DATA NOTE

The genome sequence of *Zethes insularis* Rambur, 1833 (Lepidoptera: Erebidae)

[version 1; peer review: 2 approved]

Niklas Wahlberg[1], Charlotte J. Wright[2], Joana I. Meier[2], Mark L. Blaxter[2], 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, Project Psyche Community

[1]Lund University, Lund, Sweden
[2]Tree of Life, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 09 Jan 2026, **11**:23
https://doi.org/10.12688/wellcomeopenres.25723.1
Latest published: 09 Jan 2026, **11**:23
https://doi.org/10.12688/wellcomeopenres.25723.1

Abstract

We present a genome assembly from a male specimen of *Zethes insularis* (Arthropoda; Insecta; Lepidoptera; Erebidae). The assembly contains two haplotypes with total lengths of 862.07 megabases and 869.20 megabases. Most of haplotype 1 (99.0%) is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. Most of haplotype 2 (95.76%) is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled, with a length of 15.77 kilobases. This work is part of Project Psyche, a collaborative programme generating genomes for European butterflies and moths.

Keywords

Zethes insularis; Erebidae; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 09 Jan 2026	✓ view	✓ view

1. **Iker Irisarri**, Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain
Raul Contreras, Museo Nacional de Ciencias Naturales (Ringgold ID: 16625), Madrid, Spain

2. **Vadim Pisarenco**, Universitat of Barcelona, Barcelona, Spain

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Charlotte J. Wright (cw22@sanger.ac.uk)

Author roles: Wahlberg N: Investigation, Resources, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Wright CJ**: Funding Acquisition, Project Administration, Supervision; **Meier JI**: Funding Acquisition, Project Administration, Supervision; **Blaxter ML**: Funding Acquisition, Project Administration, Supervision;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540). NW was funded by the Swedish Research Council, Grant No. 2024-04040.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Wahlberg N, Wright CJ, Meier JI *et al.* **The genome sequence of *Zethes insularis* Rambur, 1833 (Lepidoptera: Erebidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2026, **11**:23
https://doi.org/10.12688/wellcomeopenres.25723.1

First published: 09 Jan 2026, **11**:23 https://doi.org/10.12688/wellcomeopenres.25723.1

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; Noctuoidea; Erebidae; Erebinae; *Zethes*; *Zethes insularis* Rambur, 1833 (NCBI:txid1973593)

Background

The moth *Zethes insularis* Rambur, 1833 belongs to the family Erebidae, and is found in the Mediterranean region in Europe, with its distribution stretching through the Middle East to the Caucasus mountains (GBIF Secretariat, 2025). The genus comprises some seven species, but *Z. insularis* is the only European representative (Goater *et al.*, 2003).

Not much is known about the biology of *Z. insularis*. It is found in dry, hot scrub habitats of the Mediterranean region. Its larvae have been recorded feeding on *Pistacia terebinthus* (Annacardiaceae), a small tree growing in Mediterranean scrub (Wagner, 2025). The adults come to light and bait traps quite readily and can be quite common where the species is found. The adults fly from March to September in two or three generations (Goater *et al.*, 2003).

We present a chromosomally complete genome sequence for *Zethes insularis*, based on a male specimen collected near the town of Novigrad, Zadar County, Croatia (Figure 1). The genome was sequenced as part of Project Psyche, a collaborative effort to sequence all named species of Lepidoptera found in Europe.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult male *Zethes insularis* (specimen ID LUNDPSY000082, ToLID ilZetInsu1; Figure 1), collected from Novigrad, Zadar, Croatia (latitude 44.1863, longitude 15.5421) on 2024-07-10. The specimen was collected and identified by Niklas Wahlberg (Lund University).

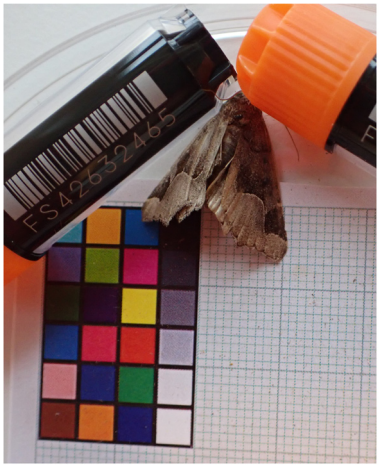

Figure 1. Voucher photograph of the *Zethes insularis* (ilZetInsu1) specimen used for genome sequencing.

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 ilZetInsu1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the thorax was homogenised by powermashing using a PowerMasher II tissue disruptor.

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol. DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor®3 for LI PacBio protocol. Sheared DNA was purified by automated SPRI (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 33.23 ng/μL and a yield of 1 561.81 ng, with a fragment size of 15.6 kb.

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), according to 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.

Specimen details, sequencing platforms, and data yields are summarised in Table 1.

Table 1. Specimen and sequencing data for BioProject PRJEB87483.

Platform	PacBio HiFi	Hi-C
ToLID	ilZetInsu1	ilZetInsu1
Specimen ID	LUNDPSY000082	LUNDPSY000082
BioSample (source individual)	SAMEA116129931	SAMEA116129931
BioSample (tissue)	SAMEA116130182	SAMEA116130176
Tissue	thorax	head
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR14835544	ERR14782893
Read count total	8.13 million	836.79 million
Base count total	81.10 Gb	126.35 Gb

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the head of the ilZetInsu1 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 Diagnocine 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 (Thermo Fisher 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 to create 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.

Specimen details, sequencing platforms, and data yields are summarised in Table 1.

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. 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 19 breaks and 66 joins. This reduced

the scaffold count by 34.2%, increased the scaffold N50 by 2.3%, and reduced the total assembly length by 2.3%. The curation process is described at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020), run in a Singularity container (Kurtzer *et al.*, 2017), was used 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 (Di Tommaso *et al.*, 2017) 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).

We used lep_busco_painter to paint Merian elements along chromosomes (Wright *et al.*, 2024). Merian elements represent the 32 ancestral linkage groups in Lepidoptera. The painting process utilised BUSCO gene locations from the lepidoptera_odb10 set (Kriventseva *et al.*, 2019) and chromosome lengths from NCBI Datasets. Each complete BUSCO gene (both single-copy and duplicated) was assigned to a Merian element based on a reference database, then coloured and plotted along chromosomes drawn to scale.

Genome sequence report

Sequence data

PacBio sequencing of the *Zethes insularis* specimen generated 81.10 Gb (gigabases) from 8.13 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 867.89 Mb, with a heterozygosity of 0.24% and repeat content of 28.14% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 91× coverage. Hi-C sequencing produced 126.35 Gb from 836.79 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. Haplotype 1 was curated to chromosome level, while

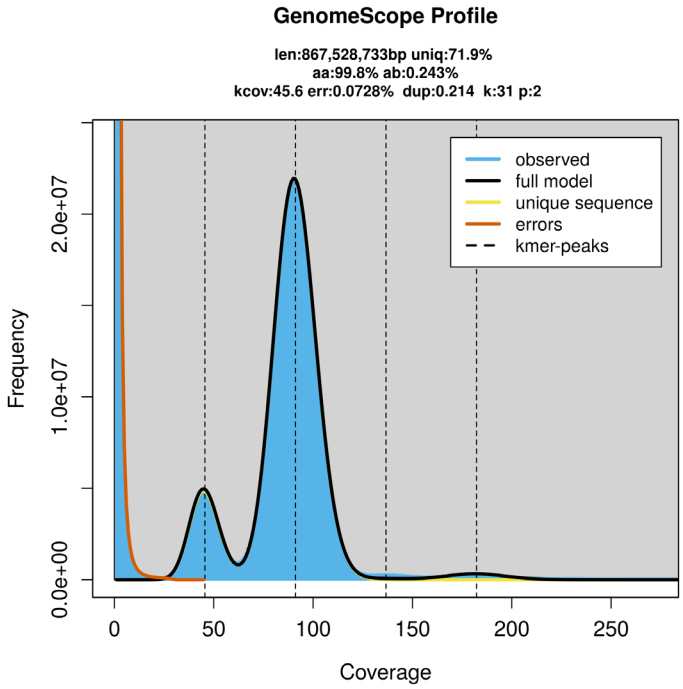

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

haplotype 2 was assembled to scaffold level. The final assembly has a total length of 862.07 Mb in 214 scaffolds, with 93 gaps, and a scaffold N50 of 29.55 Mb (Table 2).

Most of the assembly sequence (99.0%) was assigned to 31 chromosomal-level scaffolds, representing 30 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Chromosome painting with Merian elements illustrates the distribution of orthologues along chromosomes and highlights patterns of chromosomal evolution relative to Lepidopteran ancestral linkage groups (Figure 4). This genome has

Table 2. Genome assembly statistics.

Assembly name	ilZetInsu1.hap1.1	ilZetInsu1.hap2.1
Assembly accession	GCA_965239635.1	GCA_965239655.1
Assembly level	chromosome	chromosome
Span (Mb)	862.07	869.20
Number of chromosomes	31	31
Number of contigs	307	296
Contig N50	12.72 Mb	13.86 Mb
Number of scaffolds	214	200
Scaffold N50	29.55 Mb	29.75 Mb
Longest scaffold length (Mb)	40.53	40.41
Sex chromosomes	Z	Z
Organelles	Mitochondrion: 15.77 kb	-

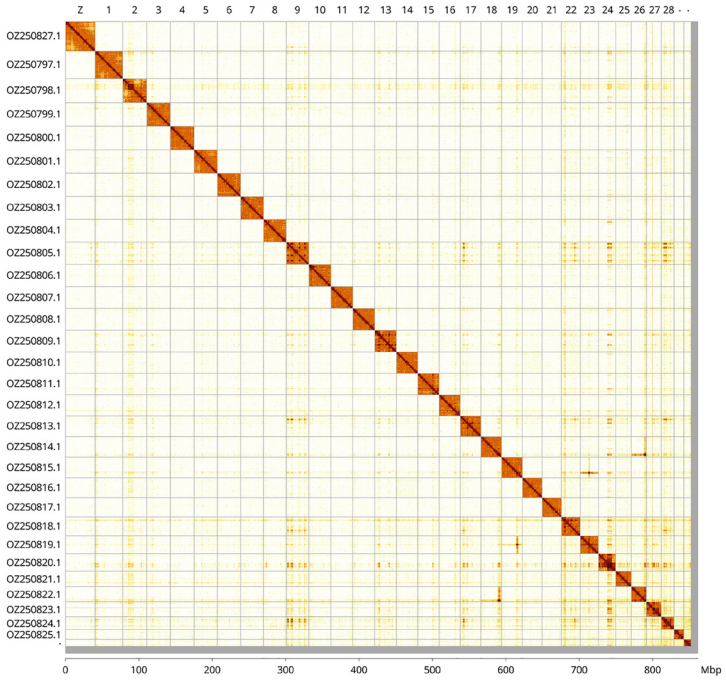

Figure 3. Hi-C contact map of the *Zethes insularis* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Zethes insularis* ilZetInsu1 (haplotype 2 also at chromosome level).

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ250797.1	1	37.50	37	M8
OZ250798.1	2	32.96	37.50	M11
OZ250799.1	3	32.10	37.50	M3
OZ250800.1	4	32.07	37.50	M2
OZ250801.1	5	32.06	37.50	M7
OZ250802.1	6	31.59	37	M9
OZ250803.1	7	31.02	37	M12
OZ250804.1	8	30.95	37.50	M17;M20
OZ250805.1	9	30.79	37.50	M23
OZ250806.1	10	30.48	37.50	M18
OZ250807.1	11	29.83	37.50	M1
OZ250808.1	12	29.66	37.50	M5
OZ250809.1	13	29.55	37.50	M14
OZ250810.1	14	29.11	37	M16
OZ250811.1	15	29.06	37.50	M21
OZ250812.1	16	28.89	37.50	M6
OZ250813.1	17	28.39	37	M13
OZ250814.1	18	28.18	37.50	M15
OZ250815.1	19	28.15	37.50	M10
OZ250816.1	20	26.94	37.50	M22
OZ250817.1	21	26.38	37.50	M4
OZ250818.1	22	25.57	37.50	M19
OZ250819.1	23	24.73	38	M24
OZ250820.1	24	24.39	38.50	M27
OZ250821.1	25	21.06	38	M26
OZ250822.1	26	20.56	39	M29
OZ250823.1	27	20.52	38	M28
OZ250824.1	28	17.29	39.50	M30
OZ250825.1	29	13.50	39	M25
OZ250826.1	30	9.67	39.50	M31
OZ250827.1	Z	40.53	36.50	MZ

been assembled using PacBio and Hi-C data and phased. The result is two curated haplotypes.Z chromosome identified by Merian painting. A haplotypic inversion was observed in the region on chromosome 1 (34.8-37.4 Mbp), chromosome 6 (22.4–31.5 Mbp), and chromosome 9 (3.5-6.1 Mbp). The exact order and orientation of the contigs on chromosome 24 (12.0–18.3 Mbp) are unknown.

The mitochondrial genome was also assembled (length 15.77 kb, OZ250828.1). This sequence is included as a contig in the

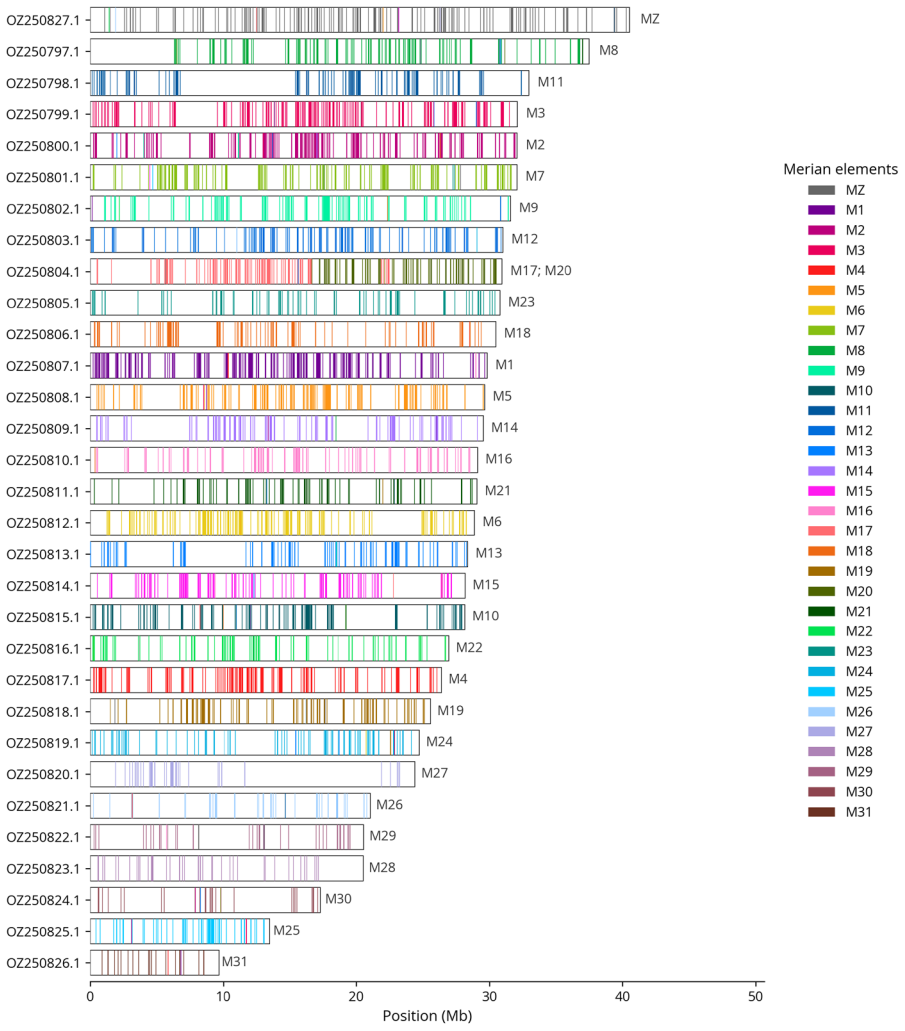

Figure 4. Merian elements painted across chromosomes in the ilZetInsu1.hap1.1 assembly of *Zethes insularis*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

For haplotype 1, the estimated QV is 62.3, and for haplotype 2, 62.3. When the two haplotypes are combined, the assembly achieves an estimated QV of 62.3. The *k*-mer completeness is 93.27% for haplotype 1, 93.33% for haplotype 2, and 99.79% for the combined haplotypes (Figure 5). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.7% of the expected gene set (single = 97.8%, duplicated = 0.9%) in haplotype 1. For haplotype 2, BUSCO analysis identified 98.7% of the expected gene set (single = 98.0%, duplicated = 0.8%).

The snail plot in Figure 6 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 7 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 **7.C.Q62**, meeting the recommended reference standard.

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Tree of Life collaborator. 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

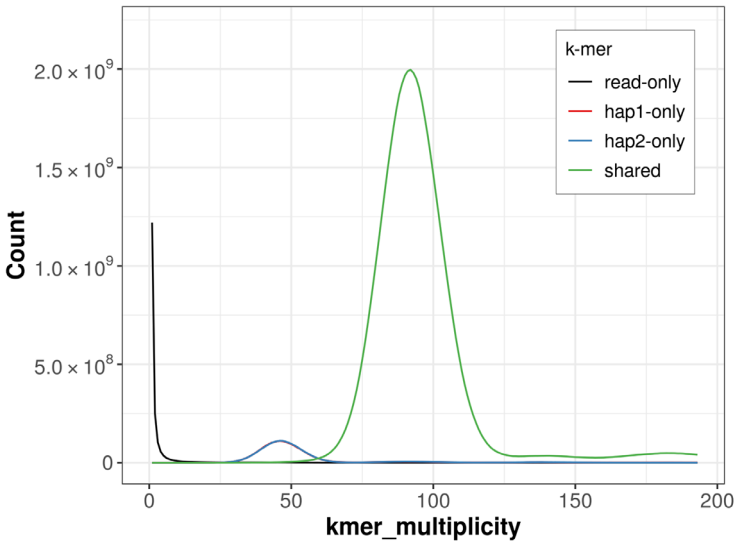

Figure 5. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve (the homozygous peak) corresponds to *k*-mers shared by both haplotypes and the red and blue curves (the heterozygous peaks) show *k*-mers found only in one of the haplotypes.

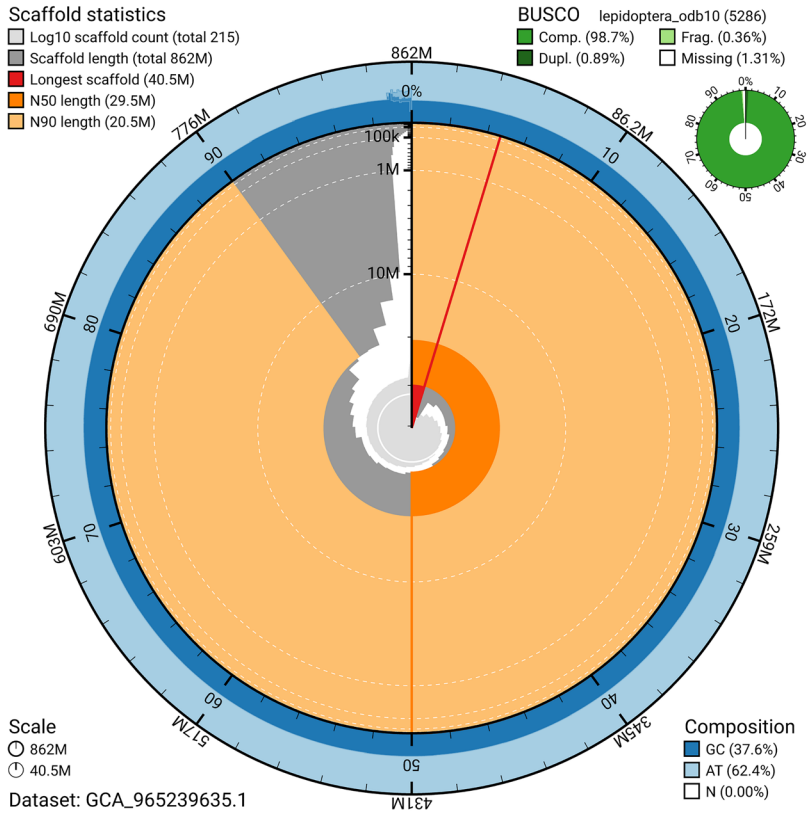

Figure 6. Assembly metrics for ilZetInsu1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

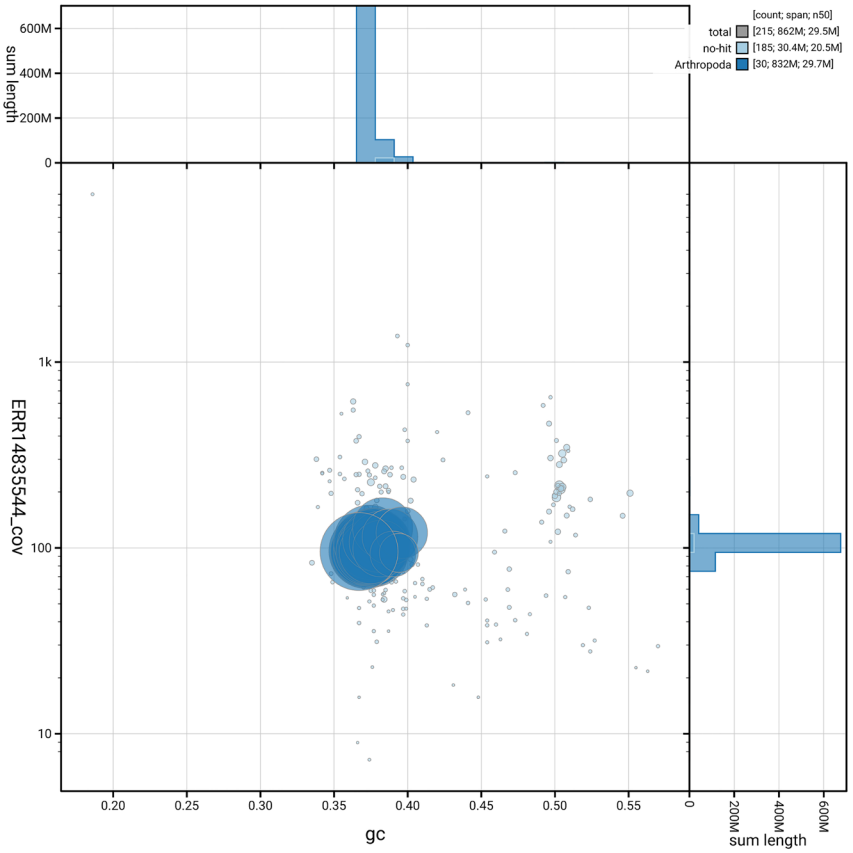

Figure 7. BlobToolKit GC-coverage plot for ilZetInsu1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Zethes insularis* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	7.C.Q62	6.C.Q40
Contig N50 length	12.72 Mb	≥ 1 Mb
Scaffold N50 length	29.55 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 62.3; haplotype 2: 62.3; combined: 62.3	≥ 40
k-mer completeness	Haplotype 1: 93.27%; Haplotype 2: 93.33%; combined: 99.79%	≥ 95%
BUSCO	C:98.7% [S:97.8%; D:0.9%]; F:0.4%; M:0.9%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99%	≥ 90%

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 undertaken according to a Research Collaboration Agreement or Material Transfer Agreement

entered into by the Tree of Life collaborator, Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances, other Tree of Life collaborators.

Data availability

European Nucleotide Archive: Zethes insularis. Accession number PRJEB87483. The genome sequence is released openly for reuse. The *Zethes insularis* genome sequencing initiative is part of 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 Ensembl at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

Pipelines used for genome assembly at the WSI Tree of Life are available at https://pipelines.tol.sanger.ac.uk/pipelines. Table 5 lists software versions used in this study.

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	4.4.5	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.7.1	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
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
lep_busco_painter	1.0.0	https://github.com/charlottewright/lep_busco_painter
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.28-r1209	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	24.10.4	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView
samtools	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.7.1	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2.2	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

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

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008.
PubMed Abstract | Publisher Full Text | Free 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

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: ***Zethes insularis* Rambur, 1833.** on GBIF Backbone Taxonomy. 2025.
Reference Source

Goater B, Ronkay L, Fibiger M: **Noctuidae Europaeae 10. Catocalinae & Plusiinae.** Soro, Denmark: Entomological Press, 2003.
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): giaa153.
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

Kriventseva EV, Kuznetsov D, Tegenfeldt F, *et al.*: **OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs.** *Nucleic Acids Res.* 2019; **47**(D1): D807–D811.
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

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
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

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2.
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.*: **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

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

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

Wagner W: ***Zethes insularis* - European Lepidoptera and their ecology**. 2025.
Reference Source

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 05 March 2026

https://doi.org/10.21956/wellcomeopenres.28321.r147526

✔ **Vadim Pisarenco**
Universitat of Barcelona, Barcelona, Spain

The authors present a high-quality, phased, chromosome-level genome assembly of the poorly studied moth species *Zethes insularis*. To my knowledge, this is the first genome sequence available for this genus. The methods are clearly described and detailed enough to allow replication by others. I have some minor comments and suggestions, outlined below:

1- In the "Assembly statistics" section, the following sentence is a bit confusing ("Haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level.") considering that both haplotypes are at chr-level as you state later ("[...] assembled using PacBio and Hi-C data and phased. The result is two curated haplotypes.")

2- Here, "Z chromosome identified by Merian painting. A haplotypic inversion was observed [...]". You might consider providing a more detailed explanation about how the inversions are detected (I understand that by comparing the distribution of orthologs along the chromosomes). You could also highlight those inversions in Figure 4.

3- (Typo?) BUSCO percentages sum 100% in Table 4, but do not in Figure 6 (100,37) due to % of Missing BUSCOs (1.31% in Figure 6 vs 0.9% in Table 4).

4- (Suggestion) If you find it appropriate, you could include a map showing the range of distribution of this species.

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

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

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

Yes

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolutionary and 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 20 February 2026

https://doi.org/10.21956/wellcomeopenres.28321.r146728

Iker Irisarri
Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain
Raul Contreras
Biodiversity and Evolutionary Biology, Museo Nacional de Ciencias Naturales (Ringgold ID: 16625), Madrid, Community of Madrid, Spain

Wahlberg and colleagues present the haplotype-resolved genome assembly of the moth Zethes insularis, which has very good contiguity and completeness statistics and reaches the quality required by EBP. This is part of the Project Psyche and was sequenced by the Wellcome Sanger Institute team.

Fig 1. The moth is not very visible because it is covered by the tubes. It might be worth checking if other, better pictures are available.

It might be a coincidence, but might be worth checking whether it is 2.3% in both cases here: “This reduced the scaffold count by 34.2%, increased the scaffold N50 by 2.3%, and reduced the total assembly length by 2.3%”

Assembly statistics: 1) It is said that haplotype 1 is curated to chromosome-level and haplotype 2 to scaffold level, but both of them recover the 31 expected chromosomes. Why is that? 2) “The final assembly has a total length of 862.07 Mb in 214 scaffolds, with 93 gaps, and a scaffold N50 of 29.55 Mb (Table 2).” Please, specify that this refers to haplotype 1.

Fig. 3 Could you please indicate in the caption which haplotype this refers to?

Please, check this sentence “Table 4 lists the assembly metric benchmarks adapted from Rhie et al.

(2021) the Earth BioGenome Project Report on Assembly Standards September 2024." Should maybe "Rhie et al. (2021)" be written after 2024?

Page 7. Space missing in "haplotypes.Z"

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

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