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

## Authors

{{ COLLECTORS }}, {{ IDENTIFER }}, $INTRO\_WRITER, {{ GAL }} Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, 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\_count }} chromosomal pseudomolecules [*if sex chromosomes present:*, 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\_NUM protein coding genes.]

[If this is a plant:] The plastid genome has been assembled and is {{ plastid\_size }} kilobases in length.

## Keywords

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

|  |  |  |  |
| --- | --- | --- | --- |
| 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 $SAMPLE\_SEX *{{ GENUS\_SPECIES }}* (Figure 1) collected from {{ COLLECTION\_LOCATION }} ({{ LATITUDE }}, {{ LONGITUDE }}). A total of {{ LR\_cov }}-fold coverage in Pacific Biosciences single-molecule HiFi long reads and {{ SR\_cov }}-fold coverage in 10X Genomics read clouds were 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%] [[*if $SCAFF\_N50\_change >0:*, and increasing the scaffold N50 by $SCAFF\_N50\_CHANGE%].

The final assembly has a total length of {{ genome\_length }} Mb in {{ num\_scaffolds }} sequence scaffolds with a scaffold N50 of {{ SCAFF\_N50 }} Mb (Table 1). [*if $CHROM\_ASSEM < 100:* Most ($CHROM\_ASSEM%)*; else:* All] of the assembly sequence was assigned to {{ chromosome\_count }} 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 (Figures 2–5; Table 2). [If $JIRA\_GN\_TEXT present: $JIRA\_GN\_TEXT.; else omit] [*if two haplotypes deposited:* 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 xxx with *k*-mer completeness of xxx%, and the assembly has a BUSCO $BUSCO\_VERSION completeness of $BUSCO% (single = xxx , duplicated = xxx), using the $BUSCO\_REF reference set (n = xxx).

Metadata for specimens, spectral estimates, sequencing runs, contaminants and pre-curation assembly statistics can be found at https://links.tol.sanger.ac.uk/species/{{ 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 }}* ({{ TOL\_ID }}) was collected from {{ COLLECTION\_LOCATION }} (latitude {{ LATITUDE }}, longitude {{ LONGITUDE }}) on {{ COLLECTOR\_DATA }}. The specimen was taken from {{ HABITAT }} by {{ COLLECTORS }} ({{ COLLECTOR\_INSTITUTE }}) using {{ coll\_method }}. The specimen was identified by {{ IDENTIFIER }} ({{ IDENTIFER\_INSTIUTE }}) and preserved by {{ preserv\_method }}.

*[If DNA extracted in the TOL lab:* DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute (WSI). The {{ TOL\_ID}} sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. {{ TISSUE\_TYPE }} tissue was *[[if powermasher used:* disrupted using a Nippi Powermasher fitted with a BioMasher pestle*; else:* cryogenically disrupted to a fine powder using a Covaris cryoPREP Automated Dry Pulveriser, receiving multiple impacts*]]*. High molecular weight (HMW) DNA was extracted using the Qiagen MagAttract HMW DNA extraction kit. Low molecular weight DNA was removed from a 20 ng aliquot of extracted DNA using the 0.8X AMpure XP purification kit prior to 10X Chromium sequencing; a minimum of 50 ng DNA was submitted for 10X sequencing. HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 30. Sheared DNA was purified by solid-phase reversible immobilisation using AMPure PB beads with a 1.8X ratio of beads to sample to remove the shorter fragments and concentrate the DNA sample. 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.*]*

*[If DNA extracted by SciOps:* DNA was extracted from {{ TISSUE\_TYPE }} tissue of {{ TOL\_ID }} at the Wellcome Sanger Institute (WSI) Scientific Operations core using the Qiagen MagAttract HMW DNA kit, according to the manufacturer’s instructions.*]*

*[If RNA-Seq performed:* RNA was extracted from $RNA\_TISSUE tissue of $RNA\_TOL\_ID in the Tree of Life Laboratory at the WSI using TRIzol, according to the manufacturer’s instructions. RNA was then eluted in 50 μl RNAse-free water and its concentration assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit RNA Broad-Range (BR) Assay kit. Analysis of the integrity of the RNA was done using 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. *[if RNA-Seq performed:* Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit.*]* DNA *[if RNA-Seq performed:* and RNA*]* sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi) *[if RNA-Seq performed:*, $RNA\_SEQ (RNA-Seq)*]* and $10X\_SEQ (10X) instruments. Hi-C data were also generated from $HIC\_TISSUE tissue of {{ HiC }} using the Arimav2 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 and Marth, 2012). The assembly was then scaffolded with Hi-C data (Rao *et al.*, 2014) using YaHS (Zhou, McCarthy and 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.*, 2022), 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 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, Muffato and Qi, 2023) and “sanger-tol/genomenote” (Surana, Muffato and Sadasivan Baby, 2023). The genome was analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Simão *et al.*, 2015; Manni *et al.*, 2021) 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.

**Legal and ethical review process for Darwin Tree of Life Partner submitted materials**

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: $BIOPROJECT\_TITLE. Accession number {{ BIOPROJECT\_ACCESSION }}; <https://identifiers.org/ena.embl/>{{ BIOPROJECT\_ACCESSION }}. (Wellcome Sanger Institute, 2022)

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 genome not annotated:* 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

[If sample provided by a non-Sanger GAL, otherwise omit]Members of the {{ GAL }} are listed here: $GAL\_DOI]

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: <https://doi.org/10.5281/zenodo.4789928>.

*[If species has been barcoded:*

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 programme are listed here: <https://doi.org/10.5281/zenodo.4783585>.

Members of Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective are listed here: <https://doi.org/10.5281/zenodo.4790455>.

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. and Mirny, L.A. (2020) ‘Cooler: Scalable storage for Hi-C data and other genomically labeled arrays’, *Bioinformatics*, 36(1), pp. 311–316. Available at: https://doi.org/10.1093/bioinformatics/btz540.

Aken, B.L. *et al.* (2016) ‘The Ensembl gene annotation system’, *Database*, 2016. Available at: https://doi.org/10.1093/database/baw093.

Allio, R. *et al.* (2020) ‘MitoFinder: Efficient automated large‐scale extraction of mitogenomic data in target enrichment phylogenomics’, *Molecular Ecology Resources*, 20(4), pp. 892–905. Available at: https://doi.org/10.1111/1755-0998.13160.

Bernt, M. *et al.* (2013) ‘MITOS: Improved de novo metazoan mitochondrial genome annotation’, *Molecular Phylogenetics and Evolution*, 69(2), pp. 313–319. Available at: https://doi.org/10.1016/j.ympev.2012.08.023.

Brůna, T. *et al.* (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. Available at: https://doi.org/10.1093/nargab/lqaa108.

Challis, R. *et al.* (2020) ‘BlobToolKit – interactive quality assessment of genome assemblies’, *G3: Genes, Genomes, Genetics*, 10(4), pp. 1361–1374. Available at: https://doi.org/10.1534/g3.119.400908.

Cheng, H. *et al.* (2021) ‘Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm’, *Nature Methods*, 18(2), pp. 170–175. Available at: https://doi.org/10.1038/s41592-020-01056-5.

Chow, W. *et al.* (2016) ‘gEVAL — a web-based browser for evaluating genome assemblies’, *Bioinformatics*, 32(16), pp. 2508–2510. Available at: https://doi.org/10.1093/bioinformatics/btw159.

Garrison, E. and Marth, G. (2012) ‘Haplotype-based variant detection from short-read sequencing’. Available at: http://arxiv.org/abs/1207.3907.

Ghurye, J. *et al.* (2019) ‘Integrating Hi-C links with assembly graphs for chromosome-scale assembly’, *PLoS Computational Biology*, 15(8). Available at: https://doi.org/10.1371/journal.pcbi.1007273.

Guan, D. *et al.* (2020) ‘Identifying and removing haplotypic duplication in primary genome assemblies’, *Bioinformatics*, 36(9), pp. 2896–2898. Available at: https://doi.org/10.1093/bioinformatics/btaa025.

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

Howe, K. *et al.* (2021) ‘Significantly improving the quality of genome assemblies through curation’, *GigaScience*. Oxford University Press. Available at: https://doi.org/10.1093/gigascience/giaa153.

Kerpedjiev, P. *et al.* (2018) ‘HiGlass: Web-based visual exploration and analysis of genome interaction maps’, *Genome Biology*, 19(1). Available at: https://doi.org/10.1186/s13059-018-1486-1.

Manni, M. *et al.* (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. Available at: https://doi.org/10.1093/molbev/msab199.

Rao, S.S.P. *et al.* (2014) ‘A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping’, *Cell*, 159(7), pp. 1665–1680. Available at: https://doi.org/10.1016/j.cell.2014.11.021.

Rhie, A. *et al.* (2020) ‘Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies’, *Genome Biology*, 21(1). Available at: https://doi.org/10.1186/s13059-020-02134-9.

Rhie, A. *et al.* (2021) ‘Towards complete and error-free genome assemblies of all vertebrate species’, *Nature*, 592(7856), pp. 737–746. Available at: https://doi.org/10.1038/s41586-021-03451-0.

Simão, F.A. *et al.* (2015) ‘BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs’, *Bioinformatics*, 31(19), pp. 3210–3212. Available at: https://doi.org/10.1093/bioinformatics/btv351.

Surana, P., Muffato, M. and Sadasivan Baby, C. (2023) ‘sanger-tol/genomenote v1.0.dev (v1.0.dev)’. Zenodo. Available at: https://doi.org/10.5281/zenodo.6785935 (Accessed: 17 April 2023).

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

Di Tommaso, P. *et al.* (2017) ‘Nextflow enables reproducible computational workflows’, *Nature Biotechnology*, 35(4), pp. 316–319. Available at: https://doi.org/10.1038/nbt.3820.

Uliano-Silva, M. *et al.* (2022) ‘MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio High Fidelity reads’, *bioRxiv* [Preprint]. Available at: https://doi.org/10.1101/2022.12.23.521667.

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

Vasimuddin, Md. *et al.* (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. Available at: https://doi.org/10.1109/IPDPS.2019.00041.

Zhou, C., McCarthy, S.A. and Durbin, R. (2023) ‘YaHS: yet another Hi-C scaffolding tool’, *Bioinformatics*. Edited by C. Alkan, 39(1). Available at: https://doi.org/10.1093/bioinformatics/btac808.

[add dataset citation]

Wellcome Sanger Institute (2022) [TITLE], European Nucleotide Archive, [dataset], accession number {{ BioProject }}

# Figures

[Figure 1: $SPEC\_IMAGE]

**Figure 1. Photograph of the *{{ species }}* ({{ tolid }}) specimen used for genome sequencing**

[Figure 2: $BTK\_FIG1]

**Figure 2: Genome assembly of *{{ species }}*, {{ assembly\_name }}: 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 *{{ species }}*, {{ assembly\_name }}: 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 *{{ species }}*, {{ assembly\_name }}: 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 *{{ species }}*, {{ assembly\_name }}: Hi-C contact map** **of the {{ assembly\_name }}** 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** ***{{ species }}*, {{ assembly\_name }}**

|  |  |  |
| --- | --- | --- |
| **Project accession data** | | |
| Assembly identifier | {{ assembly\_name }} | |
| Species | *{{ species }}* | |
| Specimen | {{ tolid }} | |
| NCBI taxonomy ID | {{ taxid }} | |
| BioProject | {{ BioProject }} | |
| BioSample ID | {{ BioSample }} | |
| Isolate information | {{ tolid }}  {{ HiC }} | |
| **Assembly metrics\*** | | *Benchmark* |
| Consensus quality (QV) |  | *≥ 50* |
| *k*-mer completeness |  | *≥ 95%* |
| BUSCO\*\* |  | *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\_reads }} | |
| 10X Genomics Illumina | {{ chrom\_reads }} | |
| Hi-C Illumina | {{ HiC\_reads }} | |
| PolyA RNA-Seq Illumina | {{ RNA\_reads }} | |
| **Genome assembly** | | |
| Assembly accession | {{ accession }} | |
| *Accession of alternate haplotype* | {{ alt\_accession }} | |
| Span (Mb) | {{ genome\_length }} | |
| Number of contigs | {{ num\_contigs }} | |
| Contig N50 length (Mb) | {{ contig\_N50 }} | |
| Number of scaffolds | {{ num\_scaffolds }} | |
| Scaffold N50 length (Mb) | {{ scaffold\_N50 }} | |
| Longest scaffold (Mb) | {{ longest\_scaff }} | |

|  |  |
| --- | --- |
| **Genome annotation** | |
| Number of protein-coding genes | $PCG |
| Number of non-coding genes | $CDS\_LENGTH |
| Number of gene transcripts | $EXONS\_PER\_TRANSC |

\* 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 $BUSCO\_VERSION. 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 *{{ species }}*, {{ tolid }}**

{{ chr\_table }}

##### **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> |
| 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> |