DATA NOTE

The genome sequence of the White-barred Clearwing, *Synanthedon spheciformis* (Denis & Schiffermüller), 1775 (Lepidoptera: Sesiidae)

[version 1; peer review: 3 approved]

Jaakko Pohjoismaki[1], Marko Mutanen[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 Eastern Finland, Joensuu, North Karelia, Finland
[2]University of Oulu, Oulu, Northern Ostrobothnia, Finland
[3]Wellcome Sanger Institute, Hinxton, England, UK

v1 **First published:** 28 Jul 2025, **10**:373
https://doi.org/10.12688/wellcomeopenres.24613.1
Latest published: 28 Jul 2025, **10**:373
https://doi.org/10.12688/wellcomeopenres.24613.1

Abstract

We present a genome assembly from a female specimen of *Synanthedon spheciformis* (White-barred Clearwing; Arthropoda; Insecta; Lepidoptera; Sesiidae). The assembly contains two haplotypes with total lengths of 388.57 megabases and 372.40 megabases. Most of haplotype 1 (99.88%) is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 19.11 kilobases.

Keywords

Keywords: Synanthedon spheciformis, White-barred Clearwing, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 1 28 Jul 2025	✓ view	✓ view	✓ view

1. **Shova Mishra**, University of Florida, Gainesville, USA
North Carolina State University, Raleigh, USA

2. **Jason Charamis**, Foundation for Research and Technology - Hellas, Irákleion, Greece

3. **Teodora Staykova**, Plovdiv University, Plovdiv, Bulgaria

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: Pohjoismaki J: Investigation, Resources, Writing – Original Draft Preparation; **Mutanen M**: Resources, Supervision; **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).
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: Pohjoismaki J, Mutanen M, Wright CJ *et al.* **The genome sequence of the White-barred Clearwing, *Synanthedon spheciformis* (Denis & Schiffermüller), 1775 (Lepidoptera: Sesiidae) [version 1; peer review: 3 approved]** Wellcome Open Research 2025, **10**:373 https://doi.org/10.12688/wellcomeopenres.24613.1

First published: 28 Jul 2025, **10**:373 https://doi.org/10.12688/wellcomeopenres.24613.1

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; Apoditrysia; Sesioidea; Sesiidae; Sesiinae; Synanthedonini; *Synanthedon*; *Synanthedon spheciformis* (Denis & Schiffermüller), 1775 (NCBI:txid233843)

Background

The White-barred Clearwing (*Synanthedon spheciformis*) is a diurnal moth in the family Sesiidae. With its slender black body, transparent wings, and a distinctive white bar across the abdomen, the moth mimics ichneumonid wasps, including their flight behaviour.

The species is widely distributed across Europe and parts of Asia (Fibiger & Kristensen, 1974; Silvonen *et al.*, 2014). It is typically found in open woodland margins, roadsides, clearcuts and powerline corridors with young alders (*Alnus* spp.) or birches (*Betula* spp.) (Vuola & Korpla, 1978).

Synanthedon spheciformis is univoltine, with adults flying between late May and July (Fibiger & Kristensen, 1974). The species is not rare, but its inconspicuous lifestyle and behaviour make it difficult to observe. Females lay eggs near the base of host alder or birch, preferring coppice shoots growing from cut stumps (Vuola & Korpla, 1978). The larvae bore into the wood, but are sap-feeders rather than true xylophages, relying on living trees for sustenance. They overwinter twice, and in spring, the full-grown larva constructs an emergence tunnel, extending 10–30 cm from the root base, in which it pupates. The larvae are heavily attacked by parasitoids, particularly by the tachinid fly *Leskia aurea* (Diptera: Tachinidae) (Pohjoismäki, 2019).

The high-quality genome assembly generated for *Synanthedon spheciformis* contributes to broader studies within the Sesiidae family, including research on the genetic basis of mimicry and metabolic pathways associated with the wood-boring lifestyle. Additionally, the genome provides a foundation for studying population genetics, host plant specialisation and evolutionary history of the Sesiidae family.

Methods

Sample acquisition

The specimen used for genome sequencing was an adult female *Synanthedon spheciformis* (specimen ID SAN28000001, ToLID ilSynSphe1; Figure 1), collected from Jaamankangas, Kontiolahti, North Karelia, Finland (latitude 62.6712, longitude 29.7359) on 15/06/2023. The specimen was collected and identified by Jaakko Pohjoismaki (University of Eastern Finland).

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 ilSynSphe1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the whole organism was homogenised by powermashing using a PowerMasher II tissue disruptor.

Figure 1. Voucher photograph of the *Synanthedon spheciformis* (ilSynSphe1) specimen used for genome sequencing.

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 30.66 ng/μL and a yield of 1,441.02 ng, with a fragment size of 16.3 kb. The 260/280 spectrophotometric ratio was 1.96, and the 260/230 ratio was 2.62. RNA was extracted from whole organism tissue of ilSynSphe1 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 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 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 ilSynSphe1 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 and equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq X.

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

RNA-seq 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)

Table 1. Specimen and sequencing data for BioProject PRJEB81618.

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	ilSynSphe1	ilSynSphe1	ilSynSphe1
Specimen ID	SAN28000001	SAN28000001	SAN28000001
BioSample (source individual)	SAMEA114539580	SAMEA114539580	SAMEA114539580
BioSample (tissue)	SAMEA114539614	SAMEA114539614	SAMEA114539614
Tissue	whole organism	whole organism	whole organism
Sequencing instrument	Revio	Illumina NovaSeq X	Illumina NovaSeq X
Run accessions	ERR13900451	ERR13907229	ERR14986720
Read count total	1.85 million	744.43 million	127.45 million
Base count total	18.65 Gb	112.41 Gb	19.24 Gb

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 to generate 150-bp 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 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 17 breaks, 27 joins, and removal of 12 haplotypic duplications. The curation process is documented 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

Chromosomal painting was performed using lep_busco_painter using Merian elements, which represent the 32 ancestral linkage groups in Lepidoptera (Wright *et al.*, 2024). Painting was based on gene locations from the lepidoptera_odb10 BUSCO analysis and chromosome lengths from the genome index produced using SAMtools faidx (Danecek *et al.*, 2021). Each complete BUSCO (including both single-copy and duplicated BUSCOs) was assigned to a Merian element using a reference database, and coloured positions were plotted along chromosomes drawn to scale.

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 implementation of the earlier Snakemake BlobToolKit pipeline (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. Simultaneously, it queries the GoaT database (Challis *et al.*, 2023) to identify relevant BUSCO lineages and runs BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO 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 package management via Conda and Bioconda (Grüning *et al.*, 2018), and containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

The genome of a specimen of *Synanthedon spheciformis* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 18.65 Gb (gigabases) from 1.85 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 379.74 Mb, with a heterozygosity of 0.24% and repeat content of 26.56%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 48× coverage. Hi-C sequencing produced 112.41 Gb from 744.43 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.

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 388.57 Mb in 43 scaffolds, with 83 gaps, and a scaffold N50 of 13.83 Mb (Table 2).

Most of the assembly sequence (99.88%) was assigned to 31 chromosomal-level scaffolds, representing 30 autosomes and the Z sex chromosome. The Z chromosome was identified by coverage and alignment to the genome of *Synanthedon formicaeformis* (GCA_945859745.1). There was no evidence of a W chromosome in this assembly, therefore it appears that this species exhibits the ZO karyotype. The order and orientation of scaffolds in the region ~4.71–5.63 Mb on Chromosome 10 is uncertain. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 2; Table 3). Chromosome painting with Merian elements

Table 2. Genome assembly statistics.

Assembly name	ilSynSphe1.hap1.1	ilSynSphe1.hap2.1
Assembly accession	GCA_964332195.1	GCA_964332305.1
Assembly level	chromosome	scaffold
Span (Mb)	388.57	372.40
Number of chromosomes	31	N/A
Number of contigs	126	141
Contig N50	6.69 Mb	6.74 Mb
Number of scaffolds	43	72
Scaffold N50	13.83 Mb	13.48 Mb
Longest scaffold length (Mb)	17.93	N/A
Sex chromosomes	Z	N/A
Organelles	Mitochondrial genome: 19.11 kb	N/A

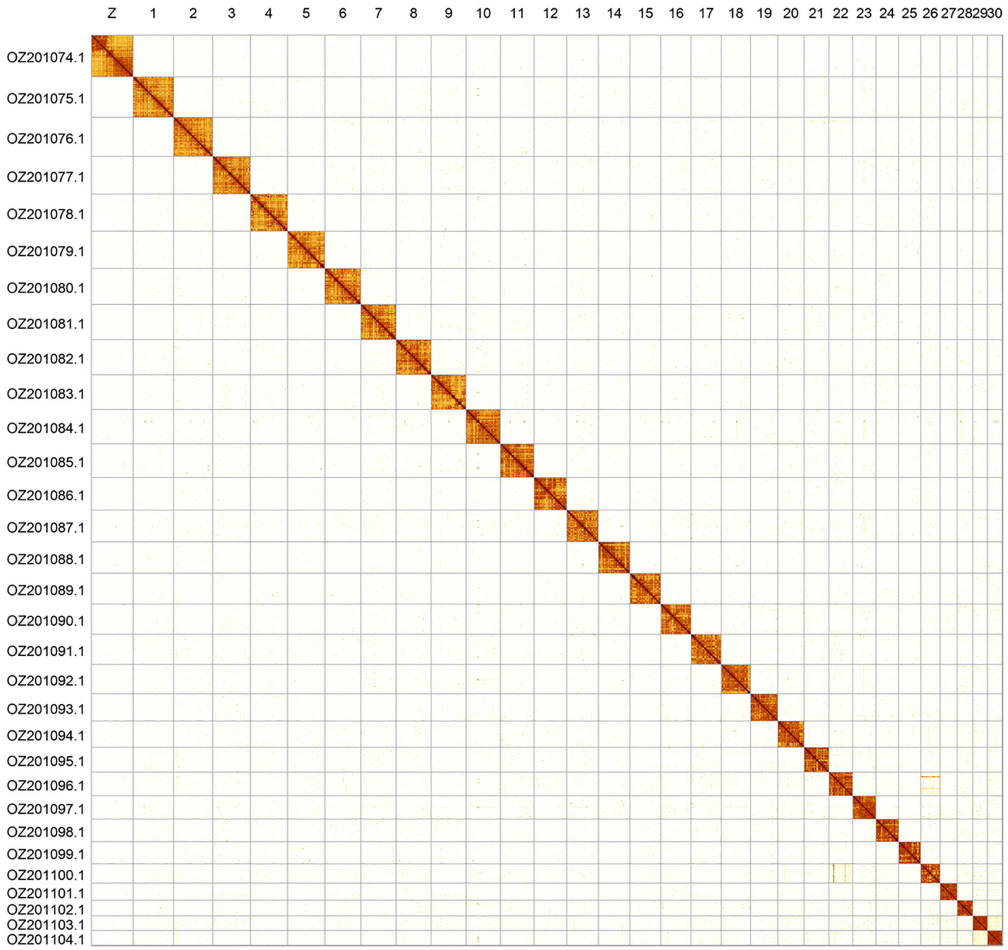

Figure 2. Hi-C contact map of the *Synanthedon spheciformis* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

illustrates the distribution of orthologues along chromosomes and highlights patterns of chromosomal evolution relative to Lepidopteran ancestral linkage groups (Figure 3).

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

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Synanthedon spheciformis* ilSynSphe1.

INSDC accession	Molecule	Length (Mb)	GC%	Assigned Merian elements
OZ201075.1	1	17.22	34.50	M2
OZ201076.1	2	16.73	35	M1
OZ201077.1	3	15.98	34.50	M8
OZ201078.1	4	15.86	35	M3
OZ201079.1	5	15.81	35	M17;M20
OZ201080.1	6	15.39	34.50	M5
OZ201081.1	7	15.06	34	M9
OZ201082.1	8	14.95	34.50	M12
OZ201083.1	9	14.76	34.50	M18
OZ201084.1	10	14.64	35.50	M7
OZ201085.1	11	14.40	35	M16
OZ201086.1	12	13.83	34.50	M6
OZ201087.1	13	13.59	35	M22
OZ201088.1	14	13.41	35	M4
OZ201089.1	15	13.16	34.50	M21
OZ201090.1	16	12.91	34.50	M15
OZ201091.1	17	12.78	35	M11
OZ201092.1	18	12.53	35	M10
OZ201093.1	19	11.63	35	M13
OZ201094.1	20	11.12	35	M14
OZ201095.1	21	10.62	35	M23
OZ201096.1	22	10.14	36	M26
OZ201097.1	23	9.85	35.50	M19
OZ201098.1	24	9.67	35	M24
OZ201099.1	25	9.37	35	M28
OZ201100.1	26	8.36	35.50	M27
OZ201101.1	27	7.23	35.50	M25
OZ201102.1	28	6.63	35.50	M29
OZ201103.1	29	6.28	36.50	M31
OZ201104.1	30	6.25	36.50	M30
OZ201074.1	Z	17.93	34.50	MZ
OZ201105.1	MT	0.02	22	N/A

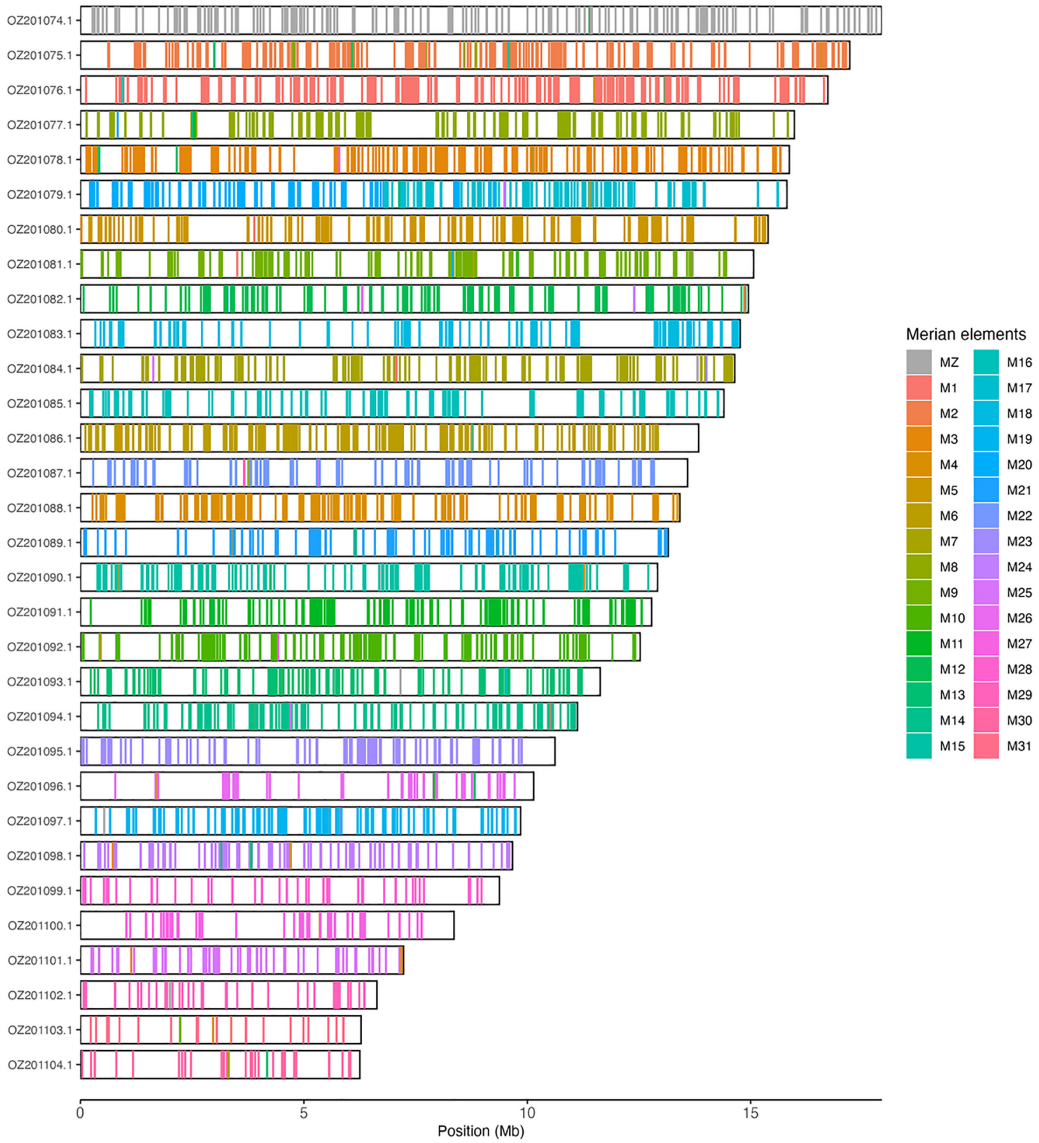

Figure 3. Merian elements painted across chromosomes in the ilSynSphe1.hap1.1 assembly of *Synanthedon spheciformis*. Chromosomes are drawn to scale, with the positions of orthologues shown as coloured bars. Each orthologue is coloured by the Merian element that it belongs to. All orthologues which could be assigned to Merian elements are shown.

Assembly quality metrics

For haplotype 1, the estimated QV is 59.4, and for haplotype 2, 59.5. When the two haplotypes are combined, the assembly achieves an estimated QV of 59.4. The *k*-mer completeness is 95.53% for haplotype 1, 91.44% for haplotype 2, and 99.59% for the combined haplotypes (Figure 4). BUSCO analysis using the lepidoptera_odb10 reference set (n = 5 286) (Kriventseva *et al.*, 2019) identified 98.1% of the expected gene set (single = 97.4%, duplicated = 0.6%) for haplotype 1. The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.Q59**, 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

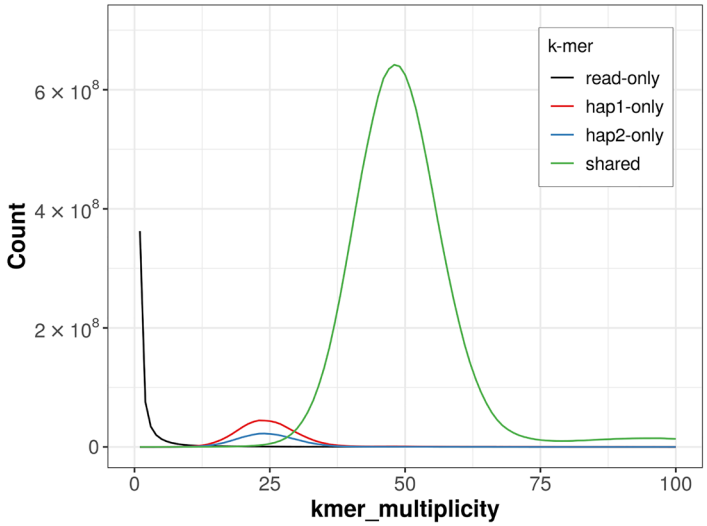

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve (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.

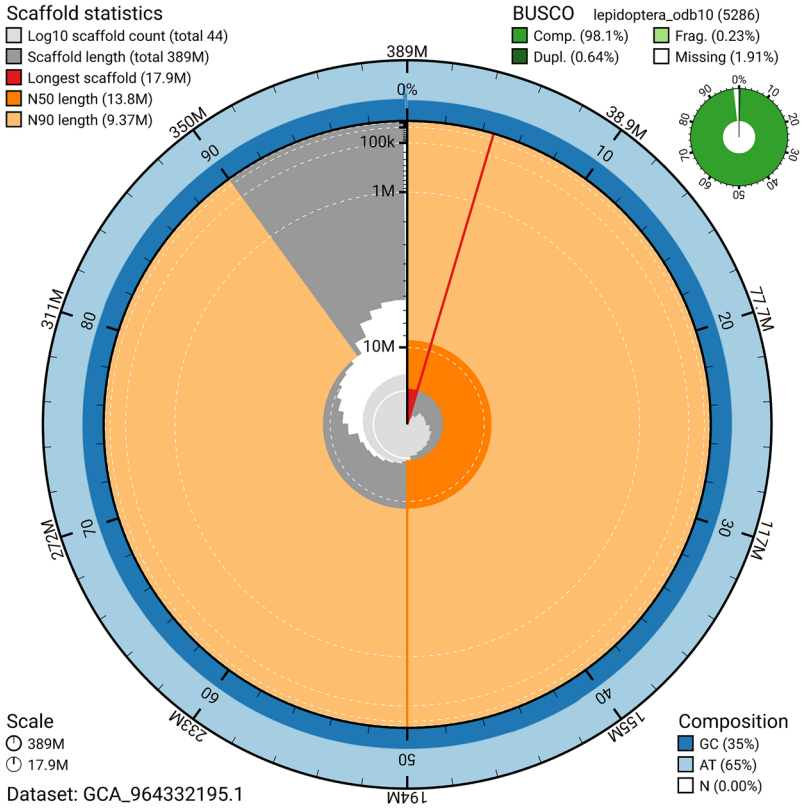

Figure 5. Assembly metrics for ilSynSphe1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost 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.

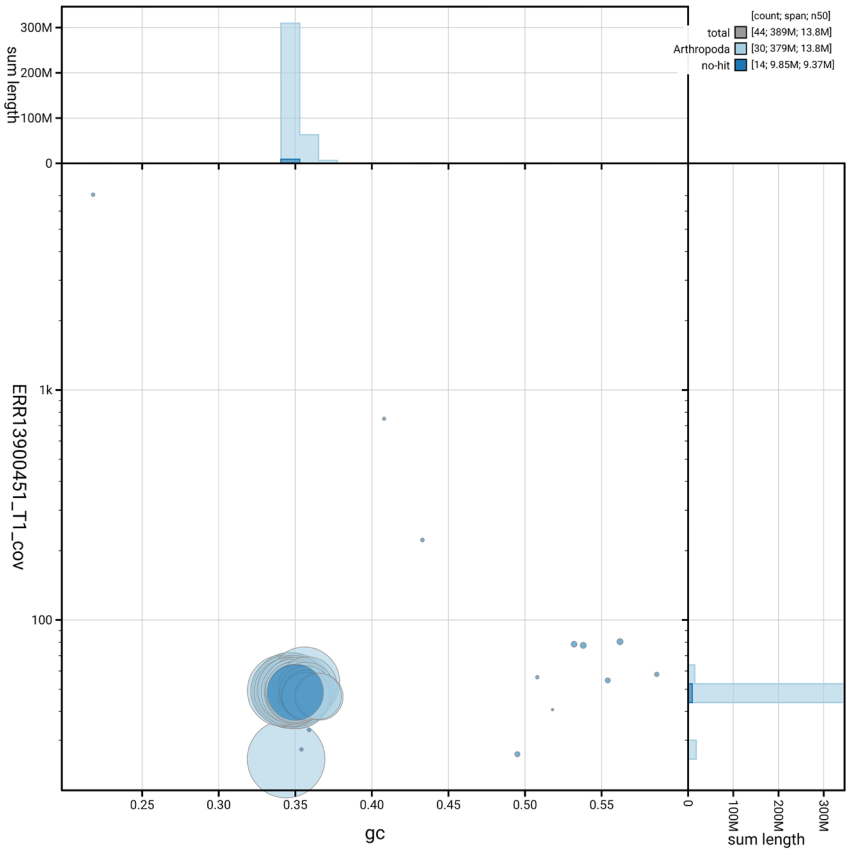

Figure 6. BlobToolKit GC-coverage plot for ilSynSphe1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Synanthedon spheciformis* assembly.

Measure (Benchmark)	Value
EBP summary (haplotype 1)	6.C.Q59
Contig N50 length (≥ 1 Mb)	6.69 Mb
Scaffold N50 length (= chromosome N50)	13.83 Mb
Consensus quality (QV) (≥ 40)	Haplotype 1: 59.4; haplotype 2: 59.5; combined: 59.4
k-mer completeness (≥ 95%)	Haplotype 1: 95.53%; Haplotype 2: 91.44%; combined: 99.59%
BUSCO* (S > 90%; D < 5%)	C:98.1%[S:97.4%,D:0.6%],F:0.2%,M:1.7%,n:5286
Percentage of assembly assigned to chromosomes (≥ 90%)	99.88%

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.

Data availability

European Nucleotide Archive: Synanthedon spheciformis (white-barred clearwing). Accession number PRJEB81618. The genome sequence is released openly for reuse. The *Synanthedon spheciformis* 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	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
lep_busco_painter	1.0.0	https://github.com/charlottewright/lep_busco_painter
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextSnapshot	N/A	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.6.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

Author information

- 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.

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, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24.
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): 1332–35.
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.
PubMed Abstract | Publisher Full Text

Fibiger M, Kristensen NP: **The Sesiidae (Lepidoptera) of Fennoscandia and Denmark.** Gadstrup: Scandinavian Science Press Ltd, 1974; **2**.
Reference Source

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

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
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): D807–11.
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

Pohjoismäki J: **Manipulation of host behavior by *Leskia aurea* (Fallén) (Tachinidae).** *Tachinid Times.* 2019; **32**: 37–39.
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

Silvonen K, Top-Jensen M, Fibiger M: **Suomen Päivä- Ja Yöperhoset - Maastokäsikirja [a Field Guide to the Butterflies and Moths of Finland. In Finnish].** Oestermarie: Bugbook Publishing, 2014.
Reference Source

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

Vuola M, Korpla S: **Suomen lasisiipisten (Sesiidae) ja puuntuhoojien (Cossidae) elintavoista (Lepidoptera). 4. Lepän lasisiipi (*Synanthedon spheciformis*) ja pohjan lasisiipi (*S. polaris*). [On the habits of Finnish clearwing moths (Sesiidae) and goat moths (Cossidae). Part 4. and *S. polaris.* In Finnish].** *Notulae Entomologicae.* 1978; **58**: 101–5.

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 30 September 2025

https://doi.org/10.21956/wellcomeopenres.27123.r131954

✔ **Teodora Staykova**
Plovdiv University, Plovdiv, Bulgaria

The genome of *Synanthedon speciformis* (Lepidoptera: Sesiidae) was studied based on a female individual. The methods used are described in detail. Two haplotypes were identified. The majority of haplotype 1 (99.88%) is scaffolded in 31 chromosomal pseudomolecules, of which 30 are autosomes and one sex chromosome. The sex Z chromosome was identified by coverage and alignment with the genome of Synanthedon formicaeformis (GCA_945859745.1). No second sex chromosome was identified during the study, leading to the conclusion that female heterogamety of the ZO type is observed in this species.
In addition to presenting new interesting genetic data for the species, the study is also significant because it provides a basis for studying population genetic characteristics, host plant specialization and the evolutionary history of the Sesiidae family.

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: Genetics, Population Genetics, Genetics of Bombyx mori

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 19 September 2025

https://doi.org/10.21956/wellcomeopenres.27123.r131949

Jason Charamis
Foundation for Research and Technology - Hellas, Irákleion, Greece

This manuscript describes the genome assembly of the White-barred Clearwing, *Synanthedon spheciformis* . The work is technically sound, and the authors have provided a very good dataset and a good rationale about the usefulness of this resource in future studies. I have no further comments or suggestions to make; the manuscript is ready for indexing.

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: Arthropod 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 16 September 2025

https://doi.org/10.21956/wellcomeopenres.27123.r130965

Shova Mishra
[1] University of Florida, Gainesville, USA
[2] North Carolina State University, Raleigh, USA

The article presents a genome sequence of *Synanthedon spheciformis,* including two haplotypes with different genome sizes. The manuscript is well written with clear introduction, methods, and results. Associated data sets are made publicly available which is very useful for the scientific community.
The paper describes the RNA-seq method briefly. However, there is no corresponding results presented, nor the data is uploaded. Keeping this in mind, it may not be necessary to include RNA-seq section in the method section.

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: Nematology, Genomics, Transcriptomics, Nematicide resitance

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.