DATA NOTE

REVISED The genome sequence of a springtail, *Orchesella flavescens* (C.Bourlet, 1839)

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

James McCulloch[1,2], Liam M. Crowley[1],
University of Oxford and Wytham Woods 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

[1]University of Oxford, Oxford, England, UK
[2]Wellcome Sanger Institute, Hinxton, England, UK

v2 **First published:** 18 Mar 2025, **10**:138
https://doi.org/10.12688/wellcomeopenres.23773.1
Latest published: 26 Sep 2025, **10**:138
https://doi.org/10.12688/wellcomeopenres.23773.2

Abstract

We present a genome assembly from a specimen of *Orchesella flavescens* (springtail; Arthropoda; Collembola; Entomobryomorpha; Orchesellidae). The genome sequence has a total length of 273.69 megabases. Most of the assembly (99.88%) is scaffolded into 6 chromosomal pseudomolecules. The mitochondrial genome has also been assembled and is 14.92 kilobases in length.

Keywords

Orchesella flavescens, springtail, genome sequence, chromosomal, Entomobryomorpha

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ? ? ✓

	1	2	3
version 2 (revision) 26 Sep 2025			✓ view
version 1 18 Mar 2025	? view	? view	? view

1. **Daoyuan Yu**, Nanjing Agricultural University, Nanjing, China
Zhihong Zhan, Nanjing Agricultural University College of Plant Protection (Ringgold ID: 214175), Nanjing, China
2. **Nerivânia Nunes Godeiro**, Shanghai Natural History Museum, Shanghai, China
3. **Michael Hiller**, Senckenberg Nature Research Society, Frankfurt Am Main, Germany

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: McCulloch J: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Crowley LM**: 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: McCulloch J, Crowley LM, University of Oxford and Wytham Woods Genome Acquisition Lab *et al.* **The genome sequence of a springtail, *Orchesella flavescens* (C.Bourlet, 1839) [version 2; peer review: 1 approved, 2 approved with reservations]** Wellcome Open Research 2025, **10**:138 https://doi.org/10.12688/wellcomeopenres.23773.2

First published: 18 Mar 2025, **10**:138 https://doi.org/10.12688/wellcomeopenres.23773.1

REVISED **Amendments from Version 1**

In Version 2 of this data note we have made corrections to the Background section in line with reviewers' comments, including correcting the family from Entomobryidae to Orchesellidae. We have changed the static view of the chromosome map in Figure 5 to a labelled map, while the interactive HiGlass version is still available in the link.

Any further responses from the reviewers can be found at the end of the article

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Collembola; Entomobryomorpha; Entomobryoidea; Orchesellidae; Orchesellinae; *Orchesella*; *Orchesella flavescens* (C.Bourlet, 1839) (NCBI:txid48711)

Background

Orchesella flavescens is a slender springtail (Collembola: Entomobryomorpha) in the family Orchesellidae (formerly treated as Entomobryidae: Orchesellinae). It is one of four species in its genus with confirmed records in the UK, the others being the very common *O. cincta* and *O. villosa* and the more localised, montane *O. alticola* (Hopkin, 2007). *Orchesella* are unique among British slender springtails in having six antennal segments, and they can also be recognised by their large size. *O. flavescens* is distinguished by the thin longitudinal bands of dark pigment running through each thoracic and abdominal segment. Reports of other longitudinally banded *Orchesella* in the UK (e.g. *O. quinquefasciata*, *O. spectabilis*) remain unconfirmed, and in both species the bands do not extend to the abdominal apex.

Orchesella flavescens is an uncommon species in the UK, with the number of records eclipsed by *O. cincta* and *O. villosa*. Indeed, it was not recorded for over 80 years between 1925 and 2009, when it then started to be recorded again in some southern woodlands, wherein it is usually found among moss or low vegetation. For example, Womersley (1924) – describing it as "the most magnificent spring-tail" – recorded it on multiple occasions in patches of dog's mercury (*Mercurialis perennis*) in a Somerset woodland.

Despite its relative scarcity, this is the first *Orchesella* species to have its genome assembled by the Darwin Tree of Life project. This genome will be useful for studies on speciation and population genomics given springtails' high rates of cryptic diversity (Timmermans *et al.*, 2005). Additionally, some springtail genomes have been found to exhibit unusual characteristics, including paternal genome elimination, accessory chromosomes, and exceptionally large chromosomes (Jaron *et al.*, 2022; Jin *et al.*, 2024). The *Orchesella* genome could harbour similarly interesting characteristics.

Genome sequence report

Sequencing data

The genome of a specimen of *Orchesella flavescens* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 24.24 Gb from 2.05 million reads. GenomeScope analysis of the PacBio HiFi data estimated the haploid genome size at 270.12 Mb, with a heterozygosity of 0.30% and repeat content of 21.38%. These values provide an initial assessment of genome complexity and the challenges anticipated during assembly. Based on this estimated genome size, the sequencing data provided approximately 84.0x coverage of the genome. Chromosome conformation Hi-C sequencing produced 98.56 Gb from 652.72 million reads. Table 1 summarises the specimen and sequencing information, including the BioProject, study name, BioSample numbers, and sequencing data for each technology.

Figure 1. Photograph of the *Orchesella flavescens* (qeOrcFlav1) specimen used for genome sequencing.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The assembly was improved by manual curation, which corrected 88 misjoins or missing joins and removed 19 haplotypic duplications. These interventions reduced the total assembly length by 0.71%, decreased the scaffold count by 70.83%, and increased the scaffold N50 by 1.32%. The final assembly has a total length of 273.69 Mb in 13 scaffolds, with 236 gaps, and a scaffold N50 of 51.76 Mb (Table 2).

The snail plot in Figure 2 provides a summary of the assembly statistics, indicating the distribution of scaffold lengths and other assembly metrics. Figure 3 shows the distribution of scaffolds by GC proportion and coverage. Figure 4 presents a cumulative assembly plot, with separate curves representing different scaffold subsets assigned to various phyla, illustrating the completeness of the assembly.

Most of the assembly sequence (99.89%) was assigned to 6 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3).

The mitochondrial genome was also assembled. This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record in GenBank.

Table 1. Specimen and sequencing data for *Orchesella flavescens*.

Project information			
Study title	Orchesella flavescens		
Umbrella BioProject	PRJEB68258		
Species	*Orchesella flavescens*		
BioSpecimen	SAMEA112232546		
NCBI taxonomy ID	48711		
Specimen information			
Technology	**ToLID**	**BioSample accession**	**Organism part**
PacBio long read sequencing	qeOrcFlav1	SAMEA112232992	whole organism
Hi-C sequencing	qeOrcFlav1	SAMEA112232992	whole organism
Sequencing information			
Platform	**Run accession**	**Read count**	**Base count (Gb)**
Hi-C Illumina NovaSeq 6000	ERR12259815	6.53e+08	98.56
PacBio Revio	ERR12257391	2.05e+06	24.24

Assembly quality metrics

The estimated Quality Value (QV) and *k*-mer completeness metrics, along with BUSCO completeness scores, were calculated for each haplotype and the combined assembly. The QV reflects the base-level accuracy of the assembly, while *k*-mer completeness indicates the proportion of expected *k*-mers identified in the assembly. BUSCO scores provide a measure of completeness based on benchmarking universal single-copy orthologues.

The primary haplotype has a QV of 63.1, and the combined primary and alternate assemblies achieve an estimated QV of 62.7. The *k*-mer completeness for the primary haplotype is 92.09%, and for the alternate haplotype it is 78.10%. The combined primary and alternate assemblies achieve a *k*-mer completeness of 98.54%. BUSCO analysis using the arthropoda_odb10 reference set (n = 1,013) indicated a completeness score of 96.2% (single = 88.6%, duplicated = 7.6%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project (EBP) Report on Assembly Standards September 2024. The assembly achieves the EBP reference standard of **6.C.Q63**.

Methods

Sample acquisition and DNA barcoding

The *Orchesella flavescens* used for sequencing (specimen ID Ox002318, ToLID qeOrcFlav1) was collected from Wytham Woods, Oxfordshire, United Kingdom (latitude 51.77, longitude –1.33) on 2022-07-22 by potting. The specimen was collected by James McCulloch and Liam Crowley (University of Oxford), identified by James McCulloch and preserved on dry ice.

The initial identification by expert ID 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) (Pereira *et al.*, 2022). 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 have been deposited on protocols.io (Beasley *et al.*, 2023).

Metadata collection for samples adhered to the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io (Denton *et al.*, 2023b). The qeOrcFlav1 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the whole organism

Table 2. Genome assembly data for *Orchesella flavescens*.

Genome assembly		
Assembly name	qeOrcFlav1.1	
Assembly accession	GCA_964034955.1	
Alternate haplotype accession	*GCA_964034975.1*	
Assembly level for primary assembly	chromosome	
Span (Mb)	273.69	
Number of contigs	249	
Number of scaffolds	13	
Longest scaffold (Mb)	58.43	
Assembly metric	**Measure**	***Benchmark***
Contig N50 length	2.42 Mb	*≥ 1 Mb*
Scaffold N50 length	51.76 Mb	*= chromosome N50*
Consensus quality (QV)	Primary: 63.1; alternate: 62.4; combined 62.7	*≥ 40*
k-mer completeness	Primary: 92.09%; alternate: 78.10%; combined: 98.54%	*≥ 95%*
BUSCO*	C:96.2%[S:88.6%,D:7.6%], F:1.2%,M:2.6%,n:1,013	*S > 90%; D < 5%*
Percentage of assembly mapped to chromosomes	99.89%	*≥ 90%*
Sex chromosomes	Not identified	*localised homologous pairs*
Organelles	Mitochondrial genome: 14.92 kb	*complete single alleles*

* BUSCO scores based on the arthropoda_odb10 BUSCO set using version 5.5.0. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison.

was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a).

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023a). For ultra-low input (ULI) PacBio sequencing, DNA was fragmented using the Covaris g-TUBE method (Oatley *et al.*, 2023b). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Strickland *et al.*, 2023). 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.

Hi-C sample preparation

Tissue from the whole organism of the qeOrcFlav1 sample was processed for Hi-C sequencing at the WSI Scientific Operations core, using the Arima-HiC v2 kit. In brief, frozen tissue (stored at –80 °C) was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde concentration. After crosslinking, the tissue was homogenised using the Diagnocine Power Masher-II and BioMasher-II tubes and pestles. Following the Arima-HiC v2 kit manufacturer's instructions, crosslinked DNA was digested using a restriction enzyme master mix. The 5'-overhangs were filled in and labelled with biotinylated nucleotides and proximally ligated. An overnight incubation was carried out for enzymes to digest remaining proteins and for crosslinks to reverse. A clean up was performed with SPRIselect beads prior to library preparation. Additionally, the biotinylation percentage was estimated using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit and Arima-HiC v2 QC beads.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core.

PacBio HiFi (ULI)

The sample requires Covaris g-TUBE shearing to approximately 10 kb prior to library preparation. Ultra-low input libraries were

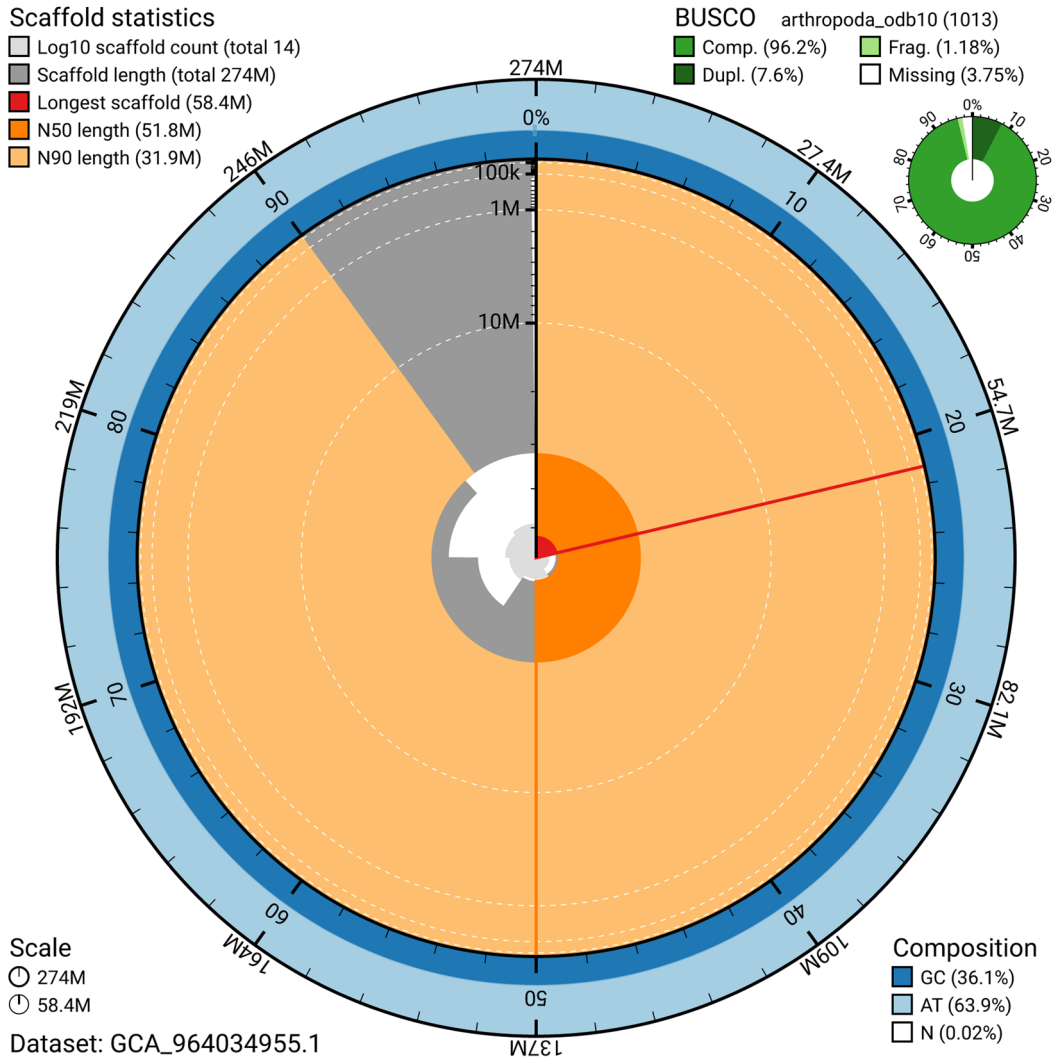

Figure 2. Genome assembly of *Orchesella flavescens*, qeOrcFlav1.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. 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 arthropoda_odb10 set is presented at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_964034955.1/dataset/GCA_964034955.1/snail.

prepared using PacBio SMRTbell® Express Template Prep Kit 2.0 and PacBio SMRTbell® gDNA Sample Amplification Kit. To begin, samples were normalised to 20 ng of DNA. Initial removal of single-strand overhangs, DNA damage repair, and end repair/A-tailing were performed per manufacturer's instructions. From the SMRTbell® gDNA Sample Amplification Kit, amplification adapters were then ligated. A 0.85X pre-PCR clean-up was performed with Promega ProNex beads and the sample was then divided into two for a dual PCR. PCR reactions A and B each followed the PCR programs as described in the manufacturer's protocol. A 0.85X post-PCR clean-up was performed with ProNex beads for PCR reactions A and B and DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit and fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and gDNA 55kb BAC analysis kit. PCR reactions A and B were then pooled, ensuring the total mass was ≥500 ng in 47.4 μl. The pooled sample then repeated the process for DNA damage repair, end repair/A-tailing and additional hairpin adapter ligation. A 1X clean-up was performed with ProNex beads and DNA concentration was quantified

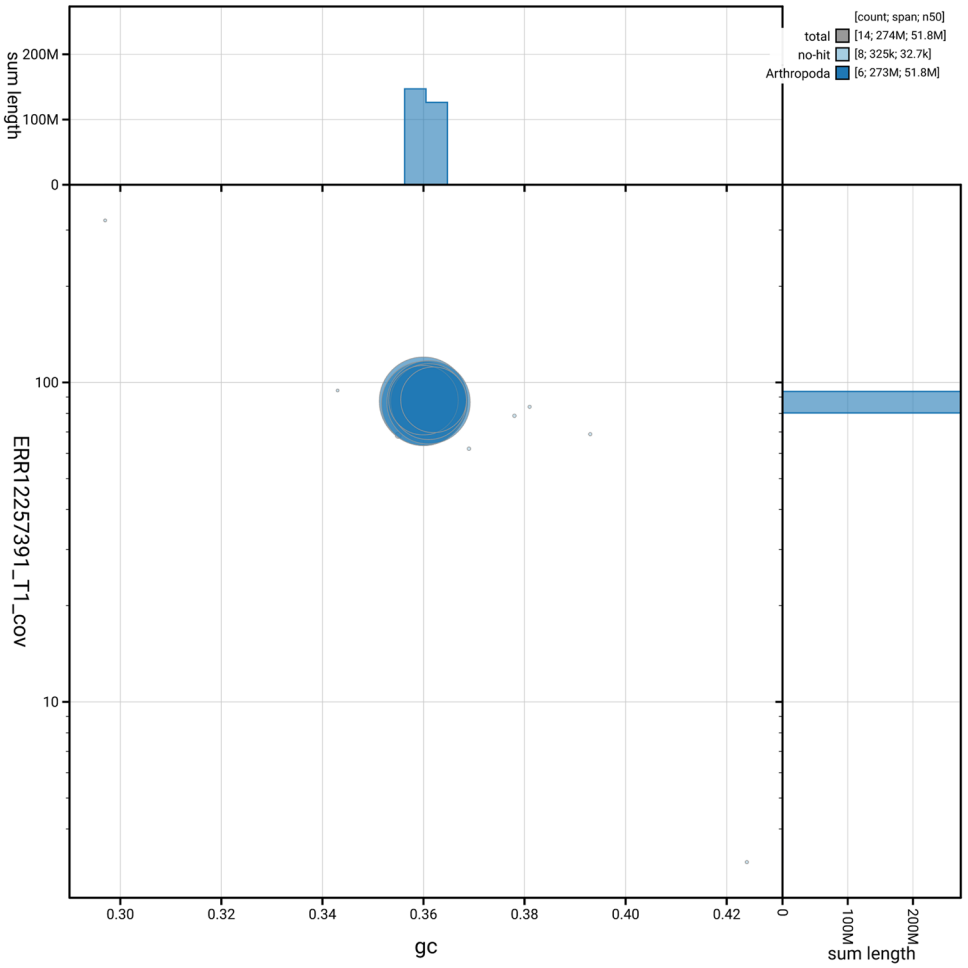

Figure 3. Genome assembly of *Orchesella flavescens*, qeOrcFlav1.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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 at https://blobtoolkit.genomehubs.org/view/GCA_964034955.1/blob.

using the Qubit and fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies). Size selection was performed using Sage Sciences' PippinHT system with target fragment size determined by analysis from the Femto Pulse, usually a value between 4000 and 9000 bp. Size selected libraries were then cleaned-up using1.0X ProNex beads and normalised to 2 nM before proceeding to sequencing.

Samples were sequenced on a Revio instrument (Pacific Biosciences, California, USA). Prepared libraries were normalised to 2 nM, and 15 µL was used for making complexes. Primers were annealed and polymerases were hybridised to create circularised complexes according to manufacturer's instructions. The complexes were purified with the 1.2X clean up with SMRTbell beads. The purified complexes were then diluted to the Revio loading concentration (in the range 200–300 pM), and spiked with a Revio sequencing internal control. Samples were sequenced on Revio 25M SMRT cells (Pacific Biosciences, California, USA). The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, as well as perform primary and secondary analysis of the data upon completion.

Hi-C

For Hi-C library preparation, DNA was fragmented using the Covaris E220 sonicator (Covaris) and size selected using SPRISelect beads to 400 to 600 bp. The DNA was then enriched using the Arima-HiC v2 kit Enrichment beads. Using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) for end repair, a-tailing, and adapter ligation. This uses a custom protocol which resembles the standard NEBNext Ultra II DNA Library Prep protocol but where library preparation occurs while DNA is bound to the Enrichment beads. For library amplification, 10 to 16 PCR cycles were required, determined by the sample biotinylation percentage. The Hi-C sequencing

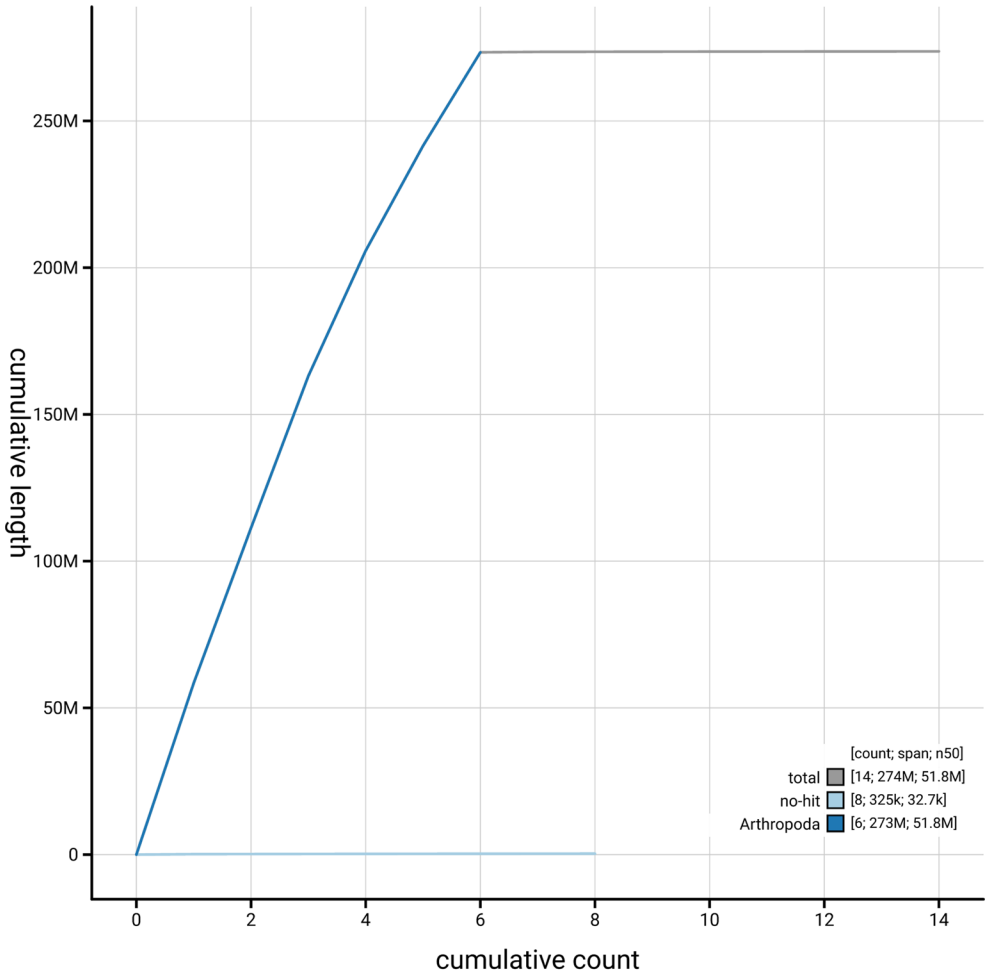

Figure 4. Genome assembly of *Orchesella flavescens,* qeOrcFlav1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_964034955.1/dataset/GCA_964034955.1/cumulative.

was performed using paired-end sequencing with a read length of 150 bp on an Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

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 first 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 were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). The contigs were further scaffolded using the provided Hi-C data (Rao *et al.*, 2014) 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 (article in preparation). Flat files and maps used in curation were generated in TreeVal (Pointon *et al.*, 2023). Manual curation was primarily

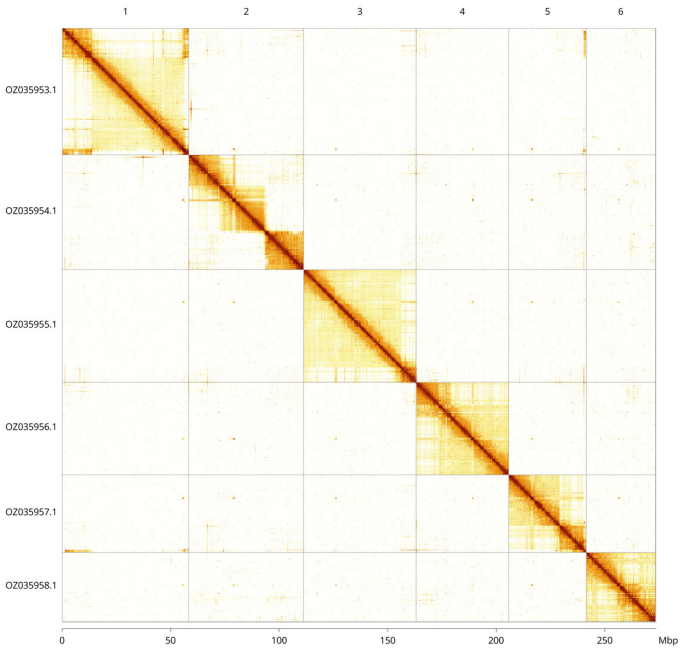

Figure 5. Genome assembly of *Orchesella flavescens*: Hi-C contact map of the qeOrcFlav1.1 assembly, visualised using PretextView. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure in HiGlass may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=dNatFGY-SvGRR3Gfrljvsg.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Orchesella flavescens*, qeOrcFlav1.

INSDC accession	Name	Length (Mb)	GC%
OZ035953.1	1	58.43	36
OZ035954.1	2	52.89	36
OZ035955.1	3	51.76	36
OZ035956.1	4	42.71	36
OZ035957.1	5	35.71	36
OZ035958.1	6	31.88	36
OZ035959.1	MT	0.01	29.5

conducted using PretextView (Harry, 2022), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023) and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were corrected, and duplicate sequences were tagged and removed. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation (article in preparation).

Assembly quality assessment

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

A Hi-C contact map was produced for the final version of the assembly. The Hi-C reads were aligned using bwa-mem2 (Vasimuddin *et al.*, 2019) and the alignment files were combined using SAMtools (Danecek *et al.*, 2021). The Hi-C alignments were converted into a contact map using BEDTools (Quinlan & Hall, 2010) and the Cooler tool suite (Abdennur & Mirny, 2020). The contact map is visualised in HiGlass (Kerpedjiev *et al.*, 2018).

The BlobToolKit pipeline is a Nextflow port of the previous Snakemake BlobToolKit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads in SAMtools and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoaT database (Challis *et al.*, 2023) to identify all matching BUSCO lineages to run BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineages, the pipeline aligns the BUSCO genes to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) with DIAMOND blastp (Buchfink *et al.*, 2021). The genome is also divided into chunks according to the density of the BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database using DIAMOND blastx. Genome sequences without a hit are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The blobtools suite combines all these 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), relying on the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), as well as the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

Table 4 contains a list of relevant software tool versions and sources.

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 here. 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

Table 4. Software tools: versions and sources.

Software tool	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
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.5-r587	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	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14, 1.17, and 1.18	https://github.com/MultiQC/MultiQC
NCBI Datasets	15.12.0	https://github.com/ncbi/datasets
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	-	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.5.1	https://github.com/sanger-tol/blobtoolkit

Software tool	Version	Source
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	1.2a.2	https://github.com/c-zhou/yahs

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: Orchesella flavescens. Accession number PRJEB68258; https://identifiers.org/ena.embl/PRJEB68258. The genome sequence is released openly for reuse. The *Orchesella flavescens* 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. 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.

Author information

Members of the University of Oxford and Wytham Woods Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.12157525.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.12158331.

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: https://doi.org/10.5281/zenodo.12162482.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: https://doi.org/10.5281/zenodo.12165051.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: https://doi.org/10.5281/zenodo.12160324.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.12205391.

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

References

Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316.
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

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

Beasley J, Uhl R, Forrest LL, *et al.*: **DNA barcoding SOPs for the Darwin Tree of Life project.** *protocols.io.* 2023; [Accessed 25 June 2024].
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, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24.
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

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

da Veiga Leprevost F, Grüning BA, Alves Aflitos S, *et al.*: **BioContainers: an open-source and community-driven framework for software standardization.** *Bioinformatics.* 2017; **33**(16): 2580–2582.
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

Denton A, Oatley G, Cornwell C, *et al.*: **Sanger Tree of Life sample homogenisation: PowerMash.** *protocols.io.* 2023a.
Publisher Full Text

Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life wet laboratory protocol collection V.1.** *protocols.io.* 2023b.
Publisher Full Text

Diesh C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of synteny and structural variation.** *Genome Biol.* 2023; **24**(1): 74.
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

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
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.
Reference Source

Hopkin SP: **A key to the Collembola (springtails) of Britain and Ireland.** Shrewsbury: Field Studies Council (FSC) Publications, 2007.
Reference Source

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

Jaron KS, Hodson CN, Ellers J, *et al.*: **Genomic evidence of paternal genome elimination in the globular springtail *Allacma fusca*.** *Genetics.* 2022; **222**(3): iyac117.
PubMed Abstract | Publisher Full Text | Free Full Text

Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life sample preparation: triage and dissection.** *protocols.io.* 2023.
Publisher Full Text

Jin J, Zhan Z, Wei X, *et al.*: **Genomic insights into the chromosomal elongation in a family of Collembola.** *Proc Biol Sci.* 2024; **291**(2018): 20232937.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
PubMed Abstract | Publisher Full Text | Free Full Text

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

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

Manni M, Berkeley MR, 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, [Accessed 2 April 2024].
Reference Source

Oatley G, Denton A, Howard C: **Sanger Tree of Life HMW DNA extraction: automated MagAttract v.2.** *protocols.io.* 2023a.
Publisher Full Text

Oatley G, Sampaio F, Kitchin L, *et al.*: **Sanger Tree of Life HMW DNA fragmentation: covaris g-TUBE for ULI PacBio.** *protocols.io.* 2023b; [Accessed 13 June 2024].
Publisher Full Text

Pereira L, Sivell O, Sivess L, *et al.*: **DToL Taxon-specific Standard Operating Procedure for the terrestrial and freshwater arthropods working group**. 2022.
Publisher Full Text

Pointon DL, Eagles W, Sims Y, *et al.*: **sanger-tol/treeval v1.0.0 – Ancient Atlantis.** 2023.
Publisher Full Text

Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–842.
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

Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *protocols.io.* 2023.
Publisher Full Text

Timmermans MJTN, Ellers J, Mariën J, *et al.*: **Genetic structure in *Orchesella cincta* (Collembola): strong subdivision of European populations inferred from mtDNA and AFLP markers.** *Mol Ecol.* 2005; **14**(7): 2017–2024.
PubMed Abstract | Publisher 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, 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

Womersley H: **The apterygotan fauna of the South West of England. II.** *Proc Bris Nat Soc.* 1924; **6**: 166–172.

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 2

Reviewer Report 27 September 2025

https://doi.org/10.21956/wellcomeopenres.27550.r134421

✓ **Michael Hiller**
Senckenberg Nature Research Society, Frankfurt Am Main, Germany

All my comments have been addressed.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: 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.

Version 1

Reviewer Report 07 April 2025

https://doi.org/10.21956/wellcomeopenres.26221.r120913

? **Michael Hiller**
Senckenberg Nature Research Society, Frankfurt Am Main, Germany

The study presents the first high-quality genome of a springtail in the Genus Orchesella. The genome is highly contiguous and complete. Also, I am pleased to see that the PacBio library preparation protocol is described in detail.

I have only one comment regarding the scaffolding: A karyotype is likely not known (?) and the HiC signal shows variation inside the chromosome-level scaffolds. E.g. scaffold 2 and 5 have a block at the bottom that equal signal to the other regions in same same and other scaffolds. Scaffold 1 also has a smaller part at the bottom that appears different. I therefore wonder if one can check if the 6 chromosomes have telomers and if the potentially problematic regions lack telomers.

What is meant by 'dubious records' in the introduction? Species misidentification?

It would be great, but optional of course, to add an image to Figure 1 that shows some of the described morphological characteristics.

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

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.

Author Response 20 Sep 2025

Tree of Life Team Sanger

Thank you for reviewing this data note. We have made changes to version 2 of the article, and respond to your main points below:

Reviewer comment: "A karyotype is likely not known (?) and the HiC signal shows variation inside the chromosome-level scaffolds. E.g. scaffold 2 and 5 have a block at the bottom that equal signal to the other regions in same same and other scaffolds. Scaffold 1 also has a smaller part at the bottom that appears different. I therefore wonder if one can check if the 6 chromosomes have telomers and if the potentially problematic regions lack telomers."

Response: We curated the assembly in PretextView with a telomere-repeat track enabled to help

verify chromosome boundaries. The regions highlighted by the reviewer do not show off-diagonal contact patterns indicative of mis-joins; the diagonals remain continuous through these areas. We have replaced the static image of the Hi-C chromosome map with a labelled PretextView map. The HiGlass image remains hyperlinked in the caption.

Reviewer comment: "What is meant by 'dubious records' in the introduction? Species misidentification?"

Response: We have clarified in the text that we are referring to potentially misidentified specimens. The Tree of Life team

Competing Interests: No competing interests were disclosed.

Reviewer Report 31 March 2025

https://doi.org/10.21956/wellcomeopenres.26221.r120912

?

Nerivânia Nunes Godeiro
Shanghai Natural History Museum, Shanghai, Shanghai, China

The authors presented the genome of *Ochesella flavescens*, a Collembola species widespread in Europe.

In the background section, the authors said that the species belongs to Entomobryidae family, but it belongs to Orchesellidae, as described in the species taxonomy section.

I have more concerns about the methodology. The first is related to the BUSCO reference set used to assess the completeness of the genome. There is a specific reference set "collembola_odb10" (n=4073) available, and the authors used a general "arthropoda_odb10" (n=1013).

I also didn't understand how the authors were able to perform all experiments using only one specimen, at least this is what they said in the sample acquisition topic. Sometimes one whole specimen is not enough to produce a good result, imagine a bit of tissue. It would be interesting to see the protocol followed for this specific case, not the general protocol cited as the reference. Collembola are tiny animals, please give more details of how much tissue was used for each step.

The mitogenome sequence submitted to NCBI was not annotated. Did you check if it's complete?

Except for these details, the article is well written and makes an important contribution.

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?
Partly

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Collembola phylogenomics and taxonomy

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.

Author Response 20 Sep 2025

Tree of Life Team Sanger

Thank you for reviewing this data note. We have made changes to version 2 of the article, and respond to your main points below:

Reviewer comment 1: In the background section, the authors said that the species belongs to Entomobryidae family, but it belongs to Orchesellidae, as described in the species taxonomy section.

Response: We have corrected this and added an explanation in the background about the change of taxonomic assignment for this species.

Reviewer comment 2: I have more concerns about the methodology. The first is related to the BUSCO reference set used to assess the completeness of the genome. There is a specific reference set "collembola_odb10" (n=4073) available, and the authors used a general "arthropoda_odb10" (n=1013).

Response: There is no officially released BUSCO lineage for Collembola in either ODB10 or ODB12; the BUSCO catalogue provides higher-level sets (e.g. arthropoda_odb10) but not a Collembola-specific dataset. Accordingly, we used the closest available official lineage on our system, which is arthropoda_odb10 for ODB10, to ensure reproducibility.

Reviewer comment 3: I also didn't understand how the authors were able to perform all experiments using only one specimen, at least this is what they said in the sample acquisition topic. Sometimes one whole specimen is not enough to produce a good result, imagine a bit of tissue. It would be interesting to see the protocol followed for this specific case, not the general protocol cited as the reference. Collembola are tiny animals, please

give more details of how much tissue was used for each step.

Response: It is difficult to obtain enough material for both PacBio and DNA sequencing from a tiny arthropod, but it is preferable to use Hi-C data from the same organism to scaffold the assembly. The Methods section indicates that we used the Ultra-low input library preparation technique and Revio sequencing to achieve good genome coverage.

Reviewer comment 4: The mitogenome sequence submitted to NCBI was not annotated. Did you check if it's complete?

Response: Yes, while we do not provide full annotation, MitoHiFi uses MitoFinder to ensure that the sequence is complete.

Competing Interests: No competing interests were disclosed.

Reviewer Report 31 March 2025

https://doi.org/10.21956/wellcomeopenres.26221.r120921

?

Daoyuan Yu
Nanjing Agricultural University, Nanjing, China
Zhihong Zhan
Nanjing Agricultural University College of Plant Protection (Ringgold ID: 214175), Nanjing, Jiangsu, China

This study provides foundational data for the field of springtail genomics.

Below are my review comments:

1. A detailed analysis of repeat sequence proportions and types should be provided.

2. BUSCO assessment shows 7.6% duplicated genes, exceeding the EBP standard of less than 5%; however, no explanation is provided.

3. No detailed discussion of how genome heterozygosity (0.3%) affected assembly.

4. A higher-resolution version of the Hi-C contact map in Figure 5 is needed.

5. Table 3 should include information on gene density for each chromosome.

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: Collembola systematics, ecology and genomics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.