DATA NOTE

The genome sequence of the spotted cranefly, *Nephrotoma appendiculata* (Pierre, 1919) [version 1; peer review: 2 approved]

Liam M. Crowley[1], Denise C. Wawman[1],
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]Department of Biology, University of Oxford, Oxford, England, UK

v1 **First published:** 15 Feb 2024, **9**:38
https://doi.org/10.12688/wellcomeopenres.20886.1
Latest published: 15 Feb 2024, **9**:38
https://doi.org/10.12688/wellcomeopenres.20886.1

Abstract

We present a genome assembly from an individual male *Nephrotoma appendiculata* (the spotted cranefly; Arthropoda; Insecta; Diptera; Tipulidae). The genome sequence is 1,138.0 megabases in span. Most of the assembly is scaffolded into 4 chromosomal pseudomolecules, including the X sex chromosome. The mitochondrial genome has also been assembled and is 17.42 kilobases in length. Gene annotation of this assembly on Ensembl identified 17,753 protein coding genes.

Keywords

Nephrotoma appendiculata, spotted cranefly, genome sequence, chromosomal, Diptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 15 Feb 2024	✓ view	✓ view

1. **Adam James Reid**, University of Cambridge, Cambridge, UK

2. **Hans-Peter Fuehrer**, University of Veterinary Medicine Vienna, Vienna, Austria

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, Writing – Review & Editing; **Wawman DC**: Writing – Original Draft Preparation;

Competing interests: No competing interests were disclosed.

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

How to cite this article: Crowley LM, Wawman DC, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the spotted cranefly, *Nephrotoma appendiculata* (Pierre, 1919) [version 1; peer review: 2 approved]** Wellcome Open Research 2024, **9**:38 https://doi.org/10.12688/wellcomeopenres.20886.1

First published: 15 Feb 2024, **9**:38 https://doi.org/10.12688/wellcomeopenres.20886.1

Species taxonomy

Eukaryota; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Nematocera; Tipulomorpha; Tipuloidea; Tipulidae; Tipulinae; *Nephrotoma*, *Nephrotoma appendiculata* (Pierre, 1919) (NCBI:txid2741127).

Background

The Spotted or Inverted-U Tiger Cranefly *Nephrotoma appendiculata* is a member of the family Tipulidae, or long-palped craneflies, and as such, it has the fairly typical "Daddy Long Legs" shape of these Diptera. Those in the genus *Nephrotoma* have black stipes on a yellow background, earning them the name tiger craneflies, and a distinctive pattern of wing venation that distinguishes them from other genera (Stubbs, 2021).

Nephrotoma appendiculata is a moderate sized cranefly with a wing length of 12–15 mm. It usually has a pale stigma, but this can be dark in a few specimens. There is a wide dull black stipe on the dorsal abdomen reaching across to the yellow sides. The determining feature is an upside-down U-shaped black mark around the base of the halteres (Stubbs, 2021).

Nephrotoma appendiculata is a common grassland species with adults flying from April to early June. It is tolerant of a range of pH and moisture levels, preferring unimproved grassland with medium to long grass, on better soils, but avoiding short turf, impoverished grassland and shade (Stubbs, 2021).

The structure of the spermatozoa of *Nephrotoma appendiculata* was found to be similar to that of several craneflies in the family Limoniidae and this has been used to support the idea that the families Tipulidae and Limoniidae should be combined (Dallai *et al.*, 2008) but they currently remain classified into two families (Chandler, 2023), and the final decision is likely to be based on phylogenetic analyses of DNA sequences.

The genome of the spotted cranefly, *Nephrotoma appendiculata*, 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. Here we present a chromosomally complete genome sequence for *Nephrotoma appendiculata*, based on one male specimen from Wytham Woods, Oxfordshire, UK.

Genome sequence report

The genome was sequenced from one male *Nephrotoma appendiculata* (Figure 1) collected from Wytham Woods, Oxfordshire, UK (51.76, –1.34). A total of 30-fold coverage in Pacific Biosciences single-molecule HiFi long reads was generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 70 missing joins or mis-joins and removed 17 haplotypic duplications, reducing the assembly length by 0.25% and the scaffold number by 12.06%, and decreasing the scaffold N50 by 45.60%.

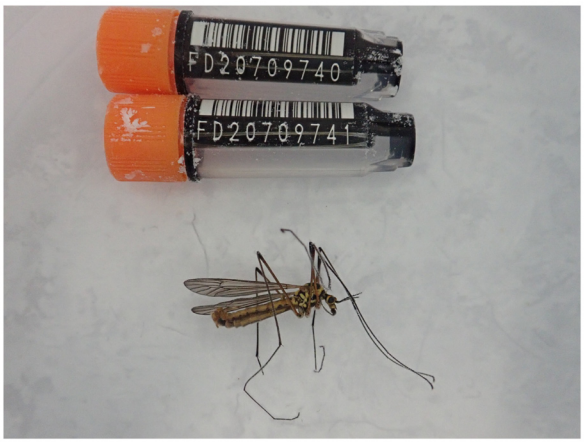

Figure 1. Photograph of the *Nephrotoma appendiculata* (idNepAppe1) specimen used for genome sequencing.

The final assembly has a total length of 1138.0 Mb in 422 sequence scaffolds with a scaffold N50 of 375.9 Mb (Table 1). The snailplot in Figure 2 provides a summary of the assembly statistics, while the distribution of assembly scaffolds on GC proportion and coverage is shown in Figure 3. The cumulative assembly plot in Figure 4 shows curves for subsets of scaffolds assigned to different phyla. Most (99.7%) of the assembly sequence was assigned to 4 chromosomal-level scaffolds, representing 3 autosomes and the X sex chromosome. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 2). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial genome was also assembled and can be found as a contig within the multifasta file of the genome submission.

The estimated Quality Value (QV) of the final assembly is 61.0 with *k*-mer completeness of 100.0%, and the assembly has a BUSCO v5.3.2 completeness of 94.4% (single = 93.0%, duplicated = 1.4%), using the diptera_odb10 reference set (n = 3,285).

Metadata for specimens, barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly statistics are given at https://links.tol.sanger.ac.uk/species/2741127.

Genome annotation report

The *Nephrotoma appendiculata* genome assembly (GCA_947310385.1) was annotated using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Nephrotoma_appendiculata_GCA_947310385.1/Info/Index). The resulting annotation includes 24,340 transcribed mRNAs from 14,126 protein-coding genes and 3,241 non-coding genes.

Methods

Sample acquisition and nucleic acid extraction

A male *Nephrotoma appendiculata* (specimen ID Ox001277, ToLID idNepAppe1) was netted in Wytham Woods, Oxfordshire (biological vice-county Berkshire), UK (latitude

Table 1. Genome data for *Nephrotoma appendiculata*, idNepAppe1.1.

Project accession data		
Assembly identifier	idNepAppe1.1	
Species	*Nephrotoma appendiculata*	
Specimen	idNepAppe1	
NCBI taxonomy ID	2741127	
BioProject	PRJEB55031	
BioSample ID	SAMEA10166758	
Isolate information	idNepAppe1, male: head and thorax (DNA and Hi-C sequencing), abdomen (RNA sequencing)	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	61.0	*≥ 50*
k-mer completeness	100.0%	*≥ 95%*
BUSCO**	C:94.4%[S:93.0%,D:1.4%], F:0.9%,M:4.7%,n:3,285	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	99.7%	*≥ 95%*
Sex chromosomes	X	*localised homologous pairs*
Organelles	Mitochondrial genome: 17.42 kb	*complete single alleles*
Raw data accessions		
PacificBiosciences SEQUEL II	ERR10008908	
Hi-C Illumina	ERR10015065	
PolyA RNA-Seq Illumina	ERR10378025	
Genome assembly		
Assembly accession	GCA_947310385.1	
Accession of alternate haplotype	GCA_947311015.1	
Span (Mb)	1,138.0	
Number of contigs	1102	
Contig N50 length (Mb)	3.2	
Number of scaffolds	422	
Scaffold N50 length (Mb)	375.9	
Longest scaffold (Mb)	384.6	
Genome annotation		
Number of protein-coding genes	14,126	
Number of non-coding genes	3,241	
Number of gene transcripts	24,340	

* Assembly metric benchmarks are adapted from column VGP-2020 of "Table 1: Proposed standards and metrics for defining genome assembly quality" from (Rhie *et al.*, 2021).

** BUSCO scores based on the diptera_odb10 BUSCO set using version 5.3.2. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/CAMZJQ01/dataset/CAMZJQ01/busco.

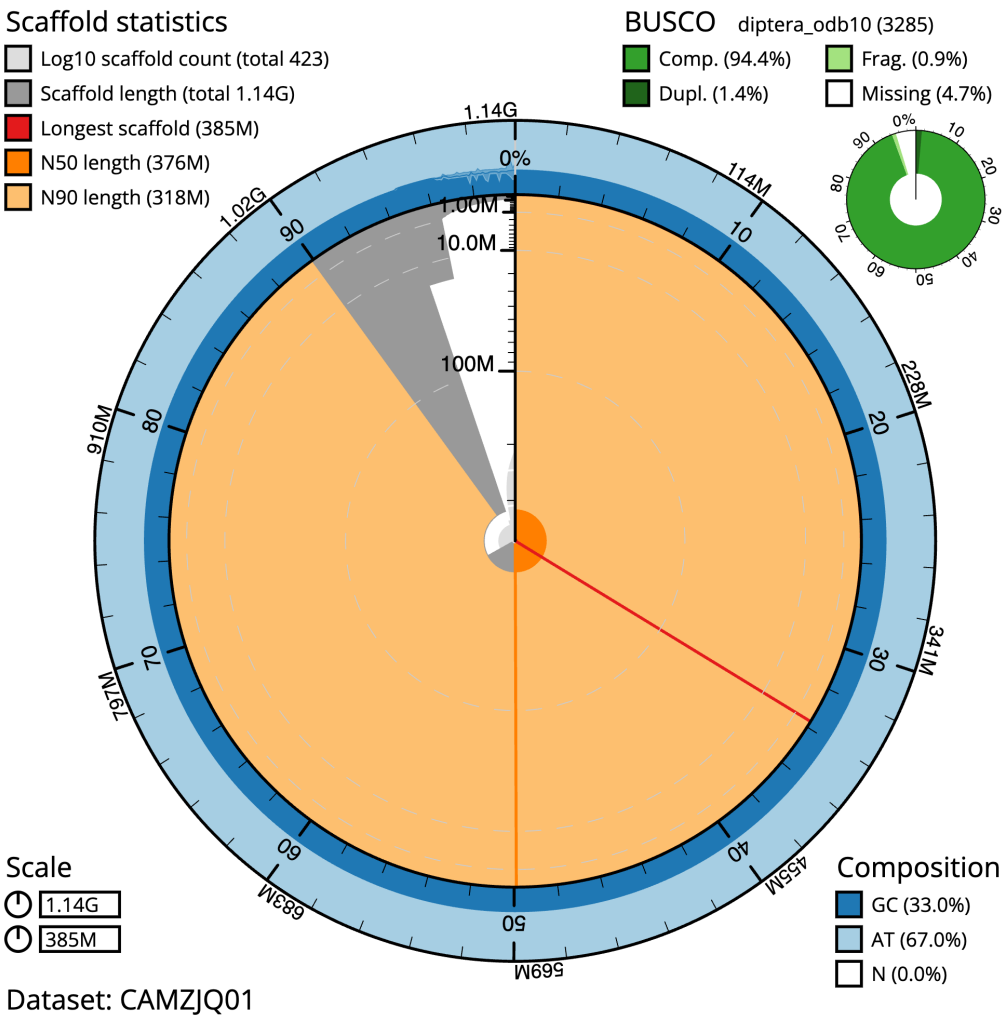

Figure 2. Genome assembly of *Nephrotoma appendiculata*, idNepAppe1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 1,138,061,071 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (384,599,746 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (375,927,842 and 317,755,181 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the diptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/CAMZJQ01/dataset/CAMZJQ01/snail.

51.76, longitude –1.34) on 2021-04-23. The specimen was collected and identified by Liam Crowley (University of Oxford) and preserved on dry ice.

Protocols developed by the Wellcome Sanger Institute (WSI) Tree of Life core laboratory have been deposited on protocols.io (Denton *et al.*, 2023b). The workflow for high molecular weight (HMW) DNA extraction at the WSI includes a sequence of core procedures: sample preparation; sample homogenisation, DNA extraction, fragmentation, and clean-up. In sample preparation, the idNepAppe1 sample was weighed and dissected on dry ice (Jay *et al.*, 2023). Tissue from the head and thorax was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a). HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023). HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 31 (Bates *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation (Strickland *et al.*, 2023): in brief, the method employs a 1.8X ratio of AMPure PB beads to sample to eliminate shorter fragments and concentrate the DNA. The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer and Qubit

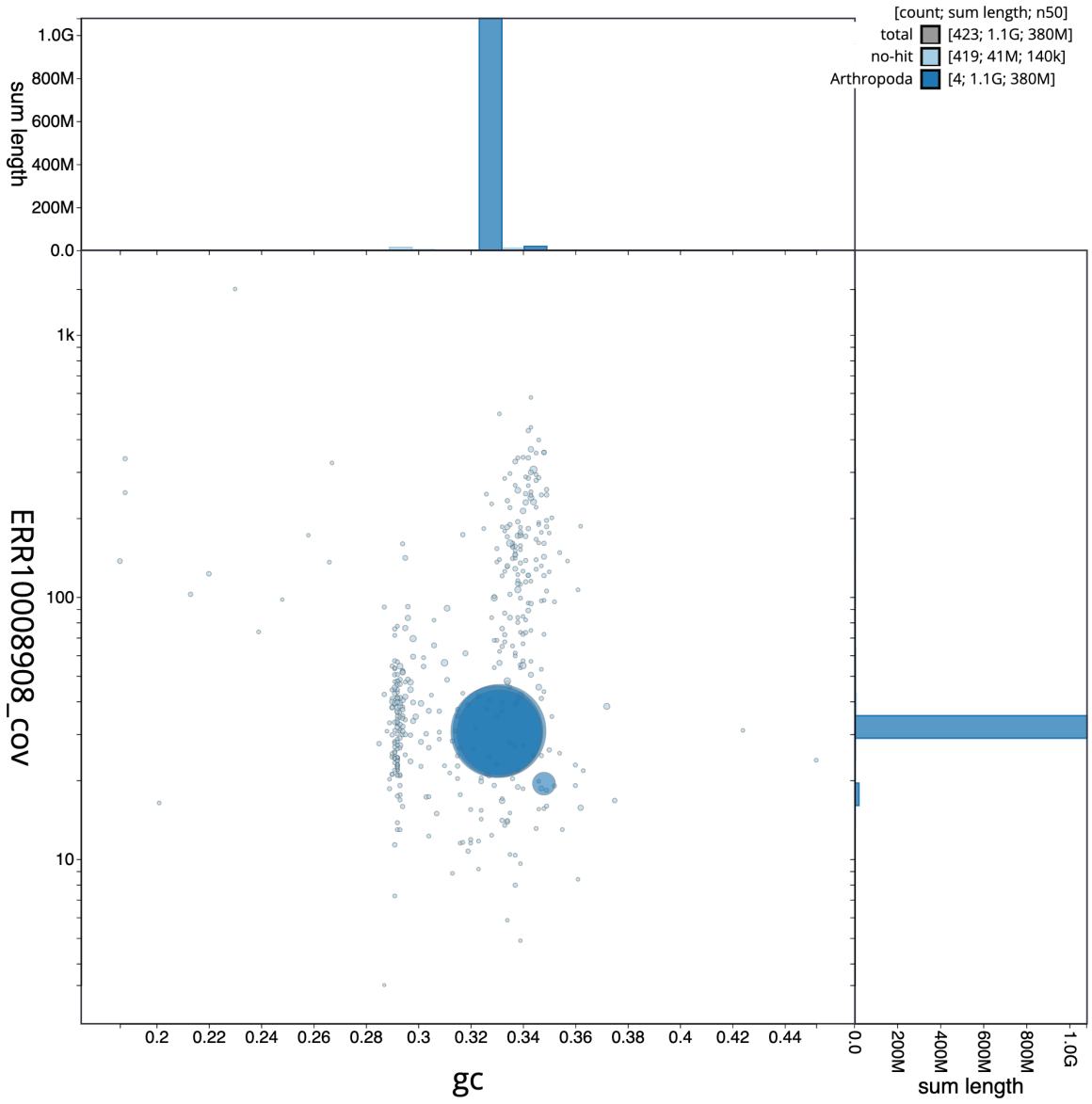

Figure 3. Genome assembly of *Nephrotoma appendiculata*, idNepAppe1.1: BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/CAMZJQ01/dataset/CAMZJQ01/blob.

dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

RNA was extracted from abdomen tissue of idNepAppe1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol (do Amaral *et al.*, 2023). 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.

Sequencing

Pacific Biosciences HiFi circular consensus DNA sequencing libraries were constructed according to the manufacturers' instructions. Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit. DNA and RNA sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi) and Illumina NovaSeq 6000 (RNA-Seq) instruments. Hi-C data were also generated from remaining head and thorax tissue of idNepAppe1 using the Arima2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

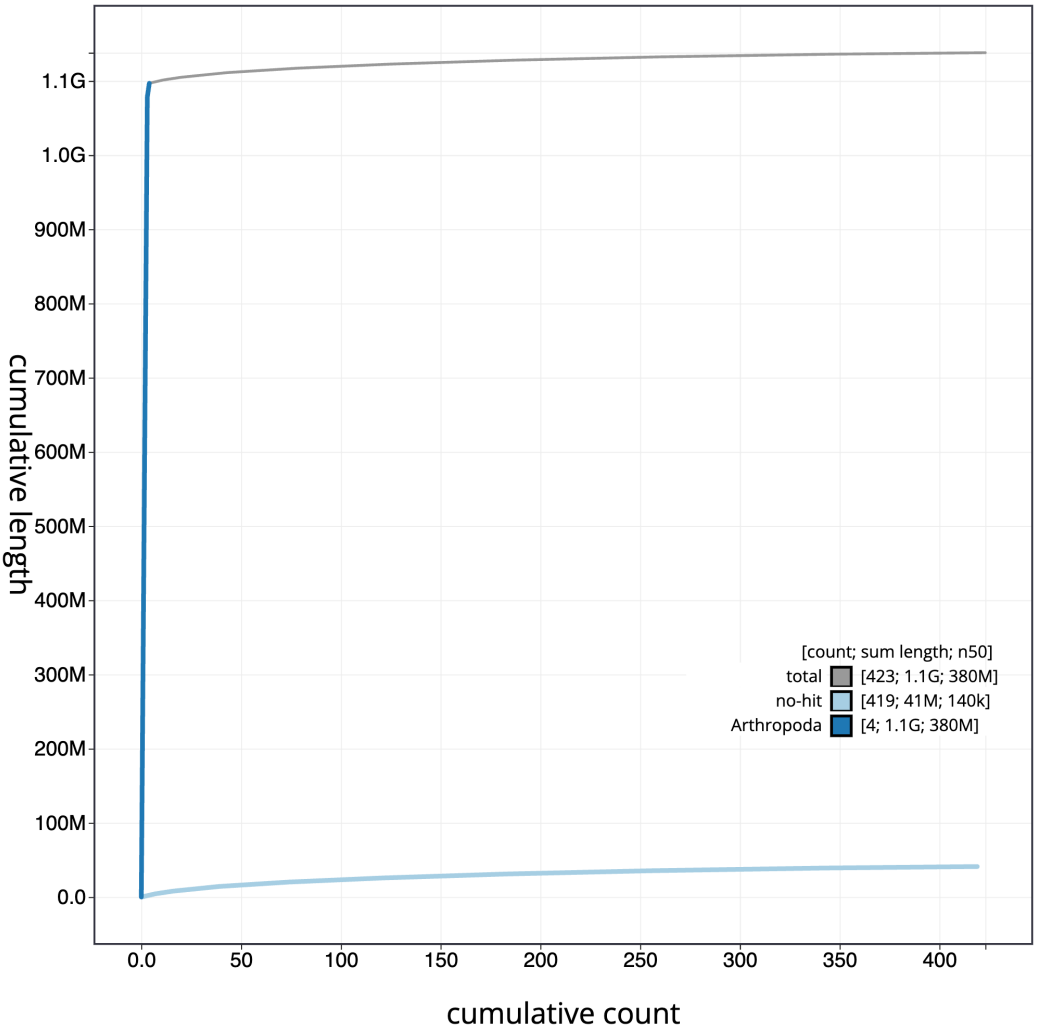

Figure 4. Genome assembly of *Nephrotoma appendiculata*, idNepAppe1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/CAMZJQ01/dataset/CAMZJQ01/cumulative.

Genome assembly, curation and evaluation

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021) and haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2020). The assembly was then scaffolded with Hi-C data (Rao *et al.*, 2014) using YaHS (Zhou *et al.*, 2023). The assembly was checked for contamination and corrected as described previously (Howe *et al.*, 2021). Manual curation was performed using HiGlass (Kerpedjiev *et al.*, 2018) and Pretext (Harry, 2022). The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) or MITOS (Bernt *et al.*, 2013) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). To assess the assembly metrics, the *k*-mer completeness and QV consensus quality values were calculated in Merqury (Rhie *et al.*, 2020). This work was done using Nextflow (Di Tommaso *et al.*, 2017) DSL2 pipelines "sanger-tol/readmapping" (Surana *et al.*, 2023a) and "sanger-tol/genomenote" (Surana *et al.*, 2023b). The genome was analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021; Simão *et al.*, 2015) were calculated.

Table 3 contains a list of relevant software tool versions and sources.

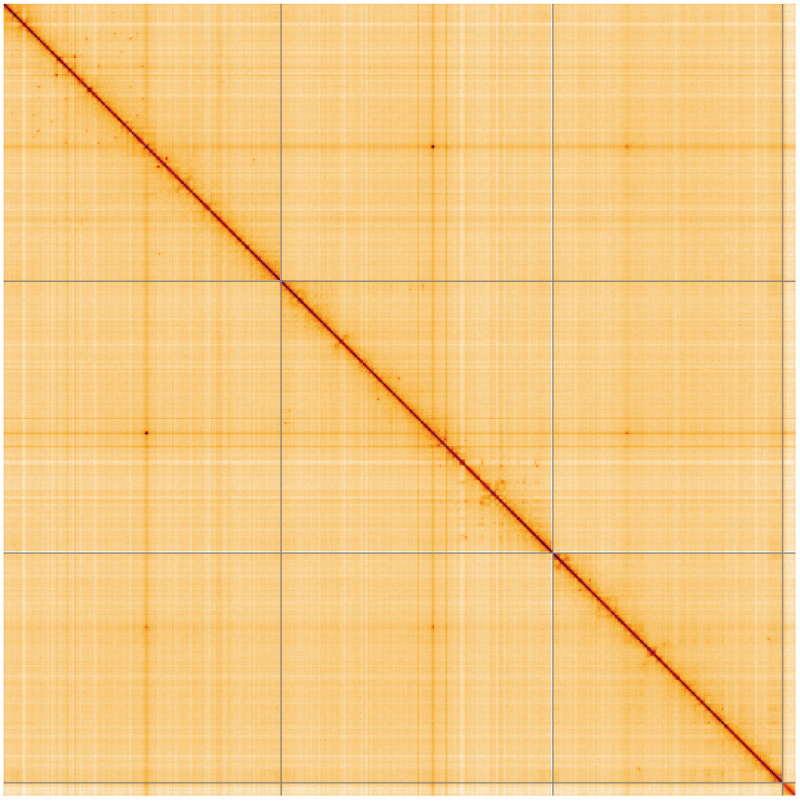

Figure 5. Genome assembly of *Nephrotoma appendiculata*, idNepAppe1.1: Hi-C contact map of the idNepAppe1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=ENODhlutRSaWfOvEvECH7A.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Nephrotoma appendiculata*, idNepAppe1.

INSDC accession	Chromosome	Length (Mb)	GC%
OX371223.1	1	384.6	33.0
OX371224.1	2	375.93	33.0
OX371225.1	3	317.76	33.0
OX371226.1	X	18.65	35.0
OX371227.1	MT	0.02	23.0

Genome annotation

The Ensembl gene annotation system (Aken *et al.*, 2016) was used to generate annotation for the *Nephrotoma appendiculata* assembly (GCA_947310385.1). Annotation was created primarily through alignment of transcriptomic data to the genome, with gap filling via protein-to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 2019).

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the **'Darwin Tree of Life Project Sampling Code of Practice'**, which can be found in full on the Darwin Tree of Life website here. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and 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

Table 3. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	4.1.7	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.3.2	https://gitlab.com/ezlab/busco
Hifiasm	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	1.11.6	https://github.com/higlass/higlass
Merqury	MerquryFK	https://github.com/thegenemyers/MERQURY.FK
MitoHiFi	2	https://github.com/marcelauliano/MitoHiFi
PretextView	0.2	https://github.com/wtsi-hpag/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
sanger-tol/genomenote	v1.0	https://github.com/sanger-tol/genomenote
sanger-tol/readmapping	1.1.0	https://github.com/sanger-tol/readmapping/tree/1.1.0
YaHS	yahs-1.1.91eebc2	https://github.com/c-zhou/yahs

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: *Nephrotoma appendiculata* (spotted cranefly). Accession number PRJEB55031; https://identifiers.org/ena.embl/PRJEB55031 (Wellcome Sanger Institute, 2022). The genome sequence is released openly for reuse. The *Nephrotoma appendiculata* 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. Raw data and assembly accession identifiers are reported in Table 1.

Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.7125292.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.4893703.

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: https://doi.org/10.5281/zenodo.10066175.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: https://doi.org/10.5281/zenodo.10043364.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: https://doi.org/10.5281/zenodo.10066637.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.5013541.

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

Aken BL, Ayling S, Barrell D, *et al.*: **The Ensembl gene annotation system.** *Database (Oxford).* 2016; **2016**: baw093.
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

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

Bernt M, Donath A, Jühling F, *et al.*: **MITOS: Improved *de novo* metazoan**

mitochondrial genome annotation. *Mol Phylogenet Evol.* 2013; **69**(2): 313–319.
PubMed Abstract | Publisher 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

Chandler P: **Checklist of Diptera of the British Isles.** 2023.
Reference Source

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

Dallai R, Lombardo BM, Mercati D, *et al.*: **Sperm structure of Limoniidae and their phylogenetic relationship with Tipulidae (*Diptera, Nematocera*).** *Arthropod Struct Dev.* 2008; **37**(1): 81–92.
PubMed Abstract | Publisher 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.** *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

do Amaral RJV, Bates A, Denton A, *et al.*: **Sanger Tree of Life RNA Extraction: Automated MagMax™ mirVana.** *protocols.io.* 2023.
Publisher 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; [Accessed 19 October 2022].
Reference Source

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* Oxford University Press, 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

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

Oatley G, Denton A, Howard C: **Sanger Tree of Life HMW DNA Extraction: Automated MagAttract v.2.** *protocols.io.* 2023.
Publisher 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

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–3212.
PubMed Abstract | Publisher Full Text

Strickland M, Cornwell C, Howard C: **Sanger Tree of Life Fragmented DNA clean up: Manual SPRI.** *protocols.io.* 2023.
Publisher Full Text

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

Surana P, Muffato M, Qi G: **sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo.* 2023a.
Reference Source

Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev).** *Zenodo.* 2023b.
Publisher 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

UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–D515.
PubMed Abstract | Publisher Full Text | Free Full Text

Vasimuddin Md, 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

Wellcome Sanger Institute: **The genome sequence of the spotted cranefly, *Nephrotoma appendiculata* (Pierre, 1919).** European Nucleotide Archive, [dataset], accession number PRJEB55031. 2022.

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 21 May 2024

https://doi.org/10.21956/wellcomeopenres.23110.r81550

✔ **Hans-Peter Fuehrer**
University of Veterinary Medicine Vienna, Vienna, Austria

The authors present a genome assembly from an individual spotted cranefly (*Nephrotoma appendiculata*; Tipulidae). Methods are described well. Overall the manuscript is of relevance and good quality.

Only two minor corrections are recommended:

- Abstract and Tab.1: Change 1,138.0 to 1,138
- Change to 1,102 at number of contigs in Tab 1.

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: Parasitology, Entomology

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 12 April 2024

https://doi.org/10.21956/wellcomeopenres.23110.r78182

Adam James Reid
University of Cambridge, Cambridge, UK

This genome note represents a very high quality genome assembly of the spotted cranefly, *Nephrotoma appendiculata*. The standards of methodological reporting and the availability of data and code are excellent. The authors have taken pains to acknowledge the large number of people involved in the (Darwin) Tree of Life endeavour. There are minor issues with some of the presentation, which could be improved in places.

In the section *Genome sequence report,* I assume that "(51.76, –1.34)" is a geographical reference, but the frame of reference is not given. Perhaps it could say "latitude = 51.76, longitude = –1.34".

Also in the section *Genome sequence report*. I wanted to check that this phrase is correct: "and decreasing the scaffold N50 by 45.60%", i.e. that after manual assembly the scaffold N50 reduced, presumably due to the breaking of very large scaffolds. The data wasn't there for me to confirm.

I'm afraid that even though I have seen them multiple times before I do not find the snail plots at all straightforward to understand. While I understand the desire to capture diverse and complicated information in a plot that can be automatically generated, I don't think the majority of readers will get much from this. Why are there three different blue colours for the GC content on the smaller contigs? Why is there not more red representing the largest contig and dark orange representing the N50? The assembly for the Tiger Cranefly (Sivell O, et.al., 2023 [Ref 1]) is quite similar, but the plots look very different. Is it possible to interpret the pale grey spiral and which orders of magnitude are represented by the white lines? The immediate solution to these problems would be to add more information to the legend to better guide the reader through what they are seeing.

In Table 1, the benchmark column is generally very useful for giving the reader a good idea of what is expected of a good assembly, however the values in the *Sex Chromosomes* and *Organelles* rows don't seem to make sense. We at least need a better explanation of what the possible values might be. Are the benchmarks met in these cases?

References

1. Sivell O, Sivell D, Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, et al.: The genome sequence of a Tiger Cranefly, Nephrotoma flavescens (Linnaeus, 1758).*Wellcome Open Res*. 2023; **8**: 148 PubMed Abstract | Publisher Full Text

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

Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, genomics, transcriptomics, epigenomics, parasitology, developmental biology

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.