





DATA NOTE

The genome sequence of a snail-killing fly, *Dichetophora*

obliterata (Fabricius, 1805)

[version 1; peer review: 2 approved]

Liam M. Crowley¹, Olga Sivell ², Ryan Mitchell³, Duncan Sivell ²,
University of Oxford and Wytham Woods Genome Acquisition Lab,
Natural History Museum Genome Acquisition Lab,
Darwin Tree of Life Barcoding collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory
team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

¹University of Oxford, Oxford, England, UK

²Natural History Museum, London, England, UK

³Independent researcher, Sligo, County Sligo, Ireland

V1 First published: 08 Apr 2025, 10:176
<https://doi.org/10.12688/wellcomeopenres.23993.1>
Latest published: 08 Apr 2025, 10:176
<https://doi.org/10.12688/wellcomeopenres.23993.1>

Abstract

We present a genome assembly from a female specimen of *Dichetophora obliterata* (snail-killing fly; Arthropoda; Insecta; Diptera; Sciomyzidae). The genome sequence has a total length of 1,312.79 megabases. Most of the assembly (99.78%) is scaffolded into 6 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 21.36 kilobases. Gene annotation of this assembly on Ensembl identified 15,139 protein-coding genes.

Keywords




Dichetophora obliterata, snail-killing fly, genome sequence, chromosomal, Diptera



This article is included in the [Tree of Life](#) gateway.

Open Peer Review

Approval Status  

	1	2
version 1 08 Apr 2025	 view	 view
1. Denise Gemmellaro , Kean University, Union, USA		
2. Taro Nakamura  , National Institute for Basic Biology, Okazaki, Japan		
Any reports and responses or comments on the article can be found at the end of the article.		

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

Author roles: Crowley LM: Investigation, Resources; Sivell O: Investigation, Resources; Mitchell R: Investigation, Resources; Sivell D: Investigation, Resources;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540) 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.

Copyright: © 2025 Crowley LM *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Crowley LM, Sivell O, Mitchell R *et al.* **The genome sequence of a snail-killing fly, *Dichetophora obliterata* (Fabricius, 1805) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, 10:176 <https://doi.org/10.12688/wellcomeopenres.23993.1>

First published: 08 Apr 2025, 10:176 <https://doi.org/10.12688/wellcomeopenres.23993.1>

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Eremoneura; Cyclorrhapha; Schizophora; Acalyptratae; Sciomyzoidea; Sciomyzidae; *Dichetophora*; *Dichetophora obliterata* (Fabricius, 1805) (NCBI:txid1226583)

Background

The chalk snailkiller, *Dichetophora obliterata*, is a species of sciomyzid fly with a widespread distribution across Europe, North Africa, and parts of the Middle East (GBIF Secretariat, 2023). It has been recorded in Scotland, England, Belgium, the Netherlands, France, Austria, Germany, Switzerland, Spain, Italy, the Czech Republic, Romania, Greece, Morocco, Turkey, Iraq, and Iran (GBIF Secretariat, 2023).

In Britain, it is largely restricted to the southern half of the country, where it is associated with calcareous habitats, including chalk downland, limestone quarries, brownfield sites, and stabilised coastal dunes. It is common in South Wales and south-west England but has not been recorded north of Teesside (Falk, 2025; NBN Atlas Partnership, 2023).

The adult fly is characterised by a grey-striped thorax, a reddish abdomen, yellow legs with orange femoral bands, and a net-like pattern on the leading edge of the wings. The species is univoltine, producing a single generation per year. There is a long flight period from mid- or late May to mid-October (Ball, 2017). Eggs are laid on the shells of living terrestrial snails. The larvae are parasitoids of terrestrial snails, with early instars developing inside smaller species such as *Lauria cylindracea* and later instars feeding on larger snails, including *Helicella* and *Theba* spp. (Ball, 2017; Gedling Conservation Trust, 2023). Overwintering occurs as a mature larva or puparium, with pupation taking place either inside the host shell or separately (Ball, 2017).

The genome of *Dichetophora obliterata* was sequenced as part of the Darwin Tree of Life Project, based on a specimen from Wytham Woods, Oxfordshire, United Kingdom (Figure 1).



Figure 1. Photograph of the *Dichetophora obliterata* (idDicObli2) specimen used for genome sequencing.

Genome sequence report

Sequencing data

The genome of an adult specimen of *Dichetophora obliterata* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 45.52 Gb from 4.42 million reads. GenomeScope analysis of the PacBio HiFi data estimated the haploid genome size at 1,288.93 Mb, with a heterozygosity of 1.01% and repeat content of 46.76%. These values provide an initial assessment of genome complexity and the challenges anticipated during assembly. Based on this estimated genome size, the sequencing data provided approximately 34.0x coverage of the genome. Chromosome conformation Hi-C sequencing produced 139.66 Gb from 924.89 million reads. Table 1 summarises the specimen and sequencing information.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The assembly was improved by manual curation, which corrected 127 misjoins or missing joins and removed 12 haplotypic duplications. These interventions decreased the scaffold count by 41.54% and increased the scaffold N50 by 2.33%. The final assembly has a total length of 1,312.79 Mb in 75 scaffolds, with 627 gaps, and a scaffold N50 of 266.48 Mb (Table 2).

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

Most of the assembly sequence (99.77%) was assigned to 6 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3). The specimen is homogametic (female). We did not identify the sex chromosome(s) as sequence data from the heterogametic sex was not available and homology is unreliable for sex chromosome identification in Diptera due to frequent sex chromosome turnover (Vicoso & Bachtrog, 2015).

The mitochondrial genome was also assembled. This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

The estimated Quality Value (QV) and *k*-mer completeness metrics, along with BUSCO completeness scores, were calculated for each haplotype and the combined assembly. The QV reflects the base-level accuracy of the assembly, while *k*-mer completeness indicates the proportion of expected *k*-mers identified in the assembly. BUSCO scores provide a measure of completeness based on benchmarking universal single-copy orthologues.

The primary haplotype has a QV of 59.4, and the combined primary and alternate assemblies achieve an estimated QV

Table 1. Specimen and sequencing data for *Dichetophora obliterata*.

Project information			
Study title	Dichetophora obliterata (chalk snailkiller)		
Umbrella BioProject	PRJEB70855		
Species	<i>Dichetophora obliterata</i>		
BioSpecimen	SAMEA10979120		
NCBI taxonomy ID	1226583		
Specimen information			
Technology	ToLID	BioSample accession	Organism part
PacBio long read sequencing	idDicObli2	SAMEA10979528	head and thorax
Hi-C sequencing	idDicObli2	SAMEA10979528	head and thorax
RNA sequencing	idDicObli1	SAMEA11025277	Whole organism
Sequencing information			
Platform	Run accession	Read count	Base count (Gb)
Hi-C Illumina NovaSeq 6000	ERR12356296	9.25e+08	139.66
PacBio Sequel IIe	ERR12353024	2.58e+06	25.01
PacBio Sequel IIe	ERR12353025	1.85e+06	20.5
RNA Illumina NovaSeq 6000	ERR12356297	7.90e+07	11.94

of 59.4. The *k*-mer recovery for the primary haplotype is 81.63%, and for the alternate haplotype it is 81.07%. The combined primary and alternate assemblies display a *k*-mer recovery of 97.24%. BUSCO analysis using the diptera_odb10 reference set (*n* = 3,285) identified 98.1% of the expected gene set (single = 95.3%, duplicated = 2.7%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project (EBP) Report on Assembly Standards September 2024. The assembly achieves the EBP reference standard of 6.C.Q59.

Genome annotation report

The *Dichetophora obliterata* genome assembly (GCA_963920525.1) was annotated at the European Bioinformatics Institute (EBI) on Ensembl Rapid Release. The resulting annotation includes 22,552 transcribed mRNAs from 15,139 protein-coding and 706 non-coding genes (Table 2; https://rapid.ensembl.org/Dichetophora_oblitterata_GCA_963920525.1/Info/Index). The average transcript length is 25,095.71. There are 1.42 coding transcripts per gene and 5.09 exons per transcript.

Methods

Sample acquisition and DNA barcoding

An adult female *Dichetophora obliterata* (specimen ID Ox001861, ToLID idDicObli2) was collected from Wytham Woods, Oxfordshire, United Kingdom (latitude 51.77,

longitude −1.34) on 2021-09-03 by netting. The specimen was collected and identified by Liam Crowley (University of Oxford) and preserved on dry ice. This specimen was used for DNA and Hi-C sequencing.

The specimen used for RNA sequencing (specimen ID NHMUK014036990, ToLID idDicObli1) was collected from Dry Sandford Pit, England, United Kingdom (latitude 51.69, longitude −1.33) on 2021-06-19, using an aerial net. The specimen was collected by Ryan Mitchell (independent researcher) and Olga Sivell (Natural History Museum), identified by Duncan Sivell (Natural History Museum) and preserved in liquid nitrogen.

The initial identification by Expert Id 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 specimen and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI) (Pereira *et al.*, 2022). 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

Table 2. Genome assembly data for *Dichetophora obliterata*.

Genome assembly		
Assembly name	idDicObli2.1	
Assembly accession	GCA_963920525.1	
Alternate haplotype accession	GCA_963920515.1	
Assembly level for primary assembly	chromosome	
Span (Mb)	1,312.79	
Number of contigs	702	
Number of scaffolds	75	
Longest scaffold (Mb)	316.54	
Assembly metrics	Measure	Benchmark
Contig N50 length	5.39 Mb	≥ 1 Mb
Scaffold N50 length	266.48 Mb	= chromosome N50
Consensus quality (QV)	Primary: 59.4; alternate: 59.5; combined 59.4	≥ 40
k-mer completeness	Primary: 81.63%; alternate: 81.07%; combined: 97.24%	$\geq 95\%$
BUSCO*	C:98.1%[S:95.3%,D:2.7%], F:0.3%,M:1.6%,n:3,285	$S > 90\%$, $D < 5\%$
Percentage of assembly mapped to chromosomes	99.77%	$\geq 90\%$
Sex chromosomes	Not identified	localised homologous pairs
Organelles	Mitochondrial genome: 21.36 kb	complete single alleles
Genome annotation of assembly GCA_963920525.1 at Ensembl		
Number of protein-coding genes	15,139	
Number of non-coding genes	706	
Number of gene transcripts	22,552	

* BUSCO scores based on the diptera_odb10 BUSCO set using version 5.5.0. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison.

of Life barcoding have been deposited on protocols.io (Beasley *et al.*, 2023).

Metadata collection for samples adhered to the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

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 procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io (Denton *et al.*, 2023b). The idDicObli2 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the head and thorax was

homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a).

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023). 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, 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.

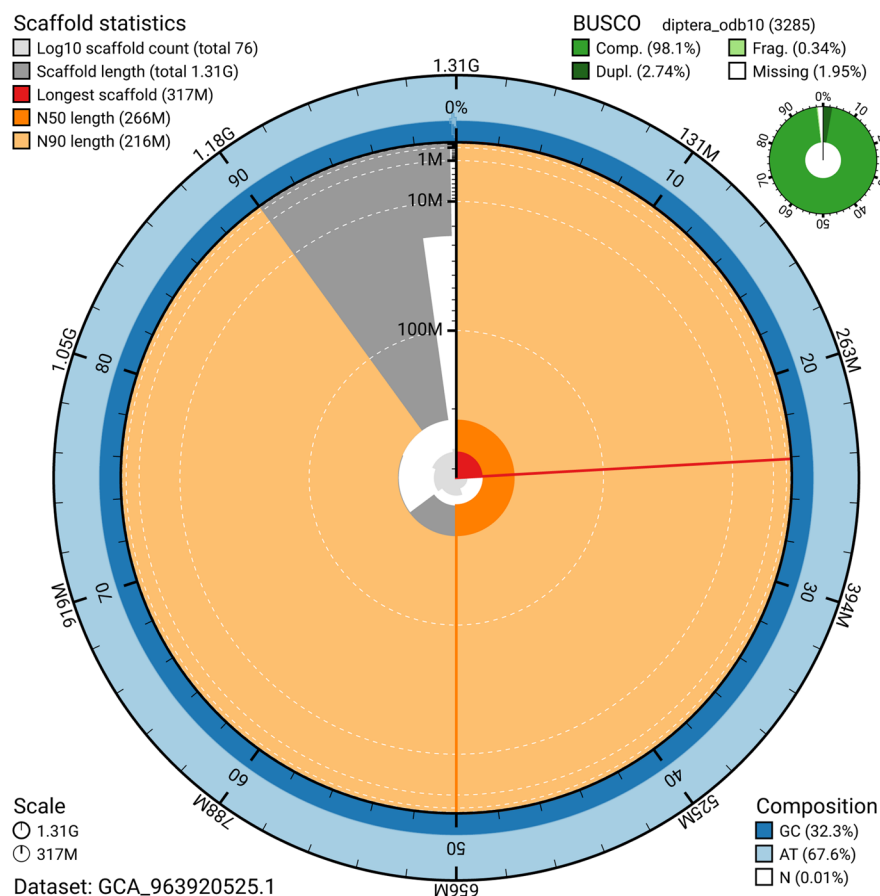


Figure 2. Genome assembly of *Dichetophora oblitterata*, idDicObli2.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the diptera_odb10 set is presented at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963920525.1/dataset/GCA_963920525.1/snail.

RNA was extracted from tissue of idDicObli1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ mirVana 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.

Hi-C sample preparation

Tissue from the head and thorax of the idDicObli2 sample was processed for Hi-C sequencing at the WSI Scientific Operations core, using the Arima-HiC v2 kit. In brief, 20–50 mg of frozen tissue (stored at -80°C) was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde concentration. After crosslinking, the tissue was homogenised using the Diagnocine Power Masher-II and BioMasher-II tubes and pestles. Following the Arima-HiC v2 kit manufacturer's instructions, crosslinked DNA was digested using a restriction

enzyme master mix. The 5'-overhangs were filled in and labelled with biotinylated nucleotides and proximally ligated. An overnight incubation was carried out for enzymes to digest remaining proteins and for crosslinks to reverse. A clean up was performed with SPRIselect beads prior to library preparation. Additionally, the biotinylation percentage was estimated using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit and Arima-HiC v2 QC beads.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core.

PacBio HiFi

At a minimum, samples were required to have an average fragment size exceeding 8 kb and a total mass over 400 ng to proceed to the low input SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences, California, USA), depending on genome size and sequencing depth required. Libraries were prepared

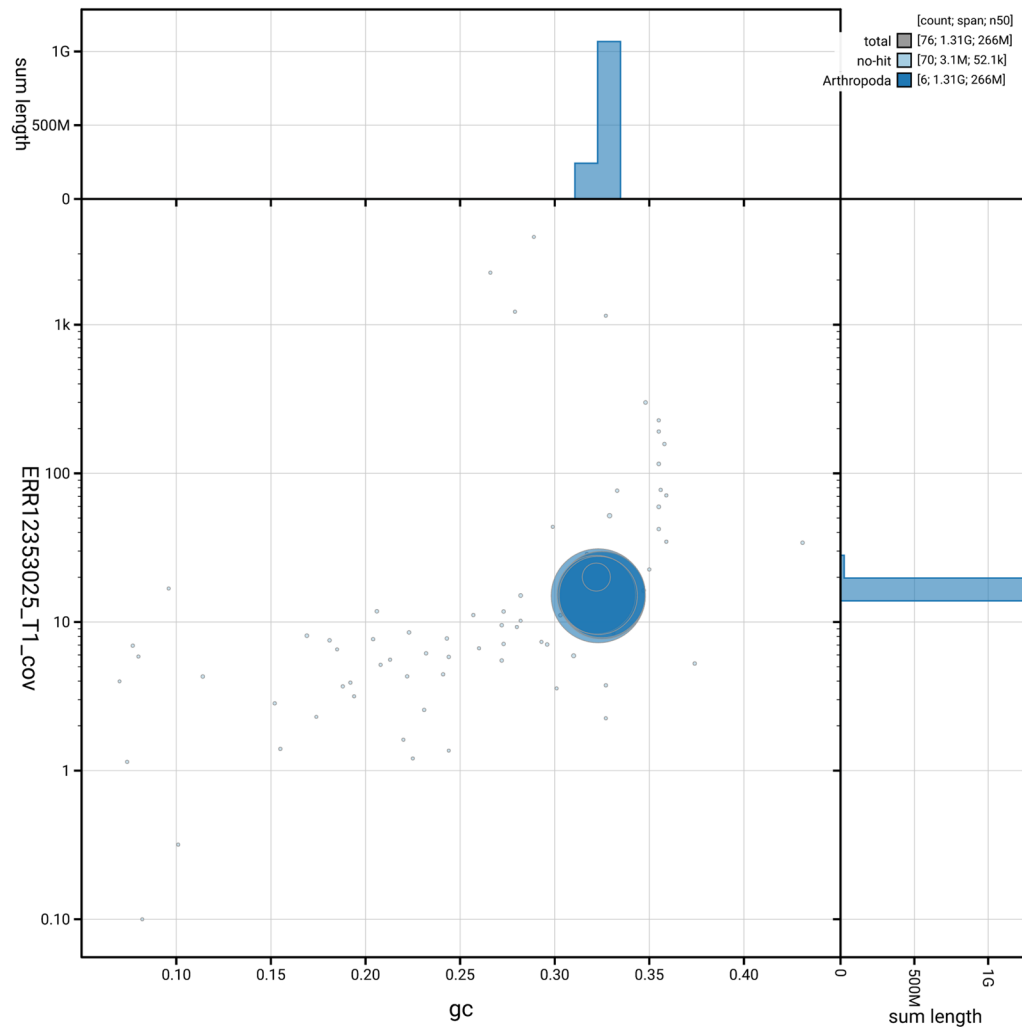


Figure 3. Genome assembly of *Dichetophora oblitterata*, idDicObli2.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963920525.1/dataset/GCA_963920525.1/blob.

using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA) as per the manufacturer's instructions. The kit includes the reagents required for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead cleanup, and nuclease treatment. Following the manufacturer's instructions, size selection and clean up was carried out using diluted AMPure PB beads (Pacific Biosciences, California, USA). DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) with Qubit 1X dsDNA HS assay kit and the final library fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and gDNA 55kb BAC analysis kit.

Samples were sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was in the range 40–135 pM. The SMRT link software, a PacBio web-based end-to-end

workflow manager, was used to set-up and monitor the run, as well as perform primary and secondary analysis of the data upon completion.

Hi-C

For Hi-C library preparation, DNA was fragmented using the Covaris E220 sonicator (Covaris) and size selected using SPRISelect beads to 400 to 600 bp. The DNA was then enriched using the Arima-HiC v2 kit Enrichment beads. Using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) for end repair, a-tailing, and adapter ligation. This uses a custom protocol which resembles the standard NEBNext Ultra II DNA Library Prep protocol but where library preparation occurs while DNA is bound to the Enrichment beads. For library amplification, 10 to 16 PCR cycles were required, determined by the sample biotinylation percentage. The Hi-C sequencing was performed using

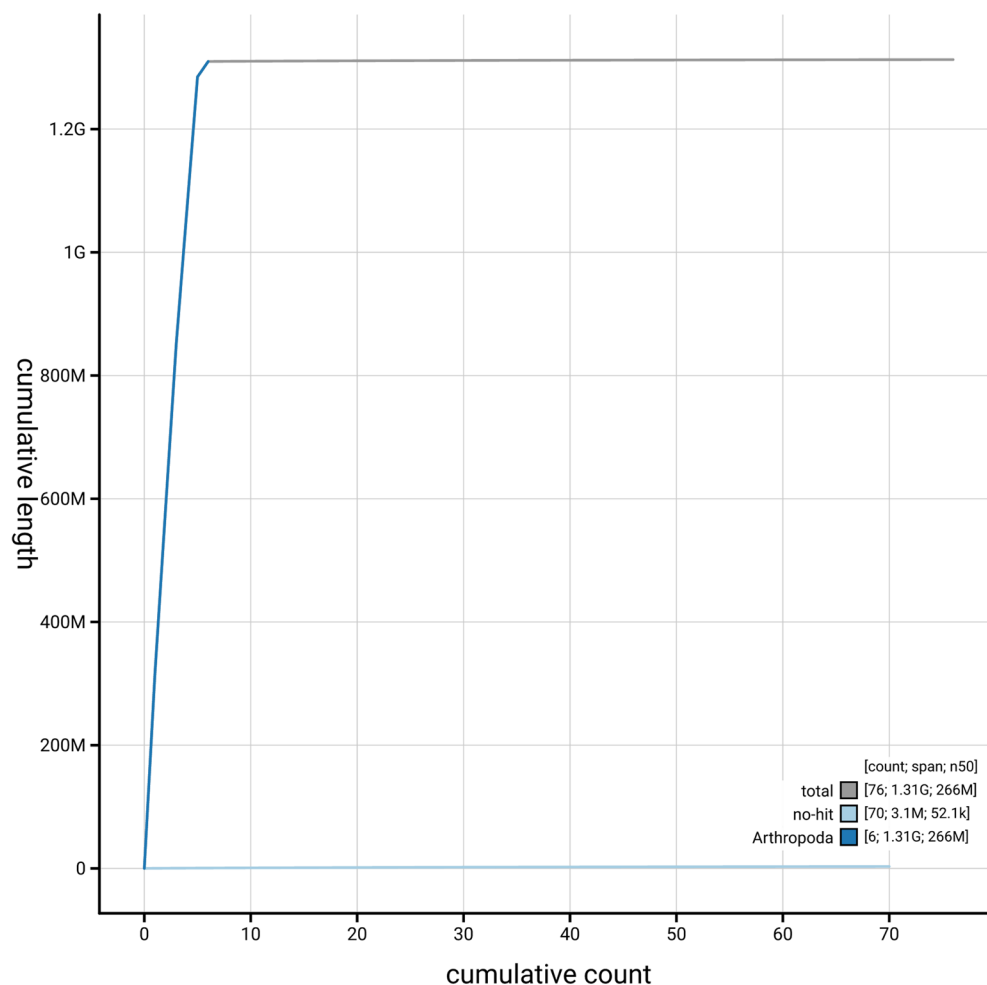


Figure 4. Genome assembly of *Dichetophora obliterated*, idDicObli2.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963920525.1/dataset/GCA_963920525.1/cumulative.

paired-end sequencing with a read length of 150 bp on an Illumina NovaSeq 6000 instrument.

RNA

Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit, following the manufacturer's instructions. RNA sequencing was performed on the Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

Assembly

Prior to assembly of the PacBio HiFi reads, a database of k -mer counts ($k = 31$) was generated from the filtered reads using FastK. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to analyse the k -mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were 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 for handling potential misassemblies. 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.

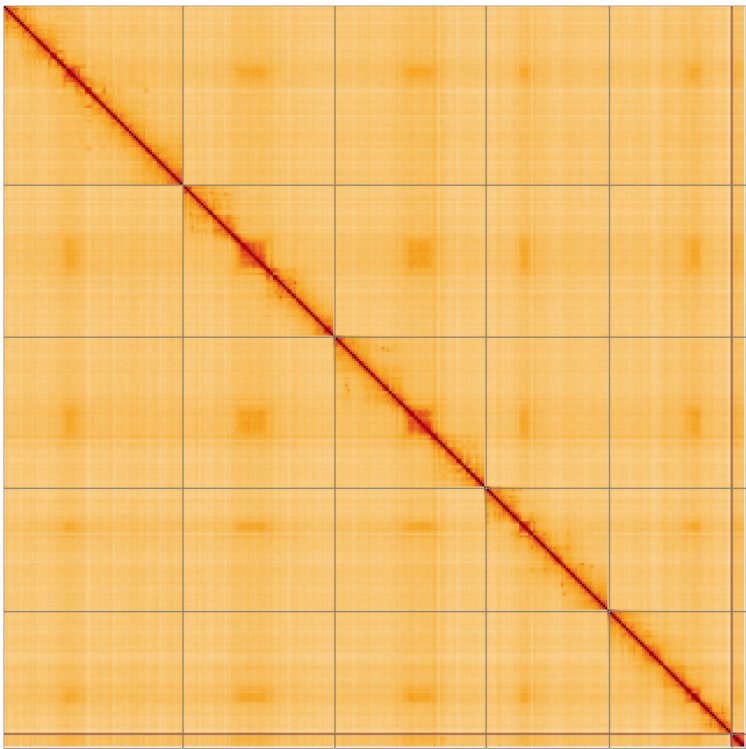


Figure 5. Genome assembly of *Dichetophora oblitterata*: Hi-C contact map of the idDicObli2.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at <https://genome-note-higlass.tol.sanger.ac.uk/l/?d=GDtD1pwnTueuLyveTVqfVw>.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Dichetophora oblitterata*, idDicObli2.

INSDC accession	Name	Length (Mb)	GC%
OY986435.1	1	316.54	32.5
OY986436.1	2	268.16	32.5
OY986437.1	3	266.48	32.5
OY986438.1	4	217.4	32
OY986439.1	5	216.24	32.5
OY986440.1	6	24.9	32
OY986441.1	MT	0.02	29

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. Flat files and maps used in curation were generated via the TreeVal pipeline (Pointon *et al.*, 2023). Manual curation was conducted primarily in PretextView (Harry, 2022) and HiGlass (Kerpedjiev *et al.*, 2018), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were amended,

and duplicate sequences were tagged and removed. The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation>.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020), run in a Singularity container (Kurtzer *et al.*, 2017), was used to evaluate *k*-mer completeness and assembly quality for the primary and alternate haplotypes using the *k*-mer databases (*k* = 31) that were computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

A Hi-C contact map was produced for the final version of the assembly. The Hi-C reads were aligned using bwa-mem2 (Vasimuddin *et al.*, 2019) and the alignment files were combined using SAMtools (Danecek *et al.*, 2021). The Hi-C alignments were converted into a contact map using BEDTools (Quinlan & Hall, 2010) and the Cooler tool suite (Abdennur & Mirny, 2020). The contact map was visualised in HiGlass (Kerpedjiev *et al.*, 2018).

The blobtoolkit pipeline is a Nextflow (Di Tommaso *et al.*, 2017) port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads in SAMtools and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoT database (Challis *et al.*, 2023) to identify all matching BUSCO lineages to run BUSCO (Manni *et al.*, 2021). For the three domain level

BUSCO lineages, the pipeline aligns the BUSCO genes to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) with DIAMOND blastp (Buchfink *et al.*, 2021). The genome is also divided into chunks according to the density of the BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database using DIAMOND blastx. Genome sequences without a hit are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The blobtools suite combines all these outputs into a blobdir for visualisation.

The blobtoolkit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), relying on the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), as well as the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

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

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	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	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	666652151335353eef2fcd58880bcef5bc2928e1	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	0.19.5-r587	https://github.com/chhy123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84aa44357826c0b6753eb28de	https://github.com/higlass/higlass
MercuryFK	d00d98157618f4e8d1a9190026b19b471055b22e	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14, 1.17, and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.04.1	https://github.com/nextflow-io/nextflow
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	-	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.5.1	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

Genome annotation

The [Ensembl Genebuild](#) annotation system (Aken *et al.*, 2016) was used to generate annotation for the *Dichetophora obliterata* assembly (GCA_963920525.1) 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).

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the ‘**Darwin Tree of Life Project Sampling Code of Practice**’, which can be found in full on the Darwin Tree of Life website [here](#). By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they 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: *Dichetophora obliterata* (chalk snailkiller). Accession number PRJEB70855; <https://identifiers.org/ena.embl/PRJEB70855>. The genome sequence is released openly for reuse. The *Dichetophora obliterata* 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 Natural History Museum Genome Acquisition Lab are listed here: <https://doi.org/10.5281/zenodo.12159242>.

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

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

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

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

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

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

References

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

Aken BL, Ayling S, Barrell D, *et al.*: **The Ensembl gene annotation system**. *Database (Oxford)*. 2016; **2016**: baw093. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool**. *J Mol*

Biol. 1990; **215**(3): 403–410.

[PubMed Abstract](#) | [Publisher Full Text](#)

Ball S: **A Key to the British Sciomyzidae**. *Dipterists Forum*. 2017. [Reference Source](#)

Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the Universal Protein Knowledgebase in 2023**. *Nucleic Acids Res*. 2023; **51**(D1): D523–D531. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor³ for LI PacBio**. *protocols.io*. 2023. [Publisher Full Text](#)

Beasley J, Uhl R, Forrest LL, *et al.*: **DNA barcoding SOPs for the Darwin Tree of**

Life project. *protocols.io*. 2023; (Accessed 25 June 2024).

[Publisher Full Text](#)

Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND**. *Nat Methods*. 2021; **18**(4): 366–368.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Challis R, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved]**. *Wellcome Open Res*. 2023; **8**: 24.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools**. *GigaScience*. 2021; **10**(2): giab008.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[Publisher Full Text](#)

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

[Publisher Full Text](#)

Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows**. *Nat Biotechnol*. 2017; **35**(4): 316–319.

[PubMed Abstract](#) | [Publisher Full Text](#)

Diesch 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](#)

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](#)

Falk S: **Dichotophora obliterata (Chalk Snailkiller)**. *Flickr*. 2025; [Accessed 7 February 2025].

[Reference Source](#)

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: **GBIF Backbone Taxonomy: *Dichotophora obliterata***. 2023.

[Reference Source](#)

Gedling Conservation Trust: **Chalk snailkiller**. 2023.

[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](#)

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies**. *Bioinformatics*. 2020; **36**(9): 2896–2898.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Harry E: **PretextView (Paired Read TEXTure Viewer): a desktop application for viewing pretext contact maps**. 2022.

[Reference Source](#)

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation**. *GigaScience*. 2021; **10**(1): giaa153.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[Publisher Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute**. *PLoS One*. 2017; **12**(5): e0177459.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Lawniczak MKN, Davey RP, Rajan J, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project [version 1; peer review: 2 approved with reservations]**. *Wellcome Open Res*. 2022; **7**: 187.

[Publisher Full Text](#)

Li H: **Minimap2: pairwise alignment for nucleotide sequences**. *Bioinformatics*. 2018; **34**(18): 3094–3100.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[Reference Source](#)

NBN Atlas Partnership: **NBN Atlas: *Dichotophora obliterata***. 2023.

[Reference Source](#)

Oatley G, Denton A, Howard C: **Sanger Tree of Life HMW DNA extraction: automated MagAttract v.2**. *protocols.io*. 2023.

[Publisher Full Text](#)

Pereira L, Sivell O, Sivess L, *et al.*: **DTOL: taxon-specific standard operating procedure for the terrestrial and freshwater arthropods working group**. 2022.

[Publisher Full Text](#)

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

[Publisher Full Text](#)

Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics*. 2010; **26**(6): 841–842.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes**. *Nat Commun*. 2020; **11**(1): 1432.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[Publisher Full Text](#)

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer review: 2 approved]**. *Wellcome Open Res*. 2024; **9**: 339.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

UniProt Consortium: **UniProt: a worldwide hub of protein knowledge**. *Nucleic Acids Res*. 2019; **47**(D1): D506–D515.

[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](#)

Vicosa B, Bachtrog D: **Numerous transitions of sex chromosomes in Diptera**. *PLoS Biol*. 2015; **13**(4): e1002078.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Zhou C, McCarthy SA, Durbin R: **YaHS: yet another Hi-C scaffolding tool**. *Bioinformatics*. 2023; **39**(1): btac808.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 26 May 2025

<https://doi.org/10.21956/wellcomeopenres.26470.r122489>

© 2025 Nakamura T. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Taro Nakamura 

National Institute for Basic Biology, Okazaki, Japan

This manuscript reports the genome sequencing of *Dichetophora obliterata*, a parasitic snail-killing fly. The assembled genome size is approximately 1.3 Gb, and high-quality chromosome-level assembly and annotation have been achieved. The genome analysis is clearly presented, demonstrating high data quality and notable scientific contribution.

The methodologies used in this study are clearly and comprehensively described, reaching a standard that allows for reproducibility by third parties.

The application of PacBio HiFi long-read sequencing combined with Hi-C technology has provided a high-quality, chromosome-level genome assembly. Quality metrics such as the QV scores and BUSCO analysis further support the high quality of the generated data.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Developmental Biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 08 May 2025

<https://doi.org/10.21956/wellcomeopenres.26470.r122172>

© 2025 Gemmellaro D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Denise Gemmellaro

Department of Biological Sciences, Kean University, Union, New Jersey, USA

The authors assembled the complete genome of a female specimen of *Dichetophora obliterata*. This is a dipteran belonging to the family Sciomyzidae. The genome sequence is 1,312.79 megabase long and 15,139 protein coding genes were identified. This is a very straight forward paper presenting a clear, complete and reproducible protocol for genome analysis, from DNA extraction to bioinformatics.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Entomology, Forensic Entomology, Diptera, Decomposition Ecology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
