DATA NOTE

The genome sequence of the Common Plume moth, *Emmelina monodactyla* (Linnaeus, 1758) [version 1; peer review: awaiting peer review]

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

[1]UK Centre for Ecology and Hydrology, Wallingford, Oxfordshire, UK
[2]Independent researcher, Shrewsbury, Shropshire, UK

[+] Deceased author

v1 **First published:** 23 Feb 2023, **8**:97
https://doi.org/10.12688/wellcomeopenres.19035.1
Latest published: 23 Feb 2023, **8**:97
https://doi.org/10.12688/wellcomeopenres.19035.1

Abstract

We present a genome assembly from an individual female *Emmelina monodactyla* (the Common Plume; Arthropoda; Insecta; Lepidoptera; Pterophoridae). The genome sequence is 312 megabases in span. Most of the assembly is scaffolded into 30 chromosomal pseudomolecules, including the assembled W and Z sex chromosomes. The mitochondrial genome has also been assembled and is 15.3 kilobases in length. Gene annotation of this assembly on Ensembl identified 13,451 protein coding genes.

Keywords

Emmelina monodactyla, the Common Plume moth, 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; **Kerry B**: Writing – Original Draft Preparation;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (206194, 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, University of Oxford and Wytham Woods Genome Acquisition Lab, Wellcome Sanger Institute Tree of Life programme *et al.* **The genome sequence of the Common Plume moth, *Emmelina monodactyla* (Linnaeus, 1758) [version 1; peer review: awaiting peer review]** Wellcome Open Research 2023, **8**:97 https://doi.org/10.12688/wellcomeopenres.19035.1

First published: 23 Feb 2023, **8**:97 https://doi.org/10.12688/wellcomeopenres.19035.1

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Pterophoroidea; Pterophoridae; Pterophorinae; *Emmelina*; *Emmelina monodactyla* (Linnaeus, 1758) (NCBI:txid467774).

Background

Emmelina monodactyla (Common Plume moth) is a member of the distinctive Pterophoridae family, which have a striking T-shaped appearance and separated, feather-like wings which give them the common name 'plume moths'. However, these clefted wings can be difficult to observe at rest, as the moths' wings are rolled and held perpendicular to its slender body. The species is endemic to Europe but is now found almost worldwide. In the United Kingdom the larvae feed on *Convolvulus* (bindweed) species, but in other countries additional food plants have been identified, including *Ipomoea batatas* (Tóth *et al.*, 2003).

E. monodactyla is found in a variety of habitats wherever *Convolvulus* is present and is consequently widespread in the United Kingdom. Larvae are produced in two overlapping generations and adults overwinter, so can therefore be found year-round in the UK. In appearance it varies from light to dark brown, with a brown-streaked pale buff coloured dorsal longitudinal band on the abdomen. The wingspan of *E. monodactyla* varies between 18–26 mm, and it possesses a pair of unequal length spurs on each of its back legs (Simpson *et al.*, 2020). Lighter colour forms including a pale grey-white have also been noted in Sweden (Tóth *et al.*, 2003). Larvae have a greenish-yellow appearance and a wide green dorsal band, with a fine interrupted yellow line down the centre. Larvae occasionally have red dorsal markings. There are 39 other Pterophoridae present in the UK, several of which have similar colouring, size and wing position to *E. monodactyla* which can make it difficult to differentiate without genital dissection, complicated by the variety of colour and markings that can be present on adults. The adults are however one of the few Pterophoridae found early in the year due to their overwintering practices (Manley, 2011).

An assembled genome sequence for *E. monodactyla* will add to the growing set of resources for understanding the biology and adaptations of Lepidoptera.

Genome sequence report

The genome was sequenced from one female *E. monodactyla* specimen (Figure 1) collected from Wytham Woods, UK (latitude 51.77, longitude –1.34) A total of 38-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 99-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 21 missing or mis-joins and removed five haplotypic duplications, reducing the assembly length by 1.04% and the scaffold number by 34.62%.

Figure 1. Photograph of the *Emmelina monodactyla* (ilEmmMono1) specimen used for genome sequencing.

The final assembly has a total length of 312.4 Mb in 34 sequence scaffolds with a scaffold N50 of 10.4 Mb (Table 1). Most (99.94%) of the assembly sequence was assigned to 30 chromosomal-level scaffolds, representing 28 autosomes and the W and Z sex chromosome. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 2–Figure 5; Table 2). The assembly has a BUSCO v5.3.2 (Manni *et al.*, 2021) completeness of 97.6% (single 96.9%, duplicated 0.7%) using the lepidoptera_odb10 reference set. While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

Genome annotation report

The *E. monodactyla* genome assembly GCA_916618145.1 was annotated using the Ensembl rapid annotation pipeline (Table 1; Ensembl accession number GCA_916618145.1). The resulting annotation includes 25,430 transcribed mRNAs from 13,451 protein-coding and 3,447 non-coding genes.

Methods

Sample acquisition and nucleic acid extraction

One female *E. monodactyla* specimen (ilEmmMono1) was collected in Wytham Woods, Oxfordshire (biological vice-county: Berkshire), UK (latitude 51.77, longitude –1.34) on 8 October 2020. The specimen was taken from woodland habitat using a light trap. The specimen was collected and identified by Douglas Boyes (University of Oxford) and snap-frozen on dry ice.

DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute (WSI). The ilEmmMono1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. Whole organism tissue was disrupted using

Table 1. Genome data for *Emmelina monodactyla*, ilEmmMono1.1.

Project accession data		
Assembly identifier	ilEmmMono1.1	
Species	*Emmelina monodactyla*	
Specimen	ilEmmMono1	
NCBI taxonomy ID	467774	
BioProject	PRJEB46324	
BioSample ID	SAMEA8603203	
Isolate information	ilEmmMono1: female	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	54.9	*≥ 50*
k-mer completeness	99.99%	*≥ 95%*
BUSCO**	C:97.6%[S:96.9%,D:0.7%], F:0.5%,M:1.9%,n:5,286	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	99.94%	*≥ 95%*
Sex chromosomes	Z and W	*localised homologous pairs*
Organelles	Mitochondrial genome assembled	*complete single alleles*
Raw data accessions		
PacificBiosciences SEQUEL II	ERR6808006	
10X Genomics Illumina	ERR6688551–ERR6688554	
Hi-C Illumina	ERR6688550	
Genome assembly		
Assembly accession	GCA_916618145.1	
Accession of alternate haplotype	GCA_916618085.1	
Span (Mb)	312.4	
Number of contigs	60	
Contig N50 length (Mb)	10.3	
Number of scaffolds	34	
Scaffold N50 length (Mb)	10.4	
Longest scaffold (Mb)	16.3	
Genome annotation		
Number of protein-coding genes	13,452	
Number of non-coding genes	3,447	
Number of gene transcripts	25,430	

* Assembly metric benchmarks are adapted from column VGP-2020 of "Table 1: Proposed standards and metrics for defining genome assembly quality" from (Rhie *et al.*, 2021).

** BUSCO scores based on the lepidoptera_odb10 BUSCO set using v5.3.2. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/ilEmmMono1.1/dataset/CAKAJY01.1/busco.

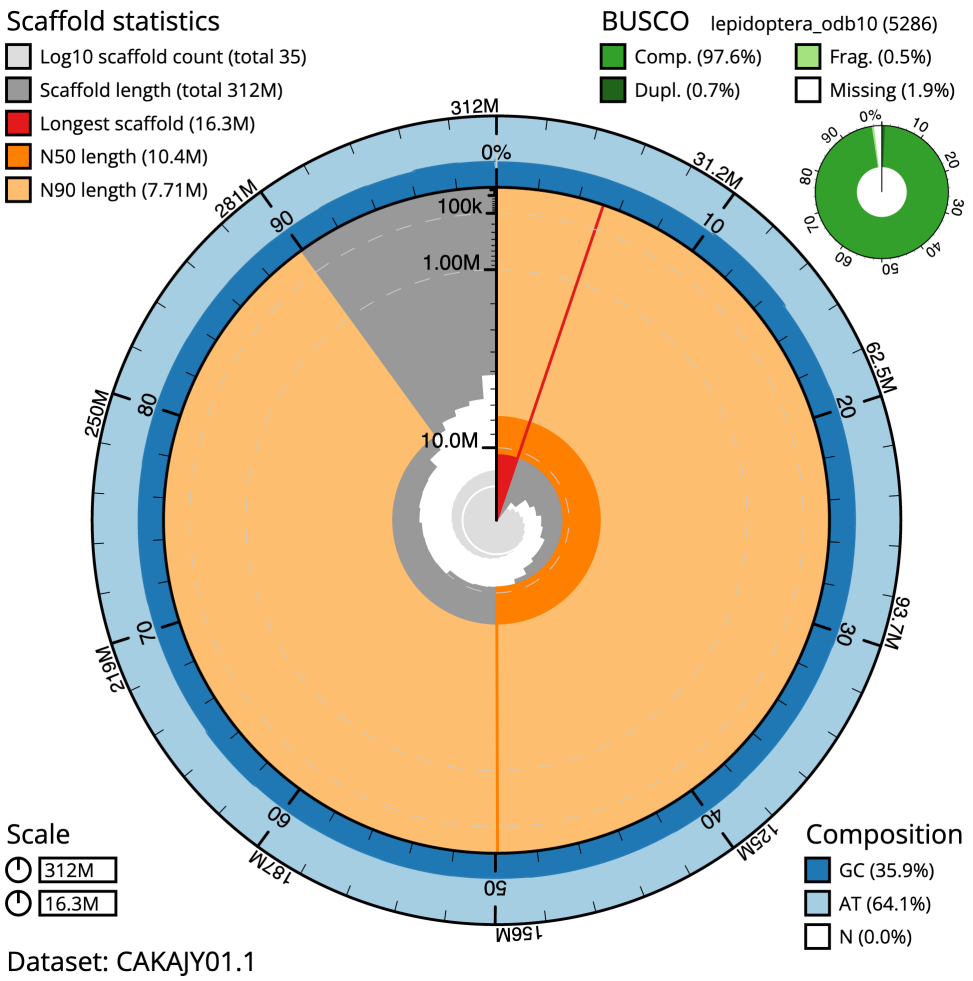

Figure 2. Genome assembly of *Emmelina monodactyla*, ilEmmMono1.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 312,410,734 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (16,349,857 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (10,389,392 and 7,713,513 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the lepidoptera_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilEmmMono1.1/dataset/CAKAJY01.1/snail.

a Nippi Powermasher fitted with a BioMasher pestle. High molecular weight (HMW) DNA was extracted using the Qiagen MagAttract HMW DNA extraction kit. Low molecular weight DNA was removed from a 20 ng aliquot of extracted DNA using 0.8X AMpure XP purification kit prior to 10X Chromium sequencing; a minimum of 50 ng DNA was submitted for 10X sequencing. HMW DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system with speed setting 30. Sheared DNA was purified by solid-phase reversible immobilisation using AMPure PB beads with a 1.8X ratio of beads to sample to remove the shorter fragments and concentrate the DNA sample. The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

Sequencing

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

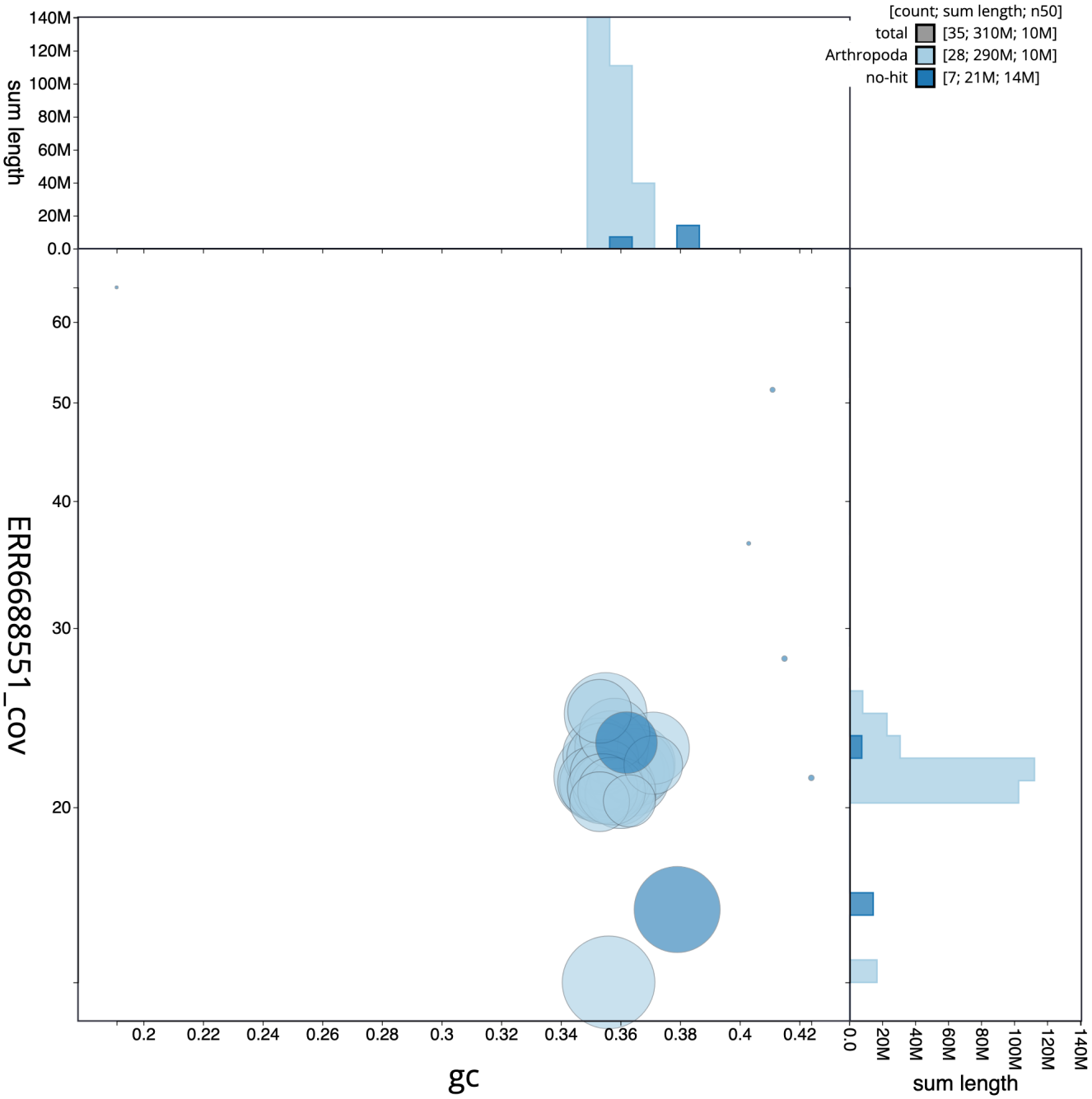

Figure 3. Genome assembly of *Emmelina monodactyla*, ilEmmMono1.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilEmmMono1.1/dataset/CAKAJY01.1/blob.

NovaSeq 6000 (10X) instruments. Hi-C data were also generated from tissue of ilEmmMono1 using the Arima v2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

Genome assembly

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021) and haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2020). One round of polishing was performed by aligning 10X Genomics read data to the assembly with Long Ranger ALIGN, calling variants with freebayes (Garrison & Marth, 2012). The assembly was then scaffolded with Hi-C data (Rao *et al.*, 2014) using SALSA2 (Ghurye *et al.*, 2019). The assembly was checked for contamination as described previously (Howe *et al.*, 2021).

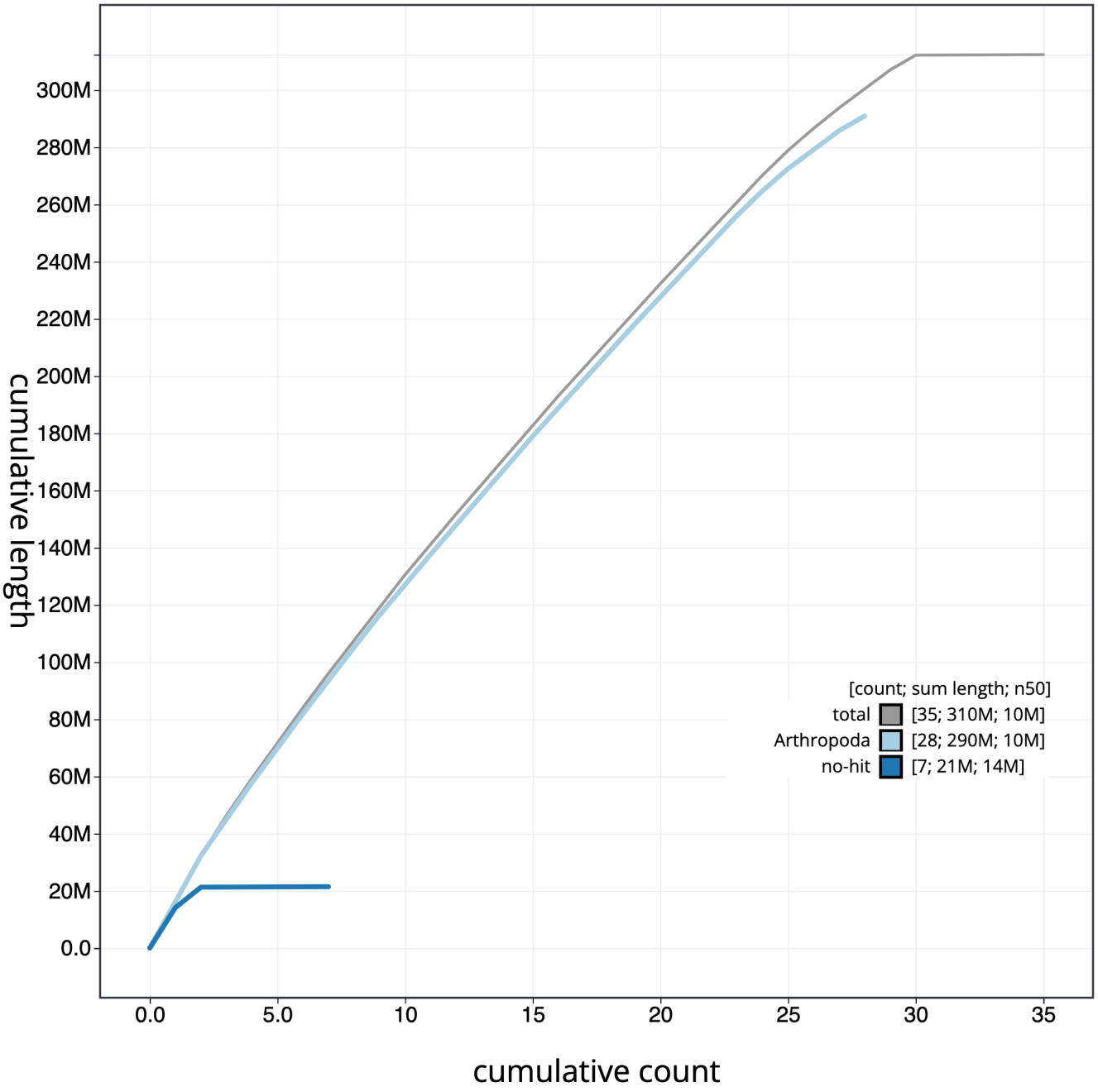

Figure 4. Genome assembly of *Emmelina monodactyla*, ilEmmMono1.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilEmmMono1.1/dataset/CAKAJY01.1/cumulative.

Manual curation was performed using HiGlass (Kerpedjiev *et al.*, 2018) and Pretext (Harry, 2022). The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2022), which performed annotation using MitoFinder (Allio *et al.*, 2020). The genome was analysed and BUSCO scores generated within the BlobToolKit environment (Challis *et al.*, 2020). Table 3 contains a list of all software tool versions used, where appropriate.

Genome annotation

The Ensembl gene annotation system (Aken *et al.*, 2016) was used to generate annotation for the *E. monodactyla*

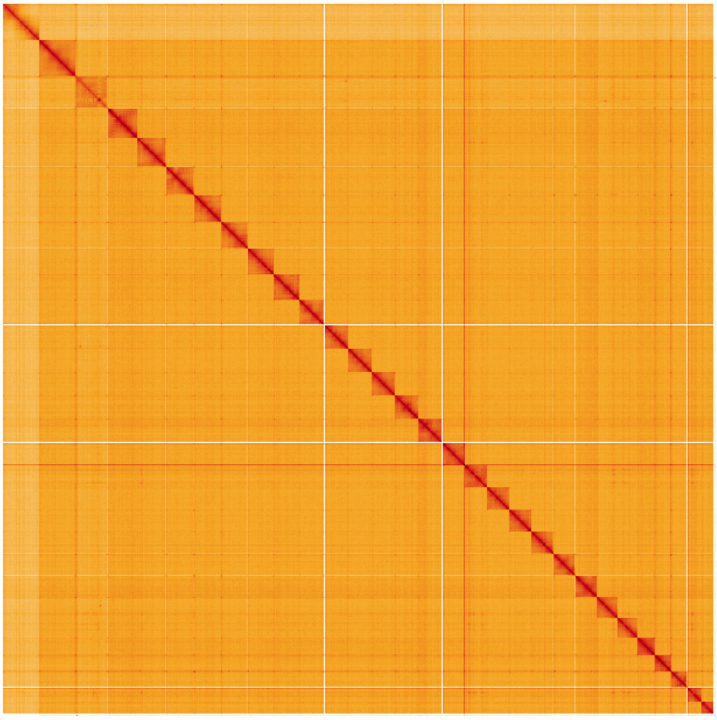

Figure 5. Genome assembly of *Emmelina monodactyla*, ilEmmMono1.1: Hi-C contact map. Hi-C contact map of the ilEmmMono1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=eDA4K_MhQXq-SBY0oHCLOw.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Emmelina monodactyla*, ilEmmMono1.

INSDC accession	Chromosome	Size (Mb)	GC%
OU745300.1	1	15.91	35.3
OU745302.1	2	12.98	35.5
OU745303.1	3	12.57	36.1
OU745304.1	4	12.17	36.3
OU745305.1	5	12	36.5
OU745306.1	6	11.58	35.7
OU745307.1	7	11.58	35.5
OU745308.1	8	11.31	36.5
OU745309.1	9	10.79	35.9
OU745310.1	10	10.44	35.8
OU745311.1	11	10.39	35.2
OU745312.1	12	10.35	35.7
OU745313.1	13	10.27	35.3
OU745314.1	14	10.27	35.7
OU745315.1	15	9.87	37.1
OU745316.1	16	9.83	35.4
OU745317.1	17	9.81	35.1
OU745318.1	18	9.77	35.6
OU745319.1	19	9.59	35.5
OU745320.1	20	9.49	36
OU745321.1	21	9.45	35.8
OU745322.1	22	9.38	35.4
OU745323.1	23	8.66	35.7
OU745324.1	24	7.71	35.3
OU745325.1	25	7.17	36.2
OU745326.1	26	6.78	35.3
OU745296.1	27	6.51	37.1
OU745297.1	28	5.15	36.3
OU745301.1	W	14.09	37.9
OU745299.1	Z	16.35	35.6
OU745298.1	MT	0.02	19.4
-	unplaced	0.18	41.4

Table 3. Software tools and versions used.

Software tool	Version	Source
BlobToolKit	3.5.0	Challis *et al.*, 2020
freebayes	1.3.1-17-gaa2ace8	Garrison & Marth, 2012
Hifiasm	0.15.3-r339	Cheng *et al.*, 2021
HiGlass	1.11.6	Kerpedjiev *et al.*, 2018
Long Ranger ALIGN	2.2.2	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines
MitoHiFi	2	Uliano-Silva *et al.*, 2022
PretextView	0.2	Harry, 2022
purge_dups	1.2.3	Guan *et al.*, 2020
SALSA	2.2	Ghurye *et al.*, 2019

assembly (GCA_916618145.1). Annotation was created primarily through alignment of transcriptomic data to the genome, with gap filling via protein to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 2019).

Ethics and compliance issues

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. 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. All efforts are undertaken to minimise the suffering of animals used for sequencing. 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: *Emmelina monodactyla (*common plume). Accession number PRJEB46324; https://identifiers.org/ena.embl/PRJEB46324. (Wellcome Sanger Institute, 2021)

The genome sequence is released openly for reuse. The *Emmelina monodactyla* genome sequencing initiative is part of the Darwin Tree of Life (DToL) project. All raw sequence data and the assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in Table 1.

Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.4789928.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.4893703.

Members of the Wellcome Sanger Institute Tree of Life programme are listed here: https://doi.org/10.5281/zenodo.4783585.

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

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.5013541.

Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783558.

References

Aken BL, Ayling S, Barrell D, *et al.*: **The Ensembl gene annotation system.** *Database (Oxford).* 2016; **2016**: baw093.
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

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

Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing.** 2012.
Publisher Full Text

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

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898.
PubMed Abstract | Publisher Full Text | Free Full Text

Harry E: **PretextView (Paired REadTEXTure Viewer): A desktop application for viewing pretext contact maps.** 2022; (Accessed: 19 October 2022).
Reference Source

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

Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HiGlass: Web-based visual exploration and analysis of genome interaction maps.** *Genome Biol.* 2018; **19**(1): 125.
PubMed Abstract | Publisher Full Text | Free Full Text

Manley C: **British Moths and Butterflies.** London: Bloomsbury Publishing, 2011.

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

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): 1665–80.
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

Simpson T, Duncan I, Williams M: **Moths of the West Midlands.** Newbury: Pisces Publication, 2020.

Tóth P, Tóthová M, Cagáň L: ***Emmelina monodactyla* (Linnaeus, 1758) (Lepidoptera: Pterophoridae), its parasitization and potential as a biological control agent of Field or Lesser Bindweed (*Convolvulus arvensis L.*).** *Entomologist's Gazette.* 2003; **54**(4): 233–241.
Reference Source

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio High Fidelity reads.** *bioRxiv.* [Preprint], 2022.
Publisher Full Text

UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–D515.
PubMed Abstract | Publisher Full Text | Free Full Text

Wellcome Sanger Institute: **The genome sequence of the Common Plume moth, *Emmelina monodactyla* (Linnaeus, 1758)**. European Nucleotide Archive. [dataset], accession number PRJEB46324, 2021.