DATA NOTE

# The genome sequence of a cranefly, *Tipula lateralis* Meigen, 1804

[version 1; peer review: 2 approved]

Duncan Sivell [1], Olga Sivell [1],
Natural History Museum Genome Acquisition Lab,
Darwin Tree of Life Barcoding collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]Natural History Museum, London, England, UK

**v1** **First published:** 19 Mar 2025, **10**:149
https://doi.org/10.12688/wellcomeopenres.23899.1
**Latest published:** 19 Mar 2025, **10**:149
https://doi.org/10.12688/wellcomeopenres.23899.1

## Abstract

We present a genome assembly from a female specimen of *Tipula lateralis* (cranefly; Arthropoda; Insecta; Diptera; Tipulidae). The genome sequence has a total length of 701.32 megabases. Most of the assembly (89.15%) is scaffolded into 4 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 16.5 kilobases. Gene annotation of this assembly on Ensembl identified 11,388 protein-coding genes.

## Keywords

Tipula lateralis, cranefly, genome sequence, chromosomal, Diptera

This article is included in the Tree of Life gateway.

**Open Peer Review**

**Approval Status** ✓ ✓

| | 1 | 2 |
|---|---|---|
| **version 1** 19 Mar 2025 | ✓ view | ✓ view |

1. **Marcus Stensmyr**, Lund University, Lund, Sweden

2. **Markus Friedrich**, Wayne State University, Michigan, 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 D**: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Sivell O**: Investigation, Resources;

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

**Grant information:** This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540) and the Darwin Tree of Life Discretionary Award [218328, https://doi.org/10.35802/218328 ].
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*



**How to cite this article:** Sivell D, Sivell O, Natural History Museum Genome Acquisition Lab *et al.* **The genome sequence of a cranefly, *Tipula lateralis* Meigen, 1804 [version 1; peer review: 2 approved]** Wellcome Open Research 2025, **10**:149 https://doi.org/10.12688/wellcomeopenres.23899.1

**First published:** 19 Mar 2025, **10**:149 https://doi.org/10.12688/wellcomeopenres.23899.1

## Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Nematocera; Tipulomorpha; Tipuloidea; Tipulidae; Tipulinae; *Tipula*; *Yamatotipula*; *Tipula lateralis* Meigen, 1804 (NCBI:txid2025111)

## Background

*Tipula lateralis* is a western Palaearctic species that is widespread across Europe reaching the Urals and the Levant in the east (GBIF Secretariat, 2024; Oosterbroek & Theowald, 1992). In Britain this species is also widespread ranging from the south coast of England to Shetland (Stubbs, 1992; Stubbs, 2021) and is one of the most common large craneflies to be found along riverbanks and by lake margins (Boardman, 2016).

*Tipula lateralis* is a relatively slender *Tipula* species with streaked wings and two broad dark lateral stripes on a pale grey abdomen. Other British *Yamatotipula* species and the much larger *Tipula* (*Acutipula*) *vittata* also have abdomens with dark lateral stripes. *Tipula lateralis* males can be identified by the shape of the outer clasper and tergite 9 (Boardman, 2016; Stubbs, 1973), while details of wing shading can be used to identify the females (Stubbs, 2021). The associated habitat may also be useful when identifying this species.

*Tipula lateralis* larvae are aquatic living in wet sediments, among vegetable debris at water margins, in seepages, ponds, ditches and other wet environments, only avoiding very acid or oligotrophic sites (Stubbs, 2021). Adults are usually found by wet habitats as a result and tend to avoid shaded environments (Stubbs, 1992). Dufour (1986) has caught adults some distance away from water bodies in Switzerland, but he interprets this as the adults having good dispersal ability. In Britain there are two generations per year with adult activity peaking in April to May and in September (Stubbs, 1972; Stubbs, 1973; Stubbs, 1992).

White (1951) made a detailed study of *Tipula lateralis* and its life cycle, prompted by infestations of watercress beds in Lincolnshire. This study showed, however, that the larvae were feeding on decaying vegetable matter and were not damaging the living crop. Larvae and pupae of *T. lateralis* have also been keyed and illustrated by Brindle (1960). Common names proposed for this species include the "common lined Tipula" (Stubbs, 1973) and the "common yam" (Stubbs, 2021).

It had long been recognised that across its range *Tipula lateralis* might contain more than one species (Lackschewitz, 1923). *Tipula barbarensis* Theowald & Oosterbroek, 1980 (Morocco), *Tipula iranensis* Theowald, 1978 (Iran, Turkey, Russia) and *Tipula intermedia* Eiroa, 1990 (Spain, Portugal) are former subspecies of *T. lateralis* that have been promoted to species level (Oosterbroek, 1994; Oosterbroek *et al.*, 2020). Stubbs (2021) considers it may be possible there is more than one taxon within the British *Tipula lateralis* population.

*Tipula intermedia* shares the same BIN as *T. lateralis* (Ferreira *et al.*, 2021), hence they cannot be reliably separated using COI barcodes and morphological characters provided by Oosterbroek *et al.* (2020) should be consulted for identification. To date *T. intermedia* has only been recorded from the Iberian Peninsula. This genome sequence of *Tipula lateralis* will aid research on the taxonomy of this and closely related species and help to better understand the status of *T. intermedia* as well as support research into the genetic identification of these species.

The high-quality genome of *Tipula lateralis* presented here was sequenced from a single specimen (NHMUK014446897, SAMEA9654302) from Cothill Fen National Nature Reserve, England, collected by Olga Sivell and Duncan Sivell (Figure 1). It was identified by Duncan Sivell following Stubbs (2021) and Boardman (2016). 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 *Tipula lateralis* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 30.29 Gb from 2.42 million reads. GenomeScope analysis of the PacBio HiFi data estimated the haploid genome size at 572.99 Mb, with a heterozygosity of 2.66% and repeat content of 34.62%. These values provide an initial assessment of genome complexity and the challenges anticipated during assembly. Based on this estimated genome size, the sequencing data provided approximately 48.0x coverage of the genome. Chromosome conformation Hi-C sequencing produced 239.47 Gb from 1,585.87 million reads. Table 1 summarises the specimen and sequencing information.

### Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC


**Figure 1. Photograph of the *Tipula lateralis* (idTipLate1) specimen used for genome sequencing.**

**Table 1. Specimen and sequencing data for *Tipula lateralis*.**

| **Project information** | | | |
|---|---|---|---|
| **Study title** | Tipula lateralis | | |
| **Umbrella BioProject** | PRJEB66744 | | |
| **Species** | *Tipula lateralis* | | |
| **BioSpecimen** | SAMEA9654302 | | |
| **NCBI taxonomy ID** | 2025111 | | |
| **Specimen information** | | | |
| **Technology** | **ToLID** | **BioSample accession** | **Organism part** |
| **PacBio long read sequencing** | idTipLate1 | SAMEA9654386 | abdomen |
| **Hi-C sequencing** | idTipLate1 | SAMEA9654388 | head and thorax |
| **Sequencing information** | | | |
| **Platform** | **Run accession** | **Read count** | **Base count (Gb)** |
| **Hi-C Illumina NovaSeq 6000** | ERR12102417 | 1.59e+09 | 239.47 |
| **PacBio Sequel IIe** | ERR12120135 | 2.42e+06 | 30.29 |

databases. The assembly was improved by manual curation, which corrected 78 misjoins or missing joins and removed 12 haplotypic duplications. These interventions reduced the total assembly length by 0.77%, decreased the scaffold count by 4.95%, and increased the scaffold N50 by 1.58%. The final assembly has a total length of 701.32 Mb in 671 scaffolds, with 377 gaps, and a scaffold N50 of 198.68 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 (88.95%) was assigned to 4 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3). The specimen is a homogametic female. We did not identify the sex chromosome(s) as sequence data from the heterogametic sex was not available and homology is unreliable for sex chromosome identification in Diptera due to frequent sex chromosome turnover (Vicoso & Bachtrog, 2015).

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

### Assembly quality metrics

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

The primary haplotype has a QV of 55.7, and the combined primary and alternate assemblies achieve an estimated QV of 56.1. The *k*-mer recovery for the primary haplotype is 65.27%, and for the alternate haplotype it is 64.54%. The combined primary and alternate assemblies display a *k*-mer recovery of 98.22%. BUSCO analysis using the diptera_odb10 reference set ($n$ = 3,285) identified 95.0% of the expected gene set (single = 93.7%, duplicated = 1.3%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project (EBP) Report on Assembly Standards September 2024. The assembly achieves the EBP reference standard of **6.8.Q55.**

## Genome annotation report

The *Tipula lateralis* genome assembly (GCA_963932295.1) was annotated externally by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 19,373 transcribed mRNAs from 11,388 protein-coding and 3,109 non-coding genes. The average transcript length is 8,869.40. There are 1.34 coding transcripts per gene and 3.72 exons per transcript. For further information about the annotation, please refer to https://beta.ensembl.org/species/7d6679ea-b5a9-4dc9-9e3e-0c4a45bc7441.

## Methods

### Sample acquisition and DNA barcoding

An adult female *Tipula lateralis* (specimen ID NHMUK014446897, ToLID idTipLate1) was collected from

**Table 2. Genome assembly data for *Tipula lateralis*.**

| **Genome assembly** | | |
|---|---|---|
| Assembly name | idTipLate1.1 | |
| Assembly accession | GCA_963932295.1 | |
| *Alternate haplotype accession* | *GCA_963932235.1* | |
| Assembly level for primary assembly | chromosome | |
| Span (Mb) | 701.32 | |
| Number of contigs | 1,048 | |
| Number of scaffolds | 671 | |
| Longest scaffold (Mb) | 200.01 | |
| **Assembly metric** | **Measure** | ***Benchmark*** |
| Contig N50 length | 3.51 Mb | *≥ 1 Mb* |
| Scaffold N50 length | 198.68 Mb | *= chromosome N50* |
| Consensus quality (QV) | Primary: 55.7; alternate: 56.5; combined 56.1 | *≥ 40* |
| *k*-mer completeness | Primary: 65.27%; alternate: 64.54%; combined: 98.22% | *≥ 95%* |
| BUSCO* | C:95.0%[S:93.7%,D:1.3%], F:0.8%,M:4.2%,n:3,285 | *S > 90%; D < 5%* |
| Percentage of assembly mapped to chromosomes | 88.95% | *≥ 90%* |
| Sex chromosomes | Not identified | *localised homologous pairs* |
| Organelles | Mitochondrial genome: 16.5 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.

Cothill Fen National Nature Reserve, England, United Kingdom (latitude 51.69, longitude –1.34), using an aerial net. The specimen was collected by Duncan Sivell and Olga Sivell (Natural History Museum), identified by Duncan Sivell and preserved by dry freezing (–80 °C).

The initial identification by morphology 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. Detailed protocols are available on protocols.io (Denton *et al.*, 2023b).

The idTipLate1 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the abdomen was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a). HMW DNA was extracted using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023a). DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Bates *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Oatley *et al.*, 2023b). The

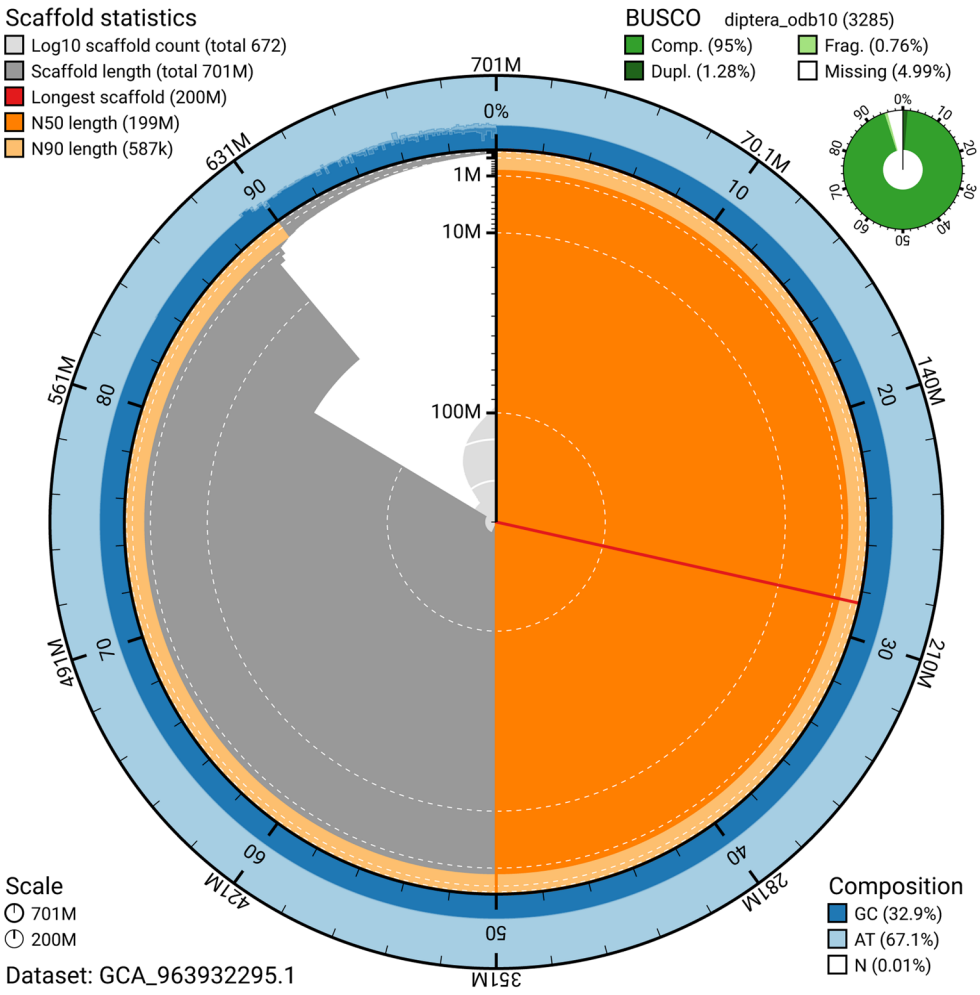


**Figure 2. Genome assembly of *Tipula lateralis*, idTipLate1.1: metrics.** The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. 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_963932295.1/dataset/GCA_963932295.1/snail.

concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

## Hi-C sample preparation and crosslinking

Tissue from the head and thorax of the idTipLate1 sample was processed for Hi-C sequencing at the WSI Scientific Operations core, using the Arima-HiC v2 kit. In brief, 20–50 mg of frozen tissue (stored at –80 °C) was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde concentration. After crosslinking, the tissue was homogenised using the Diagnocine Power Masher-II and BioMasher-II tubes and pestles. Following the Arima-HiC v2 kit manufacturer's instructions, crosslinked DNA was digested using a restriction enzyme master mix. The 5'-overhangs were filled in and labelled with biotinylated nucleotides and proximally ligated. An overnight incubation was carried out for enzymes to digest remaining proteins and for crosslinks to reverse. A clean up was performed with SPRIselect beads prior to library preparation. Additionally, the biotinylation percentage was estimated using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit and Arima-HiC v2 QC beads.

## Library preparation and sequencing

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

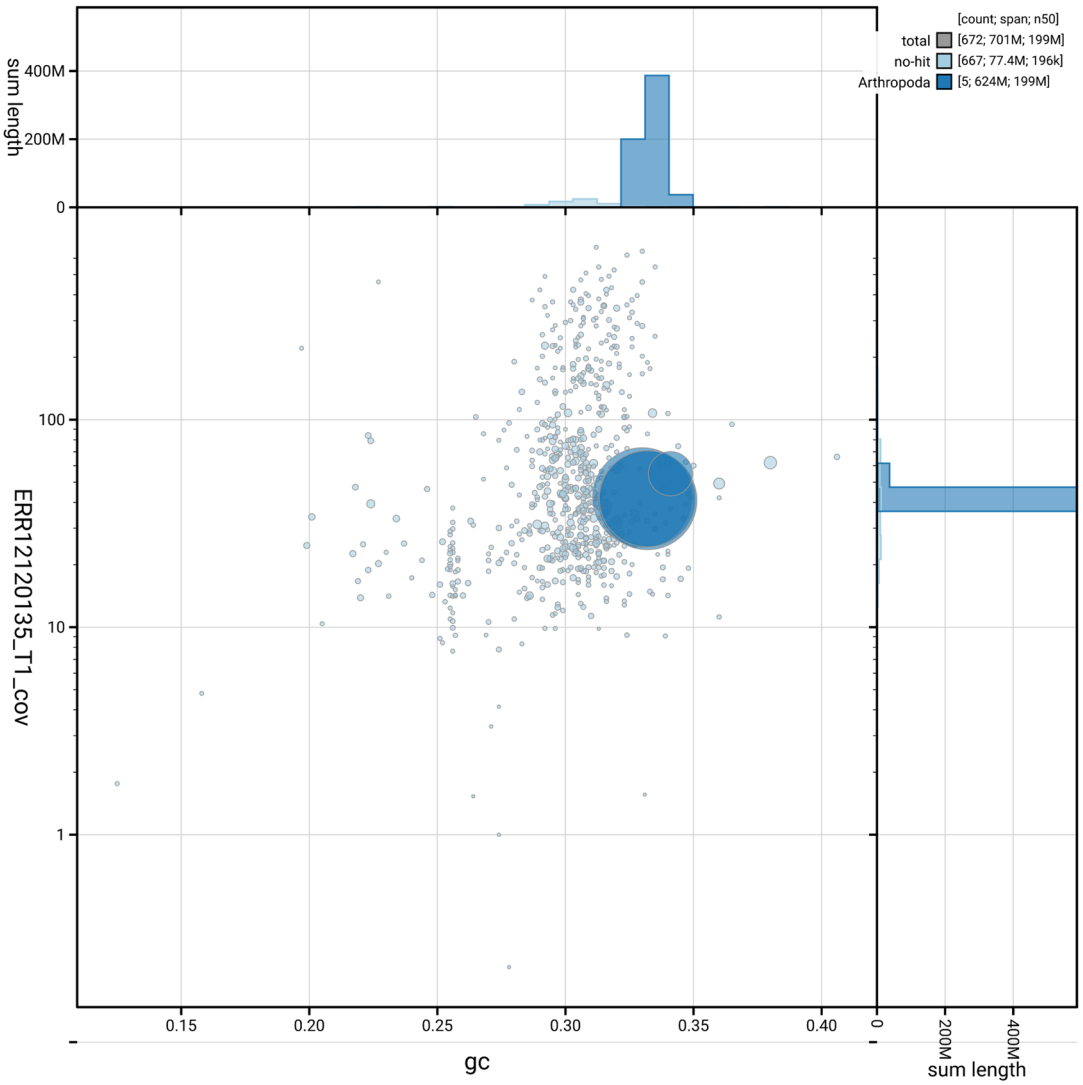


**Figure 3. Genome assembly of *Tipula lateralis*, idTipLate1.1: BlobToolKit GC-coverage plot.** Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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_963932295.1/blob.

***PacBio HiFi***

At a minimum, samples were required to have an average fragment size exceeding 8 kb and a total mass over 400 ng to proceed to the low input SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences, California, USA), depending on genome size and sequencing depth required. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA) as per the manufacturer's instructions. The kit includes the reagents required for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead cleanup, and nuclease treatment. Following the manufacturer's instructions, size selection and clean up was carried out using diluted AMPure PB beads (Pacific Biosciences, California, USA). DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) with Qubit 1X dsDNA HS assay kit and the final library fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and gDNA 55kb BAC analysis kit.

Samples were sequenced 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

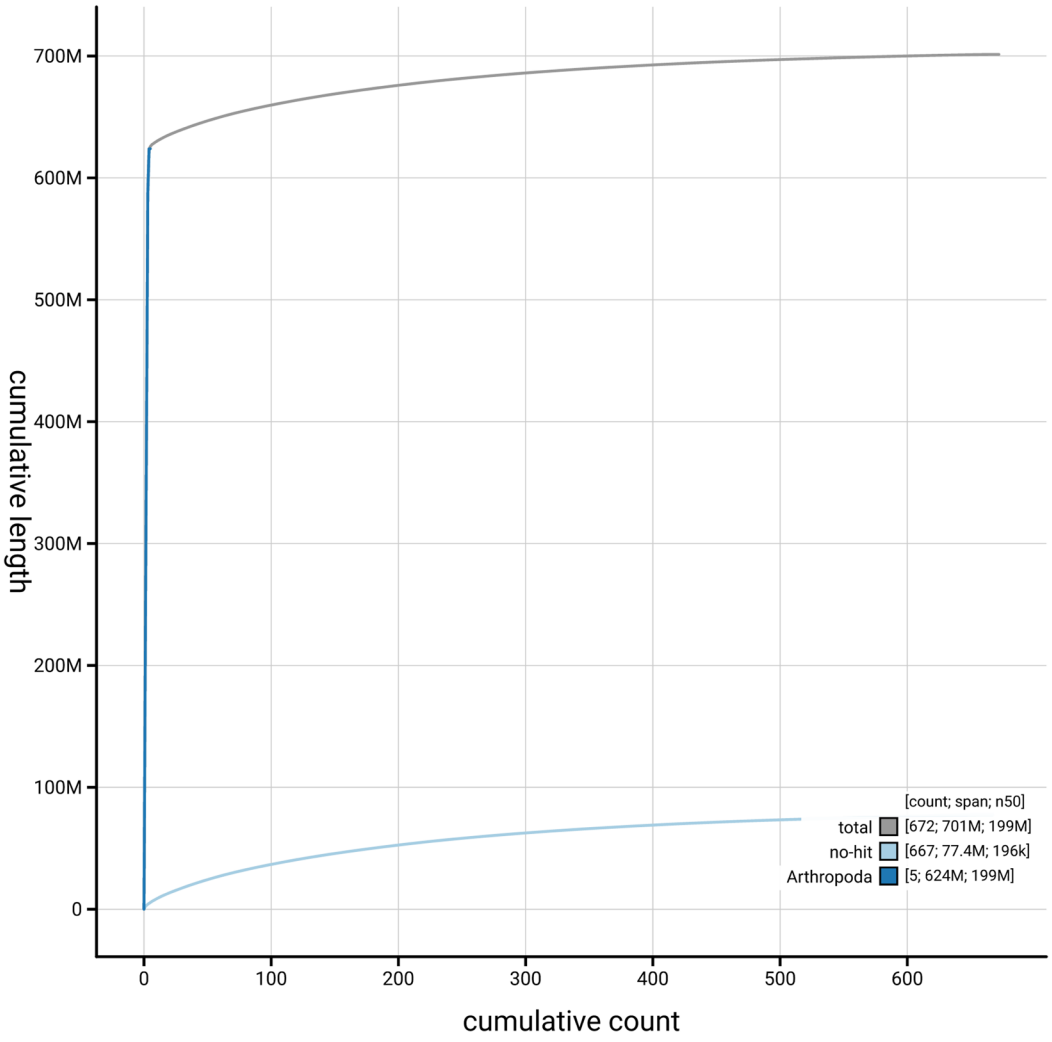


**Figure 4. Genome assembly of Tipula lateralis, idTipLate1.1: BlobToolKit cumulative sequence plot.** The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963932295.1/dataset/GCA_963932295.1/cumulative.

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

***Hi-C***

For Hi-C library preparation, DNA was fragmented using the Covaris E220 sonicator (Covaris) and size selected using SPRISelect beads to 400 to 600 bp. The DNA was then enriched using the Arima-HiC v2 kit Enrichment beads. Using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) for end repair, A-tailing, and adapter ligation. This uses a custom protocol which resembles the standard NEBNext Ultra II DNA Library Prep protocol but where library preparation occurs while DNA is bound to the Enrichment beads. For library amplification, 10 to 16 PCR cycles were required, determined by the sample biotinylation percentage. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on an Illumina NovaSeq 6000 instrument.

## Genome assembly, curation and evaluation

***Assembly***

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 were mapped to the primary contigs

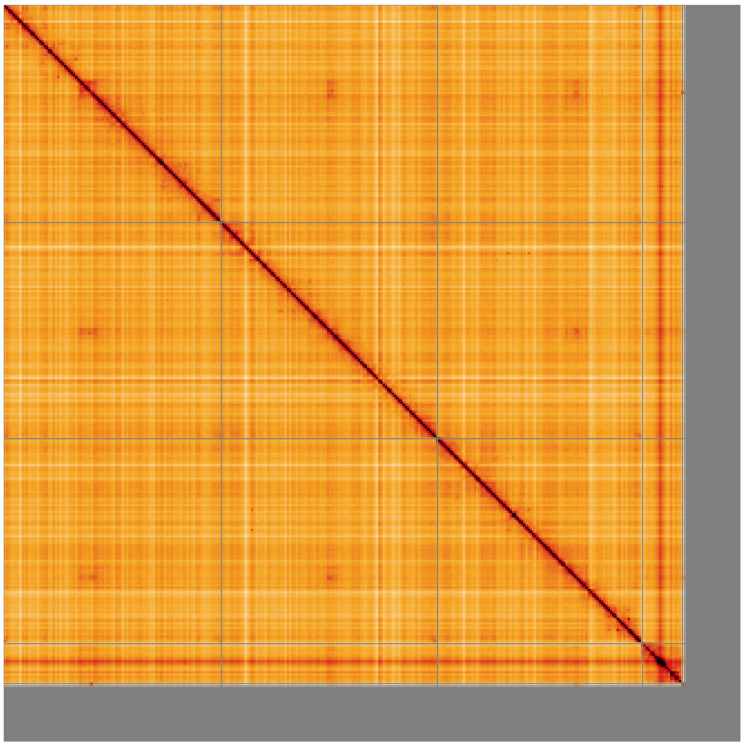

**Figure 5. Genome assembly of *Tipula lateralis:* Hi-C contact map of the idTipLate1.1 assembly, visualised using HiGlass.** Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=ChaB9SOWRM-2_lDw95dFIw.

**Table 3. Chromosomal pseudomolecules in the genome assembly of *Tipula lateralis*, idTipLate1.**

| INSDC accession | Name | Length (Mb) | GC% |
|---|---|---|---|
| OZ010736.1 | 1 | 200.01 | 33 |
| OZ010737.1 | 2 | 198.68 | 33 |
| OZ010738.1 | 3 | 188.28 | 33 |
| OZ010739.1 | 4 | 36.84 | 34 |
| OZ010740.1 | MT | 0.02 | 23 |

using bwa-mem2 (Vasimuddin *et al.*, 2019). The contigs were further scaffolded using the provided Hi-C data (Rao *et al.*, 2014) in YaHS (Zhou *et al.*, 2023) using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

***Assembly curation***

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

***Assembly quality assessment***

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

A Hi-C contact map was produced for the final version of the assembly. The Hi-C reads were aligned using bwa-mem2 (Vasimuddin *et al.*, 2019) and the alignment files were combined using SAMtools (Danecek *et al.*, 2021). The Hi-C

alignments were converted into a contact map using BEDTools (Quinlan & Hall, 2010) and the Cooler tool suite (Abdennur & Mirny, 2020). The contact map was visualised in HiGlass (Kerpedjiev *et al.*, 2018).

The blobtoolkit pipeline is a Nextflow (Di Tommaso *et al.*, 2017) port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads in SAMtools and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the 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

**Table 4. Software tools: versions and sources.**

| Software tool | Version | Source |
|---|---|---|
| BEDTools | 2.30.0 | https://github.com/arq5x/bedtools2 |
| BLAST | 2.14.0 | ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/ |
| BlobToolKit | 4.3.9 | https://github.com/blobtoolkit/blobtoolkit |
| BUSCO | 5.5.0 | https://gitlab.com/ezlab/busco |
| bwa-mem2 | 2.2.1 | https://github.com/bwa-mem2/bwa-mem2 |
| Cooler | 0.8.11 | https://github.com/open2c/cooler |
| DIAMOND | 2.1.8 | https://github.com/bbuchfink/diamond |
| fasta_windows | 0.2.4 | https://github.com/tolkit/fasta_windows |
| FastK | 666652151335353eef2fcd58880bcef5bc2928e1 | https://github.com/thegenemyers/FASTK |
| Gfastats | 1.3.6 | https://github.com/vgl-hub/gfastats |
| GoaT CLI | 0.2.5 | https://github.com/genomehubs/goat-cli |
| Hifiasm | 0.19.5-r587 | https://github.com/chhylp123/hifiasm |
| HiGlass | 44086069ee7d4d3f6f3f0012569789ec138f42b84 aa44357826c0b6753eb28de | https://github.com/higlass/higlass |
| MerquryFK | d00d98157618f4e8d1a9190026b19b471055b22e | https://github.com/thegenemyers/MERQURY.FK |
| Minimap2 | 2.24-r1122 | https://github.com/lh3/minimap2 |
| MitoHiFi | 3 | https://github.com/marcelauliano/MitoHiFi |
| MultiQC | 1.14, 1.17, and 1.18 | https://github.com/MultiQC/MultiQC |
| Nextflow | 23.04.1 | https://github.com/nextflow-io/nextflow |
| PretextView | 0.2.5 | https://github.com/sanger-tol/PretextView |
| purge_dups | 1.2.5 | https://github.com/dfguan/purge_dups |
| samtools | 1.19.2 | https://github.com/samtools/samtools |
| sanger-tol/ascc | - | https://github.com/sanger-tol/ascc |
| sanger-tol/blobtoolkit | 0.5.1 | https://github.com/sanger-tol/blobtoolkit |
| Seqtk | 1.3 | https://github.com/lh3/seqtk |
| Singularity | 3.9.0 | https://github.com/sylabs/singularity |
| TreeVal | 1.2.0 | https://github.com/sanger-tol/treeval |
| YaHS | 1.2a.2 | https://github.com/c-zhou/yahs |

of materials by a Darwin Tree of Life Partner is subject to the **'Darwin Tree of Life Project Sampling Code of Practice'**, which can be found in full on the Darwin Tree of Life website here. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project.

Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible. The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international)

Each transfer of samples is further undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Darwin Tree of Life Partner, Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances other Darwin Tree of Life collaborators.

## Data availability

European Nucleotide Archive: Tipula lateralis. Accession number PRJEB66744; https://identifiers.org/ena.embl/PRJEB66744. The genome sequence is released openly for reuse. The *Tipula lateralis* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project. All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

## Author information

Members of the Natural History Museum Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.12159242.

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

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

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


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

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

Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text

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

Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor®3 for LI PacBio.** *protocols.io.* 2023.
Publisher Full Text

Beasley J, Uhl R, Forrest LL, *et al.*: **DNA barcoding SOPs for the Darwin Tree of Life project.** *protocols.io.* 2023; [Accessed 25 June 2024].
Publisher Full Text

Boardman A: **Shropshire craneflies.** Field Studies Council Publications, 2016.
Reference Source

Brindle A: **The larvae and pupae of the British Tipulinae (Diptera: Tipulidae).** *Transactions of the Society of British Entomology.* 1960; **14**(3): 63–114.

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**

Dufour C: **Documenta Faunistica Helvetiae, 2. Les Tipulidae de Suisse (Diptera, Nematocera).** Neuchâtel: Centre Suisse de Cartographie de la Faune, 1986.
**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**

Ferreira S, Oosterbroek P, Stary J, *et al.*: **The InBIO Barcoding Initiative database: DNA barcodes of Portuguese *diptera* 02 - *Limoniidae, Pediciidae* and *Tipulidae*.** *Biodivers Data J.* 2021; **9**: e69841.
**PubMed Abstract** | **Publisher Full Text** | **Free 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: ***Tipula lateralis* Meigen, 1804.** In: *GBIF Backbone Taxonomy.* 2024; [Accessed 17 February 2025].
**Reference Source**

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

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

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

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

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

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

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

Lackschewitz P: **Der Formenkreis der Tipula lateralis Meig. (s. lat) im Ostbaltischen Gebiet.** *Arbeiten Des Naturforscher-Vereins Zu Riga.* 1923; **15**: 1–16.
**Reference Source**

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, Howard C: **Sanger Tree of Life fragmented DNA clean up: automated SPRI.** *protocols.io.* 2023b.
**Publisher Full Text**

Oosterbroek P: **Notes on western Palaearctic species of the *Tipula (Yamatotipula) lateralis* group, with the description of a new species from Turkey (Diptera: Tipulidae).** *Eur J Entomol.* 1994; **91**(4): 429–435.
**Reference Source**

Oosterbroek P, Starý J, Andrade R, *et al.*: **The Craneflies of continental Portugal (Diptera, Limoniidae, Pediciidae, Tipulidae) including 28 species new for Portugal.** *Bol Asoc Esp Entomol.* 2020; **44**(3–4): 317–358.
**Reference Source**

Oosterbroek P, Theowald B: **Family Tipulidae.** In: Soós, Á., Papp, L., and Oosterbroek, P. (eds.) *Catalogue of palaearctic diptera.* Elsevier, 1992; **1**: 56–78.

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

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

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

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

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

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

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

Stubbs AE: **Introduction to craneflies.** *The Identification of British Craneflies. AES Bulletin.* 1972; **31**: 83–93.

Stubbs AE: **Introduction to craneflies – Part III, spring and summer *Tipula* species.** *AES Bulletin.* 1973; **32**: 14–23.

Stubbs AE: **Provisional atlas of long-palped craneflies (Diptera: Tipulinae) of Britain and Ireland.** Biological Records Centre, Monks Wood, 1992.
**Reference Source**

Stubbs AE: **British craneflies.** British Entomological and Natural History Society, 2021.
**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**

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**

White JH: **Observations on the life history and biology of *Tipula lateralis* Meig.** *Ann Appl Biol.* 1951; **38**(4): 847–858.
**Publisher Full Text**

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

# Open Peer Review

## Current Peer Review Status: ✔ ✔

**Version 1**

Reviewer Report 02 May 2025

https://doi.org/10.21956/wellcomeopenres.26366.r121603



✔ **Markus Friedrich**
Wayne State University, Michigan, USA

This report describes the chromosome-level sequencing of a female specimen of the cranefly species Tipula lateralis (Meigen, 1804) for the Darwin Tree of Life initiative. Data generation, processing, and curation are state of the art quality.

**Comments:**
1. "Tipula intermedia shares the same BIN..." Define BIN
2. Review inconsistent use of Tipula lateralis vs T. lateralis throughout major text body.
3. Define INSDC
4. Fig. 4: buscogenes = BUSCO genes
5. The average transcript length is 8,869.40 ribonucleotides?
6. "At a minimum, samples were required to have an average fragment size exceeding 8 kb and a total mass over 400 ng to proceed to the low input SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences, California, USA), depending on genome size and sequencing depth required." Only one sample (abdomen of idTipLate1) was subjected to PacBio HiFi sequencing.

**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:*** Comparative genomics

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

Reviewer Report 02 April 2025

https://doi.org/10.21956/wellcomeopenres.26366.r121600



**Marcus Stensmyr**
Lund University, Lund, Sweden

The submitted manuscript presents a high quality genome assembly of a cranefly (Tipula lateralis), sequenced as part of the Darwin Tree of Life project. The genome has been sequenced, assembled, and annotated using the state-of-the-art pipeline employed by the DToL project. Kudos to the authors for adding a nice and informative introduction, which these types of papers typically lack (at least the ones I have reviewed...). The provision of an additional crane fly genome assembly will aid efforts understanding the evolutionary history of this fun group of dipterans.

**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:*** Diptera neurogenetics and 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.**