DATA NOTE

# The genome sequence of a long-legged fly, *Neurigona quadrifasciata* (Fabricius, 1781) (Diptera: Dolichopodidae)

[version 1; peer review: 3 approved]

Liam M. Crowley[1], James McCulloch[1,2], Steven Falk[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]University of Oxford, Oxford, England, UK
[2]Wellcome Sanger Institute, Hinxton, England, UK
[3]Independent researcher, Kenilworth, Warwickshire, England, UK

**v1** **First published:** 24 Oct 2025, **10**:597
https://doi.org/10.12688/wellcomeopenres.24974.1
**Latest published:** 24 Oct 2025, **10**:597
https://doi.org/10.12688/wellcomeopenres.24974.1

## Abstract

We present a genome assembly from an individual female *Neurigona quadrifasciata* (long-legged fly; Arthropoda; Insecta; Diptera; Dolichopodidae). The genome sequence has a total length of 582.53 megabases. Most of the assembly (99.61%) is scaffolded into 6 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 16.89 kilobases. 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

Neurigona quadrifasciata; long-legged 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** 24 Oct 2025 | ✓ view | ✓ view | ✓ view |

1. **Dong Mei Li**, Ministry for Primary Industries, Auckland, New Zealand
2. **Stuart J.E. Baird**, Academy of Sciences of the Czech Republic, Národní, Czech Republic
3. **Kristina Gagalova**, Curtin University, Perth, Australia

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

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

**Author roles: Crowley LM**: Investigation, Resources; **McCulloch J**: Investigation, Resources; **Falk S**: 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).
*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:** Crowley LM, McCulloch J, Falk S *et al.* **The genome sequence of a long-legged fly, *Neurigona quadrifasciata* (Fabricius, 1781) (Diptera: Dolichopodidae) [version 1; peer review: 3 approved]** Wellcome Open Research 2025, **10**:597 https://doi.org/10.12688/wellcomeopenres.24974.1

**First published:** 24 Oct 2025, **10**:597 https://doi.org/10.12688/wellcomeopenres.24974.1

## Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Eremoneura; Empidoidea; Dolichopodidae; Neurigoninae; *Neurigona*; *Neurigona quadrifasciata* (Fabricius, 1781) (NCBI:txid240896)

## Background

*Neurigona quadrifasciata* (Fabricius, 1781), a species of long-legged fly, is a neurigonine dolichopodid. In Britain the subfamily Neurigoninae comprises only *Neurigona* (Fonseca, 1978). Adults are largely yellow with dark basal bands on abdominal tergites; males have the 3rd–4th fore-tarsomeres black and plumose with the 5th white, and the discal vein shows a double bend (Fonseca, 1978). Body length is approximately 4–5 mm (Grichanov, 2010).

Males of several *Neurigona* species have been observed flying in a zig-zag up tree trunks, and adults are predatory on small insects on trunks; courtship and predation on typhlocybine leafhoppers have been noted for *Neurigona* (summarised by Fonseca, 1978). Field observations (Poland) report adults resting and hunting on tree trunks in damp woods and along streams, active mainly from May to September, with males displaying black-fringed fore tarsi (segments 3–4) and a white 5th segment (Insektarium, 2014).

*Neurigona quadrifasciata* is widespread in the West Palaearctic, recorded across much of Europe and western Russia east to the Urals, Krasnoyarsk and Baikal (Grichanov, 2010). In Britain and Ireland, records are scattered rather than abundant (NBN Atlas Partnership, 2025).

We present a chromosome-level genome sequence for *Neurigona quadrifasciata* — the first genome assembly for the genus *Neurigona* and one of nine genomes available for the family Dolichopodidae as of September 2025 (data obtained via NCBI datasets, O'Leary *et al.*, 2024). The assembly was produced using the Tree of Life pipeline from a specimen collected in Wytham Woods, Oxfordshire, United Kingdom (Figure 1). It was generated as part of the Darwin Tree of Life Project, which aims to generate high-quality reference genomes for all named eukaryotic species in Britain and Ireland to support research, conservation, and the sustainable use of biodiversity (Blaxter *et al.*, 2022).

## Methods

### Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult female *Neurigona quadrifasciata* (specimen ID Ox002306, ToLID idNeuQuad1; Figure 1), collected from Wytham Woods, Oxfordshire, UK (latitude 51.772, longitude –1.338) on 2022-07-05. The specimen was collected by James McCulloch and Liam Crowley, and formally identified by James McCulloch. A second specimen was used for Hi-C sequencing (specimen ID Ox002708, ToLID idNeuQuad2). It was collected from the same location on 2022-06-14. The specimen was collected by Steven Falk and Liam Crowley, and formally identified by Steven Falk. Specimen idNeuQuad2 was also used for RNA sequencing. For the Darwin Tree of Life sampling and metadata approach, refer to Lawniczak *et al.* (2022).


**Figure 1. Photograph of the *Neurigona quadrifasciata* (idNeuQuad1) 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 each 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 idNeuQuad1 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 in the WSI Scientific Operations core using the Automated MagAttract v2 protocol. DNA was sheared using centrifuge-mediated fragmentation to produce fragments in the 8–10 kb range, following the Covaris g-TUBE protocol for ultra-low input (ULI). Sheared DNA was purified by manual 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 1.36 ng/μL and a yield of 61.20 ng, with a fragment size of 9.3 kb. The 260/280 spectrophotometric ratio was 2.53, and the 260/230 ratio was 1.83.

RNA was extracted from whole organism tissue of idNeuQuad2 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol. The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

## 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 on a Revio instrument (Pacific Biosciences). The prepared library was normalised to 2 nM, and 15 μL was used for making complexes. Primers were annealed and polymerases bound to generate circularised complexes, following the manufacturer's instructions. Complexes were purified using 1.2X SMRTbell beads, then diluted to the Revio loading concentration (200–300 pM) and spiked with a Revio sequencing internal control. The sample was sequenced on a Revio 25M SMRT cell. The SMRT Link software (Pacific Biosciences), a web-based workflow manager, was used to configure and monitor the run and to carry out primary and secondary data analysis.

## Hi-C

### *Sample preparation and crosslinking*

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the idNeuQuad2 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. 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.

## RNA library preparation and sequencing

Libraries were prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs), following the manufacturer's instructions. Poly(A) mRNA in the total RNA solution was isolated using oligo(dT) beads, converted to cDNA, and uniquely indexed; 14 PCR cycles were performed. Libraries were size-selected to produce fragments between 100–300 bp. Libraries were quantified, normalised, pooled to a final concentration of 2.8 nM, and diluted to 150 pM for loading. Sequencing was carried out on the Illumina NovaSeq 6000 to generate 150-bp paired-end reads.

## 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. 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 72 breaks and 118 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 *Neurigona quadrifasciata* specimen generated 31.17 Gb (gigabases) from 4.47 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 605.87 Mb, with a heterozygosity of 0.10% and repeat content of 18.46% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 48× coverage. Hi-C sequencing produced 91.62 Gb from 606.77 million reads, which were used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. 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 final assembly has a total length of 582.53 Mb in 99 scaffolds, with 503 gaps, and a scaffold N50 of 135.25 Mb (Table 2).

Most of the assembly sequence (99.61%) was assigned to 6 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). We did not identify the sex chromosome(s) as sequence data from the heterogametic sex was not available and homology is unreliable for sex chromosome identification in Diptera due to frequent sex chromosome turnover (Vicoso & Bachtrog, 2015). During genome curation, we observed that chromosome 4 contains a haplotypic inversion from ~32.5–34 Mbp.

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.

The combined primary and alternate assemblies achieve an estimated QV of 62.6. The *k*-mer completeness is 91.37% for the primary assembly, 91.37% for the alternate haplotype, and 91.37% for the combined assemblies (Figure 4).

BUSCO v.5.5.0 analysis using the diptera_odb10 reference set ($n$ = 3 285) identified 93.4% of the expected gene set (single = 92.2%, duplicated = 1.2%). 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) the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the primary assembly, is **6.C.Q62**, meeting the recommended reference standard.

### 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

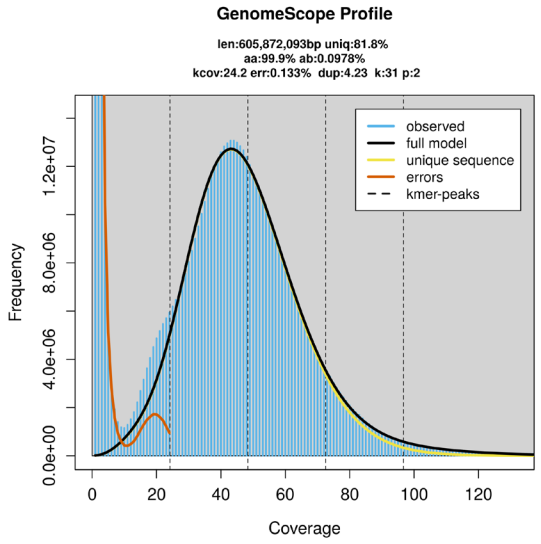


**Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2.** The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

**Table 1. Specimen and sequencing data for BioProject PRJEB78304.**

| Platform | PacBio HiFi | Hi-C | RNA-seq |
|---|---|---|---|
| **ToLID** | idNeuQuad1 | idNeuQuad2 | idNeuQuad2 |
| **Specimen ID** | Ox002306 | Ox002708 | Ox002708 |
| **BioSample (source individual)** | SAMEA112232538 | SAMEA112232875 | SAMEA112232875 |
| **BioSample (tissue)** | SAMEA112232984 | SAMEA112233383 | SAMEA112233383 |
| **Tissue** | whole organism | whole organism | whole organism |
| **Instrument** | Revio | Illumina NovaSeq 6000 | Illumina NovaSeq 6000 |
| **Run accessions** | ERR13420759 | ERR13389746 | ERR13389745 |
| **Read count total** | 4.47 million | 606.77 million | 79.45 million |
| **Base count total** | 31.17 Gb | 91.62 Gb | 12.00 Gb |

**Table 2. Genome assembly statistics.**

| | |
|---|---|
| **Assembly name** | idNeuQuad1.1 |
| **Assembly accession** | GCA_964251855.1 |
| **Alternate haplotype accession** | GCA_964248995.1 |
| **Assembly level** | chromosome |
| **Span (Mb)** | 582.53 |
| **Number of chromosomes** | 6 |
| **Number of contigs** | 602 |
| **Contig N50** | 2.42 Mb |
| **Number of scaffolds** | 99 |
| **Scaffold N50** | 135.25 Mb |
| **Sex chromosomes** | N/A |
| **Organelles** | Mitochondrion: 16.89 kb |

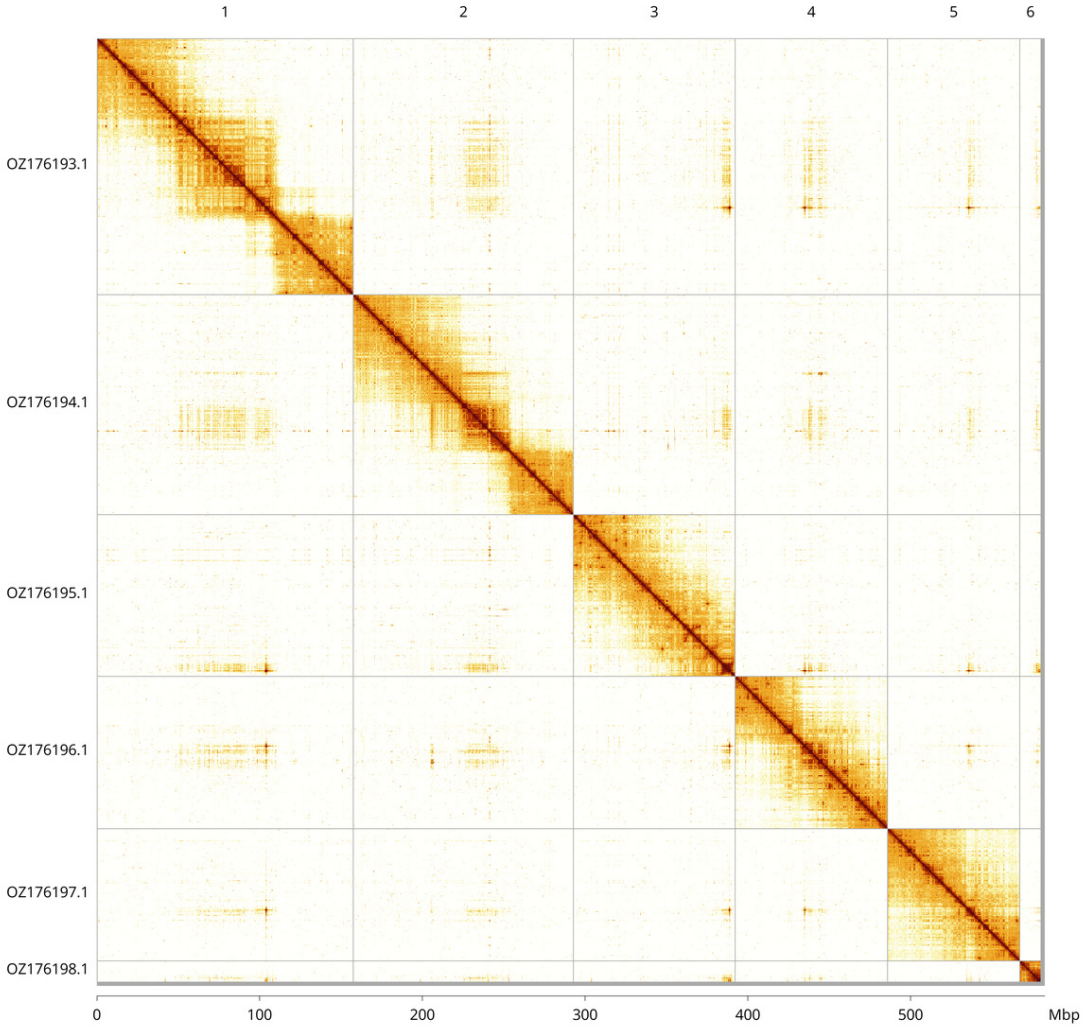


**Figure 3. Hi-C contact map of the *Neurigona quadrifasciata* genome assembly.** Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

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)

**Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Neurigona quadrifasciata* idNeuQuad1.**

| INSDC accession | Molecule | Length (Mb) | GC% |
|---|---|---|---|
| OZ176193.1 | 1 | 157.64 | 26.50 |
| OZ176194.1 | 2 | 135.25 | 26.50 |
| OZ176195.1 | 3 | 99.31 | 26.50 |
| OZ176196.1 | 4 | 93.76 | 26.50 |
| OZ176197.1 | 5 | 81.28 | 26 |
| OZ176198.1 | 6 | 13.01 | 28.50 |

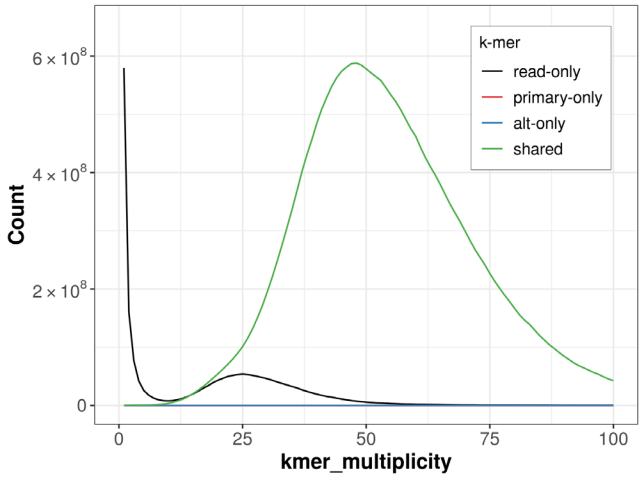


**Figure 4. Evaluation of *k*-mer completeness using MerquryFK.** This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers 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.

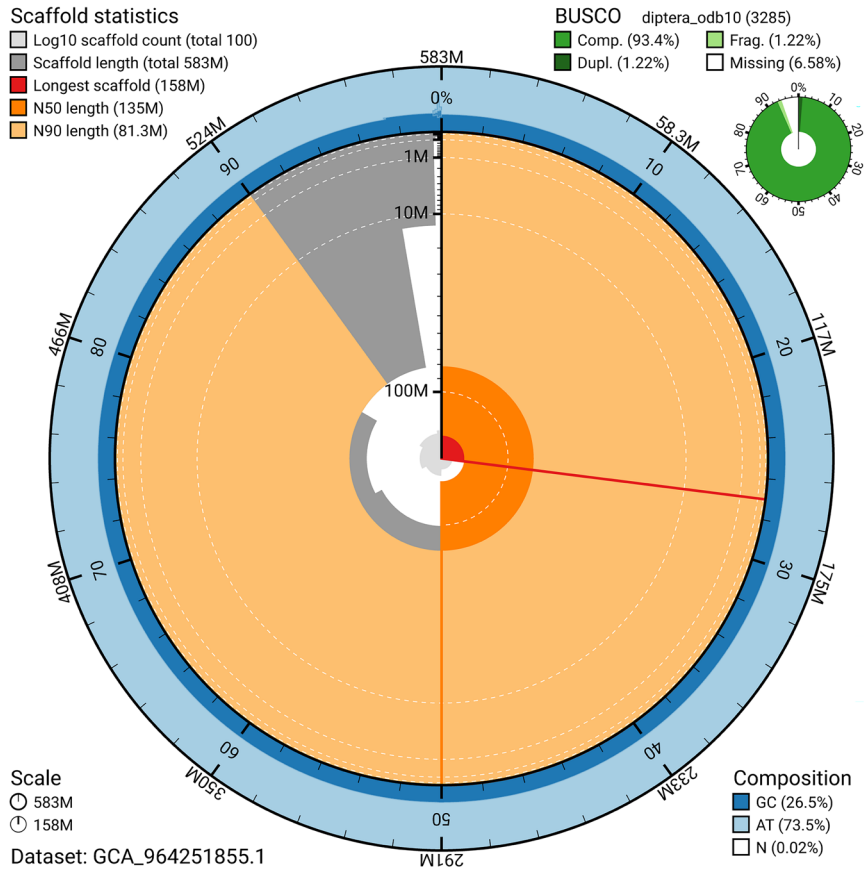


**Figure 5. Assembly metrics for idNeuQuad1.1.** The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost 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.

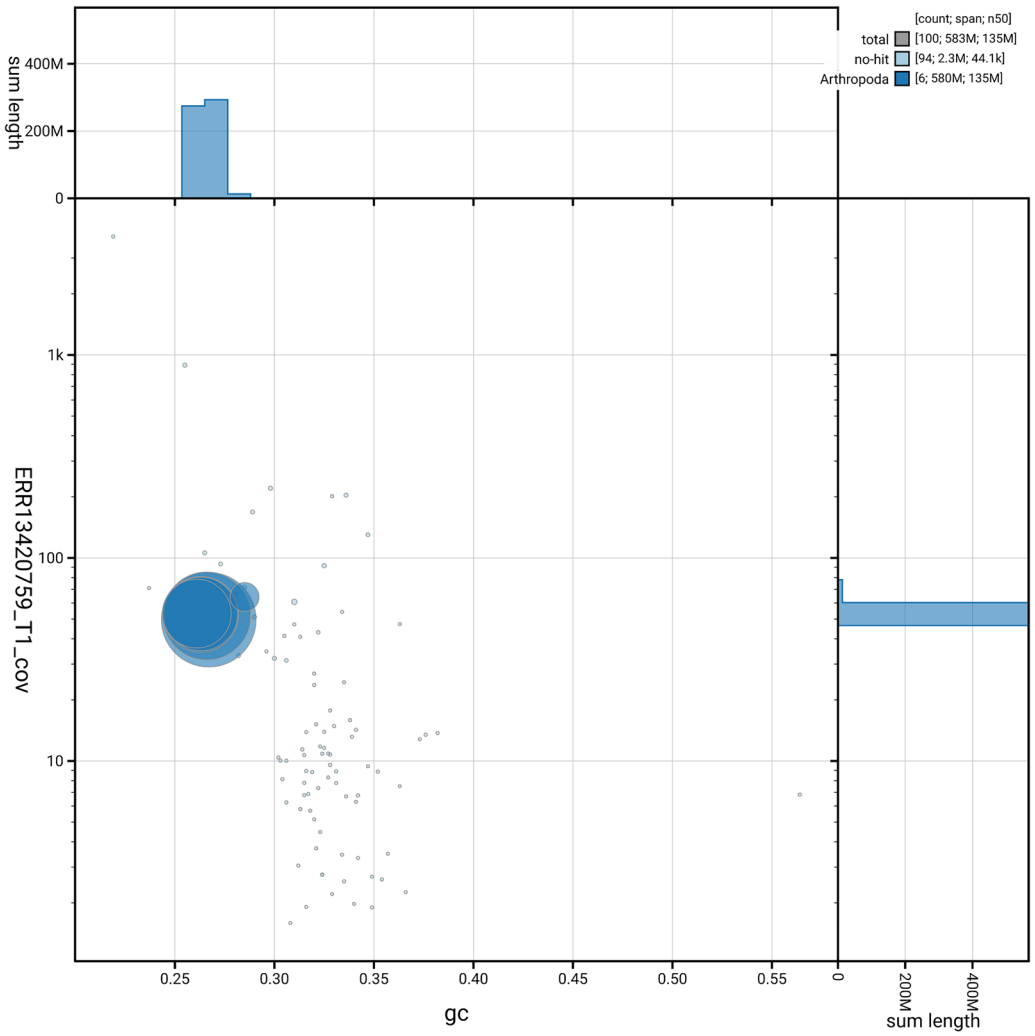


**Figure 6. BlobToolKit GC-coverage plot for idNeuQuad1.1.** Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Neurigona quadrifasciata* assembly.**

| Measure | Value | Benchmark |
|---|---|---|
| EBP summary (primary) | 6.C.Q62 | 6.C.Q40 |
| Contig N50 length | 2.42 Mb | ≥ 1 Mb |
| Scaffold N50 length | 135.25 Mb | = chromosome N50 |
| Consensus quality (QV) | Primary: 62.6; alternate: 62.6; combined: 62.6 | ≥ 40 |
| *k*-mer completeness | Primary: 91.37%; alternate: 91.37%; combined: 91.37% | ≥ 95% |
| BUSCO | C:93.4% [S:92.2%; D:1.2%]; F:1.2%; M:5.4%; n:3 285 | S > 90%; D < 5% |
| Percentage of assembly assigned to chromosomes | 99.61% | ≥ 90% |

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: Neurigona quadrifasciata. Accession number PRJEB78304. The genome sequence is released openly for reuse. The *Neurigona quadrifasciata* 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.

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 | 0.19.8-r603 | 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 | 23.10.0 | https://github.com/nextflow-io/nextflow |
| PretextSnapshot | N/A | https://github.com/sanger-tol/PretextSnapshot |

| Software | Version | Source |
|---|---|---|
| PretextView | 0.2.5 | https://github.com/sanger-tol/PretextView |
| samtools | 1.19.2 | https://github.com/samtools/samtools |
| sanger-tol/ascc | 0.1.0 | https://github.com/sanger-tol/ascc |
| sanger-tol/blobtoolkit | 0.6.0 | 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

Blaxter M, Mieszkowska N, Di Palma F, *et al.*: **Sequence locally, think globally: the Darwin Tree of Life Project.** *Proc Natl Acad Sci U S A.* 2022; **119**(4): e2115642118.
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

Fonseca ECM d'Assis: **Diptera, Orthorrhapha Brachycera: Dolichopodidae.** London: Royal Entomological Society of London, 1978; **IX**.
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

Grichanov IYa: **West-Palearctic species of the genus *Neurigona* Rondani (Diptera: Dolichopodidae).** *Russ Entomol J.* 2010; **19**(3): 249–56.
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

Insektarium: **Neurigona quadrifasciata – Szczudlarz plamkowany.** 2014.
Reference Source

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

NBN Atlas Partnership: ***Neurigona quadrifasciata* (Fabricius, 1781).** 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**

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 03 February 2026

https://doi.org/10.21956/wellcomeopenres.27509.r144108



✔ **Kristina Gagalova**
Curtin University, Perth, Australia

Thank you for submitting the manuscript "The genome sequence of a long-legged fly, *Neurigona quadrifasciata* (Fabricius, 1781) (Diptera: Dolichopodidae)"
I found the methods and results section clear. I only have a few points for clarification:
1) Is there any cytological information regarding the number of chromosomes of the species?
2) What about the sex chromosomes? Can you please state that clearly in the results?
3) In genome scope, do you assume that the genome is diploid?

**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:*** insect and plant genomics, pipeline development, omics data integration, pangenomics, structural 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 21 January 2026

https://doi.org/10.21956/wellcomeopenres.27509.r144106



**Stuart J.E. Baird**
Academy of Sciences of the Czech Republic, Národní, Staré Město, Czech Republic

The genome note is a clean example of the DToL pipeline output. I have only one comment:
In Figure 4 (Evaluation of k-mer completeness using MerquryFK) there is no evidence of the red curve for primary-only data. If it is obscured behind another curve, this should be explained. It it is absent, this should be explained.

**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:*** Admixture genomics. Spatial genetics.

**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 25 November 2025

https://doi.org/10.21956/wellcomeopenres.27509.r138603



**Dong Mei Li**
Ministry for Primary Industries, Auckland, New Zealand

This manuscript presents a chromosomal-level genome sequence for the long-legged fly, *Neurigona quadrifasciata.* In addition, a mitochondrial genome has also been assembled. The manuscript is well-written and provide valuable information for future research, as it delivers a high-quality reference genome for the species – the first reported for the genus *Neurigona*. I do not have major revisions to suggest, but I do have a few comments:

1. **Keywords:** Most of the listed keywords (e.g. *Neurigona quadrifasciata*; long-legged fly, genome sequences; Diptera) are already in the manuscript. Typically, keywords should not simply repeat words from the title but instead provide additional indexing terms that broaden discoverability. I recommend selecting alternative or complementary keywords (e.g," "mitochondrial genome," "chromosomal assembly or others).
2. **Background:** In the second paragraph, the manuscript provides detailed information on the behavior of males within the genus. However, it would be helpful to also include specific details about the behavior of *N. quadrifasciata* males, rather than only general genus-level traits. In addition, while the last sentence describes male morphology, there is no mention of female morphology. Including information on female characteristics and behavior would provide a more balanced and comprehensive background for readers.
3. **Figure 1:** The current figure would be more informative if the diagnostic features of the fly were clearly highlighted. In addition, the figure legend should indicate the specimen shown is female included the sample ID as well. This would enhance the utility of the figure for readers who may wish to compare diagnostic features across related taxa.
4. **Methods**
    - Several sections of the Methods appear to closely resemble the text from another genome sequence publication in this journal, I suggest to rephrasing these sentences. One example is the second paragraph in the "Sample acquisition and DNA barcoding" section.
    - The Qubit Fluorometer is referred to in three different ways: "Qubit Fluorometer," "Qubit Fluorometer (Thermo Fisher Scientific)," and *"Qubit Fluorometer V4 (Thermo Fisher Scientific)."* For consistency and clarity, the authors should standardize the terminology. It is recommended to use Qubit Fluorometer V4, (Thermo Fisher Scientific) when first introduced, and then refer to it consistently as Qubit Fluorometer throughout the remainder of the manuscript.
    - When Qubit dsDNA High Sensitivity Assay kit first mentioned in the Nucleic acid extraction section, the abbreviation HS should be added immediately after "High Sensitivity". Thereafter, the kit should be referred to as Qubit HS Assay kit.
5. **Methods – Sample Acquisition and DNA Barcoding**:
    - The manuscript does not provide details on how the two specimens were collected. Please specify the collection methods used (e.g., sweep netting, Malaise trap, hand collection, or other techniques).
    - In the 1$^{st}$ paragraph, the manuscript mentions that a second specimen was used for Hi-C sequencing and RNA-Seq. However, unlike the first specimen, the sex of this individual is not specified. Please indicate whether the second specimen was male or female in the Methods section, if relevant, discuss any implications for the results.
    - In the second paragraph of the *"Sample acquisition and DNA barcoding"* section, additional details should be provided for clarity. For example, the phrase "a small sample was dissected from each specimen" is vague; the authors should specify which tissue was used (e.g., leg, thoracic muscle, or abdominal tissue).
    - It is also unclear whether tissue lysis, PCR, and sequencing were performed at the Wellcome Sanger Institute or in the authors' laboratory. If these steps were not

conducted at WSI, please describe the methods used for DNA extraction, PCR amplification, and sequencing.
    - It would be useful to clarify whether identical COI sequences were obtained from both specimens, as this information supports species verification and consistency across samples.
6. **Methods – Nucleic Acid Extraction**
    - In this section, the abbreviation **WSI** should be used instead of spelling out "Wellcome Sanger Institute," since the abbreviation has already been introduced.
    - More detail is needed regarding the RNA extraction process. Please specify which tissue was used and the approximate amount of tissue (in mg) employed for RNA extraction. Providing these details will improve reproducibility and allow readers to better assess the quality and reliability of the RNA-Seq data.
7. **Methods – PacBio HiFi library preparation bd sequencing**
    - There is an inconsistency in reported DNA fragment size: 9.3 kb is mentioned near the end of page 3, while ~10 kb is noted on page 4. Please ensure consistency throughout the section.
    - The statement "A 0.85x pre-PCR clean-up was carried out with Promega ProNex beads" appears twice in consecutive paragraphs. This repetition should be avoided by consolidating the information into a single clear description.
    - The rationale for dividing the DNA into two aliquots for dual PCR is not explained. Please provide justification for this step, as it will help readers understand the methodological choices made.
8. **Methods – Hi-C:** The manuscript states that 20–50 mg of tissue was used. Since only a single specimen was processed, the exact amount of tissue should be reported rather than a range. In addition, please specify which part of the insect was used for Hi-C sequencing.
9. **Table 1.** It appears that not the whole organism was used for sequencing, since tissue was removed for DNA barcoding.
10. **Assembly statistics:** the manuscript notes that mitochondrial genome was assembled, and its length is mentioned in the Abstract. However, no details are provided in this section about the genome structure or gene annotation. Some details will assist readers to get better understanding of the mitochondrial genome.

**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:*** Molecular biology and DNA/RNA sequencing methods, genomics,

bioinformatics, Molecular Diagnostics, Mitochondrial genome, Phylogeny, Real-time PCR assay development.

**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.**