DATA NOTE

The genome sequence of the Clouded Buff, *Diacrisia sannio* (Linnaeus, 1758) (Lepidoptera: Erebidae)

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

Kay Lucek[1], Bulah Wu[2], Michal Rindoš[2], Charlotte J. Wright[3], Joana I. Meier[3], Mark L. Blaxter[3],
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]University of Neuchâtel, Neuchâtel, Switzerland
[2]Biology Centre of the Czech Academy of Sciences, Institute of Entomology, Ceske Budejovice, Czech Republic
[3]Tree of Life Programme, Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 23 Jan 2026, **11**:54
https://doi.org/10.12688/wellcomeopenres.25823.1
Latest published: 23 Jan 2026, **11**:54
https://doi.org/10.12688/wellcomeopenres.25823.1

Abstract

We present a genome assembly from a male specimen of *Diacrisia sannio* (Clouded Buff; Arthropoda; Insecta; Lepidoptera; Erebidae). The assembly contains two haplotypes with total lengths of 722.98 megabases and 724.10 megabases. Most of haplotype 1 (99.86%) is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. Most of haplotype 2 (99.89%) is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled, with a length of 15.58 kilobases. This work is part of Project Psyche, a collaborative programme generating genomes for European butterflies and moths.

Keywords

Diacrisia sannio; Clouded Buff; genome sequence; chromosomal; Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ? ✓

	1	2	3
version 1 23 Jan 2026	✓ view	? view	✓ view

1. **Jayan Duminda M Senevirathna**, Uva Wellassa University, Badulla, Sri Lanka
2. **Nhat Tan Pham**, Adam Mickiewicz University in Poznań, Poznań, Poland
3. **Annabel Whibley**, The University of Auckland, Auckland, New Zealand
Bragato Research Institute, Lincoln, New Zealand

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: Lucek K: Investigation, Resources, Supervision, Writing – Review & Editing; **Wu B**: Data Curation; **Rindoš M**: 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). KL was funded by Swiss National Science Foundation (SNSF) Grant ID 202869.
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: Lucek K, Wu B, Rindoš M *et al.* **The genome sequence of the Clouded Buff, *Diacrisia sannio* (Linnaeus, 1758) (Lepidoptera: Erebidae) [version 1; peer review: 2 approved, 1 approved with reservations]** Wellcome Open Research 2026, **11**:54 https://doi.org/10.12688/wellcomeopenres.25823.1

First published: 23 Jan 2026, **11**:54 https://doi.org/10.12688/wellcomeopenres.25823.1

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; Arctiinae; Arctiini; *Diacrisia*; *Diacrisia sannio* (Linnaeus, 1758) (NCBI:txid464721)

Background

The Clouded Buff is a moth of the Arctiinae with a Palaearctic distribution. The adults are sexually dimorphic with the absent proboscis. The body and forewings of the males are lemon-coloured, while the body of the females is brown with black spots. The forewings of the females are brown-orange. The difference between the sexes can also be seen in the pectination of the antennae. *Diacrisia sannio* is a mesophilic species that inhabits mesophilic grasslands and partially wooded biotopes (Witt *et al.*, 2011).

It is monovoltine in central and northern Europe and bivoltine in the south. The males are diurnal and nocturnal, while the females are usually only active during the night. The caterpillars are brown with red, conspicuous hairs, a white dorsal line and white stigmata. The Clouded Buff is a polyphagous species and feeds on *Epilobium* spp., *Galium* spp., *Plantago* spp., *Taraxacum* spp. and *Urctica* spp. (Pro Natura, Schweizerischer Bund für Naturschutz, 2000). The species overwinters as a caterpillar.

We present a chromosome-level genome sequence for *D. sannio*, sequenced as part of Project Psyche. The sequence data were derived from a male specimen (Figure 1) collected from Grindelwald, Switzerland.

Figure 1. Voucher photograph of the *Diacrisia sannio* (ilDiaSann1) specimen used for genome sequencing.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult male *Diacrisia sannio* (specimen ID SAN28000077, ToLID ilDiaSann1; Figure 1), collected from Grindelwald, Switzerland (latitude 46.6576, longitude 8.0999; elevation 1 975 m) on 2023-07-07. The specimen was collected and identified by Kay Lucek (University of Neuchâtel).

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 ilDiaSann1 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 62.46 ng/μL and a yield of 2 935.62 ng, with a fragment size of 15.0 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.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the abdomen; head of the ilDiaSann1 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.

Table 1. Specimen and sequencing data for BioProject PRJEB84281.

Platform	PacBio HiFi	Hi-C
ToLID	ilDiaSann1	ilDiaSann1
Specimen ID	SAN28000077	SAN28000077
BioSample (source individual)	SAMEA115109809	SAMEA115109809
BioSample (tissue)	SAMEA115109847	SAMEA115109846; SAMEA115109845
Tissue	thorax	abdomen; head
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR14151109; ERR14151110	ERR14171996; ERR14172297
Read count total	4.31 million	1 722.62 million
Base count total	41.39 Gb	260.12 Gb

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 one break and 19 joins, which reduced the scaffold count by 10.3%. 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 database (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 painted Merian elements, which represent the 32 ancestral linkage groups in Lepidoptera (Wright *et al.*, 2024), along chromosomes using lep_busco_painter. Painting used BUSCO gene locations from the lepidoptera_odb10 set (Kriventseva *et al.*, 2019) and chromosome lengths from NCBI Datasets. Each complete BUSCO (including both single-copy and duplicated BUSCOs) was assigned to a Merian element using a reference database. Positions were coloured and plotted along chromosomes drawn to scale.

Genome sequence report

Sequence data

PacBio sequencing of the *Diacrisia sannio* specimen generated 41.39 Gb (gigabases) from 4.31 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 714.84 Mb, with a heterozygosity of 1.87% and repeat content of 32.25% (Figure 2). These estimates guided expectations for the assembly. Based on the

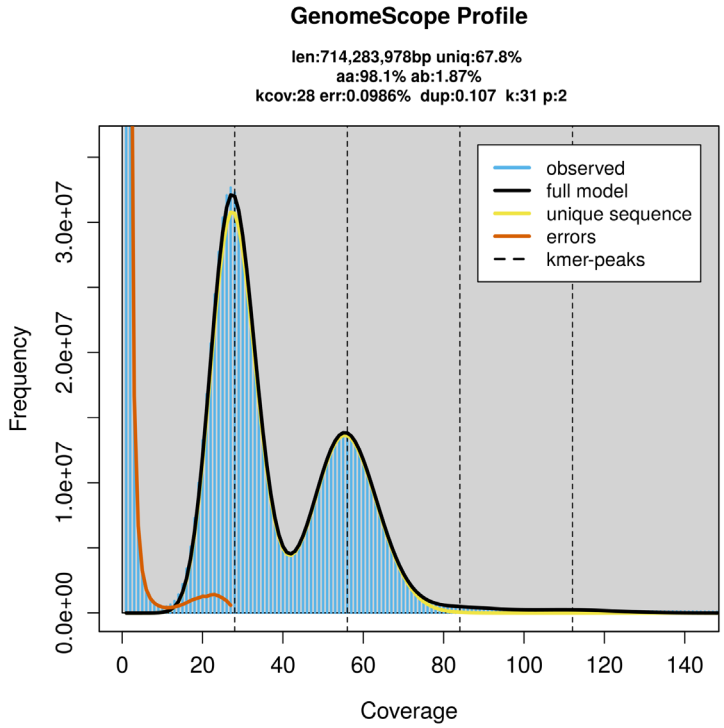

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

estimated genome size, the sequencing data provided approximately 56× coverage. Hi-C sequencing produced 260.12 Gb from 1 722.62 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 haplotype 2 was assembled to scaffold level. The final assembly has a total length of 722.98 Mb in 60 scaffolds, with 128 gaps, and a scaffold N50 of 24.83 Mb (Table 2).

Most of the assembly sequence (99.86%) 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

Table 2. Genome assembly statistics.

Assembly name	ilDiaSann1.hap1.1	ilDiaSann1.hap2.1
Assembly accession	GCA_965282705.1	GCA_965282695.1
Assembly level	chromosome	chromosome
Span (Mb)	722.98	724.10
Number of chromosomes	31	31
Number of contigs	188	172
Contig N50	8.07 Mb	8.73 Mb
Number of scaffolds	60	51
Scaffold N50	24.83 Mb	25.06 Mb
Longest scaffold length (Mb)	34.3	34.55
Sex chromosomes	Z	Z
Organelles	Mitochondrion: 15.58 kb	-

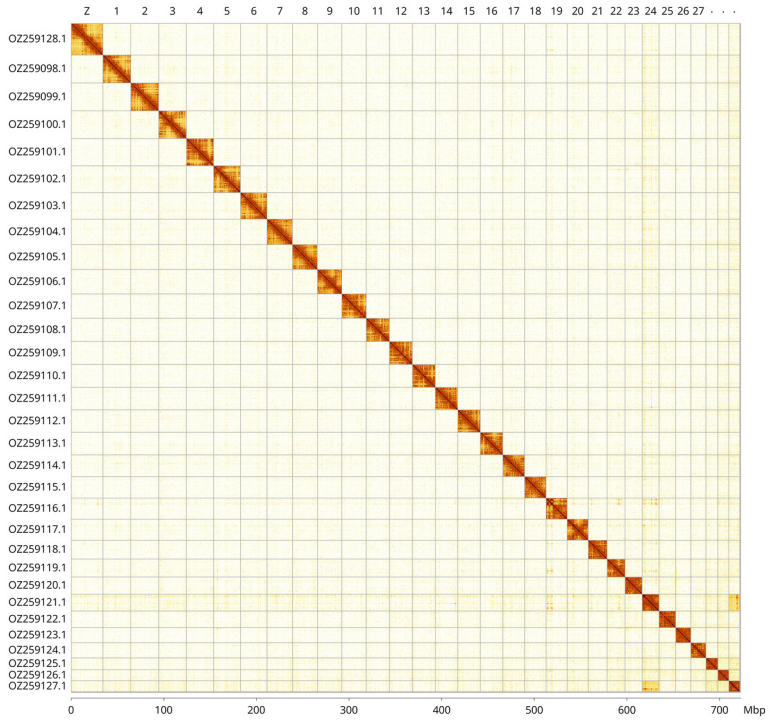

Figure 3. Hi-C contact map of the *Diacrisia sannio* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Diacrisia sannio* ilDiaSann1 (haplotype 2 also at chromosome level).

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ259098.1	1	30.19	36	M1
OZ259099.1	2	30.16	36	M2
OZ259100.1	3	29.95	36	M17;M20
OZ259101.1	4	29.25	35.50	M9
OZ259102.1	5	28.99	36	M3
OZ259103.1	6	28.79	35.50	M8
OZ259104.1	7	27.40	36	M5
OZ259105.1	8	26.92	35.50	M7
OZ259106.1	9	26.40	36	M6
OZ259107.1	10	26.25	35.50	M12
OZ259108.1	11	25.16	35.50	M18
OZ259109.1	12	24.83	36	M16
OZ259110.1	13	24.57	35.50	M21
OZ259111.1	14	24.40	36	M15
OZ259112.1	15	24.28	36	M10
OZ259113.1	16	24.26	36	M4
OZ259114.1	17	23.45	36	M11
OZ259115.1	18	23.26	35.50	M22
OZ259116.1	19	22.81	36	M23
OZ259117.1	20	22.69	36	M14
OZ259118.1	21	20.35	36	M13
OZ259119.1	22	19.34	36.50	M19
OZ259120.1	23	18.54	36.50	M24
OZ259121.1	24	18.32	38	M30
OZ259122.1	25	17.76	36.50	M26
OZ259123.1	26	16.49	36.50	M28
OZ259124.1	27	16.30	37	M27
OZ259125.1	28	12.78	37	M25
OZ259126.1	29	11.99	37.50	M29
OZ259127.1	30	11.82	39	M31
OZ259128.1	Z	34.30	35.50	MZ

illustrates the distribution of orthologues along chromosomes and highlights patterns of chromosomal evolution relative to Lepidopteran ancestral linkage groups (Figure 4). The Z chromosome was identified based on read coverage and BUSCO gene painting with Merian elements (Wright *et al.*, 2024).

The mitochondrial genome was also assembled (length 15.58 kb, OZ259129.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

For haplotype 1, the estimated QV is 66.9, and for haplotype 2, 66.8. When the two haplotypes are combined, the assembly achieves an estimated QV of 66.9. The *k*-mer completeness is 69.42% for haplotype 1, 69.49% for haplotype 2, and 99.72% for the combined haplotypes (Figure 5). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.9% of the expected gene set (single = 98.0%, duplicated = 0.9%) in haplotype 1. 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) and the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.Q66**, 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 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.

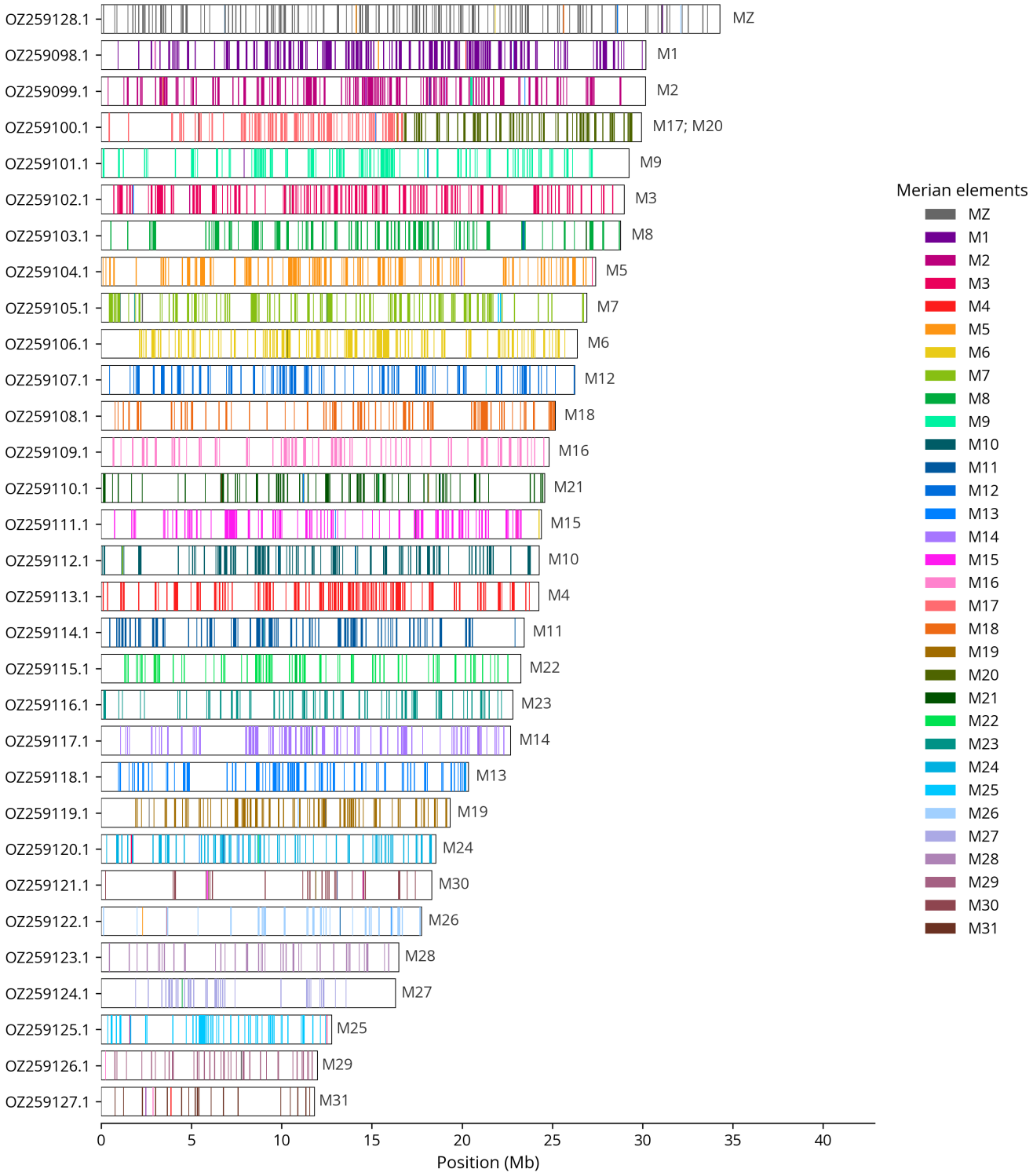

Figure 4. Merian elements painted across chromosomes in the ilDiaSann1.hap1.1 assembly of *Diacrisia sannio*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

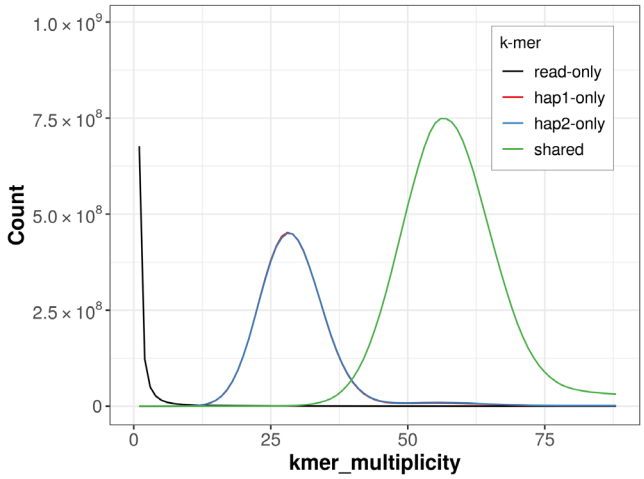

Figure 5. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers 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.

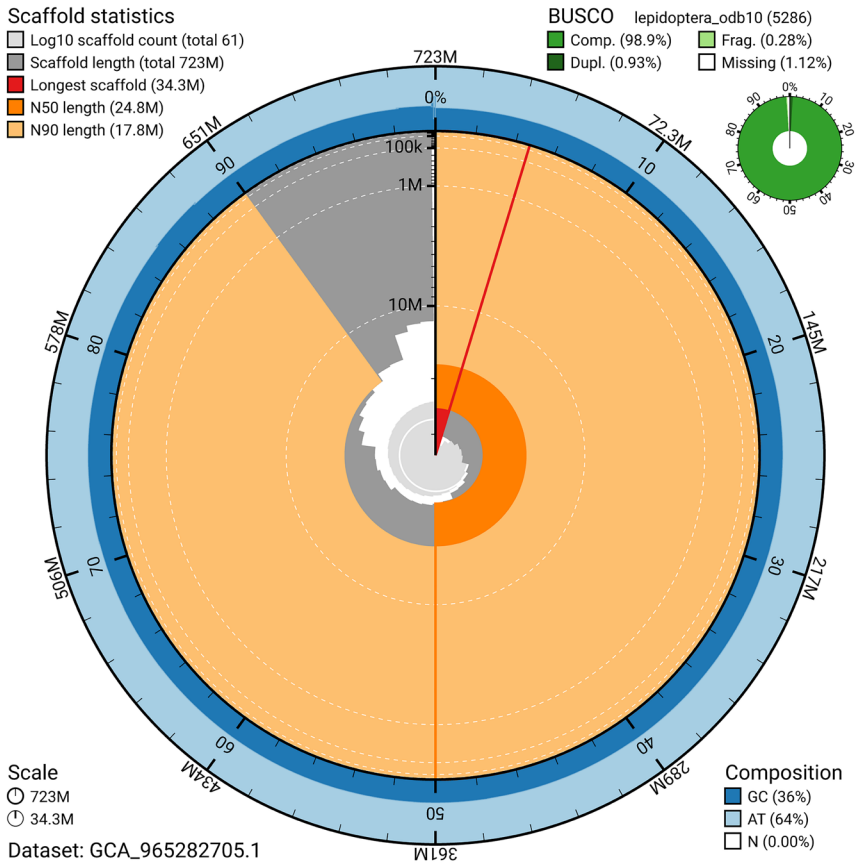

Figure 6. Assembly metrics for ilDiaSann1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

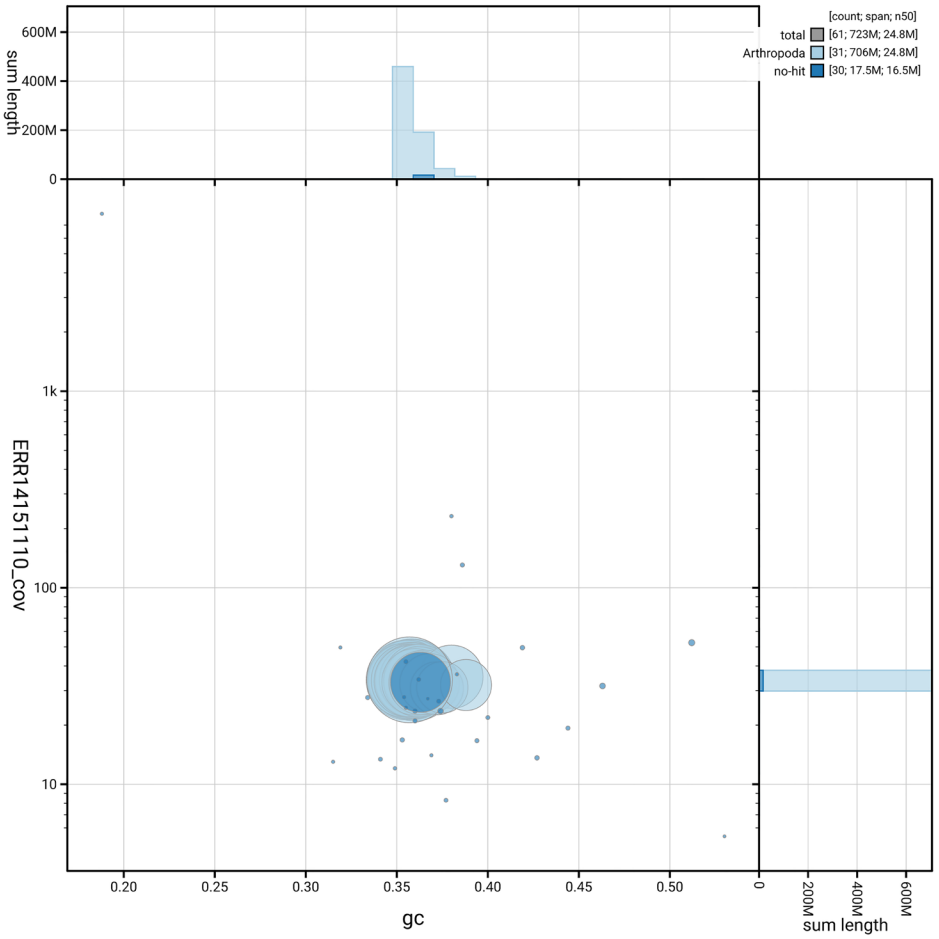

Figure 7. BlobToolKit GC-coverage plot for ilDiaSann1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Diacrisia sannio* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q66	6.C.Q40
Contig N50 length	8.07 Mb	≥ 1 Mb
Scaffold N50 length	24.83 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 66.9; haplotype 2: 66.8; combined: 66.9	≥ 40
k-mer completeness	Haplotype 1: 69.42%; Haplotype 2: 69.49%; combined: 99.72%	≥ 95%
BUSCO	C:98.9% [S:98.0%; D:0.9%]; F:0.3%; M:0.8%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.86%	≥ 90%

Notes: EBP summary uses log10(Contig N50); chromosome-level (C) or log10(Scaffold N50); Q (Merqury QV). BUSCO: C=complete; S=single-copy; D=duplicated; F=fragmented; M=missing; n=orthologues

Data availability

European Nucleotide Archive: Diacrisia sannio (clouded buff). Accession number PRJEB84281. The genome sequence is released openly for reuse. The *Diacrisia sannio* 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.

Author information

Contributors are listed at the following links:

- 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 Project Psyche Community

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

Software	Version	Source
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

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

Pro Natura, Schweizerischer Bund für Naturschutz: **Schmetterlinge Und Ihre Lebensräume, Band 3.** Egg bei Zürich: Fotorotar AG Verlag; 2000.

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

Witt TJ, Speidel W, Ronkay G, *et al.*: **Subfamilia arctiinae leach, [1815].** In: T. J. Witt and W. Speidel (eds), *Lymantriinae – arctiinae, including phylogeny and checklist of the quadrifid noctuoidea of Europe*. Sorø: Entomological Press, 2011.

Wright CJ, Stevens L, Mackintosh A, *et al.*: **Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera.** *Nat Ecol Evol.* 2024; **8**(4): 777–790.
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 13 May 2026

https://doi.org/10.21956/wellcomeopenres.28424.r146930

✔ **Annabel Whibley**
[1] The University of Auckland, Auckland, New Zealand
[2] Grapevine Improvement, Bragato Research Institute, Lincoln, New Zealand

Lucek and colleagues report a genome assembly of the Clouded Buff, *Diacrisia sannio,* with each haplotype reporting 31 pseudochromosomes (30 autosomes plus Z) and each containing >99.8 of the haplotype assembly lengths. The work was carried out as part of Project Psyche and the methods and reporting follow standard templating. I find the work to be systematically and comprehensively described, using contemporary and appropriate methods. The resulting assemblies are of excellent quality, supported by strong evaluation statistics (each haplotype assembly reaches 6.C.66). No genome annotation or RNAseq data collection is reported at this time. Public data links and metadata are fully available.
Minor comment: the repetition of “mesophilic” in the opening paragraph feels unnecessary. Would “Diacrista sannio inhabits mesophilic ….” do the same work or is there extra information conveyed by underscoring the ecological niche of the species is also mesophilic?

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, Bioinformatics

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 28 February 2026

https://doi.org/10.21956/wellcomeopenres.28424.r146924

?

Nhat Tan Pham
Adam Mickiewicz University in Poznań, Poznań, Poland

Authors describe the methods and results of genome assembly for the Clouded Buff, *Diacrisia sannio* (Linnaeus, 1758) (Lepidoptera: Erebidae). Two haplotypes were assembled using 56X PacBio data and Arima2 Hi-C data. The methods are logical and well-written. However, the method for identifying the Z chromosome has not yet been fully deciphered.
“The Z chromosome was identified based on read coverage and BUSCO gene painting with Merian element”. No data is available to verify the coverage. The authors can also check homology with the Erebidae genome (e.g., *Hyphantria cunea*).

The authors report two haplotypes from a genome assembly of *Diacrisia sannio*, using high-coverage (56X) PacBio HiFi and Hi-C sequences. Most of the assembly sequence was assigned to 30 autosomes and the Z chromosome. The authors also assembled the mitochondria and painted Merian elements in each chromosome. The method and results were well written, but a minor revision to the results is recommended to clarify how the authors identify the Z chromosome.
In detail, the authors stated, “The Z chromosome was identified based on read coverage and BUSCO gene painting with Merian elements.” However, I could not find information on coverage. It can also be based on homology with a published Erebidae genome (eg. *Anticarsia gemmatalis*). Addressing this minor issue will strengthen the paper.

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: genome assembly, Lepidoptera

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 28 Feb 2026

Tree of Life Team Sanger

We thank the peer reviewer for reviewing this data note and for the suggestion to clarify how the Z chromosome was identified. Read coverage is shown in the BlobToolKit GC–coverage plot (Figure 7) and can be explored further in the linked interactive BlobToolKit viewer. In this assembly, coverage is broadly similar across chromosome-scale scaffolds, as expected for a homogametic (male, ZZ) individual. The Z chromosome was assigned using BUSCO Merian-element painting (Figure 4), where one chromosome is painted predominantly as the Merian Z element (MZ). Given this unambiguous MZ assignment, we did not perform an additional homology-based check against other erebid assemblies.

Competing Interests: No competing interests were disclosed.

Reviewer Report 14 February 2026

https://doi.org/10.21956/wellcomeopenres.28424.r146220

Jayan Duminda M Senevirathna
Uva Wellassa University, Badulla, Uva Province, Sri Lanka

The manuscript is well written and includes the required data and methodological details. However, several minor revisions are recommended to improve clarity and presentation.
The abstract should be revised for conciseness and clarity. When describing the two haplotypes, the sentences are currently very similar and somewhat repetitive; these can be combined to improve readability and flow. In addition, the broader importance and implications of the genome analysis should be more clearly highlighted in the abstract to emphasize the scientific contribution of the study.
Furthermore, in Table 2, consistency should be maintained when presenting fragment lengths (e.g., units, formatting, decimal places), as inconsistencies may affect clarity and professionalism. Addressing these minor issues will further strengthen the manuscript.

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: Aquatic Biotechnology

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.