DATA NOTE

The genome sequence of the Silver-barred Sober moth, *Aproaerema taeniolella* (Zeller, 1839) [version 1; peer review: awaiting peer review]

Douglas Boyes[1+], Finley Hutchinson[2], Liam M. Crowley[3], Clare Boyes[4],
University of Oxford and Wytham Woods Genome Acquisition Lab,
Darwin Tree of Life Barcoding collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]UK Centre for Ecology & Hydrology, Wallingford, England, UK
[2]University of Exeter, Penryn, England, UK
[3]University of Oxford, Oxford, England, UK
[4]Independent researcher, Welshpool, Wales, UK

+ Deceased author

v1 **First published:** 03 Sep 2024, **9**:500
https://doi.org/10.12688/wellcomeopenres.22890.1

Latest published: 03 Sep 2024, **9**:500
https://doi.org/10.12688/wellcomeopenres.22890.1

Abstract

We present a genome assembly of a female Silver-barred Sober moth *Aproaerema taeniolella* (Arthropoda; Insecta; Lepidoptera; Gelechiidae). The genome sequence has a length of 636.60 megabases. Most of the assembly is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled and is 15.19 kilobases in length. Gene annotation of this assembly on Ensembl identified 22,274 protein-coding genes.

Keywords

Aproaerema taeniolella, Silver-barred Sober moth, genome sequence, chromosomal, Lepidoptera

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

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

This article is included in the Tree of Life gateway.

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

Author roles: Boyes D: Investigation, Resources; **Hutchinson F**: Investigation, Resources; **Crowley LM**: Investigation, Resources; **Boyes C**: Writing – Original Draft Preparation;

Competing interests: No competing interests were disclosed.

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

How to cite this article: Boyes D, Hutchinson F, Crowley LM *et al.* **The genome sequence of the Silver-barred Sober moth, *Aproaerema taeniolella* (Zeller, 1839) [version 1; peer review: awaiting peer review]** Wellcome Open Research 2024, **9**:500 https://doi.org/10.12688/wellcomeopenres.22890.1

First published: 03 Sep 2024, **9**:500 https://doi.org/10.12688/wellcomeopenres.22890.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Amphiesmenoptera; Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Gelechioidea; Gelechiidae; Anacampsinae; *Aproaerema*; *Aproaerema taeniolella* (Zeller, 1839) (NCBI:txid2566309).

Background

Aproaerema taeniolella (Silver-barred Sober) is a micro-moth in the family Geleciidae which is local in England and Wales, becoming rare in Scotland (Sterling *et al.*, 2023). It is found throughout Europe although it is often local (GBIF Secretariat, 2024).

Aproaerema taeniolla is a blackish moth and is one of several similar species which could be readily confused. The adult moth has a wingspan of between 11–14 mm and often has a whitish band about two-thirds of the way along the forewing but sometime this is absent. The adult has one generation a year flying between June and August. It readily comes to light, and during daytime adults can sometimes be disturbed from the late afternoon (Emmet & Langmaid, 2002). Little is known about the life cycle of this moth. It is assumed to lay its eggs on the foodplants, *Lotus corniculatus* or less usually *Lotus pedunculatus*. The stage in which it overwinters is unknown. Larvae are occasionally found during mid to late April in leaf and shoot spinnings. They pupate in the ground litter (Langmaid *et al.*, 2018).

The genome of *Aproaerema taeniolella* 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 *Aproaerema taeniolella* based on one female specimen from Wytham Woods, Oxfordshire, UK.

Genome sequence report

The genome of an adult female *Aproaerema taeniolella* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of 20.65 Gb (gigabases) from 2.13 million reads, providing approximately 33-fold coverage. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data, which produced 83.06 Gbp from 550.06 million reads, yielding an approximate coverage of 130-fold. Specimen and sequencing information is summarised in Table 1.

Manual assembly curation corrected 151 missing joins or mis-joins and 33 haplotypic duplications, reducing the assembly length by 0.85% and the scaffold number by 18.12%, and increasing the scaffold N50 by 2.17%. The final assembly has a total length of 636.60 Mb in 383 sequence scaffolds, with 1,121 gaps. The scaffold N50 is 20.4 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 (96.03%) of the assembly sequence was assigned to 31 chromosomal-level scaffolds, representing 30 autosomes and the Z sex chromosome. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 3). 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.

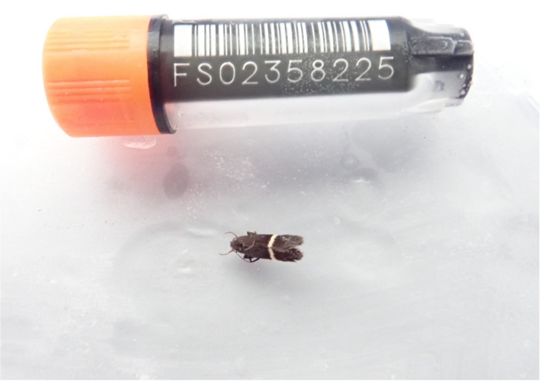

Figure 1. Photograph of the *Aproaerema taeniolella* (ilAprTaen1) specimen used for genome sequencing.

The estimated Quality Value (QV) of the final assembly is 59.1 with *k*-mer completeness of 100.0%, and the assembly has a BUSCO v5.3.2 completeness of 96.6% (single = 95.5%, duplicated = 1.1%), using the lepidoptera_odb10 reference set (*n* = 5,286).

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/2566309.

Genome annotation report

The *Aproaerema taeniolella* genome assembly (GCA_949987775.1) was annotated at the European Bioinformatics Institute (EBI) on Ensembl Rapid Release. The resulting annotation includes 22,449 transcribed mRNAs from 22,274 protein-coding genes (Table 2; https://rapid.ensembl.org/Aproaerema_taeniolella_GCA_949987775.1/Info/Index). The average transcript length is 7,802.54. There are 1.01 coding transcripts per gene and 4.88 exons per transcript.

Methods

Sample acquisition

The *Aproaerema taeniolella* specimens used for genome sequencing (specimen ID Ox000657, ToLID ilAprTaen1) and Hi-C scaffolding (specimen ID Ox000658, ToLID ilAprTaen2) were adult specimens from Wytham Woods, Oxfordshire (biological

Table 1. Specimen and sequencing data for *Aproaerema taeniolella*.

Project information			
Study title	*Aproaerema taeniolella* (silver-barred sober)		
Umbrella BioProject	PRJEB56804		
Species	*Aproaerema taeniolella*		
BioSample	SAMEA7701519		
NCBI taxonomy ID	2566309		
Specimen information			
Technology	**ToLID**	**BioSample accession**	**Organism part**
PacBio long read sequencing	ilAprTaen1	SAMEA7701701	Whole organism
Hi-C sequencing	ilAprTaen2	SAMEA7701702	Whole organism
RNA sequencing	ilAprTaen3	SAMEA113425860	Whole organism
Sequencing information			
Platform	**Run accession**	**Read count**	**Base count (Gb)**
Hi-C Illumina NovaSeq 6000	ERR10395985	5.50e+08	83.06
PacBio Sequel IIe	ERR10395979	2.13e+06	20.65
RNA Illumina NovaSeq 6000	ERR12245527	5.99e+07	9.04

vice-county Berkshire). UK (latitude 51.77, longitude –1.34). The specimens were collected and identified by Douglas Boyes (University of Oxford) and preserved on dry ice.

The specimen used for RNA sequencing (specimen ID Ox003071, ToLID ilAprTaen3) was an adult specimen collected from the same location on 2022-07-22, using a light trap. The specimen was collected by Finley Hutchinson and Liam Crowley and identified by Finley Hutchinson and then preserved on dry ice.

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

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of core procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io (Denton *et al.*, 2023b).

In sample preparation, the ilAprTaen1 sample was weighed and dissected on dry ice (Jay *et al.*, 2023). Tissue from the whole organism was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a). HMW DNA was extracted using the Automated MagAttract v1 protocol (Sheerin *et al.*, 2023). DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 30 (Todorovic *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Strickland *et al.*, 2023). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

RNA was extracted from whole organism tissue of ilAprTaen3 in the Tree of Life Laboratory at the WSI using the RNA

Table 2. Genome assembly data for *Aproaerema taeniolella*, ilAprTaen1.1.

Genome assembly		
Assembly name	ilAprTaen1.1	
Assembly accession	GCA_949987775.1	
Accession of alternate haplotype	*GCA_950005065.1*	
Span (Mb)	636.60	
Number of contigs	1,505	
Contig N50 length (Mb)	1.0	
Number of scaffolds	383	
Scaffold N50 length (Mb)	20.4	
Longest scaffold (Mb)	39.05	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	59.1	*≥ 50*
k-mer completeness	100.0%	*≥ 95%*
BUSCO**	C:96.6%[S:95.5%,D:1.1%], F:0.7%,M:2.7%,n:5,286	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	96.03%	*≥ 95%*
Sex chromosomes	Z	*localised homologous pairs*
Organelles	Mitochondrial genome: 15.19 kb	*complete single alleles*
Genome annotation of assembly GCA_949987775.1 at Ensembl		
Number of protein-coding genes	22,274	
Number of gene transcripts	22,449	

* 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 lepidoptera_odb10 BUSCO set using version 5.3.2. 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 https://blobtoolkit.genomehubs.org/view/ilAprTaen1_1/dataset/ilAprTaen1_1/busco.

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 DNA sequencing libraries were constructed according to the manufacturers' instructions. Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit. DNA and RNA sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences Sequel IIe (HiFi) and Illumina NovaSeq 6000 (RNA-Seq) instruments. Hi-C data were also generated from whole organism tissue of ilAprTaen2 using the Arima-HiC v2 kit. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on the Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

Assembly

The HiFi reads were first assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Cheng *et al.*, 2021). 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

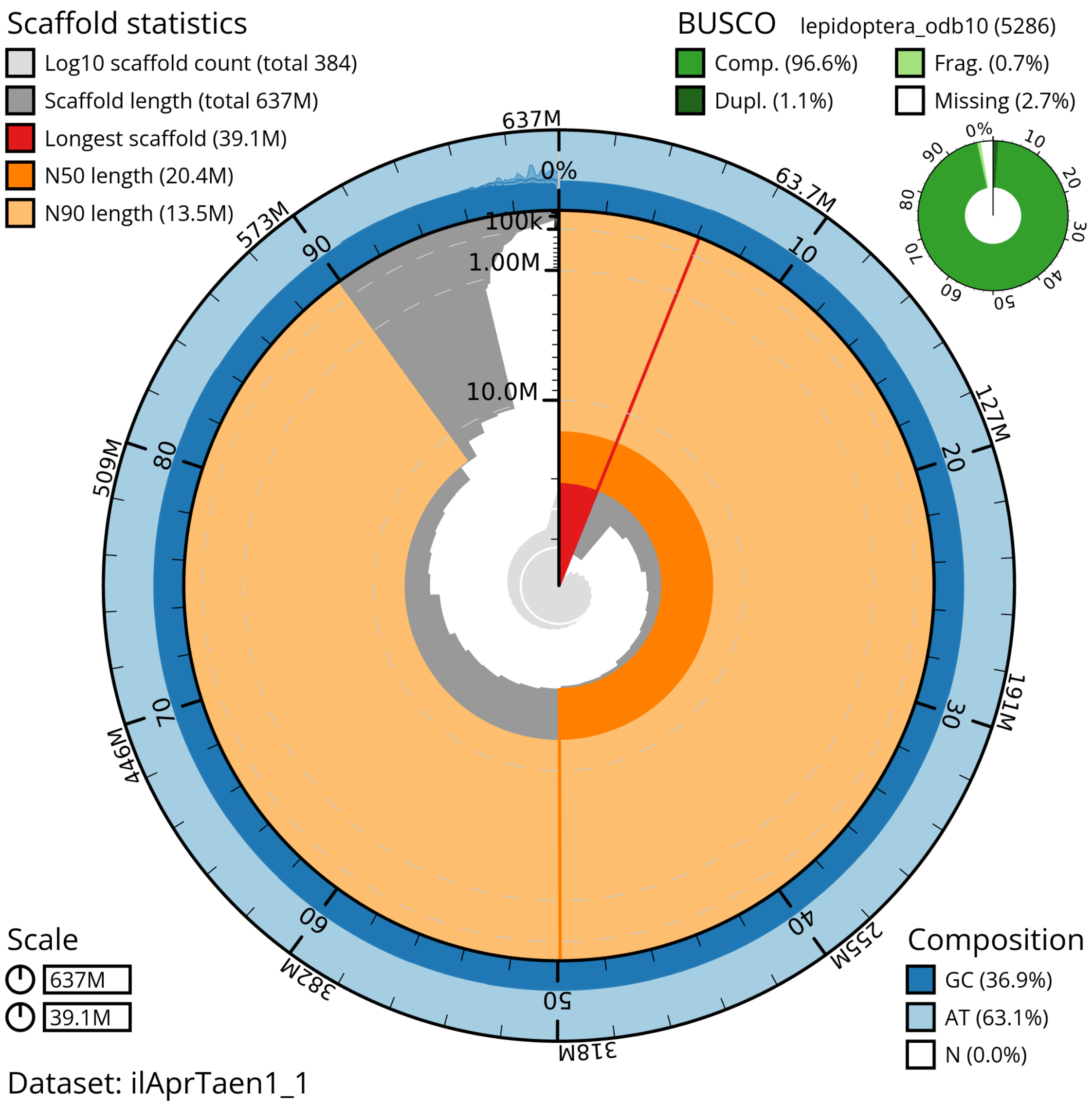

Figure 2. Genome assembly of *Aproaerema taeniolella*, ilAprTaen1.1: metrics. The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 636,615,290 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (39,052,235 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (20,393,386 and 13,545,650 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the lepidoptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilAprTaen1_1/dataset/ilAprTaen1_1/snail.

(Rao *et al.*, 2014) in YaHS (Zhou *et al.*, 2023) with 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

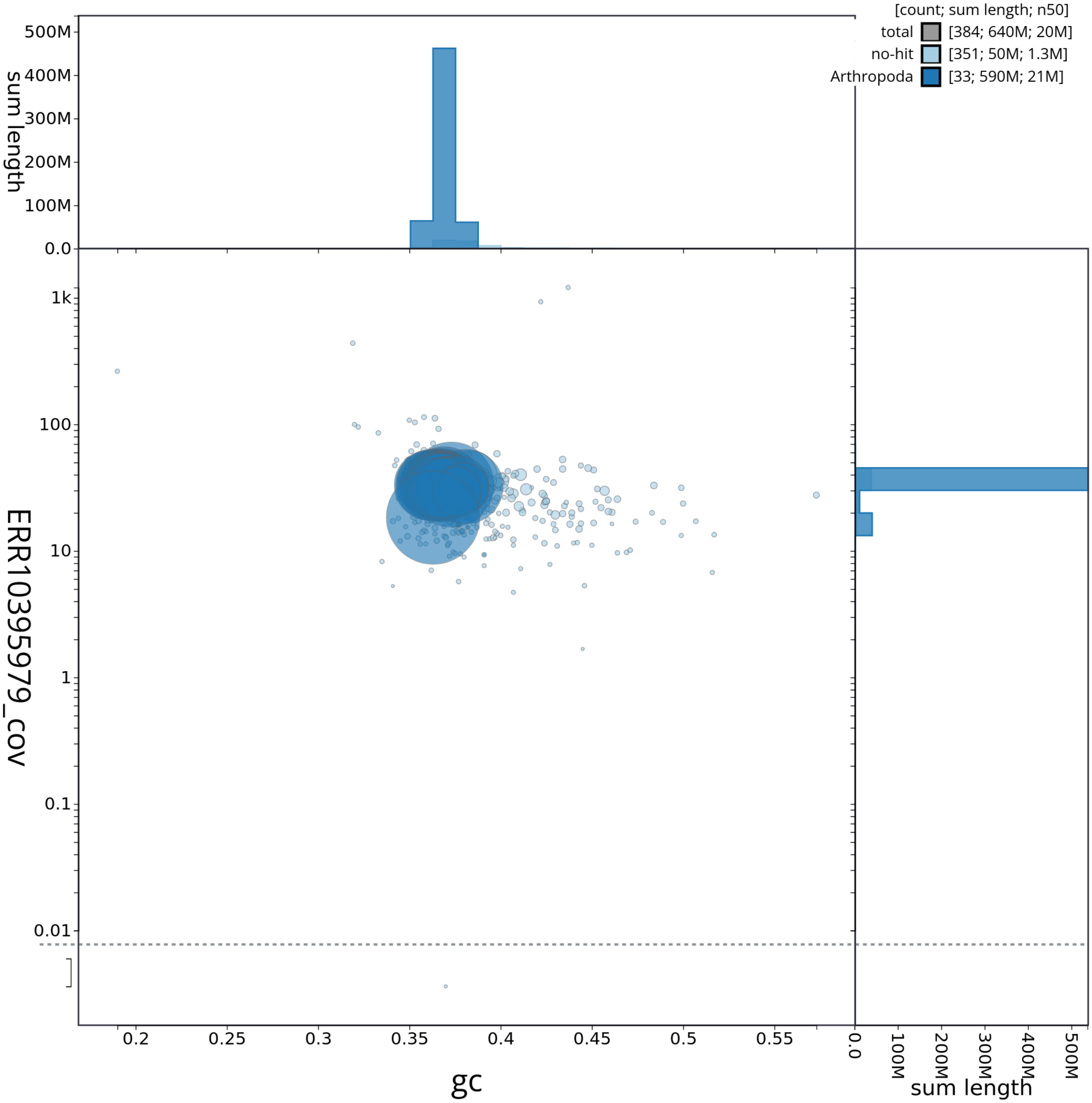

Figure 3. Genome assembly of *Aproaerema taeniolella*, ilAprTaen1.1: BlobToolKit GC-coverage plot. Blob plot of base coverage in ERR10395979 against GC proportion for sequences in assembly ilAprTaen1_1. Sequences are coloured by phylum. Circles are sized in proportion to sequence length. Histograms show the distribution of sequence length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilAprTaen1_1/dataset/ilAprTaen1_1/blob.

Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were corrected, and duplicate sequences were tagged and removed. The sex chromosome was identified by read coverage statistics. The process is documented at https://gitlab.com/wtsi-grit/rapid-curation (article in preparation).

Evaluation of the final assembly

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

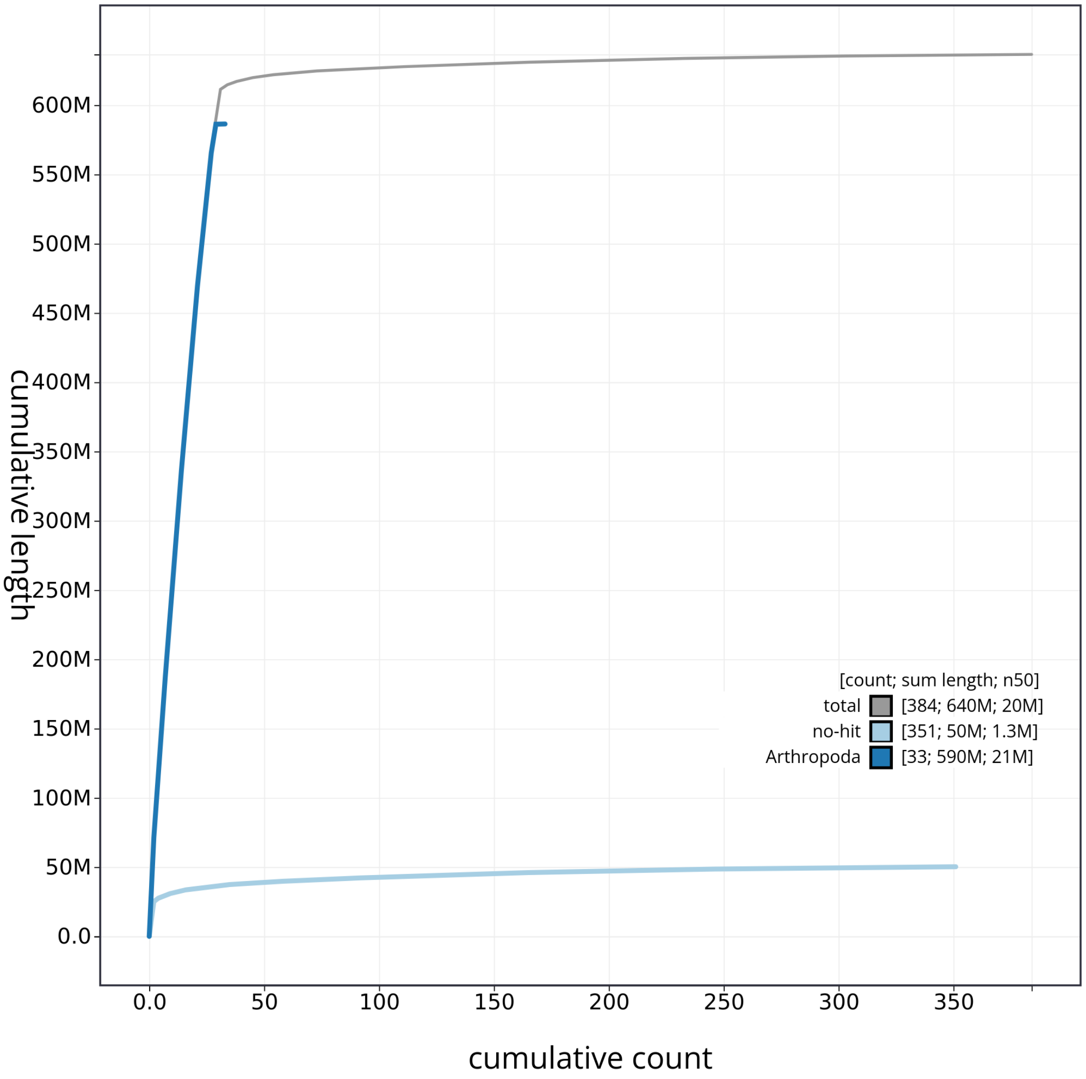

Figure 4. Genome assembly of *Aproaerema taeniolella* ilAprTaen1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all sequences. Coloured lines show cumulative lengths of sequences assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilAprTaen1_1/dataset/ilAprTaen1_1/cumulative.

work was done using the “sanger-tol/readmapping” (Surana *et al.*, 2023a) and “sanger-tol/genomenote” (Surana *et al.*, 2023b) pipelines. The genome readmapping pipelines were developed using the nf-core tooling (Ewels *et al.*, 2020), use MultiQC (Ewels *et al.*, 2016), and make extensive use of the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), and the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions. The genome was also analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021; Simão *et al.*, 2015) were calculated.

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

Genome annotation

The BRAKER2 pipeline (Brůna *et al.*, 2021) was used in the default protein mode to generate annotation for the *Aproaerema taeniolella* assembly (GCA_949987775.1) in Ensembl Rapid Release at the EBI.

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

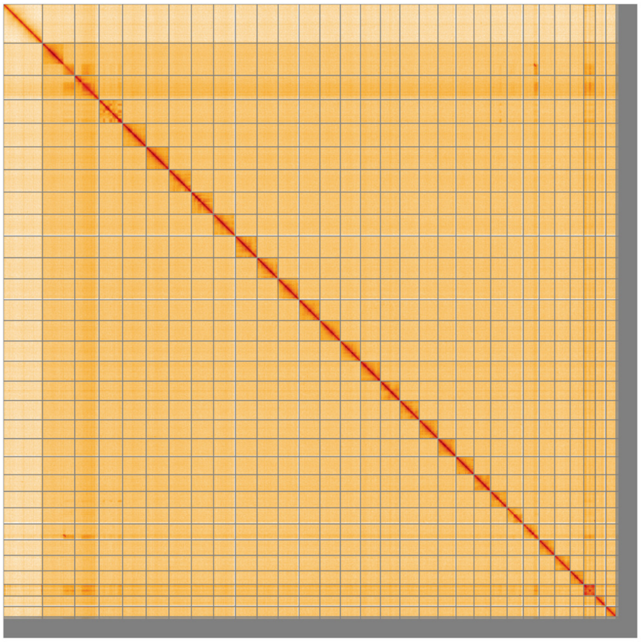

Figure 5. Genome assembly of *Aproaerema taeniolella* ilAprTaen1.1: Hi-C contact map of the ilAprTaen1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=aaRWp2vBRhuTj2_XDD5weA,.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Aproaerema taeniolella*, ilAprTaen1.

INSDC accession	Name	Length (Mb)	GC%
OX465243.1	1	32.14	37.5
OX465244.1	2	24.26	38.0
OX465245.1	3	23.52	37.0
OX465246.1	4	23.48	37.0
OX465247.1	5	22.74	36.5
OX465248.1	6	22.41	36.5
OX465249.1	7	22.11	36.0
OX465250.1	8	21.8	36.5
OX465251.1	9	21.63	36.0
OX465252.1	10	21.02	36.5
OX465253.1	11	20.88	36.5
OX465254.1	12	20.78	36.5
OX465255.1	13	20.39	36.5
OX465256.1	14	20.14	36.0
OX465257.1	15	19.89	36.5
OX465258.1	16	19.39	36.5
OX465259.1	17	19.25	36.5
OX465260.1	18	18.77	36.5
OX465261.1	19	17.98	36.5
OX465262.1	20	17.97	37.0
OX465263.1	21	16.68	37.0
OX465264.1	22	16.37	37.0
OX465265.1	23	16.08	37.0
OX465266.1	24	15.86	37.5
OX465267.1	25	15.52	36.5
OX465268.1	26	15.47	37.0
OX465269.1	27	13.55	37.0
OX465270.1	28	11.57	38.5
OX465271.1	29	10.45	38.0
OX465272.1	30	10.22	37.5
OX465242.1	Z	39.05	36.5
OX465273.1	MT	0.02	19.5

Table 4. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	4.2.1	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.3.2	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Hifiasm	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	1.11.6	https://github.com/higlass/higlass
Merqury.FK	d00d98157618f4e8d1a9190026b19b471055b22e	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

submission of materials by a Darwin Tree of Life Partner is subject to the **'Darwin Tree of Life Project Sampling Code of Practice'**, which can be found in full on the Darwin Tree of Life website here. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they 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: *Aproaerema taeniolella* (silver-barred sober). Accession number PRJEB56804; https://identifiers.org/ena.embl/PRJEB56804 (Wellcome Sanger Institute, 2023). The genome sequence is released openly for reuse. The *Aproaerema taeniolella* 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. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

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

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

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

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

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

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

References

Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316.
PubMed Abstract | Publisher Full Text | Free Full Text

Allio R, Schomaker-Bastos A, Romiguier J, *et al.*: **MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics.** *Mol Ecol Resour.* 2020; **20**(4): 892–905.
PubMed Abstract | Publisher Full Text | Free Full Text

Beasley J, Uhl R, Forrest LL, *et al.*: **DNA barcoding SOPs for the Darwin Tree of Life project.** *Protocols.io.* 2023. [Accessed 25 June 2024].
Publisher Full Text

Brůna T, Hoff KJ, Lomsadze A, *et al.*: **BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database.** *NAR Genom Bioinform.* 2021; **3**(1): lqaa108.
PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374.
PubMed Abstract | Publisher Full Text | Free Full Text

Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175.
PubMed Abstract | Publisher Full Text | Free Full Text

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial arthropod fauna [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 123.
PubMed Abstract | Publisher Full Text | Free Full Text

da Veiga Leprevost F, Grüning BA, Alves Aflitos S, *et al.*: **BioContainers: an open-source and community-driven framework for software standardization.** *Bioinformatics.* 2017; **33**(16): 2580–2582.
PubMed Abstract | Publisher Full Text | Free Full Text

Denton A, Oatley G, Cornwell C, *et al.*: **Sanger Tree of Life sample homogenisation: PowerMash.** *protocols.io.* 2023a.
Publisher Full Text

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

Diesh C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of synteny and structural variation.** *Genome Biol.* 2023; **24**(1): 74.
PubMed Abstract | Publisher Full Text | Free Full Text

do Amaral RJV, Bates A, Denton A, *et al.*: **Sanger Tree of Life RNA extraction: automated MagMax™ mirVana.** *Protocols.io.* 2023.
Publisher Full Text

Emmet AM, Langmaid JR: **The moths and butterflies of great Britain and Ireland**. Gelechiidae. Colchester: Harley Books, 2002.
Reference Source

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
PubMed Abstract | Publisher Full Text

Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
PubMed Abstract | Publisher Full Text | Free Full Text

GBIF Secretariat: ***Aproaerema taeniolella* (Zeller, 1839).** *GBIF Backbone Taxonomy.* 2024; [Accessed 16 July 2024].
Reference Source

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
PubMed Abstract | Publisher Full Text | Free Full Text

Harry E: **PretextView (Paired Read Texture Viewer): a desktop application for viewing pretext contact maps.** 2022; [Accessed 19 October 2022].
Reference Source

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *Gigascience.* 2021; **10**(1): giaa153.
PubMed Abstract | Publisher Full Text | Free Full Text

Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life sample preparation: triage and dissection.** *protocols.io.* 2023.
Publisher Full Text

Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HiGlass: web-based visual exploration and analysis of genome interaction maps.** *Genome Biol.* 2018; **19**(1): 125.
PubMed Abstract | Publisher Full Text | Free Full Text

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
PubMed Abstract | Publisher Full Text | Free Full Text

Langmaid JR, Palmer S, Young MR: **A field guide to the smaller moths of great Britain and Ireland. 3rd ed**. British Entomological and Natural History Society, 2018.
Reference Source

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | Publisher Full Text | Free Full Text

Merkel D: **Docker: lightweight linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2.
Reference Source

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

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

Sheerin E, Sampaio F, Oatley G, *et al.*: **Sanger Tree of Life HMW DNA extraction: automated MagAttract v.1.** *Protocols.io.* 2023. [Accessed 21 November 2023].
Publisher Full Text

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–3212.
PubMed Abstract | Publisher Full Text

Sterling P, Parsons M, Lewington R: **Field guide to the micro-moths of great Britain and Ireland, Second Edition**. London: Bloomsbury Publishing, 2023.
Reference Source

Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *Protocols.io.* 2023.
Publisher Full Text

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

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

Todorovic M, Sampaio F, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor®3 for PacBio HiFi.** *Protocols.io.* 2023.
Publisher Full Text

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer

review: awaiting peer review]. *Wellcome Open Res.* 2024; **9**: 339.
Publisher Full Text

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems**. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2019; 314–324.
Publisher Full Text

Wellcome Sanger Institute: **The genome sequence of the Silver-barred Sober moth, *Aproaerema taeniolella* (Zeller, 1839).** European Nucleotide Archive. [dataset], accession number PRJEB56804, 2023.

Zhou C, McCarthy SA, Durbin R: **YaHS: yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**