DATA NOTE

The genome sequence of the freshwater leech, *Erpobdella octoculata* (Linnaeus, 1758) (Hirudinida: Erpobdellidae)

[version 1; peer review: 2 approved, 1 approved with reservations]

Susan J. Skipp[1], Andrew Morrise[2],
Natural History Museum Genome Acquisition Lab,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]Environment Agency, West Malling, Kent, England, UK
[2]Environment Agency, Warrington, Cheshire, England, UK

v1 **First published:** 24 Nov 2025, **10**:660
https://doi.org/10.12688/wellcomeopenres.25215.1
Latest published: 24 Nov 2025, **10**:660
https://doi.org/10.12688/wellcomeopenres.25215.1

Abstract

We present a genome assembly from an individual *Erpobdella octoculata* (leeches; Annelida; Clitellata; Hirudinida; Erpobdellidae). The genome sequence has a total length of 386.12 megabases. Most of the assembly (98.81%) is scaffolded into 10 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 19.12 kilobases. Gene annotation of this assembly on Ensembl identified 19 465 protein-coding genes. 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

Erpobdella octoculata; leeches; genome sequence; chromosomal; Hirudinida

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ? ✓

	1	2	3
version 1 24 Nov 2025	✓ view	? view	✓ view

1. **Thomas D. Lewin**, University of Oxford, Oxford, UK
2. **Gonghua Lin**, Jinggangshan University, Jiangxi, China
3. **Jun Kim**, Chungnam National University, Daejeon, South Korea

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: Skipp SJ: Investigation, Resources, Writing – Review & Editing; **Morrise A**: Investigation, Resources;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (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.

How to cite this article: Skipp SJ, Morrise A, Natural History Museum Genome Acquisition Lab *et al.* **The genome sequence of the freshwater leech, *Erpobdella octoculata* (Linnaeus, 1758) (Hirudinida: Erpobdellidae) [version 1; peer review: 2 approved, 1 approved with reservations]** Wellcome Open Research 2025, **10**:660 https://doi.org/10.12688/wellcomeopenres.25215.1

First published: 24 Nov 2025, **10**:660 https://doi.org/10.12688/wellcomeopenres.25215.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Spiralia; Lophotrochozoa; Annelida; Clitellata; Hirudinea; Hirudinida; Erpobdelliformes; Erpobdellidae; *Erpobdella*; *Erpobdella octoculata* (Linnaeus, 1758) (NCBI:txid39305)

Background

Erpobdella octoculata (Linnaeus, 1758) (Figure 1) is a common and widespread freshwater leech in Europe. It is found throughout Britain and Ireland, but is relatively uncommon in northern Scotland. It occurs in many freshwater habitats and can be abundant in organically enriched streams (Murphy & Learner, 1982).

Field gut-content studies of *E. octoculata* show that its diet is dominated by chironomid larvae and oligochaetes (Beracko & Rogánska, 2014). It is a macrophagous predator that often swallows invertebrate prey whole, especially chironomid larvae. It also sucks fluids from wounded or dead invertebrates and will scavenge vertebrate carrion. Newly hatched juveniles feed by sucking fluids from wounded prey (Kutschera, 2003).

Life history of *E. octoculata* varies along stream gradients. Upstream populations are usually biennial and iteroparous; downstream they are annual and semelparous with faster growth and higher reproductive effort (Beracko & Rogánska, 2014).

We report a chromosomally complete genome sequence for *Erpobdella octoculata*, produced as part of the Darwin Tree of Life project. This is the first whole genome sequence for the family Erpobdellidae as of October 2025 (data obtained via NCBI datasets, O'Leary *et al.*, 2024).

Methods

Sample acquisition

The specimen used for genome sequencing was an adult *Erpobdella octoculata* (specimen ID NHMUK014361447, ToLID wkErpOcto2), collected from River Weaver Hunts Lock, UK (latitude 53.251, longitude –2.5189) on 2019-03-18. A second specimen was used for Hi-C sequencing (specimen ID NHMUK014360727, ToLID wkErpOcto1), collected from Shropshire Union Canal, Bagley Lane, UK (latitude 52.9754, longitude –2.5097) on 2019-03-18. Sample metadata were collected in line with the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

Figure 1. Photograph of *Erpobdella octoculata* by WWalas (not the specimen used for genome sequencing).

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 wkErpOcto2 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the whole organism was homogenised by powermashing using a PowerMasher II tissue disruptor. HMW DNA was extracted 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 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 15.0 ng/μL and a yield of 6 000.00 ng.

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

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen head and terminal body tissue of the wkErpOcto1 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). Dependingon 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 were created. 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 HiSeq X Ten.

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 (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). The Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019), and the contigs were scaffolded in YaHS (Zhou *et al.*, 2023) with 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 57 breaks, 54 joins, and removal of 20 haplotypic duplications. This reduced the scaffold count by 13.6%, reduced the scaffold N50 by 1.6%, and reduced the total assembly length by 2.6%. 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 the primary and alternate haplotypes using the *k*-mer databases (k = 31) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

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 *Erpobdella octoculata* specimen generated 22.86 Gb (gigabases) from 1.86 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 370.71 Mb, with a heterozygosity of 1.10% and repeat content of 39.78% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 58× coverage. Hi-C sequencing produced 82.44 Gb from 545.93 million reads, which were used to scaffold the assembly. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC

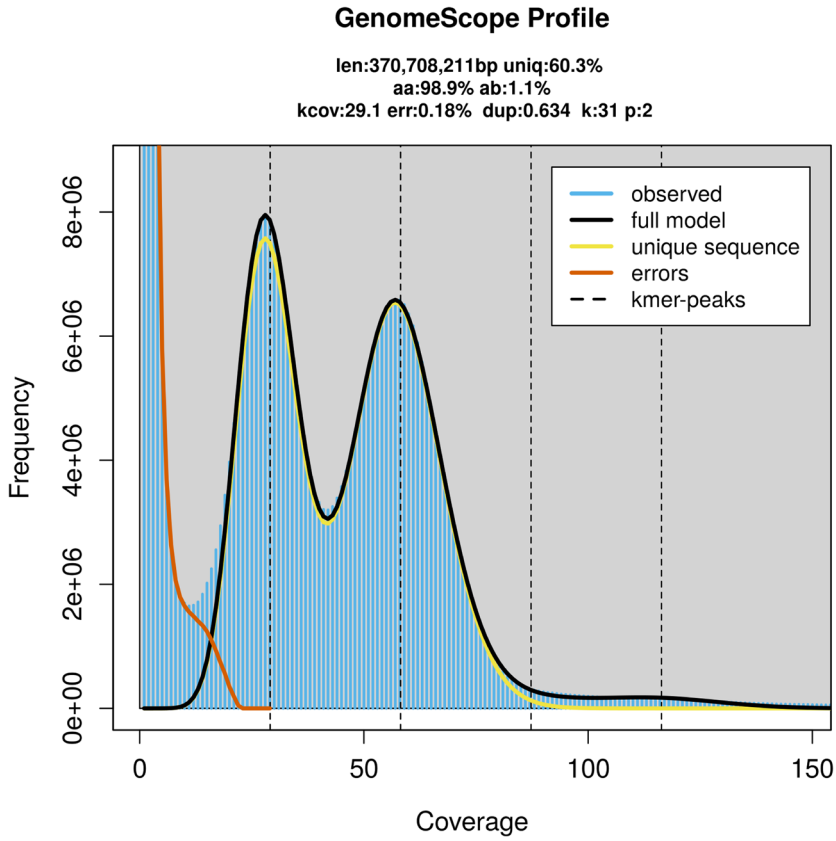

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB73391.

Platform	PacBio HiFi	Hi-C
ToLID	wkErpOcto2	wkErpOcto1
Specimen ID	NHMUK014361447	NHMUK014360727
BioSample (source individual)	SAMEA7520983	SAMEA7520811
BioSample (tissue)	SAMEA7521083	SAMEA7520867
Tissue	whole organism	head | terminal body
Instrument	Sequel IIe	HiSeq X Ten
Run accessions	ERR12721044	ERR12723455
Read count total	1.86 million	545.93 million
Base count total	22.86 Gb	82.44 Gb

databases. The final assembly has a total length of 386.12 Mb in 101 scaffolds, with 604 gaps, and a scaffold N50 of 39.24 Mb (Table 2).

Most of the assembly sequence (98.81%) was assigned to 10 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). There are a number of heterozygous inversions at the following locations: Chromosome 1 ~50.5–52 Mbp, Chromosome 2 ~23.13–24.33 Mbp, Chromosome 2 ~32.6–33.9 Mbp, Chromosome 3 ~17.5–18.8 Mbp, Chromosome 4 ~33.64–36.02 Mbp, Chromosome 10 ~11.11–12.07.

Table 2. Genome assembly statistics.

Assembly name	wkErpOcto2.1
Assembly accession	GCA_963978895.1
Alternate haplotype accession	GCA_963978845.1
Assembly level	chromosome
Span (Mb)	386.12
Number of chromosomes	10
Number of contigs	705
Contig N50	1.14 Mb
Number of scaffolds	101
Scaffold N50	39.24 Mb
Organelles	Mitochondrion: 19.12 kb

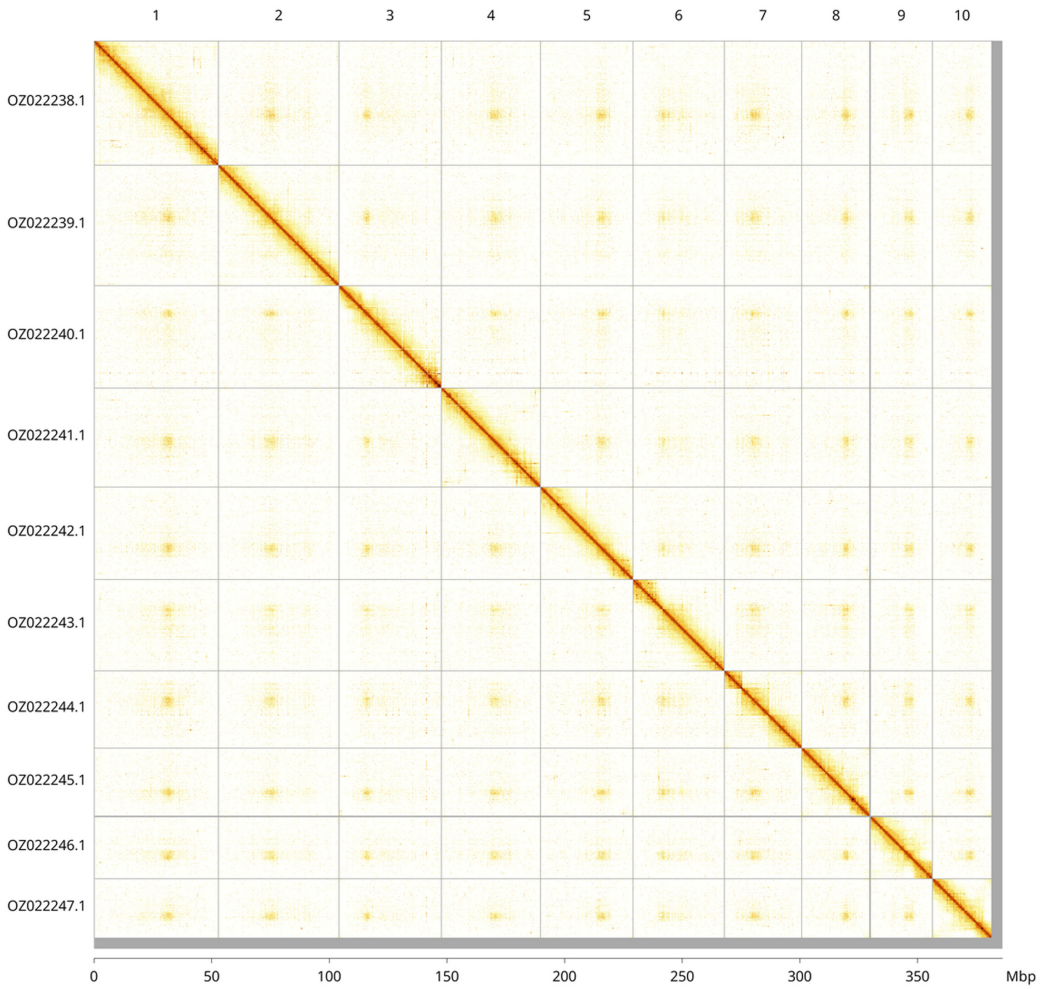

Figure 3. Hi-C contact map of the *Erpobdella octoculata* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

The mitochondrial genome was also assembled (length 19.12 kb, OZ022248.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

The combined primary and alternate assemblies achieve an estimated QV of 51.6. The *k*-mer completeness is 79.25% for the primary assembly, 78.38% for the alternate haplotype, and 98.24% for the combined assemblies (Figure 4).

BUSCO v.5.5.0 analysis using the metazoa_odb10 reference set (n = 954) identified 86.7% of the expected gene set (single = 81.6%, duplicated = 5.1%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for the primary assembly. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage.

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, calculated for the primary assembly, is **6.C.Q51**, meeting the recommended reference standard.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Erpobdella octoculata* wkErpOcto2.

INSDC accession	Molecule	Length (Mb)	GC%
OZ022238.1	1	52.90	34
OZ022239.1	2	51.22	34.50
OZ022240.1	3	43.52	34
OZ022241.1	4	42.20	34.50
OZ022242.1	5	39.24	34
OZ022243.1	6	38.81	34
OZ022244.1	7	32.92	34.50
OZ022245.1	8	29.14	34.50
OZ022246.1	9	26.46	34
OZ022247.1	10	25.13	34.50

Genome annotation report

The *Erpobdella octoculata* genome assembly (GCA_963978895.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 30 689 transcribed mRNAs from 19 465 protein-coding and 4 219 non-coding genes. The average transcript length is 5 597.31 bp, with an average of 1.30 coding transcripts per gene and 5.50 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

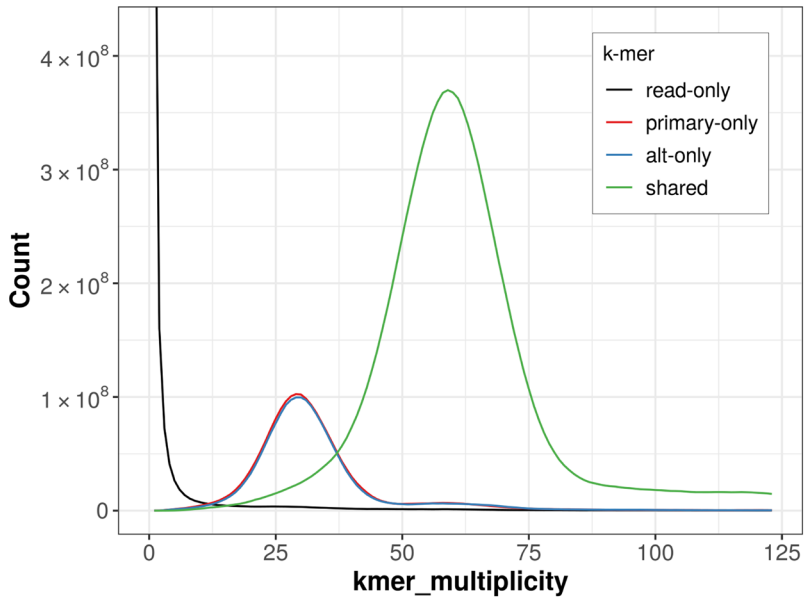

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

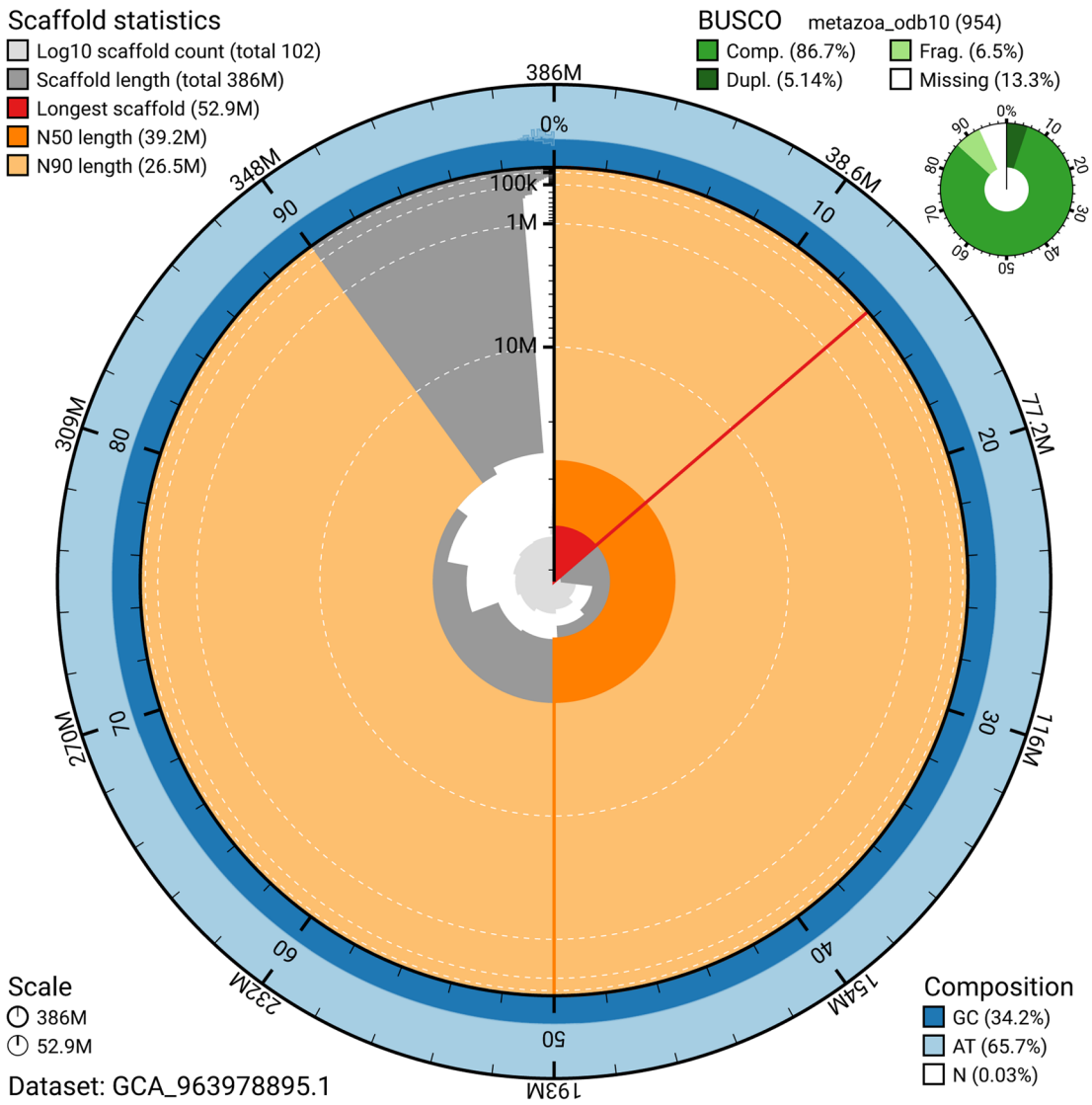

Figure 5. Assembly metrics for wkErpOcto2.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue 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 metazoa_odb10 set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

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 diligenceis 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

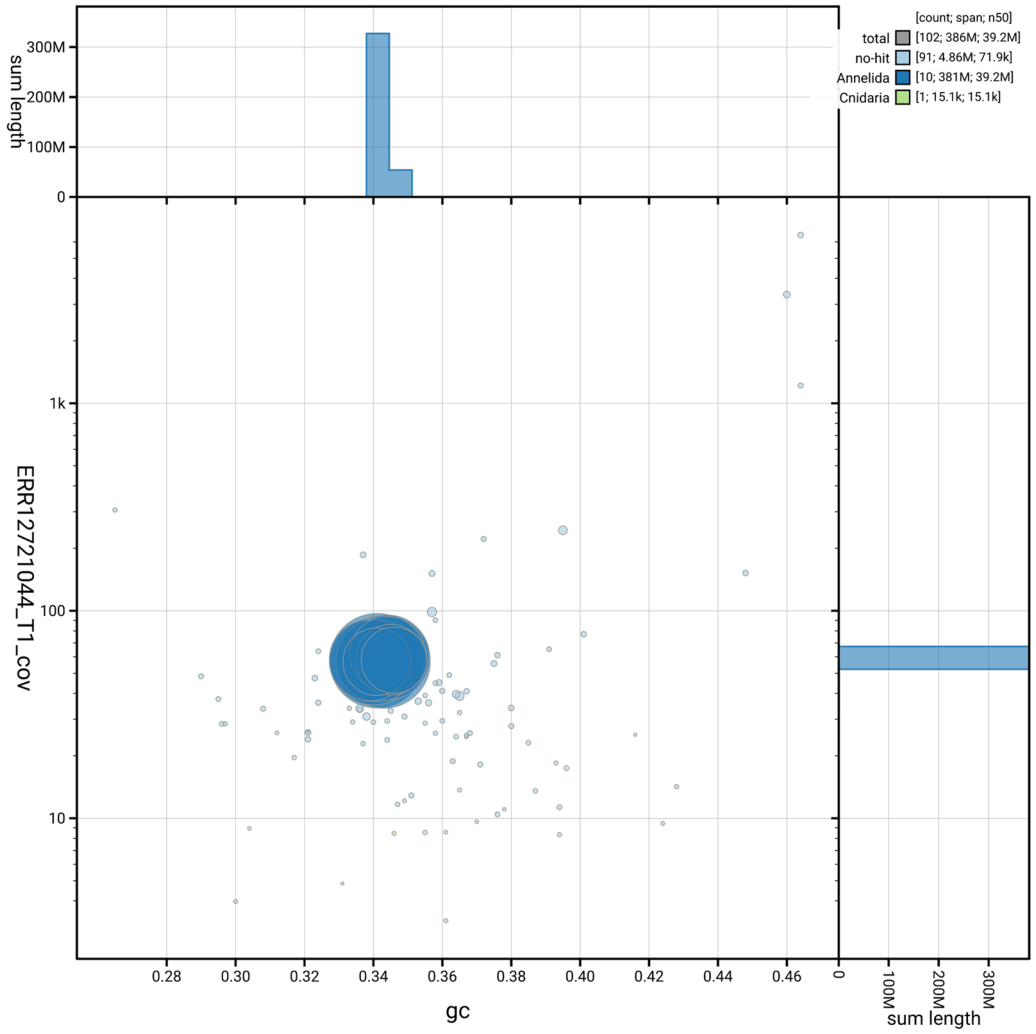

Figure 6. BlobToolKit GC-coverage plot for wkErpOcto2.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display 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 *Erpobdella octoculata* assembly.

Measure	Value	Benchmark
EBP summary (primary)	6.C.Q51	6.C.Q40
Contig N50 length	1.14 Mb	≥ 1 Mb
Scaffold N50 length	39.24 Mb	= chromosome N50
Consensus quality (QV)	Primary: 51.6; alternate: 53.0; combined: 51.6	≥ 40
k-mer completeness	Primary: 79.25%; alternate: 78.38%; combined: 98.24%	≥ 95%
BUSCO	C:86.7% [S:81.6%; D:5.1%]; F:6.5%; M:6.8%; n:954	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	98.81%	≥ 90%

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: Erpobdella octoculata (leech). Accession number PRJEB73391. The genome sequence is released openly for reuse. The *Erpobdella octoculata* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665) and the Sanger Institute Tree of Life Programme (PRJEB43745). 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 Natural History Museum Genome Acquisition Lab
- 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	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
Hifiasm	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.24-r1122	https://github.com/lh3/minimap2
MitoHiFi	2	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC

Software	Version	Source
Nextflow	23.04.1	https://github.com/nextflow-io/nextflow
PretextSnapshot	0.0.5	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.4.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.1a.2	https://github.com/c-zhou/yahs

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

Beracko P, Rogánska A: **Intra- and interspecific variations in life strategies of *Erpobdella octoculata* and *Erpobdella vilnensis* in different habitats along the longitudinal gradient of stream.** *Limnologica.* 2014; **48**: 28–38.
Publisher 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

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

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

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): 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

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

Kutschera U: **The feeding strategies of the leech *Erpobdella octoculata* (L.): a laboratory study.** *Int Rev Hydrobiol.* 2003; **88**(1): 94–101.
Publisher Full Text

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

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

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

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

Murphy PM, Learner MA: **The life history and production of the leech *Erpobdella octoculata* (Hirudinea Erpobdellidae) in the River Ely, South Wales.** *J Anim Ecol.* 1982; **51**(1): 57–67.
Publisher Full Text

O'Leary NA, Cox E, Holmes JB, *et al.*: **Exploring and retrieving sequence and metadata for species across the Tree of Life with NCBI datasets.** *Sci Data.* 2024; **11**(1): 732.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free**

genome assemblies of all vertebrate species. *Nature.* 2021; **592**(7856): 737–746.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

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

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**

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

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

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

Open Peer Review

Current Peer Review Status: ✔ ? ✔

Version 1

Reviewer Report 06 February 2026

https://doi.org/10.21956/wellcomeopenres.27791.r144171

✔ **Jun Kim**
Chungnam National University, Daejeon, Daejeon, South Korea

As this work was conducted by the Darwin Tree of Life Consortium, the dataset appears to be of high quality and carefully curated. The analytical workflow follows well-established standardized procedures, and potential errors seem to have been appropriately addressed.

The only point on which I would appreciate further clarification concerns the presence of telomeric repeats. If the assemblies indeed represent complete chromosomes, one expects clusters of telomeric repeats (typically, TTAGGG or its variants) to be present at both ends of each scaffold. While it is possible that such signals may be absent due to limitations in assembly quality, it would be helpful if the authors could clarify whether telomeric repeat clusters are observed at the ends of the current chromosome-level scaffolds. Specifically, if present, it would be better to clarify how many such clusters were identified. Given a chromosome number of n = 10 in this manuscript, one might expect approximately 20 telomeric repeat clusters at the ends; it would be informative to know whether a comparable number was observed. In addition, it would be helpful to clarify whether any telomeric repeat clusters were detected internally within scaffolds (excluding interstitial telomeric sequences and considering only those located at contig or scaffold termini). This is because this internal location of telomeric repeat clusters could be a signature of scaffolding errors.

Finally, if available, I would appreciate it if the authors could indicate whether independent experimental validation has been performed, such as karyotype analyses using DAPI staining or related approaches, and whether those results are consistent with the current inference of n = 10 chromosomes.

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: Genomics, genome assembly, long-read sequencing, DNA damage and variant formation

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 13 January 2026

https://doi.org/10.21956/wellcomeopenres.27791.r144170

?

Gonghua Lin
Jinggangshan University, Jiangxi, China

This study presents a chromosome-level genome of Erpobdella octoculata along with preliminary structural annotations. In recent years, many leech genomes have been elucidated, but the focus has largely been on the Hirudiniformes suborder. The genome of Erpobdella octoculata, belonging to the Erpobdelliformes suborder, contributes to understanding the evolutionary characteristics of the entire class Hirudinea and merits indexing. The authors are recommended to improve the manuscript in the following aspects:
(1) In the Background section, the authors provide only a brief description of the research subject, which diminishes the appeal of the paper. Given the substantial number of leech genomes reported in recent years, the authors could cite relevant studies to argue for the importance of leech genome research.
(2) For the utilization of the genome, structural annotation information is as crucial as the genome sequence itself. While the authors provide detailed descriptions of genome sequencing, assembly, quality assessment, and sequence information, the section on genome annotation is overly brief, affecting the academic and potential application value of the paper. It is recommended that the authors enhance the annotation-related content. For example, the Methods section should elaborate on how Ensembl automatic annotation was performed, including software and versions used, parameter settings, etc. The Results section should include annotation outcomes, such as length distribution, N50, and GC content of CDS. Additionally, functional annotation of these CDS could be considered, such as functional predictions by aligning with databases like NR, TrEMBL, and KEGG.
(3) In the Genome annotation report section, the authors mention “This annotation includes 30,689 transcribed mRNAs” and “The average transcript length is 5,597.31 bp, ...”. These refer to

RNA-Seq data generated during the annotation process rather than annotation results.
(4) In the Data Availability section, the authors state, "The genome will be annotated using available RNA-Seq data ...". Is this statement necessary here, or might it require rephrasing?

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?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Functional Genomics

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, however I have significant reservations, as outlined above.

Reviewer Report 17 December 2025

https://doi.org/10.21956/wellcomeopenres.27791.r140991

Thomas D. Lewin
Department of Biology, University of Oxford, Oxford, England, UK

Skipp et al. 2025 R1

This Data Note describes the sequencing and assembly of a genome for the freshwater leech *Erpobdella octoculata*. As a standard Darwin Tree of Life report, the methods are detailed and well-described. The genome itself is of high quality: chromosome-level, with two haplotypes deposited, and good scores for the quality metrics used. Raw and assembled data is all publicly available. I have no concerns about the data or the manuscript and have only two small suggestions that the authors may wish to consider.

First, while genome annotations by Ensembl are reported in the results, the annotation process is not described in the methods. I think at least a brief description of this process should be included

in the manuscript so readers have a clear idea how this was performed and do not have to go fishing through Ensembl websites to find the information.

Second, I think some more information is needed in the introduction about why this genome specifically could be useful. Indeed, it is a very exciting addition for a number of reasons. For instance, leech genomes are currently of interest due to a potential whole genome duplication in the lineage (refer to 1), something that is relatively rare in invertebrates. They also made headlines last year when three papers simultaneously reported the complete loss of macrosynteny in leeches and earthworms (refer to 2,3,4) ; and see summary in (refer to 5). Leech genomes have also been studied to elucidate adaptations to a blood-sucking lifestyle (refer to 6). Therefore, this genome may play a key role in the study of whole genome duplication, genome rearrangements and ecological adaptations to unusual lifestyles like hematophagy.

References

1. Vargas-Chávez et al. 2025 Nat. Ecol. Evol. https://www.nature.com/articles/s41559-025-02728-1.
2. Vargas-Chávez et al. 2025 Nat. Ecol. Evol. https://www.nature.com/articles/s41559-025-02728-1.
3. Schultz et al. 2024 BioRxiv https://www.biorxiv.org/content/10.1101/2024.05.12.593736v1.full.
4. Lewin et al. 2024 Mol. Biol. Evol. https://academic.oup.com/mbe/article/41/9/msae172/7733614.
5. Vargas-Chávez et al. 2025 Trends. Genet.
https://www.sciencedirect.com/science/article/pii/S0168952525001374.
6. Zheng et al. 2023 Gigascience.
https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad023/7113332.

Is the rationale for creating the dataset(s) clearly described?

Partly

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: Evolutionary Genomics

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.