DATA NOTE

The genome sequence of the Red-belted Clearwing, *Synanthedon myopaeformis* (Borkhausen, 1789) [version 1; peer review: 1 approved, 1 approved with reservations]

William B. V. Langdon[1], Peter W. H. Holland[1],
University of Oxford and Wytham Woods 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]University of Oxford, Oxford, England, UK

v1 **First published:** 20 Mar 2024, **9**:162
https://doi.org/10.12688/wellcomeopenres.21111.1
Latest published: 20 Mar 2024, **9**:162
https://doi.org/10.12688/wellcomeopenres.21111.1

Abstract

We present a genome assembly from an individual male *Synanthedon myopaeformis* (the Red-belted Clearwing; Arthropoda; Insecta; Lepidoptera; Sesiidae). The genome sequence is 295.7 megabases in span. Most of the assembly is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled and is 15.18 kilobases in length. Gene annotation of this assembly on Ensembl identified 15,959 protein coding genes.

Keywords

Synanthedon myopaeformis, Red-belted Clearwing, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ? ✓

	1	2
version 1 20 Mar 2024	? view	✓ view

1. **Josephine Paris**, Marche Polytechnic University, Ancona, Italy

2. **Kyle M Benowitz**, Austin Peay State University, Clarksville, USA

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: Langdon WBV: Investigation, Resources; **Holland PWH**: Writing – Original Draft Preparation, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute [206194, https://doi.org/10.35802/206194] 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: Langdon WBV, Holland PWH, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Red-belted Clearwing, *Synanthedon myopaeformis* (Borkhausen, 1789) [version 1; peer review: 1 approved, 1 approved with reservations]** Wellcome Open Research 2024, **9**:162 https://doi.org/10.12688/wellcomeopenres.21111.1

First published: 20 Mar 2024, **9**:162 https://doi.org/10.12688/wellcomeopenres.21111.1

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 myopaeformis* (Borkhausen, 1789) (NCBI: txid1108570).

Background

The clearwing moths, family Sesiidae, are a group of Lepidoptera known for their transparent scale-free regions on their wings and for their visual and behavioural mimicry of wasps and bees. *Synanthedon myopaeformis* is a widely distributed member of the family; it is commonly called the Red-belted Clearwing in Britain and the Apple Clearwing in the United States and Canada. The moth has been recorded across much of Europe, with a concentration of records from northern and central regions and sporadic southern records from Italy, Greece and Spain including Gran Canaria (GBIF Secretariat, 2023). In Britain, the species is commonest in south-eastern counties, where it may be observed in summer resting on the trunks of trees in the mornings or flying on sunny afternoons (South, 1961); it is most readily found by using pheromone lures (Clifton, no date; Larsson, 2016). There are also scattered records from Asia, including from Kazakhstan, Kyrgyzstan and Russia (GBIF Secretariat, 2023). Since 2005, there have also been records from North America following accidental introductions (Beaton & Carter, 2006; Judd & Philip, 2006).

The larvae of *S. myopaeformis* feed beneath the bark of apple and other Rosaceaeous trees, developing for up to two years and forming extensive feeding galleries. When adults emerge, the empty pupal case is often left projecting from the exit hole (South, 1961). The larvae can cause significant damage to the host tree, causing weakening and allowing secondary pathogen infections and further pest invasion. Consequently, the species has become a serious pest of apple orchards in Bulgaria (Németh-Major *et al.*, 2017), Turkey (Özpinar *et al.*, 2009) and Jordan (Ateyyat & Al-Antary, 2006), amongst other countries. Since the species was introduced into Canada and the United States around 2005, it has also become an established pest of apple orchards (Beaton & Carter, 2006; Judd & Philip, 2006; LaGasa, 2009).

Here we report a complete genome sequence for the Red-belted Clearwing *Synanthedon myopaeformis* determined as part of the Darwin Tree of Life project. The genome sequence of *S. myopaeformis* may prove beneficial in designing control strategies and will contribute to the growing set of resources for studying the evolution of Lepidoptera.

Genome sequence report

The genome was sequenced from one male *Synanthedon myopaeformis* (Figure 1) collected from Wytham Woods, Oxfordshire, UK (51.78, –1.32). A total of 77-fold coverage in Pacific Biosciences single-molecule HiFi long reads was generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected two missing joins or mis-joins.

Figure 1. Photograph of the *Synanthedon myopaeformis* (ilSynMyop1) specimen used for genome sequencing.

The final assembly has a total length of 295.7 Mb in 32 sequence scaffolds with a scaffold N50 of 10.8 Mb (Table 1). The snail plot in Figure 2 provides a summary of the assembly statistics, while the distribution of assembly scaffolds on GC proportion and coverage is shown in Figure 3. The cumulative assembly plot in Figure 4 shows curves for subsets of scaffolds assigned to different phyla. Most (99.97%) of the assembly sequence was assigned to 31 chromosomal-level scaffolds, representing 30 autosomes and the Z sex chromosome. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 2). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial genome was also assembled and can be found as a contig within the multifasta file of the genome submission.

The estimated Quality Value (QV) of the final assembly is 61.9 with *k*-mer completeness of 100.0%, and the assembly has a BUSCO v5.3.2 completeness of 98.0% (single = 97.4%, duplicated = 0.5%), using the lepidoptera_odb10 reference set (n = 5,286).

Metadata for specimens, barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly statistics are given at https://links.tol.sanger.ac.uk/species/1108570.

Genome annotation report

The *Synanthedon myopaeformis* genome assembly (GCA_944738685.1) was annotated at the European Bioinformatics Institute (EBI) on Ensembl Rapid Release. The resulting annotation includes 16,108 transcribed mRNAs from 15,929 protein-coding genes (Table 1; https://rapid.ensembl.org/Synanthedon_myopaeformis_GCA_944738685.1/Info/Index).

Table 1. Genome data for *Synanthedon myopaeformis*, ilSynMyop1.1.

Project accession data		
Assembly identifier	ilSynMyop1.1	
Species	*Synanthedon myopaeformis*	
Specimen	ilSynMyop1	
NCBI taxonomy ID	1108570	
BioProject	PRJEB53610	
BioSample ID	SAMEA10979197	
Isolate information	ilSynMyop1, male: head and thorax (DNA and Hi-C sequencing); abdomen (RNA sequencing)	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	61.9	*≥ 50*
k-mer completeness	100.0%	*≥ 95%*
BUSCO**	C:98.0%[S:97.4%,D:0.5%], F:0.5%,M:1.6%,n:5,286	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	99.97%	*≥ 95%*
Sex chromosomes	ZZ	*localised homologous pairs*
Organelles	Mitochondrial genome: 15.18 kb	*complete single alleles*
Raw data accessions		
PacificBiosciences SEQUEL II	ERR9854839	
Hi-C Illumina	ERR9866452	
PolyA RNA-Seq Illumina	ERR10123709	
Genome assembly		
Assembly accession	GCA_944738685.1	
Accession of alternate haplotype	GCA_944738625.1	
Span (Mb)	295.7	
Number of contigs	37	
Contig N50 length (Mb)	10.4	
Number of scaffolds	32	
Scaffold N50 length (Mb)	10.8	
Longest scaffold (Mb)	13.68	
Genome annotation		
Number of protein-coding genes	15,929	
Number of gene transcripts	16,108	

* Assembly metric benchmarks are adapted from column VGP-2020 of "Table 1: Proposed standards and metrics for defining genome assembly quality" from Rhie *et al.* (2021).

** BUSCO scores based on the lepidoptera_odb10 BUSCO set using version 5.3.2. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/idConQuad1_1/dataset/idConQuad1_1/busco.

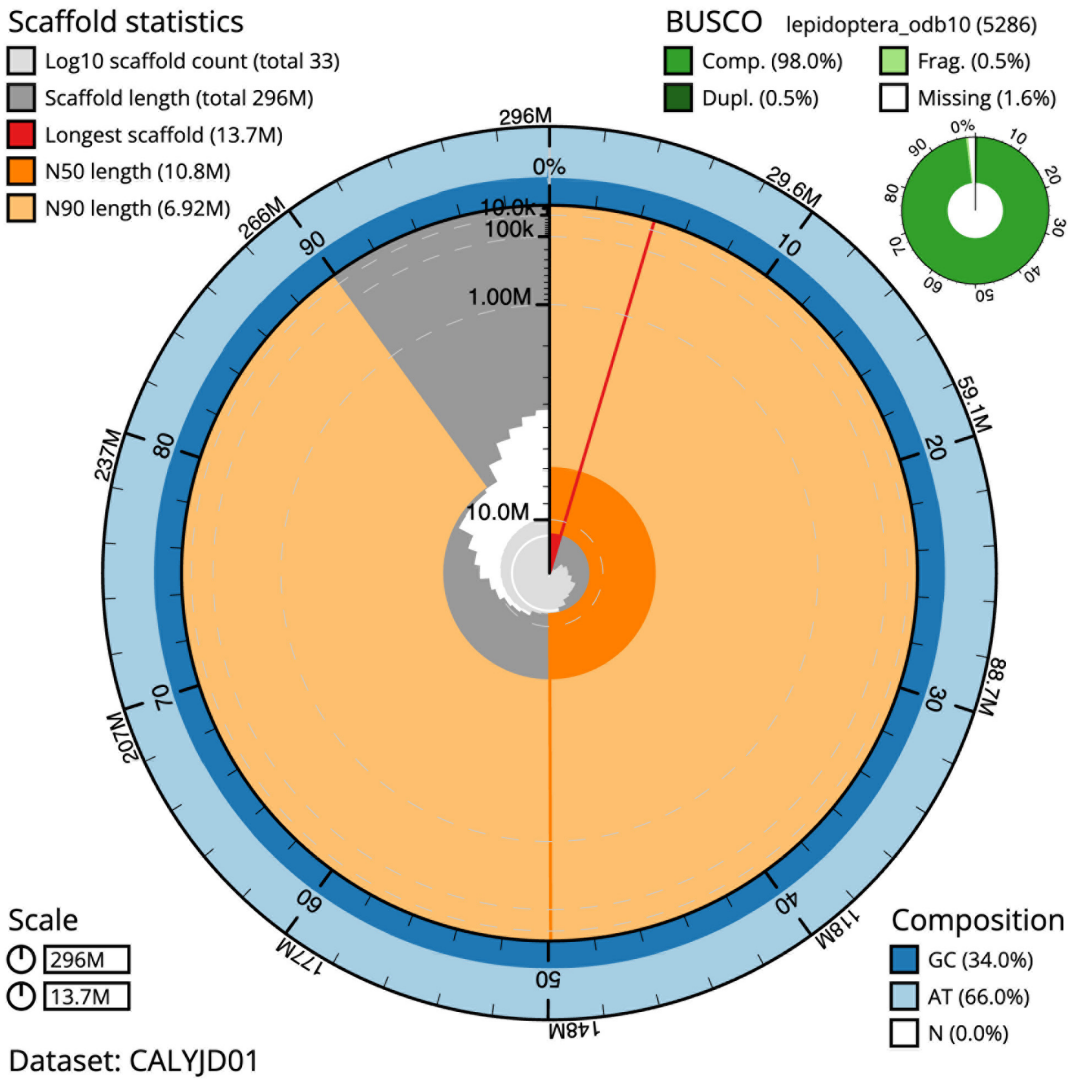

Figure 2. Genome assembly of *Synanthedon myopaeformis*, ilSynMyop1.1: metrics. The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 295,689,421 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (13,676,935 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (10,766,502 and 6,917,070 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the lepidoptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/CALYJD01/dataset/CALYJD01/snail.

Methods

Sample acquisition and nucleic acid extraction

A male *Synanthedon myopaeformis* (specimen ID Ox001934, ToLID ilSynMyop1) was collected from Wytham Woods, Oxfordshire (biological vice-county Berkshire), UK (latitude 51.78, longitude –1.32) on 2021-06-16 using a pheromone lure. The specimen was collected and identified by Will Langdon (University of Oxford) and the preserved on dry ice.

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) includes a sequence of core procedures: sample preparation; sample homogenisation, DNA extraction, fragmentation, and clean-up. In sample preparation, the ilSynMyop1 sample was weighed and dissected on dry ice (Jay *et al.*, 2023). Tissue from the head and thorax was homogenised using a PowerMasher II tissue disruptor, setting aside tissue for Hi-C seqiencing (Denton *et al.*, 2023a). HMW DNA was extracted using the Automated MagAttract v1 protocol (Sheerin *et al.*, 2023). DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Todorovic *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation (Strickland *et al.*, 2023): in brief, the method employs a 1.8X ratio of AMPure PB beads to sample to eli

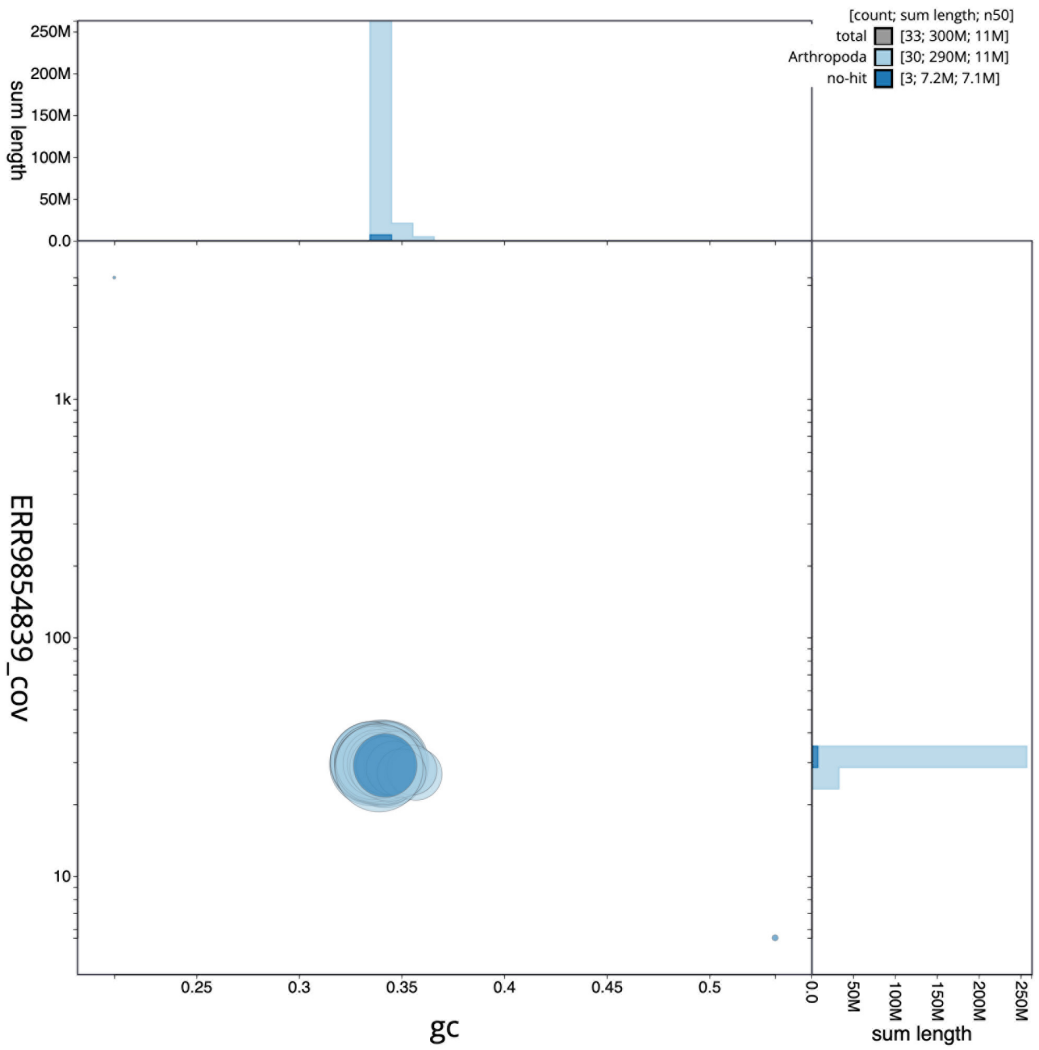

Figure 3. Genome assembly of *Synanthedon myopaeformis*, ilSynMyop1.1: BlobToolKit GC-coverage plot. Sequences are coloured by phylum. Circles are sized in proportion to sequence length. Histograms show the distribution of sequence length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/CALYJD01/dataset/CALYJD01/blob.

fragments and concentrate the DNA. The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

RNA was extracted from abdomen tissue of ilSynMyop1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol (do Amaral *et al.*, 2023). 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.

Protocols developed by the WSI Tree of Life laboratory are publicly available on protocols.io (Denton *et al.*, 2023b).

Sequencing

Pacific Biosciences HiFi circular consensus DNA sequencing libraries were constructed according to the manufacturers' instructions. Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit. DNA and RNA sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi) and Illumina NovaSeq 6000 (RNA-Seq) instruments. Hi-C data were also generated from remaining head and thorax tissue of ilSynMyop1 using the Arima2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021) and haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2020). The assembly was then scaffolded with Hi-C data (Rao *et al.*, 2014) using YaHS

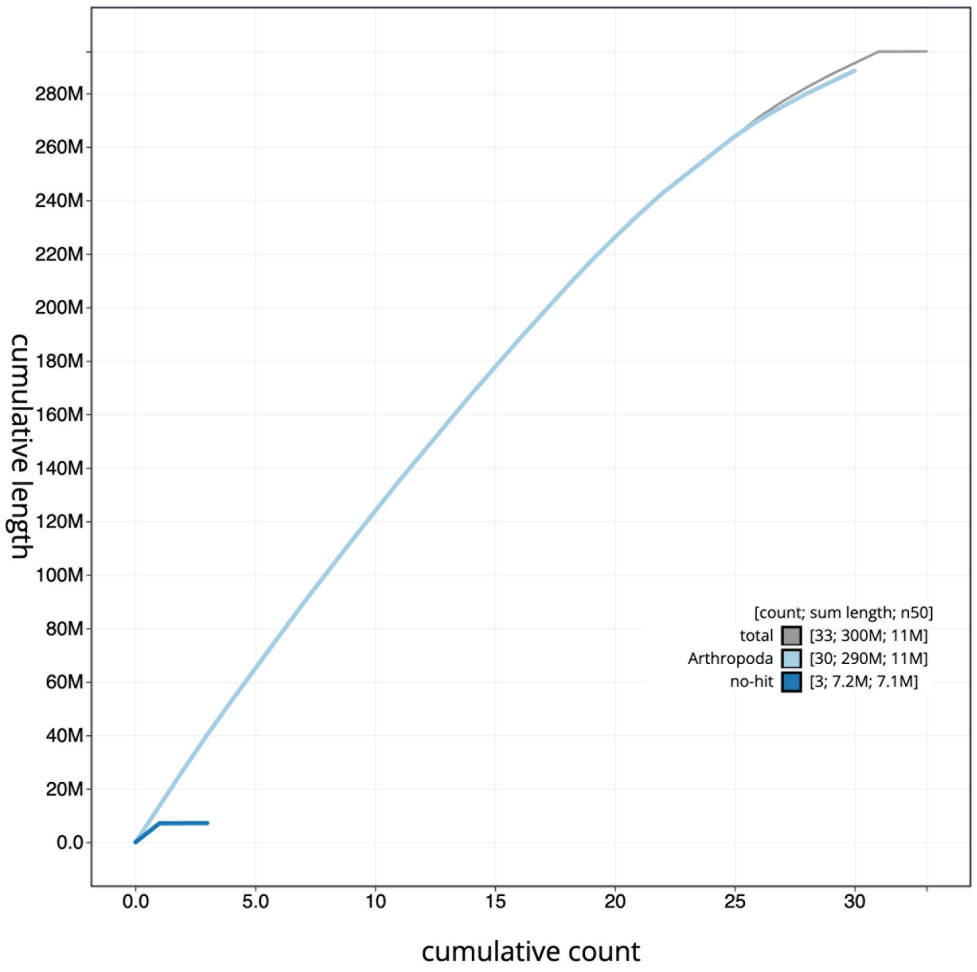

Figure 4. Genome assembly of *Synanthedon myopaeformis*, ilSynMyop1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all sequences. Coloured lines show cumulative lengths of sequences assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/CALYJD01/dataset/CALYJD01/cumulative.

(Zhou *et al.*, 2023). The assembly was checked for contamination and corrected as described previously (Howe *et al.*, 2021). Manual curation was performed using HiGlass (Kerpedjiev *et al.*, 2018) and PretextView (Harry, 2022). The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) or MITOS (Bernt *et al.*, 2013) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). To assess the assembly metrics, the *k*-mer completeness and QV consensus quality values were calculated in Merqury (Rhie *et al.*, 2020). This work was done using Nextflow (Di Tommaso *et al.*, 2017) DSL2 pipelines "sanger-tol/readmapping" (Surana *et al.*, 2023a) and "sanger-tol/genomenote" (Surana *et al.*, 2023b). The genome was analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021; Simão *et al.*, 2015) were calculated.

Table 3 contains a list of relevant software tool versions and sources.

Genome annotation

The BRAKER2 pipeline (Bruna *et al.*, 2021) was used in the default protein mode to generate annotation for the *Synanthedon*

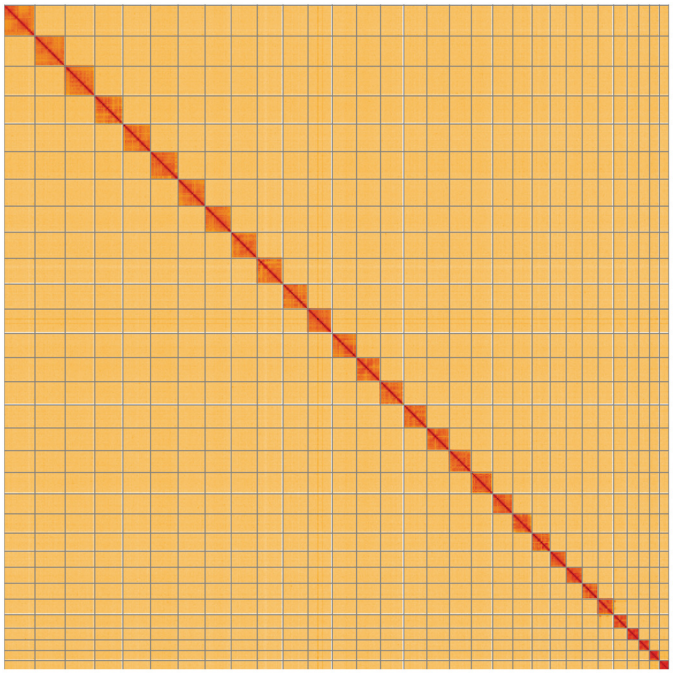

Figure 5. Genome assembly of *Synanthedon myopaeformis*, ilSynMyop1.1: Hi-C contact map of the ilSynMyop1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=VKNqRfW5R0aiUV6wt18nOw.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Synanthedon myopaeformis*, ilSynMyop1.

INSDC accession	Chromosome	Length (Mb)	GC%
OX122914.1	1	13.45	34.0
OX122915.1	2	13.18	34.0
OX122916.1	3	12.51	34.0
OX122917.1	4	12.28	33.5
OX122918.1	5	12.26	34.0
OX122919.1	6	11.99	33.5
OX122920.1	7	11.61	33.5
OX122921.1	8	11.6	33.5
OX122922.1	9	11.47	33.5
OX122923.1	10	11.11	34.0
OX122924.1	11	10.82	34.0
OX122925.1	12	10.77	34.0
OX122926.1	13	10.69	34.0
OX122927.1	14	10.36	33.5
OX122928.1	15	10.21	34.0
OX122929.1	16	10.07	33.5
OX122930.1	17	9.75	34.0
OX122931.1	18	9.49	33.5
OX122932.1	19	8.92	34.0
OX122933.1	20	8.58	34.5
OX122934.1	21	8.03	34.0
OX122935.1	22	7.13	34.0
OX122936.1	23	7.1	34.5
OX122937.1	24	7.09	34.0
OX122938.1	25	6.92	34.0
OX122939.1	26	5.98	34.0
OX122940.1	27	5.2	34.5
OX122941.1	28	4.78	35.5
OX122942.1	29	4.4	35.5
OX122943.1	30	4.19	35.0
OX122913.1	Z	13.68	34.0
OX122944.1	MT	0.02	21.5

Table 3. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	3.5.2	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.3.2	https://gitlab.com/ezlab/busco
Hifiasm	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	1.11.6	https://github.com/higlass/higlass
Merqury	MerquryFK	https://github.com/thegenemyers/MERQURY.FK
MitoHiFi	2	https://github.com/marcelauliano/MitoHiFi
PretextView	0.2	https://github.com/wtsi-hpag/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
sanger-tol/genomenote	v1.0	https://github.com/sanger-tol/genomenote
sanger-tol/readmapping	1.1.0	https://github.com/sanger-tol/readmapping/tree/1.1.0
YaHS	yahs-1.1.91eebc2	https://github.com/c-zhou/yahs

myopaeformis assembly (GCA_944738685.1) in Ensembl Rapid Release at the EBI.

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 here. 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: *Synanthedon myopaeformis* (red-belted clearwing). Accession number PRJEB53610; https://identifiers.org/ena.embl/PRJEB53610 (Wellcome Sanger Institute, 2022). The genome sequence is released openly for reuse. The *Synanthedon myopaeformis* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project. All raw sequence data and the assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in Table 1.

Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.7125292.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.4893703.

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: https://doi.org/10.5281/zenodo.10066175.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: https://doi.org/10.5281/zenodo.10043364.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: https://doi.org/10.5281/zenodo.10066637.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.5013541.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783558.

References

Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Ateyyat MA, Al-Antary TM: **Management and within-tree spatial distribution of the small red-belted clearwing borer, *Synanthedon myopaeformis* (Borkhausen) (Lepidoptera: Sesiidae), infesting dwarfing apple orchards in southern Jordan.** *J Entomol Soc B Col.* 2006; **103**: 11–17.
Reference Source

Beaton D, Carter K: **Apple clearwing moth (*Synanthedon myopaeformis*) – a new pest in Ontario.** *Ontario Ministry of Agriculture, Food and Rural Affairs Newsletter. Hort Matters.* 2006; **6**: 1.

Bernt M, Donath A, Jühling F, *et al.*: **MITOS: improved *de novo* metazoan mitochondrial genome annotation.** *Mol Phylogenet Evol.* 2013; **69**(2): 313–319.
PubMed Abstract | Publisher Full Text

Brůna T, Hoff KJ, Lomsadze A, *et al.*: **BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database.** *NAR Genom Bioinform.* 2021; **3**(1): lqaa108.
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

Clifton J: **Separation of Red-belted and Large Red-belted Clearwing**. [no date], [Accessed 13 February 2024].
Reference Source

Denton A, Oatley G, Cornwell C, *et al.*: **Sanger Tree of Life Sample Homogenisation: PowerMash.** *protocols.io.* 2023a.
Publisher Full Text

Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life Wet Laboratory Protocol Collection V.1.** *protocols.io.* 2023b.
Publisher 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

do Amaral RJV, Bates A, Denton A, *et al.*: **Sanger Tree of Life RNA Extraction: Automated MagMax™ mirVana.** *Protocols.Io.* 2023.
Publisher Full Text

GBIF Secretariat: ***Synanthedon myopaeformis* (Borkhausen, 1789).** *GBIF Backbone Taxonomy.* 2023; [Accessed 13 February 2024].
Reference Source

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898.
PubMed Abstract | Publisher Full Text | Free Full Text

Harry E: **PretextView (Paired REad TEXTure Viewer): A desktop application for viewing pretext contact maps**. 2022; [Accessed 19 October 2022].
Reference Source

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *Gigascience.* Oxford University Press, 2021; **10**(1): giaa153.
PubMed Abstract | Publisher Full Text | Free Full Text

Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life Sample Preparation: Triage and Dissection.** *protocols.io.* 2023.
Publisher Full Text

Judd G, Philip H: **Apple Clearwing Moth, *Synanthedon myopaeformis* (Borkhausen): Recent Introduction to Pacific Northwest Pest Complex**. In: *Proceedings of 80th Annual Western Orchard Pest and Disease Management Conference*. 2006; 123.

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

LaGasa E: **Apple Clearwing Moth Survey in Washington State and Northwestern Washington Detections, 2008**. In: *Proceedings of 68th Pacific Northwest Insect Management Conference*. 2009.
Reference Source

Larsson MC: **Pheromones and Other Semiochemicals for Monitoring Rare and Endangered Species.** *J Chem Ecol.* 2016; **42**(9): 853–868.
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

Németh-Major B, Ábrahám RR, Redeczki R: **The importance of Red-belted clearwing (*Synanthedon myopaeformis* Borkhausen) in organic apple orchards.** *Növényvédelem.* 2017; **53**: 59–66.
Reference Source

Özpinar A, ?ahin AK, Polat B: **Population development of apple clearwing moth (*Synanthedon myopaeformis* Borkh., Lepidoptera: Sesiidae) in Çanakkale province.** *TABAD: Tarim Bilimleri Arastirma Dergisi.* 2009; **2**: 115–118.
Reference Source

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

Sheerin E, Sampaio F, Oatley G, *et al.*: **Sanger Tree of Life HMW DNA Extraction: Automated MagAttract v.1.** *protocols.io.* 2023.
Publisher Full Text

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–3212.
PubMed Abstract | Publisher Full Text

South R: **Moths of the British Isles**. New edition. London: Frederick Warne and Co, 1961.
Reference Source

Strickland M, Cornwell C, Howard C: ***Sanger Tree of Life Fragmented DNA clean up: Manual SPRI.*** *protocols.io.* 2023.
Publisher Full Text

Surana P, Muffato M, Qi G: **sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo.* 2023a.
Publisher Full Text

Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev).** *Zenodo.* 2023b.
Publisher Full Text

Todorovic M, Sampaio F, Howard C: **Sanger Tree of Life HMW DNA Fragmentation: Diagenode Megaruptor®3 for PacBio HiFi.** *protocols.io.* 2023.
Publisher 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

Wellcome Sanger Institute: **The genome sequence of the red-belted clearwing, *Synanthedon myopaeformis* (Borkhausen, 1789).** European Nucleotide Archive, [dataset], accession number PRJEB53610, 2022.

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 15 July 2024

https://doi.org/10.21956/wellcomeopenres.23351.r85998

✓ **Kyle M Benowitz**
Austin Peay State University, Clarksville, Tennessee, USA

The manuscript "The genome sequence of the Red-belted Clearwing, *Synanthedon myopaeformis* (Borkhausen, 1789)" reports a high quality genome assembly for a European pest moth species. The paper is well written with a clear justification for the utility of a genome sequence. The methods are appropriate and easy to follow, and the results for the assembly and annotation are presented clearly. I just have a few minor comments for the authors to address prior to final publication.

Minor comments:

Abstract: The abstract lists 15,959 coding genes whereas the results and table 1 list 15,929. Please correct.

Results: I'm a little surprised not to see a BUSCO analysis of the annotation. This is usually performed as a quality assessment of the annotation. It will be very easy for the authors to do, and is especially important given that the annotation doesn't appear to utilize any RNA-seq data from the species.

Methods: At the end of the first paragraph of the methods, the sentence should be changed to "... identified by Will Langdon (University of Oxford) and preserved on dry ice."

Methods: I agree with the other reviewer that the author should specify which software (MitoFinder or MITOS) was used with MitoHiFi for the mitochondrial genome assembly.

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, evolution, behavior

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 May 2024

https://doi.org/10.21956/wellcomeopenres.23351.r79103

?

Josephine Paris
Department of Life and Environmental Science (Di.S.V.A.), Marche Polytechnic University, Ancona, Italy

This report provides details on the assembly and annotation of the Red-belted Clearwing, a moth species native to Europe, and recently introduced to North America. The species is considered a pest in some countries, as the larvae can cause damage to apple (and other Rosaceaeous) trees. The genome has been assembled and annotated as part of the Darwin Tree of Life Initiative. The genome (295.7 Mb, scaffold N50 =10.8 Mb) was assembled using the latest sequencing technologies (PacBio Hifi data, scaffolded with Hi-C chromosome conformation data). The assembly is of good quality and a high-quality annotation has been provided by Ensembl. The datasets and raw data are easily found via the links provided on the ENA, and the annotation is available via Ensembl's Rapid Release.

Minor comments:
1. The rationale for this genome is clearly explained, with ample detail on the species' distribution and its biology. I only have one comment:
"sporadic southern records" reads a bit odd. How about "sporadic records in southern Europe, including Italy, Greece, and Spain ...”

2. Mostly sufficient details are provided in the Methods, except I think the amount of Hi-C data generated (and its coverage) should be included as metrics to allow replication by others.

3. I would also be more quantitative about how much of the assembly sequence was assigned to

the scaffolds in the Abstract (99.97%), rather than just saying "most".

4. Finally, for the mitochondrial assembly using MitoHiFi, it's not clear which algorithm was used - MitoFinder or MITOS?

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?
Partly

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Population and conservation genomics. Genome assembly and annotation on non-model species.

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.