DATA NOTE

The genome sequence of the Engrailed moth, *Ectropis crepuscularia* (Denis & Schiffermüller), 1775

[version 1; peer review: awaiting peer review]

Douglas Boyes[1+], Liam M. Crowley[2],
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]UK Centre for Ecology & Hydrology, Wallingford, England, UK
[2]University of Oxford, Oxford, England, UK

[+] Deceased author

v1 **First published:** 08 Apr 2025, **10**:175
https://doi.org/10.12688/wellcomeopenres.23995.1
Latest published: 08 Apr 2025, **10**:175
https://doi.org/10.12688/wellcomeopenres.23995.1

Abstract

We present a genome assembly from a male specimen of *Ectropis crepuscularia* (Engrailed; Arthropoda; Insecta; Lepidoptera; Geometridae). The genome sequence has a total length of 878.53 megabases. Most of the assembly (99.28%) is scaffolded into 32 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled, with a length of 15.4 kilobases. Gene annotation of this assembly on Ensembl identified 14,903 protein-coding genes.

Keywords

Ectropis crepuscularia, Engrailed, genome sequence, chromosomal, Lepidoptera

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

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: Boyes D: Investigation, Resources; **Crowley LM**: 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, 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: Boyes D, Crowley LM, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of the Engrailed moth, *Ectropis crepuscularia* (Denis & Schiffermüller), 1775 [version 1; peer review: awaiting peer review]** Wellcome Open Research 2025, **10**:175 https://doi.org/10.12688/wellcomeopenres.23995.1

First published: 08 Apr 2025, **10**:175 https://doi.org/10.12688/wellcomeopenres.23995.1

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; Geometroidea; Geometridae; Ennominae; *Ectropis*; *Ectropis crepuscularia* (Denis & Schiffermüller), 1775 (NCBI:txid572747)

Background

The genome of the Engrailed moth, *Ectropis crepuscularia*, 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 *Ectropis crepuscularia*, based on a specimen from Wytham Woods, Oxfordshire, United Kingdom (Figure 1).

Genome sequence report

Sequencing data

The genome of a specimen of *Ectropis crepuscularia* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 25.40 Gb (gigabases) from 2.01 million reads. GenomeScope analysis of the PacBio HiFi data estimated the haploid genome size at 867.43 Mb, with a heterozygosity of 1.75% and repeat content of 42.21%. These values provide an initial assessment of genome complexity and the challenges anticipated during assembly. Based on this estimated genome size, the sequencing data provided approximately 28.0x coverage of the genome. Chromosome conformation Hi-C sequencing produced 124.16 Gb from 822.27 million reads. Table 1 summarises the specimen and sequencing information

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The assembly was improved by manual curation, which corrected 11 misjoins or missing joins and removed four haplotypic duplications. The final assembly has a total length of 878.53 Mb in 90 scaffolds, with 33 gaps, and a scaffold N50 of 29.7 Mb (Table 2).

Figure 1. Photograph of the *Ectropis crepuscularia* (ilEctCrep1) specimen used for genome sequencing.

The snail plot in Figure 2 provides a summary of the assembly statistics, indicating the distribution of scaffold lengths and other assembly metrics. Figure 3 shows the distribution of scaffolds by GC proportion and coverage. Figure 4 presents a cumulative assembly plot, with separate curves representing different scaffold subsets assigned to various phyla, illustrating the completeness of the assembly.

Most of the assembly sequence (99.28%) was assigned to 32 chromosomal-level scaffolds, representing 31 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3). During curation, chromosome Z was assigned by synteny to the genome of *Agriphila straminella* (GCA_950108535.1) (Boyes *et al.*, 2024).

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

Assembly quality metrics

The estimated Quality Value (QV) and *k*-mer completeness metrics, along with BUSCO completeness scores, were calculated for each haplotype and the combined assembly. The QV reflects the base-level accuracy of the assembly, while *k*-mer completeness indicates the proportion of expected *k*-mers identified in the assembly. BUSCO scores provide a measure of completeness based on benchmarking universal single-copy orthologues.

The combined primary and alternate assemblies achieve an estimated QV of 66.3. The *k*-mer recovery for the primary haplotype is 70.29%, and for the alternate haplotype 69.18%; the combined primary and alternate assemblies have a *k*-mer recovery of 98.70%. BUSCO v.5.5.0 analysis using the lepidoptera_odb10 reference set (n = 5,286) identified 98.3% of the expected gene set (single = 97.6%, duplicated = 0.7%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project (EBP) Report on Assembly Standards September 2024. The assembly achieves the EBP reference standard of **7.C.Q66**.

Genome annotation report

The *Ectropis crepuscularia* genome assembly (GCA_963693475.1) was annotated externally by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 32,654 transcribed mRNAs from 14,903 protein-coding and 7,193 non-coding genes. The average transcript length is 19,772.39. There are 1.48 coding transcripts per gene and 6.05 exons per transcript. For further information about the annotation, please refer to https://rapid.ensembl.org/Ectropis_crepuscularia_GCA_963693475.1/Info/Index.

Table 1. **Specimen and sequencing data for *Ectropis crepuscularia*.**

Project information			
Study title	Ectropis crepuscularia (the engrailed)		
Umbrella BioProject	PRJEB67599		
Species	*Ectropis crepuscularia*		
BioSpecimen	SAMEA7701526		
NCBI taxonomy ID	572747		
Specimen information			
Technology	**ToLID**	**BioSample accession**	**Organism part**
PacBio long read sequencing	ilEctCrep1	SAMEA7701710	whole organism
Hi-C sequencing	ilEctCrep2	SAMEA113425891	head and thorax
Sequencing information			
Platform	**Run accession**	**Read count**	**Base count (Gb)**
Hi-C Illumina NovaSeq 6000	ERR12143994	8.22e+08	124.16
PacBio Sequel IIe	ERR12205258	9.57e+02	0.01
PacBio Sequel IIe	ERR12205259	2.00e+06	25.39

Table 2. **Genome assembly data for *Ectropis crepuscularia*.**

Genome assembly		
Assembly name	ilEctCrep1.1	
Assembly accession	GCA_963693475.1	
Alternate haplotype accession	*GCA_963693465.1*	
Assembly level for primary assembly	chromosome	
Span (Mb)	878.53	
Number of contigs	123	
Number of scaffolds	90	
Longest scaffold (Mb)	42.98	
Assembly metric	**Measure**	***Benchmark***
Contig N50 length	29.17 Mb	*≥ 1 Mb*
Scaffold N50 length	29.7 Mb	*= chromosome N50*
Consensus quality (QV)	Primary: 65.9; alternate: 66.7; combined: 66.3	*≥ 40*
k-mer completeness	Primary: 70.29%; alternate: 69.18%; combined: 98.70%	*≥ 95%*
BUSCO*	C:98.3%[S:97.6%,D:0.7%], F:0.4%,M:1.3%,n:5,286	*S > 90%; D < 5%*
Percentage of assembly mapped to chromosomes	99.28%	*≥ 90%*
Sex chromosomes	Z	*localised homologous pairs*
Organelles	Mitochondrial genome: 15.4 kb	*complete single alleles*

* BUSCO scores based on the lepidoptera_odb10 BUSCO set using version 5.5.0. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison.

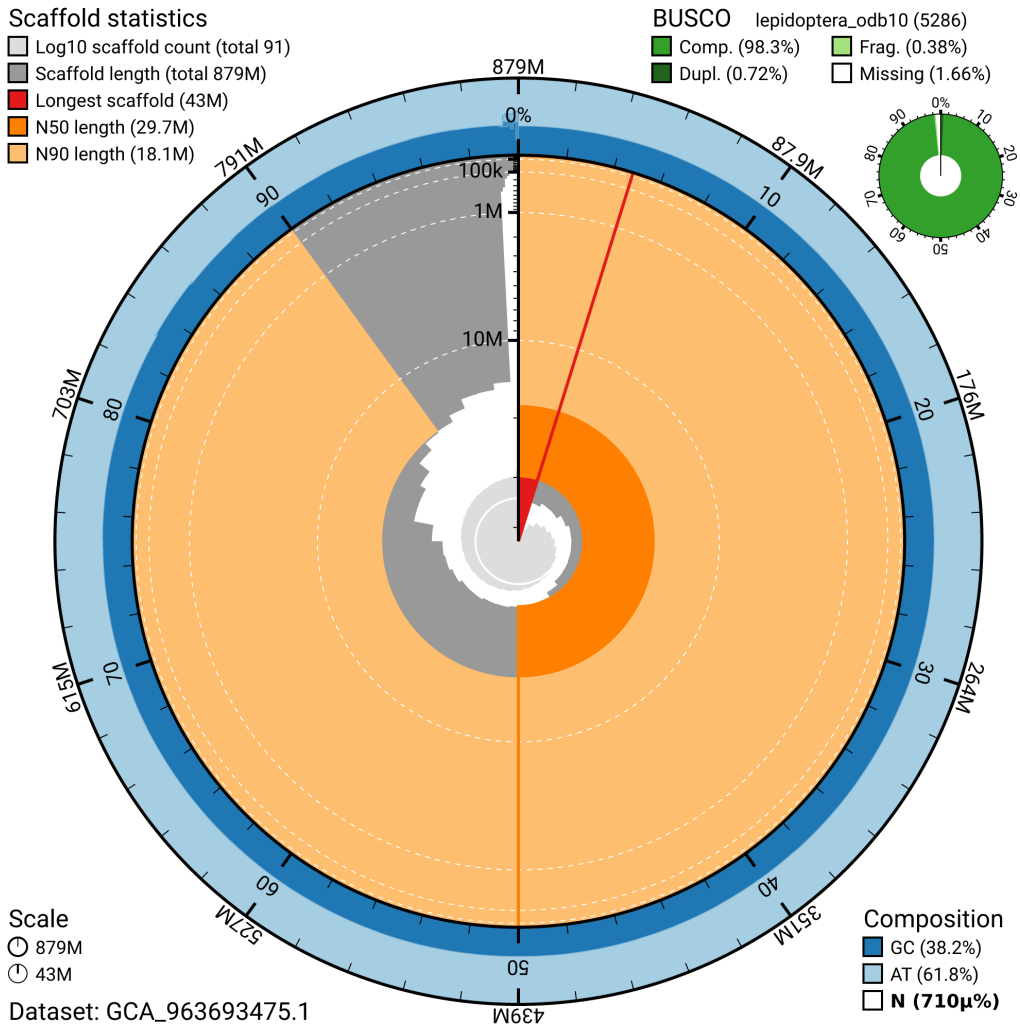

Figure 2. Genome assembly of *Ectropis crepuscularia*, ilEctCrep1.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the lepidoptera_odb10 set is presented at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963693475.1/dataset/GCA_963693475.1/snail.

Methods

Sample acquisition and DNA barcoding

An adult male *Ectropis crepuscularia* (specimen ID Ox000665, ToLID ilEctCrep1) was collected from Wytham Woods, Oxfordshire, United Kingdom (latitude 51.77, longitude -1.34) on 2020-07-20, using a light trap. The specimen was collected and identified by Douglas Boyes (University of Oxford) and preserved on dry ice. This specimen was used for PacBio HiFi sequencing. The specimen used for Hi-C sequencing (specimen ID Ox003115, ToLID ilEctCrep2) collected from the same location on 2022-08-04, using a light trap. The specimen was collected and identified by Liam Crowley (University of Oxford) and preserved by on dry ice.

The initial identification was verified by an additional DNA barcoding process according to the framework developed by Twyford *et al.* (2024). A small sample was dissected from the specimen and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI) (Pereira *et al.*, 2022). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species

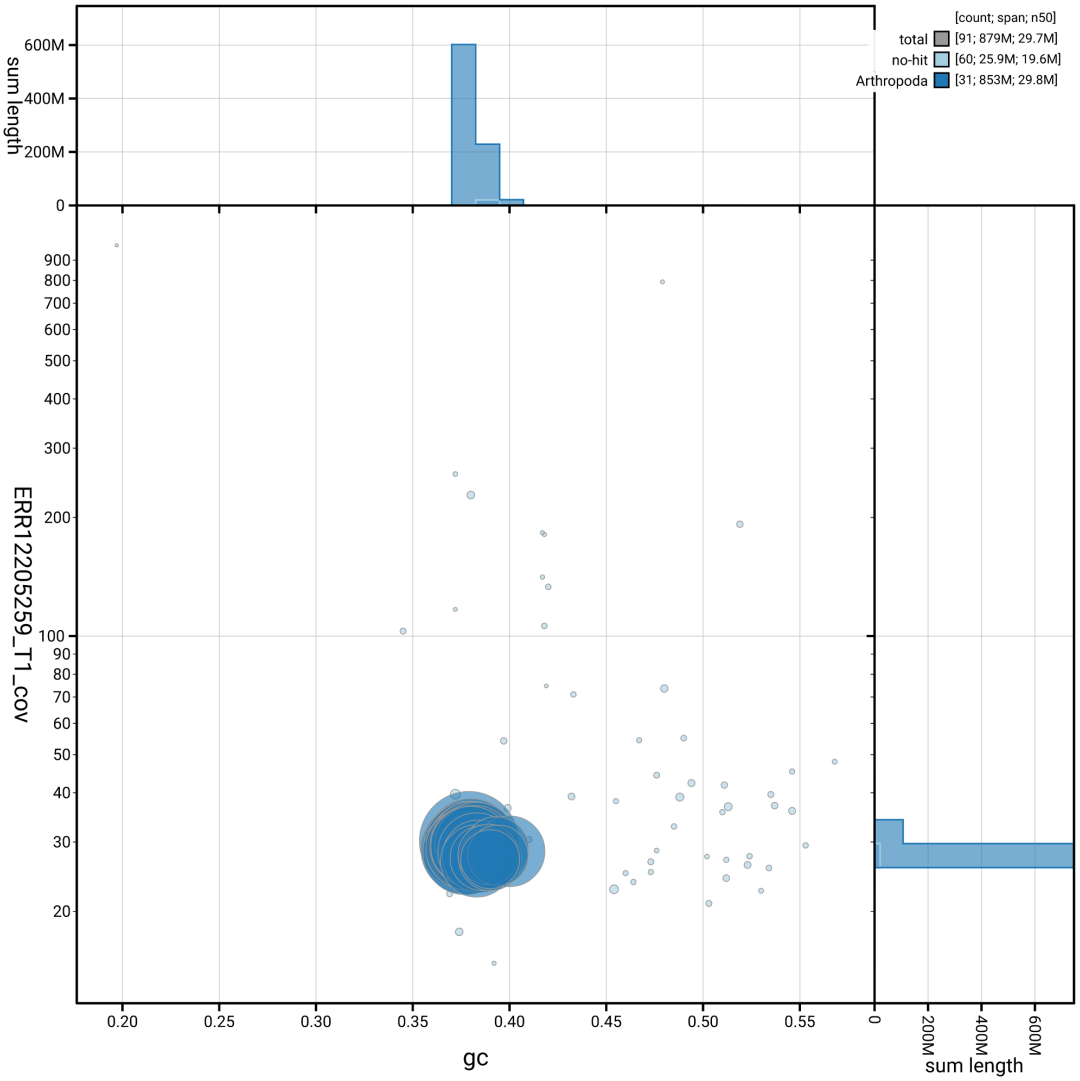

Figure 3. Genome assembly of *Ectropis crepuscularia*, ilEctCrep1.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963693475.1/dataset/GCA_963693475.1/blob.

identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding have been deposited on protocols.io (Beasley *et al.*, 2023).

Metadata collection for samples adhered to the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io (Denton *et al.*, 2023b). The ilEctCrep1 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (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 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, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA

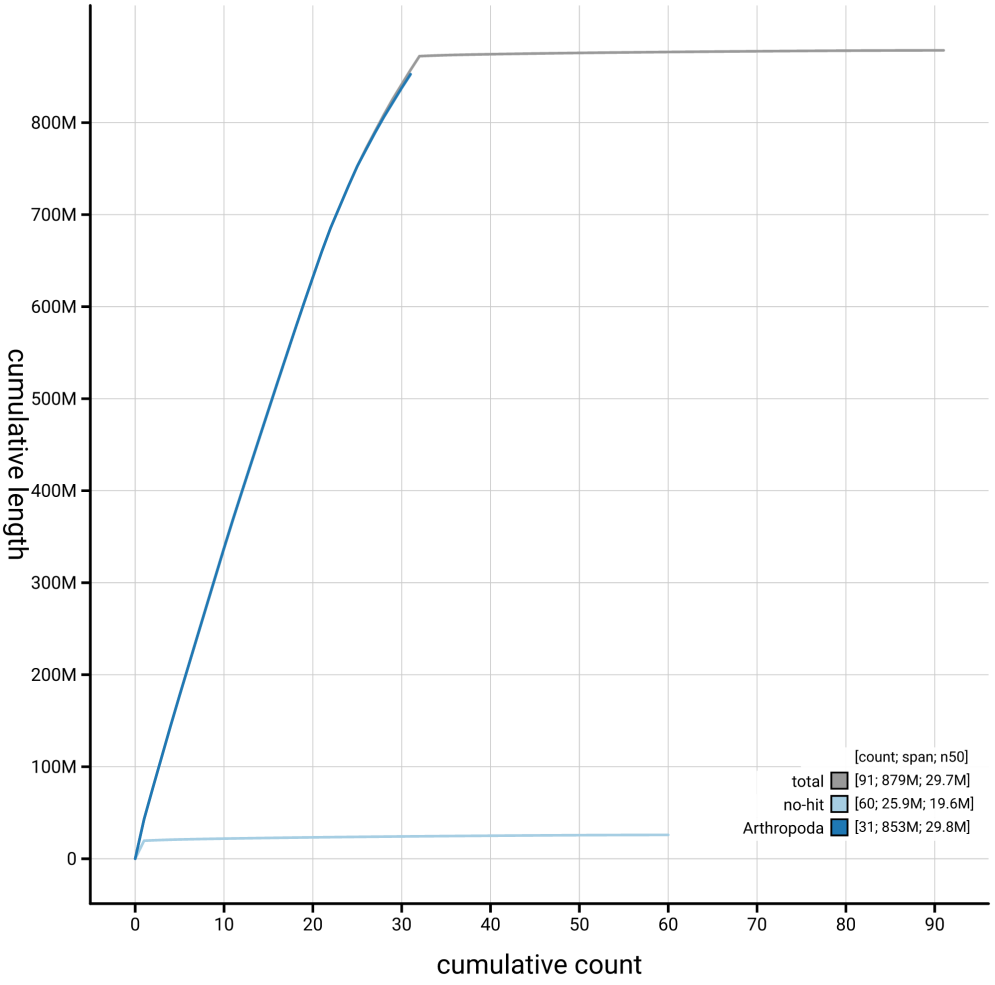

Figure 4. Genome assembly of *Ectropis crepuscularia,* ilEctCrep1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963693475.1/dataset/GCA_963693475.1/cumulative.

(Strickland *et al.*, 2023). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. The fragment size distribution was evaluated by running the sample on the FemtoPulse system.

Hi-C sample preparation and cross-linking

Tissue from the head and thorax of the ilEctCrep2 sample was processed for Hi-C sequencing at the WSI Scientific Operations core, using the Arima-HiC v2 kit. In brief, 20–50 mg of frozen tissue (stored at –80 °C) was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde concentration. After crosslinking, the tissue was homogenised using the Diagnocine Power Masher-II and BioMasher-II tubes and pestles. Following the Arima-HiC v2 kit manufacturer's instructions, crosslinked DNA was digested using a restriction enzyme master mix. The 5'-overhangs were filled in and labelled with biotinylated nucleotides and proximally ligated. An overnight incubation was carried out for enzymes to digest remaining proteins and for crosslinks to reverse. A clean up was performed with SPRIselect beads prior to library preparation. Additionally, the biotinylation percentage was estimated using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit and Arima-HiC v2 QC beads.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core.

PacBio HiFi

At a minimum, samples were required to have an average fragment size exceeding 8 kb and a total mass over 400 ng to proceed to the low input SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences, California, USA), depending on genome size and sequencing depth required. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences,

California, USA) as per the manufacturer's instructions. The kit includes the reagents required for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead cleanup, and nuclease treatment. Following the manufacturer's instructions, size selection and clean up was carried out using diluted AMPure PB beads (Pacific Biosciences, California, USA). DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) with Qubit 1X dsDNA HS assay kit and the final library fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and gDNA 55kb BAC analysis kit.

Samples were sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was in the range 40–135 pM. The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, as well as perform primary and secondary analysis of the data upon completion.

Hi-C

For Hi-C library preparation, DNA was fragmented using the Covaris E220 sonicator (Covaris) and size selected using SPRISelect beads to 400 to 600 bp. The DNA was then enriched using the Arima-HiC v2 kit Enrichment beads. Using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) for end repair, a-tailing, and adapter ligation. This uses a custom protocol which resembles the standard NEBNext Ultra II DNA Library Prep protocol but where library preparation occurs while DNA is bound to the Enrichment beads. For library amplification, 10 to 16 PCR cycles were required, determined by the sample biotinylation percentage. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on an Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

Assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts ($k = 31$) was generated from the filtered reads using FastK. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were first assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). The Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019), and the contigs were scaffolded using YaHS (Zhou *et al.*, 2023) using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. Flat files and maps used in curation were generated via the TreeVal pipeline (Pointon *et al.*, 2023). Manual curation was conducted primarily in PretextView (Harry, 2022) and HiGlass (Kerpedjiev *et al.*, 2018), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were amended, and duplicate sequences were tagged and removed. Sex chromosomes were identified by synteny analysis. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020), run in a Singularity container (Kurtzer *et al.*, 2017), was used to evaluate *k*-mer completeness and assembly quality for the primary and alternate haplotypes using the *k*-mer databases ($k = 31$) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

A Hi-C contact map was produced for the final version of the assembly. The Hi-C reads were aligned using bwa-mem2 (Vasimuddin *et al.*, 2019) and the alignment files were combined using SAMtools (Danecek *et al.*, 2021). The Hi-C alignments were converted into a contact map using BEDTools (Quinlan & Hall, 2010) and the Cooler tool suite (Abdennur & Mirny, 2020). The contact map was visualised in HiGlass (Kerpedjiev *et al.*, 2018).

The blobtoolkit pipeline is a Nextflow (Di Tommaso *et al.*, 2017) port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads in SAMtools and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoaT database (Challis *et al.*, 2023) to identify all matching BUSCO lineages to run BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineages, the pipeline aligns the BUSCO genes to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) with DIAMOND blastp (Buchfink *et al.*, 2021). The genome is also divided into chunks according to the density of the BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database using DIAMOND blastx. Genome sequences without a hit are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The blobtools suite combines all these outputs into a blobdir for visualisation.

The blobtoolkit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), relying on the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), as well as the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

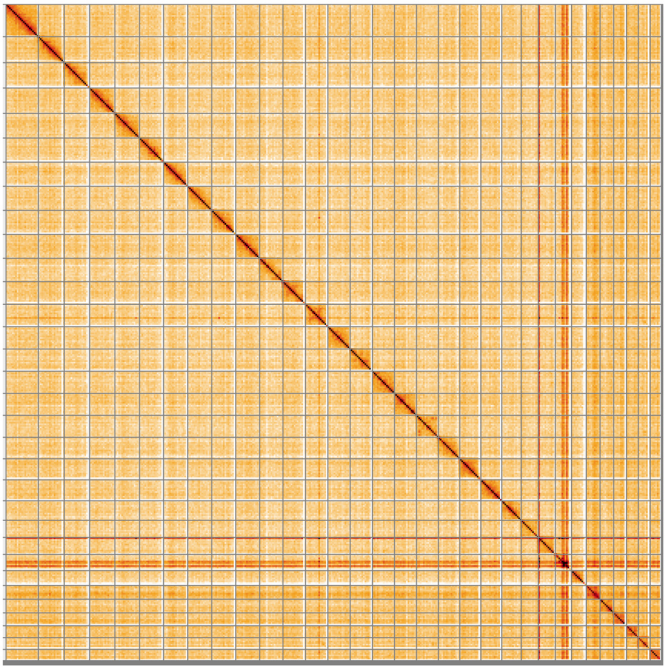

Figure 5. Genome assembly of *Ectropis crepuscularia:* Hi-C contact map of the ilEctCrep1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=KGu-kArIQWafKNkQE5nzSg.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Ectropis crepuscularia*, ilEctCrep1.

INSDC accession	Name	Length (Mb)	GC%
OY856339.1	1	34.48	38
OY856340.1	2	33.75	37.5
OY856341.1	3	33.69	38
OY856342.1	4	32.73	38
OY856343.1	5	32.11	38
OY856344.1	6	32.06	38
OY856345.1	7	32.0	38
OY856346.1	8	31.91	38
OY856347.1	9	31.86	38
OY856348.1	10	31.01	37.5
OY856349.1	11	30.04	38
OY856350.1	12	29.75	38
OY856351.1	13	29.7	38
OY856352.1	14	29.54	38
OY856353.1	15	29.49	38
OY856354.1	16	29.17	38
OY856355.1	17	29.11	38.5
OY856356.1	18	28.55	38.5
OY856357.1	19	28.04	38
OY856358.1	20	27.81	38
OY856359.1	21	25.95	38.5
OY856360.1	22	22.67	38.5
OY856361.1	23	22.66	38.5
OY856362.1	24	21.63	40
OY856363.1	25	19.63	38.5
OY856364.1	26	18.32	39.5
OY856365.1	27	18.1	38.5
OY856366.1	28	17.1	39
OY856367.1	29	15.78	39
OY856368.1	30	15.73	39.5
OY856369.1	31	14.91	39
OY856338.1	Z	42.98	38
OY856370.1	MT	0.02	20

Table 4 contains a list of relevant software tool versions and sources.

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 4. Software tools: versions and sources.

Software tool	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	666652151335353eef2fcd58880bcef5bc2928e1	https://github.com/thegenemyers/FASTK
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.5-r587	https://github.com/chhylp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84a a44357826c0b6753eb28de	https://github.com/higlass/higlass
MerquryFK	d00d98157618f4e8d1a9190026b19b471055b22e	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14, 1.17, and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	23.04.1	https://github.com/nextflow-io/nextflow
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	-	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.4.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

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: Ectropis crepuscularia (the engrailed). Accession number PRJEB67599; https://identifiers.org/ena.embl/PRJEB67599. The genome sequence is released openly for reuse. The *Ectropis crepuscularia* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project (PRJEB40665) and Project Psyche (PRJEB71705). All raw sequence data and the assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.12157525.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.12158331.

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: https://doi.org/10.5281/zenodo.12162482.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: https://doi.org/10.5281/zenodo.12165051.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: https://doi.org/10.5281/zenodo.12160324.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.12205391.

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

Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text

Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the universal protein knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531.
PubMed Abstract | Publisher Full Text | Free Full Text

Beasley J, Uhl R, Forrest LL, *et al.*: **DNA barcoding SOPs for the Darwin Tree of Life project.** *protocols.io.* 2023; [Accessed 25 June 2024].
Publisher Full Text

Boyes D, Young MR, University of Oxford and Wytham Woods Genome Acquisition Lab, *et al.*: **The genome sequence of the Straw Grass-veneer moth, *Agriphila straminella* (Denis & Schiffermüller), 1775 [version 1; peer review: 3 approved].** *Wellcome Open Res.* 2024; **9**: 433.
PubMed Abstract | Publisher Full Text | Free Full Text

Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374.
PubMed Abstract | Publisher Full Text | Free Full Text

Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175.
PubMed Abstract | Publisher Full Text | Free Full Text

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial arthropod fauna [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 123.
PubMed Abstract | Publisher Full Text | Free Full Text

da Veiga Leprevost F, Grüning BA, Alves Aflitos S, *et al.*: **BioContainers: an open-source and community-driven framework for software standardization.** *Bioinformatics.* 2017; **33**(16): 2580–2582.
PubMed Abstract | Publisher Full Text | Free Full Text

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Diesh C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of synteny and structural variation.** *Genome Biol.* 2023; **24**(1): 74.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
PubMed Abstract | Publisher Full Text

Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
PubMed Abstract | Publisher Full Text | Free Full Text

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
PubMed Abstract | Publisher Full Text | Free Full Text

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.
Reference Source

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

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

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
PubMed Abstract | Publisher Full Text | Free Full Text

Lawniczak MKN, Davey RP, Rajan J, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project [version 1; peer review: 2 approved with reservations].** *Wellcome Open Res.* 2022; **7**: 187.
Publisher Full Text

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text | Free Full Text

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | Publisher Full Text | Free Full Text

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2, [Accessed 2 April 2024].
Reference Source

Pereira L, Sivell O, Sivess L, *et al.*: **DToL Taxon-specific Standard Operating Procedure for the terrestrial and freshwater arthropods working group**. 2022.
Publisher Full Text

Pointon DL, Eagles W, Sims Y, *et al.*: **sanger-tol/treeval v1.0.0 – Ancient Atlantis.** 2023.
Publisher Full Text

Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–842.
PubMed Abstract | Publisher Full Text | Free Full Text

Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat Commun.* 2020; **11**(1): 1432.
PubMed Abstract | Publisher Full Text | Free Full Text

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *protocols.io.* 2023.
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

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2024; **9**: 339.
PubMed Abstract | Publisher Full Text | Free Full Text

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.
PubMed Abstract | Publisher Full Text | Free Full Text

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.
Publisher Full Text

Zhou C, McCarthy SA, Durbin R: **YaHS: yet another Hi-C scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
PubMed Abstract | Publisher Full Text | Free Full Text