



DATA NOTE

The genome sequence of the Knapweed Pearl, *Paratalanta hyalinalis* (Hübner, 1796) (Lepidoptera: Crambidae)

[version 1; peer review: 2 approved]

David C. Lees ¹, Natural History Museum Genome Acquisition Lab, 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

¹Natural History Museum, London, England, UK

v1 First published: 06 Oct 2025, 10:548
<https://doi.org/10.12688/wellcomeopenres.24962.1>
Latest published: 06 Oct 2025, 10:548
<https://doi.org/10.12688/wellcomeopenres.24962.1>

Abstract

We present a genome assembly from an individual female *Paratalanta hyalinalis* (Knapweed Pearl; Arthropoda; Insecta; Lepidoptera; Crambidae). The assembly contains two haplotypes with total lengths of 529.45 megabases and 506.05 megabases. Most of haplotype 1 (99.95%) is scaffolded into 30 chromosomal pseudomolecules, including the Z sex chromosome, while most of haplotype 2 (99.85%) is scaffolded into 29 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 15.2 kilobases. This assembly was generated as part of the Darwin Tree of Life project, which produces reference genomes for eukaryotic species found in Britain and Ireland.

Keywords

Paratalanta hyalinalis; Knapweed Pearl; genome sequence; chromosomal; Lepidoptera



This article is included in the [Tree of Life](#) gateway.

Open Peer Review

Approval Status  

	1	2
version 1 06 Oct 2025	 view	 view
1. Yvonne Jing Mei Liew , Universiti Malaya, Kuala Lumpur, Malaysia		
2. João Pimenta  , CIBIO-InBIO, Research Centre in Biodiversity and Genetic Resources, Porto, Portugal		

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: Lees DC: Investigation, Resources, 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 (220540) 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: © 2025 Lees DC *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: Lees DC, Natural History Museum Genome Acquisition Lab, Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team *et al.* **The genome sequence of the Knapweed Pearl, *Paratalanta hyalinalis* (Hübner, 1796) (Lepidoptera: Crambidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, 10:548 <https://doi.org/10.12688/wellcomeopenres.24962.1>

First published: 06 Oct 2025, 10:548 <https://doi.org/10.12688/wellcomeopenres.24962.1>

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Amphiesmenoptera; Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Pyraloidea; Crambidae; Pyraustinae; *Paratalanta*; *Paratalanta hyalinalis* (Hübner, 1796) (NCBI:txid1594322)

Background

Paratalanta hyalinalis (Hübner, 1796), also known as the Knapweed Pearl, is a moth in the family Crambidae with a forewing length of 14–16 mm (Sterling *et al.*, 2023) or about 28–35 mm wingspan (Parsons & Clancy, 2023). The wings are light yellow with a faint gloss, with light brown crosslines which are faintly marked, including one outcurved around the outer reniform stigma of the forewing, and an inner convex one. In the Bordered Pearl *P. pandalis* (Hübner, (1825)), the subterminal fascia of the forewing is more distinct, whilst the confusable Lesser Pearl *Sitochroa verticalis* (L.) has blackish ventral markings.

The moth has a flight period from June to July; adults are readily disturbed by day and attracted to light at night or found at flowers (Parsons & Clancy, 2023; Sterling *et al.*, 2023). The larva feeds on *Centaurea* spp. (Asteraceae) including Black Knapweed *C. nigra* L.; Barrett (1904: 191–192) states that the first instar eats small holes either side of the midrib within a silken web, and feeds from August to June mostly in a gallery towards the stem, and hibernates in a whitish cocoon. The larva is also reputed to eat Great Mullein *Verbascum thapsus* L. and Stinging Nettle *Urtica dioica* L. but such reports traceable to Carl Reutti in 1898 are more likely due to misidentifications of the moth (Lepiforum, 2025).

Restricted to south-east England (40 records on NBN Gateway, 2025), the Knapweed Pearl is an uncommon species in the United Kingdom, also occurring in the Channel Islands, not only localised but with indications of decline, and is graded as Nationally Scarce B (Parsons & Clancy, 2023). It is confined to chalk substrates, including downland and clearings in woodland (Sterling *et al.*, 2023). In the Palearctic the species ranges from Great Britain (excluding Ireland) and western Europe as far east as central Asia and from southern Scandinavia south to the shores of the Mediterranean (GBIF Secretariat, 2025).

The DNA barcode from the *P. hyalinalis* mitogenome assembly (OZ204433.1) represents the COI-5P cluster Barcode Index Number (BIN) BOLD:AAF7132 and is identical to the most common ha plotype on BOLD (31/07/2025). Nearest neighbours which are 4.3–5.7% *p*-distant are *Paratalanta 'maximalis'*, (BOLD:AAL6047), *P. ussuralis* (Bremer, 1864) (BOLD:AAL6046, BOLD:ACE5889), and *P. cultralis* (Staudinger, 1867) (BOLD:AAL6479). However, a single representative of another BIN from Turkey (BOLD:ADC6940), also identified as *P. hyalinalis*, is 3.07% *p*-distant to the barcode of the *P. hyalinalis* assembly. *Paratalanta* Meyrick, 1890 is a genus of the tribe

Pyraustini with nine described species (Mally *et al.*, 2019) which seems not to have been included in phylogenetic studies to date, a purpose for which this genome should prove useful. The most closely related genome released so far is probably that of *Anania crocealis* (Lees *et al.*, 2023) whose DNA barcode region is about 7% *p*-distant.

The assembly presented here was produced using the Tree of Life pipeline from a *P. hyalinalis*, specimen collected in Trottscliffe Bank, England, United Kingdom (Figure 1), as part of the Darwin Tree of Life project.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult female *Paratalanta hyalinalis* (specimen ID NHMUK014584864, ToLID ilParHyal1; Figure 1), collected from Trottscliffe Bank, England, United Kingdom (latitude 51.32, longitude 0.34) on 2022-06-03. The specimen was collected and identified by David Lees (Natural History Museum London). Sample metadata were collected in line with the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

The initial identification was verified by an additional DNA barcoding process according to the framework developed by Twyford *et al.* (2024). A small sample was dissected from the specimen and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI) (see the protocol). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the



Figure 1. Photograph of the *Paratalanta hyalinalis* (ilParHyal1) specimen used for genome sequencing.

WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding are available on protocols.io.

Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The ilParHyal1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the head and thorax was homogenised by [powermashing](https://protocols.io) using a PowerMasher II tissue disruptor. HMW DNA was extracted in the WSI Scientific Operations core using the [Automated MagAttract v2](https://protocols.io) protocol. DNA was sheared into an average fragment size of 12–20 kb following the [Megaruptor®3 for LI PacBio](https://protocols.io) protocol. Sheared DNA was purified by [automated SPRI](https://protocols.io) (solid-phase reversible immobilisation). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 24 ng/μL and a yield of 1 128.00 ng, with a fragment size of 15.2 kb. The 260/280 spectrophotometric ratio was 1.95, and the 260/230 ratio was 2.37.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), following the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

The sample was sequenced on a Revio instrument (Pacific Biosciences). The prepared library was normalised to 2 nM, and 15 μL was used for making complexes. Primers were annealed and polymerases bound to generate circularised complexes, following the manufacturer's instructions. Complexes were purified using 1.2X SMRTbell beads, then diluted to the Revio loading concentration (200–300 pM) and spiked with a Revio sequencing internal control. The sample was sequenced on a Revio 25M SMRT cell. The SMRT Link software (Pacific Biosciences), a web-based workflow manager, was used to configure and monitor the run and to carry out primary and secondary data analysis.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen head and thorax tissue of the ilParHyal1 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's

instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagenode Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRIselect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (ThermoFisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

Biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size selected to 400–600 bp using SPRIselect beads. DNA was enriched with Arima-HiC v2 kit Enrichment beads. End repair, A-tailing, and adapter ligation were carried out with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs), following a modified protocol where library preparation occurs while DNA remains bound to the Enrichment beads. Library amplification was performed using KAPA HiFi HotStart mix and a custom Unique Dual Index (UDI) barcode set (Integrated DNA Technologies). Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, libraries were amplified with 10–16 PCR cycles. Post-PCR clean-up was performed with SPRIselect beads. Libraries were quantified using the AccuClear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech).

Prior to sequencing, libraries were normalised to 10 ng/μL. Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq X.

Genome assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts (*k* = 31) was generated from the filtered reads using [FastK](https://github.com/raeser/FastK). GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were assembled using Hifiasm in Hi-C phasing mode (Cheng *et al.*, 2021; Cheng *et al.*, 2022), producing two haplotypes. Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). Contigs were further scaffolded with Hi-C data in YaHS (Zhou *et al.*, 2023), using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

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

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included 12 breaks and 54 joins. The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation>. PretextViewSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

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

The genome was analysed using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified from the NCBI Taxonomy (Schoch *et al.*, 2020). For the three domain-level lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the

density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all outputs into a blobdir for visualisation. The BlobToolKit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), with containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

PacBio sequencing of the *Paratantalant hyalinalis* specimen generated 29.54 Gb (gigabases) from 3.27 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 534.34 Mb, with a heterozygosity of 1.15% and repeat content of 28.07% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 54× coverage. Hi-C sequencing produced 122.75 Gb from 812.95 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The genome was assembled into two haplotypes using Hi-C phasing. For haplotype 1, the final assembly has a total length of 529.45 Mb in 44 scaffolds, with 101 gaps, and a scaffold N50 of 19.94 Mb, while haplotype 2 has a total length of 506.05 Mb in 67 scaffolds, with a scaffold N50 of 19.83 Mb (Table 2).

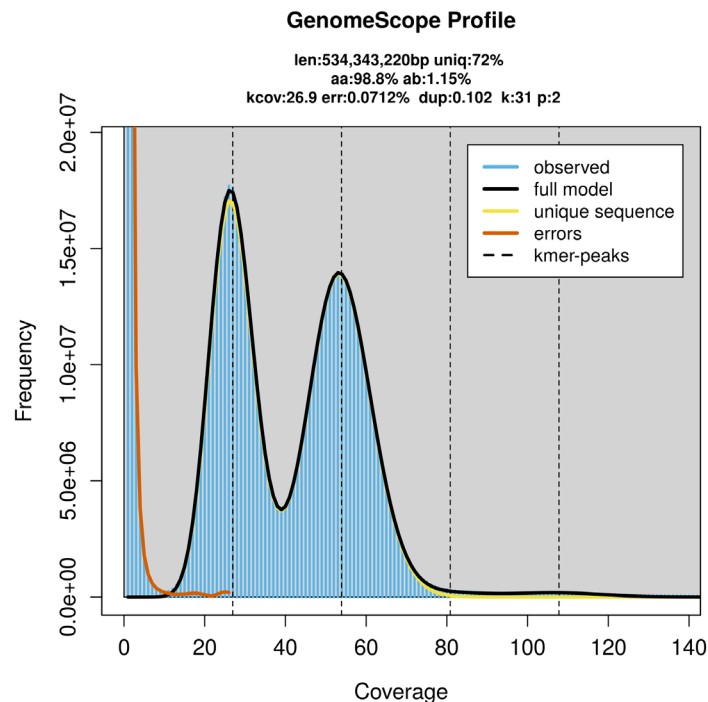


Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB82159.

Platform	PacBio HiFi	Hi-C
ToLID	ilParHyal1	ilParHyal1
Specimen ID	NHMUK014584864	NHMUK014584864
BioSample (source individual)	SAMEA115574494	SAMEA115574494
BioSample (tissue)	SAMEA115599495	SAMEA115599495
Tissue	head and thorax	head and thorax
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13946457	ERR13947476
Read count total	3.27 million	812.95 million
Base count total	29.54 Gb	122.75 Gb

Table 2. Genome assembly statistics.

Assembly name	ilParHyal1.hap1.1	ilParHyal1.hap2.1
Assembly accession	GCA_964340875.1	GCA_964340925.1
Assembly level	chromosome	chromosome
Span (Mb)	529.45	506.05
Number of chromosomes	30	29
Number of contigs	145	163
Contig N50	8.51 Mb	7.22 Mb
Number of scaffolds	44	67
Scaffold N50	19.94 Mb	19.83 Mb
Longest scaffold length (Mb)	29.92	23.86
Sex chromosomes	Z	N/A
Organelles	Mitochondrion: 15.2 kb	N/A

For haplotype 1, most of the assembly sequence (99.95%) was assigned to 30 chromosomal-level scaffolds, representing 29 autosomes and the Z sex chromosome. Haplotype 2 was also curated to chromosome level. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). The Z chromosome was identified based BUSCO gene painting with ancestral Merian elements (Wright *et al.*, 2024). No W chromosome could be identified and the species appears to be a ZO male. During curation we also observed that the exact order and orientation of the contigs on chromosome 27 (4 150–4 750 Kbp) are unknown.

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

For haplotype 1, the estimated QV is 67.3, and for haplotype 2, 68.2. When the two haplotypes are combined, the assembly

achieves an estimated QV of 67.7. The *k*-mer completeness is 76.22% for haplotype 1, 73.12% for haplotype 2, and 98.27% for the combined haplotypes (Figure 4).

BUSCO analysis using the lepidoptera_odb10 reference set (*n* = 5 286) identified 94.1% of the expected gene set (single = 93.9%, duplicated = 0.2%) for haplotype 1. The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.Q67**, meeting the recommended reference standard.

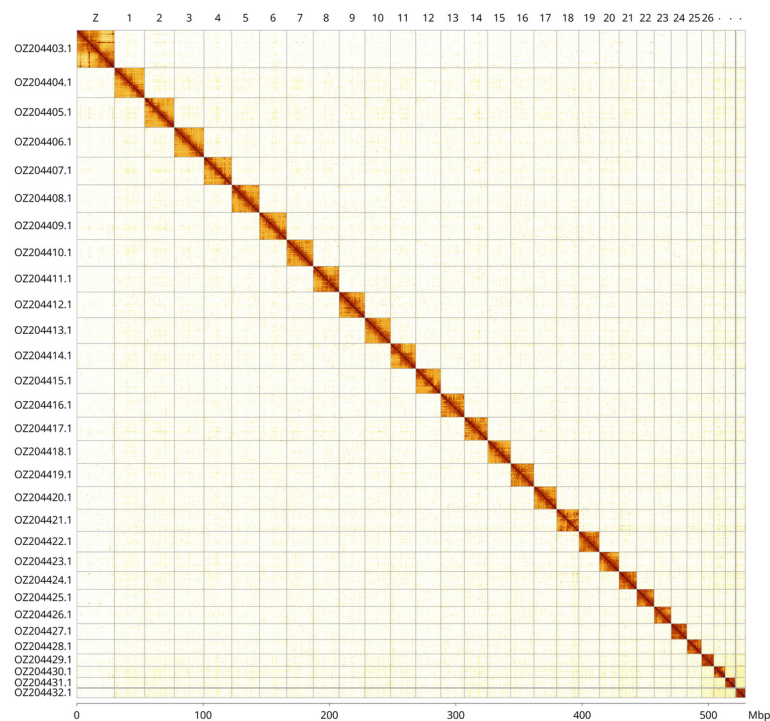


Figure 3. Hi-C contact map of the *Paratalanta hyalinialis* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in both haplotypes of the genome assembly of *Paratalanta hyalinialis*.

Haplotype 1				Haplotype 2			
INSDC accession	Name	Length (Mb)	GC%	INSDC accession	Name	Length (Mb)	GC%
OZ204404.1	1	23.80	37	OZ204343.1	1	23.86	37
OZ204405.1	2	23.49	37	OZ204344.1	2	23.39	36.50
OZ204406.1	3	23.47	36.50	OZ204345.1	3	23.26	37
OZ204407.1	4	22.07	37	OZ204346.1	4	22.19	36.50
OZ204408.1	5	21.86	36.50	OZ204347.1	5	21.80	36.50
OZ204409.1	6	21.57	36.50	OZ204348.1	6	21.51	36.50
OZ204410.1	7	21.04	36.50	OZ204349.1	7	21.01	36.50
OZ204411.1	8	20.54	36.50	OZ204350.1	8	20.37	36.50
OZ204412.1	9	20.35	36.50	OZ204351.1	9	20.34	36.50
OZ204413.1	10	20.34	36.50	OZ204352.1	10	20.01	36.50
OZ204414.1	11	19.94	36.50	OZ204353.1	11	19.89	36.50
OZ204415.1	12	19.82	36.50	OZ204354.1	12	19.83	36.50
OZ204416.1	13	18.66	36.50	OZ204355.1	13	19.81	36.50
OZ204417.1	14	18.47	37	OZ204356.1	14	18.74	36.50

Haplotype 1				Haplotype 2			
INSDC accession	Name	Length (Mb)	GC%	INSDC accession	Name	Length (Mb)	GC%
OZ204418.1	15	18.28	37	OZ204357.1	15	18.69	37
OZ204419.1	16	18.26	37	OZ204358.1	16	18.43	37
OZ204420.1	17	18.09	37	OZ204359.1	17	18.13	37
OZ204421.1	18	17.57	37	OZ204360.1	18	17.59	37
OZ204422.1	19	16.23	37	OZ204361.1	19	17.50	37
OZ204423.1	20	15.57	37.50	OZ204362.1	20	16.21	37
OZ204424.1	21	13.97	37.50	OZ204363.1	21	16.04	37.50
OZ204425.1	22	13.68	37	OZ204364.1	22	13.98	37.50
OZ204426.1	23	13.55	37	OZ204365.1	23	13.72	37
OZ204427.1	24	12.49	37.50	OZ204366.1	24	12.54	37.50
OZ204428.1	25	11.58	37	OZ204367.1	25	12.01	37.50
OZ204429.1	26	9.69	38	OZ204368.1	26	9.69	38
OZ204430.1	27	9.05	39	OZ204369.1	27	9.03	39
OZ204431.1	28	8.44	38.50	OZ204370.1	28	8.39	38.50
OZ204432.1	29	7.39	38.50	OZ204371.1	29	7.34	38.50
OZ204403.1	Z	29.92	36.50				

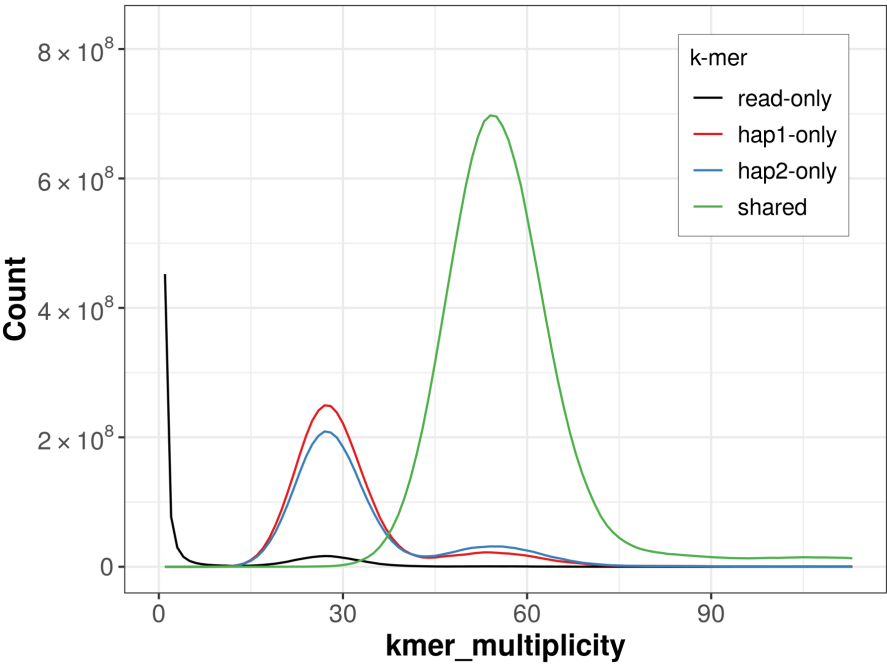


Figure 4. Evaluation of *k*-mer completeness using MerquyFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve corresponds to *k*-mers shared by both haplotypes, and the red and blue curves show *k*-mers found only in one of the haplotypes.

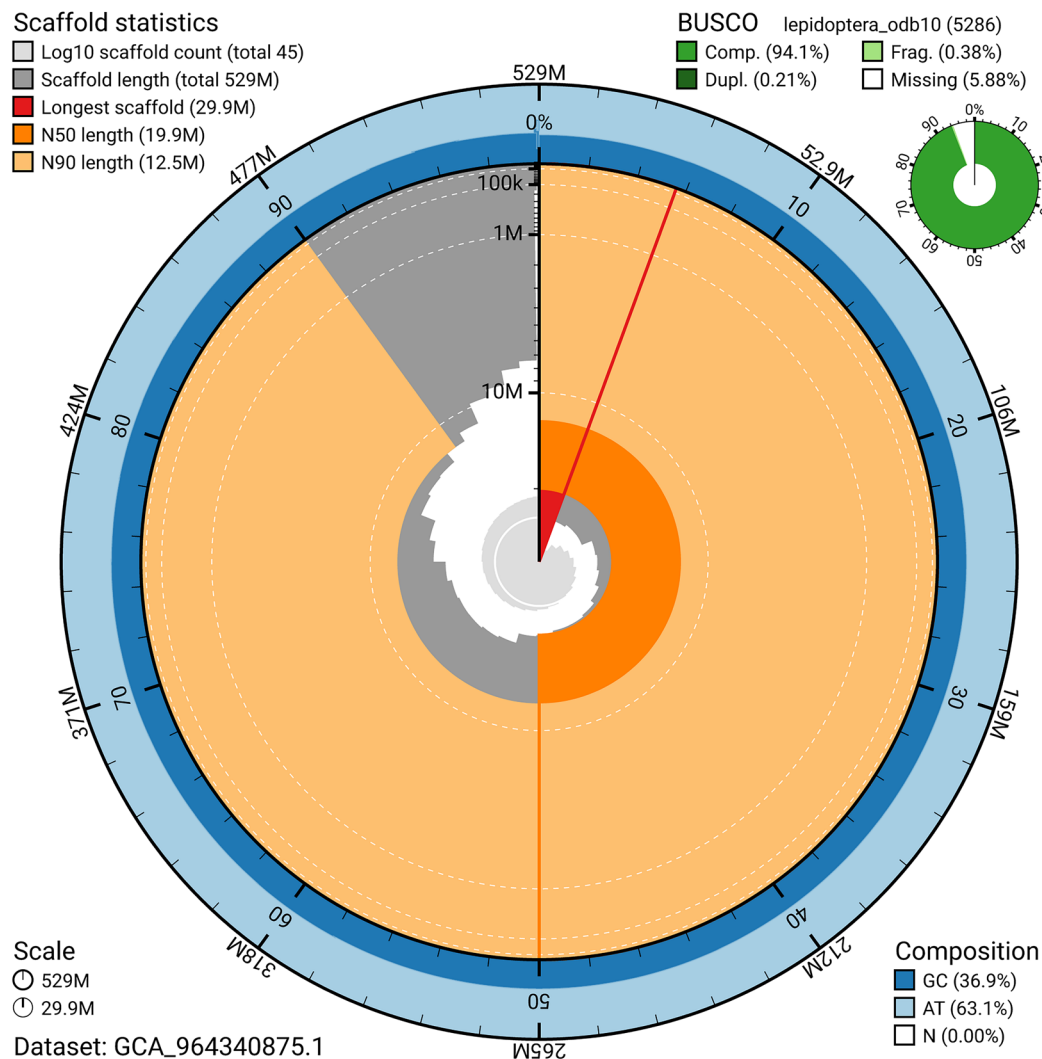


Figure 5. Assembly metrics for ilParHyal1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the set is presented at the top right. An interactive version of this figure can be accessed on the [BlobToolKit viewer](#).

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](#). By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for,

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

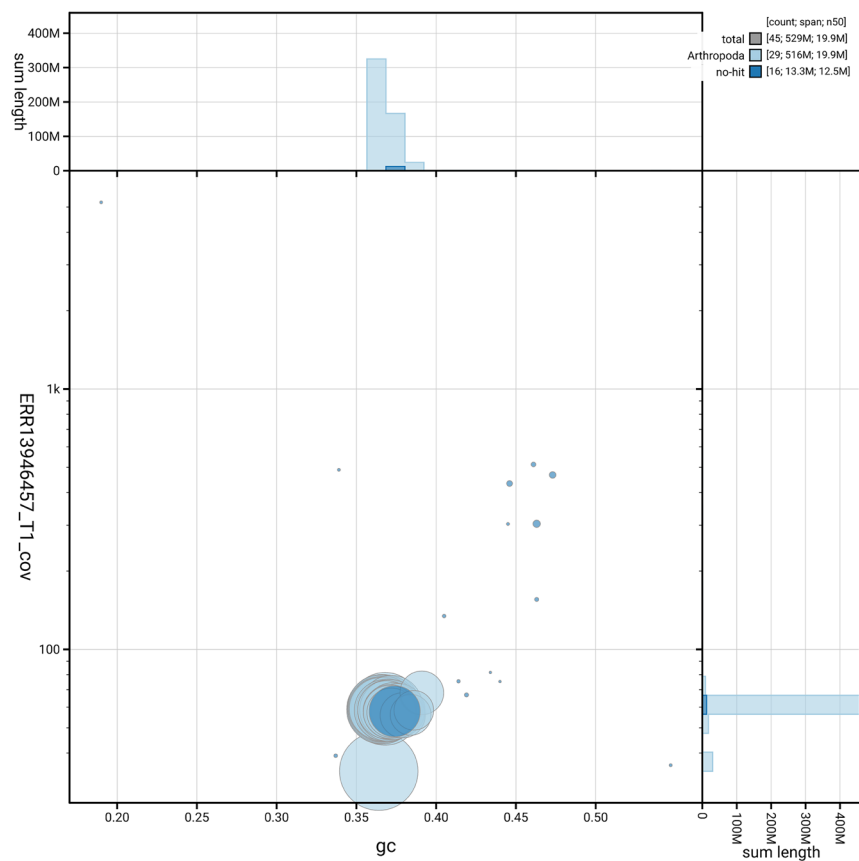


Figure 6. BlobToolKit GC-coverage plot for *ilParHyal1.hap1.1*. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the [BlobToolKit viewer](#).

Table 4. Earth Biogenome Project summary metrics for the *Paratalanta hyalinalis* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q67	6.C.Q40
Contig N50 length	8.51 Mb	≥ 1 Mb
Scaffold N50 length	19.94 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 67.3; haplotype 2: 68.2; combined: 67.7	≥ 40
k-mer completeness	Haplotype 1: 76.22%; Haplotype 2: 73.12%; combined: 98.27%	≥ 95%
BUSCO	C:94.1% [S:93.9%; D:0.2%]; F:0.4%; M:5.5%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.95%	≥ 90%

- 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: *Paratalanta hyalinalis*. Accession number [PRJEB82159](#). The genome sequence is released openly for reuse. The *Paratalanta hyalinalis* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665), the Sanger Institute Tree of Life Programme (PRJEB43745) and Project Psyche (PRJEB71705). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the [Ensembl](#) pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in [Table 1](#) and [Table 2](#).

Production code used in genome assembly at the WSI Tree of Life is available at <https://github.com/sanger-tol>. [Table 5](#) lists software versions used in this study.

Author information

Contributors are listed at the following links:

- Members of the [Natural History Museum Genome Acquisition Lab](#)
- Members of the [Darwin Tree of Life Barcoding collective](#)
- Members of the [Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team](#)
- Members of [Wellcome Sanger Institute Scientific Operations – Sequencing Operations](#)
- Members of the [Wellcome Sanger Institute Tree of Life Core Informatics team](#)
- Members of the [Tree of Life Core Informatics collective](#)
- Members of the [Darwin Tree of Life Consortium](#)

Table 5. Software versions and sources

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast/
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Goat CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextSnapshot	N/A	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView

Software	Version	Source
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.6.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

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](#)
- Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–410. [PubMed Abstract](#) | [Publisher Full Text](#)
- Barrett CG: **The Lepidoptera of the British islands 9 heterocera. Geometrina—Pyralidina.** London: Lovell Reeve; Co, 1904. [Reference Source](#)
- Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the Universal Protein Knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Challis R, 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](#)
- Cheng H, Jarvis ED, Fedrigo O, *et al.*: **Haplotype-resolved assembly of diploid genomes without parental data.** *Nat Biotechnol.* 2022; **40**(9): 1332–1335. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial Arthropod fauna [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 123. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): gjab008. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278. [PubMed Abstract](#) | [Publisher Full Text](#)
- Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- GBIF Secretariat: ***Paratalanta hyalinalis* (Hübner, 1796) Knapweed Pearl.** 2025. [Reference Source](#)
- Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025. [Publisher Full Text](#)
- Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): gaa153. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HiGlass: web-based visual exploration and analysis of genome interaction maps.** *Genome Biol.* 2018; **19**(1): 125. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lawnczak MKN, Davey RP, Rajan J, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life Project [version 1; peer review: 2 approved with reservations].** *Wellcome Open Res.* 2022; **7**: 187. [Publisher Full Text](#)
- Lees DC, Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, *et al.*: **The genome sequence of the Ochreous Pearl, *Anania crocealis* (Hübner, 1796) [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 568. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lepiforum: ***Paratalanta hyalinalis* (Hübner, 1796).** 2025. [Reference Source](#)
- Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mally R, Hayden JE, Neinhuis C, *et al.*: **The phylogenetic systematics of Spilomelinae and Pyraustinae (Lepidoptera: Pyraloidea: Crambidae) inferred from DNA and morphology.** *Arthropod Systematics & Phylogeny.* 2019; **77**(1): 141–204. [Publisher Full Text](#)
- Manni M, Berkeley MR, Seppely M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2. [Reference Source](#)
- NBN Gateway: ***Paratalanta hyalinalis* (Hübner, 1796) Knapweed Pearl.** 2025. [Reference Source](#)
- Parsons M, Clancy S: **A guide to the pyralid and crambid moths of Britain and Ireland.** Cornwall: Atropos Publishing, 2023. [Reference Source](#)
- Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat Commun.* 2020; **11**(1): 1432. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rhie A, Walenz BP, Koren S, *et al.*: **Merquy: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome*

Biol. 2020; **21**(1): 245.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI taxonomy: a comprehensive update on curation, resources and tools.** *Database (Oxford)*. 2020; **2020**: baaa062.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Sterling P, Parsons M, Lewington R: **Field guide to the micro-moths of Great Britain and Ireland.** London: Bloomsbury Publishing, 2023.

[Reference Source](#)

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life Project [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2024; **9**: 339.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Uliano-Silva M, Ferreira JGRN, Krashenninnikova 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](#)

Wright CJ, Stevens L, Mackintosh A, *et al.*: **Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera.** *Nat Ecol Evol.* 2024; **8**(4): 777–90.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 18 November 2025

<https://doi.org/10.21956/wellcomeopenres.27497.r137314>

© 2025 Pimenta J. 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.



João Pimenta 

CIBIO-InBIO, Research Centre in Biodiversity and Genetic Resources, Porto, Portugal

This genome note is overall well-structured, follows the standard Darwin Tree of Life reporting format, and presents a high-quality genome assembly for *Paratalanta hyalinalis*. The content is technically detailed and methodologically clear. Background information on the species is extensive and informative.

The Methods section is comprehensive and follows DTOL protocols. The assembly pipeline description is complete and up to current standards.

Comments:

The authors mention that sample metadata was collected following DTOL standards. However, there is no indication of where the metadata can be accessed. I'm aware that the metadata is deposited in COPO and accessible through ENA, following Lawniczak et al. (2022), but a sentence should be added to clarify where the data has been deposited and how it can be accessed.

Typos:

ha plotype - Background section - "...identical to the most common ha plotype on BOLD..."

There are some italics missing from species names throughout the text

"In the Palaearctic the species ranges from Great Britain (excluding Ireland)..." - comment: Great Britain already excludes Ireland, maybe change to: "British Isles (excluding Ireland)"

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: Evolutionary 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 03 November 2025

<https://doi.org/10.21956/wellcomeopenres.27497.r135540>

© 2025 Liew Y. 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.



Yvonne Jing Mei Liew

Universiti Malaya, Kuala Lumpur, Kuala Lumpur, Malaysia

1. Correct typo haplotype - and is identical to the most common ha plotype on BOLD (31/07/2025).
2. Figure 1 - a scale should be provided to indicate the size of the specimen
3. Table 3 should include the information for mitochondrial genome sequence
4. Can the author explain the reason why mitochondrial genome is not assembled from Illumina sequence data for Haplotype 2?

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 (mitogenome, NGS)

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.
