# The genome sequence of the {{ COMMON\_NAME }}, *{{ GENUS\_SPECIES }}* {{ TAXONOMY\_AUTHORITY }}

## Authors

{{ AUTHORS }}, $INTRO\_WRITER, {{ GAL }} Genome Acquisition Lab, {{ HIC\_GAL }} 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

## Abstract

{% if sex\_chromosomes is defined and sex\_chromosomes %}We present a genome assembly from an individual $SAMPLE\_SEX {{ GENUS\_SPECIES }} (the {{ COMMON\_NAME }}; {{ PHYLUM }}; {{ CLASS }}; {{ ORDER }}; {{ FAMILY }}). The genome sequence spans {{ GENOME\_LENGTH }} megabases. Most of the assembly is scaffolded into {{ CHROMOSOME\_NUMBER }} chromosomal pseudomolecules, including the {{ sex\_chromosomes }} sex chromosome. {% else %}We present a genome assembly from an individual $SAMPLE\_SEX {{ GENUS\_SPECIES }} (the {{ COMMON\_NAME }}; {{ PHYLUM }}; {{ CLASS }}; {{ ORDER }}; {{ FAMILY }}). The genome sequence spans {{ GENOME\_LENGTH }} megabases. Most of the assembly is scaffolded into {{ CHROMOSOME\_NUMBER }} chromosomal pseudomolecules.{% endif %}{% if not length\_plastid\_kb %} The mitochondrial genome has also been assembled and is {{ MITO\_SIZE }} kilobases in length.{% else %} The mitochondrial and plastid genome assemblies have lengths of {{ MITO\_SIZE }} kilobases and {{ length\_plastid\_kb }} kilobases, respectively.{% endif %} {% if annot\_url %}Gene annotation of this assembly on Ensembl identified {{ PCG }} protein-coding genes.{% endif %}

## Keywords

{{ GENUS\_SPECIES }}, {{ COMMON\_NAME }}, genome sequence, chromosomal, {{ ORDER }}

|  |  |  |  |
| --- | --- | --- | --- |
| Name | email | affiliation | [Roles](https://credit.niso.org/) |
|  |  |  |  |
|  |  |  |  |

# Main body

## Species taxonomy

{{ TAX\_STRING }} *{{ GENUS\_SPECIES }} {{* TAXONOMY\_AUTHORITY }} (NCBI:txid{{ NCBI\_TAXID }}).

## Background

The genome of the {{ COMMON\_NAME }}, {{ GENUS\_SPECIES }}, 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 {{ GENUS\_SPECIES }}, based on $SAMPLE\_NUMBER $SAMPLE\_SEX specimen[s] from {{ COLLECTION\_LOCATION }}.

## Genome sequence report

The genome of a[n] {{ LIFESTAGE }} {{ observed\_sex }} {{ GENUS\_SPECIES }}(Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of {{ pacbio\_base\_count\_gb }} Gb (gigabases) from {{ pacbio\_read\_count\_millions }} million reads, providing approximately {{ LR\_cov }}-fold coverage. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data, which produced {{ hic\_base\_count\_gb }} Gbp from {{ hic\_read\_count\_millions }} million reads, yielding an approximate coverage of {{ hic\_cov }}-fold. Specimen and sequencing information is summarised in Table 1.

Manual assembly curation corrected {% if manual\_breaks\_and\_joins != 0 %}{{ manual\_breaks\_and\_joins }} missing joins or mis-joins{% endif %} {% if manual\_haplotig\_removals != 0 %} and {{ manual\_haplotig\_removals }} haplotypic duplications{% endif %}, reducing the assembly length by {{ scaffold\_total\_length\_change }}% {% if scaffold\_count\_change != 0 %} and the scaffold number by {{ scaffold\_count\_change }}%{% endif %}, and increasing the scaffold N50 by {{ scaffold\_N50\_change }}%. The final assembly has a total length of {{ GENOME\_LENGTH }} Mb in {{ SCAFF\_NUMBER }} sequence scaffolds, with {{ gap\_count }} gaps, and a scaffold N50 of {{ SCAFF\_N50 }} Mb (Table 2). 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 ({{ perc\_assem }}%) of the assembly sequence was assigned to {{ CHROMOSOME\_NUMBER }} chromosomal-level scaffolds{% if sex\_chromosomes %}, representing {{ autosomes }} autosomes and the {{ sex\_chromosomes }} sex chromosome{% endif %}. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 3). {% if genome\_note%} {{ genome\_note }}.{% endif %} 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 {{ QV }} with *k*-mer completeness of {{ KMER }}%, and the assembly has a BUSCO v5.4.3 completeness of {{ BUSCO\_C }}% (single = {{ BUSCO\_S }}%, duplicated = {{ BUSCO\_D }}%), using the {{ BUSCO\_REF }} reference set (n = {{ BUSCO\_N }}).

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/{{ NCBI\_TAXID }}.

### {% if annot\_url %}Genome annotation report

The {{ GENUS\_SPECIES }} genome assembly ({{ annot\_accession }}) was annotated at the European Bioinformatics Institute (EBI) on Ensembl Rapid Release. The resulting annotation includes {{ transcripts }} transcribed mRNAs from {{ prot\_genes }} protein-coding {% if non\_coding != 0 %}and {{ non\_coding }} non-coding {% endif %}genes (Table 2; {{ annot\_url }}). The average transcript length is {{ av\_transcript\_length }}. There are {{ av\_transc }} coding transcripts per gene and {{ av\_exon }} exons per transcript.{% endif %}

# Methods

## Sample acquisition

A[n] {{ observed\_sex }} {{ LIFESTAGE }} {{ GENUS\_SPECIES }}(specimen ID {{ SPECIMEN\_ID }}, ToLID {{ TOLID }}) was collected from {{ COLLECTION\_LOCATION }} (latitude {{ LATITUDE }}, longitude {{ LONGITUDE }}) on {{ COLLECTION\_DATE }} by {{ COLLECTION\_METHOD }}. The specimen was collected by {{ COLLECTORS }} ({{ COLLECTOR\_INSTITUTE }}) and identified by {{ IDENTIFIER }} ({{ IDENTIFIER\_INSTITUTE }}) and preserved by {{ PRESERVATION\_METHOD }}.

{% if HIC\_TOLID and HIC\_TOLID != TOLID %}The specimen used for Hi-C sequencing (specimen ID {{ HIC\_SPECIMEN\_ID }}, ToLID {{ HIC\_TOLID }}) was a[n] {{ HIC\_LIFESTAGE }} specimen collected from {{ HIC\_COLLECTION\_LOCATION }} (latitude {{ HIC\_LATITUDE }}, longitude {{ HIC\_LONGITUDE }} on {{ HIC\_COLLECTION\_DATE }} by {{ HIC\_COLLECTION\_METHOD }}. The specimen was collected by {{ HIC\_COLLECTORS }} ({{ HIC\_COLLECTOR\_INSTITUTE }}) and identified by {{ HIC\_IDENTIFIER }} ({{ HIC\_IDENTIFIER\_INSTITUTE }}) and preserved by {{ HIC\_PRESERVATION\_METHOD }}. {% endif %}

{% if RNA\_TOLID and RNA\_TOLID != TOLID %}The specimen used for RNA sequencing (specimen ID {{ RNA\_SPECIMEN\_ID }}, ToLID {{ RNA\_TOLID }}) was a[n] {{ RNA\_LIFESTAGE }} specimen collected from {{ RNA\_COLLECTION\_LOCATION }} (latitude {{ RNA\_LATITUDE}}, longitude {{ RNA\_LONGITUDE }}) on {{ RNA\_COLLECTION\_DATE }} by {{ RNA\_COLLECTION\_METHOD }}. The specimen was collected by {{ RNA\_COLLECTORS }} ({{ RNA\_COLLECTOR\_INSTITUTE }}) and identified by {{ RNA\_IDENTIFIER }} ({{ RNA\_IDENTIFIER\_INSTITUTE }}) and preserved by {{ RNA\_PRESERVATION\_METHOD }}. {% endif %}

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 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 is 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 {{ TOLID }} sample was weighed and dissected on dry ice (Jay *et al.*, 2023).

[1: sample homogenisation method]

For sample homogenisation, {{ ORGANISM\_PART }} tissue was cryogenically disrupted using the Covaris cryoPREP® Automated Dry Pulverizer (Narváez-Gómez *et al.*, 2023).

OR

Tissue from the {{ ORGANISM\_PART }} was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a).

[2 Core lab:

* MagAttract v1 was used for all non-plant/fungi samples up until January 2023 - from then on, MagAttract v2 is used.]

[3 Core lab]

* For MagAttract v1, Sanger Tree of Life HMW DNA Fragmentation: Diagenode Megaruptor®3 for PacBio HiFi shearing protocol is used.
* For MagAttract v2, Plant MagAttract v3, Plant MagAttract v4 and POE - Sanger Tree of Life HMW DNA Fragmentation: Diagenode Megaruptor®3 for LI PacBio shearing protocol is used.]

[4 core lab]

* SPRI went from being routinely manual to routinely automated on September 2022. All ULI samples have an automated SPRI.

[🡪Core Lab Automatic Magattract v1, with Megaruptor®3 for PacBio HiFi and manual SPRI:]

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.*, 2023a). 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.

[🡪Core Lab Manual Magattract v1, with Megaruptor®3 for PacBio HiFi and manual SPRI:]

HMW DNA was extracted using the Manual MagAttract v1 protocol (Strickland *et al.*, 2023b). 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.*, 2023a). 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.

[🡪Core Lab Magattract v2 with Megaruptor®3 for LI PacBio and automated SPRI:]

HMW DNA was extracted using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023a). DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Bates *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Oatley *et al.*, 2023b). 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.

[2, 3 & 4 SciOps

* MagAttract v2 used for all non-plant/fungi samples.
* Sanger Tree of Life HMW DNA Fragmentation: Diagenode Megaruptor®3 for LI PacBio shearing protocol is used.
* The SPRI performed is routinely manual.]

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023a). The DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Bates *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation (Strickland *et al.*, 2023a): in brief, the method employs a 1.8X ratio of AMPure PB beads to sample 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.

[3 For ULI, Covaris G-tube fragmentation. ]

For ULI PacBio sequencing, DNA was fragmented using the Covaris g-TUBE method (Oatley et al., 2023c).

[5 RNA -always core lab, usually automated]

{% if RNA\_TOLID %} RNA was extracted from {{ RNA\_ORGANISM\_PART }} tissue of {{ RNA\_TOLID }} 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. {% endif %}

## Sequencing

{% if RNA\_TOLID %} 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 {{ pacbio\_instrument\_model }} (HiFi) and {{ rna\_instrument\_model }} (RNA-Seq) instruments. Hi-C data were also generated from {{ HIC\_ORGANISM\_PART }} tissue of {{ HIC\_TOLID }} 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 {{ hic\_instrument\_model }} instrument. {% endif %}

{% if not RNA\_TOLID %} Pacific Biosciences HiFi circular consensus DNA sequencing libraries were constructed according to the manufacturers’ instructions. DNA sequencing was performed by the Scientific Operations core at the WSI on a Pacific Biosciences {{ pacbio\_instrument\_model }} instrument. Hi-C data were also generated from {{ HIC\_ORGANISM\_PART }} tissue of {{ HIC\_TOLID }} 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 {{ hic\_instrument\_model }} instrument. {% endif %}

## 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. {include sex chromosome identification step where relevant} The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation> (article in preparation).

### Evaluation of the final assembly

[old BTK version – done by Rich]

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur and Mirny, 2020). To assess the assembly metrics, the k-mer completeness and QV consensus quality values were calculated in Merqury (Rhie *et al.*, 2020). This work was done using Nextflow (Di Tommaso *et al.*, 2017) DSL2 pipelines “sanger-tol/readmapping” (Surana *et al.*, 2023a) and “sanger-tol/genomenote” (Surana *et al.*, 2023b). The genome was analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021; Simão *et al.*, 2015) were calculated.

[new BTK version, next flow pipeline]

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 and Hall, 2010) and the Cooler tool suite (Abdennur and Mirny, 2020), which is then visualised with HiGlass (Kerpedjiev *et al.*, 2018). It also provides statistics about 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.

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](https://www.anaconda.com/) 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.

{% if annot\_url %}

## Genome annotation

{% if annot\_method == "ensembl" %}The [Ensembl Genebuild](https://rapid.ensembl.org/info/genome/genebuild/full_genebuild.html) annotation system (Aken *et al.*, 2016) was used to generate annotation for the {{ GENUS\_SPECIES }}assembly ({{ ASSEMBLY\_ACCESSION }}) in Ensembl Rapid Release at the EBI. Annotation was created primarily through alignment of transcriptomic data to the genome, with gap filling via protein-to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 2019). {% elif annot\_method == "braker" %}The [BRAKER2](https://rapid.ensembl.org/info/genome/genebuild/braker.html) pipeline (Brůna *et al.*, 2021) was used in the default protein mode to generate annotation for the {{ GENUS\_SPECIES }}assembly ({{ ASSEMBLY\_ACCESSION }}) in Ensembl Rapid Release at the EBI. {% endif %}{% endif %}

### 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](https://www.darwintreeoflife.org/project-resources/). 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 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: {{ ENA\_BIOPROJECT\_TITLE }}. Accession number {{ ENA\_BIOPROJECT\_ACCESSION }}; <https://identifiers.org/ena.embl/>{{ ENA\_BIOPROJECT\_ACCESSION }} (Wellcome Sanger Institute, {{ ENA\_FIRST\_PUBLIC }}). The genome sequence is released openly for reuse. The {{ GENUS\_SPECIES }}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. {% if not annot\_url %}The genome will be annotated using available RNA-Seq data and presented through the [Ensembl](https://www.ensembl.org/) pipeline at the European Bioinformatics Institute.{% endif %} Raw data and assembly accession identifiers are reported in Tables 1 and 2.

## Grant information

This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (206194) and the Darwin Tree of Life Discretionary Award (218328).

## Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: <https://doi.org/10.5281/zenodo.12157525>.

Members of the Natural History Museum Genome Acquisition Lab are listed here: <https://doi.org/10.5281/zenodo.12159242>.

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](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. and Mirny, L. A. (2020) Cooler: Scalable storage for Hi-C data and other genomically labeled arrays, *Bioinformatics*, 36 (1), pp. 311–316. DOI:10.1093/bioinformatics/btz540.

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., *et al.* (2016) The Ensembl gene annotation system., *Database: The Journal of Biological Databases and Curation*, 2016, pp. baw093. DOI:10.1093/database/baw093.

Allio, R., Schomaker‐Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B. and Delsuc, F. (2020) MitoFinder: Efficient automated large‐scale extraction of mitogenomic data in target enrichment phylogenomics, *Molecular Ecology Resources*, 20 (4), pp. 892–905. DOI:10.1111/1755-0998.13160.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool, *Journal of Molecular Biology*, 215 (3), pp. 403–410. DOI:10.1016/S0022-2836(05)80360-2.

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., *et al.* (2023) UniProt: the Universal Protein Knowledgebase in 2023, *Nucleic Acids Research*, 51 (D1), pp. D523–D531. DOI:10.1093/nar/gkac1052.

Bates, A., Clayton-Lucey, I. and Howard, C. (2023) Sanger Tree of Life HMW DNA Fragmentation: Diagenode Megaruptor®3 for LI PacBio, *protocols.io*. DOI:10.17504/protocols.io.81wgbxzq3lpk/v1.

Beasley, J., Uhl, R., Forrest, L. L., Bell, D., Hart, M., Yahr, R. and Kilias, E. (2023) DNA barcoding SOPs for the Darwin Tree of Life Project, *protocols.io*. Available from: https://dx.doi.org/10.17504/protocols.io.261ged91jv47/v1 [Accessed 25 June 2024].

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. and Borodovsky, M. (2021) BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database, *NAR Genomics and Bioinformatics*, 3 (1), pp. 1–11. DOI:10.1093/nargab/lqaa108.

Buchfink, B., Reuter, K. and Drost, H.-G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND, *Nature Methods*, 18 (4), pp. 366–368. DOI:10.1038/s41592-021-01101-x.

Challis, R., Kumar, S., Sotero-Caio, C., Brown, M. and Blaxter, M. (2023) Genomes on a Tree (GoaT): A versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life, *Wellcome Open Research*, 8, pp. 24. DOI:10.12688/wellcomeopenres.18658.1.

Challis, R., Richards, E., Rajan, J., Cochrane, G. and Blaxter, M. (2020) BlobToolKit – interactive quality assessment of genome assemblies, *G3: Genes, Genomes, Genetics*, 10 (4), pp. 1361–1374. DOI:10.1534/g3.119.400908.

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. and Li, H. (2021) Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm, *Nature Methods*, 18 (2), pp. 170–175. DOI:10.1038/s41592-020-01056-5.

Crowley, L., Allen, H., Barnes, I., Boyes, D., Broad, G. R., Fletcher, C., *et al.* (2023) A sampling strategy for genome sequencing the British terrestrial arthropod fauna, *Wellcome Open Research*, 8, pp. 123. DOI:10.12688/wellcomeopenres.18925.1.

da Veiga Leprevost, F., Grüning, B. A., Alves Aflitos, S., Röst, H. L., Uszkoreit, J., Barsnes, H., *et al.* (2017) BioContainers: an open-source and community-driven framework for software standardization, *Bioinformatics*, 33 (16), pp. 2580–2582. DOI:10.1093/bioinformatics/btx192.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., *et al.* (2021) Twelve years of SAMtools and BCFtools, *GigaScience*, 10 (2). DOI:10.1093/gigascience/giab008.

Denton, A., Oatley, G., Cornwell, C., Quail, M. and Howard, C. (2023a) Sanger Tree of Life Sample Homogenisation: PowerMash, *protocols.io*. DOI:10.17504/protocols.io.5qpvo3r19v4o/v1.

Denton, A., Yatsenko, H., Jay, J., Houliston, K. and Howard, C. (2023b) Sanger Tree of Life Wet Laboratory Protocol Collection, *protocols.io*. DOI:10.17504/protocols.io.8epv5xxy6g1b/v1.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E. and Notredame, C. (2017) Nextflow enables reproducible computational workflows, *Nature Biotechnology*, 35 (4), pp. 316–319. DOI:10.1038/nbt.3820.

Diesh, C., Stevens, G. J., Xie, P., De Jesus Martinez, T., Hershberg, E. A., Leung, A., *et al.* (2023) JBrowse 2: a modular genome browser with views of synteny and structural variation, *Genome Biology*, 24 (1), pp. 74. DOI:10.1186/s13059-023-02914-z.

do Amaral, R. J. V., Bates, A., Denton, A., Yatsenko, H., Jay, J. and Howard, C. (2023) Sanger Tree of Life RNA Extraction: Automated MagMaxTM mirVana, *protocols.io*. DOI:10.17504/protocols.io.6qpvr36n3vmk/v1.

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P. and Nahnsen, S. (2020) The nf-core framework for community-curated bioinformatics pipelines, *Nature Biotechnology*, 38 (3), pp. 276–278. DOI:10.1038/s41587-020-0439-x.

Ewels, P., Magnusson, M., Lundin, S. and Käller, M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report, *Bioinformatics*, 32 (19), pp. 3047–3048. DOI:10.1093/bioinformatics/btw354.

Formenti, G., Abueg, L., Brajuka, A., Brajuka, N., Gallardo-Alba, C., Giani, A., Fedrigo, O. and Jarvis, E. D. (2022) Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs, *Bioinformatics*, 38 (17), pp. 4214–4216. DOI:10.1093/bioinformatics/btac460.

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R. and Köster, J. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences, *Nature Methods*, 15 (7), pp. 475–476. DOI:10.1038/s41592-018-0046-7.

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y. and Durbin, R. (2020) Identifying and removing haplotypic duplication in primary genome assemblies., *Bioinformatics (Oxford, England)*, 36 (9), pp. 2896–2898. DOI:10.1093/bioinformatics/btaa025.

Harry, E. (2022) PretextView (Paired REad TEXTure Viewer): A desktop application for viewing pretext contact maps. Available from: https://github.com/wtsi-hpag/PretextView [Accessed 19 October 2022].

Howe, K., Chow, W., Collins, J., Pelan, S., Pointon, D. L., Sims, Y., Torrance, J., Tracey, A. and Wood, J. (2021) Significantly improving the quality of genome assemblies through curation, *GigaScience*, 10 (1). DOI:10.1093/gigascience/giaa153.

Jay, J., Yatsenko, H., Narváez-Gómez, J. P., Mbye, H., Morra, M., Strickland, M., Cornwell, C. and Howard, C. (2023) Sanger Tree of Life Sample Preparation: Triage and Dissection, *protocols.io*. DOI:10.17504/protocols.io.x54v9prmqg3e/v1.

Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., *et al.* (2018) HiGlass: web-based visual exploration and analysis of genome interaction maps, *Genome Biology*, 19 (1), pp. 125. DOI:10.1186/s13059-018-1486-1.

Kurtzer, G. M., Sochat, V. and Bauer, M. W. (2017) Singularity: Scientific containers for mobility of compute, *PLOS ONE*, 12 (5), pp. e0177459. DOI:10.1371/journal.pone.0177459.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences, *Bioinformatics*, 34 (18), pp. 3094–3100. DOI:10.1093/bioinformatics/bty191.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. and Zdobnov, E. M. (2021) BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes, *Molecular Biology and Evolution*, 38 (10), pp. 4647–4654. DOI:10.1093/molbev/msab199.

Merkel, D. (2014) Docker: lightweight Linux containers for consistent development and deployment, *Linux Journal*, 2014 (239), pp. 2. Available from: https://dl.acm.org/doi/10.5555/2600239.2600241 [Accessed 2 April 2024].

Muffato, M., Butt, Z., Challis, R., Kumar, S., Qi, G., Ramos Díaz, A. and Surana, P. (2024) sanger-tol/blobtoolkit: v0.3.0 – Poliwag. DOI:https://doi.org/10.5281/zenodo.10649272.

Narváez-Gómez, J. P., Mbye, H., Oatley, G., Park, N., Strickland, M. and Howard, C. (2023) Sanger Tree of Life Sample Homogenisation: Covaris cryoPREP® Automated Dry Pulverizer, *protocols.io*. DOI:dx.doi.org/10.17504/protocols.io.eq2lyjp5qlx9/v1.

Oatley, G., Denton, A. and Howard, C. (2023a) *Sanger Tree of Life HMW DNA Extraction: Automated MagAttract v.2*, *protocols.io*. DOI:10.17504/protocols.io.kxygx3y4dg8j/v1.

Oatley, G., Sampaio, F. and Howard, C. (2023b) Sanger Tree of Life Fragmented DNA clean up: Automated SPRI, *protocols.io*. DOI:10.17504/protocols.io.q26g7p1wkgwz/v1.

Oatley, G., Sampaio, F., Kitchin, L., do Amaral, R. J. V. and Howard, C. (2023c) Sanger Tree of Life HMW DNA Fragmentation: Covaris g-TUBE for ULI PacBio, *protocols.io*. Available from: https://dx.doi.org/10.17504/protocols.io.q26g7pm81gwz/v1 [Accessed 13 June 2024].

Pointon, D.-L., Eagles, W., Sims, Y., Muffato, M. and Surana, P. (2023) sanger-tol/treeval v1.0.0 – Ancient Atlantis. DOI:10.5281/zenodo.10047653.

Quinlan, A. R. and Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, 26 (6), pp. 841–842. DOI:10.1093/bioinformatics/btq033.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, *Cell*, 159 (7), pp. 1665–1680. DOI:10.1016/j.cell.2014.11.021.

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species, *Nature*, 592 (7856), pp. 737–746. DOI:10.1038/s41586-021-03451-0.

Rhie, A., Walenz, B. P., Koren, S. and Phillippy, A. M. (2020) Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies, *Genome Biology*, 21 (1). DOI:10.1186/s13059-020-02134-9.

Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Sherry, S. T., Yankie, L. and Karsch-Mizrachi, I. (2024) GenBank 2024 Update, *Nucleic Acids Research*, 52 (D1), pp. D134–D137. DOI:10.1093/nar/gkad903.

Sheerin, E., Sampaio, F., Oatley, G., Todorovic, M., Strickland, M., do Amaral, R. J. V. and Howard, C. (2023) Sanger Tree of Life HMW DNA Extraction: Automated MagAttract v.1, *protocols.io*. DOI:10.17504/protocols.io.x54v9p2z1g3e/v1.

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

Strickland, M., Cornwell, C. and Howard, C. (2023a) Sanger Tree of Life Fragmented DNA clean up: Manual SPRI, *protocols.io*. DOI:10.17504/protocols.io.kxygx3y1dg8j/v1.

Strickland, M., Moll, R., Cornwell, C., Smith, M. and Howard, C. (2023b) Sanger Tree of Life HMW DNA Extraction: Manual MagAttract, *protocols.io*. DOI:10.17504/protocols.io.6qpvr33novmk/v1.

Surana, P., Muffato, M. and Qi, G. (2023a) sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0). Zenodo. DOI:10.5281/zenodo.7755669.

Surana, P., Muffato, M. and Sadasivan Baby, C. (2023b) sanger-tol/genomenote (v1.0.dev). Zenodo. DOI:10.5281/zenodo.6785935.

Todorovic, M., Sampaio, F. and Howard, C. (2023) Sanger Tree of Life HMW DNA Fragmentation: Diagenode Megaruptor®3 for PacBio HiFi, *protocols.io*. DOI:10.17504/protocols.io.8epv5x2zjg1b/v1.

Twyford, A. D., Beasley, J., Barnes, I., Allen, H., Azzopardi, F., Bell, D., *et al.* (2024) A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life Project, *Wellcome Open Research*, 9, pp. 339. DOI:10.12688/wellcomeopenres.21143.1.

Uliano-Silva, M., Ferreira, J. G. R. N., Krasheninnikova, K., Blaxter, M., Mieszkowska, N., Hall, N., *et al.* (2023) MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads, *BMC Bioinformatics*, 24 (1), pp. 288. DOI:10.1186/s12859-023-05385-y.

UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge, *Nucleic Acids Research*, 47 (D1), pp. D506–D515. DOI:10.1093/nar/gky1049.

Vasimuddin, Md., Misra, S., Li, H. and Aluru, S. (2019) Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems, in: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, pp. 314–324.

Zhou, C., McCarthy, S. A. and Durbin, R. (2023) YaHS: yet another Hi-C scaffolding tool, *Bioinformatics*, 39 (1). DOI:10.1093/bioinformatics/btac808.

[add dataset citation]

Wellcome Sanger Institute ({{ ENA\_FIRST\_PUBLIC }}) [TITLE], European Nucleotide Archive, [dataset], accession number {{ ENA\_BIOPROJECT\_ACCESSION }}

# Figures

**Figure 1. Photograph of the {{ *GENUS\_SPECIES* }}({{ TOLID }}) specimen used for genome sequencing**

{% if Fig\_2\_Snail\_png %} {{ Fig\_2\_Snail\_png }} {% endif %}

**Figure 2: Genome assembly of {{ *GENUS\_SPECIES* }}, {{ ASSEMBLY\_ID }}: metrics.** The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. $BTK\_SNAIL\_LEG An interactive version of this figure is available at {{ BTK\_SNAIL\_URL }}.

{% if Fig\_3\_Blob\_png %} {{ Fig\_3\_Blob\_png }} {% endif %}

**Figure 3: Genome assembly of {{ *GENUS\_SPECIES* }}, {{ ASSEMBLY\_ID }}: BlobToolKit GC-coverage plot.** 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 {{ BTK\_BLOB\_URL }}.

{% if Fig\_4\_Cumulative\_png %} {{ Fig\_4\_Cumulative\_png }} {% endif %}

**Figure 4: Genome assembly of {{ *GENUS\_SPECIES* }}, {{ ASSEMBLY\_ID }}: 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 buscogenes taxrule. An interactive version of this figure is available at {{ BTK\_CUMULATIVE\_URL }}.

**Figure 5: Genome assembly of {{ *GENUS\_SPECIES* }}, {{ ASSEMBLY\_ID }}: Hi-C contact map** **of the {{ ASSEMBLY\_ID }}** 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 {{ HIGLASS\_URL }}

# Tables

**Table 1: Specimen and sequencing data for** ***{{ GENUS\_SPECIES }}***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Project information** | | | | |
| **Study title** | {{ ENA\_BIOPROJECT\_TITLE }} | | | |
| **Umbrella BioProject** | {{ ENA\_BIOPROJECT\_ACCESSION }} | | | |
| **Species** | {{ *SPECIES* }} | | | |
| **BioSample** | {{ PROJECT\_BIOSAMPLE\_ACCESSION }} | | | |
| **NCBI taxonomy ID** | {{ NCBI\_TAXID }} | | | |
| **Specimen information** | | | | |
| **Technology** | **ToLID** | **BioSample accession** | | **Organism part** |
| **PacBio long read sequencing** | {{ TOLID }} | {{ BIOSAMPLE\_ACCESSION }} | | {{ ORGANISM\_PART }} |
| **Hi-C sequencing** | {{ HIC\_TOLID }} | {{ HIC\_BIOSAMPLE\_ACCESSION }} | | {{ HIC\_ORGANISM\_PART }} |
| **RNA sequencing** | {{ RNA\_TOLID }} | {{ RNA\_BIOSAMPLE\_ACCESSION }} | | {{ RNA\_ORGANISM\_PART }} |
| **Sequencing information** | | | | |
| **Platform** | **Run accession** | **Read count** | **Base count (Gb)** | |
| {% if seq\_data is defined and seq\_data %}{% for technology, runs in seq\_data.items() %}{% for run in runs %}**{{ technology }} {{ run.instrument\_model }}** | {{ run.read\_accession }} | {{ run.read\_count }} | {{ run.base\_count\_gb }} | |
| {% endfor %}{% endfor %}{% endif %} |  |  |  | |

**Table 2: Genome assembly data for** ***{{ GENUS\_SPECIES }}*, {{ ASSEMBLY\_ID }}**

|  |  |  |
| --- | --- | --- |
| **Genome assembly** | | |
| Assembly name | {{ ASSEMBLY\_ID }} | |
| Assembly accession | {{ ASSEMBLY\_ACCESSION }} | |
| *Accession of alternate haplotype* | {{ alternate\_accession }} | |
| Span (Mb) | {{ GENOME\_LENGTH }} | |
| Number of contigs | {{ CONTIG\_NUMBER }} | |
| Contig N50 length (Mb) | {{ CONTIG\_N50 }} | |
| Number of scaffolds | {{ SCAFF\_NUMBER }} | |
| Scaffold N50 length (Mb) | {{ SCAFF\_N50 }} | |
| Longest scaffold (Mb) | {{ LONGEST\_SCAFF }} | |
| **Assembly metrics\*** | | Benchmark |
| Consensus quality (QV) | {{ QV }} | ≥ 50 |
| *k*-mer completeness | {{ KMER }}% | ≥ 95% |
| BUSCO\*\* | {{ BUSCO\_STRING}} | C ≥ 95% |
| Percentage of assembly mapped to chromosomes | {{ perc\_assem }}% | ≥ 95% |
| Sex chromosomes | {{ sex\_chromosomes }} | localised homologous pairs |
| Organelles | Mitochondrial genome: {{ MITO\_SIZE }} kb {% if length\_plastid\_kb %} Plastid genome: {{ length\_plastid\_kb }} kb {% endif %} | complete single alleles |
| **Genome annotation of assembly {{ ASSEMBLY\_ACCESSION }} at Ensembl** | | |
| Number of protein-coding genes | {{ prot\_genes }} | |
| Number of non-coding genes | {{ non\_coding }} | |
| Number of gene transcripts | {{ transcripts }} | |

\* 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 {{ BUSCO\_REF }} 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 {{ BTK\_BUSCO\_URL }}.

**Table 3: Chromosomal pseudomolecules in the genome assembly of *{{ GENUS\_SPECIES }}*, {{ TOLID }}**

|  |  |  |  |
| --- | --- | --- | --- |
| **INSDC accession** | **Name** | **Length (Mb)** | **GC%** |
| {%tr for chr in CHR\_TABLE %} | | | |
| {{ chr.get(‘Accession’) }} | {{ chr.get(‘Chromosome’) }} | {{ chr.get(‘Length’) }} | {{ chr.get(‘GC’) }} |
| {%tr endfor %} | | | |

**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 | 0.19.8-r603 | <https://github.com/chhylp123/hifiasm> |
| HiGlass | 44086069ee7d4d3f6f3f0012569789ec138f42b84aa44357826c0b6753eb28de | <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> |