DATA NOTE

REVISED The genome sequence of the lesser worm flesh fly, *Sarcophaga* (*Sarcophaga*) *subvicina* Rohdendorf, 1937

[version 2; peer review: 4 approved, 2 approved with reservations]

Steven Falk[1], University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, John F. Mulley[2], Darwin Tree of Life Consortium

[1]independent Researcher, Kenilworth, Warwickshire, UK
[2]School of Natural Sciences, Bangor University, Bangor, Wales, UK

v2 **First published:** 08 Feb 2023, **8**:65
https://doi.org/10.12688/wellcomeopenres.18717.1
Latest published: 04 Sep 2024, **8**:65
https://doi.org/10.12688/wellcomeopenres.18717.2

Abstract

We present a genome assembly from an individual male *Sarcophaga subvicina* (the lesser worm flesh fly; Arthropoda; Insecta; Diptera; Sarcophagidae). The genome sequence is 71 megabases in span. Most of the assembly (95.91%) is scaffolded into six chromosomal pseudomolecules, with the X sex chromosome assembled. The mitochondrial genome has also been assembled and is 16.7 kilobases in length. Gene annotation of this assembly on Ensembl identified 16,793 protein coding genes.

Keywords

Sarcophaga subvicina, lesser worm flesh fly, genome sequence, chromosomal, Diptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ? ✓ ✓ ? ✓ ✓

	1	2	3	4	5	6
version 2 (revision) 04 Sep 2024			✓ view	? view	✓ view	✓ view
version 1 08 Feb 2023	? view	✓ view	? view			

1. **Paul Mireji**, Kenya Agricultural and Livestock Research Organization, Kikuyu, Kenya
2. **Carel Oosthuizen**, University of Pretoria, Pretoria, South Africa
3. **Andrzej Grzywacz**, Nicolaus Copernicus University, Toruń, Poland
4. **Jaakko Pohjoismäki**, University of Eastern Finland, Joensuu, Finland
5. **Jason Charamis**, Foundation for Research and Technology - Hellas, Irákleion, Greece

6. **Ruiqi Li**, University of Colorado Boulder, Boulder, 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: Falk S: Investigation, Resources; **Mulley JF**: Writing – Original Draft Preparation, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

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

How to cite this article: Falk S, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective *et al.* **The genome sequence of the lesser worm flesh fly, *Sarcophaga* (*Sarcophaga*) *subvicina* Rohdendorf, 1937 [version 2; peer review: 4 approved, 2 approved with reservations]** Wellcome Open Research 2024, **8**:65
https://doi.org/10.12688/wellcomeopenres.18717.2

First published: 08 Feb 2023, **8**:65 https://doi.org/10.12688/wellcomeopenres.18717.1

REVISED **Amendments from Version 1**

The following changes have been made to the article:

- The species taxonomic authority has been corrected to Rohdendorf, 1937 throughout.
- We have included a link to the TOLQC page providing assembly metadata for this sequencing project.
- We have added a chromosome grid to the Hi-C map for Figure 5.
- We have added information on species identification.
- We corrected the text on the sample homogenisation step of DNA extraction.

Any further responses from the reviewers can be found at the end of the article

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Oestroidea; Sarcophagidae; Sarcophaga; *Sarcophaga*; *Sarcophaga subvicina* Rohdendorf, 1937 (NCBI txid:236850).

Background

Sarcophaga subvicina (Diptera: Sarcophagidae) is a relatively large (up to 8–15 mm (van Emden, 1954)) flesh fly with a Palearctic distribution (Pape, 1996). *S. subvicina* show the characteristic patterning of the Sarocophaga genus, with an overall blackish/greyish colouration, a checked abdomen, three longitudinal stripes on the thorax, and large red/orange eyes, and so can be difficult to separate from other members of the genus without examination of male genitalia or DNA barcoding (Jordaens *et al.*, 2013; Szpila *et al.*, 2015). *Sarcophaga* is a large genus, and the nearly 900 species contained within it are classified into 169 subgenera (Buenaventura *et al.*, 2017), with *S. subvicina* placed in the *Sarcophaga* subgenus along with over 20 other species (Pape, 1996). The relative species-richness of this subgenus stands in stark contrast to the majority of sarcophagid subgenera, which are monotypic. The *Sarcophaga* subgenus contains three of the roughly 65 currently recognised UK Sarcophagid species (*S. carnaria*, *S. variegata*, and *S. subvicina*), in what is often termed the "carnaria group".

Sarcophaga subvicina is found across the UK, with a range that extends to the north of Scotland, and is most abundant between May and September (see: https://species.nbnatlas.org/species/NBNSYS0000030329). It has been reported as favouring open (urban/grassland) habitats (Fremdt & Amendt, 2014; Hwang & Turner, 2005), and adults have been attracted to large carcasses (Szpila *et al.*, 2015). Larvae have been reported only from small mammal carcasses, and reared in captivity on meat and dead slugs (Blackith & Blackith, 1994; Pape, 1987), but this species seems to more likely represent an earthworm specialist. All Sarcophagids examined to date have a diploid chromosome number of 12, with an XY sex determination system and males the heterogametic sex (Srivastava & Gaur, 2015).

The genome of the lesser worm flesh fly *S. subvicina* 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 *S. subvicina* based on an individual male specimen from Wytham Woods, Berkshire.

Genome sequence report

The genome was sequenced from one male *S. subvicina* specimen collected in Wytham Woods, Berkshire (Figure 1). A total of 65-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 51-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 95 missing/misjoins and removed four haplotypic duplications, reducing the assembly length by 0.57% and the scaffold number by 16.97%, and increasing the scaffold N50 by 4.73%.

The final assembly has a total length of 714 Mb in 274 sequence scaffolds with a scaffold N50 of 123 Mb (Table 1). Most (95.91%) of the assembly sequence was assigned to six chromosomal-level scaffolds, representing 5 autosomes and the X sex chromosome (Figure 2–Figure 5; Table 2). Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size. This is a male specimen with known XY sex determination system, however we have been unable to identify Y sequences. The X chromosome is assembled from scaffolds of undetermined order and orientation. The assembly has a BUSCO 5.3.2 (Manni *et al.*, 2021) completeness of 99.2% (single 98.5%, duplicated 0.7%), using the diptera_odb10 reference set (n = 3,285). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

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/236850.

Figure 1. Image of the *Sarcophaga subvicina* (idSarSubv1) specimen used for genome sequencing.

Table 1. Genome data for idSarSubv1.1.

Project accession data	
Assembly identifier	idSarSubv1.1
Species	*Sarcophaga subvicina*
Specimen	idSarSubv1
NCBI taxonomy ID	236850
BioProject	PRJEB51465
BioSample ID	SAMEA7746447
Isolate information	male, thorax tissue (genomic DNA), head tissue (Hi-C)
Assembly metrics*	
Base pair QV	52.9 (Benchmark: ≥50)
k-mer completeness	99.99% (Benchmark: ≥95%)
BUSCO**	C:99.2%[S:98.5%,D:0.7%],F:0.2%,M:0.6%,n:3285 (Benchmark: C ≥ 95%)
Percentage of assembly mapped to chromosomes	95.91% (Benchmark: ≥95%)
Sex chromosomes	X chromosome identified (Benchmark: localised homologous pairs)
Organelles	Mitochondrion genome assembled (Benchmark: complete single alleles)
Raw data accessions	
PacificBiosciences SEQUEL II	ERR9284049, ERR9284050
10X Genomics Illumina	ERR9248453–ERR9248456
Hi-C Illumina	ERR9248452
Genome assembly	
Assembly accession	GCA_936449025.1
Accession of alternate haplotype	GCA_936440885.1
Span (Mb)	714.2
Number of contigs	445
Contig N50 length (Mb)	102.9
Number of scaffolds	274
Scaffold N50 length (Mb)	122.7
Longest scaffold (Mb)	159.5
Genome annotation	
Number of protein-coding genes	16,793

* 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 v5.3.2. 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/idSarSubv1.1/dataset/CAKZFR01/busco.

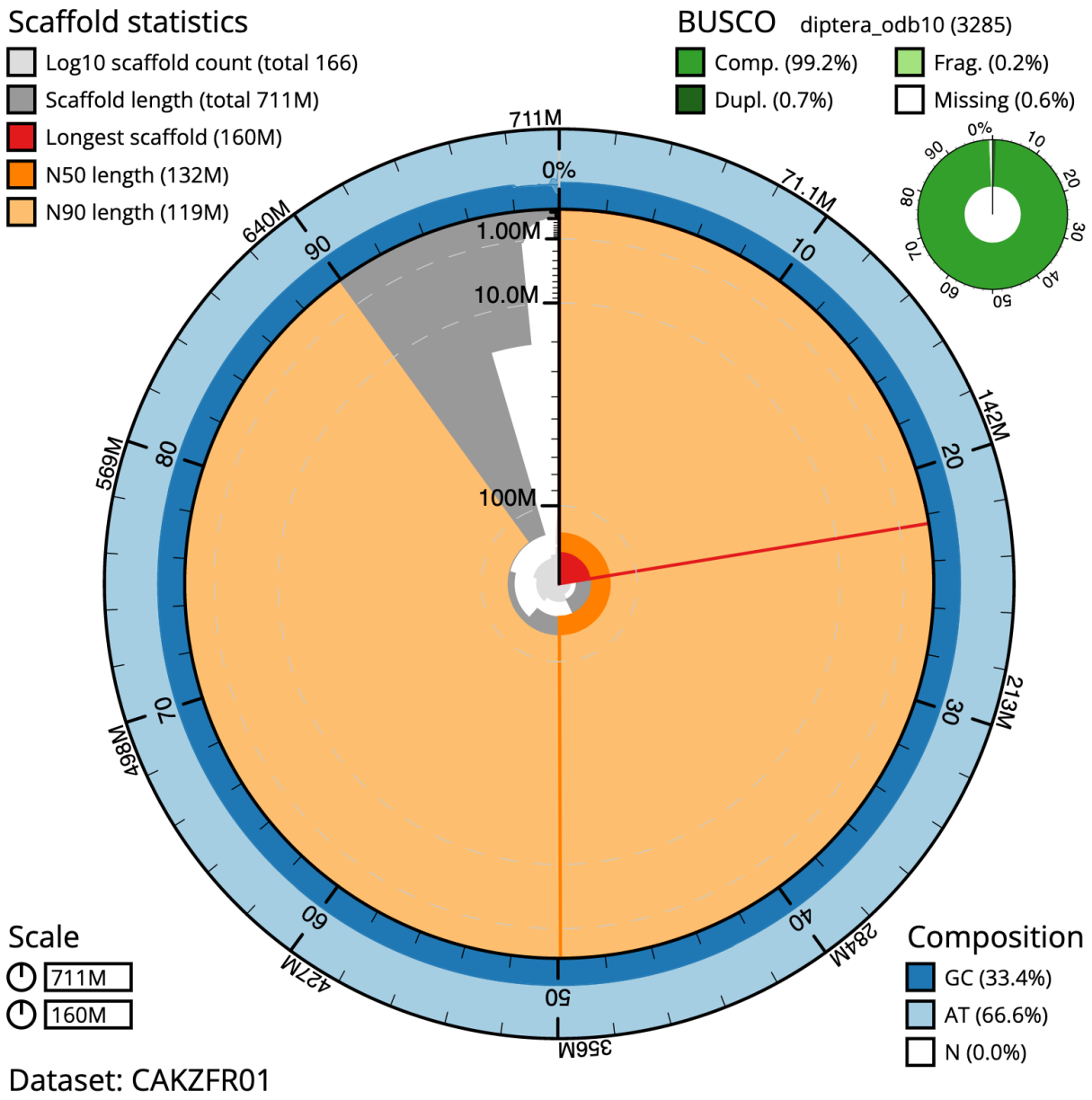

Figure 2. Genome assembly of *Sarcophaga subvicina*, idSarSubv1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 711,151,016 bp assembly. The distribution of chromosome lengths is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly (159,501,612bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (132,242,496 and 118,606,681bp), respectively. The pale grey spiral shows the cumulative chromosome 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/idSarSubv1.1/dataset/CAKZFR01/snail.

Genome annotation report

The idSarSubv1.1 genome was annotated using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Sarcophaga_subvicina_GCA_936449025.1/). The resulting annotation includes 39,250 transcribed mRNAs from 16,793 protein-coding and 11,903 non-coding genes.

Methods

Sample acquisition and nucleic acid extraction

A male *S. subvicina* (idSarSubv1) was collected and identified by Steven Falk (independent researcher). The species was identified using the latest keys to the identification of Sarcophagidae (https://osf.io/preprints/osf/vf5r6), and species

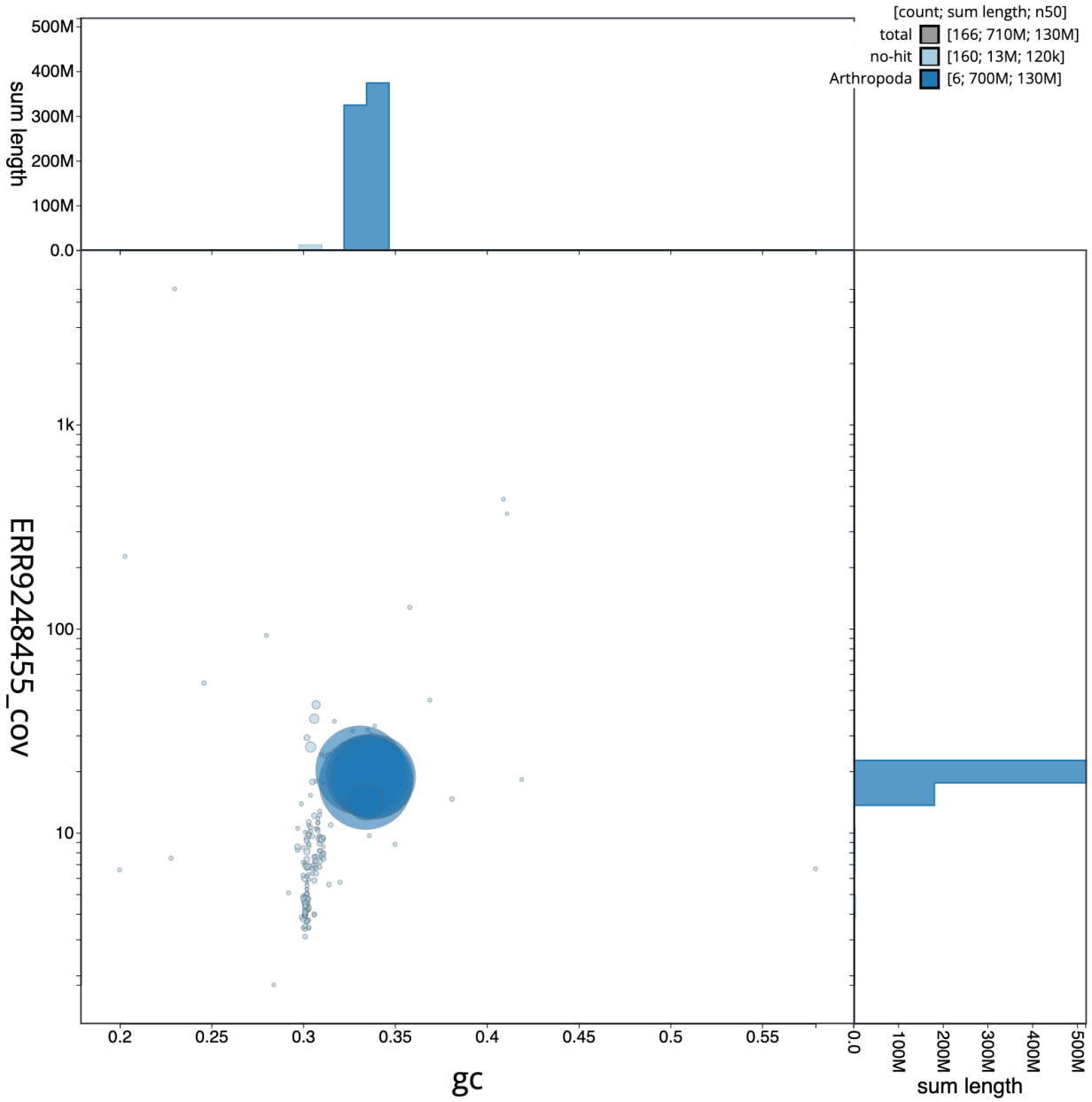

Figure 3. Genome assembly of *Sarcophaga subvicina*, idSarSubv1.1: GC coverage. BlobToolKit GC-coverage plot. Chromosomes are coloured by phylum. Circles are sized in proportion to chromosome length. Histograms show the distribution of chromosome length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/idSarSubv1.1/dataset/CAKZFR01/blob.

identification was also confirmed by COI barcode. The specimen was collected using a net in Wytham Woods, Berkshire (latitude 51.766, longitude –1.309) and snap-frozen on dry ice.

DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute. The idSarSubv1 sample was weighed and dissected on dry ice with head tissue set aside for Hi-C sequencing. Thorax tissue was disrupted using a Nippi Powermasher fitted with a BioMasher pestle. High molecular weight (HMW) DNA was extracted using the Qiagen MagAttract HMW DNA extraction kit. Low molecular weight DNA was removed from a 20 ng aliquot of extracted DNA using 0.8X AMpure XP purification kit prior to 10X Chromium sequencing; a minimum of 50 ng DNA was submitted for 10X sequencing. HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 30. Sheared DNA was purified by solid-phase reversible immobilisation using AMPure PB beads with a 1.8X ratio of beads to sample to remove the shorter fragments and concentrate the DNA sample. The concentration of the sheared and purified DNA was assessed using a Nanodrop

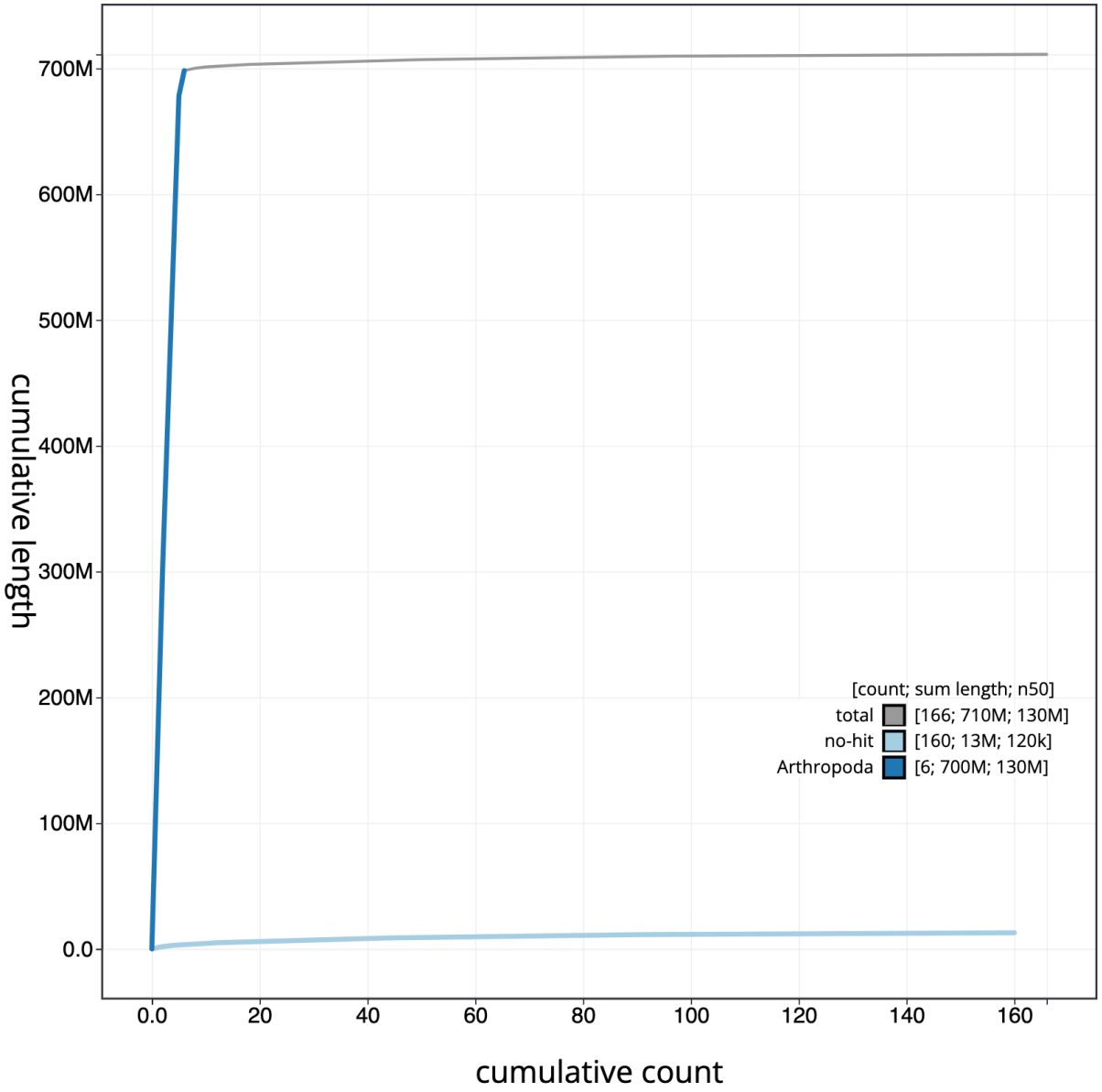

Figure 4. Genome assembly of *Sarcophaga subvicina*, idSarSubv1.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all chromosomes. Coloured lines show cumulative lengths of chromosomes assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/idSarSubv1.1/dataset/CAKZFR01/cumulative.

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

Sequencing

Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud DNA sequencing libraries were constructed according to the manufacturers' instructions. DNA sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi) and Illumina NovaSeq 6000 (10X) instruments. Hi-C data were also generated from head tissue of idSarSubv1 using the Arima v2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

Genome assembly

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021) and haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2020). One round of polishing was performed by aligning 10X Genomics read data to the assembly with Long Ranger ALIGN, calling variants with

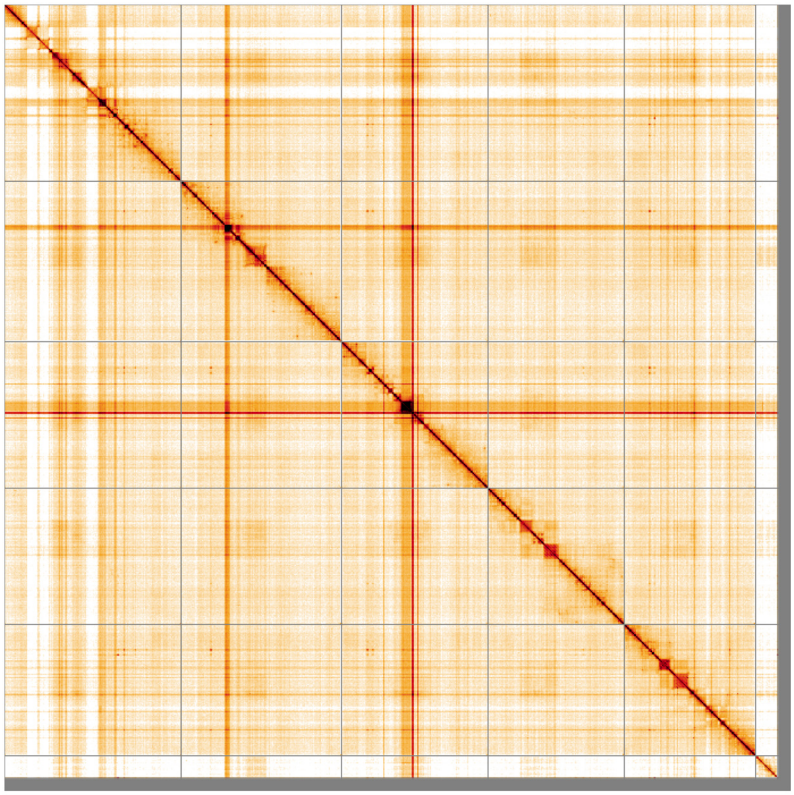

Figure 5. Genome assembly of *Sarcophaga subvicina*, idSarSubv1.1: Hi-C contact map. Hi-C contact map of the idSarSubv1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=fjqWR98ySXisVhiBypHNoA.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Sarcophaga subvicina*, idSarSubv1.

INSDC accession	Chromosome	Size (Mb)	GC%
OW388080.1	1	159.5	33.4
OW388081.1	2	144.87	33.1
OW388082.1	3	132.24	33.8
OW388083.1	4	122.97	33.5
OW388084.1	5	118.61	33.8
OW388085.1	X	20.1	33.4
OW388086.1	MT	0.02	23

freebayes (Garrison & Marth, 2012). The assembly was then scaffolded with Hi-C data (Rao *et al.*, 2014) using YaHS (Zhou *et al.*, 2022). The assembly was checked for contamination 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.*, 2021), which performed annotation using MitoFinder (Allio *et al.*, 2020). The genome was analysed and BUSCO scores generated within the BlobToolKit environment (Challis *et al.*, 2020). Table 3 contains a list of all software tool versions used, where appropriate.

Genome annotation

The Ensembl gene annotation system (Aken *et al.*, 2016) at the European Bioinformatics Institute (EBI) was used to generate annotation for the *S. subvicina* assembly (GCA_936449025.1). 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).

Ethics/compliance issues

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. By agreeing with and signing up to the Sampling Code of Practice,

Table 3. Software tools and versions used.

Software tool	Version	Source
BlobToolKit	3.4.0	Challis *et al.*, 2020
freebayes	1.3.1-17- gaa2ace8	Garrison & Marth, 2012
Hifiasm	0.15.3	Cheng *et al.*, 2021
HiGlass	1.11.6	Kerpedjiev *et al.*, 2018
Long Ranger ALIGN	2.2.2	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines
MitoHiFi	2.0	Uliano-Silva *et al.*, 2021
PretextView	0.2.x	https://github.com/wtsi-hpag/PretextView
purge_dups	1.2.3	Guan *et al.*, 2020
YaHS	yahs-1.1.91eebc2	Zhou *et al.*, 2022

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. 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: *Sarcophaga subvicina*. Accession number PRJEB51465; https://identifiers.org/ena.embl/PRJEB51465 (Wellcome Sanger Institute, 2022).

The genome sequence is released openly for reuse. The *Sarcophaga subvicina* 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.4789928.

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 programme are listed here: https://doi.org/10.5281/zenodo.4783585.

Members of Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.4790455.

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

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

Blackith RM, Blackith RE: **A check-list of Irish flesh-flies (Diptera: Sarcophagidae: Sarcophagini) and their known distribution.** *Irish Naturalists' Journal.* 1994; **24**(11): 427–434.
Reference Source

Buenaventura E, Whitmore D, Pape T: **Molecular phylogeny of the hyperdiverse genus *Sarcophaga* (Diptera: Sarcophagidae), and comparison between algorithms for identification of rogue taxa.** *Cladistics.* 2017; **33**(2): 109–133.
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

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

Fremdt H, Amendt J: **Species composition of forensically important blow flies (Diptera: Calliphoridae) and flesh flies (Diptera: Sarcophagidae)**

through space and time. *Forensic Sci Int.* 2014; **236**: 1–9.
PubMed Abstract | Publisher Full Text

Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing**. 2012.
Reference Source

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

Hwang C, Turner BD: **Spatial and temporal variability of necrophagous *Diptera* from urban to rural areas.** *Med Vet Entomol.* 2005; **19**(4): 379–391.
PubMed Abstract | Publisher Full Text

Jordaens K, Sonet G, Richet R, *et al.*: **Identification of forensically important *Sarcophaga* species (Diptera: Sarcophagidae) using the mitochondrial *COI* gene.** *Int J Legal Med.* 2013; **127**(2): 491–504.
PubMed Abstract | 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

Pape T: **The Sarcophagidae (Diptera) of Fennoscandia and Denmark.** *Caryologia.* Leiden, Netherlands: Brill, 1987.
Reference Source

Pape T: **Catalogue of the Sarcophagidae of the world (Insecta: Diptera).** Associated Publishers, 1996.
Reference Source

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

Srivastava R, Gaur P: **Revelation of heterochromatin heterogeneity in Sarcophagid chromosomes using DNA ligand Mithramycin.** *Caryologia.* 2015; **68**(1): 55–60.
Publisher Full Text

Szpila K, Mądra A, Jarmusz M, *et al.*: **Flesh flies (Diptera: Sarcophagidae) colonising large carcasses in Central Europe.** *Parasitol Res.* 2015; **114**(6): 2341–2348.
PubMed Abstract | Publisher Full Text | Free Full Text

Uliano-Silva M, *et al.*: **MitoHiFi**. 2021. Accessed: 19 October 2022.
Reference Source

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

van Emden FI: **Diptera Cyclorrhapha, Calyptrata (I) Section (a). Tachinidae and Calliphoridae.** In: *Handbooks for the Identification of British Insects*'. In. Entomological Society of London, 1954.
Reference Source

Wellcome Sanger Institute: **The genome sequence of the lesser worm flesh fly, *Sarcophaga* (*Sarcophaga*) *subvicina* (Baranov, 1937).** European Nucleotide Archive. [Dataset]. 2022.
https://identifiers.org/ena.embl/PRJEB51465

Zhou C, McCarthy SA, Durbin R: **YaHS: yet another Hi-C scaffolding tool.** *bioRxiv.* 2022.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✓ ✓ ? ✓ ✓

Version 2

Reviewer Report 20 November 2024

https://doi.org/10.21956/wellcomeopenres.24707.r109727

✓ **Ruiqi Li**
University of Colorado Boulder, Boulder, Colorado, USA

The genome of *Sarcophaga subvicina* is high-quality. I only have a few minor comments:

1. It would be great if the authors could provide a paragraph describing what biological questions this genome could help us answer. See some examples here: Li et al 2024a,b

2. Add a scale to Fig. 1?

References

1. Li R, Li J, Lopez JV, Oatley G, et al.: The genome sequence of the giant clam, Tridacna gigas (Linnaeus, 1758).*Wellcome Open Res*. 2024; **9**: 145 PubMed Abstract | Publisher Full Text
2. Li R, Li J, Lopez J, Oatley G, et al.: The genome sequence of the giant clam, Tridacna crocea (Lamarck, 1819). *Wellcome Open Research*. 2023; **8**. 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: 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 05 November 2024

https://doi.org/10.21956/wellcomeopenres.24707.r109725

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

This study presents the genome assembly and annotation of *Sarcophaga subvicina*, a representative of lesser worm flesh flies. The work is technically sound and the generated genome assembly seems of great quality.

I have only a few requests before acceptance for indexing:

1. "The genome sequence is 71 megabases in span." : Typo should be corrected in the following sentence from Abstract - assembly size is 714 Mb

2. "Annotation was created primarily through alignment of transcriptomic data to the genome": the transcriptomic evidence used as hints for gene prediction are not mentioned, while new RNA sequencing data are not produced within the frame of this study

3. The number of predicted protein-coding genes is within the typical range for Diptera. To my view, what is currently missing is a BUSCO assessment of the ENSEMBL-produced gene annotation and a comparison with other reference dipterans, such as Drosophila melanogaster and/or other related species.

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

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

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

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Arthropod Comparative Genomics

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

Reviewer Report 04 November 2024

https://doi.org/10.21956/wellcomeopenres.24707.r108255

?

Jaakko Pohjoismäki
University of Eastern Finland, Joensuu, Finland

The data note by Steven Falk and others with the DToL consortium presents the reference-quality genome assembly of *Sarcophaga subvicina*. The assembly meets the EBP standards and relevant data related to its assembly, annotation and QC, as well as sample metadata are provided under links.

The note has been already reviewed once and the authors have made amendments accordingly. I have only few minor comments.

Background
1) Twice)) in the first reference.
2) Besides its *COI*-barcode,*Sarcophaga subvicina* can be reliably distinguished from the otherwise similar *carnaria*-group species only by the morphology of male genitalia. While I do not doubt the identification, especially as the barcoding was done, I would be curious to know how the snap-frozen specimen was handled, as the genitalia would need to be examined somehow. This is relevant because one needs to usually examine dozens of more common species (such as *subvicina*) to find one of the less abundant ones. I'm fine if the identification was made post-hoc, just wanted to know if the authors had found some solution for the trade-off between the requirement for time consuming morphological determination of unfrozen specimens, which can be manipulated to show the necessary details vs snap freezing them for the genome work.
3) Species in the *carnaria*-group should all be obligate predators/parasitoids of earthworms, records from carrions are likely misidentifications. While this is mentioned, it is not really referenced, but rather the focus is on trivia (attracted to large carcasses, breeding in captivity).

Methods
A point raised by an earlier review: It is mentioned that RNA-seq data was used to aid the annotation, but this has not been explained in the methods. I see from the sample metadata that head, thorax and abdomen were treated as separate samples. Methods state that the head was used for Hi-C and the thorax for the PacBio sequencing. I assume that the abdomen was then used for the RNA-seq?

As a note to the previous reviews, I do not find it necessary to find justifications for sequencing a

genome for a species nor detailed analysis of the genome structure etc. Production of genome assemblies will provide an excellent resource for any such future work and these genome consortia can focus only on their production and has value as such.

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: Molecular biology; biodiversity genetics; taxonomy

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, however I have significant reservations, as outlined above.

Reviewer Report 19 September 2024

https://doi.org/10.21956/wellcomeopenres.24707.r97290

Andrzej Grzywacz
Nicolaus Copernicus University, Toruń, Poland

The authors appropriately addressed the issues raised during the previous revision round.

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: Systematics, entomology, 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.

Version 1

Reviewer Report 17 August 2023

https://doi.org/10.21956/wellcomeopenres.20755.r63100

?

Andrzej Grzywacz
Nicolaus Copernicus University, Toruń, Poland

The report describes the results of extraction, sequencing and assembly of genome of *Sarcophaga subvicina*. The protocols of molecular work are well described. Some remarks consider other aspects of this work.

An error is present in the title of this report. *Sarcophaga* (*Sarcophaga*) *subvicina* has been described by Rohdendorf, not by Baranov. The correct title should include:
"Sarcophaga (*Sarcophaga*) *subvicina* Rohdendorf, 1937".

According to the catalogue of Pape (1996) it is a species of Palaearctic distribution. Please delete "Neararctic" [sic!].
Please correct cases where taxon names are not italicised.

Please change "Sarcophagid" to "sarcophagid".

Please provide some potential examples of studies where generated dataset can be employed.

I recommend adding more information concerning specimen identification. In particular, a reference to a taxonomic key used for species delimitation, i.e., latest species-level taxonomic literature that contains the currently accepted species concept (Meier 2017).

In the Methods section, please check whether a sentence "Thorax tissue was *[[if powermasher used:* disrupted using a Nippi Powermasher fitted with a BioMasher pestle; *else:* cryogenically disrupted to a fine powder using a Covaris cryoPREP Automated Dry Pulveriser, receiving multiple impacts*]]*."

is correct.

References

1. Pape T: Catalogue of the Sarcophagidae of the world (Insecta: Diptera). *Associated Publishers*. 1996.
2. MEIER R: Citation of taxonomic publications: the why, when, what and what not. *Systematic Entomology*. 2017; **42** (2): 301-304 Publisher Full Text

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

No

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: Systematics, entomology, 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, however I have significant reservations, as outlined above.

Author Response 09 May 2024

Tree of Life Team Sanger

Thank you for the helpful comments on this data note - we have revised the article in line with your suggestions.

Competing Interests: No competing interests were disclosed.

Reviewer Report 01 August 2023

https://doi.org/10.21956/wellcomeopenres.20755.r62041

Carel Oosthuizen

Zoology and Entomology, University of Pretoria, Pretoria, Gauteng, South Africa

This article reports on the genome assembly of *Sarcophaga subvicina.* Six chromosomes, 5 autosomes and one sex chromosome (X), were defined by assembling the sequence reads.

I think this is a great report on a study that is well performed with high quality results. The date is well explained and easily accessible in the different forms that it was reported in. Well done to the authors.

The only suggestion that I would like to make is that there be a paragraph included on the possible reasons why the Y chromosome sequences could not be identified as the authors already know that it should be present.

Once again, well done.

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: Population genetics, epigenetics, population 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 05 July 2023

https://doi.org/10.21956/wellcomeopenres.20755.r59681

?

Paul Mireji
Biotechnology Research Institute, Kenya Agricultural and Livestock Research Organization, Kikuyu, Kenya

In this manuscript, the authors isolated nuclear and mitochondrial DNA from single male one

male *Sarcophaga subvicina* and sequenced the DNA using Pacific Biosciences and 10X Genomics platforms, both of which generated long reads most appropriate for genome assembly due to their better (51-65) fold coverage than other technologies such as illimina generated short reads. The assembly was successfully achieved to a high (chromosome) level followed by curation using an appropriate suite of bioinformatic tools and annotations based on transcriptomic data and protein from UniProt database.

The sample collection, DNA extraction, sequencing and assembly are well described and the results well presented. I however have the following concerns

1. The importance of assembling this genome is not adequately presented, other than as part of a routine process in the as part of the Darwin Tree of Life Project and that that this species is attracted to large carcasses. Is there any other economic, medical/ veterinary etc importance of this species? What knowledge will better be understanding of the advance?

2. While the transcriptomic data was used in the annotation of the genome, information on generation and nature of this data is missing. Was it from the same species? Was it tissue specific or whole body? The nature (source) of this data influences the gene families that will be preferentially annotated. For example, midgut derived RNA seq data will provide better annotation of midgut associated genes such as digestive genes than olfaction genes associated with the antennae.

3. The assembly information would have been improved by information on genome arrangements such as synteny, intergenic sequences, transposon/repetitive sequence expansions, number of exons and their average size and orthologs in relation to closest relative among others.

4. Information on nature of the genes annotated has not been provided. The information would yield important insights with broad implications important aspects of the flesh fly biology. The approach would identify gene families involved in important biological aspects of the fly such as its olfaction, flesh feeding/nutrition, immunity, microbiome, reproduction and developmental biology among other important aspects of this fly

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

Partly

Are the protocols appropriate and is the work technically sound?

Yes

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

Yes

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Medical Entomology, Bioinformatics 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, however I have significant reservations, as outlined above.

Author Response 09 May 2024

Tree of Life Team Sanger

Thank you for your comments on this data note. We have made a variety of changes in response to reviewers' comments. In response to points 2, 3 and 4, we would like to point out that we have given links to an external source of annotation of the genome, provided by the European Bioinformatics Institute. We do not provide the annotation, and do not analyse the annotation in this data note. We do not have access to information such as the source of transcriptomic data used, as we have not presented RNA sequencing for this species.

Competing Interests: No competing interests were disclosed.