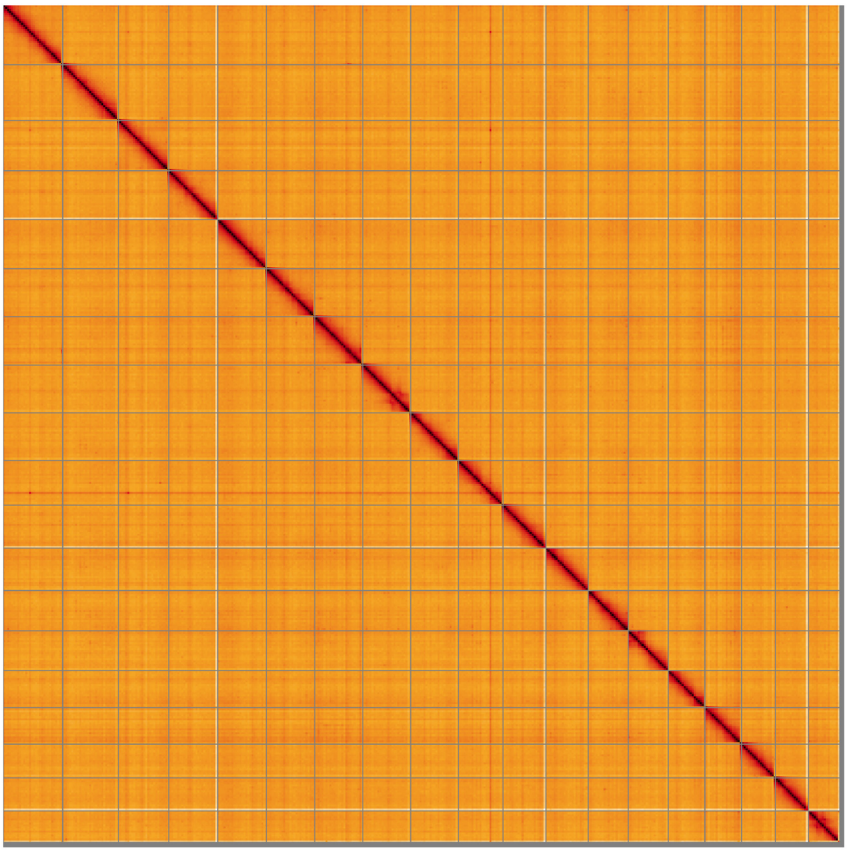

Figure 5. Genome assembly of *Azorinus chamasolen* xbAzoCham1.1: Hi-C contact map of the xbAzoCham1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=DUjPZ72NRw6BmRjSOVOkgA.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Azorinus chamasolen*, xbAzoCham1.

INSDC accession	Name	Length (Mb)	GC%
OY755235.1	1	122.08	37.0
OY755236.1	2	114.09	36.5
OY755237.1	3	102.72	36.5
OY755238.1	4	100.06	37.0
OY755239.1	5	100.06	37.0
OY755240.1	6	98.58	37.0
OY755241.1	7	98.54	37.0
OY755242.1	8	97.72	36.5
OY755243.1	9	97.67	36.5
OY755244.1	10	91.08	37.0
OY755245.1	11	87.99	37.0
OY755246.1	12	86.79	37.0
OY755247.1	13	81.91	37.0
OY755248.1	14	81.4	36.5
OY755249.1	15	75.54	37.0
OY755250.1	16	75.07	37.0
OY755251.1	17	69.16	37.0
OY755252.1	18	67.49	37.0
OY755253.1	19	64.14	36.5
OY755254.1	MT	0.02	38.5

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	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84 aa44357826c0b6753eb28de	https://github.com/higlass/higlass
Merqury.FK	d00d98157618f4e8d1a9190026b19b471055b22e	https://github.com/thegenemyers/MERQURY.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

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 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: *Azorinus chamasolen*. Accession number PRJEB62728; https://identifiers.org/ena.embl/PRJEB62728. The genome sequence is released openly for reuse. The *Azorinus chamasolen* 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 Natural History Museum Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.12159242.

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; [Accessed 25 June 2024].
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, Oatley G, Cornwell C, *et al.*: **Sanger Tree of Life sample homogenisation: PowerMash.** *protocols.io.* 2023a.
Publisher Full Text

Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life wet laboratory protocol collection V.1.** *protocols.io.* 2023b.
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

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

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.
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.
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**

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, 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**

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2, [Accessed 2 April 2024].
Reference Source

Muffato M, Butt Z, Challis R, *et al.*: **sanger-tol/blobtoolkit: v0.3.0 – Poliwag.** 2024.
Publisher Full Text

Pointon DL, Eagles W, Sims Y, *et al.*: **sanger-tol/treeval v1.0.0 – Ancient Atlantis.** 2023.
Publisher 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.*: **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**

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

Surana P, Muffato M, Qi G: **Sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo.* 2023a.
Publisher Full Text

Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev).** *Zenodo.* 2023b.
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

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: 2 approved].** *Wellcome Open Res.* 2024; **9**: 339.
PubMed Abstract | **Publisher Full Text** | **Free 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**

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

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**