DATA NOTE

The genome sequence of the dumpy grass hoverfly, *Melanostoma mellinum* (Linnaeus, 1758) [version 1; peer review: 2 approved]

William Hawkes¹, Karl Wotton 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, Darwin Tree of Life Consortium

¹Centre for Ecology and Conservation, Department of Biosciences, University of Exeter, Penryn, UK

V1 First published: 15 Feb 2022, 7:59
<https://doi.org/10.12688/wellcomeopenres.17615.1>
Latest published: 15 Feb 2022, 7:59
<https://doi.org/10.12688/wellcomeopenres.17615.1>

Abstract

We present a genome assembly from an individual male *Melanostoma mellinum* (the dumpy grass hoverfly; Arthropoda; Insecta; Diptera; Syrphidae). The genome sequence is 731 megabases in span. The majority of the assembly (99.67%) is scaffolded into five chromosomal pseudomolecules, with the X and Y sex chromosomes assembled. The complete mitochondrial genome was also assembled and is 16.1 kilobases in length.

Keywords

Melanostoma mellinum, dumpy grass hoverfly, 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 2022	view	view

1. **Thomas Pauli**, University of Freiburg, Freiburg, Germany
2. **Renata Schama** , Instituto Nacional de Ciência e Tecnologia em Entomologia Molecular-INCT-EM, Rio de Janeiro, Brazil
Fiocruz, Rio de Janeiro, Brazil

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: **Hawkes W:** Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Wotton K:** Investigation, Supervision, 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) and the Darwin Tree of Life Discretionary Award (218328).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Hawkes W *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Hawkes W, Wotton K, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the dumpy grass hoverfly, *Melanostoma mellinum* (Linnaeus, 1758) [version 1; peer review: 2 approved]** Wellcome Open Research 2022, 7:59 <https://doi.org/10.12688/wellcomeopenres.17615.1>

First published: 15 Feb 2022, 7:59 <https://doi.org/10.12688/wellcomeopenres.17615.1>

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Syrphoidea; Syrphidae; Syrphinae; Melanostomini; Melanostoma; *Melanostoma mellinum* (Linnaeus, 1758) (NCBI:txid653684).

Background

Melanostoma mellinum is a small hoverfly with yellow and black markings, perhaps using this aposematic colourway to loosely mimic wasps and so gain protection from predators (Ball & Morris, 2015; van Veen, 2010). These hoverflies have dark faces and dark scutellums. Female *Melanostoma* species have distinctive inverted yellow triangle abdominal markings and can be separated from the other common species *M. scalare* by the narrow dust spots on the otherwise shining frons (Ball & Morris, 2015). Male *M. mellinum* (Figure 1) have square yellow abdominal markings with tergite 2 and 3 being as long as wide, while *M. scalare* males have longer tergites (Ball & Morris, 2015; van Veen, 2010). Colour and morphological variation is high and *M. mellinum* may comprise a species complex (Haarto & Ståhls, 2014). *M. mellinum* is an abundant grassland species in the UK and can be found in greatest numbers during summer, feeding on the pollen of wind pollinated flowers such as grasses, plantain and sedges (Ball & Morris, 2015). The larvae of this species feed on aphids in the leaf litter and ground layer and are incredibly generalist predators, preying on at least 32 different aphid species (Dziok, 2005)

and gall forming Psyllidae (Hodkinson & Flint, 1971). Like some other generalist hoverflies, *M. mellinum* adults display seasonal migratory behaviour (Aubert *et al.*, 1976; Gatter & Schmid, 1990). This species seems to be heavily affected by the *Entomophora* fungal pathogen (Jensen & Eilenberg, 2001) and dead *M. mellinum* individuals can be often found hanging beneath flower heads with fungal spores exuding (WH, personal observation; Figure 1). Migration in this species may function to reduce the negative effects of this fungus on the population by allowing escape from contaminated habitats and by culling infected animals, both phenomena seen in other migrants (see Satterfield *et al.* (2018)). This is the first production of a high quality *M. mellinum* genome and we believe that the sequence described here, generated as part of the Darwin Tree of Life project, will further aid understanding of the biology and ecology of this hoverfly.

Genome sequence report

The genome was sequenced from a single male *M. mellinum* collected from Wytham Great Wood, Oxfordshire (Biological vice-county: Berkshire), UK (latitude 51.769, longitude -1.330) (Figure 1). A total of 26-fold coverage in Pacific Biosciences single-molecule long reads and 55-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 332 missing/misjoins and removed 35 haplotypic duplications, reducing the assembly size by 2.84% and the scaffold number by 74.75%, and increasing the scaffold N50 by 395.90%.

Figure 1. Example images of *Melanostoma mellinum*. Top: female (left) and male (right) examples of *M. mellinum*; photographs provided by WH. Bottom left: examples of dead *M. mellinum* individuals that have succumbed to the *Entomophora* fungal pathogen; photograph provided by WH. Bottom right: the male *M. mellinum* specimen used for genome sequencing; photograph provided by Liam Crowley, University of Oxford.

The final assembly has a total length of 731 Mb in 76 sequence scaffolds with a scaffold N50 of 235 Mb (Table 1). The majority, 99.67%, of the assembly sequence was assigned to five chromosomal-level scaffolds, representing three autosomes (numbered by sequence length), and the X and Y sex chromosomes (Figure 2–Figure 5; Table 2). The assembly has a BUSCO v5.1.2 (Simão *et al.*, 2015) completeness of 96.3% (single 94.8%, duplicated 1.5%) using the diptera_odb10 reference set (n=3285). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

Methods

Sample acquisition, DNA extraction and sequencing

One male *M. mellinum* sample, idMelMell2, was collected from Wytham Great Wood, Oxfordshire, (Biological vice-county:

Berkshire), UK (latitude 51.769, longitude -1.330) by Will Hawkes, University of Exeter on 7 August 2019. The specimen was caught with a net, identified by the same individual, snap-frozen on dry ice and stored using a CoolRack.

DNA was extracted from the head/thorax of idMelMell2 at the Wellcome Sanger Institute (WSI) Scientific Operations core from the whole organism using the Qiagen MagAttract HMW DNA kit, according to the manufacturer's instructions. Pacific Biosciences HiFi circular consensus and 10X Genomics Chromium read cloud sequencing libraries were constructed according to the manufacturers' instructions. Sequencing was performed by the Scientific Operations core at the Wellcome Sanger Institute on Pacific Biosciences SEQUEL II (HiFi) and Illumina HiSeq X (10X) instruments. Hi-C data were generated from abdomen tissue of idMelMell2 using the Arima

Table 1. Genome data for *Melanostoma mellinum*, idMelMell2.1.

Project accession data	
Assembly identifier	idMelMell2.1
Species	<i>Melanostoma mellinum</i>
Specimen	idMelMell2
NCBI taxonomy ID	653684
BioProject	PRJEB46300
BioSample ID	SAMEA7520051
Isolate information	Male, head/thorax (genome assembly), abdomen (Hi-C)
Raw data accessions	
PacificBiosciences SEQUEL II	ERR6636092, ERR6636093
10X Genomics Illumina	ERR6688425-ERR6688428
Hi-C Illumina	ERR6688429
Genome assembly	
Assembly accession	GCA_914767635.1
Accession of alternate haplotype	GCA_914767615.1
Span (Mb)	731
Number of contigs	479
Contig N50 length (Mb)	4.6
Number of scaffolds	76
Scaffold N50 length (Mb)	235
Longest scaffold (Mb)	266
BUSCO* genome score	C:96.3%[S:94.8%,D:1.5%],F:0.9%,M:2.8%,n:3285

*BUSCO scores based on the diptera_odb10 BUSCO set using v5.1.2. 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/idMelMell2.1/dataset/CAJZBU01/busco>.

Figure 2. Genome assembly of *Melanostoma mellinum*, idMelMell2.1: 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 731,040,377 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 (265,601,162 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (235,131,548 and 204,347,527 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/idMelMell2.1/dataset/CAJZBU01/snail>.

v2 Hi-C kit in the Tree of Life laboratory and sequenced at the Scientific Operations core on an Illumina NovaSeq 6000 instrument.

Genome assembly

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021); 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 longranger align, calling variants with freebayes (Garrison & Marth, 2012). The assembly was then scaffolded with Hi-C data (Rao *et al.*, 2014) using SALSA2 (Ghurye *et al.*, 2019). The assembly was checked for contamination as described previously (Howe *et al.*, 2021). Manual curation was performed

Figure 3. Genome assembly of *Melanostoma mellinum*, idMelMell2.1: GC coverage. 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/idMelMell2.1/dataset/CAJZBU01/blob>.

using HiGlass (Kerpedjiev *et al.*, 2018) and Pretext. The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2021), which performs 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.

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, the Darwin Tree of Life Partner agrees they will meet

Figure 4. Genome assembly of *Melanostoma mellinum*, idMelMell2.1: cumulative sequence. 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/idMelMell2.1/dataset/CAJZBU01/cumulative>.

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.

Figure 5. Genome assembly of *Melanostoma mellinum*, idMelMell2.1: Hi-C contact map. Hi-C contact map of the idMelMell2.1 assembly, visualised in HiGlass. Chromosomes are arranged in size order from left to right and top to bottom.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Melanostoma mellinum*, idMelMell2.1.

INSDC accession	Chromosome	Size (Mb)	GC%
OU612058.1	1	265.60	36.7
OU612059.1	2	235.13	37.0
OU612060.1	3	204.35	37.4
OU612061.1	X	20.89	36.7
OU612062.1	Y	1.11	36.8
OU612063.1	MT	0.02	18.7
-	Unplaced	3.94	37.0

Table 3. Software tools used.

Software tool	Version	Source
Hifiasm	0.15.2	Cheng et al., 2021
purge_dups	1.2.3	Guan et al., 2020
SALSA2	2.2	Ghurye et al., 2019
longranger align	2.2.2	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines
freebayes	1.3.1-17-gaa2ace8	Garrison & Marth, 2012
MitoHiFi	2.0	Uliano-Silva et al., 2021
HiGlass	1.11.6	Kerpedjiev et al., 2018
PretextView	0.2.x	https://github.com/wtsi-hpag/PretextView
BlobToolKit	2.6.4	Challis et al., 2020

Data availability

European Nucleotide Archive: *Melanostoma mellinum* (dumpy grass hoverfly). Accession number [PRJEB46300](https://identifiers.org/ena.embl/PRJEB46300); <https://identifiers.org/ena.embl/PRJEB46300>.

The genome sequence is released openly for reuse. The *M. mellinum* genome sequencing initiative is part of the [Darwin Tree of Life](https://doi.org/10.5281/zenodo.5744840) (DTOL) project. All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated and presented through the Ensembl pipeline at the European Bioinformatics Institute. 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.5746938>.

Members of the Darwin Tree of Life Barcoding collective are listed here: <https://doi.org/10.5281/zenodo.5744972>.

Members of the Wellcome Sanger Institute Tree of Life programme are listed here: <https://doi.org/10.5281/zenodo.5744840>.

Members of Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective are listed here: <https://doi.org/10.5281/zenodo.5746904>.

Members of the Tree of Life Core Informatics collective are listed here: <https://doi.org/10.5281/zenodo.5743293>.

Members of the Darwin Tree of Life Consortium are listed here: <https://doi.org/10.5281/zenodo.5638618>.

References

- Allio R, Schomaker-Bastos A, Romiguier J, *et al.*: **MitoFinder: Efficient Automated Large-Scale Extraction of Mitogenomic Data in Target Enrichment Phylogenomics.** *Mol Ecol Resour.* 2020; **20**(4): 892–905. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Aubert J, Aubert JJ, Others: **Douze Ans de Captures Systematiques de Syrphides (Diptères) Au Col de Bretolet (Alpes Valaisannes).** 1976. [Reference Source](#)
- Ball S, Morris R: **Britain's Hoverflies: A Field Guide - Revised and Updated Second Edition.** Princeton University Press. 2015. [Reference Source](#)
- Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit - Interactive Quality Assessment of Genome Assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–74. [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): 170–75. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Dziok F: **Evolution of Prey Specialization in Aphidophagous Syrphids of the Genera *Melanostoma* and *Platycheirus* (Diptera: Syrphidae) 1. Body Size, Development and Prey Traits.** *Eur J Entomol.* 2005; **102**(3): 413–421. [Publisher Full Text](#)
- Garrison E, Marth G: **Haplotype-Based Variant Detection from Short-Read Sequencing.** *arXiv: 1207.3907.* 2012. [Reference Source](#)
- Gatter W, Schmid U: **Hoverfly Migration (Diptera Syrphidae) at Randecker Maar Southwest Germany.** *Spixiana.* 1990; **15**: 1–100.
- Ghurye J, Rhie A, Walenz BP, *et al.*: **Integrating Hi-C Links with Assembly Graphs for Chromosome-Scale Assembly.** *PLoS Comput Biol.* 2019; **15**(8): e1007273. [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): 2896–98. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Haarto A, Ståhls G: **When mtDNA COI Is Misleading: Congruent Signal of ITS2 Molecular Marker and Morphology for North European *Melanostoma Schiner, 1860* (Diptera, Syrphidae).** *Zookeys.* 2014; (431): 93–134. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hodkinson ID, Flint PWH: **Some Predators from the Galls of *Psyllopsis Fraxini* L.(Hem., Psyllidae).** *Entomologist's Monthly Magazine.* 1971; **107**: 11–12. [Reference Source](#)
- Howe K, Chow W, Collins J, *et al.*: **Significantly Improving the Quality of Genome Assemblies through Curation.** *GigaScience.* 2021; **10**(1): g1aa153. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Jensen AB, Eilenberg J: **Genetic Variation within the Insect-Pathogenic Genus *Entomophthora*, Focusing on the *E. Muscae* Complex, Using PCR–RFLP of the ITS II and the LSU rDNA.** *Mycol Res.* 2001; **105**(3): 307–12. [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](#)
- Rao SS, 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): 1665–80. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Satterfield DA, Marra PP, Sillett TS, *et al.*: **Responses of Migratory Species and Their Pathogens to Supplemental Feeding.** *Philos Trans R Soc Lond B Biol Sci.* 2018; **373**(1745): 20170094. [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): 3210–12. [PubMed Abstract](#) | [Publisher Full Text](#)
- Uliano-Silva M, Nunes JGF, Krasheninnikova K, *et al.*: **marcelauliano/MitoHiFi: mitohifi_v2.0.** 2021. [Publisher Full Text](#)
- Van Veen MP: **Hoverflies of Northwest Europe: Identification Keys to the Syrphidae.** BRILL. 2010. [Reference Source](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 12 December 2022

<https://doi.org/10.21956/wellcomeopenres.19486.r53457>

© 2022 Schama R. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Renata Schama

¹ Instituto Nacional de Ciência e Tecnologia em Entomologia Molecular-INCT-EM, Rio de Janeiro, Brazil

² Fiocruz, Rio de Janeiro, Brazil

The Data Note describes how the genome sequence of the dumpy grass hoverfly, *Melanostoma millenium*, was sequenced. It clearly describes all the methodologies used, from collection, DNA extraction, DNA sequencing, and computational methods. Although they did DNA extraction of a whole body part the sequencing technology and posterior analyses were optimal in cleaning up the data and making sure they could build the different chromosomes. All quality metrics were good and the final assembly is well-supported with high completeness. The data is readily accessible, making sure the computational genomics part of the work could be replicated and the final result is freely available for use. Overall this is a clear and well-performed work.

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, and gene family evolution with a focus on arthropod genomes (particularly mosquitoes).

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 16 February 2022

<https://doi.org/10.21956/wellcomeopenres.19486.r48743>

© 2022 Pauli T. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Thomas Pauli

Institute of Medical Bioinformatics and Systems Medicine, Faculty of Medicine, Medical Center-University of Freiburg, University of Freiburg, Freiburg, Germany

The data note article written by Hawkes and Wotton describes the sequencing and subsequent assembly of the *Melanostoma mellinum* genome using long-read sequencing on a PacBio platform.

In my estimation, the authors succeeded in generating a well-documented and accessible dataset. I only have minor criticism regarding the data note's writing style and presentation.

Background

- *Melanostoma mellinum* is a small hoverfly with yellow and black markings, perhaps using this aposematic colourway to loosely mimic wasps and so gain protection from predators

This sentence is odd for the first sentence of a background. (1) Many hoverflies are assumed to mimic wasps not just *M. mellinum*. (2) The word 'perhaps' suggests that this is rather a musing instead of a factual description.

I would (1) prefer to start describing a trait that stands out for this species in particular and (2) move factual descriptions towards the start of the paragraph and include assumptions towards the end of the paragraph.

- *Migration in this species may function to reduce*

"Migration in this species might reduce" would be a better sentence in my opinion.

- *This is the first production of a high quality *M. mellinum* genome and we believe that the sequence described here, generated as part of the [Darwin Tree of Life project](#), will further aid understanding of the biology and ecology of this hoverfly.*

I encourage the authors to elaborate on this. Do you think your dataset will assist researchers in answering specific research questions regarding the biology and ecology of this species? Do you expect (or even observe) duplication or loss events for certain genes or genes of certain gene families in comparison to other closely related taxa?

The public availability of new molecular biologicals dataset is in itself valuable, but I would wish for

a better description of the project's rationale.

Figures

Figure 3 should be edited to make it more accessible:

1. I consider my eyes in good condition but found the labels, legends and scale numbers in this figure too small to read. The figure uses a lot of white space. Changing the font sizes might reduce white space and make the plot more appealing to look at.
2. I prefer figures that are self-explanatory. This figure could be labeled more clearly. Specifically: the x-axis is labeled: gc, a more appropriate label would be "GC content", the y-Axis is labeled: ERR6688428_cov, a more appropriate label would be "Coverage".
3. It took me a while to understand what was meant by the word 'phylum' in the figure description. Firstly, the colors were difficult to distinguish from each other; secondly, the labels in the upper right corner were badly visible; thirdly, the method of using buscogenes taxrule to classify phyla is described in the description of Figure 4. It would be better to mention it in figure 3 as well. Furthermore, I find the term 'phylum' inaccurate here. Since the only classes are 'Arthropoda' and 'no-hit' (and the third class: 'total', which is a composite of 'Arthropoda'+ 'no-hit'.

I am aware that this plot is automatically generated and is supposed to look similar to the plot in the provided interactive version to which you linked. But I suggest to either polish the presentation of this Figure, or move it to the supplement.

You might want to consider some of these suggestions for Figure 4 as well. It looks better than Figure 3, but I found the legend describing the 'phyla' and also the numbers on the scale too small to read.

In summary, I think this data does a good job in sufficiently describing the generation of the associated dataset and I approve of this article with minor revisions.

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: I am a bioinformatician who has primarily worked with Illumina sequencing

data of Arthropods before (primarily Hymenoptera, but I have also worked with sequencing data from hover flies). More recently, my professional interests shifted towards biomedical informatics. I analyze the Illumina sequencing data of tumor biopsies from human patients. I have theoretical, but no practical experience with long-read sequencing data and while I'm experienced in bioinformatics, I cannot evaluate the laboratory protocols used.

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.