DATA NOTE

The genome sequence of the Scarce Hook-tip, *Sabra harpagula* (Esper, 1786) (Lepidoptera: Drepanidae)

[version 1; peer review: 3 approved]

Mark J. Sterling[1], Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding Collective, Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team, Wellcome Sanger Institute Scientific Operations: Sequencing Operations, Wellcome Sanger Institute Tree of Life Core Informatics team, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]Natural History Museum, London, England, UK

v1 **First published:** 26 May 2026, **11**:308
https://doi.org/10.12688/wellcomeopenres.26632.1
Latest published: 26 May 2026, **11**:308
https://doi.org/10.12688/wellcomeopenres.26632.1

Abstract

We present a genome assembly from an individual male *Sabra harpagula* (Scarce Hook-tip; Arthropoda; Insecta; Lepidoptera; Drepanidae). The genome sequence has a total length of 377.24 megabases. Most of the assembly (99.98%) is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled, with a length of 15.78 kilobases. This assembly was generated as part of the Darwin Tree of Life project, which produces genomes for eukaryotic species found in Britain and Ireland.

Keywords

Sabra harpagula, Scarce Hook-tip, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 1 26 May 2026	✓ view	✓ view	✓ view

1. **Rajib Deb**, ICAR-National Research Centre, Guwahat, India
2. **Fedor Čiampor**, Slovak Academy of Sciences, Bratislava, Slovakia
3. **Gulab D Khedkar**, Dr Babasaheb Ambedkar Marathwada University, Aurangabad, India

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: Sterling MJ: Investigation, Resources;

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, https://doi.org/10.35802/218328].
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: Sterling MJ, Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding Collective *et al.* **The genome sequence of the Scarce Hook-tip, *Sabra harpagula* (Esper, 1786) (Lepidoptera: Drepanidae) [version 1; peer review: 3 approved]** Wellcome Open Research 2026, **11**:308 https://doi.org/10.12688/wellcomeopenres.26632.1

First published: 26 May 2026, **11**:308 https://doi.org/10.12688/wellcomeopenres.26632.1

Species taxonomy

Eukaryota; Metazoa; Arthropoda; Insecta; Lepidoptera; Drepanidae; *Sabra*; *Sabra harpagula* (Esper, 1786) (NCBI: txid721158).

Background

Sabra harpagula (Esper, 1786), the Scarce Hook-tip, is a local moth in Britain and Ireland, with established populations mainly in the Wye Valley area of western England and south Wales. It is also recorded as a suspected immigrant. Adults have strongly hooked forewings, with a second projection about halfway along the outer edge of the forewing. The brown and gold central blotch, together with lilac and black outer markings, help separate it from similar hook-tip moths (Waring *et al.*, 2017).

In Britain, *S. harpagula* has one main adult generation, flying in June and July. A second brood is more usual in mainland Europe, where adults may also occur in August. Second-generation moths are reported to be smaller. In Britain, a partial second generation may be possible, but appears to be unusual and has been recorded mainly in captivity (Kimber, 2026; Waring *et al.*, 2017).

The larva feeds on small-leaved lime (*Tilia cordata*). It is present from late July to late September, sometimes into mid-October, and probably spends much of its time high in the canopy, which may partly explain why it is rarely encountered by beating lower branches. The species overwinters as a pupa in a thin cocoon spun in a curled leaf, which later falls to the ground (Waring *et al.*, 2017).

We present a chromosome-level genome sequence for *Sabra harpagula*, the Scarce Hook-tip. This assembly is the first high-quality genome for the genus *Sabra* and one of 16 genomes available for the family Drepanidae as of April 2026 (data obtained via NCBI datasets, O'Leary *et al.*, 2024).

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Sabra harpagula* (specimen ID NHMUK014438727, ToLID ilSabHarp2; Figure 1), collected from Lower Wyndcliff Car Park, Wales, United Kingdom (latitude 51.67, longitude −2.69) on 2022-04-24. A second specimen, collected on the same occasion, was used for Hi-C and RNA sequencing (specimen ID NHMUK014438725, ToLID ilSabHarp1). The specimens were collected by Mark Sterling and David Lees and identified by Mark Sterling.

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

Nucleic acid extraction

Detailed protocols for nucleic acid extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The ilSabHarp2 sample was weighed and triaged to determine the appropriate extraction protocol. For HMW DNA extraction, 14 mg of tissue from the abdomen was used. The tissue was homogenised by powermashing using a PowerMasher II tissue disruptor. High molecular weight (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 manual 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 38.69 ng/μL and a yield of 1 779.74 ng, with a fragment size of 14.6 kb. The Genomic Quality Number (GQN) was 7.5.

RNA was extracted from whole organism tissue of ilSabHarp1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol. The RNA concentration was assessed using a Nanodrop

Figure 1. Photographs of the *Sabra harpagula* (ilSabHarp2) specimen used for genome sequencing.

spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

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 25 M 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 whole organism tissue of the ilSabHarp1 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.

RNA library preparation and sequencing

Libraries were prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs), following the manufacturer's instructions. Poly(A) mRNA in the total RNA solution was isolated using oligo (dT) beads, converted to cDNA, and uniquely indexed; 14 PCR cycles were performed. Libraries were size-selected to produce fragments between 100–300 bp. Libraries were quantified, normalised, pooled to a final concentration of 2.8 nM, and diluted to 150 pM for loading. Sequencing was carried out on the Illumina NovaSeq X, generating paired-end reads.

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 (Cheng *et al.*, 2021) with the --primary option. The Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019), and the contigs were scaffolded in YaHS (Zhou *et al.*, 2023) with the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MerquryFK (Rhie *et al.*, 2020).

The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023).

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 nine breaks, 11 joins, and removal of two haplotypic duplications. This reduced the scaffold count by 7.9%. 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 MerquryFK tool (Rhie *et al.*, 2020) was run in a Singularity container (Kurtzer *et al.*, 2017) to evaluate *k*-mer completeness and assembly quality for the primary and alternate 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 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 *Sabra harpagula* specimen generated 21.23 Gb (gigabases) from 2.52 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 374.35 Mb, with a heterozygosity of 0.16% and repeat content of 21.87% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 55× coverage. Hi-C sequencing produced 117.13 Gb from 387.84 million reads, which were used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. Table 1 summarises the specimen and sequencing details.

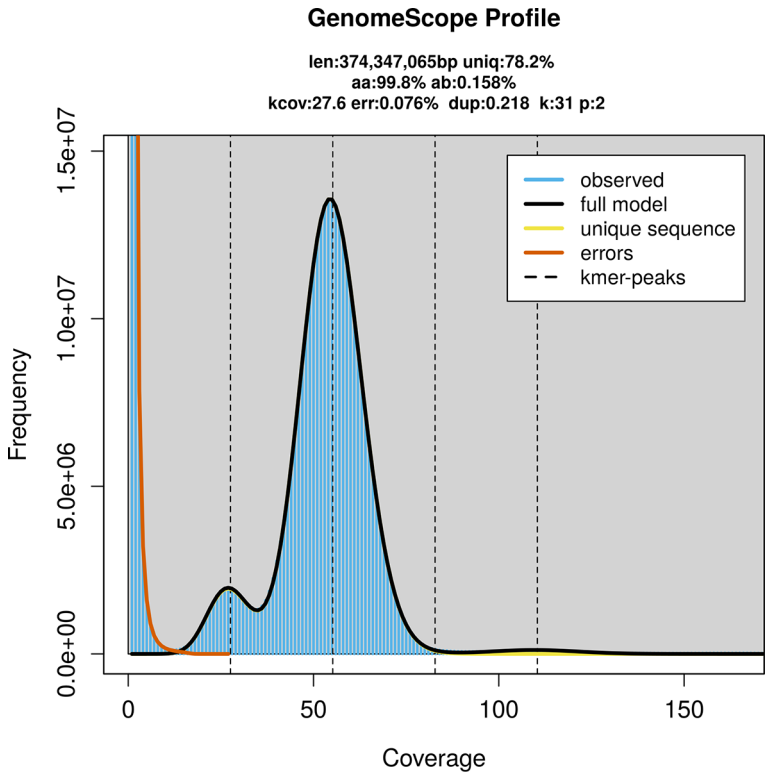

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. **Specimen and sequencing data for *Sabra harpagula* (BioProject PRJEB75058).**

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	ilSabHarp2	ilSabHarp1	ilSabHarp1
Specimen ID	NHMUK014438727	NHMUK014438725	NHMUK014438725
BioSample (source individual)	SAMEA112964414	SAMEA112964413	SAMEA112964413
BioSample (tissue)	SAMEA112975585	SAMEA112975583	SAMEA112975583
Tissue	abdomen	whole organism	whole organism
Instrument	Revio	Illumina NovaSeq X	Illumina NovaSeq X
Run accessions	ERR12921334	ERR12945495	ERR14792830
Read count total	2.52 million	387.84 million	38.34 million
Base count total	21.23 Gb	117.13 Gb	11.58 Gb

Table 2. **Genome assembly data for *Sabra harpagula*.**

Genome assembly	Primary assembly
Assembly name	ilSabHarp2.1
Assembly accession	GCA_964059465.1
Alternate haplotype accession	GCA_964059455.1
Assembly level	chromosome
Span (Mb)	377.24
Number of chromosomes	31
Number of contigs	115
Contig N50	5.45 Mb
Number of scaffolds	34
Scaffold N50	13.21 Mb
Sex chromosomes	Z
Organelles	Mitochondrion: 15.78 kb

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The final assembly has a total length of 377.24 Mb in 34 scaffolds, with 81 gaps, and a scaffold N50 of 13.21 Mb (Table 2).

Most of the assembly sequence (99.98%) 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). Z chromosome identified based on synteny with Drepana falcataria (GCA_945859725.1).

The mitochondrial genome was also assembled (length 15.78 kb, OZ060604.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

Assembly quality metrics

The combined primary and alternate assemblies achieve an estimated QV of 64.0. The *k*-mer completeness is 95.74% for the primary assembly, 70.34% for the alternate haplotype, and 99.77% for the combined assemblies (Figure 4).

BUSCO v.5.5.0 analysis using the lepidoptera_odb10 reference set (n = 5 286) identified 98.7% of the expected gene set (single = 98.4%, duplicated = 0.2%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for the primary assembly. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage.

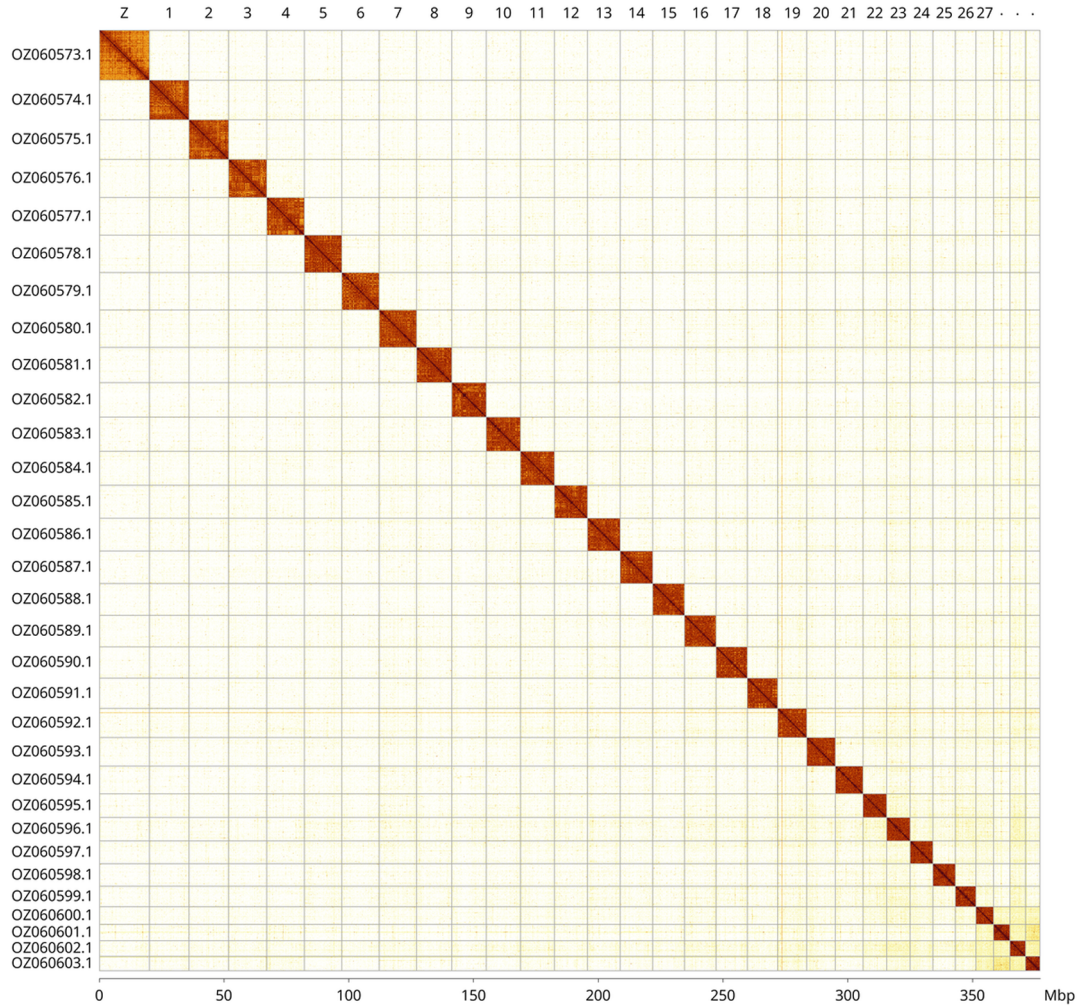

Figure 3. Hi-C contact map of the *Sabra harpagula* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Sabra harpagula* ilSabHarp2.

INSDC accession	Molecule	Length (Mb)	GC%
OZ060574.1	1	15.88	37.50
OZ060575.1	2	15.85	38
OZ060576.1	3	15.27	38
OZ060577.1	4	15.15	38
OZ060578.1	5	15	37.50
OZ060579.1	6	14.99	37.50
OZ060580.1	7	14.96	37.50
OZ060581.1	8	14.10	37.50
OZ060582.1	9	13.82	37.50
OZ060583.1	10	13.70	37.50
OZ060584.1	11	13.65	38
OZ060585.1	12	13.21	38

Table 3. *Continued*

INSDC accession	Molecule	Length (Mb)	GC%
OZ060586.1	13	13.10	38
OZ060587.1	14	13.07	38
OZ060588.1	15	12.69	38
OZ060589.1	16	12.66	38
OZ060590.1	17	12.58	38
OZ060591.1	18	12.15	38.50
OZ060592.1	19	11.63	38.50
OZ060593.1	20	11.48	38
OZ060594.1	21	11.09	38.50
OZ060595.1	22	9.46	38.50
OZ060596.1	23	9.42	39
OZ060597.1	24	9.16	39
OZ060598.1	25	9.02	39
OZ060599.1	26	8.20	39
OZ060600.1	27	7.10	39.50
OZ060601.1	28	6.49	40.50
OZ060602.1	29	6.29	40
OZ060603.1	30	5.84	40.50
OZ060573.1	Z	20.14	37.50

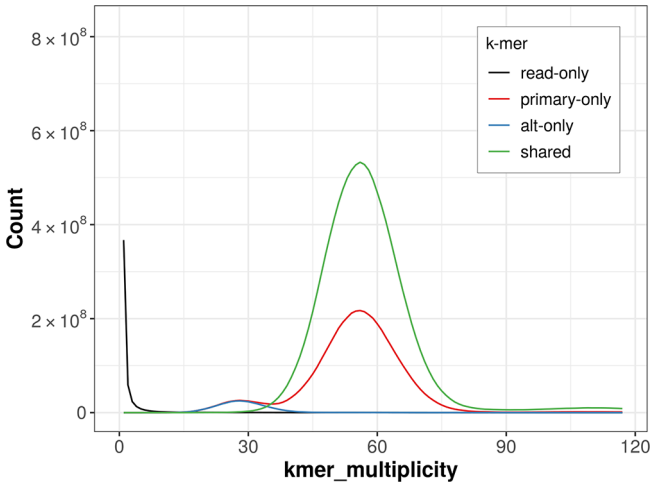

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards. The EBP metric, calculated for the primary assembly, is **6.C.Q65**, meeting the recommended reference standard.

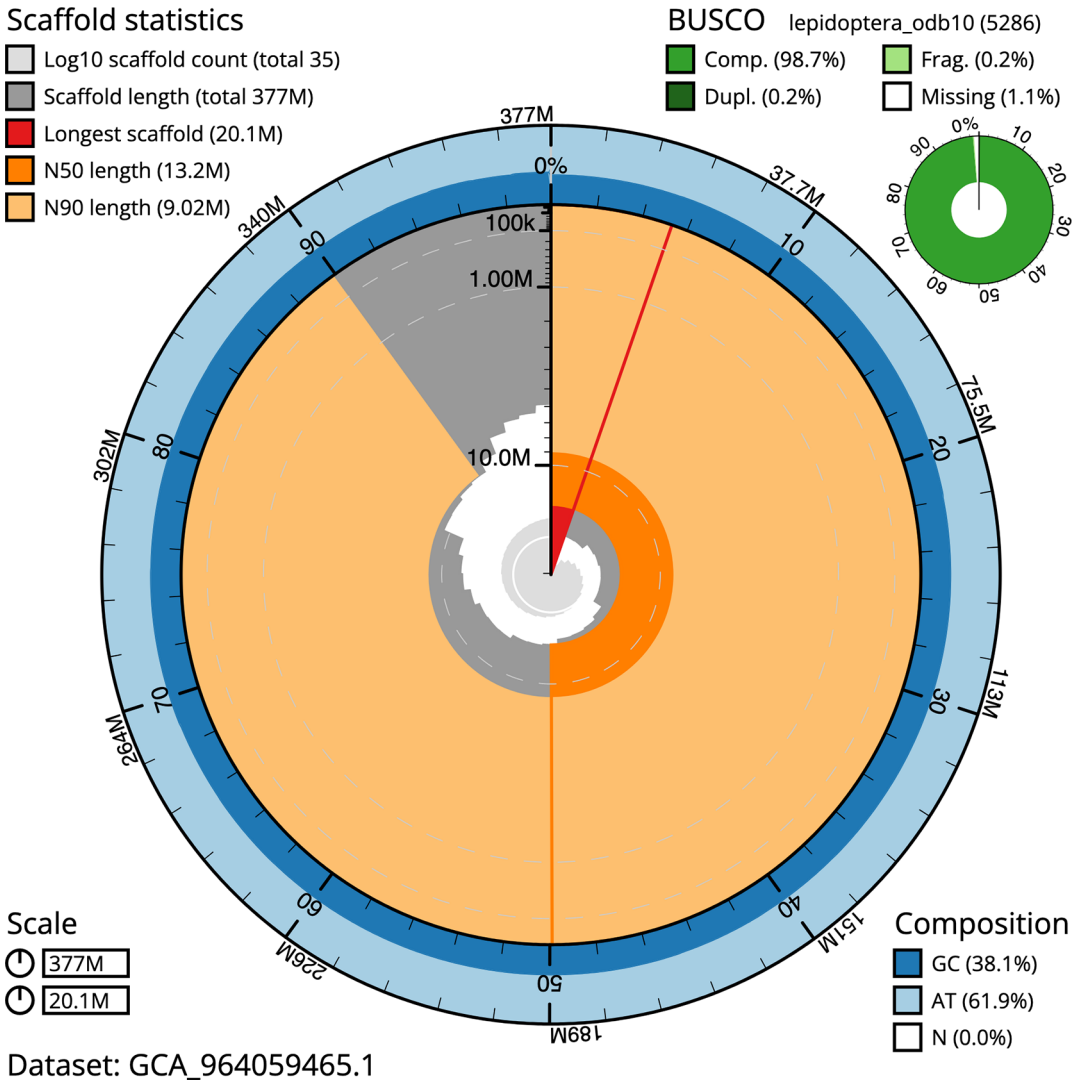

Figure 5. Assembly metrics for ilSabHarp2.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue 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 lepidoptera_odb10 set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

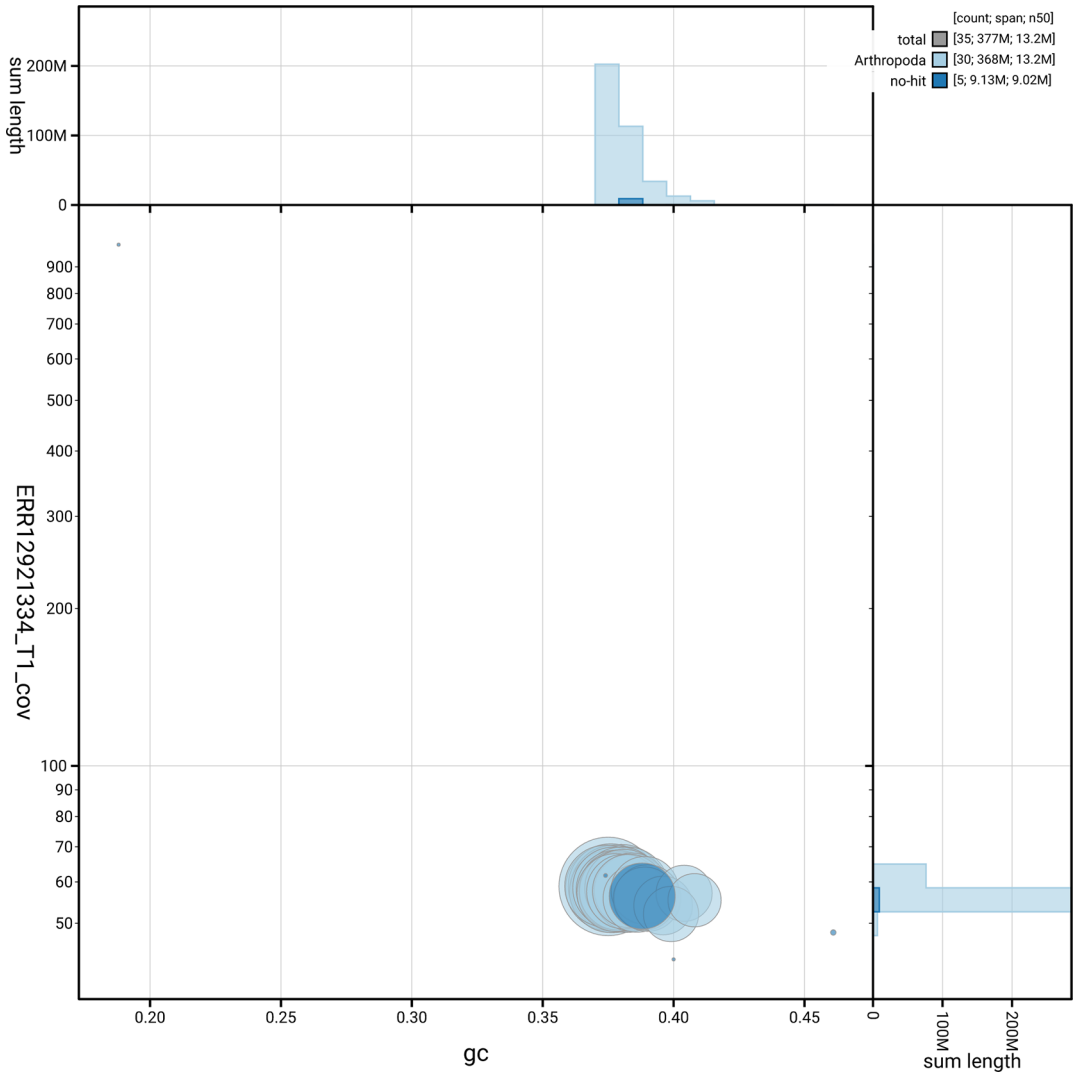

Figure 6. BlobToolKit blob plot for ilSabHarp2.1. The plot shows base coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total 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 *Sabra harpagula* assembly.

Measure	Value	Benchmark
EBP summary (primary)	6.C.Q65	6.C.Q40
Contig N50 length	5.45 Mb	≥ 1 Mb
Scaffold N50 length	13.21 Mb	= chromosome N50
Consensus quality (QV)	Primary: 65.5; alternate: 62.6; combined: 64.0	≥ 40
k-mer completeness	Primary: 95.74%; alternate: 70.34%; combined: 99.77%	≥ 95%
BUSCO	C:98.7% [S:98.4%, D:0.2%], F:0.2%, M:1.1%, n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.98%	≥ 90%

Notes: The 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.

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

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:

- 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: Sabra harpagula. Accession number PRJEB75058. The genome sequence is released openly for reuse. The *Sabra harpagula* 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 Tables 1 and 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.

Table 5. Software versions and sources used for *Sabra harpagula*.

Software	Version	Source
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
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond

Table 5. *Continued*

Software	Version	Source
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
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	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	1.0.3	https://github.com/sanger-tol/PretextView
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
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	1.2a.2	https://github.com/c-zhou/yahs

References

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 M-J, 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 H-G: **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 Genes|Genomes|Genetics.* 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

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial Arthropod fauna.** *Wellcome Open Research.* 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): giab008.
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.
Publisher Full Text | Free 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

Kimber I: **Scarce hook-tip *Sabra harpagula*.** 2026.
Reference Source

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.
Publisher Full Text

O'Leary NA, Cox E, Holmes JB, *et al.*: **Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets.** *Sci. Data.* 2024; **11**(1): 732.
PubMed Abstract | Publisher Full Text | Free Full Text

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.* 2020; **2020**: baaa062.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life Project.** *Wellcome Open Res.* 2024; **9**: 339.
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.** *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–24.
Publisher Full Text

Waring P, Townsend M, Tunmore M, *et al.*: *Field Guide to the Moths of Great Britain and Ireland.* London: Bloomsbury Wildlife; 2017.

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 June 2026

https://doi.org/10.21956/wellcomeopenres.29328.r156832

✔ **Gulab D Khedkar**
Dr Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India

Data note title:
The genome sequence of the Scarce Hook-tip, Sabra harpagula (Esper, 1786) (Lepidoptera: Drepanidae)

Peer review report
This manuscript presents a high-quality chromosome-level genome assembly for *Sabra harpagula*. The study provides the first high-quality genomic resource for the genus *Sabra* and represents a valuable contribution to the growing genomic resources available for Lepidoptera. The manuscript (Data Note) is well written, clearly organized, and provides sufficient biological background on the species. The sequencing strategy, assembly workflow, scaffolding, curation procedures, and quality assessment methods are comprehensive and follow current best practices in genome assembly. The authors provide detailed methodological descriptions, ensuring transparency and reproducibility.
The assembly metrics are excellent, with near-complete chromosomal assignment, high BUSCO completeness, strong consensus quality values, and high k-mer completeness. The incorporation of PacBio HiFi sequencing, Hi-C scaffolding, and extensive quality-control analyses has resulted in a robust reference-quality genome assembly. Furthermore, the public availability of the raw sequencing data and assembly enhances the utility of this resource for the broader scientific community.
I have only one minor suggestion. The manuscript could briefly expand on the potential applications of this genome resource for future studies of Drepanidae evolution, comparative genomics, adaptation, and conservation biology. However, this is optional and does not affect the scientific quality of the work.
Overall, the manuscript is scientifically sound, technically rigorous, and provides an important genomic resource. I recommend the manuscript for indexing in its current form.

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, population genetics

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 June 2026

https://doi.org/10.21956/wellcomeopenres.29328.r156824

Fedor Čiampor
Slovak Academy of Sciences, Bratislava, Slovakia

This data note presents a high-quality chromosome-level genome assembly for Sabra harpagula, the Scarce Hook-tip moth. The study is clearly structured and provides short background on the species, including its distribution, life cycle, host plant, and taxonomic placement. The methodology is detailed, covering sample collection, DNA barcoding, nucleic acid extraction, PacBio HiFi sequencing, Hi-C scaffolding, RNA sequencing, assembly, and quality assessment.
The note is a concise but valuable. Its main contribution is the production of the first high-quality genome for the genus Sabra, adding an important resource for Lepidoptera genomics, biodiversity research, and future comparative studies.
I have two comments on the manuscript. The first sentence of the "Background" section sounds as if this were a species found only in Britain (or at most in Europe, according to the following paragraph). Actually, it is apparently widespread throughout the Palearctic. I think a mention of this would be appropriate in the manuscript.
Why is the purified DNA yield listed, but not the RNA yield?

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: animal DNA metabarcoding

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 02 June 2026

https://doi.org/10.21956/wellcomeopenres.29328.r156825

Rajib Deb
ICAR-National Research Centre, Guwahat, Assam, India

The manuscript presents a high-quality chromosome-level genome assembly for *Sabra harpagula* generated as part of the Darwin Tree of Life project. The rationale for dataset generation is clearly described, and the assembly provides a valuable genomic resource as the first high-quality genome for the genus *Sabra*.
The sequencing, assembly, scaffolding, curation, and quality assessment protocols are appropriate and technically sound. The assembly achieves excellent quality metrics, including high BUSCO completeness, strong consensus quality, and near-complete chromosome assignment.
Methods are described in sufficient detail to support reproducibility, with clear reporting of sample collection, laboratory procedures, software tools, and accession numbers. The raw data and assembled genome are publicly available and presented in a usable and accessible format.
I have no major concerns. The manuscript is scientifically sound and suitable for indexing in its current form. Minor optional improvements could include additional discussion of future genome annotation plans and chromosome nomenclature, but these are not required for publication.

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

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.