DATA NOTE

# The genome sequence of the muscid fly *Eudasyphora cyanicolor* (Zetterstedt, 1845)

[version 1; peer review: 2 approved]

Steven Falk[1], Liam M. Crowley[2], Ruth Y. Akinmusola[3],
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
[3]University of Bath, Bath, England, UK

**v1** **First published:** 10 Feb 2025, **10**:56
https://doi.org/10.12688/wellcomeopenres.23675.1
**Latest published:** 10 Feb 2025, **10**:56
https://doi.org/10.12688/wellcomeopenres.23675.1

## Abstract

We present a genome assembly from an individual female specimen of *Eudasyphora cyanicolor* (Arthropoda; Insecta; Diptera; Muscidae). The genome sequence has a total length of 1,450.40 megabases. Most of the assembly (99.68%) is scaffolded into 5 chromosomal pseudomolecules. The mitochondrial genome has also been assembled and is 19.34 kilobases in length.

## Keywords

Eudasyphora cyanicolor, muscid fly, genome sequence, chromosomal, Diptera

This article is included in the Tree of Life gateway.

**Open Peer Review**

**Approval Status** ✓ ✓

| | 1 | 2 |
|---|---|---|
| **version 1** 10 Feb 2025 | ✓ view | ✓ view |

1. **Jason Charamis**, Foundation for Research and Technology - Hellas, Irákleion, Greece
2. **Rakesh K Mishra**, Tata Institute for Genetics and Society, Bengaluru, India

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; **Akinmusola RY**: Writing – Original Draft Preparation;

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

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



**How to cite this article:** Falk S, Crowley LM, Akinmusola RY *et al.* **The genome sequence of the muscid fly *Eudasyphora cyanicolor* (Zetterstedt, 1845) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, **10**:56
https://doi.org/10.12688/wellcomeopenres.23675.1

**First published:** 10 Feb 2025, **10**:56 https://doi.org/10.12688/wellcomeopenres.23675.1

## 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; Calyptratae; Muscoidea; Muscidae; Muscinae; Muscini; *Eudasyphora*; *Eudasyphora cyanicolor* (Zetterstedt, 1845) (NCBI:txid1643974)

## Background

*Eudasyphora cyanicolor* (Zetterstedt, 1845) belongs to the Muscidae family of houseflies. Muscid flies are diverse, with different members exhibiting varied feeding mechanisms such as phytophagy, saprophagy, hematophagy, and carnivory (Kutty *et al.*, 2014; Moon, 2019). These flies are synanthropes, thriving on various substrates and habitats influenced by human activities and the environment (Moon, 2019).

Some Muscidae members are important vectors of pathogens and parasites with negative impacts on the public hygiene of humans and livestock (Cook, 2020; Issa, 2019; Onwugamba *et al.*, 2018). However, most muscid flies are ecologically relevant in the breakdown and recycling of organic waste (Moon, 2019). The members of the *Eudasyphora* genus are frequent visitors to our homes and sometimes overwinter indoors (Logunov & Dennis, 2016). *E. cyanicolor* is typically associated with woodlands, scrublands, swamps, and animal dung (Taylor, 1999; Tridrikh & Sorokina, 2021).

*E. cyanicolor* can be easily identified by its deep metallic blue body and a stripe of whitish dusting at the front of the mesonotum (Falk, 2016; Grzywacz *et al.*, 2017). These unique features distinguish it from *E. cyanella*, its closest relative, which possesses a green colour (Falk, 2016; Nihei & De Carvalho, 2007). *E. cyanicolor* is widespread in Europe, with the highest records in the United Kingdom and Finland (GBIF Secretariat, 2024). The UK records show adult emergence from March to November, with peak flight activity in summer (NBN Atlas Partnership, 2024).

This article presents a chromosomal-level genome assembly for *E. cyanicolor*, sequenced as part of the Darwin Tree of Life Project. It will contribute to future studies on the biodiversity, genomics and evolution of this species.

## Genome sequence report

The genome of *Eudasyphora cyanicolor* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of 45.53 Gb (gigabases) from 4.39 million reads, providing an estimated 36-fold coverage. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data, which produced 138.68 Gb from 918.42 million reads. Specimen and sequencing details are summarised in Table 1.

This reduced the assembly length by 3.24%, the scaffold number by 31.82%, and the scaffold N50 by 0.82%. The final assembly has a total length of 1,450.40 Mb in 74 sequence scaffolds, with 370 gaps, and a scaffold N50 of 293.4 Mb (Table 2).


**Figure 1. Photograph of the *Eudasyphora cyanicolor* (idEudCyai1) specimen used for genome sequencing.**

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 (99.68%) was assigned to 5 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by the Hi-C data, are named in order of size (Figure 5; Table 3). During manual curation it was noted that there was insufficient information to identify the sex chromosomes. 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).

While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial genome was also assembled and can be found as a contig within the multifasta file of the genome submission, and as a standalone sequence on GenBank.

The primary haplotype has a QV of 61.4, and the combined primary and alternate assemblies achieve an estimated QV of 61.4. The *k*-mer completeness for the primary haplotype is 66.92%, and for the alternate haplotype it is 66.27%. The combined primary and alternate assemblies achieve a *k*-mer completeness of 99.09%. BUSCO analysis using the diptera_odb10 reference set ($n$ = 3,285) indicated a completeness score of 98.7% (single = 97.8%, duplicated = 0.9%). The assembly achieves the EBP reference standard of 6.C.61. Other quality metrics are given in Table 2.

**Table 1.** **Specimen and sequencing data for *Eudasyphora cyanicolor*.**

| **Project information** | | | |
|---|---|---|---|
| **Study title** | Eudasyphora cyanicolor | | |
| **Umbrella BioProject** | PRJEB67607 | | |
| **Species** | *Eudasyphora cyanicolor* | | |
| **BioSample** | SAMEA113425446 | | |
| **NCBI taxonomy ID** | 1643974 | | |
| **Specimen information** | | | |
| **Technology** | **ToLID** | **BioSample accession** | **Organism part** |
| **PacBio long read sequencing** | idEudCyai1 | SAMEA113425543 | Whole organism |
| **Hi-C sequencing** | idEudCyai1 | SAMEA113425543 | Whole organism |
| **Sequencing information** | | | |
| **Platform** | **Run accession** | **Read count** | **Base count (Gb)** |
| **Hi-C Illumina NovaSeq 6000** | ERR12143998 | 9.18e+08 | 138.68 |
| **PacBio Revio** | ERR12205263 | 4.39e+06 | 45.53 |

**Table 2.** **Genome assembly data for *Eudasyphora cyanicolor*, idEudCyai1.1.**

| **Genome assembly** | | |
|---|---|---|
| Assembly name | idEudCyai1.1 | |
| Assembly accession | GCA_963930755.1 | |
| *Accession of alternate haplotype* | *GCA_963930745.1* | |
| Span (Mb) | 1,450.40 | |
| Number of contigs | 445 | |
| Number of scaffolds | 74 | |
| Longest scaffold (Mb) | 365.32 | |
| **Assembly metrics*** | | ***Benchmark*** |
| Contig N50 length (Mb) | 6.9 | *≥ 1 Mb* |
| Scaffold N50 length (Mb) | 293.4 | *= chromosome N50* |
| Consensus quality (QV) | Primary: 61.4; alternate: 61.5; combined 61.4 | *≥ 40* |
| *k*-mer completeness | Primary: 66.92%; alternate: 66.27%; combined: 99.09% | *≥ 95%* |
| BUSCO** | C:98.7%[S:97.8%,D:0.9%], F:0.6%,M:0.7%,n:3,285 | *S > 90%, D < 5%* |
| Percentage of assembly mapped to chromosomes | Primary: 66.92%; alternate: 66.27%; combined: 99.09% | *≥ 90%* |
| Sex chromosomes | Not identified | *localised homologous pairs* |
| Organelles | Mitochondrial genome: 19.34 kb | *complete single alleles* |

* Assembly metric benchmarks are adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024.

** BUSCO scores based on the diptera_odb10 BUSCO set using version 5.4.3. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/Eudasyphora_cyanicolor/dataset/GCA_963930755.1/busco.

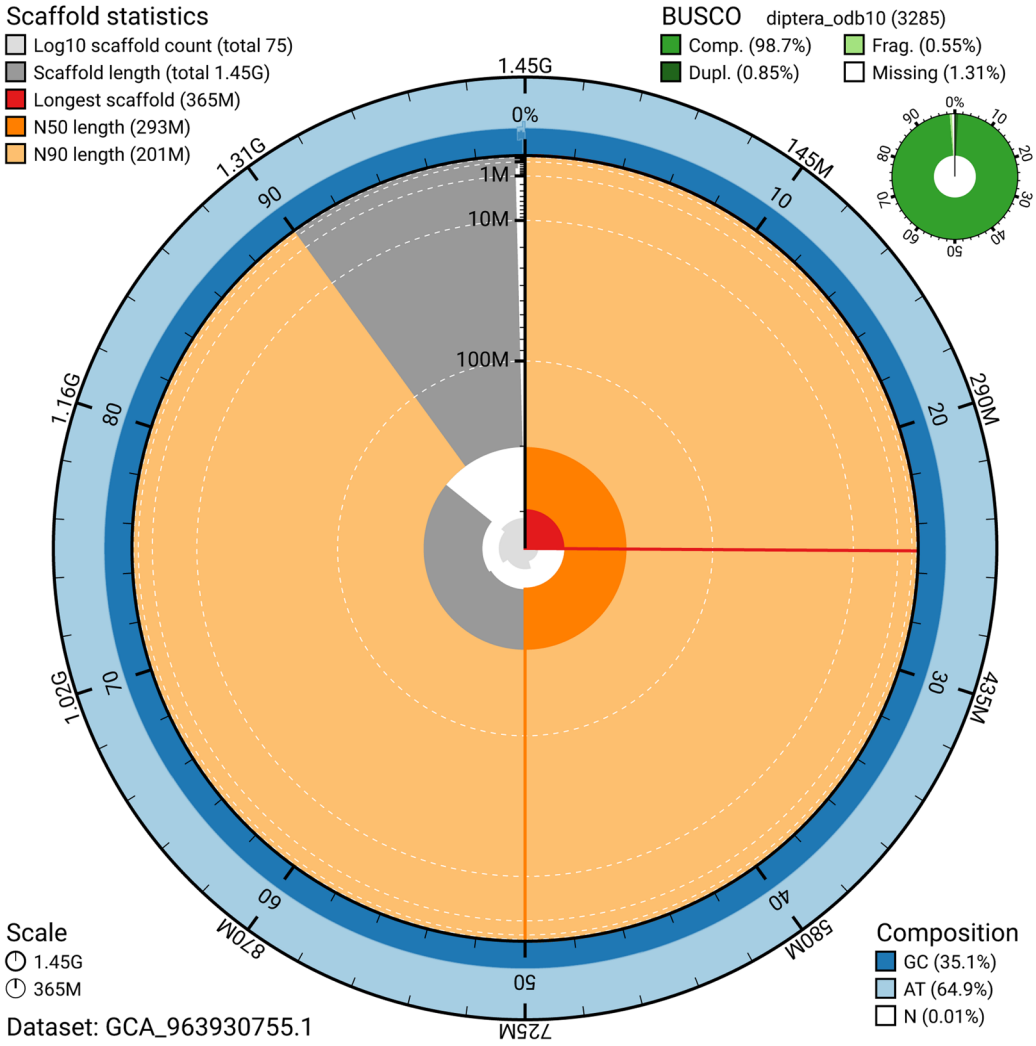


**Figure 2. Genome assembly of *Eudasyphora cyanicolor*, idEudCyai1.1: metrics.** The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around 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_963930755.1/dataset/GCA_963930755.1/snail.

## Methods

### Sample acquisition and DNA barcoding

An adult female *Eudasyphora cyanicolor* (specimen ID Ox002809, ToLID idEudCyai1) was collected from Wytham Woods, Oxfordshire, United Kingdom (latitude 51.77, longitude –1.33) on 2022-07-14 by netting. The specimen was collected by Steven Falk (independent researcher) and Liam Crowley (University of Oxford), identified by Steven Falk and preserved on dry ice.

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 specimens and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding have been deposited on protocols.io (Beasley *et al.*, 2023).

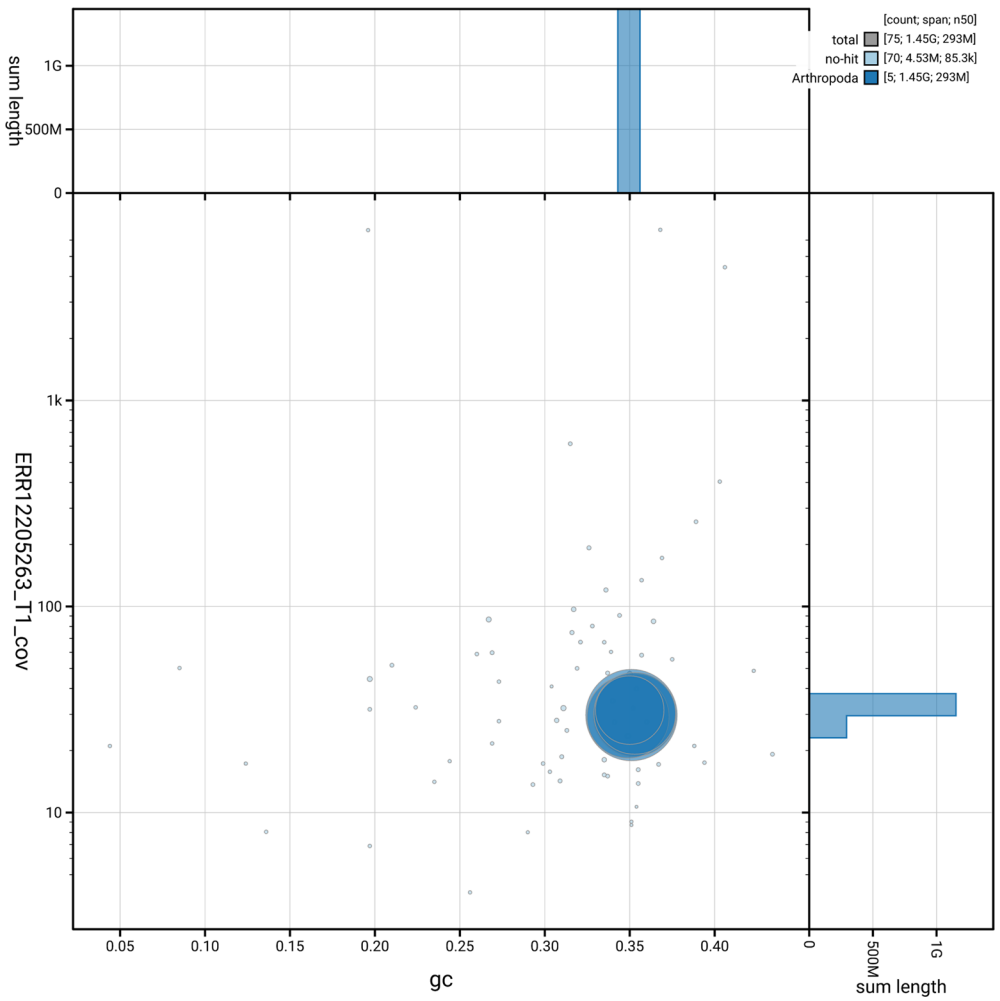


**Figure 3. Genome assembly of *Eudasyphora cyanicolor*, idEudCyai1.1: BlobToolKit GC-coverage plot showing sequence coverage (vertical axis) and GC content (horizontal axis).** The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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_963930755.1/dataset/GCA_963930755.1/blob.

## 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 idEudCyai1 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023), and 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.

## Hi-C sample preparation

Tissue from the idEudCyai1 sample was processed at the WSI Scientific Operations core, using the Arima-HiC v2 kit. Tissue (stored at –80 °C) was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde. After crosslinking, the tissue was homogenised using the Diagnocine Power Masher-II and BioMasher-II tubes and pestles. Following the kit manufacturer's instructions, crosslinked DNA was digested using a restriction enzyme master mix. The 5'-overhangs were then 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.

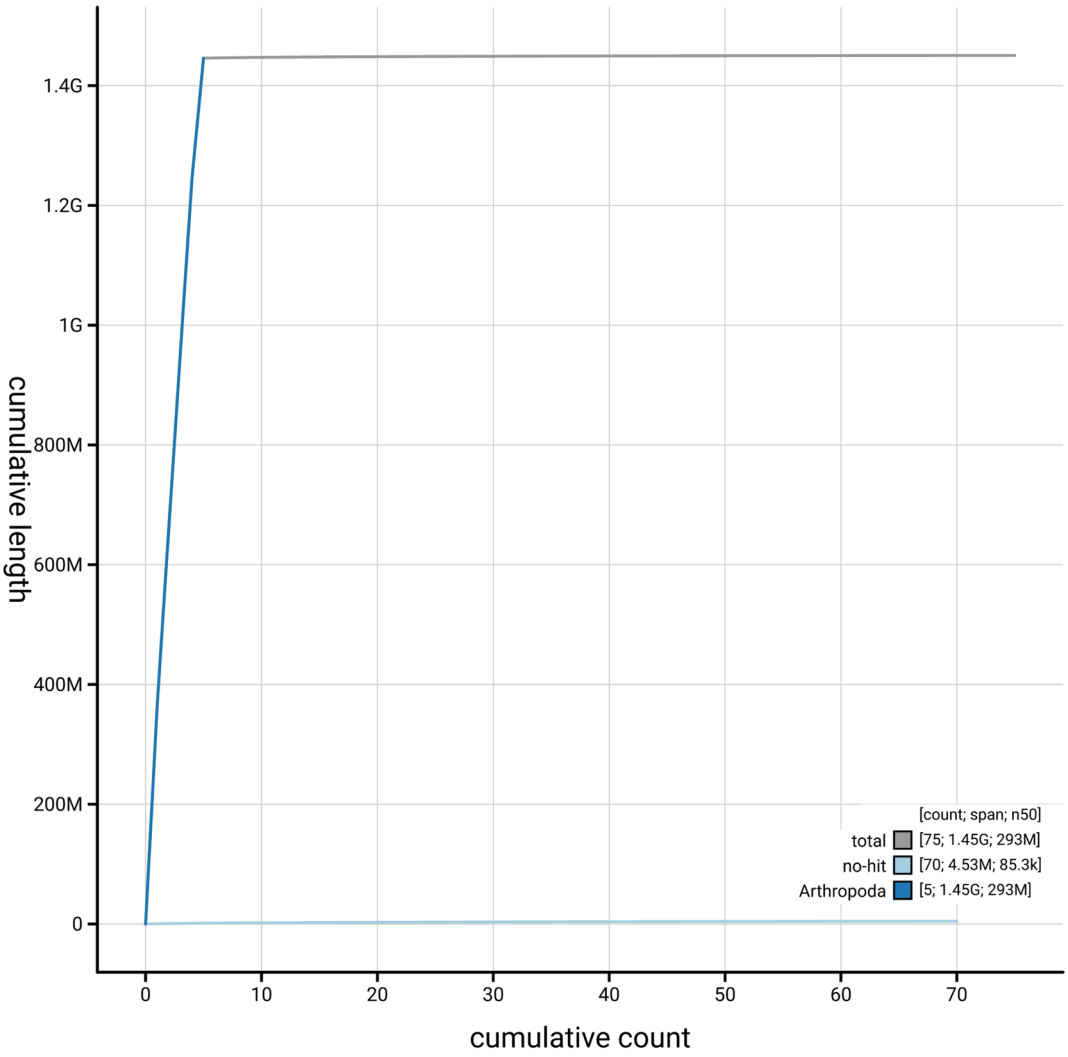


**Figure 4. Genome assembly of *Eudasyphora cyanicolor* idEudCyai1.1: BlobToolKit cumulative sequence plot.** The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963930755.1/dataset/GCA_963930755.1/cumulative.

## Library preparation and sequencing

***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 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 on a Revio instrument (Pacific Biosciences, California, USA). Prepared libraries were normalised to 2 nM, and 15 µL was used for making complexes. Primers were annealed and polymerases were hybridised to create circularised complexes according to manufacturer's instructions. The complexes were purified with the 1.2X clean up with SMRTbell beads. The purified complexes were then diluted to the Revio loading concentration (in the range 200–300 pM), and spiked with a Revio sequencing internal control. Samples were sequenced on Revio 25M SMRT cells (Pacific Biosciences,

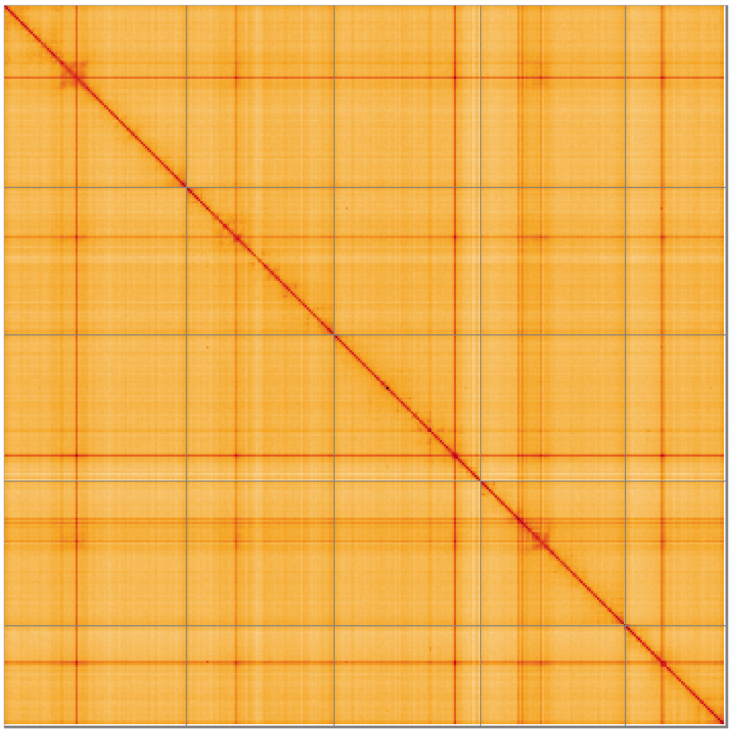

**Figure 5. Genome assembly of *Eudasyphora cyanicolor* idEudCyai1.1: Hi-C contact map of the idEudCyai1.1 assembly, visualised using HiGlass.** Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=VRyzap9aQCqFDjjPjdSuQA.

**Table 3. Chromosomal pseudomolecules in the genome assembly of *Eudasyphora cyanicolor*, idEudCyai1.**

| INSDC accession | Name | Length (Mb) | GC% |
|---|---|---|---|
| OZ005734.1 | 1 | 365.32 | 35.0 |
| OZ005735.1 | 2 | 295.72 | 35.5 |
| OZ005736.1 | 3 | 293.42 | 35.0 |
| OZ005737.1 | 4 | 290.28 | 35.5 |
| OZ005738.1 | 5 | 201.16 | 35.0 |
| OZ005739.1 | MT | 0.02 | 19.5 |

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

## Genome assembly, curation and evaluation

***Assembly***

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.

***Assembly curation***

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline (article in preparation). Flat files and maps used in curation were generated in TreeVal (Pointon *et al.*, 2023). Manual curation was primarily conducted using PretextView (Harry, 2022), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023) and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were corrected, and duplicate sequences were tagged and removed. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation (article in preparation).

***Evaluation of the final assembly***

The Merqury.FK tool (Rhie *et al.*, 2020) 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 pre-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, public 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 is visualised in HiGlass (Kerpedjiev *et al.*, 2018).

The blobtoolkit pipeline is a Nextflow 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 GoaT 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 genome assembly and evaluation pipelines were 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.

## 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'**,

**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 | 4.3.7 | https://github.com/blobtoolkit/blobtoolkit |
| BUSCO | 5.4.3 and 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 | 427104ea91c78c3b8b8b49f1a7d6bbeaa869ba1c | 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 | 0.19.8-r587 | https://github.com/chhylp123/hifiasm |
| HiGlass | 44086069ee7d4d3f6f3f0012569789ec138f42b84aa44357826c0b6753eb28de | https://github.com/higlass/higlass |

| Software tool | Version | Source |
|---|---|---|
| Merqury.FK | d00d98157618f4e8d1a9190026b19b471055b22e | https://github.com/thegenemyers/MERQURY.FK |
| MitoHiFi | 3 | https://github.com/marcelauliano/MitoHiFi |
| MultiQC | 1.14, 1.17, and 1.18 | https://github.com/MultiQC/MultiQC |
| NCBI Datasets | 15.12.0 | https://github.com/ncbi/datasets |
| Nextflow | 23.04.0-5857 | 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 | 1.16.1, 1.17, and 1.18 | https://github.com/samtools/samtools |
| sanger-tol/ascc | - | https://github.com/sanger-tol/ascc |
| sanger-tol/blobtoolkit | 0.6.0 | https://github.com/sanger-tol/blobtoolkit |
| Seqtk | 1.3 | https://github.com/lh3/seqtk |
| Singularity | 3.9.0 | https://github.com/sylabs/singularity |
| TreeVal | 1.0.0 | https://github.com/sanger-tol/treeval |
| YaHS | 1.2a.2 | https://github.com/c-zhou/yahs |

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: Eudasyphora cyanicolor. Accession number PRJEB67607; https://identifiers.org/ena.embl/PRJEB67607. The genome sequence is released openly for reuse. The *Eudasyphora cyanicolor* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project. All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

### Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.12157525.

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

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

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

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

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

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

## References


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

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

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

Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor®3 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

Cook D: **A historical review of management options used against the stable fly (Diptera: Muscidae).** *Insects.* 2020; **11**(5): 313.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

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

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

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: **British blow flies (Calliphoridae) and woodlouse flies (Rhinophoridae).** Draft key to British Calliphoridae and Rhinophoridae, 2016.
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: ***Eudasyphora cyanicolor* (Zetterstedt, 1845).** *GBIF Backbone Taxonomy.* 2024.
Reference Source

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

Grzywacz A, Ogiela J, Tofilski A: **Identification of Muscidae (Diptera) of medico-legal importance by means of wing measurements.** *Parasitol Res.* 2017; **116**(5): 1495–1504.
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

Issa R: ***Musca domestica* acts as transport vector hosts.** *Bull Natl Res Cent.* 2019; **43**(1): 73.
Publisher 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

Kutty SN, Pont AC, Meier R, *et al.*: **Complete tribal sampling reveals basal split in Muscidae (Diptera), confirms saprophagy as ancestral feeding mode, and reveals an evolutionary correlation between instar numbers and carnivory.** *Mol Phylogenet Evol.* 2014; **78**: 349–364.
PubMed Abstract | 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

Logunov DV, Dennis RL: **The insects and other invertebrates of alderley edge.** In: *The story of Alderley Edge: living with the edge.* First edition. Manchester University Press, 2016; 220–299.
Reference Source

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

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

Moon RD: **Muscid flies (Muscidae).** In: *Med Vet Entomol.* Academic Press, 2019; 345–368.
Publisher Full Text

NBN Atlas Partnership : ***Eudasyphora cyanicolor* (Zetterstedt, 1845).** *NBN Atlas.* 2024.
Reference Source

Nihei SS, De Carvalho CJB: **Phylogeny and classification of Muscini (Diptera, Muscidae).** *Zool J Linn Soc.* 2007; **149**(4): 493–532.
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

Onwugamba FC, Fitzgerald JR, Rochon K, *et al.*: **The role of 'filth flies' in the spread of antimicrobial resistance.** *Travel Med Infect Dis.* 2018; **22**: 8–17.
PubMed Abstract | 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

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**

Taylor SM: **Sheep scab—environmental considerations of treatment with doramectin.** *Vet Parasitol.* 1999; **83**(3–4): 309–317.
**PubMed Abstract** | **Publisher Full Text**

Tridrikh NN, Sorokina VS: **Muscid fly (Diptera, Muscidae) distribution in river floodplain, forest and swamp biotopes of Northern Okhotia (Magadan Province, Russia).** *Entomol Rev.* 2021; **101**(6): 820–836.
**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**

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**

Vicoso 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 27 February 2025

https://doi.org/10.21956/wellcomeopenres.26118.r118406



✔ **Rakesh K Mishra**
Tata Institute for Genetics and Society, Bengaluru, Karnataka, India

Authors in this report present genome sequencing and assembly of muscid fly. They also use NGS based approach to assemble the sequence data in five molecules (chromosomes). The description of the process and approach is described in sufficient details. This will be a very useful information in the fast growing field of genome biology of insects for understanding basic biology and human health perspective as the information is of potential use in genome based technologies of vector control.

**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:*** Genome organisation, comparative genomics, genetics and epigenetic regulation 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 February 2025

https://doi.org/10.21956/wellcomeopenres.26118.r119418



**Jason Charamis**
Foundation for Research and Technology - Hellas, Irákleion, Greece

This article describes the genome sequencing of the housefly *Eudasyphora cyanicolor*. The study produced a highly contiguous and highly accurate genome assembly of all main chromosomes, as demonstrated by the BUSCO analysis and the k-mer based QV metric, except for the sex chromosomes. This is a common challenge in dipteran genome assemblies, due to the frequent evolutionary turnover of sex chromosomes in this order.

The manuscript is mostly ready for indexing. My only request is to better describe the relevance of sequencing of this species for future ecological studies. For example, because of its synanthropic behavior, this species has been probably exposed to pesticides for several decades. In this context, genome sequencing would be useful for studying insecticide resistance in non-target insects. I would like to see the Introduction section expanded to better establish the relevance of this genome assembly.

**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:*** Arthropod Comparative Genomics

**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.**