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

## Authors

{{ AUTHORS }} $INTRO\_WRITER, {{ 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

We present a genome assembly from an individual $SAMPLE\_SEX {{ GENUS\_SPECIES }} (the {{ COMMON\_NAME }}; {{ PHYLUM }}; {{ CLASS }}; {{ORDER }}; {{ FAMILY }}). The genome sequence is {{ GENOME\_LENGTH }} megabases in span. Most of the assembly is scaffolded into {{ CHROMOSOME\_NUMBER }} chromosomal pseudomolecules, including the $SEX\_CHROMOSOMES sex chromosome. The mitochondrial genome has also been assembled and is {{ MITO\_SIZE }} kilobases in length. [*if Ensembl has completed annotation:* Gene annotation of this assembly on Ensembl identified $PCG protein coding genes.]

## Keywords

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

|  |  |  |  |
| --- | --- | --- | --- |
| Name | email | affiliation | [Roles](https://credit.niso.org/) |
| Douglas Boyes  + Deceased author | wytham.moths@biology.ox.ac.uk | UK Centre for Ecology and Hydrology, Wallingford | Resources, Investigation |
| Peter W.H. Holland | peter.holland@biology.ox.ac.uk | Department of Biology, University of Oxford | Writing |

# Main body

## Species taxonomy

{{ TAX\_STRING }} $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 was sequenced from $SAMPLE\_NUMBER $LIFE\_STAGE $SAMPLE\_SEX *{{ GENUS\_SPECIES }}* (Figure 1) collected from {{ COLLECTION\_LOCATION }} ({{ LATITUDE }}, {{ LONGITUDE }}). A total of $LR\_FOLD\_COV-fold coverage in Pacific Biosciences single-molecule HiFi long reads [and $SR\_FOLD\_COV-fold coverage in 10X Genomics read clouds] was generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected $MISSING\_MISJOINS missing joins or mis-joins and removed $HAP\_DUPS haplotypic duplications, reducing the assembly length by $ASSEMBLY\_CHANGE% [if $SCAFF\_CHANGE >0: and the scaffold number by $SCAFF\_CHANGE%, and increasing the scaffold N50 by $SCAFF\_N50\_CHANGE%.

The final assembly has a total length of {{ GENOME\_LENGTH }} Mb in {{ SCAFF\_NUMBER }} sequence scaffolds with a scaffold N50 of {{ SCAFF\_N50 }} Mb (Table 1). The snailplot 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 ($CHROM\_ASSEM%) of the assembly sequence was assigned to {{ CHROMOSOME\_NUMBER }} chromosomal-level scaffolds, representing $AUTOSOME\_NUMBER autosomes and the $SEX\_CHROMOSOMES sex chromosome. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 2). 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 v5.3.2 completeness of {{ BUSCO }}% (single = xxx%, duplicated = xxx%), using the {{ BUSCO\_REF }} reference set (n = xxx).

Metadata for specimens, barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly statistics are given at https://links.tol.sanger.ac.uk/species/{{ NCBI\_TAXID }}.

## Genome annotation report

The *{{ GENUS\_SPECIES }}* genome assembly ({{ ASSEMBLY\_ACCESSION }}) was annotated using the Ensembl rapid annotation pipeline (Table 1; $ANNOTATION\_URL). The resulting annotation includes $TRANSC\_MRNA transcribed mRNAs from $PCG protein-coding and $NCG non-coding genes. There are $CDS\_PER\_GENE coding transcripts per gene and $EXONS\_PER\_TRANSC exons per transcript.

# Methods

## Sample acquisition and nucleic acid extraction

A $SAMPLE\_SEX *{{ GENUS\_SPECIES }}* (specimen ID {{ SPECIMEN\_ID }}, ToLID {{ TOLID }}) was collected from {{ COLLECTION\_LOCATION }} (latitude {{ LATITUDE }}, longitude {{ LONGITUDE }}) on {{ COLLECTION\_DATE }} by $SAMPLING\_METHOD. The specimen was taken from {{ HABITAT }} by {{ COLLECTORS }} ({{ COLLECTOR\_INSTITUTE }}) The specimen was collected by {{ COLLECTORS }} ({{ COLLECTOR\_INSTITUTE }}) and identified by {{ IDENTIFIER }} ({{ IDENTIFIER\_INSTITUTE }}) and preserved by $PRESERV\_METHOD in $PRESERV\_SOLUTION.

HiC specimen: The specimen used for Hi-C sequencing (specimen ID $HIC\_SPECIMEN\_ID, ToLID $HIC) was collected from $HIC\_COLLECTION\_LOCATION (latitude $HIC\_LATITIUDE, longitude $HIC\_LONGITUDE) on $HIC\_COLLECTION\_DATE by $HIC\_SAMPLING\_METHOD. The specimen was collected by $HIC\_COLLECTORS ($HIC\_COLLECTOR\_INSTITUTE) and identified by $HIC\_IDENTIFIER ($ HIC\_IDENTIFIER\_INSTIUTE) and preserved on $HIC\_PRESERV\_SOLUTION.

RNA specimen: The specimen used for RNA sequencing (specimen ID $RNA\_SPECIMEN\_ID, ToLID $RNA) was collected from $RNA\_COLLECTION\_LOCATION (latitude $RNA\_LATITUDE, longitude $RNA\_LONGITUDE) on $RNA\_COLLECTION\_DATE by $RNA\_SAMPLING\_METHOD. The specimen was collected by $RNA\_COLLECTORS ($RNA\_COLLECTOR\_INSTITUTE) and identified by $RNA\_IDENTIFIER ($RNA\_IDENTIFIER\_INSTITUTE) and preserved on $RNA\_PRESERV\_SOLUTION.

Protocols developed by the Wellcome Sanger Institute (WSI) Tree of Life core laboratory have been deposited on protocols.io (Denton et al., 2023b). The workflow for high molecular weight (HMW) DNA extraction at the WSI includes a sequence of core procedures: sample preparation; sample homogenisation, DNA extraction, fragmentation, and clean-up. In sample preparation, the {{ TOLID}} sample was weighed and dissected on dry ice (Jay et al., 2023).

[1: sample homogenisation method]

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

OR

Tissue from {{ TISSUE\_TYPE }} 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 Magattract v1, with Megaruptor®3 for PacBio HiFi and manual SPRI:]

HMW DNA was extracted using the Automated MagAttract v1 protocol (Oatley, Denton & Howard, 2023a). HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 30 (Todorovic, Sampaio & Howard, 2023). Sheared DNA was purified by solid-phase reversible immobilisation (Strickland, Cornwell & Howard, 2023): 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 and 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, Denton & Howard, 2023b). HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 31 (Bates, Clayton-Lucey & Howard, 2023). Sheared DNA was purified by solid-phase reversible immobilisation (Oatley, Sampaio & Howard, 2023): 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 and 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 Laboratory using the Automated MagAttract v2 protocol (Oatley, Denton & Howard, 2023b). HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 31 (Bates, Clayton-Lucey & Howard, 2023). Sheared DNA was purified by solid-phase reversible immobilisation (Strickland, Cornwell & Howard, 2023): 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 and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

[5 RNA - automated]

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

## Sequencing

Pacific Biosciences HiFi circular consensus [and 10X Genomics read cloud] 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 II (HiFi)*,* $RNA\_SEQ (RNA-Seq) and $10X\_SEQ (10X) instruments. Hi-C data were also generated from $HIC\_TISSUE tissue of $HIC using the Arima2 kit and sequenced on the $HIC\_SEQ instrument.

## Genome assembly, curation and evaluation

Assembly was carried out with Hifiasm (Cheng et al., 2021) and haplotypic duplication was identified and removed with purge\_dups (Guan et al., 2020). [One round of polishing was performed by aligning 10X Genomics read data to the assembly with Long Ranger ALIGN, calling variants with FreeBayes (Garrison & Marth, 2012).] The assembly was then scaffolded with Hi-C data (Rao et al., 2014) using YaHS (Zhou, McCarthy & Durbin, 2023) [OR SALSA2 (Ghurye et al., 2019) ]. The assembly was checked for contamination and corrected using the gEVAL system (Chow et al., 2016) as described previously (Howe et al., 2021). Manual curation was performed using gEVAL,HiGlass (Kerpedjiev et al., 2018) and Pretext (Harry, 2022). The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva et al., 2023), which runs MitoFinder (Allio et al., 2020) or MITOS (Bernt et al., 2013) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin et al., 2019) in the Cooler file format (Abdennur & 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, Muffato & Qi, 2023) and “sanger-tol/genomenote” (Surana, Muffato & Sadasivan Baby, 2023). 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.

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

## Genome annotation

The Ensembl gene annotation system (Aken et al., 2016) was used to generate annotation for the *{{ GENUS\_SPECIES }}* assembly ({{ ASSEMBLY\_ACCESSION }}). 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).

OR

## Genome annotation

The BRAKER2 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.

**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. 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. Raw data and assembly accession identifiers are reported in Table 1.

## 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.7125292>.

Members of the Natural History Museum Genome Acquisition Lab are listed here: <https://doi.org/10.5281/zenodo.7139035>.

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.4893703.](https://doi.org/10.5281/zenodo.4893703)

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

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

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

Members of the Tree of Life Core Informatics collective are listed here: <https://doi.org/10.5281/zenodo.5013541>.

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

# References

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

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., et al. (2016) The Ensembl gene annotation system. *Database*. 2016. doi:10.1093/database/baw093.

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

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

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

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M. & Stadler, P.F. (2013) MITOS: Improved *de novo* metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*. 69 (2), 313–319. doi:10.1016/j.ympev.2012.08.023.

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

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

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

Chow, W., Brugger, K., Caccamo, M., Sealy, I., Torrance, J. & Howe, K. (2016) gEVAL — a web-based browser for evaluating genome assemblies. *Bioinformatics*. 32 (16), 2508–2510. doi:10.1093/bioinformatics/btw159.

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

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

Garrison, E. & Marth, G. (2012) *Haplotype-based variant detection from short-read sequencing*. https://arxiv.org/abs/1207.3907.

Ghurye, J., Rhie, A., Walenz, B.P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A.M. & Koren, S. (2019) Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Computational Biology*. 15 (8). doi:10.1371/journal.pcbi.1007273.

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

Harry, E. (2022) *PretextView (Paired REad TEXTure Viewer): A desktop application for viewing pretext contact maps.* https://github.com/wtsi-hpag/PretextView.

Howe, K., Chow, W., Collins, J., Pelan, S., Pointon, D.L., Sims, Y., Torrance, J., Tracey, A. & 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. & Howard, C. (2023) Sanger Tree of Life Sample Preparation: Triage and Dissection. *protocols.io*. doi:dx.doi.org/10.17504/protocols.io.x54v9prmqg3e/v1.

Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J.M., Ouellette, S.B., Azhir, A., Kumar, N., Hwang, J., Lee, S., Alver, B.H., Pfister, H., Mirny, L.A., Park, P.J. & Gehlenborg, N. (2018) HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biology*. 19 (1), 125. doi:10.1186/s13059-018-1486-1.

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A. & 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), 4647–4654. doi:10.1093/molbev/msab199.

Narváez-Gómez, J.P., Mbye, H., Oatley, G., Park, N., Strickland, M. & 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. & Howard, C. (2023a) Sanger Tree of Life HMW DNA Extraction: Automated MagAttract v.2. *protocols.io*. doi:dx.doi.org/10.17504/protocols.io.kxygx3y4dg8j/v1.

Oatley, G., Denton, A. & Howard, C. (2023b) Sanger Tree of Life HMW DNA Extraction: Automated MagAttract v.2. *protocols.io*. doi:dx.doi.org/10.17504/protocols.io.kxygx3y4dg8j/v1.

Oatley, G., Sampaio, F. & Howard, C. (2023) Sanger Tree of Life Fragmented DNA clean up: Automated SPRI. *protocols.io https://dx.doi.org/10.17504/protocols.io.q26g7p1wkgwz/v1*.

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

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

Rhie, A., Walenz, B.P., Koren, S. & 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.

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

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

Surana, P., Muffato, M. & Qi, G. (2023) *sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0)*. https://doi.org/10.5281/zenodo.7755665.

Surana, P., Muffato, M. & Sadasivan Baby, C. (2023) *sanger-tol/genomenote (v1.0.dev)*. https://zenodo.org/record/6785935.

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

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

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

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

Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. (2019) Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. May 2019 IEEE. pp. 314–324. doi:10.1109/IPDPS.2019.00041.

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

[add dataset citation]

Wellcome Sanger Institute ({{ ENA\_FIRST\_PUBLIC }}) The genome sequence of the {{ ENA\_BIOPROJECT\_TITLE }}, *{{ TAXONOMY\_AUTHORITY }}*, European Nucleotide Archive, [dataset], accession number {{ ENA\_BIOPROJECT\_ACCESSION }}

# Figures

[Figure 1: $SPEC\_IMAGE]

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

[Figure 2: $BTK\_FIG1]

**Figure 2: Genome assembly of *{{ GENUS\_SPECIES }}*, {{ ASSEMBLY\_ID }}: metrics.** The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. $BTK\_SNAIL\_LEG An interactive version of this figure is available at $BTK\_FIG1\_URL.

[Figure 3: $BTK\_FIG2]

**Figure 3: Genome assembly of *{{ GENUS\_SPECIES }}*, {{ ASSEMBLY\_ID }}: BlobToolKit GC-coverage plot.** Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at $BTK\_FIG2\_URL.

[Figure 4: $BTK\_FIG3]

**Figure 4: Genome assembly of *{{ GENUS\_SPECIES }}*, {{ ASSEMBLY\_ID }}: 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 $BTK\_FIG3\_URL.

[Figure 5: $HIC\_CONTACT]

**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

# Tables

**Table 1: Genome data for** ***{{ GENUS\_SPECIES }}*, {{ ASSEMBLY\_ID }}**

|  |  |  |
| --- | --- | --- |
| **Project accession data** | | |
| Assembly identifier | {{ ASSEMBLY\_ID }} | |
| Species | *{{ GENUS\_SPECIES }}* | |
| Specimen | {{ TOLID }} | |
| NCBI taxonomy ID | {{ NCBI\_TAXID }} | |
| BioProject | {{ ENA\_BIOPROJECT\_ACCESSION }} | |
| BioSample ID | {{ BIOSAMPLE\_ACCESSION }} | |
| Isolate information | {{ TOLID }}  {{ HIC }}  {{ RNA }} | |
| **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\_MAPPED | *≥ 95%* |
| Sex chromosomes |  | *localised homologous pairs* |
| Organelles |  | *complete single alleles* |
| **Raw data accessions** | | |
| PacificBiosciences SEQUEL II | $PACBIO\_ACCESSIONS | |
| 10X Genomics Illumina | $10X\_ACCESSIONS | |
| Hi-C Illumina | $HIC\_ACCESSIONS | |
| PolyA RNA-Seq Illumina | $RNA\_ACCESSIONS | |
| **Genome assembly** | | |
| Assembly accession | {{ ASSEMBLY\_ACCESSION }} | |
| *Accession of alternate haplotype* | $ALT\_HAP\_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 }} | |
| **Genome annotation** | | |
| Number of protein-coding genes | $PCG | |
| Number of non-coding genes | $NCG | |
| Number of gene transcripts | $TRANSC\_MRNA | |

\* 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 v5.3.2. 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 2: Chromosomal pseudomolecules in the genome assembly of *{{ GENUS\_SPECIES }}*, {{ TOLID }}**

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

##### **Table 3. Software tools: versions and sources**

|  |  |  |
| --- | --- | --- |
| **Software tool** | **Version** | **Source** |
| BlobToolKit | 4.1.7 | <https://github.com/blobtoolkit/blobtoolkit> |
| BUSCO | 5.3.2 | <https://gitlab.com/ezlab/busco> |
| Hifiasm | 0.16.1-r375 | <https://github.com/chhylp123/hifiasm> |
| HiGlass | 1.11.6 | <https://github.com/higlass/higlass> |
| Merqury | MerquryFK | <https://github.com/thegenemyers/MERQURY.FK> |
| MitoHiFi | 2 | <https://github.com/marcelauliano/MitoHiFi> |
| PretextView | 0.2 | <https://github.com/wtsi-hpag/PretextView> |
| purge\_dups | 1.2.3 | <https://github.com/dfguan/purge_dups> |
| sanger-tol/genomenote | v1.0 | <https://github.com/sanger-tol/genomenote> |
| sanger-tol/readmapping | 1.1.0 | <https://github.com/sanger-tol/readmapping/tree/1.1.0> |
| YaHS | yahs-1.1.91eebc2 | <https://github.com/c-zhou/yahs> |

[Older tools]

##### **Table 3. Software tools: versions and sources**

|  |  |  |
| --- | --- | --- |
| **Software tool** | **Version** | **Source** |
| BlobToolKit | 4.0.7 | <https://github.com/blobtoolkit/blobtoolkit> |
| BUSCO | 5.3.2 | <https://gitlab.com/ezlab/busco> |
| FreeBayes | 1.3.1-17-gaa2ace8 | <https://github.com/freebayes/freebayes> |
| gEVAL | N/A | <https://geval.org.uk/> |
| Hifiasm | 0.12 | <https://github.com/chhylp123/hifiasm> |
| HiGlass | 1.11.6 | <https://github.com/higlass/higlass> |
| Long Ranger ALIGN | 2.2.2 | <https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines> |
| Merqury | MerquryFK | <https://github.com/thegenemyers/MERQURY.FK> |
| MitoHiFi | 2 | <https://github.com/marcelauliano/MitoHiFi> |
| PretextView | 0.2 | <https://github.com/wtsi-hpag/PretextView> |
| purge\_dups | 1.2.3 | <https://github.com/dfguan/purge_dups> |
| SALSA | 2.2 | <https://github.com/salsa-rs/salsa> |
| sanger-tol/genomenote | v1.0 | <https://github.com/sanger-tol/genomenote> |
| sanger-tol/readmapping | 1.1.0 | <https://github.com/sanger-tol/readmapping/tree/1.1.0> |