DATA NOTE

The genome sequence of the Broad Centurion soldierfly, *Chloromyia formosa* (Scopoli, 1763)

[version 1; peer review: 4 approved]

Olga Sivell[1], Duncan Sivell[1], 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]Natural History Museum, London, England, UK

v1 **First published:** 23 May 2025, **10**:266
https://doi.org/10.12688/wellcomeopenres.24255.1
Latest published: 23 May 2025, **10**:266
https://doi.org/10.12688/wellcomeopenres.24255.1

Abstract

We present a genome assembly from a male specimen of *Chloromyia formosa* (Broad Centurion soldierfly; Arthropoda; Insecta; Diptera; Stratiomyidae). The genome sequence has a total length of 790.81 megabases. Most of the assembly (98.79%) is scaffolded into 7 chromosomal pseudomolecules, including the X sex chromosome. The mitochondrial genome has also been assembled, with a length of 17.69 kilobases.

Keywords

Chloromyia formosa, Broad Centurion soldierfly, genome sequence, chromosomal, Diptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ✓ ✓

	1	2	3	4
version 1 23 May 2025	✓ view	✓ view	✓ view	✓ view

1. **Kyle M Benowitz**, Arizona State University, Tempe, USA
2. **Komal Kumar Raja**, Baylor College of Medicine, Houston, USA
3. **Panagiotis Ioannidis**, Foundation for Research & Technology - Hellas, Crete, Greece
4. **Terrence Sylvester**, University of memphis, memphis, USA

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: Sivell O: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Sivell D**: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing;

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.

How to cite this article: Sivell O, Sivell D, Darwin Tree of Life Barcoding collective *et al.* **The genome sequence of the Broad Centurion soldierfly, *Chloromyia formosa* (Scopoli, 1763) [version 1; peer review: 4 approved]** Wellcome Open Research 2025, **10**:266 https://doi.org/10.12688/wellcomeopenres.24255.1

First published: 23 May 2025, **10**:266 https://doi.org/10.12688/wellcomeopenres.24255.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Stratiomyomorpha; Stratiomyidae; Sarginae; *Chloromyia*; *Chloromyia formosa* (Scopoli, 1763) (NCBI:txid343672)

Background

Chloromyia formosa (Scopoli, 1763) is a very characteristic and widely distributed species from the family Stratiomyidae, commonly called soldierflies. It is medium sized measuring 7–9 mm in length, with a broad abdomen. The thorax of the fly is metallic green (Figure 1), the abdomen bluish in females and golden-brown in males. It has densely hairy eyes which are holoptic in males (touching), the legs are mostly black. It is the only species from the genus *Chloromyia* Duncan, 1837 occurring in Britain. Only six other species are known, with only one of them, *C. speciosa* (Macquart, 1834) occur widely in the Palaearctic, one described from China (Yang *et al.*, 2014) (and others being Afrotropical (Woodley, 2001).

The adults are on the wing from early May to the end of August, occasionally to the beginning of October, peaking in early July (Drake, 1991). *Chloromyia formosa* can be found in a range of habitats, usually damp and with fertile soil, grassland and woodland. Adults frequent flowers, particularly umbellifers such as hogweed, on which they feed. The larvae develop in rotting vegetation, compost or cow dung (Drake, 1991; Zeegers & Schulten, 2022). The larva is similar in appearance to *Sargus* Fabricius, 1798, but with more numerous longitudinal stripes (7–9), and one central dark stripe (Stubbs & Drake, 2014). The larvae can be found in their breeding substrate, e.g. compost, or in grass tussocks. Males have been occasionally observed hovering above ground (Stubbs & Drake, 2014) and swarming behaviour has also been noted in this species (Baldock & Early, 2015; Stubbs & Drake, 2014).

Figure 1. Photograph of *Chloromyia formosa* by Keith Edkins (not the specimen used for genome sequencing).

This Palaearctic species is widespread in Europe and north Africa (GBIF Secretariat, 2023; Yimlahi *et al.*, 2024), in Britain it is very common and widely dispersed, becoming less frequent in northern Scotland (Drake, 1991). It has been introduced and become established in the United States (Hoebeke & Pechuman, 1982; Pechuman, 1974).

The high-quality genome of *Chloromyia formosa* presented here was sequenced from a single specimen (SAMEA7524274) from Wigmore Park, Luton, England, collected by Olga Sivell and identified by Duncan Sivell. The high-quality genome was sequenced as part of the Darwin Tree of Life Project, a collaborative effort to sequence all named eukaryotic species in the Atlantic Archipelago of Britain and Ireland.

Genome sequence report

Sequencing data

The genome of a specimen of *Chloromyia formosa* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 26.30 Gb (gigabases) from 2.41 million reads, which were used to assemble the genome. GenomeScope analysis estimated the haploid genome size at 724.51 Mb, with a heterozygosity of 1.22% and repeat content of 39.68%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 34 coverage. Hi-C sequencing produced 81.44 Gb from 539.32 million reads, used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

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 65 misjoins or missing joins and removed 5 haplotypic duplications. These interventions decreased the scaffold count by 16.79% and increased the scaffold N50 by 98.92%. The final assembly has a total length of 790.81 Mb in 113 scaffolds, with 264 gaps, and a scaffold N50 of 156.03 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 (98.79%) was assigned to 7 chromosomal-level scaffolds, representing 6 autosomes and the X sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3). During curation, the X chromosome was identified based on PacBio read coverage. The Y chromosome may be present in the unassembled contigs, but cannot be confidently identified.

Table 1. **Specimen and sequencing data for *Chloromyia formosa*.**

Project information			
Study title	Chloromyia formosa		
Umbrella BioProject	PRJEB70502		
Species	*Chloromyia formosa*		
BioSpecimen	SAMEA7524274		
NCBI taxonomy ID	343672		
Specimen information			
Technology	**ToLID**	**BioSample accession**	**Organism part**
PacBio long read sequencing	idChlForm1	SAMEA7524373	thorax
Hi-C sequencing	idChlForm1	SAMEA7524373	thorax
Sequencing information			
Platform	**Run accession**	**Read count**	**Base count (Gb)**
Hi-C Illumina NovaSeq 6000	ERR12321263	5.39e+08	81.44
PacBio Sequel IIe	ERR12340371	1.66e+06	19.66
PacBio Sequel II	ERR12373373	7.52e+05	6.64

Table 2. **Genome assembly data for *Chloromyia formosa*.**

Genome assembly		
Assembly name	idChlForm1.1	
Assembly accession	GCA_964017055.1	
Alternate haplotype accession	*GCA_964017035.1*	
Assembly level for primary assembly	chromosome	
Span (Mb)	790.81	
Number of contigs	377	
Number of scaffolds	113	
Longest scaffold (Mb)	171.21	
Assembly metric	**Measure**	***Benchmark***
Contig N50 length	5.53 Mb	*≥ 1 Mb*
Scaffold N50 length	156.03 Mb	*= chromosome N50*
Consensus quality (QV)	Primary: 60.3; alternate: 58.4; combined: 58.9	*≥ 40*
k-mer completeness	Primary: 79.45%; alternate: 76.13%; combined: 99.10%	*≥ 95%*
BUSCO*	C:95.0%[S:94.4%,D:0.6%], F:1.3%,M:3.7%,n:3,285	*S > 90%; D < 5%*
Percentage of assembly assigned to chromosomes	98.79%	*≥ 90%*
Sex chromosomes	X	*localised homologous pairs*
Organelles	Mitochondrial genome: 17.69 kb	*complete single alleles*

* 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.

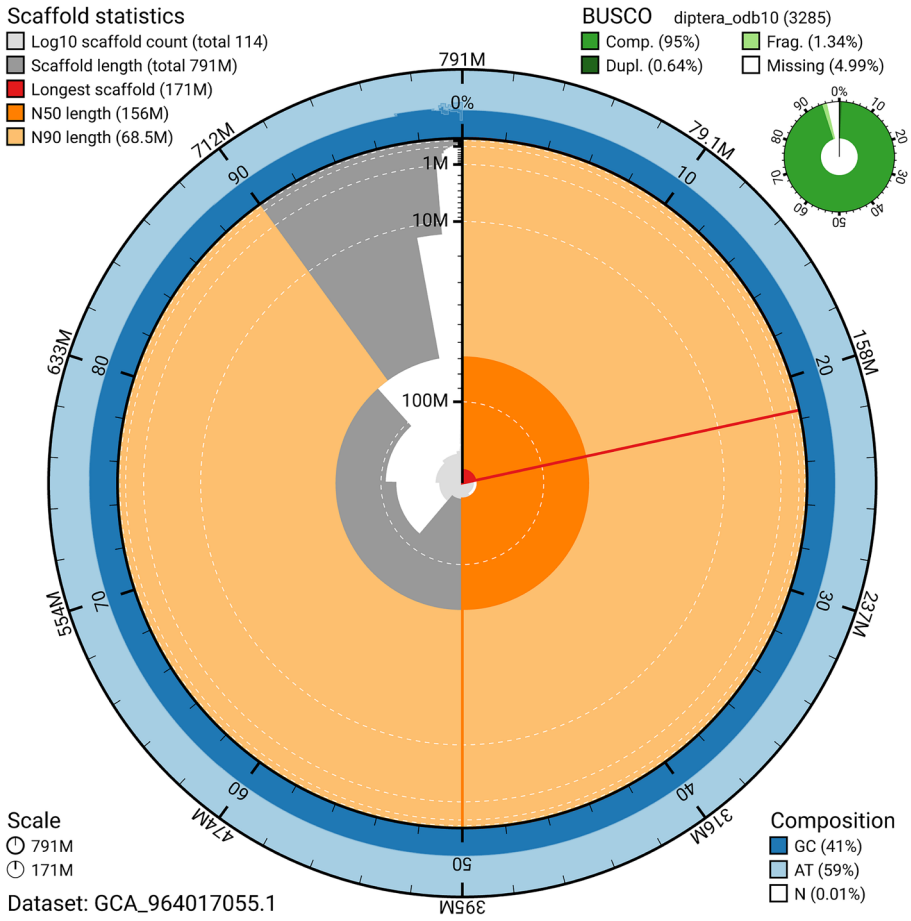

Figure 2. Genome assembly of *Chloromyia formosa, idChlForm1.1: metrics.* The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. 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_964017055.1/dataset/GCA_964017055.1/snail.

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 combined primary and alternate assemblies achieve an estimated QV of 58.9. The *k*-mer completeness is 79.45% for the primary haplotype and 76.13% for the alternate haplotype; and 99.10% for the combined primary and alternate assemblies. BUSCO v.5.5.0 analysis using the diptera_odb10 reference set (n = 3,285) identified 95.0% of the expected gene set (single = 94.4%, duplicated = 0.6%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The primary assembly achieves the EBP reference standard of **6.C.Q60**.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Chloromyia formosa* (specimen ID SAN0001252, ToLID idChlForm1), collected from Wigmore Park, Percival Way,

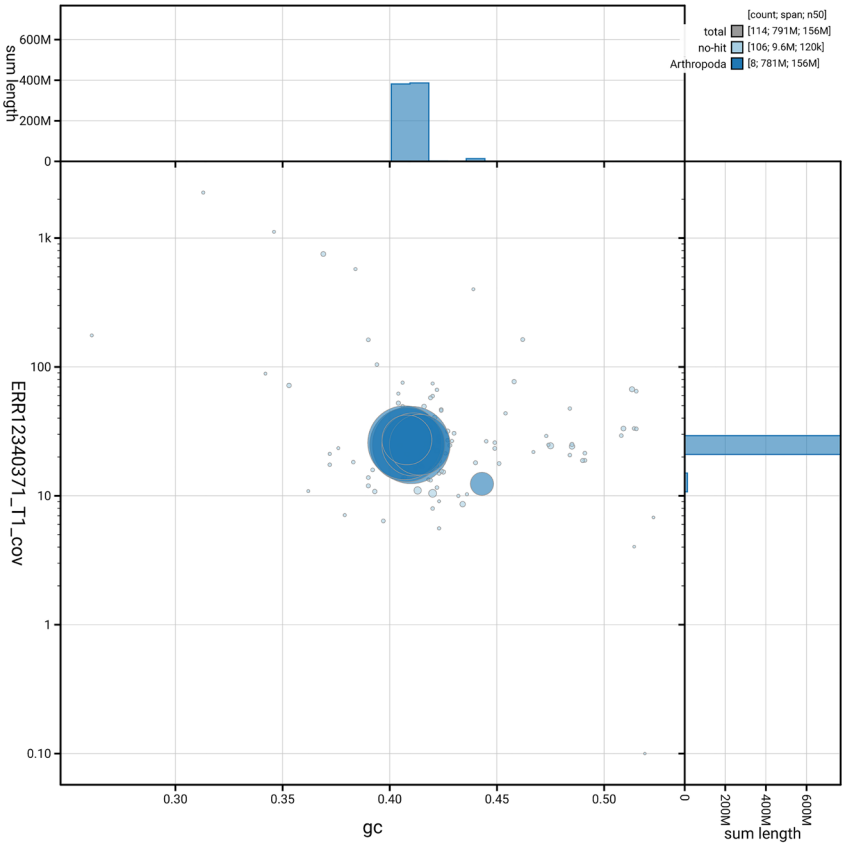

Figure 3. Genome assembly of *Chloromyia formosa*, idChlForm1.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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_964017055.1/dataset/GCA_964017055.1/blob.

Wigmore, Luton, United Kingdom (latitude 51.8838, longitude –0.3686) on 2020-05-20 by netting in an urban park. The specimen was collected by Olga Sivell (Natural History Museum), identified by Duncan Sivell (Natural History Museum) 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 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 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 (Howard *et al.*, 2025). Detailed protocols are available on protocols.io (Denton *et al.*, 2023b). The idChlForm1 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the 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.*, 2023a). For ultra-low input (ULI) PacBio sequencing, DNA was fragmented using the Covaris g-TUBE method (Oatley *et al.*, 2023b). Sheared DNA was purified by solid-phase reversible

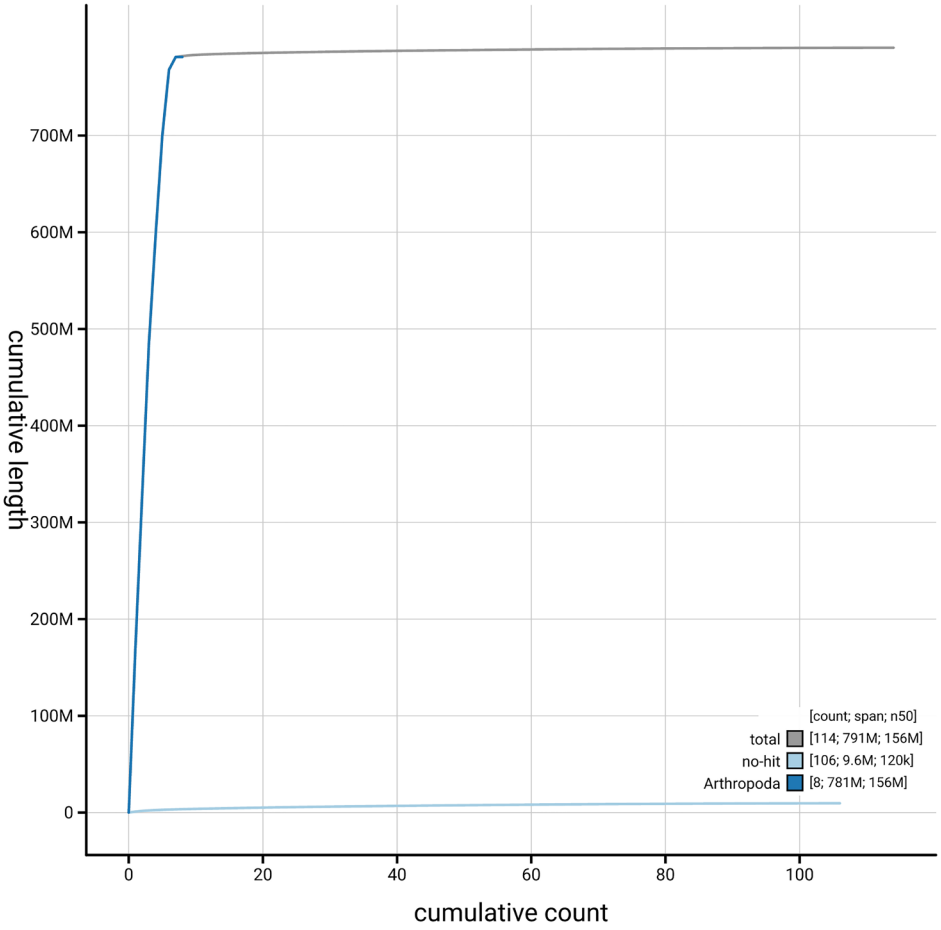

Figure 4. Genome assembly of *Chloromyia formosa,* idChlForm1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_964017055.1/dataset/GCA_964017055.1/cumulative.

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.

PacBio HiFi (ULI) library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. The sample requires Covaris g-TUBE shearing to approximately 10 kb prior to library preparation. Ultra-low input libraries were prepared using PacBio SMRTbell® Express Template Prep Kit 2.0 and PacBio SMRTbell® gDNA Sample Amplification Kit. To begin, samples were normalised to 20 ng of DNA. Initial removal of single-strand overhangs, DNA damage repair, and end repair/A-tailing were performed per manufacturer's instructions. From the SMRTbell® gDNA Sample Amplification Kit, amplification adapters were then ligated. A 0.85X pre-PCR clean-up was performed with Promega ProNex beads and the sample was then divided into two for a dual PCR. PCR reactions A and B each followed the PCR programs as described in the manufacturer's protocol. A 0.85X post-PCR clean-up was performed with ProNex beads for PCR reactions A and B and DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit and fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and gDNA 55kb BAC analysis kit. PCR reactions A and B were then pooled, ensuring the total mass was ≥500 ng in 47.4 μl. The pooled sample then repeated the process for DNA damage repair, end repair/A-tailing and additional hairpin adapter ligation. A 1X clean-up was performed with ProNex beads and DNA concentration was quantified using the Qubit and fragment size analysis was carried out using the Agilent Femto Pulse

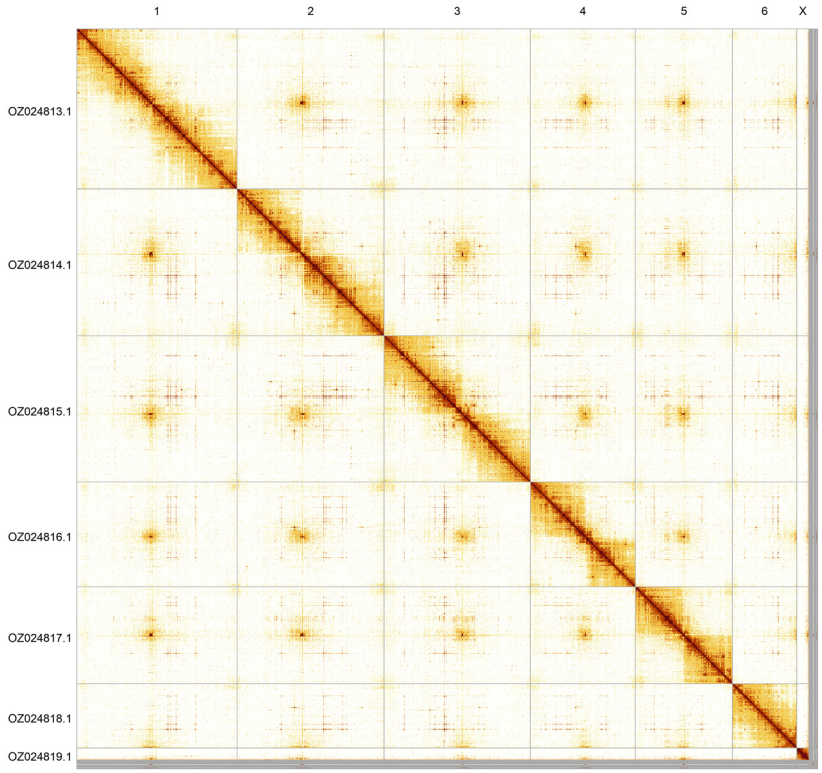

Figure 5. Genome assembly of *Chloromyia formosa*. Hi-C contact map of the idChlForm1.1 assembly, generated using PretextSnapshot. Chromosomes are shown in order of size and labelled with chromosome numbers (top) and chromosome accession numbers (left).

Table 3. Chromosomal pseudomolecules in the genome assembly of *Chloromyia formosa*, idChlForm1.

INSDC accession	Name	Length (Mb)	GC%
OZ024813.1	1	171.21	41
OZ024814.1	2	156.83	41
OZ024815.1	3	156.03	40.5
OZ024816.1	4	111.97	41
OZ024817.1	5	103.58	41.5
OZ024818.1	6	68.48	41
OZ024819.1	X	13.11	44.5
OZ024820.1	MT	0.02	26

Automated Pulsed Field CE Instrument (Agilent Technologies). Size selection was performed using Sage Sciences' PippinHT system with target fragment size determined by analysis from the Femto Pulse, usually a value between 4000 and 9000 bp. Size selected libraries were then cleaned-up using 1.0X ProNex beads and normalised to 2 nM before proceeding to sequencing.

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

Sample preparation and crosslinking

Hi-C data were generated from the thorax of the idChlForm1 sample using the Arima-HiC v2 kit (Arima Genomics) with 20–50 mg of frozen tissue (stored at –80 °C). As per manufacturer's instructions, tissue was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde concentration, and a final formaldehyde concentration of 2%. The tissue was then homogenised using the Diagnocine Power Masher-II. The crosslinked DNA was digested using a restriction enzyme master mix, then biotinylated and ligated. A clean up was performed with SPRIselect beads prior to library preparation. DNA

concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit, and sample biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

For Hi-C library preparation, the biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size-selected to 400–600 bp using SPRISelect beads. DNA was then enriched using Arima-HiC v2 Enrichment beads. The NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) was used for end repair, A-tailing, and adapter ligation, following a modified protocol in which library preparation is carried out while the DNA remains bound to the enrichment beads. PCR amplification was performed using KAPA HiFi HotStart mix and custom dual-indexed adapters (Integrated DNA Technologies) in a 96-well plate format. Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, samples were amplified for 10–16 PCR cycles. Post-PCR clean-up was carried out using SPRISelect beads. The libraries were quantified using the Accuclear Ultra High Sensitivity dsDNA Standards Assay kit (Biotium) and normalised to 10 ng/μL before sequencing. Hi-C sequencing was performed on the Illumina NovaSeq 6000 instrument using 150 bp paired-end reads.

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 first 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) 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. 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 sex chromosome was identified by read coverage. 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 ($k = 31$) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

The genome was analysed in the blobtoolkit pipeline, 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 (Danecek *et al.*, 2021) 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 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.

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

Table 4. Software tools: versions and sources.

Software tool	Version	Source
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
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/chhylp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b8 4aa44357826c0b6753eb28de	https://github.com/higlass/higlass
MerquryFK	d00d98157618f4e8d1a9190026b19b471055b 22e	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
PretextSnapshot	-	https://github.com/sanger-tol/PretextSnapshot
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
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
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

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: Chloromyia formosa. Accession number PRJEB70502; https://identifiers.org/ena.embl/PRJEB70502. The genome sequence is released openly for reuse. The *Chloromyia formosa* 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. 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 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.14870789.

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

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

Baldock DW, Early JP: **Soldierflies, their allies and conopidae of surrey.** Surrey Wildlife Trust, 2015.
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

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 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

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

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

Drake CM: **Provisional atlas of the larger Brachycera (Diptera) of Britain and Ireland.** Biological Records Centre, Institute of Terrestrial Ecology, 1991.
Reference Source

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

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

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

GBIF Secretariat: ***Chloromyia formosa* (Scopoli, 1763)**. In: *GBIF Backbone Taxonomy.* 2023; [Accessed 6 May 2025].
Publisher Full Text

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

Hoebeke ER, Pechuman LL: **New North American records for the Palearctic soldier fly, *Chloromyia formosa* (Scopoli) (Diptera: Stratiomyidae).** *Ent News.* 1982; **93**(5): 177–179.
Reference Source

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**

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, 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

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

Oatley G, Sampaio F, Kitchin L, *et al.*: **Sanger Tree of Life HMW DNA fragmentation: Covaris g-TUBE for ULI PacBio.** *protocols.io.* 2023b; [Accessed 13 June 2024].
Publisher Full Text

Pechuman LL: ***Chloromyia formosa* established in the United States (Diptera: Stratiomyidae).** *Ent News.* 1974; **85**: 54–55.
Reference Source

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

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

Stubbs A, Drake M: **British soldierflies and their allies.** 2nd ed. British Entomological and Natural History Society, 2014.
Reference Source

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

Woodley NE: **A world catalog of the Stratiomyidae (Insecta: Diptera).** *Myia.* 2001; **11**.
Reference Source

Yang D, Zhang TT, Li Z: **Stratiomyoidea of China.** Beijing: China Agricultural University Press, 2014.

Yimlahi D, Belqat B, Üstüner T, *et al.*: **New records of the Moroccan soldier flies (Diptera: Stratiomyidae) with the description of the male of *Oxycera flava* (Lindner, 1938).** *J Entomol Res Soc.* 2024; **26**(1): 93–112.
Publisher Full Text

Zeegers T, Schulten A: **Field guide to flies with three pulvilli: families of Homeodactyla of Northwest Europe.** Jeugdbondsuitgeverij, 2022.
Reference Source

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 25 June 2025

https://doi.org/10.21956/wellcomeopenres.26752.r124267

✔ **Terrence Sylvester**
University of memphis, memphis, USA

The authors provide a high-quality genome assembly of the Broad Centurion soldierfly, *Chloromyia formosa*.
Authors provide detailed background information on its morphology, life history and geographic distribution.

The authors conducted a preliminary genome size estimation using k-mer-based analysis of the HiFi reads, which served as a benchmark to assess the completeness and plausibility of the final assembly. While the final assembly size is somewhat larger than the estimated genome size, this discrepancy is not unusual and is likely due to the inherent limitations of k-mer-based estimators, which can both over- and underestimate genome size depending on heterozygosity, repeat content, and sequencing depth.

The resulting genome assembly appears to be of high quality, supported by a range of standard genome quality metrics (e.g., contiguity, completeness, and error correction). The authors also note that the assembly was curated to minimise structural and base-level errors, which further increases its reliability as a reference genome.

This genome will likely serve as a valuable resource for the broader research community, facilitating studies in comparative genomics, functional genomics, and evolutionary biology. To further enhance the utility of this resource, I recommend that the authors include additional information on the quality and characteristics of the DNA used for PacBio sequencing. Specifically, the inclusion of quality control metrics such as Nanodrop spectra, Femto Pulse or Bioanalyzer fragment size distributions, and fluorometric quantifications (e.g., Qubit) would be highly beneficial. These data would provide researchers with a reference point for the level of input DNA quality required to obtain similar assembly outcomes, which is particularly helpful for those working on non-model or difficult-to-extract species.

Beyond this, I do not have any major concerns regarding the manuscript. The authors have done a commendable job in assembling and describing this genome, and I support its publication after

the minor suggestion above is addressed.

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

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 23 June 2025

https://doi.org/10.21956/wellcomeopenres.26752.r124883

Panagiotis Ioannidis
Foundation for Research & Technology - Hellas, Crete, Greece

This paper by Sivell et al describes the sequencing and assembly of the genome of the soldierfly Chloromyia formosa. All quality metrics demonstrate that the genome assembly is of excellent quality. The only thing missing is an RNAseq library which would help in gene prediction.

Minor comments:

"...sequencing data provided approximately 34 coverage" should be "...sequencing data provided approximately 34x coverage"

In "Assembly statistics" the authors mention: "These interventions decreased the scaffold count by 16.79% and increased the scaffold N50 by 98.92%". I think that mentioning the absolute numbers instead of the percentages is much more useful for the reader.

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: insects genomics, bioinformatics

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 14 June 2025

https://doi.org/10.21956/wellcomeopenres.26752.r124877

Komal Kumar Raja
Baylor College of Medicine, Houston, Texas, USA

The article presents the genome sequence of the Broad Centurion soldierfly, *Chloromyia formosa* with a total length of 790.81 megabases. The authors briefly present the morphological and ecological features of this species in the introduction. The assembly quality metrics are great with 95% BUSCO and k-mer completeness of 99.1%. The Hi-C contact map looks quite good with clean separation of distinct blocks as expected in Hi-C maps of well-separated chromsomes. Overall, the data is a good genomic resource.

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: Single cell genomics, bioinformatics, genomics and Neurobiology

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 05 June 2025

https://doi.org/10.21956/wellcomeopenres.26752.r124881

Kyle M Benowitz
Arizona State University, Tempe, Arizona, USA

The manuscript presents a chromosome-level genome assembly of a dipteran species using standard methods of HiFi and Hi-C sequencing. The assembly is clearly of excellent quality, and I have no issues with the description of the methods or the presentation of the assembly. It's worth noting that there is no annotation provided; this could limit the usefulness of the genome for certain comparative analyses.

I do wish there was a bit more rationale provided for the sequencing of this species. The only justification given is that, as a species native to Britain, it falls under the remit of the DToL project to sequence all species in the British islands. This is fine, but nonetheless, I am curious to know whether there are any biological or phylogenetic reasons why having a genome for this species is important. What specific questions can this resource be helpful in addressing?

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

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: Insect genomics; behavior; evolution

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.