DATA NOTE

# The genome sequence of the hybotid fly, *Hybos culiciformis* (Fabricius, 1775) (Diptera: Hybotidae)

[version 1; peer review: 2 approved, 1 approved with reservations]

Steven Falk[1], Liam M. Crowley[2],
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]Independent researcher, Kenilworth, Warwickshire, England, UK
[2]University of Oxford, Oxford, England, UK

**v1** **First published:** 05 Nov 2025, **10**:617
https://doi.org/10.12688/wellcomeopenres.25056.1
**Latest published:** 05 Nov 2025, **10**:617
https://doi.org/10.12688/wellcomeopenres.25056.1

## Abstract

We present a genome assembly from a male specimen of *Hybos culiciformis* (hybotid fly; Arthropoda; Insecta; Diptera; Hybotidae). The genome sequence has a total length of 342.56 megabases. Most of the assembly (90.73%) is scaffolded into 5 chromosomal pseudomolecules, including the X and Y sex chromosomes. The mitochondrial genome has also been assembled, with a length of 17.59 kilobases. Gene annotation of this assembly on Ensembl identified 20 872 protein-coding genes. This assembly was generated as part of the Darwin Tree of Life project, which produces reference genomes for eukaryotic species found in Britain and Ireland.

## Keywords

Hybos culiciformis; hybotid fly; genome sequence; chromosomal; Diptera

This article is included in the Tree of Life gateway.

**Open Peer Review**

**Approval Status** ✓ ? ✓

| | 1 | 2 | 3 |
|---|---|---|---|
| **version 1** 05 Nov 2025 | ✓ view | ? view | ✓ view |

1. **Steven Fiddaman**, University of Oxford, Oxford, UK
2. **Jean-Baka Domelevo Entfellner**, International Livestock Research Institute (Ringgold ID: 54661), Nairobi, Kenya
3. **Saverio Brogna**, University of Birmingham, Birmingham, UK

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: Falk S**: Investigation, Resources; **Crowley LM**: 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, <a href=https://doi.org/10.35802/218328>https://doi.org/10.35802/218328 </a>].
*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:** Falk S, Crowley LM, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the hybotid fly, *Hybos culiciformis* (Fabricius, 1775) (Diptera: Hybotidae) [version 1; peer review: 2 approved, 1 approved with reservations]** Wellcome Open Research 2025, **10**:617 https://doi.org/10.12688/wellcomeopenres.25056.1

**First published:** 05 Nov 2025, **10**:617 https://doi.org/10.12688/wellcomeopenres.25056.1

## Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Eremoneura; Empidoidea; Hybotidae; Hybotinae; *Hybos*; *Hybos culiciformis* (Fabricius, 1775) (NCBI:txid1262239)

## Background

*Hybos culiciformis* is a small dance fly, measuring around 3.5–5.5 mm in length, with a dark body, black legs, swollen and bristly hind femora, red eyes that meet, and a humped thorax (Gedling Conservation Trust, 2025). It is one of three British *Hybos* species, differing from *H. femoratus*, which has red legs except for the hind pair (NatureSpot, 2025). The species occurs on woodland edges, hedgerows, and sometimes reed-beds. Adults prey on small flying insects and larvae feed on soil invertebrates (Gedling Conservation Trust, 2025). It is widespread across Europe and the Near East, and in Britain it is fairly common and widespread, often seen from June to September (NBN Atlas Partnership, 2025).

We present a chromosome-level genome sequence for *Hybos culiciformis*. This is the first genome for the genus *Hybos* and the family Hybotidae as of September 2025 (data obtained via NCBI datasets, O'Leary *et al.*, 2024). The assembly was generated as part of the Darwin Tree of Life Project, using the Tree of Life pipeline from a specimen collected in Wytham Woods, Oxfordshire, United Kingdom (Figure 1).

## Methods

### Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Hybos culiciformis* (specimen ID Ox000747, ToLID idHybCuli1; Figure 1), collected from Wytham Woods, Oxfordshire, UK (latitude 51.766, longitude –1.309) on 2020-08-03. The specimen was collected and identified by Steven Falk. A second specimen was used for Hi-C sequencing (specimen ID Ox003225, ToLID idHybCuli3). It was collected from Wytham Woods, Oxfordshire, UK (latitude 51.764, longitude –1.337) on 2022-10-03. The specimen was collected by Steven Falk and Liam Crowley and identified by Steven Falk. For the Darwin Tree of Life sampling and metadata approach, refer to Lawniczak *et al.* (2022).


**Figure 1. Photograph of the *Hybos culiciformis* (idHybCuli1) specimen used for genome sequencing.**

The initial identification was verified by an additional DNA barcoding process according to the framework developed by Twyford *et al.* (2024). A small sample was dissected from the specimen and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI) (see the protocol). 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 are available on protocols.io.

### Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The idHybCuli1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the whole organism was homogenised by powermashing using a PowerMasher II tissue disruptor. HMW DNA was extracted using the Automated MagAttract v2 protocol. We used centrifuge-mediated fragmentation to produce DNA fragments in the 8–10 kb range, following the Covaris g-TUBE protocol for ultra-low input (ULI). Sheared DNA was purified by automated SPRI (solid-phase reversible immobilisation). 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. For this sample, the final post-shearing DNA had a Qubit concentration of 1.28 ng/μL and a yield of 499.20 ng.

### PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Prior to library preparation, the DNA was fragmented to ~10 kb. Ultra-low-input (ULI) libraries were prepared using the PacBio SMRTbell® Express Template Prep Kit 2.0 and gDNA Sample Amplification Kit. Samples were normalised to 20 ng DNA. Single-strand overhang removal, DNA damage repair, and end-repair/A-tailing were performed according to the manufacturer's instructions, followed by adapter ligation. A 0.85× pre-PCR clean-up was carried out with Promega ProNex beads.

The DNA was evenly divided into two aliquots for dual PCR (reactions A and B), both following the manufacturer's protocol. A 0.85× post-PCR clean-up was performed with ProNex beads. DNA concentration was measured using a Qubit Fluorometer v4.0 (Thermo Fisher Scientific) with the Qubit HS Assay Kit, and fragment size was assessed on an Agilent

Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit. PCR reactions A and B were then pooled, ensuring a total mass of ≥500 ng in 47.4 μl.

The pooled sample underwent another round of DNA damage repair, end-repair/A-tailing, and hairpin adapter ligation. A 1× clean-up was performed with ProNex beads, followed by DNA quantification using the Qubit and fragment size analysis using the Agilent Femto Pulse. Size selection was performed on the Sage Sciences PippinHT system, with target fragment size determined by Femto Pulse analysis (typically 4–9 kb). Size-selected libraries were cleaned with 1.0× ProNex beads and normalised to 2 nM before sequencing.

The sample was 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, and to perform primary and secondary analysis of the data upon completion.

## Hi-C

### *Sample preparation and crosslinking*

The Hi-C sample was prepared from 20–50 mg of frozen tissue of the idHybCuli3 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

### *Hi-C library preparation and sequencing*

Biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size selected to 400–600 bp using SPRISelect beads. DNA was enriched with Arima-HiC v2 kit Enrichment beads. End repair, A-tailing, and adapter ligation were carried out with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs), following a modified protocol where library preparation occurs while DNA remains bound to the Enrichment beads. Library amplification was performed using KAPA HiFi HotStart mix and a custom Unique Dual Index (UDI) barcode set (Integrated DNA Technologies). Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, libraries were amplified with 10–16 PCR cycles. Post-PCR clean-up was performed with SPRISelect beads. Libraries were quantified using the AccuClear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech).

Prior to sequencing, libraries were normalised to 10 ng/μL. Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools were created. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq 6000.

## Genome 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 (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019), and the contigs were scaffolded in YaHS (Zhou *et al.*, 2023) with 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.

## Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included 109 breaks and 224 joins. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

## Assembly quality assessment

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

The genome was analysed using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified from the NCBI Taxonomy (Schoch *et al.*, 2020). For the three domain-level lineages, BUSCO

genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all 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), with containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

## Genome sequence report

### Sequence data

PacBio sequencing of the *Hybos culiciformis* specimen generated 25.16 Gb (gigabases) from 2.80 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 420.56 Mb, with a heterozygosity of 0.59% and repeat content of 59.85% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 54× coverage. Hi-C sequencing produced 112.01 Gb from 741.76 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

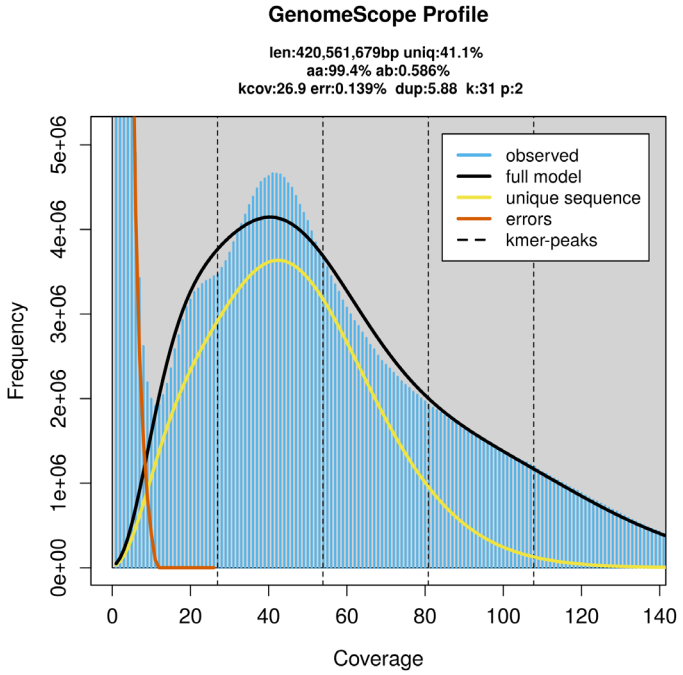


**Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2.** The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

**Table 1. Specimen and sequencing data for BioProject PRJEB67601.**

| Platform | PacBio HiFi | Hi-C |
|---|---|---|
| **ToLID** | idHybCuli1 | idHybCuli3 |
| **Specimen ID** | Ox000747 | Ox003225 |
| **BioSample (source individual)** | SAMEA7746460 | SAMEA113425793 |
| **BioSample (tissue)** | SAMEA7746532 | SAMEA113425977 |
| **Tissue** | whole organism | whole organism |
| **Instrument** | Sequel IIe | Illumina NovaSeq 6000 |
| **Run accessions** | ERR12205260 | ERR12143995 |
| **Read count total** | 2.80 million | 741.76 million |
| **Base count total** | 25.16 Gb | 112.01 Gb |

### Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The final assembly has a total length of 342.56 Mb in 568 scaffolds, with 1 155 gaps, and a scaffold N50 of 106.42 Mb (Table 2).

Most of the assembly sequence (90.73%) was assigned to 5 chromosomal-level scaffolds, representing 3 autosomes and the X and Y sex chromosomes. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Chromsomes X and Y were assigned based on read and HiC coverage data.

The mitochondrial genome was also assembled (length 17.59 kb, OZ023278.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

The combined primary and alternate assemblies achieve an estimated QV of 57.2. The *k*-mer completeness is 85.48% for the primary assembly, 82.48% for the alternate haplotype, and 98.21% for the combined assemblies (Figure 4).

BUSCO v.5.5.0 analysis using the diptera_odb10 reference set ($n$ = 3 285) identified 94.1% of the expected gene set (single = 92.2%, duplicated = 1.9%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for the primary assembly. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the primary assembly, is **5.C.Q56**.

## Genome annotation report

The *Hybos culiciformis* genome assembly (GCA_964007475.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 21 343 transcribed mRNAs from 20 872 protein-coding genes. The average transcript length is 4 253.07 bp, with an average of 4.17 exons per transcript. For further information about the annotation, please refer to the annotation page on Ensembl.

### 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. 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

**Table 2. Genome assembly statistics.**

| | |
|---|---|
| **Assembly name** | idHybCuli1.1 |
| **Assembly accession** | GCA_964007475.1 |
| **Alternate haplotype accession** | GCA_964007365.1 |
| **Assembly level** | chromosome |
| **Span (Mb)** | 342.56 |
| **Number of chromosomes** | 5 |
| **Number of contigs** | 1 723 |
| **Contig N50** | 0.43 Mb |
| **Number of scaffolds** | 568 |
| **Scaffold N50** | 106.42 Mb |
| **Sex chromosomes** | X and Y |
| **Organelles** | Mitochondrion: 17.59 kb |

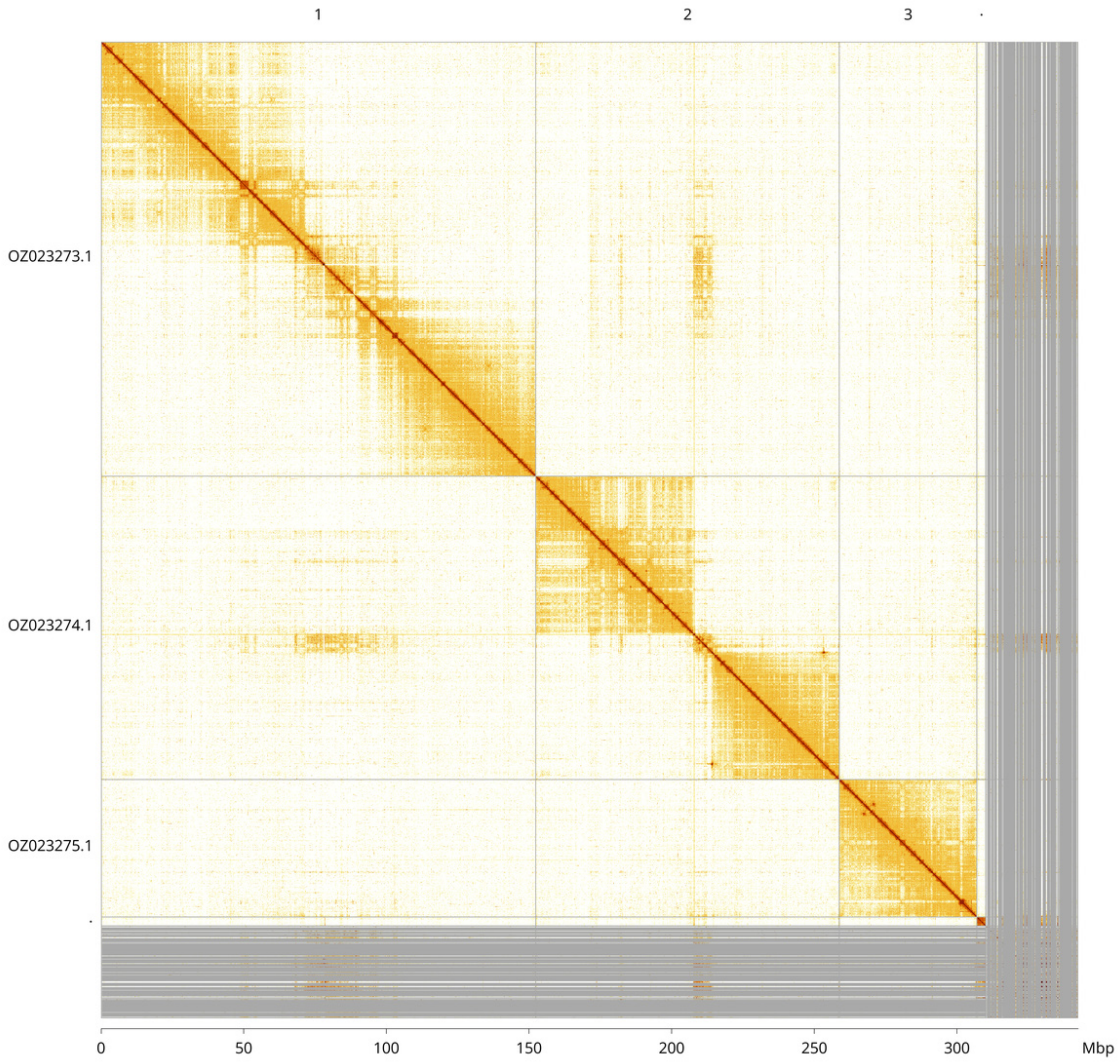


**Figure 3. Hi-C contact map of the *Hybos culiciformis* genome assembly.** Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. Because the X and Y chromosomes are markedly smaller, they are not labelled. The plot was generated using PretextSnapshot.

**Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Hybos culiciformis* idHybCuli1.**

| INSDC accession | Molecule | Length (Mb) | GC% |
|---|---|---|---|
| OZ023273.1 | 1 | 152.39 | 25 |
| OZ023274.1 | 2 | 106.42 | 24.50 |
| OZ023275.1 | 3 | 48.25 | 24 |
| OZ023276.1 | X | 3.21 | 32 |
| OZ023277.1 | Y | 0.53 | 28.50 |

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

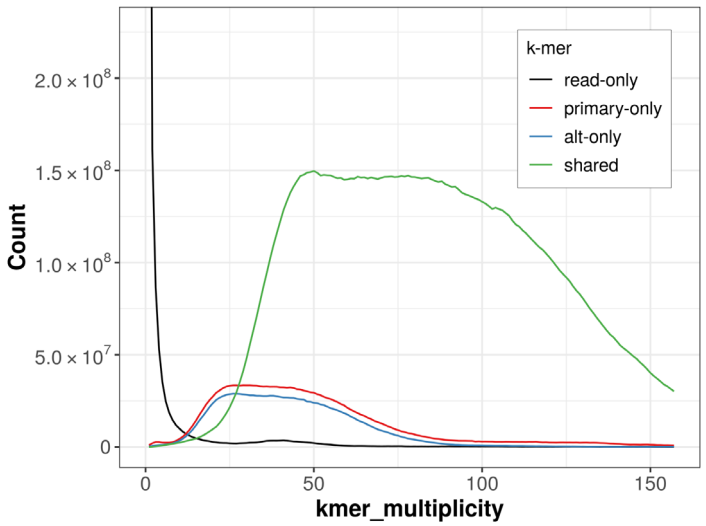


**Figure 4. Evaluation of *k*-mer completeness using MerquryFK.** This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve corresponds to *k*-mers shared by both haplotypes, and the red and blue curves show *k*-mers found only in one of the haplotypes.

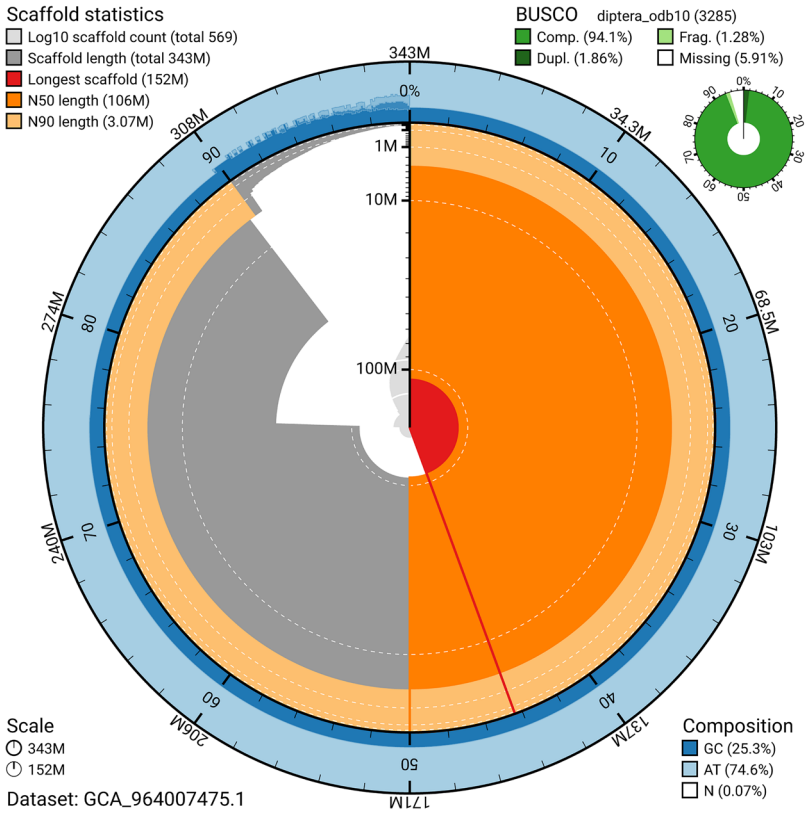


**Figure 5. Assembly metrics for idHybCuli1.1.** The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost 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 can be accessed on the BlobToolKit viewer.

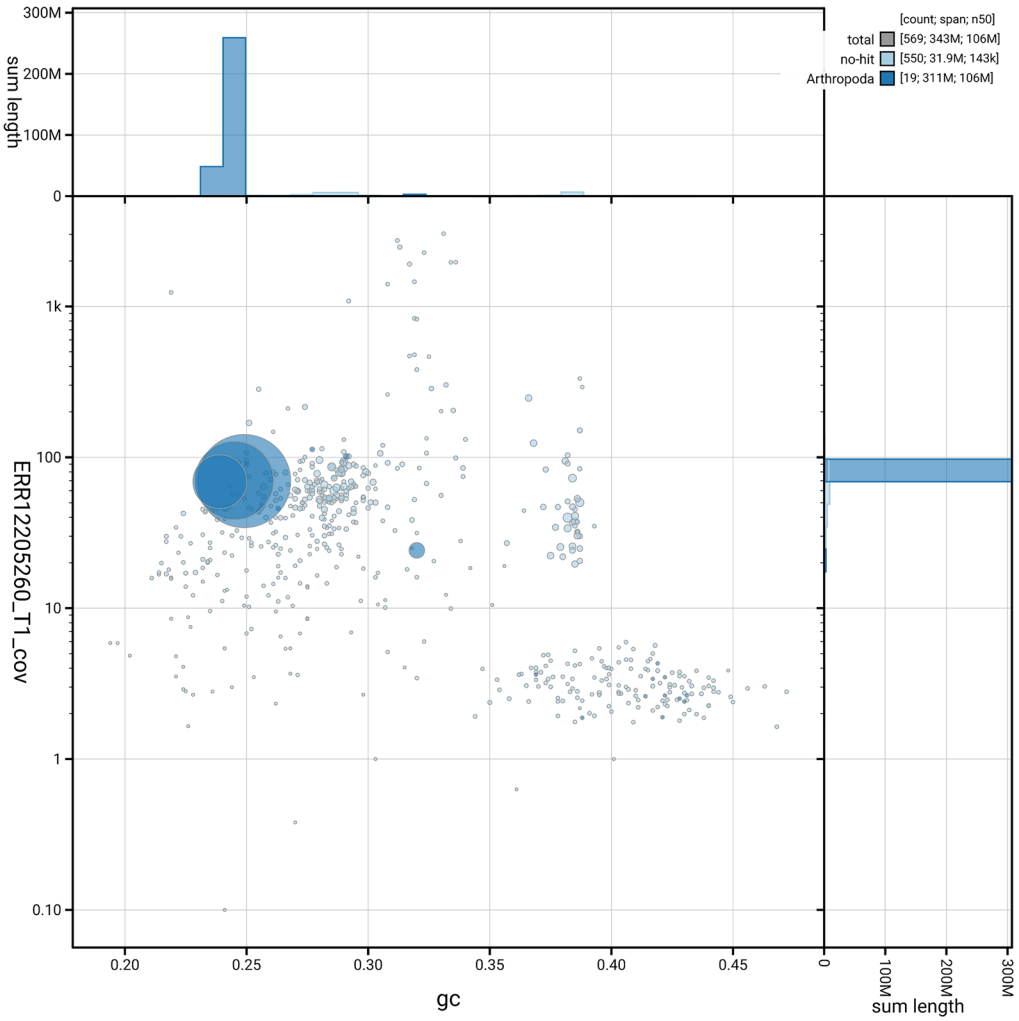


**Figure 6. BlobToolKit GC-coverage plot for idHybCuli1.1.** Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the BlobToolKit viewer.

**Table 4. Earth Biogenome Project summary metrics for the *Hybos culiciformis* assembly.**

| Measure | Value | Benchmark |
|---|---|---|
| EBP summary (primary) | 5.C.Q56 | 6.C.Q40 |
| Contig N50 length | 0.43 Mb | ≥ 1 Mb |
| Scaffold N50 length | 106.42 Mb | = chromosome N50 |
| Consensus quality (QV) | Primary: 56.5; alternate: 58.0; combined: 57.2 | ≥ 40 |
| *k*-mer completeness | Primary: 85.48%; alternate: 82.48%; combined: 98.21% | ≥ 95% |
| BUSCO | C:94.1% [S:92.2%; D:1.9%]; F:1.3%; M:4.6%; n:3 285 | S > 90%; D < 5% |
| Percentage of assembly assigned to chromosomes | 90.73% | ≥ 90% |

Institute), and in some circumstances, other Darwin Tree of Life collaborators.

## Data availability

European Nucleotide Archive: Hybos culiciformis. Accession number PRJEB67601. The genome sequence is released openly for reuse. The *Hybos culiciformis* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665) and the Sanger Institute Tree of Life Programme (PRJEB43745). 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.

Production code used in genome assembly at the WSI Tree of Life is available at https://github.com/sanger-tol. Table 5 lists software versions used in this study.

## Author information

Contributors are listed at the following links:

- Members of the University of Oxford and Wytham Woods Genome Acquisition Lab
- Members of the Darwin Tree of Life Barcoding collective
- Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team
- Members of Wellcome Sanger Institute Scientific Operations – Sequencing Operations
- Members of the Wellcome Sanger Institute Tree of Life Core Informatics team
- Members of the Tree of Life Core Informatics collective
- Members of the Darwin Tree of Life Consortium

**Table 5. Software versions and sources.**

| Software | 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 | 1.1 | https://github.com/thegenemyers/FASTK |
| GenomeScope2.0 | 2.0.1 | https://github.com/tbenavi1/genomescope2.0 |
| 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/chhylp123/hifiasm |
| HiGlass | 1.13.4 | https://github.com/higlass/higlass |
| MerquryFK | 1.1.2 | 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 |
| PretextSnapshot | - | https://github.com/sanger-tol/PretextSnapshot |
| PretextView | 0.2.5 | https://github.com/sanger-tol/PretextView |
| purge_dups | 1.2.5 | https://github.com/dfguan/purge_dups |

| Software | Version | Source |
|---|---|---|
| samtools | 1.19.2 | https://github.com/samtools/samtools |
| sanger-tol/ascc | 0.1.0 | https://github.com/sanger-tol/ascc |
| sanger-tol/blobtoolkit | 0.4.0 | https://github.com/sanger-tol/blobtoolkit |
| sanger-tol/curationpretext | 1.4.2 | https://github.com/sanger-tol/curationpretext |
| Seqtk | 1.3 | https://github.com/lh3/seqtk |
| Singularity | 3.9.0 | https://github.com/sylabs/singularity |
| TreeVal | 1.4.0 | https://github.com/sanger-tol/treeval |
| YaHS | 1.2a.2 | https://github.com/c-zhou/yahs |

## References


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

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

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, 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

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

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

Gedling Conservation Trust: **Hybos culiciformis.** 2025.
Reference Source

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

Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025.
Publisher Full Text

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

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, 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

NatureSpot: **Hybos culiciformis.** 2025.
Reference Source

NBN Atlas Partnership: ***Hybos culiciformis* on the NBN Atlas.** 2025.
Reference Source

O'Leary NA, Cox E, Holmes JB, *et al.*: **Exploring and retrieving sequence and metadata for species across the Tree of Life with NCBI datasets.** *Sci Data.* 2024; **11**(1): 732.
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

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI taxonomy: a comprehensive update on curation, resources and tools.** *Database (Oxford).* 2020; **2020**: baaa062.
PubMed Abstract | Publisher Full Text | Free 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**

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**

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 27 December 2025

https://doi.org/10.21956/wellcomeopenres.27622.r138821



✔ **Saverio Brogna**
University of Birmingham, Birmingham, England, UK

This work presents the genome assembly of a species of Hybos (Diptera). The genome sequence, together with the Hi-C data, is of high quality and provides valuable information. Overall, this represents a valuable contribution to the Darwin Tree of Life project.

I have only two minor comments for the authors. It would have been useful to perform RNA sequencing from the same specimen (particularly in light of the very low AT content of the genome), and it would have been informative to specify which tissue(s) were used.

**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:*** My research focuses on gene expression in eukaryotes, using mostly Drosophila as a model organism. I also have an interest in molecular evolution of genes.

**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 26 December 2025

https://doi.org/10.21956/wellcomeopenres.27622.r139739



?

**Jean-Baka Domelevo Entfellner**
Data and Research Methods Unit, International Livestock Research Institute (Ringgold ID: 54661), Nairobi, Nairobi County, Kenya

This work presents the first chromosome-level assembly for a dipteran insect, *Hybos culiciformis*. The assembly was performed using adequate sequencing and genome assembly technologies: the authors used PacBio HiFi reads as well as a Hi-C technique for chromosome conformation mapping. The results are correctly presented and overall, there is a good amount of detail on the samples used as well as the methodology followed.

Nevertheless, I want to point a few flaws that need fixing.

(1) On the sex of the specimens, there is some amount of uncertainty. The text says that idHybCuli1 (which was used for PacBio sequencing) is “an adult male” (section “Sample acquisition and DNA barcoding”), while the biosample record says that the sex was “NOT COLLECTED” (see e.g. https://www.ebi.ac.uk/ena/browser/view/SAMEA7746460 or https://tolqc.cog.sanger.ac.uk/darwin/insects/Hybos_culiciformis). More importantly, the second specimen, idHybCuli3 (which was used for the Hi-C work) appears to be a **female**. At least, its biosample is registered in ENA as such ( https://www.ebi.ac.uk/biosamples/samples/SAMEA113425977). This discrepancy and the uncertainty about the sex of the specimens used makes the reader feel a bit uncomfortable, especially since the allegedly assembled X and Y chromosomes are extremely short (respectively 3.21 Mb and 0.53 Mb). I am no expert of sex chromosomes in insects, but such values are surprising to me, especially since they are not contrasted with similar or diverging values obtained on closely-related species.

Of course, it would have been best to run both the PacBio and the Hi-C work on tissue from the same individual, but we understand that operational constraints may have led to the necessity to use two different individuals. Yet, since proximity maps are essential here to confirm the scaffolding results from PacBio reads assembled with YaHS, one would hope that at least, same-sex individuals were used in this study.

In Figure 3, the authors present the Hi-C contact map, but unfortunately, there is no zoom-in on the scaffolds corresponding to the alleged sex chromosomes. In the absence of any form of comparison to sex chromosomes from other Dipterans, I believe the characterisation of the X and Y scaffolds here is insufficient. The authors only state: “Chromosomes X and Y were assigned based on read and HiC coverage data.” If any form of comparative genomics with other taxa was done, it should be mentioned.

Also to mention regarding the alleged X and Y chromosomes: while autosomal chromosomes are found to have similar GC-content percentage ("25", "24.50" and "24" reported in Table 3 -> by the way, I advise the authors to harmonise the presentation of the results in Table 3, with the same precision for all figures), we have markedly different values for X (32%) and Y (28.5%). Is this common in insects, to have markedly different CG content for sex vs. autosomal chromosomes? The authors did remove bacterial contigs using the Assembly Screen for Cobionts and Contaminants (ASCC) after assembly, but some bacteria, fungi or viruses may have failed to be detected by the pipeline. The 0.53 Mb scaffold could very well be made of such foreign bacterial content. Upon zooming in onto the snail plot in Figure 5, in the area corresponding to the scaffolds beyond the first three, the way lines in the blue crown (the one for GC content) are a further indication of potential foreign material.

(2) On the number of chromosomes found, I would expect a bit of a discussion (even though I understand this is only a "data note" paper). Dipterans most commonly have n=6. Unfortunately, the authors of (Ref no. 2) didn't study Hybotidae, so one cannot know for sure, but overall, I would welcome at least a small paragraph contrasting this genome data with one or two closest sequenced species.

(3) Let me now mention some other points, mistakes and things that call for clarification.

3.1 In the second paragraph of section "Sample acquisition and DNA barcoding", one can read: "A small sample was dissected from the specimen and stored in ethanol". We wonder why, and which part of the insect that was.

3.2 In the first paragraph of section "PacBio HiFi library preparation and sequencing", one can read: "Prior to library preparation, the DNA was fragmented to ~10 kb." The reader is puzzled, since it seems it was already the case arising from the earlier use of the Covaris G-tube ("We used centrifuge-mediated fragmentation to produce DNA fragments in the 8-10 kb range", in the paragraph immediately above). Please clarify whether the two chunks of text refer to the same operation, of if two successive rounds of fragmentation were carried out, in which case, the reader is allowed to ask why such was done.

3.3 In "The DNA was evenly divided into two aliquots for dual PCR (reactions A and B), both following the manufacturer's protocol", it is unclear which manufacturer we are talking about. One guesses "PacBio", but perhaps it's better to name them here, since it is in a new paragraph.

3.4 There seems to be a typo in the text, on "with a heterozygosity of 0.59% and repeat content of 59.85%". The GenomeScope profile in Figure 2 gives a percentage of unique sequences at 41.1%. Since 100 - 41.1 = 58.9, I guess the authors meant to write in the text "58.85%" and not "59.85%", pertaining to the repeat content.

3.5 In Table 2 and in the text ("These chromosome-level scaffolds, confirmed by Hi-C data, are named [...]"), whether the output of the assembly based on PacBio reads only yielded the 1,723 contigs in 568 scaffolds. This is my understanding, but the figures in Table 2 could also include some further scaffolding based on the contact maps obtained from the Hi-C data. Please clarify around the word "confirmed" that could also be "obtained", in the above-mentioned chunk of text.

3.6 The number of chromosomes set to 5 (and, especially, the issue around the sex chromosomes)

is a bit “shaky” if the only consideration leading to this number was “we clearly got three autosomal chromosomes over 10 Mb each, and then we assign X and Y to the following two scaffolds ranked in decreasing order of length”. The sex chromosome system may be more complex (see e.g. Ref no.1)

3.7 I suggest the authors harmonise the many occurrences of “Hi-C” and “HiC” throughout the text, into a single, constant spelling. I prefer “Hi-C”.

3.8 I suggest the authors include in Table 3 the data pertaining to the assembled mitochondrion.

3.9 The curves in Figure 4 seem to signal a higher percentage of heterozygosity than the one reported in Figure 2. The difference is that in Figure 4, the k-mer appearing in reads that were not assembled in the 5 scaffolds (?) are rejected into the black curve on the left, but a quick back-of-the envelope calculation I performed based on the green, red and blue curves from Figure 4 yields much higher a percentage of heterozygosity than the 0.586% reported by GenomeScope 2 in Figure 2. Happy to discuss this further with the authors, as I think this deserves a bit of further examination, and readers would be glad to have a final, ascertained figure for the heterozygosity in this genome assembly.

3.10 In the first paragraph under “Hi-C library preparation and sequencing”, it is not clear (at least to readers deprived of expertise on the specific protocol used) why you build several libraries, to then pool them in equimolar concentrations. A word of explanation on this would be welcome.

3.11 Just before the “Genome assembly” section, when authors mention the Illumina NovaSeq 6000, please add a mention of the specific Illumina reagent kit used (e.g. an SP, S1, S2 or S4 flow cell). Related to this, readers can be surprised to see that authors sequenced their Hi-C library to such a high depth (112.01 Gb, i.e. some 266x of the tentative genome size given by GenoScope 2, and some 327x of the final reported genome size).

3.12 In the first paragraph under “Assembly quality assessment”, “databases” to be modified to “ **database**” (there is only one database of k-mers associated with a fixed k=31).

3.13 In the first paragraph under “Genome annotation report”, the authors should qualify the number of transcripts they got from the automated Ensemble annotation. Since no RNA-Seq experiment was conducted here, I suggest writing “21 343 **predicted** mRNAs” rather than “21 343 transcribed mRNAs”.

**References**

1. Vicoso B, Bachtrog D: Numerous Transitions of Sex Chromosomes in Diptera. *PLOS Biology*. 2015; **13** (4). Publisher Full Text
2. Morelli M, Blackmon H, Hjelmen C: Diptera and Drosophila Karyotype Databases: A Useful Dataset to Guide Evolutionary and Genomic Studies. *Frontiers in Ecology and Evolution*. 2022; **10**. Publisher Full Text

**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?**
Partly

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

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** bioinformatics, long-read genomics, phylogenetics

**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, however I have significant reservations, as outlined above.**

Reviewer Report 11 December 2025

https://doi.org/10.21956/wellcomeopenres.27622.r139738



**Steven Fiddaman**
University of Oxford, Oxford, England, UK

This Data Note presents a chromosome-level genome assembly for the hybotid fly, *Hybos culiciformis*, as part of the Darwin Tree of Life project. Overall, the data provided is comprehensive, and the assembly quality is high.

The methodology for this approach to genome sequencing is well established and appears to have been followed correctly.

I only have a couple of very minor points:

1. Methods/Sample acquisition and DNA barcoding: “the tissue was lysed” – by what method? Mechanical disruption or chemical dissociation?.
2. Methods/Sample acquisition and DNA barcoding:”COI” not defined.
3. Methods/Sample acquisition and DNA barcoding:”BOLD” not defined.

**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:*** Genomics, genetics, immunogenetics, immunology, virology

**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.**