DATA NOTE

The genome sequence of the Brown Scallop, *Philereme vetulata* (Denis and Schiffermüller, 1775) [version 1; peer review: awaiting peer review]

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

[1]UK Centre for Ecology & Hydrology, Wallingford, Oxfordshire, UK
[2]Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

[+] Deceased author

v1 **First published:** 06 Feb 2023, **8**:58
https://doi.org/10.12688/wellcomeopenres.18948.1
Latest published: 06 Feb 2023, **8**:58
https://doi.org/10.12688/wellcomeopenres.18948.1

Abstract

We present a genome assembly from an individual female *Philereme vetulata* (the Brown Scallop; Arthropoda; Insecta; Lepidoptera; Geometridae). The genome sequence is 771 megabases in span. Most of the assembly is scaffolded into 68 chromosomal pseudomolecules, including the assembled Z sex chromosome. The mitochondrial genome has also been assembled and is 16.3 kilobases in length. Gene annotation of this assembly on Ensembl has identified 18,096 protein coding genes.

Keywords

Phillereme vetulata, Brown Scallop, 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; **Tyler-Smith C**: Writing – Original Draft Preparation;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (206194, 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, Darwin Tree of Life Barcoding collective *et al.* **The genome sequence of the Brown Scallop, *Philereme vetulata* (Denis and Schiffermüller, 1775) [version 1; peer review: awaiting peer review]** Wellcome Open Research 2023, **8**:58 https://doi.org/10.12688/wellcomeopenres.18948.1

First published: 06 Feb 2023, **8**:58 https://doi.org/10.12688/wellcomeopenres.18948.1

Species taxonomy

Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Lepidoptera; Glossata; Ditrysia; Geometroidea; Geometridae; Larentiinae; *Philereme*; *Philereme vetulata* (NCBI:txid934888).

Background

The Brown Scallop (*Philereme vetulata*) is a geometrid moth of medium size (adult wingspan 13–16 mm). As implied by the English name, the predominant appearance of the wings of both sexes is brown with numerous faint crosslines and a scalloped edge to the hindwings (Lewis, 2019; Waring *et al.*, 2017).

In Britain, the adults fly at night in June and July (Randle *et al.*, 2019); the winter is spent as an egg, and the larvae develop in May and June, feeding at night exclusively on buckthorn (*Rhamnus cathartica*) before pupating in a cocoon in the ground (Henwood *et al.*, 2020). The flight period in 2000–2016 was about two weeks earlier in the year than in the 1970s (Randle *et al.*, 2019), probably because of the warming climate.

The moth has a broad palaearctic distribution (NBN Atlas, 2021) including the British Isles, where it is widespread but local in southern England, and present but very local in south Wales and Ireland. The distribution has declined since 1990 (Randle *et al.*, 2019), but its status is still classified as 'Least Concern' (Fox *et al.*, 2019).

Here we present a chromosomally complete genome sequence for *Philereme vetulata*, based on a single specimen from Wytham Woods, Oxfordshire. The genome sequence will provide a starting point for understanding its geographical variation and genetic history, and the molecular basis of phenotypes such as larval tolerance of the generally toxic buckthorn foodplant.

Genome sequence report

The genome was sequenced from one female *Philereme vetulata* (Figure 1) collected from Wytham Woods, UK (latitude 51.77, longitude –1.34). A total of 29-fold coverage in Pacific Biosciences single-molecule HiFi long reads and 38-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 94 missing joins or mis-joins and removed three haplotypic duplications, reducing the scaffold number by 34.17%, and increasing the scaffold N50 by 25.06%.

The final assembly has a total length of 771.1 Mb in 29 sequence scaffolds with a scaffold N50 of 12.9 Mb (Table 1). Most (99.93%) of the assembly sequence was assigned to 68 chromosomal-level scaffolds, representing 67 autosomes. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 2–Figure 5, Table 2). The specimen has half-coverage of Hi-C reads mapped across the Z-chromosome, and no scaffolds for a W chromosome, so this specimen is a female sample with ZO karyotype. The assembly has a BUSCO v5.3.2 (Manni *et al.*, 2021) completeness of 98.4% (single 97.5%, duplicated 0.8%) using the lepidoptera_odb10 reference set (n = 5,286). 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 *P. vetulata* genome assembly GCA_918857605.1 was annotated using the Ensembl rapid annotation pipeline (Table 1). The resulting annotation includes 18,269 transcribed mRNAs from 18,096 protein-coding genes.

Methods

Sample acquisition and nucleic acid extraction

A *Philereme vetulata* specimen (ilPhiVetu1) was collected in Wytham Woods, Oxfordshire (biological vice-county: Berkshire), UK (latitude 51.77, longitude –1.34) on 25 June 2020 using a light trap. Douglas Boyes (University of Oxford) collected and identified the specimen. The specimen was snap-frozen on dry ice.

DNA was extracted at the Tree of Life laboratory, Wellcome Sanger Institute. The ilPhiVetu1 sample was weighed and dissected on dry ice with tissue set aside for Hi-C sequencing. Abdomen tissue was disrupted using a Nippi Powermasher fitted with a BioMasher pestle. High molecular weight (HMW) DNA was extracted using the Qiagen MagAttract HMW DNA extraction kit. Low molecular weight DNA was removed from a 20 ng aliquot of extracted DNA using 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 NovaSeq 6000 (10X) instruments. Hi-C data were also generated from head and thorax tissue of ilPhiVetu1 using the Arima v2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

Table 1. Genome data for *Philereme vetulata*, ilPhiVetu1.2.

Project accession data		
Assembly identifier	ilPhiVetu1.2	
Species	*Philereme vetulata*	
Specimen	ilPhiVetu1	
NCBI taxonomy ID	934888	
BioProject	PRJEB46330	
BioSample ID	SAMEA7701300	
Isolate information	ilPhiVetu1: abdomen (PacBio and 10X), head and thorax (Hi-C)	
Assembly metrics*		***Benchmark***
Consensus quality (QV)	55.2	*≥ 50*
k-mer completeness	99.99%	*≥ 95%*
BUSCO**	C:98.4%[S:97.5%,D:0.8%], F:0.5%,M:1.2%,n:5,286	*C ≥ 95%*
Percentage of assembly mapped to chromosomes	99.93%	*≥ 95%*
Sex chromosomes	Z chromosome	*localised homologous pairs*
Organelles	Mitochondrial genome assembled	*complete single alleles*
Raw data accessions		
PacificBiosciences SEQUEL II	ERR6808012	
10X Genomics Illumina	ERR6688580–ERR6688583	
Hi-C Illumina	ERR6688584	
Genome assembly		
Assembly accession	GCA_918857605.2	
Accession of alternate haplotype	GCA_918811585.1	
Span (Mb)	771.1	
Number of contigs	214	
Contig N50 length (Mb)	6.3	
Number of scaffolds	79	
Scaffold N50 length (Mb)	12.9	
Longest scaffold (Mb)	27.7	
Genome annotation		
Number of protein-coding genes	18,096	
Number of gene transcripts	18,269	

* 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/ilPhiVetu1.1/dataset/CAKKOV01.1/busco.

Figure 1. Photograph of the *Philereme vetulata* (ilPhiVetu1) specimen used for genome sequencing.

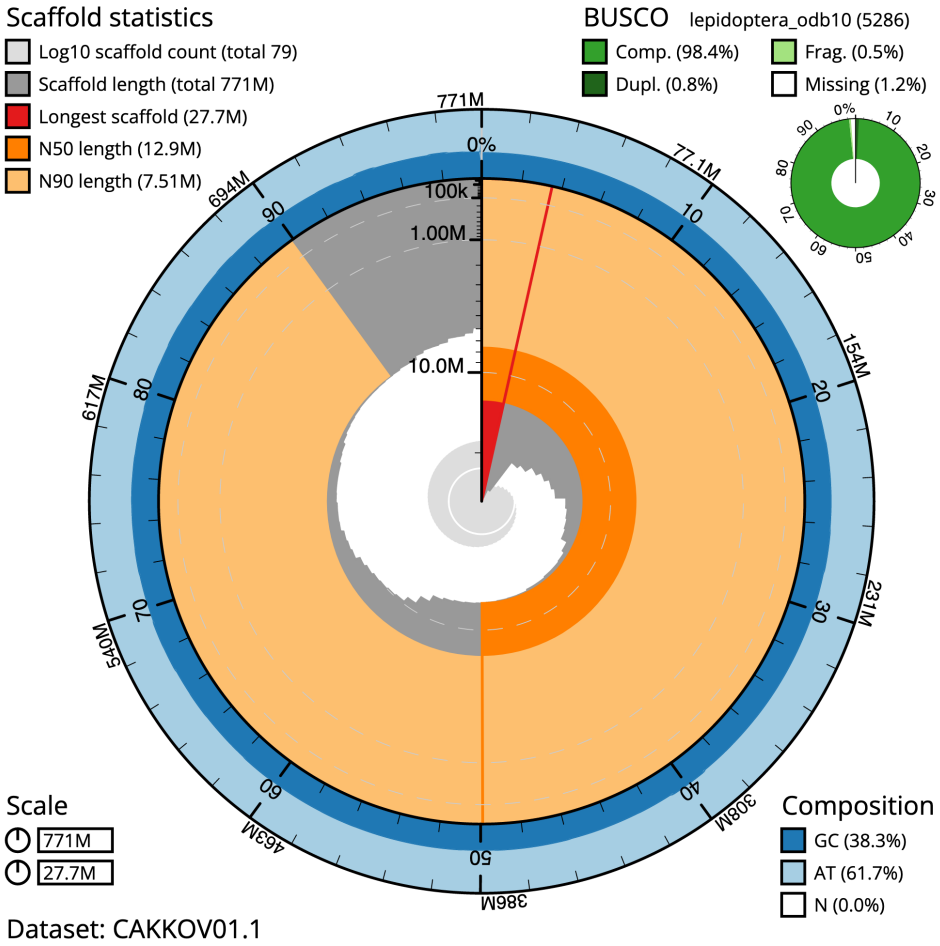

Figure 2. Genome assembly of *Philereme vetulata*, ilPhiVetu1.2: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 771,080,484 bp assembly. The distribution of sequence lengths is shown in dark grey with the plot radius scaled to the longest sequence present in the assembly (27,679,861 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 sequence lengths (12,931,513 and 7,511,951 bp), respectively. The pale grey spiral shows the cumulative sequence 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/ilPhiVetu1.1/dataset/CAKKOV01.1/snail.

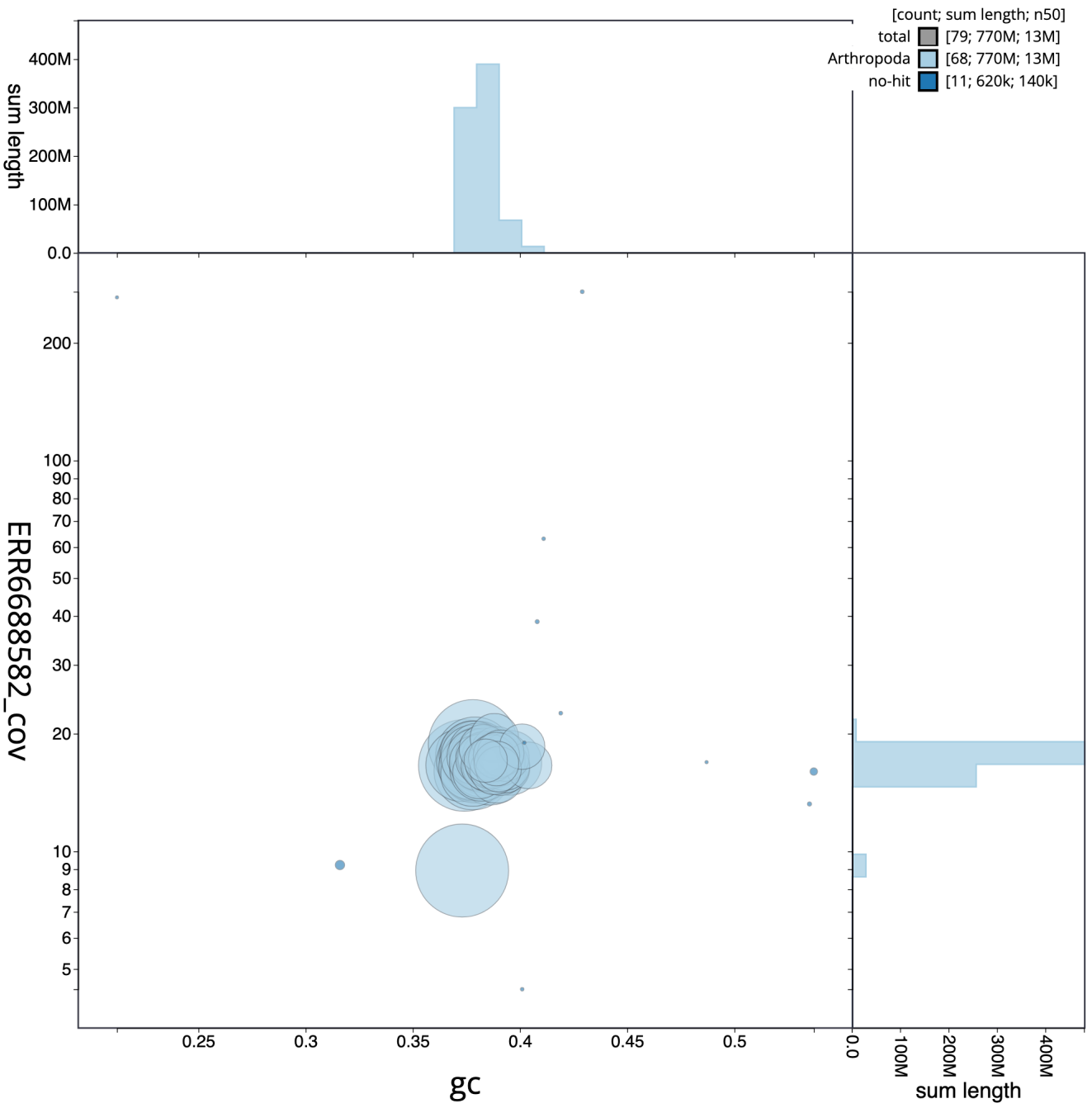

Figure 3. Genome assembly of *Philereme vetulata*, ilPhiVetu1.2: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilPhiVetu1.1/dataset/CAKKOV01.1/blob.

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). 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 were generated within the BlobToolKit environment (Challis *et al.*, 2020). Table 3 contains a list of all software tool versions used, where appropriate.

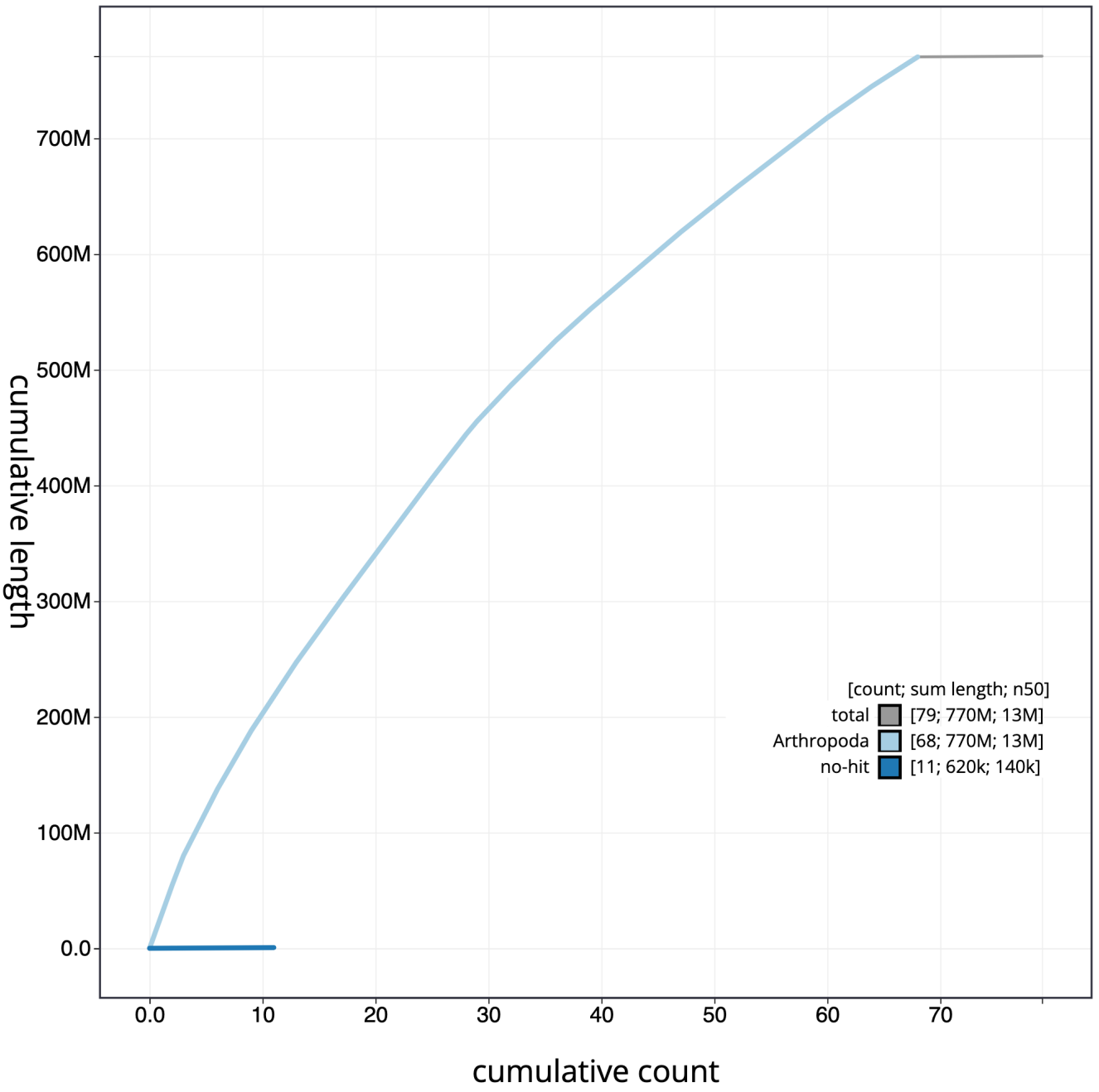

Figure 4. Genome assembly of *Philereme vetulata*, ilPhiVetu1.2: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/ilPhiVetu1.1/dataset/CAKKOV01.1/cumulative.

Genome annotation

The BRAKER2 pipeline (Brůna *et al.*, 2021) was used in the default protein mode to generate annotation for the *Philereme vetulata* assembly (GCA_918857605.1) in Ensembl Rapid Release.

Ethics/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. Each transfer of samples is further undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by

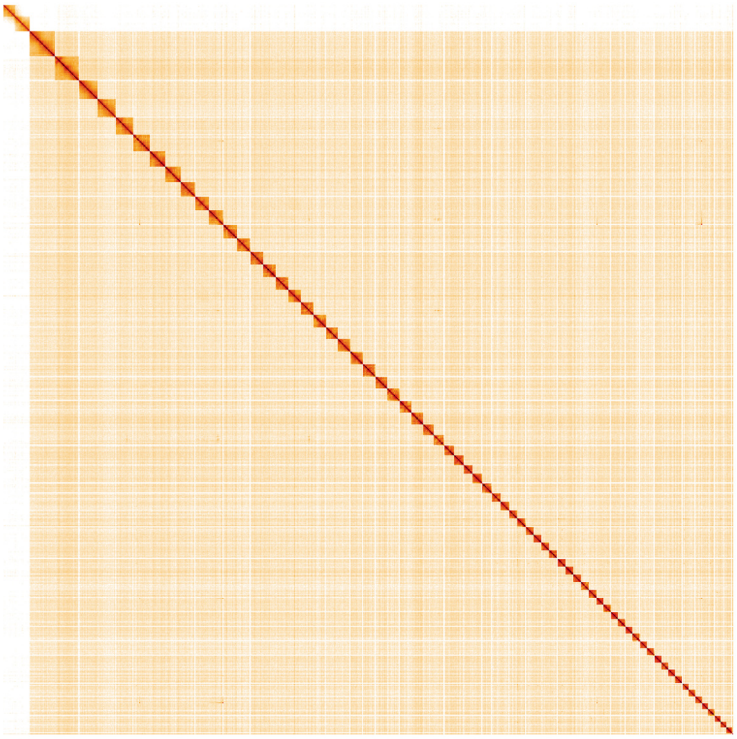

Figure 5. Genome assembly of *Philereme vetulata*, ilPhiVetu1.2: Hi-C contact map. Hi-C contact map of the ilPhiVetu1.2 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=PYCZAfJKRkGZCHDW5dydkw.

Table 2. Chromosomal pseudomolecules in the genome assembly of*Philereme vetulata*, ilPhiVetu1.

INSDC accession	Chromosome	Size (Mb)	GC%
OU975418.1	1	26.9	37.4
OU975419.1	2	25.42	37.8
OU975420.1	3	19.87	37.9
OU975421.1	4	19.28	37.8
OU975422.1	5	18.22	38
OU975423.1	6	17.48	38.3
OU975424.1	7	16.82	37.3
OU975425.1	8	16.42	37.9
OU975426.1	9	15.09	37.9
OU975427.1	10	14.87	38.6
OU975428.1	11	14.73	37.7
OU975429.1	12	14.58	37.8
OU975430.1	13	13.81	37.8
OU975431.1	14	13.49	37.9
OU975432.1	15	13.4	37.8
OU975433.1	16	13.4	37.9
OU975434.1	17	13.28	39.5
OU975435.1	18	13.27	38
OU975436.1	19	13.11	37.7
OU975437.1	20	13.1	37.8
OU975438.1	21	13.07	38.3
OU975439.1	22	13.01	38.3
OU975440.1	23	12.93	37.8
OU975441.1	24	12.88	38.2
OU975442.1	25	12.8	38.2
OU975443.1	26	12.67	38.2
OU975444.1	27	12.43	38.4
OU975445.1	28	11.65	38.2
OU975446.1	29	10.75	38.2
OU975447.1	30	10.2	38.7
OU975448.1	31	10.07	38
OU975449.1	32	9.95	37.9

INSDC accession	Chromosome	Size (Mb)	GC%
OU975450.1	33	9.85	38.3
OU975451.1	34	9.67	38
OU975452.1	35	9.44	38.9
OU975453.1	36	9	39.2
OU975454.1	37	8.86	38.1
OU975455.1	38	8.75	38.3
OU975456.1	39	8.47	38.9
OU975457.1	40	8.43	38.5
OU975458.1	41	8.42	38.6
OU975459.1	42	8.32	38.9
OU975460.1	43	8.29	38.7
OU975461.1	44	8.29	38.3
OU975462.1	45	8.29	38.2
OU975463.1	46	8.13	39.1
OU975464.1	47	8.07	38.9
OU975465.1	48	7.86	38.5
OU975466.1	49	7.75	39.3
OU975467.1	50	7.69	38.9
OU975468.1	51	7.64	38.4
OU975469.1	52	7.59	38.6
OU975470.1	53	7.59	39.3
OU975471.1	54	7.58	38.7
OU975472.1	55	7.56	39.1
OU975473.1	56	7.51	38.3
OU975474.1	57	7.47	38.8
OU975475.1	58	7.37	38.8
OU975476.1	59	7.35	39.3
OU975477.1	60	7.03	38.7
OU975478.1	61	6.85	40.4
OU975479.1	62	6.84	39.1
OU975480.1	63	6.76	38.9
OU975481.1	64	6.47	39
OU975482.1	65	6.46	40.1
OU975483.1	66	6.27	38.9
OU975484.1	67	5.89	38.4
OU975417.1	Z	27.68	37.3
OU975485.2	MT	0.02	21.3
-	-	0.6	41.3

Table 3. Software tools and versions used.

Software tool	Version	Source
BlobToolKit	3.4.0	Challis *et al.*, 2020
freebayes	1.3.1-17-gaa2ace8	Garrison & Marth, 2012
Hifiasm	0.15.3	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

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: *Philereme vetulata* (brown scallop). Accession number PRJEB46330; https://identifiers.org/ena.embl/PRJEB46330. (Wellcome Sanger, 2022)

The genome sequence is released openly for reuse. The *Philereme vetulata* 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

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

Brůna T, Hoff KJ, Lomsadze A, *et al.*: **BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database.** *NAR Genom Bioinform.* 2021; **3**(1): lqaa108.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

Fox R, Parsons MS, Harrower CA: **A review of the status of the macro-moths of Great Britain.** *Doreset,UK: Butterly Conservation.* 2019.
Reference Source

Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing.** *arXiv.* 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 REad TEXTure Viewer): A desktop application for viewing pretext contact maps**. 2022; (Accessed: 19 October 2022).
Reference Source

Henwood B, Sterling P, Lewington R: **Field Guide to the Caterpillars of Great Britain and Ireland.** London: Bloomsbury. 2020.
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

Lewis C: **70.118 *Philereme vetulata* (Brown Scallop).** British Lepidoptera. 2019; (Accessed: 17 January 2023).
Reference Source

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

NBN Atlas: ***Philereme vetulata* ([Denis & Schiffermuller], 1775) Brown Scallop.** National Biodiversity Atlas, 2021; (Accessed: 17 January 2023).
Reference Source

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

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

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

Wellcome Sanger Institute: **The genome sequence of the Brown Scallop, *Philereme vetulata* (Denis & Schiffermuller, 1775)**. European Nucleotide Archive. [dataset], accession number PRJEB46330, 2022.