Check for updates

DATA NOTE

# The genome sequence of the Marbled White Spot, *Protodeltote pygarga* (Hufnagel, 1766)

[version 1; peer review: 2 approved, 2 approved with reservations]

Douglas Boyes[1+], Owen T. Lewis [2],
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]UK Centre for Ecology & Hydrology, Wallingford, England, UK
[2]University of Oxford, Oxford, England, UK

[+] Deceased author

## Abstract

We present a genome assembly from an individual male *Protodeltote pygarga* (the Marbled White Spot; Arthropoda; Insecta; Lepidoptera; Noctuidae). The genome sequence is 421.1 megabases in span. Most of the assembly is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled and is 15.48 kilobases in length. Gene annotation of this assembly on Ensembl identified 17,784 protein coding genes.

## Keywords

Protodeltote pygarga, Marbled White Spot, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

## Open Peer Review

**Approval Status** ? ✔ ? ✔

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **version 1**<br>08 Jan 2024 | ?<br>view | ✔<br>view | ?<br>view | ✔<br>view |

1. **Balaji Chattopadhyay**, Ashoka University, Sonipat, India

2. **Alex Makunin** [iD], Wellcome Sanger Institute, Hinxton, UK

3. **Maria Nilsson** [iD], Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany

4. **Tyler Alioto** [iD], Fundacion Centro Nacional de Analisi Genomico, Barcelona, Spain

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: Boyes D**: Investigation, Resources; **Lewis OT**: Writing – Original Draft Preparation;

**How to cite this article:** Boyes D, Lewis OT, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Marbled White Spot, *Protodeltote pygarga* (Hufnagel, 1766) [version 1; peer review: 2 approved, 2 approved with reservations]** Wellcome Open Research 2024, **9**:7 https://doi.org/10.12688/wellcomeopenres.20571.1

**First published:** 08 Jan 2024, **9**:7 https://doi.org/10.12688/wellcomeopenres.20571.1

## Species taxonomy

Eukaryota; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Amphiesmenoptera; Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Noctuoidea; Noctuidae; Eustrotiinae; *Protodeltote*; *Protodeltote pygarga* (Hufnagel, 1766) (NCBI:txid708063).

## Background

The Marbled White Spot (*Protodeltote pygarga*) is a small noctuid moth. This species has undergone a marked range expansion in Britain and Ireland in recent decades. It now occurs widely in grassland, heathland, woodland and moorland habitats, but is absent from Scotland and northern England (Randle *et al.*, 2019). Its distribution extends across much of Europe to western and central Asia, and it is also recorded from Japan, Korea and China (GBIF Secretariat, 2023).

The larval host plants are grasses. This species has a single annual generation and overwinters as a pupa underground (Waring *et al.*, 2017).

The genome of the marbled white spot, *Protodeltote pygarga*, 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 *Protodeltote pygarga*, based on one male specimen from Wytham Woods, Oxfordshire.

## Genome sequence report

The genome was sequenced from one male *Protodeltote pygarga* (Figure 1) collected from Wytham Woods, Oxfordshire, UK (51.77, –1.34). A total of 92-fold coverage in Pacific Biosciences single-molecule HiFi long reads was generated. Primary assembly contigs were scaffolded with chromosome



**Figure 1. Photograph of the *Protodeltote pygarga* (ilProPyga1) specimen used for genome sequencing.**

conformation Hi-C data. Manual assembly curation corrected 7 missing joins or mis-joins and removed 2 haplotypic duplications, reducing the assembly length by 0.45%.

The final assembly has a total length of 421.1 Mb in 33 sequence scaffolds with a scaffold N50 of 14.9 Mb (Table 1). The snailplot in Figure 2 provides a summary of the assembly statistics, while the distribution of reads 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 (99.98 %) of the assembly sequence was assigned to 31 chromosomal-level scaffolds, representing 30 autosomes and the Z 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 69.4 with *k*-mer completeness of 100%, and the assembly has a BUSCO v5.3.2 completeness of 99.0% (single = 98.8%, duplicated = 0.2%), using the lepidoptera_odb10 reference set (*n* = 5,286).

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/708063.

## Genome annotation report

The *Protodeltote pygarga* genome assembly (GCA_936450705.2) was annotated using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Protodeltote_pygarga_GCA_936450705.2/Info/Index). The resulting annotation includes 17,784 transcribed mRNAs from 17,968 protein-coding genes.

## Methods

### Sample acquisition and nucleic acid extraction

A male *Protodeltote pygarga* (specimen ID Ox000464, ToLID ilProPyga1) was collected from Wytham Woods, Oxfordshire (biological vice-county Berkshire), UK (latitude 51.77, longitude –1.34) on 2020-06-13, using a light trap. The specimen was collected and identified by Douglas Boyes (University of Oxford) and preserved on dry ice.

High molecular weight (HMW) DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute (WSI), following a sequence of core procedures: sample preparation; sample homogenisation; HMW DNA extraction; DNA fragmentation; and DNA clean-up. In sample preparation, the ilProPyga1 sample was weighed and dissected on dry ice (Jay *et al.*, 2023). Tissue from the abdomen was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a). HMW DNA was extracted using the Automated MagAttract v1 protocol (Sheerin *et al.*, 2023). DNA was sheared

**Table 1. Genome data for *Protodeltote pygarga*, ilProPyga1.2.**

| Project accession data | | |
|---|---|---|
| Assembly identifier | ilProPyga1.2 | |
| Species | *Protodeltote pygarga* | |
| Specimen | ilProPyga1 | |
| NCBI taxonomy ID | 708063 | |
| BioProject | PRJEB51452 | |
| BioSample ID | SAMEA7631560 | |
| Isolate information | ilProPyga1, male: thorax (Hi-C data) | |
| **Assembly metrics*** | | ***Benchmark*** |
| Consensus quality (QV) | 69.4 | *≥50* |
| *k*-mer completeness | 100% | *≥95%* |
| BUSCO** | C:99.0%[S:98.8%,D:0.2%], F:0.3%,M:0.7%,n:5,286 | *C ≥95%* |
| Percentage of assembly mapped to chromosomes | 99.98% | *≥95%* |
| Sex chromosomes | Z chromosome | *localised homologous pairs* |
| Organelles | Mitochondrial genome assembled | *complete single alleles* |
| **Raw data accessions** | | |
| PacificBiosciences SEQUEL II | ERR9284042, ERR9284043 | |
| Hi-C Illumina | ERR9248439 | |
| **Genome assembly** | | |
| Assembly accession | GCA_936450705.2 | |
| *Accession of alternate haplotype* | GCA_936440985.2 | |
| Span (Mb) | 421.1 | |
| Number of contigs | 44 | |
| Contig N50 length (Mb) | 14.7 | |
| Number of scaffolds | 33 | |
| Scaffold N50 length (Mb) | 14.9 | |
| Longest scaffold (Mb) | 21.8 | |
| **Genome annotation** | | |
| Number of protein-coding genes | 17,784 | |
| Number of gene transcripts | 17,968 | |

\* 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 lepidoptera_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/Protodeltote%20pygarga/dataset/CAMPPH01/busco.

into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 30 (Todorovic *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
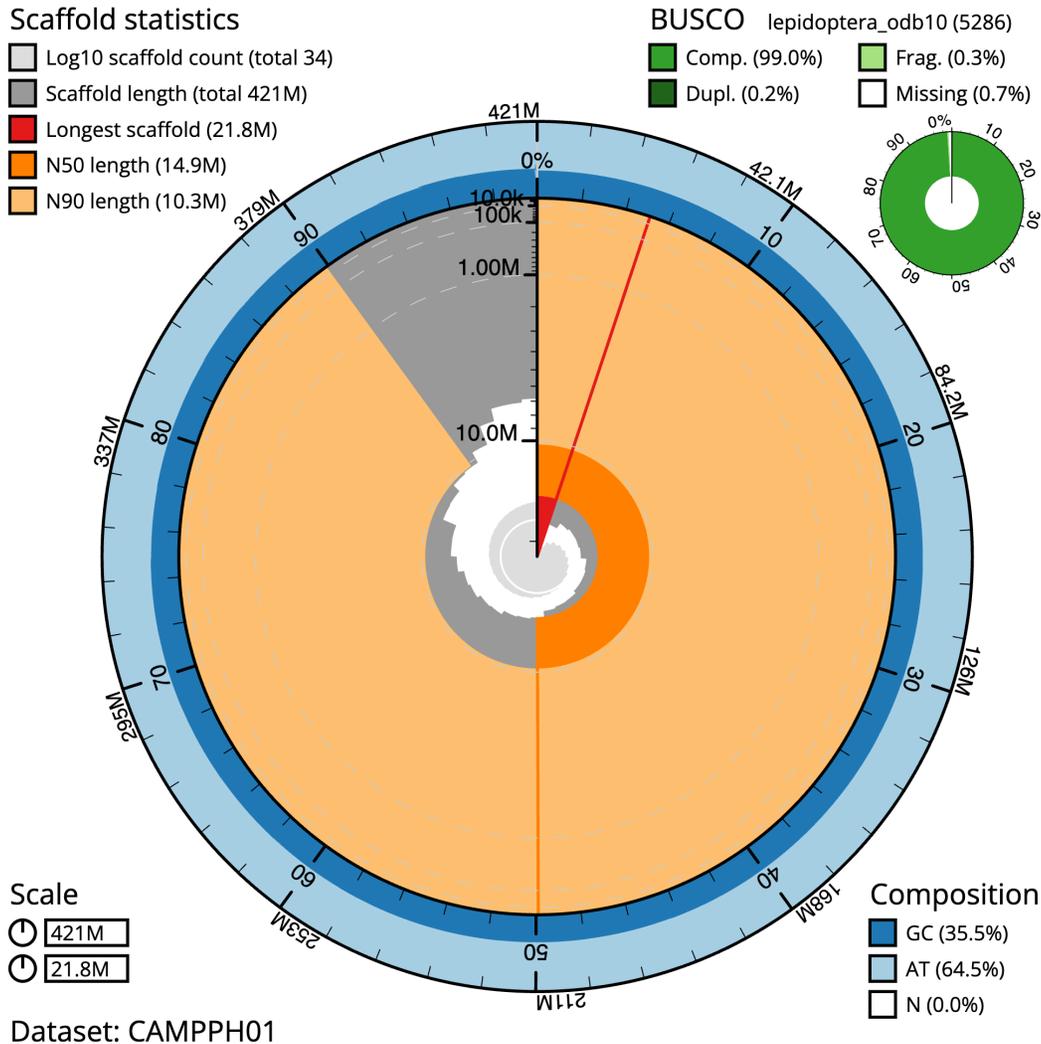
**Figure 2. Genome assembly of *Protodeltote pygarga*, ilProPyga1.2: metrics.** The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 421,153,487 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (21,803,012 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (14,901,961 and 10,302,340 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 lepidoptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/Protodeltote%20pygarga/dataset/CAMPPH01/snail.

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

Protocols developed by the Wellcome Sanger Institute (WSI) Tree of Life core laboratory have been deposited on protocols.io (Denton *et al.*, 2023b).

### Sequencing
Pacific Biosciences HiFi circular consensus 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 thorax tissue of ilProPyga1 using the Arima2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

### 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
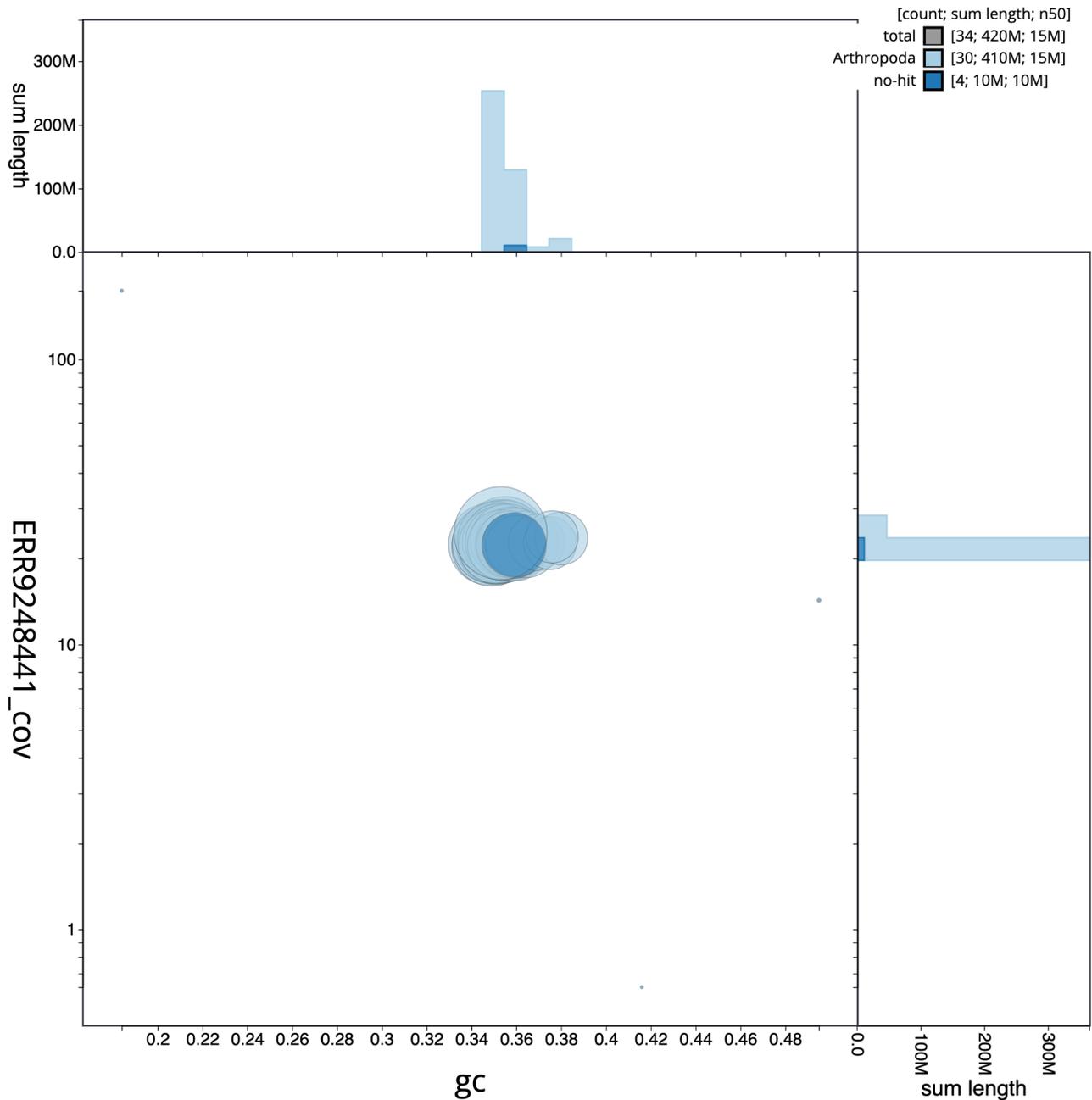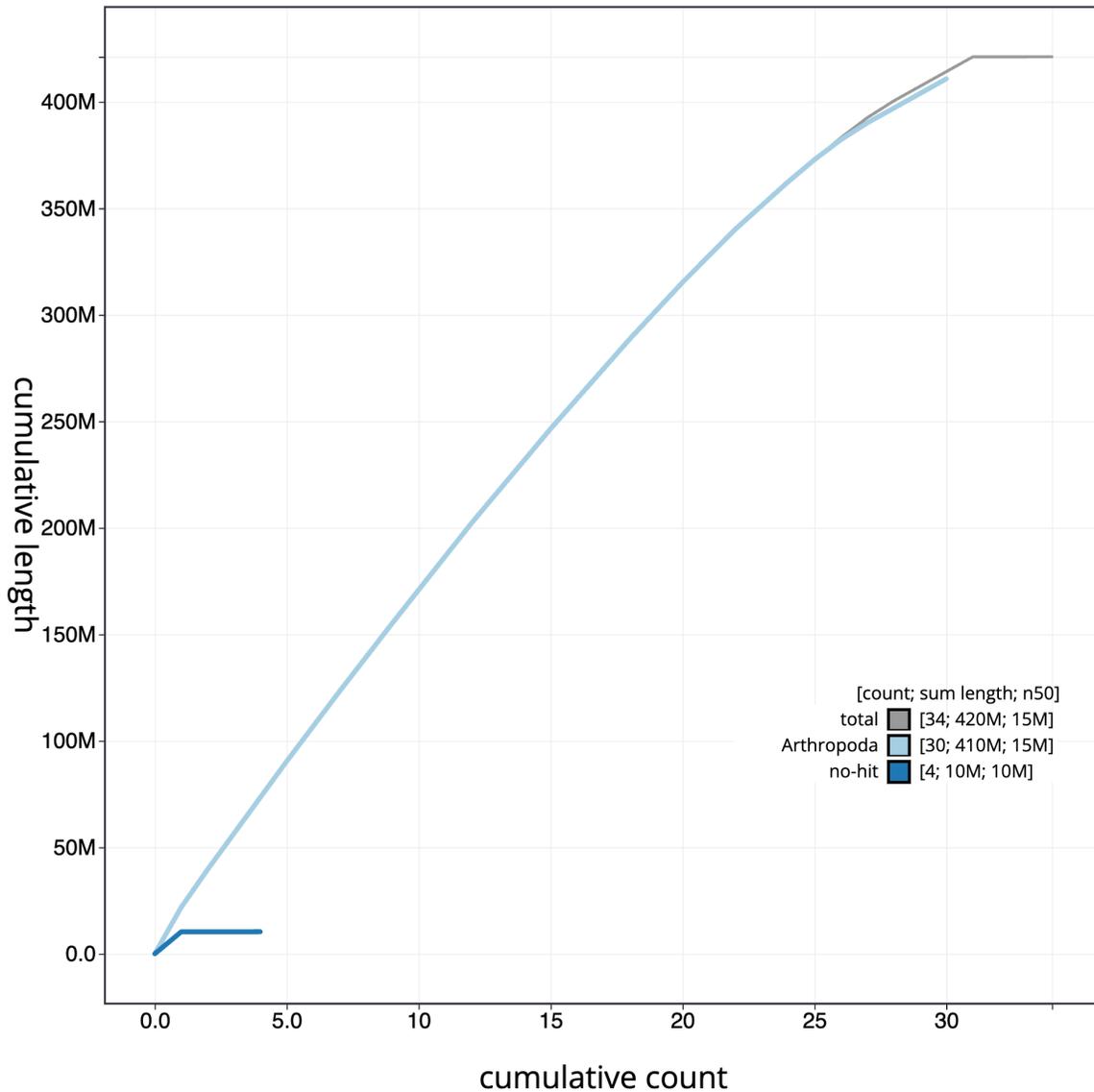
**Figure 3. Genome assembly of *Protodeltote pygarga*, ilProPyga1.2: BlobToolKit GC-coverage plot.** Scaffolds are coloured by phylum. Circles are sized in proportion to sequence length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/Protodeltote%20pygarga/dataset/CAMPPH01/blob.

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

**Figure 4. Genome assembly of *Protodeltote pygarga*, ilProPyga1.2: 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/Protodeltote%20pygarga/dataset/CAMPPH01/cumulative.

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.

### Genome annotation
The BRAKER2 pipeline (Brůna *et al.*, 2021) was used in the default protein mode to generate annotation for the *Protodeltote*

*pygarga* assembly (GCA_936450705.2) in Ensembl Rapid Release.

### 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
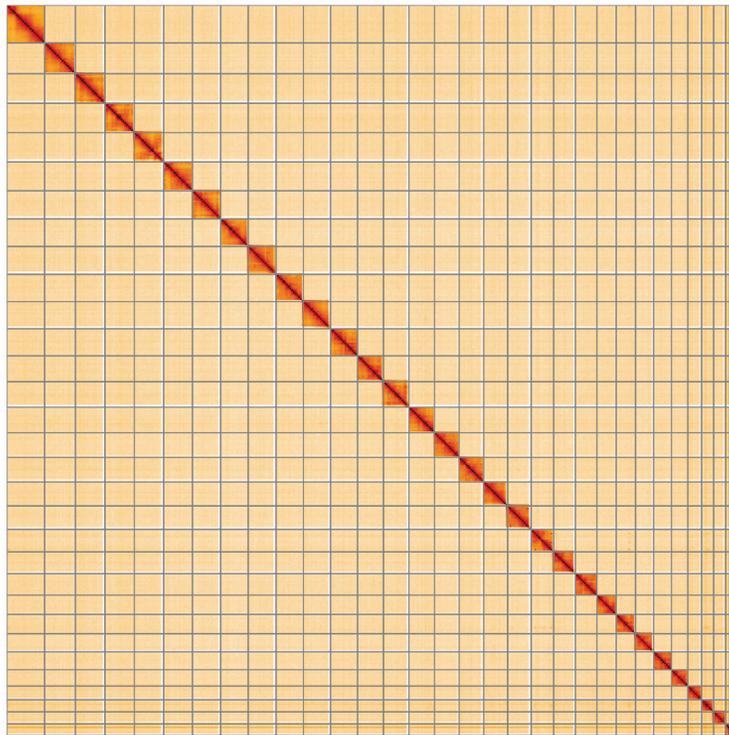
**Figure 5. Genome assembly of *Protodeltote pygarga*, ilProPyga1.2: Hi-C contact map of the ilProPyga1.2 assembly, visualised using HiGlass.** Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=NRR27qs_Twah2X0btFEQTQ.

**Table 2. Chromosomal pseudomolecules in the genome assembly of *Protodeltote pygarga*, ilProPyga1.**

| INSDC accession | Chromosome | Length (Mb) | GC% |
|---|---|---|---|
| OW388249.2 | 1 | 17.74 | 35.0 |
| OW388250.2 | 2 | 17.08 | 35.5 |
| OW388251.2 | 3 | 16.72 | 35.0 |
| OW388252.2 | 4 | 16.89 | 35.5 |
| OW388253.2 | 5 | 16.86 | 35.5 |
| OW388254.2 | 6 | 16.11 | 35.0 |
| OW388255.2 | 7 | 15.99 | 35.0 |
| OW388256.2 | 8 | 16.0 | 35.0 |
| OW388257.2 | 9 | 15.73 | 35.0 |
| OW388258.2 | 10 | 15.69 | 35.5 |
| OW388259.2 | 11 | 15.53 | 35.0 |
| OW388260.2 | 12 | 14.9 | 35.0 |
| OW388261.2 | 13 | 14.78 | 35.5 |
| OW388262.2 | 14 | 14.66 | 35.5 |
| OW388263.2 | 15 | 14.25 | 35.0 |
| OW388264.2 | 16 | 14.14 | 35.5 |
| OW388265.2 | 17 | 13.78 | 35.5 |
| OW388266.2 | 18 | 13.58 | 36.0 |
| OW388267.2 | 19 | 13.02 | 36.0 |
| OW388268.2 | 20 | 12.48 | 35.5 |
| OW388269.2 | 21 | 12.43 | 36.0 |
| OW388270.2 | 22 | 11.16 | 35.5 |
| OW388271.2 | 23 | 11.11 | 36.5 |
| OW388272.2 | 24 | 10.43 | 36.0 |
| OW388273.2 | 25 | 10.3 | 36.0 |
| OW388274.2 | 26 | 9.37 | 36.0 |
| OW388275.2 | 27 | 7.87 | 37.0 |
| OW388276.2 | 28 | 6.97 | 37.5 |
| OW388277.2 | 29 | 6.98 | 38.0 |
| OW388278.2 | 30 | 6.76 | 37.5 |
| OW388248.2 | Z | 21.8 | 35.5 |
| OW388279.2 | MT | 0.02 | 18.5 |

**Table 3. Software tools: versions and sources.**

| Software tool | Version | Source |
|---|---|---|
| BlobToolKit | 4.0.7 | 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 |

of all samples acquired for, and supplied to, the Darwin Tree of Life Project.

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

- Ethical review of provenance and sourcing of the material

- Legality of collection, transfer and use (national and international)

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

## Data availability

European Nucleotide Archive: *Protodeltote pygarga* (marbled white spot). Accession number PRJEB51452; https://identifiers.org/ena.embl/PRJEB51452 (Wellcome Sanger Institute, 2022). The genome sequence is released openly for reuse. The *Protodeltote pygarga* 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**

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**

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**

Brůna T, Hoff KJ, Lomsadze A: **BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database.** *NAR Genom Bioinform.* 2021; **3**(1): lqaa108.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit - interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

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**

GBIF Secretariat: **Protodeltote pygarga (Hufnagel, 1766).** *gbif.org.* 2023; [Accessed 30 October 2023].
**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**

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**

Randle Z, Evans-Hill LJ, Parsons MS, *et al.*: **Atlas of Britain & Ireland's Larger Moths.** Newbury: NatureBureau, 2019.
**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**

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**

Sheerin E, Sampaio F, Oatley G, *et al.*: **Sanger Tree of Life HMW DNA Extraction: Automated MagAttract v.1.** *protocols.io.* 2023; [Accessed 21 November 2023].
**Publisher 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**

Surana P, Muffato M, Qi G: **sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo.* 2023a; [Accessed 21 July 2023].
**Publisher Full Text**

Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev).** *Zenodo.* 2023b; [Accessed 21 July 2023].
**Publisher Full Text**

Todorovic M, Sampaio F, Howard C: **Sanger Tree of Life HMW DNA Fragmentation: Diagenode Megaruptor®3 for PacBio HiFi.** *protocols.io.* 2023.
**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**

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.
**Publisher Full Text**

Waring P, Townsend M, Lewington R: **Field Guide to the Moths of Great Britain and Ireland: Third Edition.** Bloomsbury Wildlife Guides, 2017.
**Reference Source**

Wellcome Sanger Institute: **The genome sequence of the Marbled White Spot, *Protodeltote pygarga* (Hufnagel, 1766).** European Nucleotide Archive. [dataset], accession number PRJEB51452, 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 01 August 2024

https://doi.org/10.21956/wellcomeopenres.22770.r78591

✔ **Tyler Alioto** (iD)

Fundacion Centro Nacional de Analisi Genomico, Barcelona, Catalonia, Spain

The genome note submitted by Boyes *et al.* reports the genome sequence of the Marbled White Spot, Proto Deltore pygarga. P. pygarga is a small moth that has increased its range across southern Britain and Ireland. It is distributed widely in Europe and Asia as well. It has a single annual generation, pupating underground over winter and the larval host plants are grasses. The male specimen sequenced was collected from Wytham Woods. 92-fold coverage of HiFi reads were generated and assembled with Hifiasm, run through purge_dups, scaffolded with 141 Gb (335x) of Arima2 Hi-C data using YaHS, and manually curated with few interventions necessary. The assembly is chromosome-scale and of high quality, meeting the minimum standards recommended by the Earth Biogenome Project. All protocols are appropriate and well-documented. All data conforms with FAIR principles, with read data and assemblies being available in the ENA. Blob toolkit figures are interactive. Additional QC data are provided at https://tolqc.cog.sanger.ac.uk/, supplementing the figures published in the data note.

No major comments.

Minor comments:
1. A ZZ-male was collected and sequenced. I believe it is usually standard practice to try to sequence the heterogametic sex. The authors should provide an explanation for why they chose a ZZ male. Were there difficulties in collecting females? Are ZW females too hard to assemble? It would help the rest of the reference genome sequencing community to know the rationale for choices made or be made aware of difficulties encountered.

2. A general comment on the genome-note engine. I would like to see the Table 1 comparison to the now-old VGP standards "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)." changed to a comparison to the latest EBP metric standards. Especially for non-vertebrate genomes.

3. Is there a way to indicate what protein set was used to annotate the genome with BRAKER2?

In order to reproduce the annotation, this information would seem necessary. Please find a way to include such info (from Ensembl rapid) in the genome note template.

**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, genome assembly, gene prediction, annotation

**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 11 June 2024

https://doi.org/10.21956/wellcomeopenres.22770.r78595

**Maria Nilsson**

Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany

The genome of the Marbled White Spot (*Protodeltote pygarga*) has been sequenced from a single male individual captured using a light trap. The Marbled White Spot is a commonly occurring moth that is found throughout Europe and has seen a range expansion in Britain. The genome assembly is 421 Mb large and is scaffolded to 33 scaffolds. The genome could not be fully phased. The QV value is 69.4 and the BUSCO gene completeness at 99.0, which indicate a very complete assembly. The contamination screen found minimal contamination and gene annotation identified 17,784 genes.

Minor comments:
1) The mitochondrial genome has been assembled, but seems to not be extracted/submitted separately. It would be helpful to have the mt genome submitted separate. It would be beneficial for those that are interested in the mitochondrial genome.
2) Please clarify for what analyzes that the abdomen DNA was used for. The material and methods describe DNA isolation from the abdomen, while all of a sudden thorax DNA is used for HiC. This

contradiction should be clarified.
3)The data set used for the blob plot in Figure 2, ERR9248441_cov, is not listed anywhere in table 1.

Conclusion: the high quality genome assembly of the common Marbled White Spot offers possibilities to study the range expansion of evolutionary properties of the moth family Noctuidae.

**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:* phylogenomics, transposable elements, mitogenomics

**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 07 May 2024

https://doi.org/10.21956/wellcomeopenres.22770.r78587

✔  **Alex Makunin** iD

Wellcome Sanger Institute, Hinxton, England, UK

The manuscript by Boyes et al present the chromosomal genome assembly for Protodeltote pygarga moth species. This study is seemingly filling a gap in taxonomic sampling by providing the first reference genome assembly for *Eustrotiinae* subfamily of *Noctuidae* family.

Overall, the genome note is cohesive and well structured. I can see two minor issues:
- DNA extraction is described in detail only for abdomen, which seemingly was used for PacBio sequencing, but not for thorax which was used for HI-C.
- The reasoning for Z chromosome assignment in a homogametic sample is not provided - was it

based on size (largest chromosome)?

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

Yes

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

Yes

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

Yes

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

Yes

*Competing Interests:* No competing interests were disclosed.

*Reviewer Expertise:* comparative genomics

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

Reviewer Report 24 April 2024

https://doi.org/10.21956/wellcomeopenres.22770.r78593

?

**Balaji Chattopadhyay**

Ashoka University, Sonipat, Haryana, India

The authors present a high quality genome assembly of an individual male of the noctuid moth Marbled White spot. The authors performed long-read and Hi-C sequencing and assembly to reconstruct chromosomal pseudomolecules of both nuclear (autosomes and sex chromosomes) and mitochondrial genomes.
While the overall methodology to generate and analyze the data is robust, I feel that the authors should mention:
1. details of each method
2. rationale for using specific methods
For scientists to follow a similar methodology it is important that the advantages and limitation of the methods are clearly stated. Similarly, for each methods it is also quite important to state the parameters that were used and the reason behind using these particular set of parameters.

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

Yes

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

Yes

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

Partly

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

Yes

*Competing Interests:* No competing interests were disclosed.

*Reviewer Expertise:* Genomics, Conservation genomics, Biogeography

**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.**