



## DATA NOTE

# The genome sequence of the Treble Lines, *Charanyca trigrammica* (Hufnagel, 1766) (Lepidoptera: Noctuidae)

[version 1; peer review: awaiting peer review]

Douglas Boyes<sup>1+</sup>, Clare Boyes<sup>2</sup>,  
University of Oxford and Wytham Woods 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

<sup>1</sup>UK Centre for Ecology & Hydrology, Wallingford, England, UK

<sup>2</sup>Independent researcher, Welshpool, Wales, UK

+ Deceased author

**V1** First published: 18 Sep 2025, 10:520  
<https://doi.org/10.12688/wellcomeopenres.24867.1>  
Latest published: 18 Sep 2025, 10:520  
<https://doi.org/10.12688/wellcomeopenres.24867.1>

## 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.

## Abstract

We present a genome assembly from an individual male *Charanyca trigrammica* (Treble Lines; Arthropoda; Insecta; Lepidoptera; Noctuidae). The assembly contains two haplotypes with total lengths of 546.43 megabases and 546.58 megabases. Most of haplotype 1 (99.97%) is scaffolded into 31 chromosomal pseudomolecules, including the Z sex chromosome. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 15.44 kilobases. This assembly was generated as part of the Darwin Tree of Life project, which produces reference genomes for eukaryotic species found in Britain and Ireland.

## Keywords

*Charanyca trigrammica*; Treble Lines; genome sequence; chromosomal; Lepidoptera



This article is included in the [Tree of Life](#) gateway.

**Corresponding author:** Darwin Tree of Life Consortium ([mark.blaxter@sanger.ac.uk](mailto: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 (220540) 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.*

**Copyright:** © 2025 Boyes D *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Boyes D, Boyes C, University of Oxford and Wytham Woods Acquisition Lab *et al.* **The genome sequence of the Treble Lines, *Charanyca trigrammica* (Hufnagel, 1766) (Lepidoptera: Noctuidae) [version 1; peer review: awaiting peer review]** Wellcome Open Research 2025, 10:520 <https://doi.org/10.12688/wellcomeopenres.24867.1>

**First published:** 18 Sep 2025, 10:520 <https://doi.org/10.12688/wellcomeopenres.24867.1>

## 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; Noctuoidea; Noctuidae; Xyleninae; *Charanyca*; *Charanyca trigrammica* (Hufnagel, 1766) (NCBI:txid987901)

## Background

*Charanyca trigrammica* (Treble Lines) is a moth in the family Noctuidae. It is a common species in the southern half of Britain where it occurs in a variety of open habitats (Waring *et al.*, 2017). It has significantly increased its range and abundance in the UK since the 1970s (Randle *et al.*, 2019) and it occurs throughout Europe (GBIF Secretariat, 2025).

As its common name suggests, this medium sized moth (forewing length 15–17 mm) is distinguished by three straight dark-brown crosslines across the forewing. The ground colour is pale brown although there is some variation. There is one generation a year and it flies in May and June (Waring *et al.*, 2017). The larvae feed at night on a variety of herbaceous plants including Knapweeds and Greater Plantain. The larvae eat the plants at ground level and it pupates underground (Heath & Emmet, 1983).

We present a chromosome-level genome sequence for *Charanyca trigrammica*, the Treble Lines. The assembly was produced using the Tree of Life pipeline from a specimen collected in Wytham Woods, Oxfordshire, United Kingdom (Figure 1). This assembly was generated as part of the Darwin Tree of Life Project, which aims to generate high-quality reference genomes for all named eukaryotic species in Britain and Ireland to support research, conservation, and the sustainable use of biodiversity (Blaxter *et al.*, 2022).

## Methods

### Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Charanyca trigrammica* (specimen ID Ox001875, ToLID ilChaTrig1; Figure 1), collected from Wytham Woods,



**Figure 1.** Photograph of the *Charanyca trigrammica* (ilChaTrig1) specimen used for genome sequencing.

Oxfordshire, United Kingdom (latitude 51.774, longitude –1.324) on 2021-05-28. The specimen was collected and identified by Douglas Boyes. For the Darwin Tree of Life sampling and metadata approach, refer to Lawniczak *et al.* (2022).

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) (see the protocol). 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 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 are available on protocols.io.

### Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The ilChaTrig1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the thorax was homogenised by powermashing using a Power-Masher II tissue disruptor. HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol. DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor®3 for LI PacBio protocol. Sheared DNA was purified by manual SPRI (solid-phase reversible immobilisation). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 28.6 ng/μL and a yield of 1344.20 ng, with a fragment size of 14.1 kb. The 260/280 spectrophotometric ratio was 1.91, and the 260/230 ratio was 3.38.

### PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), following the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

The sample was sequenced on a Revio instrument (Pacific Biosciences). The prepared library was normalised to 2 nM,

and 15 µL was used for making complexes. Primers were annealed and polymerases bound to generate circularised complexes, following the manufacturer's instructions. Complexes were purified using 1.2X SMRTbell beads, then diluted to the Revio loading concentration (200–300 pM) and spiked with a Revio sequencing internal control. The sample was sequenced on a Revio 25M SMRT cell. The SMRT Link software (Pacific Biosciences), a web-based workflow manager, was used to configure and monitor the run and to carry out primary and secondary data analysis.

## Hi-C

### Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the head of the *ilChaTrig1* sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

### Hi-C library preparation and sequencing

Biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size selected to 400–600 bp using SPRISelect beads. DNA was enriched with Arima-HiC v2 kit Enrichment beads. End repair, A-tailing, and adapter ligation were carried out with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs), following a modified protocol where library preparation occurs while DNA remains bound to the Enrichment beads. Library amplification was performed using KAPA HiFi HotStart mix and a custom Unique Dual Index (UDI) barcode set (Integrated DNA Technologies). Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, libraries were amplified with 10 to 16 PCR cycles. Post-PCR clean-up was performed with SPRISelect beads. Libraries were quantified using the AccuClear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech).

Prior to sequencing, libraries were normalised to 10 ng/µL. Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq 6000.

## Genome assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts (*k* = 31) was generated from the filtered reads using FastK. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was

used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were assembled using Hifiasm in Hi-C phasing mode (Cheng *et al.*, 2021; Cheng *et al.*, 2022), producing two haplotypes. Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). Contigs were further scaffolded with Hi-C data in YaHS (Zhou *et al.*, 2023), using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQUERY.FK (Rhie *et al.*, 2020).

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

## Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included 5 breaks, 20 joins, and removal of 13 haplotypic duplications. The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation>. PretextView-Snapshot was used to generate a Hi-C contact map of the final assembly.

## Assembly quality assessment

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

The genome was analysed using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified from the NCBI Taxonomy (Schoch *et al.*, 2020). For the three domain-level lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all outputs into a blobdir for visualisation. The BlobToolKit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), with

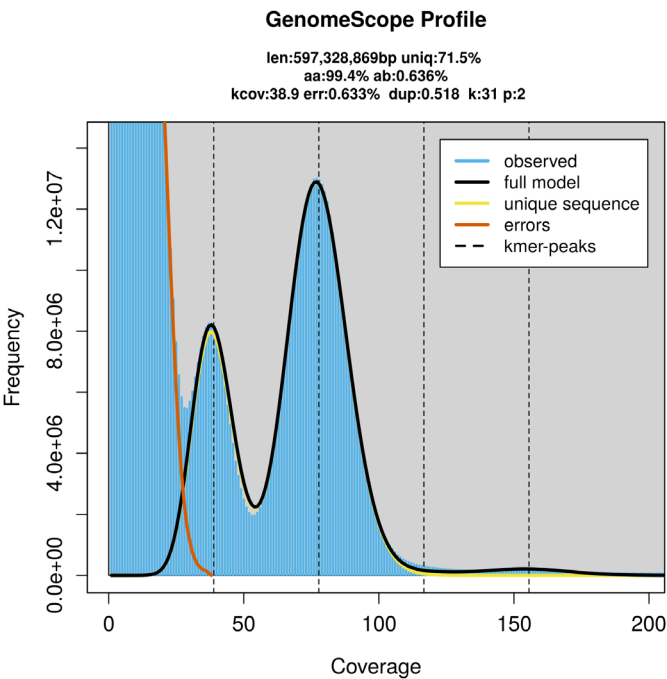
containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

PacBio sequencing of the *Charanyca trigrammica* specimen generated 56.74 Gb (gigabases) from 5.41 million reads, which were used to assemble the genome. GenomeScope2.0 analysis

estimated the haploid genome size at 597.33 Mb, with a heterozygosity of 0.64% and repeat content of 28.73% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 78× coverage. Hi-C sequencing produced 138.94 Gb from 920.12 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.



**Figure 2.** Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

**Table 1.** Specimen and sequencing data for BioProject PRJEB85024.

Platform	PacBio HiFi	Hi-C
ToLID	ilChaTrig1	ilChaTrig1
Specimen ID	Ox001875	Ox001875
BioSample (source individual)	SAMEA10979135	SAMEA10979135
BioSample (tissue)	SAMEA10979548	SAMEA10979547
Tissue	thorax	head
Instrument	Revio	Illumina NovaSeq 6000
Run accessions	ERR14209139; ERR14209138	ERR14224619
Read count total	5.41 million	920.12 million
Base count total	56.74 Gb	138.94 Gb



Assembly statistics

The genome was assembled into two haplotypes using Hi-C phasing. Haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The final assembly has a total length of 546.43 Mb in 38 scaffolds, with 133 gaps, and a scaffold N50 of 18.96 Mb (Table 2).

Most of the assembly sequence (99.97%) was assigned to 31 chromosomal-level scaffolds, representing 30 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). The Z chromosome was identified based on BUSCO gene painting with ancestral Merian elements (Wright *et al.*, 2024).

Table 2. Genome assembly statistics.

Assembly name	ilChaTrig1.hap1.1	ilChaTrig1.hap2.1
Assembly accession	GCA_965151675.1	GCA_965151685.1
Assembly level	chromosome	scaffold
Span (Mb)	546.43	546.58
Number of chromosomes	31	N/A
Number of contigs	171	176
Contig N50	6.0 Mb	6.25 Mb
Number of scaffolds	38	58
Scaffold N50	18.96 Mb	18.85 Mb
Longest scaffold length (Mb)	28.41	N/A
Sex chromosomes	Z	N/A
Organelles	Mitochondrion: 15.44 kb	N/A

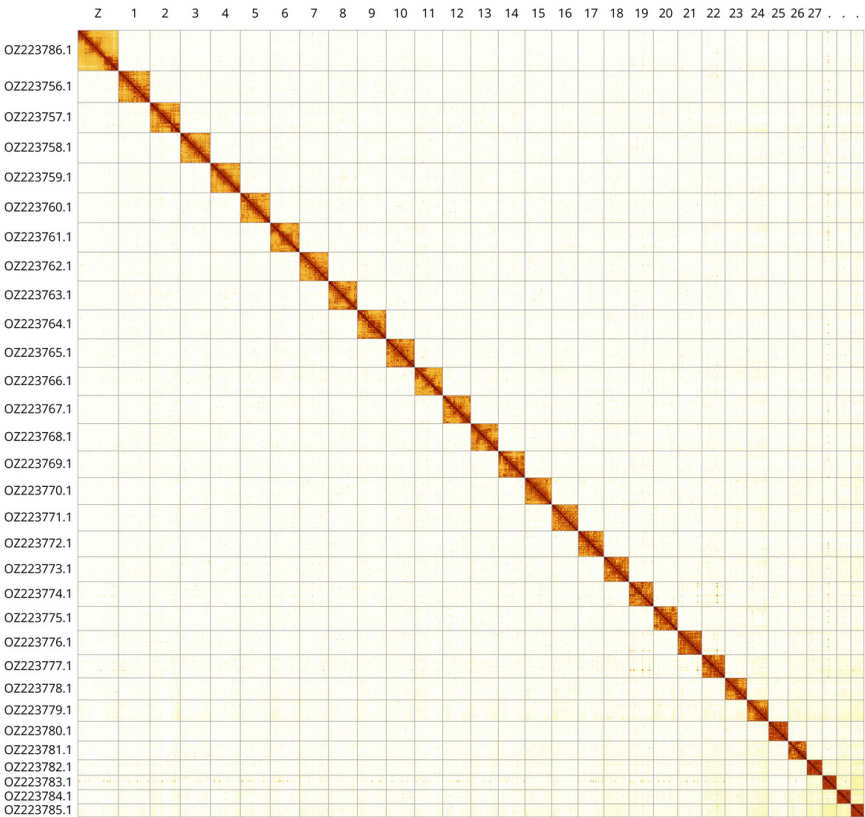


Figure 3. Hi-C contact map of the *Charanyca trigrammica* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes. The plot was generated using PretextSnapshot.

**Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Charanyca trigrammica* ilChaTrig1.**

INSDC accession	Molecule	Length (Mb)	GC%
OZ223756.1	1	21.93	37
OZ223757.1	2	20.98	37.50
OZ223758.1	3	20.90	37
OZ223759.1	4	20.82	37.50
OZ223760.1	5	20.73	37
OZ223761.1	6	20.42	37
OZ223762.1	7	20.09	37
OZ223763.1	8	20.06	37
OZ223764.1	9	20.03	37
OZ223765.1	10	19.75	37
OZ223766.1	11	19.56	37
OZ223767.1	12	19.50	37
OZ223768.1	13	18.96	37
OZ223769.1	14	18.61	37
OZ223770.1	15	18.57	37.50
OZ223771.1	16	18.40	37
OZ223772.1	17	17.89	37
OZ223773.1	18	17.44	37.50
OZ223774.1	19	16.99	37.50
OZ223775.1	20	16.86	37.50
OZ223776.1	21	16.80	37
OZ223777.1	22	16.10	37.50
OZ223778.1	23	15.18	37.50
OZ223779.1	24	14.94	38.50
OZ223780.1	25	13.71	37.50
OZ223781.1	26	12.90	37.50
OZ223782.1	27	10.87	38
OZ223783.1	28	9.89	39.50
OZ223784.1	29	9.85	38.50
OZ223785.1	30	9.15	38
OZ223786.1	Z	28.41	37

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.

For haplotype 1, the estimated QV is 67.0, and for haplotype 2, 67.6. When the two haplotypes are combined, the assembly achieves an estimated QV of 67.3. The *k*-mer completeness is 35.64% for haplotype 1, 35.63% for haplotype 2, and 41.22% for the combined haplotypes (Figure 4). The low completeness values are due to removal of contaminant sequences prior to assembly.

BUSCO analysis using the lepidoptera\_odb10 reference set ( $n = 5\,286$ ) identified 99.0% of the expected gene set (single = 98.7%, duplicated = 0.3%) for haplotype 1. The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

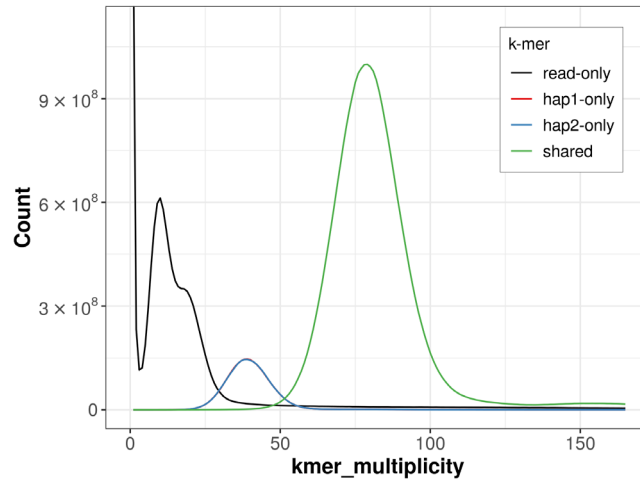
Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.C.Q67**, meeting the recommended reference standard.

#### Wellcome Sanger Institute – Legal and Governance

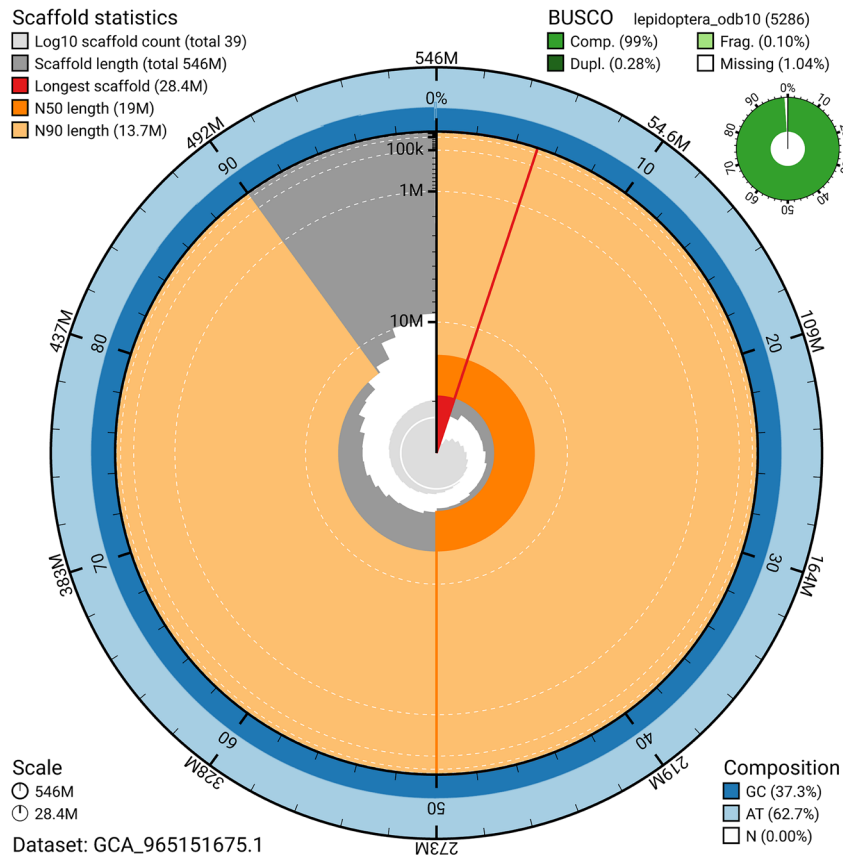
The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the ‘**Darwin Tree of Life Project Sampling Code of Practice**’, which can be found in full on the [Darwin Tree of Life website](#). By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project. Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible. The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international)

Each transfer of samples is further undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Darwin Tree of Life Partner, Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances, other Darwin Tree of Life collaborators.

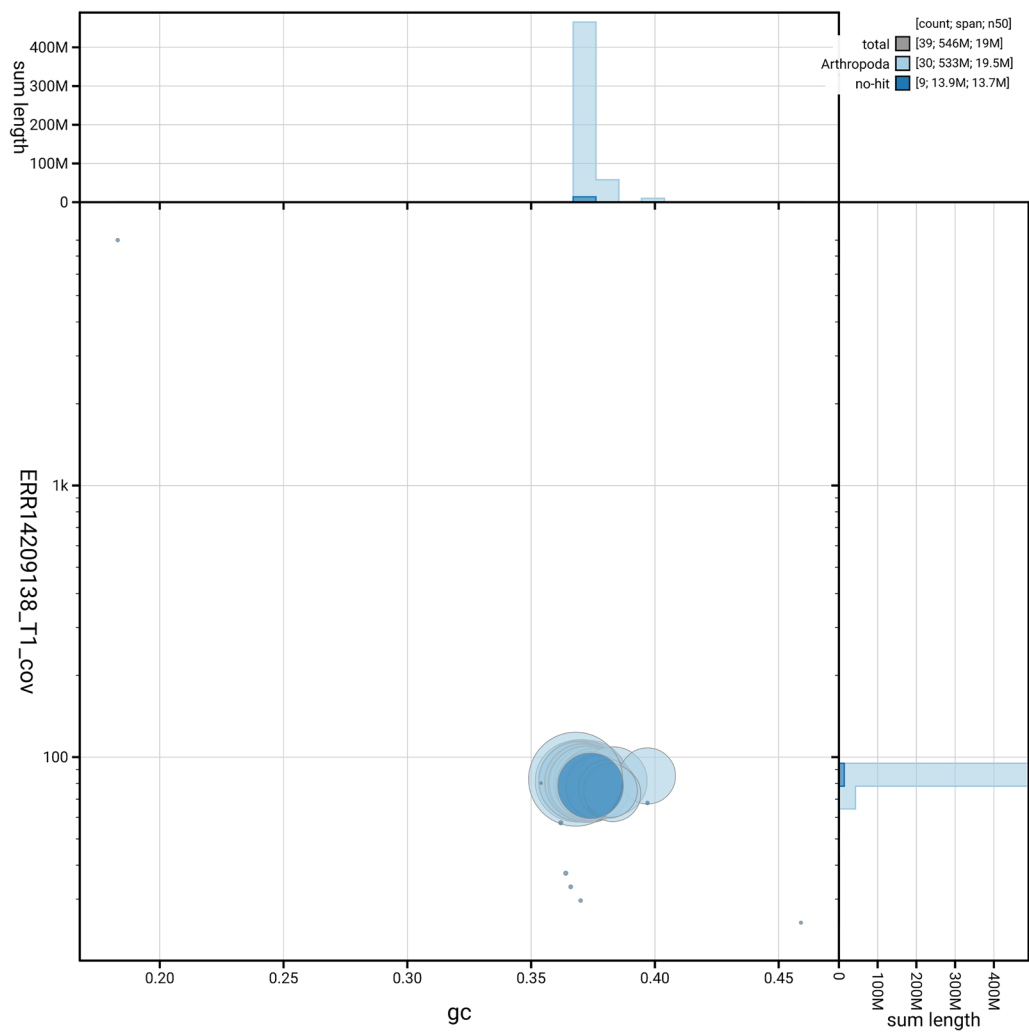


**Figure 4. Evaluation of *k*-mer completeness using MerquyFK.** This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve corresponds to *k*-mers shared by both haplotypes, and the red and blue curves show *k*-mers found only in one of the haplotypes.



**Figure 5. Assembly metrics for *ilChaTrig1.hap1.1*.** The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the set is presented at the top right. An interactive version of this figure can be accessed on the [BlobToolKit viewer](#).





**Figure 6. BlobToolKit GC-coverage plot for ilChaTrig1.hap1.1.** Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the [BlobToolKit viewer](#).

**Table 4. Earth Biogenome Project summary metrics for the *Charanyca trigrammica* assembly.**

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q67	6.C.Q40
Contig N50 length	6 Mb	≥ 1 Mb
Scaffold N50 length	18.96 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 67.0; haplotype 2: 67.6; combined: 67.3	≥ 40
k-mer completeness	Haplotype 1: 35.64%; Haplotype 2: 35.63%; combined: 41.22%	≥ 95%
BUSCO	C:99.0% [S:98.7%; D:0.3%]; F:0.1%; M:0.9%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.97%	≥ 90%

Data availability

European Nucleotide Archive: *Charanyca trigrammica* (treble lines). Accession number [PRJEB85024](#). The genome sequence is released openly for reuse. The *Charanyca trigrammica* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665), the Sanger Institute Tree of Life Programme (PRJEB43745) and Project Psyche (PRJEB71705). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the [Ensembl](#) pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in [Table 1](#) and [Table 2](#).

Production code used in genome assembly at the WSI Tree of Life is available at <https://github.com/sanger-tol>. [Table 5](#) lists software versions used in this study.

Author information

Contributors are listed at the following links:

- Members of the [University of Oxford and Wytham Woods Genome Acquisition Lab](#)
- Members of the [Darwin Tree of Life Barcoding collective](#)
- Members of the [Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team](#)
- Members of [Wellcome Sanger Institute Scientific Operations – Sequencing Operations](#)
- Members of the [Wellcome Sanger Institute Tree of Life Core Informatics team](#)
- Members of the [Tree of Life Core Informatics collective](#)
- Members of the [Darwin Tree of Life Consortium](#)

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	<a href="https://github.com/arq5x/bedtools2">https://github.com/arq5x/bedtools2</a>
BLAST	2.14.0	<a href="ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/">ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/ </a>
BlobToolKit	4.3.9	<a href="https://github.com/blobtoolkit/blobtoolkit">https://github.com/blobtoolkit/blobtoolkit</a>
BUSCO	5.5.0	<a href="https://gitlab.com/ezlab/busco">https://gitlab.com/ezlab/busco</a>
bwa-mem2	2.2.1	<a href="https://github.com/bwa-mem2/bwa-mem2">https://github.com/bwa-mem2/bwa-mem2</a>
Cooler	0.8.11	<a href="https://github.com/open2c/cooler">https://github.com/open2c/cooler</a>
DIAMOND	2.1.8	<a href="https://github.com/bbuchfink/diamond">https://github.com/bbuchfink/diamond</a>
fasta_windows	0.2.4	<a href="https://github.com/tolkit/fasta_windows">https://github.com/tolkit/fasta_windows</a>
FastK	1.1	<a href="https://github.com/thegenemyers/FASTK">https://github.com/thegenemyers/FASTK</a>
GenomeScope2.0	2.0.1	<a href="https://github.com/tbenavi1/genomescope2.0">https://github.com/tbenavi1/genomescope2.0</a>
Gfastats	1.3.6	<a href="https://github.com/vgl-hub/gfastats">https://github.com/vgl-hub/gfastats</a>
GoaT CLI	0.2.5	<a href="https://github.com/genomehubs/goat-cli">https://github.com/genomehubs/goat-cli</a>
Hifiasm	0.19.8-r603	<a href="https://github.com/chhy123/hifiasm">https://github.com/chhy123/hifiasm</a>
HiGlass	1.13.4	<a href="https://github.com/higlass/higlass">https://github.com/higlass/higlass</a>
MercuryFK	1.1.2	<a href="https://github.com/thegenemyers/MERQURY.FK">https://github.com/thegenemyers/MERQURY.FK</a>
Minimap2	2.24-r1122	<a href="https://github.com/lh3/minimap2">https://github.com/lh3/minimap2</a>
MitoHiFi	3	<a href="https://github.com/marcelauliano/MitoHiFi">https://github.com/marcelauliano/MitoHiFi</a>
MultiQC	1.14; 1.17 and 1.18	<a href="https://github.com/MultiQC/MultiQC">https://github.com/MultiQC/MultiQC</a>
Nextflow	23.10.0	<a href="https://github.com/nextflow-io/nextflow">https://github.com/nextflow-io/nextflow</a>
PretextSnapshot	N/A	<a href="https://github.com/sanger-tol/PretextSnapshot">https://github.com/sanger-tol/PretextSnapshot</a>
PretextView	0.2.5	<a href="https://github.com/sanger-tol/PretextView">https://github.com/sanger-tol/PretextView</a>
samtools	1.19.2	<a href="https://github.com/samtools/samtools">https://github.com/samtools/samtools</a>

Software	Version	Source
sanger-tol/ascc	0.1.0	<a href="https://github.com/sanger-tol/ascc">https://github.com/sanger-tol/ascc</a>
sanger-tol/blobtoolkit	0.6.0	<a href="https://github.com/sanger-tol/blobtoolkit">https://github.com/sanger-tol/blobtoolkit</a>
sanger-tol/curationpretext	1.4.2	<a href="https://github.com/sanger-tol/curationpretext">https://github.com/sanger-tol/curationpretext</a>
Seqtk	1.3	<a href="https://github.com/lh3/seqtk">https://github.com/lh3/seqtk</a>
Singularity	3.9.0	<a href="https://github.com/sylabs/singularity">https://github.com/sylabs/singularity</a>
TreeVal	1.4.0	<a href="https://github.com/sanger-tol/treeval">https://github.com/sanger-tol/treeval</a>
YaHS	1.2.2	<a href="https://github.com/c-zhou/yaHS">https://github.com/c-zhou/yaHS</a>

References

Allio R, Schomaker-Bastos A, Romiguier J, *et al.*: **MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics.** *Mol Ecol Resour.* 2020; **20**(4): 892–905.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the Universal Protein Knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Blaxter M, Mieszkowska N, Di Palma F, *et al.*: **Sequence locally, think globally: the Darwin Tree of Life Project.** *Proc Natl Acad Sci U S A.* 2022; **119**(4): e2115642118.  
[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, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

Cheng H, Jarvis ED, Fedrigo O, *et al.*: **Haplotype-resolved assembly of diploid genomes without parental data.** *Nat Biotechnol.* 2022; **40**(9): 1332–1335.  
[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](#)

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

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

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

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](#)

GBIF Secretariat: *Charanyca trigrammica* (Hufnagel, 1766). [WWW document]. 2025.  
[Reference Source](#)

Heath J, Emmet AM: **The moths and butterflies of Great Britain and Ireland: noctuidae (Part LI) and Agaristidae.** Colchester: Harley Books, 1983.  
[Reference Source](#)

Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025.  
[Publisher Full Text](#)

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): giab153.  
[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](#)

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

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

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](#)

Randle Z, Evans-Hill LJ, Parsons MS, *et al.*: **Atlas of Britain and Ireland's larger Moths.** Newbury: Pisces Publications, 2019.  
[Reference Source](#)

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

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

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

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI taxonomy: a comprehensive update on curation, resources and tools.** *Database (Oxford).* 2020; **2020**: baab062.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free 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, Krashennikova 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.** London, UK: Bloomsbury, 2017.

[Reference Source](#)

Wright CJ, Stevens L, Mackintosh A, *et al.*: **Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera.** *Nat Ecol Evol.* 2024; **8**(4): 777–790.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)