DATA NOTE

The genome sequence of the White-barred Knot-horn, *Elegia similella* (Zincken, 1818) [version 1; peer review: 2 approved, 1 approved with reservations]

James Hammond[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:** 12 Apr 2024, **9**:186
https://doi.org/10.12688/wellcomeopenres.21213.1
Latest published: 12 Apr 2024, **9**:186
https://doi.org/10.12688/wellcomeopenres.21213.1

Abstract

We present a genome assembly from an individual male *Elegia similella* (the White-barred Knot-horn; Arthropoda; Insecta; Lepidoptera; Pyralidae). The genome sequence is 780.4 megabases in span. Most of the assembly is scaffolded into 30 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled and is 15.3 kilobases in length. Gene annotation of this assembly on Ensembl identified 18,805 protein coding genes.

Keywords

Elegia similella, White-barred Knot-horn, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ? ✓ ✓

	1	2	3
version 1 12 Apr 2024	? view	✓ view	✓ view

1. **Nataliia Kopchak**, University of Manitoba (Ringgold ID: 124615), Winnipeg, Canada
   Jeffrey Marcus, University of Manitoba, Winnipeg, Canada
2. **Bryan Brunet**, Ottawa Research and Development Centre, Ontario, Canada
3. **Panagiotis Ioannidis**, Foundation for Research & Technology - Hellas, Crete, Greece

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: Hammond J: 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 (206194) and the Darwin Tree of Life Discretionary Award (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: Hammond J, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective *et al.* **The genome sequence of the White-barred Knot-horn, *Elegia similella* (Zincken, 1818) [version 1; peer review: 2 approved, 1 approved with reservations]** Wellcome Open Research 2024, **9**:186 https://doi.org/10.12688/wellcomeopenres.21213.1

First published: 12 Apr 2024, **9**:186 https://doi.org/10.12688/wellcomeopenres.21213.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Amphiesmenoptera; Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Pyraloidea; Pyralidae; Phycitinae; *Elegia; Elegia similella* (Zincken, 1818) (NCBI:txid1101167).

Background

The genome of the white-barred knot-horn, *Elegia similella*, was sequenced as part of the Darwin Tree of Life Project, a collaborative effort to sequence all named eukaryotic species in the Atlantic Archipelago of Britain and Ireland. Here we present a chromosomally complete genome sequence for *Elegia similella*, based on one male specimen from Wytham Woods, Oxfordshire, UK.

Genome sequence report

The genome was sequenced from one male *Elegia similella* (Figure 1) collected from Wytham Woods, Oxfordshire, UK (51.77, –1.34). A total of 32-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 4 missing joins or mis-joins and removed 4 haplotypic duplications, reducing the assembly length by 0.63% and the scaffold number by 2.86%.

The final assembly has a total length of 780.4 Mb in 33 sequence scaffolds with a scaffold N50 of 28.7 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 (99.99%) of the assembly sequence was assigned to 30 chromosomal-level scaffolds, representing 29 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.

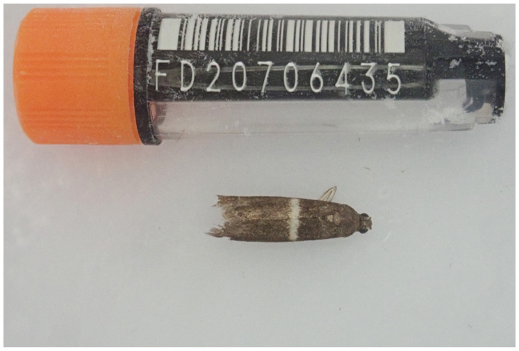

Figure 1. Photograph of the *Elegia similella* (ilEleSimi1) specimen used for genome sequencing.

The estimated Quality Value (QV) of the final assembly is 66.4 with *k*-mer completeness of 100.0%, and the assembly has a BUSCO v5.3.2 completeness of 98.8% (single = 98.3%, 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/1101167.

Genome annotation report

The *Elegia similella* genome assembly (GCA_947532085.1) was annotated using the Ensembl rapid annotation pipeline at the European Bioinformatics Institute (EBI). The resulting annotation includes 18,942 transcribed mRNAs from 18,805 protein-coding genes (Table 1; https://rapid.ensembl.org/Elegia_similella_GCA_947532085.1/Info/Index).

Methods

Sample acquisition and nucleic acid extraction

A male *Elegia similella* (specimen ID Ox001596, ToLID ilEleSimi1) was collected from Wytham Woods, Oxfordshire (biological vice-county Berkshire), UK (latitude 51.77, longitude –1.34) on 2021-06-30 using a light trap. The specimen was collected and identified by James Hammond (University of Oxford) and snap-frozen on dry ice.

Protocols developed by the Wellcome Sanger Institute (WSI) Tree of Life core laboratory have been deposited on protocols.io (Denton *et al.*, 2023b). The workflow for high molecular weight (HMW) DNA extraction at the WSI includes a sequence of core procedures: sample preparation; sample homogenisation, DNA extraction, fragmentation, and clean-up. In sample preparation, the ilEleSimi1 sample was weighed and dissected on dry ice, with tissue set aside for Hi-C sequencing (Jay *et al.*, 2023). Tissue from the whole organism was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a). HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023). HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 31 (Bates *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 eliminate shorter 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.

Table 1. Genome data for *Elegia similella*, ilEleSimi1.1.

Project accession data		
Assembly identifier	ilEleSimi1.1	
Species	*Elegia similella*	
Specimen	ilEleSimi1	
NCBI taxonomy ID	1101167	
BioProject	PRJEB56060	
BioSample ID	SAMEA10978763	
Isolate information	ilEleSimi1, male: whole organism (DNA and Hi-C sequencing)	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	66.4	*≥ 50*
k-mer completeness	100.0%	*≥ 95%*
BUSCO**	C:98.8%[S:98.3%,D:0.5%], F:0.4%,M:0.8%,n:5,286	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	99.99%	*≥ 95%*
Sex chromosomes	ZZ	*localised homologous pairs*
Organelles	Mitochondrial genome: 15.3 kb	*complete single alleles*
Raw data accessions		
PacificBiosciences SEQUEL II	ERR10224929	
Hi-C Illumina	ERR10297823	
Genome assembly		
Assembly accession	GCA_947532085.1	
Accession of alternate haplotype	GCA_947532095.1	
Span (Mb)	780.4	
Number of contigs	50	
Contig N50 length (Mb)	23.1	
Number of scaffolds	33	
Scaffold N50 length (Mb)	28.7	
Longest scaffold (Mb)	56.26	
Genome annotation		
Number of protein-coding genes	18,805	
Number of gene transcripts	18,942	

* 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/CANNWO01/dataset/CANNWO01/busco.

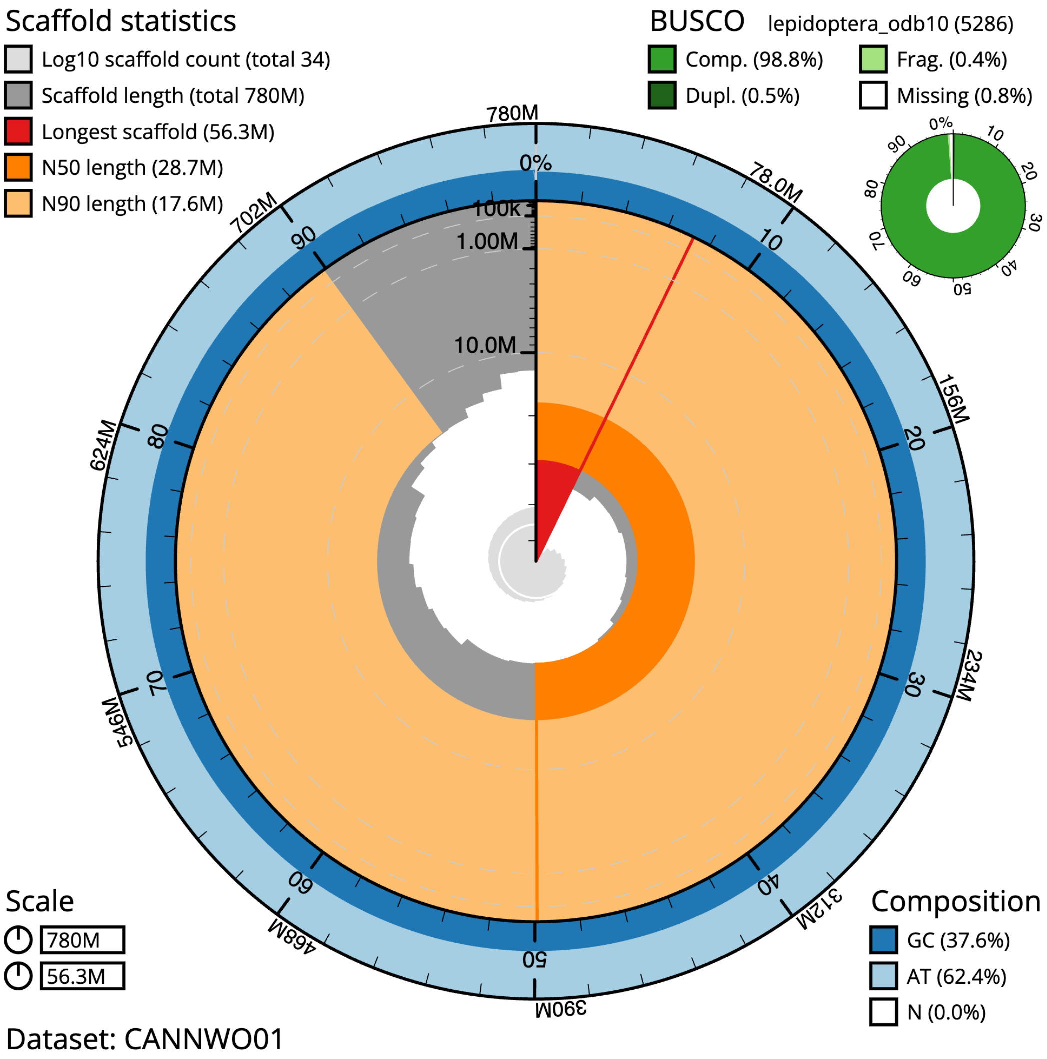

Figure 2. Genome assembly of *Elegia similella*, ilEleSimi1.1: metrics. The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 780,464,592 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (56,259,732 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (28,718,319 and 17,594,767 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/CANNWO01/dataset/CANNWO01/snail.

Sequencing

Pacific Biosciences HiFi circular consensus DNA sequencing libraries were constructed according to the manufacturers' instructions. DNA sequencing was performed by the Scientific Operations core at the WSI on a Pacific Biosciences SEQUEL II instruments. Hi-C data were also generated from remaining tissue of ilEleSimi1 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 (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)

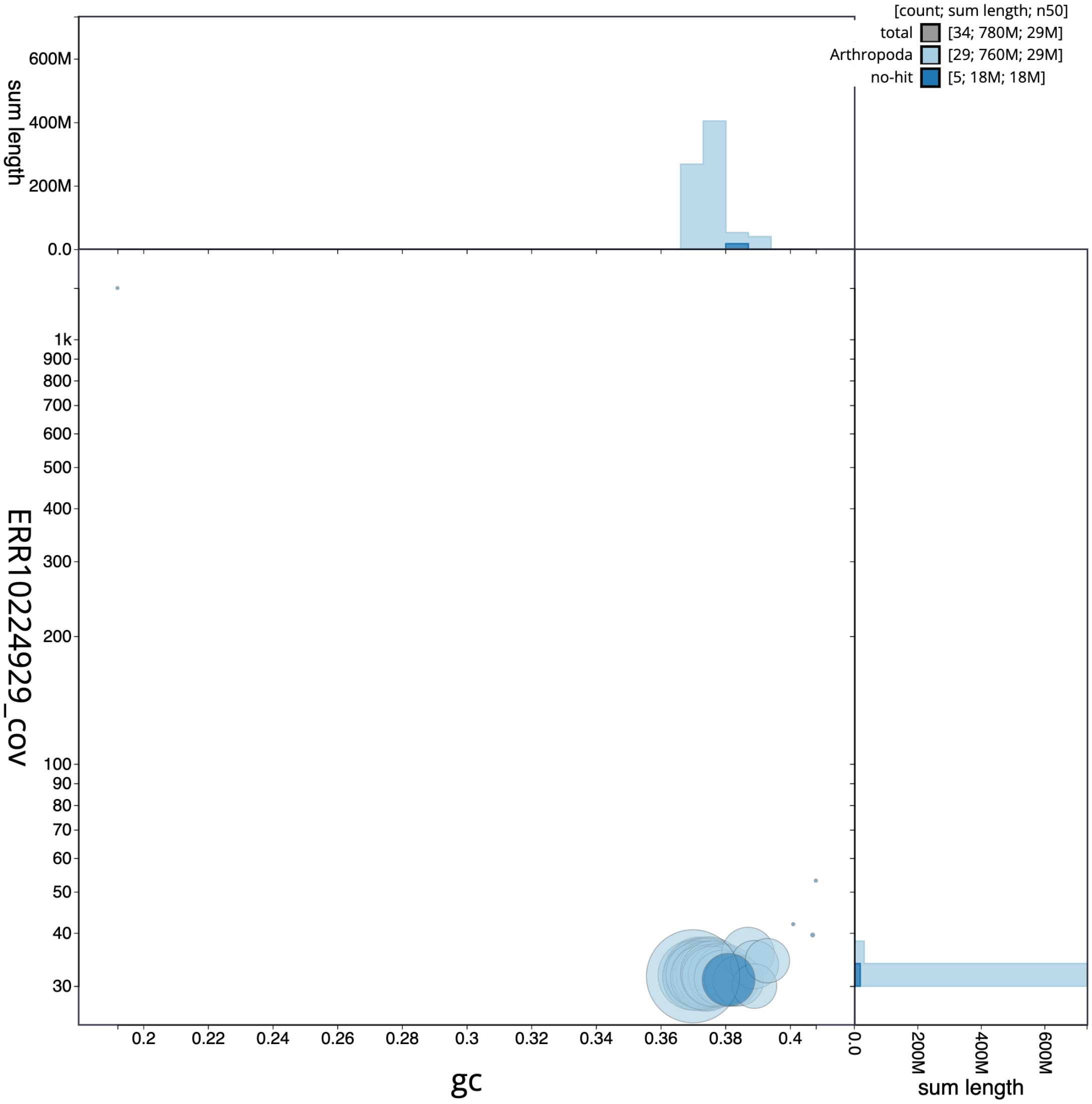

Figure 3. Genome assembly of *Elegia similella*, ilEleSimi1.1: BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/CANNWO01/dataset/CANNWO01/blob.

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.

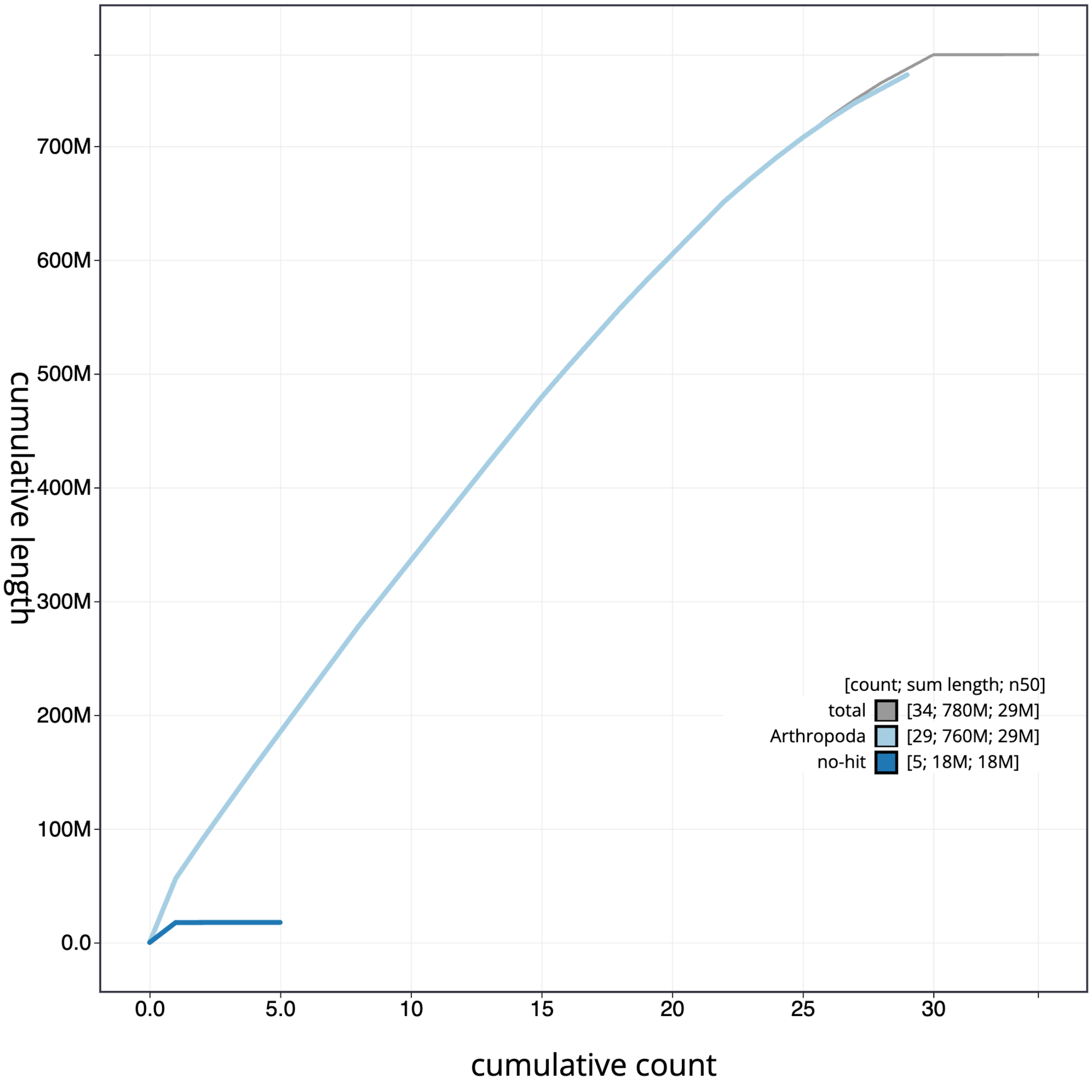

Figure 4. Genome assembly of *Elegia similella*, ilEleSimi1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/CANNWO01/dataset/CANNWO01/cumulative.

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

Genome annotation

The BRAKER2 pipeline (Brůna *et al.*, 2021) was used in the default protein mode to generate annotation for the *Elegia similella* assembly (GCA_947532085.1) 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

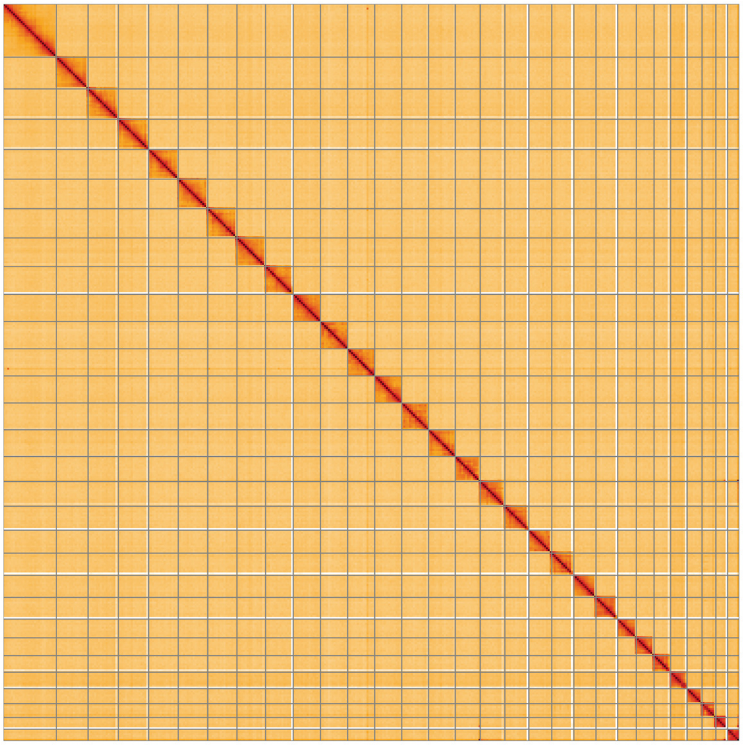

Figure 5. Genome assembly of *Elegia similella*, ilEleSimi1.1: Hi-C contact map of the ilEleSimi1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=APAUOQonQm-CcdQ6gbWXEQ.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Elegia similella*, ilEleSimi1.

INSDC accession	Chromosome	Length (Mb)	GC%
OX383926.1	1	33.62	37.0
OX383927.1	2	32.15	37.5
OX383928.1	3	32.05	37.5
OX383929.1	4	31.37	37.0
OX383930.1	5	31.33	37.0
OX383931.1	6	30.94	37.0
OX383932.1	7	30.41	37.5
OX383933.1	8	29.3	37.5
OX383934.1	9	28.87	37.5
OX383935.1	10	28.86	37.0
OX383936.1	11	28.72	37.0
OX383937.1	12	28.67	37.5
OX383938.1	13	28.36	37.5
OX383939.1	14	28.35	37.5
OX383940.1	15	26.52	37.5
OX383941.1	16	26.14	37.5
OX383942.1	17	25.36	37.5
OX383943.1	18	24.3	38.0
OX383944.1	19	23.48	38.0
OX383945.1	20	23.39	38.0
OX383946.1	21	23.0	37.5
OX383947.1	22	19.86	38.0
OX383948.1	23	18.77	38.0
OX383949.1	24	17.59	38.0
OX383950.1	25	17.43	38.5
OX383951.1	26	15.88	38.5
OX383952.1	27	14.64	39.0
OX383953.1	28	12.43	39.0
OX383954.1	29	12.32	39.5
OX383925.1	Z	56.26	37.0
OX383955.1	MT	0.02	19.5

Table 3. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	4.1.7	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

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: *Elegia similella* (white-barred knot-horn). Accession number PRJEB56060; https://identifiers.org/ena.embl/PRJEB56060 (Wellcome Sanger Institute, 2022). The genome sequence is released openly for reuse. The *Elegia similella* 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**

Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA**

Fragmentation: Diagenode Megaruptor®3 for LI PacBio. *protocols.io.* 2023.
Publisher Full Text

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**

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**

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

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**

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**

Oatley G, Denton A, Howard C: **Sanger Tree of Life HMW DNA Extraction: Automated MagAttract v.2.** *protocols.io.* 2023.
Publisher 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**

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**

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

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 White-barred Knot-horn, *Elegia similella* (Zincken, 1818).** European Nucleotide Archive, [dataset], accession number PRJEB56060, 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 02 November 2024

https://doi.org/10.21956/wellcomeopenres.23460.r104972

✓

Panagiotis Ioannidis
Foundation for Research & Technology - Hellas, Crete, Greece

This manuscript presents the sequencing, assembly and annotation of a lepidopteran insect. The quality of the genome assembly is very good (chromosome level).

I have to say that I'm happy with the fact that the genome assemblies published in this journal, now include a predicted gene set. However, it is absolutely necessary that the authors also measure the quality of this gene using BUSCO. It is very important for whoever wants to use this genome to know how good this gene set is, both in absolute terms (e.g. a gene set having 70% complete BUSCOs can't be very good), as well as in relative terms (i.e. how good the gene set is, compared to the corresponding genome assembly). I would suggest that the authors run BUSCO (using the same lineage they used for the evaluation of the genome assembly) and add the BUSCO scores in the manuscript (Table 1 seems the most fitting place).

An additional point has to do with the number of chromosomes. The authors say that they have assembled 30 chromosomes, but it would be nice to add whether this is expected. For example, is the number of chromosomes known for other closely related Lepidoptera?

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: Insect 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 01 November 2024

https://doi.org/10.21956/wellcomeopenres.23460.r104971

✓

Bryan Brunet
Ottawa Research and Development Centre, Ontario, Canada

The manuscript entitled “The genome sequence of the White-barred Knot-horn, Elegia similella (Zincken, 1818)” by Hammond *et al.* presents a chromosomal genome assembly for *Elegia similella* from a single male individual originating from the UK. The authors use a combination of Pacific Biosciences HiFi long reads and Hi-C data to assemble 30 chromosome-scale scaffolds for this species at a total genome size of 780.4 Mb. The data were also used to recover and assemble the mitogenome. The genome has very good assembly metrics, and all molecular and most bioinformatic methods employed have been sufficiently detailed either in the manuscript itself or with appropriate references to methods described in other publications. Aside from a few minor criticisms noted below, the genome should be a useful contribution to the field and is of scientific soundness and rigor.

- The authors present a photograph of the specimen in Figure 1. It is unknown whether this is sufficient to diagnose the accuracy of the identification, but at very least the authors should identify where morphological vouchers (if any) remain and/or have been deposited so that identification can be confirmed should it be necessary.

- Manual curation steps are described but the precise corrections that were made are not detailed in the manuscript. Such information should be made available in order to allow for repeatability of these steps.

- Annotation was completed using BRAKER2 in protein mode but no information is provided on what proteomic data and/or taxa were used to make these predictions. It would be useful to know the taxonomic extent for the set proteins used for annotation.

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: I am an aphid systematist with expertise in genomic approaches in the context of phylogenetics and population genetics. I've been involved in several genome assembly projects.

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 08 October 2024

https://doi.org/10.21956/wellcomeopenres.23460.r100551

?

Nataliia Kopchak
University of Manitoba (Ringgold ID: 124615), Winnipeg, Manitoba, Canada

Jeffrey Marcus
University of Manitoba, Winnipeg, Canada

In this manuscript, the author describes the sequencing and assembly of the *Elegia similella* genome using DNA from an adult male specimen collected in the UK. The primary genome sequence assembly includes proposed chromosomal pseudomolecule sequences for 29 autosomes, the Z sex chromosome, and a complete mitochondrial genome. On the whole, this is a useful contribution to the scientific literature, but please see our comments below, especially regarding the title, background, identification of the specimen, and details of mitogenome assembly.

Some suggestions to the author:

1. **Title:** The title indicates the species' genome sequence, which implies the complete genome, including that autosomes and both Z and W sex chromosomes are being analyzed. However, the paper only discusses the Z chromosome data and doesn't include an analysis of the W chromosome from females. To name the title more accurately with the content, one might write that the study focuses specifically on the male genome. We also note that in the metadata found at https://links.tol.sanger.ac.uk/spe- cies/1101167 , the sex of the specimen is listed as “not collected”, which appears to be incorrect.

2. **Background:** We recommend extending background information about the species

(morphology, ecology, distribution, larval host plants, etc.) to provide more context about *Elegia similella*. This will help readers unfamiliar with the species better understand this organism.

3. **Method of Specimen identification:** The specimen identification was named, but which keys/species descriptions were consulted and the morphological characters used for the identification have not been included in the manuscript.

4. **Figure #1:** It would be useful to add a scale bar on the photo, to better understand the size of the organism.

5. The author wrote: “Manual assembly curation corrected 4 missing joins or mis-joins and removed 4 haplotypic duplications, reducing the assembly length....” It would be beneficial to describe how 4 missing joins or mis-joins were identified and how they were corrected.

6. The authors describe how “The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva et al. 2022) which runs MitoFinder (Allio et al. 2020) or MITOS (Bernt et al. 2013) and uses these annotations to select the final mitochondrial contig...”. The authors do not describe which of the algorithms generated the selected contig for the mitogenome.

7. **Genome Annotation:** In addition to the total number of protein-coding genes, additional description of the preliminary details of the number/proportion of novel genes identified through this analysis should be included in the text of the manuscript.

8. **Reference sequence:** Any reference genomes or sequences used in this genome assembly (for example for the mitochondrial genome assembly) should be described in the main text of the manuscript and formally cited.

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

Partly

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: Evolutionary biology of insects, phylogenomics

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