DATA NOTE

The genome sequence of the Scalloped Oak, *Crocallis elinguaria* (Linneas, 1758) [version 1; peer review: 2 approved]

Douglas Boyes[1+], Clare Boyes[2],
University of Oxford and Wytham Woods Genome Acquisition Lab,
Darwin Tree of Life Barcoding collective,
Wellcome Sanger Institute Tree of Life programme,
Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]UK Centre for Ecology & Hydrology, Wallingford, England, UK
[2]Independent researcher, Welshpool, Wales, UK

[+] Deceased author

v1 **First published:** 06 Oct 2023, **8**:426
https://doi.org/10.12688/wellcomeopenres.20081.1
Latest published: 06 Oct 2023, **8**:426
https://doi.org/10.12688/wellcomeopenres.20081.1

Abstract

We present a genome assembly from an individual female *Crocallis elinguaria* (the Scalloped Oak; Arthropoda; Insecta; Lepidoptera; Geometridae). The genome sequence is 430.4 megabases in span. Most of the assembly is scaffolded into 17 chromosomal pseudomolecules, including the Z and W sex chromosomes. The mitochondrial genome has also been assembled and is 16.86 kilobases in length. Gene annotation of this assembly on Ensembl identified 17,741 protein coding genes.

Keywords

Crocallis elinguaria, Scalloped Oak, genome sequence, chromosomal, Lepidoptera

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 06 Oct 2023	✓ view	✓ view

1. **Laurence Despres**, Univ. Grenoble-Alpes, Grenoble, France

2. **Yuttapong Thawornwattana**, Harvard University, Cambridge, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Tree of Life gateway.

Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: Boyes D: Investigation, Resources; **Boyes C**: Writing – Original Draft Preparation;

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: Boyes D, Boyes C, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Scalloped Oak, *Crocallis elinguaria* (Linneas, 1758) [version 1; peer review: 2 approved]** Wellcome Open Research 2023, **8**:426 https://doi.org/10.12688/wellcomeopenres.20081.1

First published: 06 Oct 2023, **8**:426 https://doi.org/10.12688/wellcomeopenres.20081.1

Species taxonomy

Eukaryota; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Amphiesmenoptera; Lepidoptera; Glossata; Neolepidoptera; Heteroneura; Ditrysia; Obtectomera; Geometroidea; Geometridae; Ennominae; *Crocallis*; *Crocallis elinguaria* (Linneas, 1758) (NCBI:txid934829).

Background

Crocallis elinguaria (Scalloped Oak) is a macro-moth in the family Geometridae. The species is widespread and common throughout Britain and Europe, and occurs sporadically as far east as Siberia (GBIF Secretariat, 2023). The moth has undergone a large decline in abundance since the 1970s (Fox, 2013). Over the last 50 years, there has been a shift in its flight period, such that it now emerges earlier, with a peak occurrence in July rather than August (Randle *et al.*, 2019).

The moth is found in a varied range of habitats, from woodlands to urban areas (Waring *et al.*, 2017). At rest *C. elinguaria* holds its wings flat, appearing triangular in outline. The forewing edges are only slightly scalloped; the forewing length is 18–22 mm. The moth is very variable in colour, but the most common form is yellowish-buff with a darker brown central bar ornamented by a single black spot. There are forms which are orange-brown, dark brown or even blackish, with darker forms being more common in the north-west. *C. elinguaria* has one generation a year: it flies between July and August and regularly comes to light (Waring *et al.*, 2017). The female lays her eggs on tree bark where they overwinter. The eggs hatch in early spring and the larvae are omnivorous, feeding on a wide range of woody plants (Leraut, 2009).

A genome sequence from *C. elinguaria* will be useful for research into colour variation in moths, and more generally for comparative studies across the Lepidoptera. The genome of *C. elinguaria* 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 *C. elinguaria* based on the female specimen from Wytham Woods, Oxfordshire, UK.

Genome sequence report

The genome was sequenced from one female *Crocallis elinguaria* (Figure 1) collected from Wytham Woods, Oxfordshire, UK (51.77, –1.34). A total of 44-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 103-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 10 missing joins or mis-joins, reducing the scaffold number by 29.41%.

The final assembly has a total length of 430.4 Mb in 24 sequence scaffolds with a scaffold N50 of 28.5 Mb (Table 1).

Figure 1. Photograph of the *Crocallis elinguaria* (ilCroElin1) specimen used for genome sequencing.

A summary of the assembly statistics is shown in Figure 2, 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 17 chromosomal-level scaffolds, representing 15 autosomes and the Z and W sex chromosomes. 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 50.2 with *k*-mer completeness of 98.55%, and the assembly has a BUSCO v5.3.2 completeness of 98.4% (single = 97.9%, duplicated = 0.5%), using the lepidoptera_odb10 reference set (n = 5,286).

Metadata for specimens, spectral estimates, sequencing runs, contaminants and pre-curation assembly statistics can be found at https://links.tol.sanger.ac.uk/species/934829.

Genome annotation report

The *Crocallis elinguaria* genome assembly (GCA_907269065.1) was annotated using the Ensembl rapid annotation pipeline (Table 1; https://rapid.ensembl.org/Crocallis_elinguaria_GCA_907269065.1/Info/Index). The resulting annotation includes 17,905 transcribed mRNAs from 17,741 protein-coding genes.

Methods

Sample acquisition and nucleic acid extraction

A female *Crocallis elinguaria* (specimen ID Ox000666, individual ilCroElin1) was collected from Wytham Woods, Oxfordshire, UK (latitude 51.77, longitude –1.34) on 2020-07-20

Table 1. **Genome data for *Crocallis elinguaria*, ilCroElin1.1.**

Project accession data		
Assembly identifier	ilCroElin1.1	
Species	*Crocallis elinguaria*	
Specimen	ilCroElin1	
NCBI taxonomy ID	934829	
BioProject	PRJEB44984	
BioSample ID	SAMEA7701527	
Isolate information	ilCroElin1, female: abdomen (DNA sequencing); head and thorax (Hi-C scaffolding)	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	50.2	*≥ 50*
k-mer completeness	98.55%	*≥ 95%*
BUSCO**	C:98.4%[S:97.9%,D:0.5%], F:0.4%,M:1.2%,n:5,286	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	99.99%	*≥ 95%*
Sex chromosomes	Z and W chromosomes	*localised homologous pairs*
Organelles	Mitochondrial genome assembled	*complete single alleles*
Raw data accessions		
PacificBiosciences SEQUEL II	ERR6436373, ERR6807986	
10X Genomics Illumina	ERR6054752, ERR6054751, ERR6054753, ERR6054754	
Hi-C Illumina	ERR6054750	
Genome assembly		
Assembly accession	GCA_907269065.1	
Accession of alternate haplotype	GCA_907269135.1	
Span (Mb)	430.4	
Number of contigs	36	
Contig N50 length (Mb)	28.5	
Number of scaffolds	24	
Scaffold N50 length (Mb)	28.5	
Longest scaffold (Mb)	38.8	
Genome annotation		
Number of protein-coding genes	17,741	
Number of gene transcripts	17,905	

* 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 v5.3.2. 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/ilCroElin1.1/dataset/CAJSMC01.1/busco.

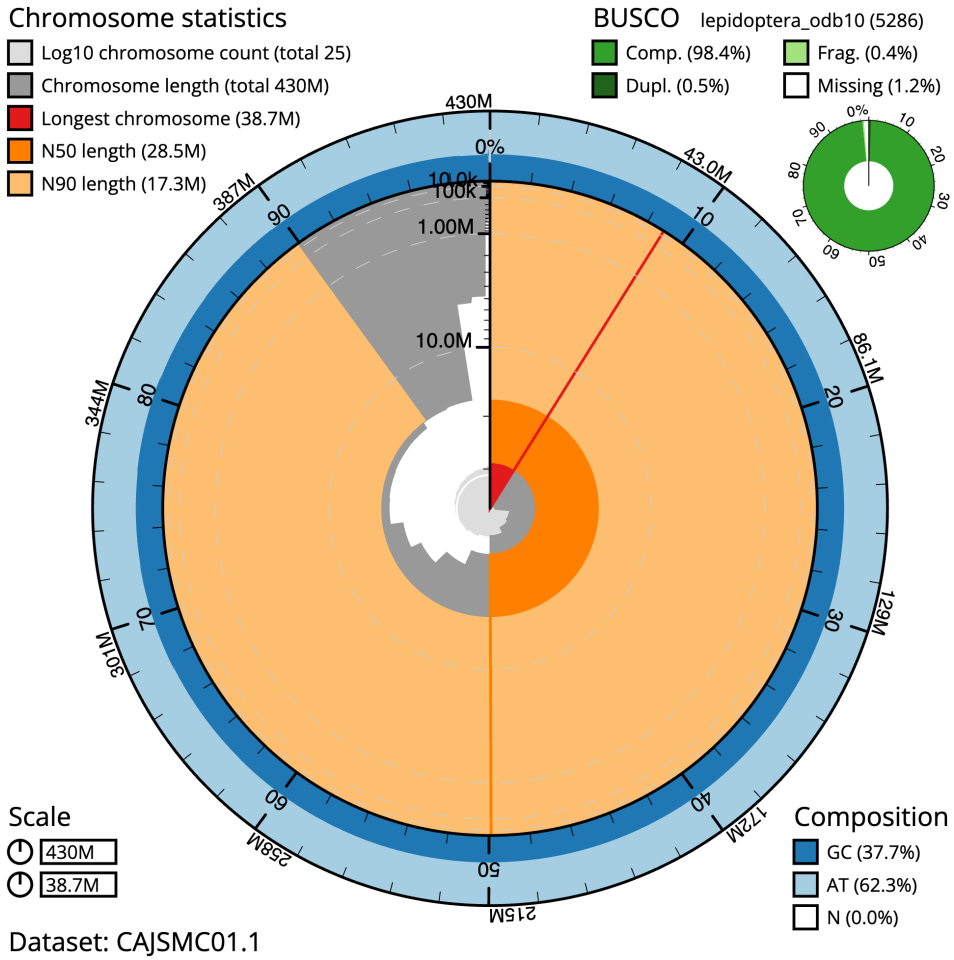

Figure 2. Genome assembly of *Crocallis elinguaria*, ilCroElin1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 430,399,360 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (38,734,590 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (28,485,152 and 17,265,692 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/ilCroElin1.1/dataset/CAJSMC01.1/snail.

using a light trap. The specimen was collected and identified by Douglas Boyes (University of Oxford) and preserved on dry ice.

DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute (WSI). The ilCroElin1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. Abdomen tissue was disrupted using a Nippi Powermasher fitted with a BioMasher pestle. High molecular weight (HMW) DNA was extracted using the Qiagen MagAttract HMW DNA extraction kit. Low molecular weight DNA was removed from a 20 ng aliquot of extracted DNA using the 0.8X AMpure XP purification kit prior to 10X Chromium sequencing; a minimum of 50 ng DNA was submitted for 10X sequencing. HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 30. Sheared DNA was purified by solid-phase reversible immobilisation using AMPure PB beads with a 1.8X ratio of beads to sample to remove the shorter fragments and concentrate the DNA sample. 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.

Sequencing

Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud DNA sequencing libraries were constructed according to the manufacturers' instructions. DNA a sequencing was performed by the Scientific Operations core at the WSI on

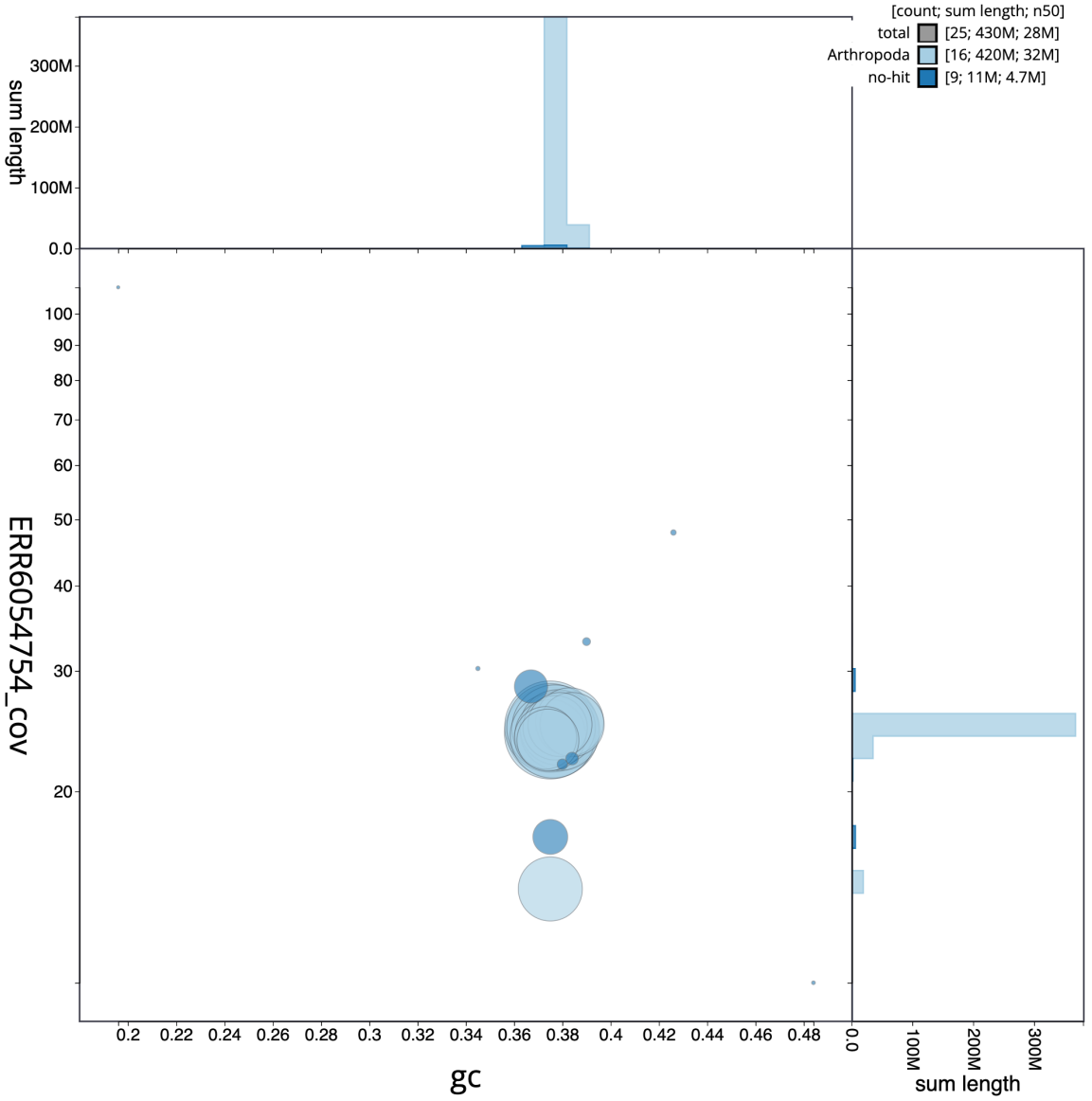

Figure 3. Genome assembly of *Crocallis elinguaria*, ilCroElin1.1: BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilCroElin1.1/dataset/CAJSMC01.1/blob.

Pacific Biosciences SEQUEL II (HiFi) and Illumina NovaSeq 6000 (10X) instruments. Hi-C data were also generated from head and thorax tissue of ilCroElin1 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). One round of polishing was performed by aligning 10X Genomics read data to the assembly with Long Ranger ALIGN, calling variants with FreeBayes (Garrison & Marth, 2012). The assembly was then scaffolded with Hi-C data (Rao *et al.*, 2014) using SALSA2 (Ghurye *et al.*, 2019). The assembly was checked for contamination and corrected using the gEVAL system (Chow *et al.*, 2016) as described previously (Howe *et al.*, 2021). Manual curation was performed using gEVAL, HiGlass (Kerpedjiev *et al.*, 2018) and Pretext (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

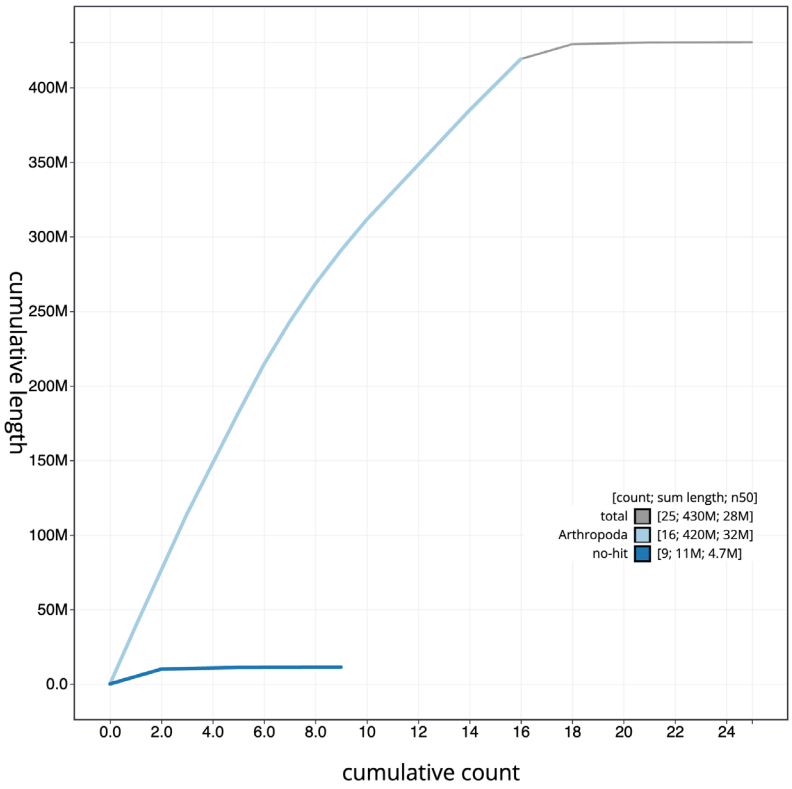

Figure 4. Genome assembly of *Crocallis elinguaria*, ilCroElin1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilCroElin1.1/dataset/CAJSMC01.1/cumulative.

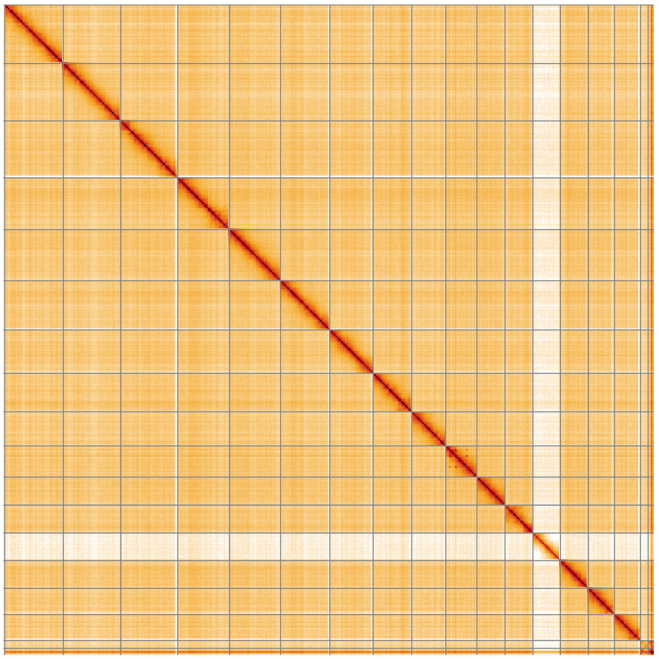

Figure 5. Genome assembly of *Crocallis elinguaria*, ilCroElin1.1: Hi-C contact map of the ilCroElin1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=fiJsohyJR0eJ6OJ1REeuiw.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Crocallis elinguaria*, ilCroElin1.

INSDC accession	Chromosome name	Length (Mb)	GC%
OU026065.1	1	38.73	37.5
OU026066.1	2	37.9	37
OU026067.1	3	37.51	37
OU026068.1	4	34.11	37.5
OU026069.1	5	33.68	37.5
OU026070.1	6	32.48	37.5
OU026071.1	7	28.49	37.5
OU026072.1	8	25.44	37.5
OU026073.1	9	22.49	37.5
OU026074.1	10	20.58	38
OU026075.1	11	18.46	37.5
OU026076.1	12	18.35	38
OU026078.1	13	18.26	37.5
OU026079.1	14	17.27	37
OU026080.1	15	17.06	37
OU026081.1	W	5.23	37
OU026077.1	Z	18.28	37.5
OU026082.1	MT	0.02	19.5

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 (Brůna *et al.*, 2021) was used in the default protein mode to generate annotation for the *Crocallis elinguaria* assembly (GCA_907269065.1) in Ensembl Rapid Release.

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.

Table 3. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	4.0.7	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.3.2	https://gitlab.com/ezlab/busco
FreeBayes	1.3.1-17-gaa2ace8	https://github.com/freebayes/freebayes
gEVAL	N/A	https://geval.org.uk/
Hifiasm	0.12	https://github.com/chhylp123/hifiasm
HiGlass	1.11.6	https://github.com/higlass/higlass
Long Ranger ALIGN	2.2.2	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines
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
SALSA	2.2	https://github.com/salsa-rs/salsa
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

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: *Crocallis elinguaria* (scalloped oak). Accession number PRJEB44984; https://identifiers.org/ena.embl/PRJEB44984. (Wellcome Sanger Institute, 2021)

The genome sequence is released openly for reuse. The *Crocallis elinguaria* 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.4789928.

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 programme are listed here: https://doi.org/10.5281/zenodo.4783585.

Members of Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.4790455.

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

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

Chow W, Brugger K, Caccamo M, *et al.*: **gEVAL - a web-based browser for evaluating genome assemblies.** *Bioinformatics.* 2016; **32**(16): 2508–10.
PubMed Abstract | Publisher Full Text | Free Full Text

Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319.
PubMed Abstract | Publisher Full Text

Fox R: **The decline of moths in Great Britain: a review of possible causes.** *Insect Conservation and Diversity.* 2013; **6**(1): 5–19.
Publisher Full Text

Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing.** 2012; [Accessed 26 July 2023].
Publisher Full Text

GBIF Secretariat: ***Crocallis elinguaria* (Linnaeus, 1758).** *GBIF.org.* 2023; [Accessed 14 September 2023].
Reference Source

Ghurye J, Rhie A, Walenz BP, *et al.*: **Integrating Hi-C links with assembly graphs for chromosome-scale assembly.** *PLoS Comput Biol.* 2019; **15**(8): e1007273.
PubMed Abstract | Publisher Full Text | Free 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

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

Leraut P: **Moths of Europe - Geometrid moths.** Verrières-le-Buisson: N.A.P. Editions, 2009.

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

Randle Z, Evans-Hill LJ, Parsons MS, *et al.*: **Atlas of Britain & Ireland's Larger Moths.** Newbury: NatureBureau, 2019.

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**

Surana P, Muffato M, Qi G: **sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo.* 2023a; [Accessed 21 July 2023].
Publisher Full Text

Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev).** *Zenodo.* 2023b; [Accessed 21 July 2023].
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

Waring P, Townsend M, Lewington R: **Field Guide to the Moths of Great Britain and Ireland: Third Edition.** Bloomsbury Wildlife Guides, 2017.
Reference Source

Wellcome Sanger Institute: **The genome sequence of the Scalloped Oak, *Crocallis elinguaria* (Linneas, 1758).** European Nucleotide Archive. [dataset], accession number PRJEB44984, 2021.

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 19 April 2024

https://doi.org/10.21956/wellcomeopenres.22239.r78876

✔ **Yuttapong Thawornwattana**
Harvard University, Cambridge, Massachusetts, USA

Boyes & Boyes present a chromosomally complete assembly of *Crocallis elinguaria* (Linneas, 1758) (Lepidoptera: Geometridae), a moth collected from Wytham Woods, England. It was generated primarily PacBio HiFi, 10X and Hi-C data.

The assembly is of high quality and will be useful to the research community. It contains 15 autosomes, the Z and W chromosomes and the mitochondrial genome. It is impressive to have the W chromosome in the assembly, but it is unclear in the Hi-C contact map (fig 5) which scaffold might corresponds to the W. Having chromosome labels in the figure would be useful. It would be nice to mention the expected haploid chromosome number, for example from cytology studies, to confirm that the chromosomes recovered here represent the complete genome.

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: evolutionary genomics, phylogenetics, insect evolution

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 03 April 2024

https://doi.org/10.21956/wellcomeopenres.22239.r76939

Laurence Despres
Univ. Grenoble-Alpes, Grenoble, France

In this data note, the authors report on the chromosome-level genome assembly of the Scalloped Oak, *Crocallis elinguaria* (Lepidoptera; Geometridae). a common and widely spread yet declining Eurasian wood moth. The assembly appears to be of good quality using appropriate methods : Pac Bio HiFi long reads (44-fold coverage), 10X genomics (103 fold-coverage) and Hi-C data to confirm scaffolding, manual curation of the final assembly, and following the high standard of the Darwin Tree of Life Project for results presentation.
The final assembly has a total length of 430.4 Mb and includes 17 chromosomes: 15 autosomes and the sexual Z and W chromosomes, as expected given that the specimen sequenced is a female. The mitogenome was also assembled. The quality of the assembly was further assessed by the proportion of complete BUSCO genes recovered, which is very high (>98% of Lepidoptera database BUSCO genes). Annotation was performed using BREAKER3 and 17,741 protein-coding genes were predicted.
I have only one minor suggestion: Given the large range of colour variation described for this species you might highlight in the introduction that the specimen sequenced corresponds to the most common form by referring to the photo of the specimen presented in figure 1 at the end of this sentence: '... but the most common form is yellowish-buff with a darker brown central bar ornamented by a single black spot' (see figure 1).

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: population 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.