



DATA NOTE

The genome sequence of the Pine Carpet, *Thera firmata* Hübner, 1822 (Lepidoptera: Geometridae)

[version 1; peer review: 2 approved]

Natural History Museum 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

v1 First published: 21 Oct 2025, 10:586
<https://doi.org/10.12688/wellcomeopenres.25046.1>
Latest published: 21 Oct 2025, 10:586
<https://doi.org/10.12688/wellcomeopenres.25046.1>

Abstract

We present a genome assembly from an individual male *Thera firmata* (Pine Carpet; Arthropoda; Insecta; Lepidoptera; Geometridae). The assembly contains two haplotypes with total lengths of 301.47 megabases and 300.34 megabases. Most of haplotype 1 (98.51%) is scaffolded into 19 chromosomal pseudomolecules, including the Z sex chromosome. Haplotype 2 was assembled to scaffold level. Gene annotation of this assembly on Ensembl identified 11 832 protein-coding genes. The mitochondrial genome has also been assembled, with a length of 16.59 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

Thera firmata; Pine Carpet; genome sequence; chromosomal; Lepidoptera



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

Open Peer Review

Approval Status

	1	2
version 1 21 Oct 2025	 view	 view
1. Alena Suchackova Bartonova , Biology Centre of the Czech Academy of Sciences, České Budějovice, Czech Republic		
2. Sara Goodwin , Cold Spring Harbor Laboratory, New York, USA		
Any reports and responses or comments on the article can be found at the end of the article.		

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

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 Natural History Museum Genome Acquisition Lab *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: Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding Collective, Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team *et al.* **The genome sequence of the Pine Carpet, *Thera firmata* Hübner, 1822 (Lepidoptera: Geometridae) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, 10:586 <https://doi.org/10.12688/wellcomeopenres.25046.1>

First published: 21 Oct 2025, 10:586 <https://doi.org/10.12688/wellcomeopenres.25046.1>

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; Larentiinae; *Thera*; *Thera firmata* Hübner, 1822 (NCBI:txid326957)

Background

Thera firmata, *T. britannica*, *T. vetustata* and *T. obeliscata* in Britain and Ireland, together with *T. variata* and *T. cembrae* from mainland Europe, form a group of species within *Thera* that are closely related. Whether each named species is genetically distinct remains debated, and taxonomic uncertainty persists (Waring & Townsend, 2017).

Thera firmata (often treated as *Pennithera firmata*) is the Pine Carpet, a resident moth of conifer woodland and plantations (Kimber, 2025). The species overwinters on the foodplant as an egg or tiny larva; feeding continues through autumn and spring, with timing quite variable, and larvae are usually present from autumn to July. Pupation takes place among pine needles, either on the tree or in the litter (Waring & Townsend, 2017). The species has one generation, and adults fly from late July to November (Waring & Townsend, 2017). It resembles Grey Pine Carpet (*Thera obeliscata*) but has a more sharply indented inner edge to the dark forewing band. Found locally throughout Britain wherever Scots pine occurs, it comes to light and may be seen by day when disturbed (Kimber, 2025).

We present a chromosome-level genome sequence for *Thera firmata*, the Pine Carpet. The assembly was produced using the Tree of Life pipeline from a specimen collected in Beinn Eighe National Nature Reserve, Scotland, UK (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).



Figure 1. Photograph of the *Thera firmata* (ilTheFirm3) specimen used for genome sequencing.

This genome sequence for *T. firmata*, along with the genomes of *T. britannica* (Boyes *et al.*, 2023) and *T. obeliscata* (Lees *et al.*, 2023), will be useful for finding variable genetic loci to help resolve taxonomy in this genus and for studying the genetic basis of adaptation to coniferous larval food plants.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult male *Thera firmata* (specimen ID NHMUK014543810, ToLID ilTheFirm3; Figure 1), collected from Beinn Eighe National Nature Reserve National Nature Reserve, Scotland, UK (latitude 57.63, longitude −5.35) on 2021-09-09. The specimen was collected and identified by David Lees (Natural History Museum). 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 ilTheFirm3 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the abdomen was homogenised by powermashing using a PowerMasher II tissue disruptor. HMW DNA was extracted using the Automated MagAttract v2 protocol. We used centrifuge-mediated fragmentation to produce DNA fragments in the 8–10 kb range, following the Covaris g-TUBE protocol for ultra-low input (ULI). Sheared DNA was purified by automated 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 2.36 ng/μL and a yield of 306.80 ng.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Prior to library preparation, the DNA was fragmented to ~10 kb. Ultra-low-input (ULI) libraries were prepared using the PacBio SMRTbell® Express

Template Prep Kit 2.0 and gDNA Sample Amplification Kit. Samples were normalised to 20 ng DNA. Single-strand overhang removal, DNA damage repair, and end-repair/A-tailing were performed according to the manufacturer's instructions, followed by adapter ligation. A 0.85× pre-PCR clean-up was carried out with Promega ProNex beads.

The DNA was evenly divided into two aliquots for dual PCR (reactions A and B), both following the manufacturer's protocol. A 0.85× post-PCR clean-up was performed with ProNex beads. DNA concentration was measured using a Qubit Fluorometer v4.0 (Thermo Fisher Scientific) with the Qubit HS Assay Kit, and fragment size was assessed on an Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit. PCR reactions A and B were then pooled, ensuring a total mass of ≥500 ng in 47.4 µL.

The pooled sample underwent another round of DNA damage repair, end-repair/A-tailing, and hairpin adapter ligation. A 1× clean-up was performed with ProNex beads, followed by DNA quantification using the Qubit and fragment size analysis using the Agilent Femto Pulse. Size selection was performed on the Sage Sciences PippinHT system, with target fragment size determined by Femto Pulse analysis (typically 4–9 kb). Size-selected libraries were cleaned with 1.0× ProNex beads and normalised to 2 nM before sequencing.

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 and thorax of the iTheFirm3 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 10–16 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 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. 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 6 breaks and 13 joins. The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation>. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The Merqury.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 *Thera firmata* specimen generated 26.25 Gb (gigabases) from 3.15 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 294.85 Mb, with a heterozygosity of 0.70% and repeat content of 18.31% ([Figure 2](#)). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 85× coverage. Hi-C sequencing produced 109.86 Gb from 727.56 million reads, which were used to scaffold the assembly. [Table 1](#) summarises the specimen and sequencing details.

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 301.47 Mb in 172 scaffolds, with 314 gaps, and a scaffold N50 of 16.26 Mb ([Table 2](#)).

Most of the assembly sequence (98.51%) was assigned to 19 chromosomal-level scaffolds, representing 18 autosomes and

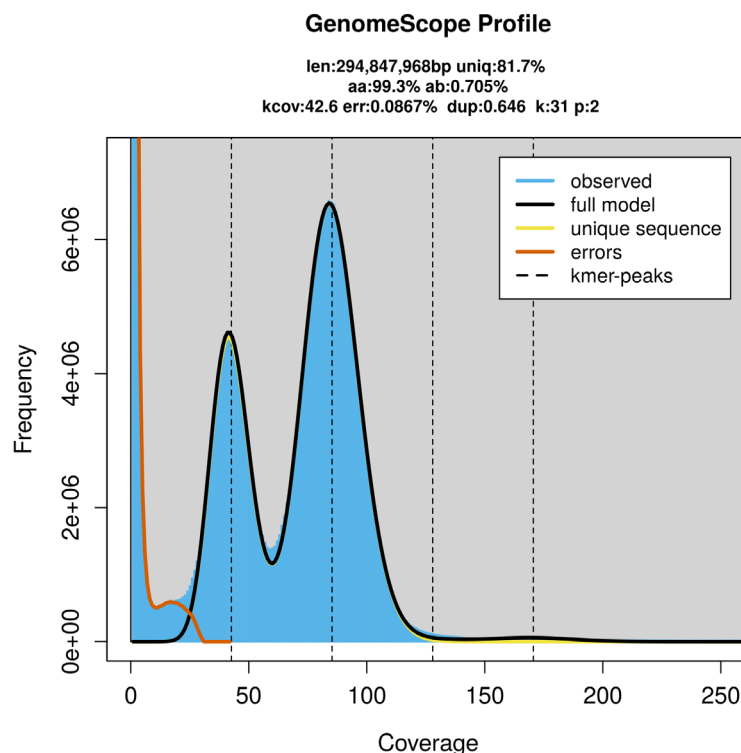


Figure 2. Frequency distribution of k -mers generated using GenomeScope2. The plot shows observed and modelled k -mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB74679.

Platform	PacBio HiFi	Hi-C
ToLID	ilTheFirm3	ilTheFirm3
Specimen ID	NHМУK014543810	NHМУK014543810
BioSample (source individual)	SAMEA14448145	SAMEA14448145
BioSample (tissue)	SAMEA14448224	SAMEA14448222
Tissue	abdomen	head and thorax
Instrument	Revio	Illumina NovaSeq 6000
Run accessions	ERR12875209; ERR12875210	ERR12893016
Read count total	3.15 million	727.56 million
Base count total	26.25 Gb	109.86 Gb

Table 2. Genome assembly statistics.

Assembly name	ilTheFirm3.hap1.1	ilTheFirm3.hap2.1
Assembly accession	GCA_964188155.1	GCA_964188035.1
Assembly level	chromosome	scaffold
Span (Mb)	301.47	300.34
Number of chromosomes	19	Scaffold-level
Number of contigs	486	398
Contig N50	1.74 Mb	1.97 Mb
Number of scaffolds	172	114
Scaffold N50	16.26 Mb	16.3 Mb
Longest scaffold length (Mb)	25.77	-
Sex chromosomes	Z	-
Organelles	Mitochondrion: 16.59 kb	-

the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Chromosome Z was assigned by synteny with the genome of *Thera obeliscata* (GCA_947578465.1) (Lees et al., 2023).

The mitochondrial genome was also assembled (length 16.59 kb, OZ076684.1). 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 64.6, and for haplotype 2, 64.9. When the two haplotypes are combined, the assembly achieves an estimated QV of 64.7. The *k*-mer completeness

is 81.93% for haplotype 1, 81.98% for haplotype 2, and 95.51% for the combined haplotypes (Figure 4).

BUSCO analysis using the lepidoptera_odb10 reference set (*n* = 5 286) identified 97.9% of the expected gene set (single = 97.0%, duplicated = 0.9%) 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) and the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric,

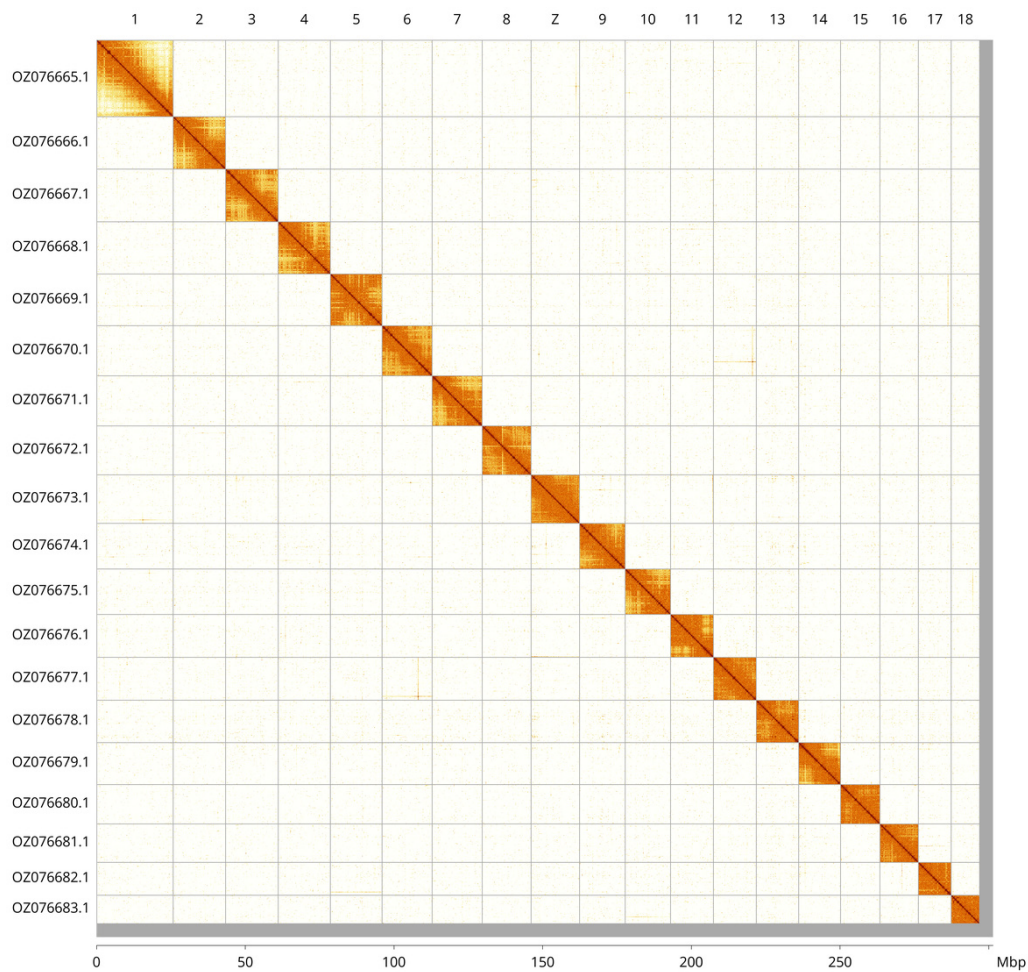


Figure 3. Hi-C contact map of the *Thera firmata* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Thera firmata* ilTheFirm3.

INSDC accession	Molecule	Length (Mb)	GC%
OZ076665.1	1	25.77	37.50
OZ076666.1	2	17.67	37.50
OZ076667.1	3	17.64	38
OZ076668.1	4	17.57	38
OZ076669.1	5	17.43	37.50
OZ076670.1	6	16.83	38
OZ076671.1	7	16.76	37.50
OZ076672.1	8	16.49	38

INSDC accession	Molecule	Length (Mb)	GC%
OZ076674.1	9	15.39	38
OZ076675.1	10	15.25	38
OZ076676.1	11	14.42	38
OZ076677.1	12	14.37	38
OZ076678.1	13	14.23	38
OZ076679.1	14	14.11	38.50
OZ076680.1	15	13.25	38
OZ076681.1	16	12.90	38
OZ076682.1	17	11.03	38
OZ076683.1	18	9.61	38
OZ076673.1	Z	16.26	38

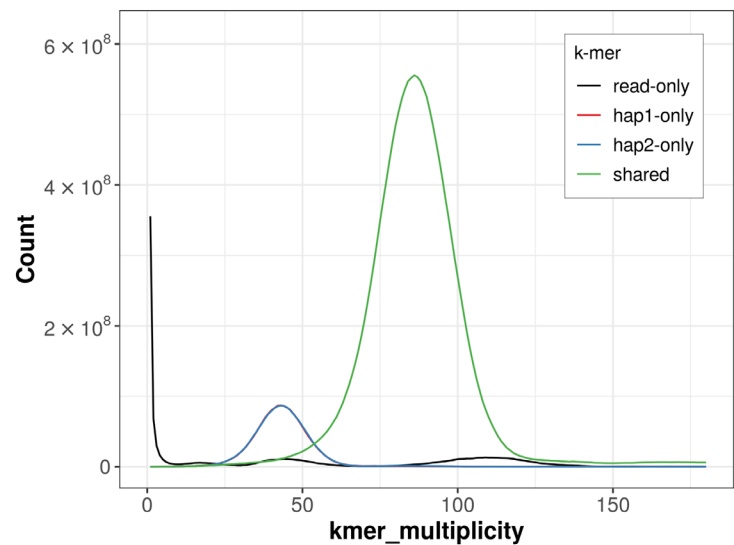


Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers 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.

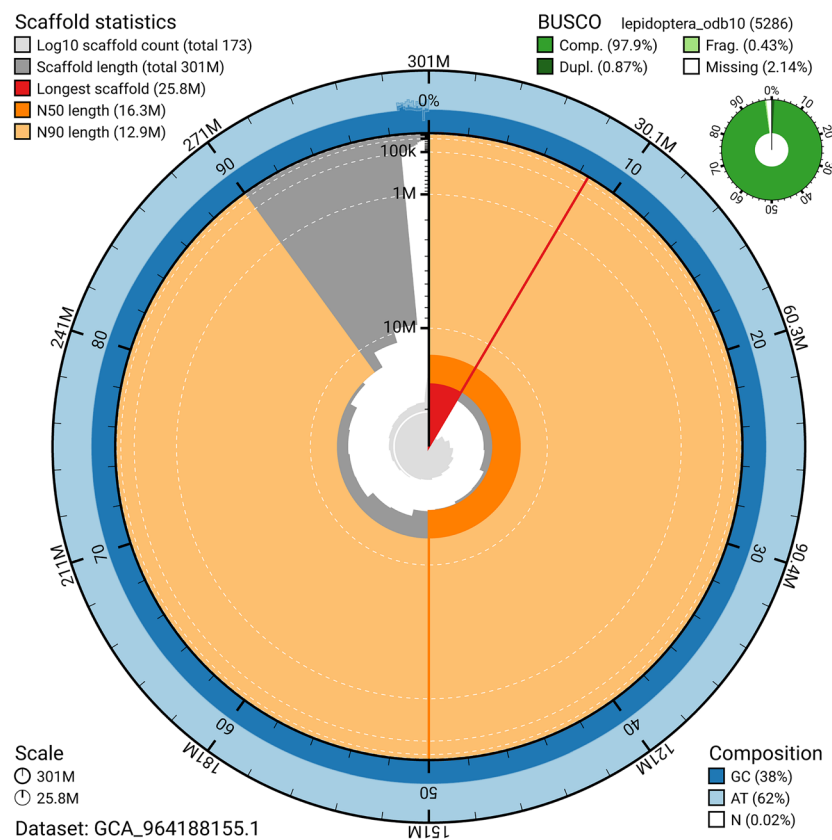


Figure 5. Assembly metrics for ilTheFirm3.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost 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](#).

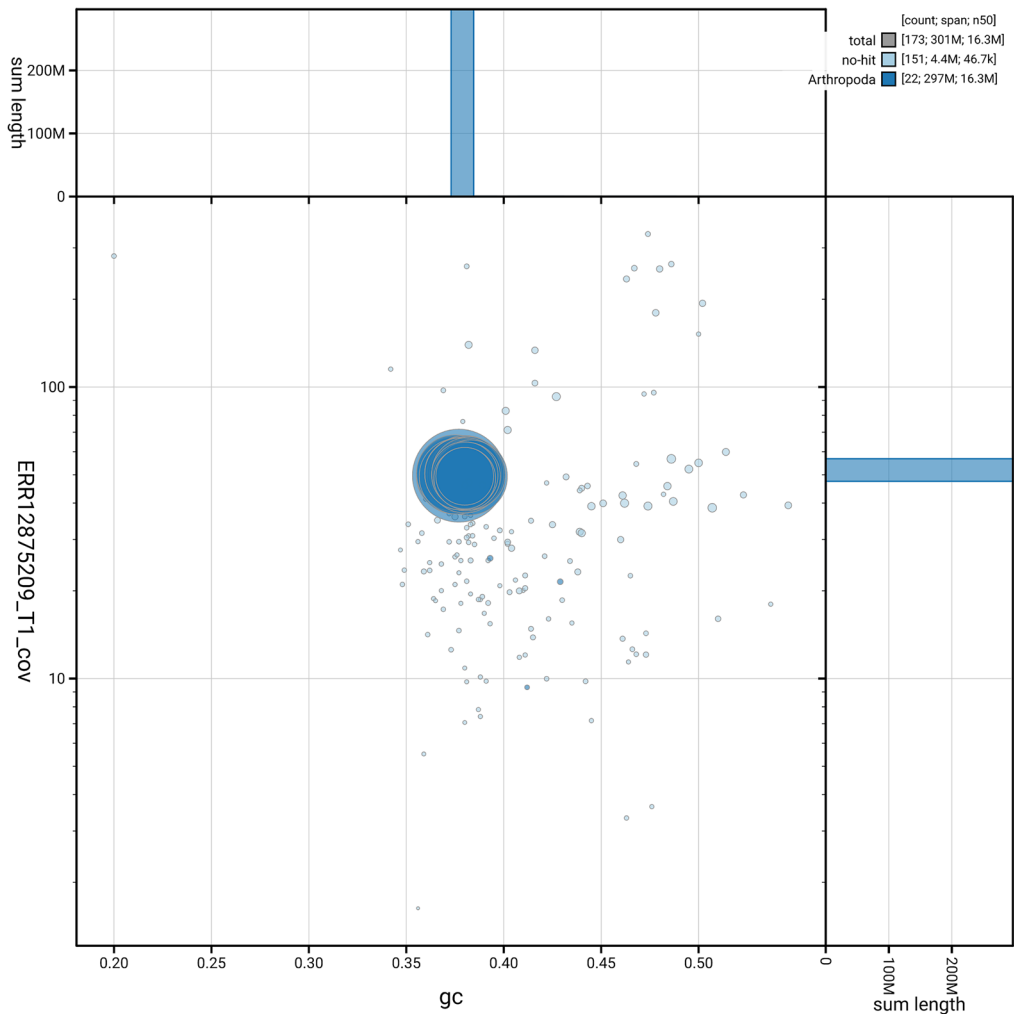


Figure 6. BlobToolKit GC-coverage plot for ilTheFirm3.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Thera firmata* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.C.Q64	6.C.Q40
Contig N50 length	1.74 Mb	≥ 1 Mb
Scaffold N50 length	16.26 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 64.6; haplotype 2: 64.9; combined: 64.7	≥ 40
k-mer completeness	Haplotype 1: 81.93%; Haplotype 2: 81.98%; combined: 95.51%	≥ 95%
BUSCO	C:97.9% [S:97.0%; D:0.9%]; F:0.4%; M:1.7%; n:5 286	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	98.51%	≥ 90%

calculated for the haplotype 1, is **6.C.Q64**, meeting the recommended reference standard.

Genome annotation report

The *Thera firmata* genome assembly (GCA_964188155.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 21 432 transcribed mRNAs from 11 832 protein-coding and 1 841 non-coding genes. The average transcript length is 10 912.07 bp, with an average of 1.57 coding transcripts per gene and 6.92 exons per transcript. For further information about the annotation, please refer to the [Ensembl annotation page](#).

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.

Data availability

European Nucleotide Archive: *Thera firmata*. Accession number [PRJEB74679](#). The genome sequence is released openly for reuse. The *Thera firmata* 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.

Table 5. Software versions and sources.

Software	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	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
Goat CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.8-r603	https://github.com/chhy123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass

Software	Version	Source
MercuryFK	1.1.2	https://github.com/thegenemyers/MERQUERY.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	23.10.0	https://github.com/nextflow-io/nextflow
PretextView	-	https://github.com/sanger-tol/PretextView
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.6.0	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

Author information

Contributors are listed at the following links:

- Members of the [Natural History Museum 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](#)

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](#)

Boyce D, Holland PWH; Darwin Tree of Life Consortium, et al.: **The genome sequence of the Spruce Carpet Moth, *Thera britannica* (Turner, 1925) [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 114.
[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): gjab008.

[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](#)

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): gjaa153.

[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](#)

Kimber I: **Pine carpet — *Pennithera firmata*.** 2025.

[Reference Source](#)

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](#)

Lees DC, Natural History Museum Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, *et al.*: **The genome sequence of the Grey Pine Carpet, *Thera obeliscata* (Hübner, 1787) [version 1; peer review: 2 approved].** *Wellcome Open Res*. 2023; **8**: 408.

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

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.*: **Merquy: 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**: baaa062.

[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, Krashenninnikova 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: **Field Guide to the Moths of Great Britain and Ireland.** London: Bloomsbury Wildlife, 2017.

[Reference Source](#)

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

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

Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 05 January 2026

<https://doi.org/10.21956/wellcomeopenres.27612.r140058>

© 2026 Goodwin S. This is an open access peer review report 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.



Sara Goodwin

Cold Spring Harbor Laboratory, New York, USA

Review of "The genome sequence of the Pine Carpet , *Thera firmata* (Hubner, 1822) (Lepidoptera Geometridae)

"

The is a data in the in The Darwin Tree of Life Project. It is a genome assembly from a single male individual. The report notes a high-quality, near chromosome-scale assembly for the Pine Carpet moth. Both PacBio HiFi Arima Hi-C were used to generate a phased assembly with two haplotype. Haplotype 1 achieved 98.51% scaffolding across 19 pseudo molecules, while haplotype 2 is assembled to the scaffold level. . Raw data and assemblies have been deposited at ENA under PRJEB74679

It may be useful to explain why the k-mer completeness of individual haplotypes ~81.9% but 95.51% combined. While unsurprising, it may be confusing for some readers.

There is a typo in the first paragraph of the methods section "Beinn Eighe National Nature Reserve National Nature Reserve"

Review of "The genome sequence of the Pine Carpet , *Thera firmata* (Hubner, 1822) (Lepidoptera Geometridae)

"

The is a data in the in The Darwin Tree of Life Project. It is a genome assembly from a single male individual. The report notes a high-quality, near chromosome-scale assembly for the Pine Carpet moth. Both PacBio HiFi Arima Hi-C were used to generate a phased assembly with two haplotype. Haplotype 1 achieved 98.51% scaffolding across 19 pseudo molecules, while haplotype 2 is assembled to the scaffold level. . Raw data and assemblies have been deposited at ENA under PRJEB74679

It may be useful to explain why the k-mer completeness of individual haplotypes ~81.9% but 95.51% combined. While unsurprising, it may be confusing for some readers.

There is a typo in the first paragraph of the methods section “Beinn Eighe National Nature Reserve National Nature Reserve”

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: Plant and Animal genomics and sequencing technologies

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 02 January 2026

<https://doi.org/10.21956/wellcomeopenres.27612.r140060>

© 2026 Bartonova A. This is an open access peer review report 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.



Alena Suchackova Bartonova

Biology Centre of the Czech Academy of Sciences, České Budějovice, Czech Republic

This is a note for the genome sequence of the Pine Carpet moth, *Thera firmata*, produced by the Darwin Tree of Life initiative. The sample preparation and genome production follow standard procedures. Links to all relevant protocols and the produced data are provided. I have not spotted anything to change.

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: entomology, butterflies, ecology, biogeography, population genetics

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.
