DATA NOTE

The genome sequence of the European ground squirrel, *Spermophilus citellus* (Linnaeus, 1766)

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

Dimitra-Lida Rammou[1], Dionisios Youlatos[1,2], Alexandros Triantafyllidis[3,4], 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

[1]Department of Zoology, School of Biology, Aristotle University of Thessaloniki, Thessaloniki, Greece
[2]International Center for Biodiversity and Primate Conservation, Dali University, Dali, Yunnan, China
[3]Department of Genetics, Development & Molecular Biology, School of Biology, Aristotle University of Thessaloniki, Thessaloniki, Greece
[4]Genomics and Epigenomics Translational Research, Center for Interdisciplinary Research and Innovation, Balkan Center, Thessaloniki, Greece

v1 **First published:** 08 Apr 2025, **10**:184
https://doi.org/10.12688/wellcomeopenres.23974.1
Latest published: 08 Apr 2025, **10**:184
https://doi.org/10.12688/wellcomeopenres.23974.1

Abstract

We present a genome assembly from a female *Spermophilus citellus* (European ground squirrel; Chordata; Mammalia; Rodentia; Sciuridae). The genome sequence has a total length of 3,090.03 megabases. Most of the assembly (95.47%) is scaffolded into 20 chromosomal pseudomolecules, including the X sex chromosome. The mitochondrial genome has also been assembled, with a length of 16.45 kilobases.

Keywords

Spermophilus citellus, European ground squirrel, genome sequence, chromosomal, Rodentia

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ? ✓

	1	2	3
version 1 08 Apr 2025	✓ view	? view	✓ view

1. **Camila do Nascimento Moreira**, Universidade Federal de Mato Grosso Do Sul, Campo Grande, Brazil
2. **Gerrit Wehrenberg**, University of Oulu, Oulu, Finland
3. **Pawel Michalak**, University of Haifa, Haifa, Israel
Edward Via College of Osteopathic Medicine (Ringgold ID: 41066), Monroe, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team (Mark.Blaxter@sanger.ac.uk)

Author roles: Rammou DL: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Youlatos D**: Investigation, Resources, Writing – Review & Editing; **Triantafyllidis A**: Investigation, Resources, Supervision, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540).
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: Rammou DL, Youlatos D, Triantafyllidis A *et al.* **The genome sequence of the European ground squirrel, *Spermophilus citellus* (Linnaeus, 1766) [version 1; peer review: 2 approved, 1 approved with reservations]** Wellcome Open Research 2025, **10**:184 https://doi.org/10.12688/wellcomeopenres.23974.1

First published: 08 Apr 2025, **10**:184 https://doi.org/10.12688/wellcomeopenres.23974.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Glires; Rodentia; Sciuromorpha; Sciuridae; Xerinae; Marmotini; *Spermophilus*; *Spermophilus citellus* (Linnaeus, 1766) (NCBI:txid9997)

Background

Spermophilus citellus (Figure 1), the European ground squirrel, is a semifossorial rodent with an elongated body between 18–24 cm, short tail measuring 20–40% of head-body length, grayish-yellow fur, and small ears. Its body mass varies from 145 g to 520 g, depending on age, sex, life cycle and environment (Matějů, 2008; Ramos-Lara *et al.*, 2014; Ružić, 1978). It inhabits grasslands and agricultural ecosystems up to 2,500 m, forming colonies and constructing burrows for hibernation, reproduction, and shelter. As a keystone species, it supports predators, enhances vegetation diversity and soil structure, and provides habitat for reptiles, birds, and arthropods in abandoned burrows.

Spermophilus citellus is endemic to Central and Southeastern Europe, with its distribution divided by the Carpathian Mountains (Kryštufek *et al.*, 2009; Říčanová *et al.*, 2013). Its northwestern range includes Czechia, Austria, Poland, Slovakia, Hungary, northern Serbia, and western Romania, while the southeastern range spans southern Serbia, North Macedonia, Bulgaria, southern Romania, Greece, Turkish Thrace, Moldova, and Ukraine (Ćosić *et al.*, 2024). Classified as Endangered [A3c ver 3.1], it is protected under the Bern Convention (Appendix II) and the EU Habitats and Species Directive (Annexes II and IV) (Ćosić *et al.*, 2024). Its populations are in continuous decline, primarily due to habitat degradation and loss, threatening its long-term survival.

Figure 1. A juvenile *Spermophilus citellus* from Thessaloniki, Greece (photo by Dionisios Youlatos) (not the specimen used for genome sequencing).

The recent sequencing project represents the first high-profile genomic initiative for this species, providing a crucial resource for investigating adaptive responses to environmental changes and human pressures. The broad latitudinal and altitudinal distribution of the target species offers a unique opportunity to examine adaptation to shifting environmental conditions, particularly in the context of climate change. Additionally, improving tools for evidence-based conservation management is essential. Integrating genetic diversity data into biodiversity conservation strategies remains a gap for this species, necessitating the development of efficient indicators to assess the adaptive potential of the species populations.

Genome sequence report

Sequencing data

The genome of a specimen of *Spermophilus citellus* was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 129.19 Gb (gigbases) from 10.02 million reads. GenomeScope analysis of the PacBio HiFi data estimated the haploid genome size at 3,115.01 Mb, with a heterozygosity of 0.18% and repeat content of 31.24%. 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 30.0x coverage of the genome. Chromosome conformation Hi-C sequencing produced 472.86 Gb from 3,131.52 million reads. Table 1 summarises the specimen and sequencing information.

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 218 misjoins or missing joins and removed 12 haplotypic duplications. These interventions reduced the total assembly length by 0.84%, decreased the scaffold count by 26.66%, and increased the scaffold N50 by 13.24%. The final assembly has a total length of 3,090.03 Mb in 497 scaffolds, with 1,436 gaps, and a scaffold N50 of 155.99 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 (95.48%) was assigned to 20 chromosomal-level scaffolds, representing 19 autosomes and the X sex chromosome. These chromosome-level scaffolds, confirmed

Table 1. Specimen and sequencing data for *Spermophilus citellus*.

Project information			
Study title	Spermophilus citellus (European suslik)		
Umbrella BioProject	PRJEB73447		
Species	*Spermophilus citellus*		
BioSpecimen	SAMEA10332752		
NCBI taxonomy ID	9997		
Specimen information			
Technology	**ToLID**	**BioSample accession**	**Organism part**
PacBio long read sequencing	mSpeCit3	SAMEA10332755	blood
Hi-C sequencing	mSpeCit2	SAMEA10332754	skin
Sequencing information			
Platform	**Run accession**	**Read count**	**Base count (Gb)**
Hi-C Illumina NovaSeq 6000	ERR12723499	3.13e+09	472.86
PacBio Sequel IIe	ERR12721086	2.44e+06	31.51
PacBio Revio	ERR12721087	7.57e+06	97.68

by Hi-C data, are named according to size (Figure 5; Table 3). During curation, it was noted that the order and orientation of contigs in Chromosome 12 between approximately 23.6–68.8Mb are uncertain. Chromosome X was identified by alignment to the genomes of *Sciurus vulgaris* (GCA_902686455.2) (Mead *et al.*, 2020b) and *Sciurus carolinensis* (GCA_902686445.2) (Mead *et al.*, 2020a).

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.

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 combined primary and alternate assemblies achieve an estimated QV of 59.9. The *k*-mer recovery for the primary haplotype is 95.78%, and for the alternate haplotype 62.61%; the combined primary and alternate assemblies have a *k*-mer recovery of 99.41%. BUSCO analysis using the glires_odb10 reference set (n = 13,798) identified 96.1% of the expected gene set (single = 92.2%, duplicated = 3.8%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The assembly achieves the EBP reference standard of **6.C.Q59.**

Methods

Sample acquisition

On September 1, 2021, four samples were collected from two living *Spermophilus citellus* specimens belonging to a single population in Thessaloniki, Greece (latitude: N 23.00, longitude: E 40.53, elevation: 25 m). Due to the species' endangered and protected status, only non-lethal sampling was performed. The first specimen (ID: ERGA_AT_GR_011, SAMEA10332751, ToLID: mSpeCit2) was an adult female from which skin tissue (SAMEA10332754) and blood (SAMEA10332753) were collected. The second specimen (ID: ERGA_AT_GR_012, SAMEA10332752, ToLID: mSpeCit3) was a juvenile female from which skin tissue (SAMEA10332756) and blood (SAMEA10332755) were obtained. Immediately after collection, all samples were placed in Eppendorf tubes, transferred to a liquid nitrogen container, and subsequently stored in a –80°C freezer for preservation. Following sampling, both individuals were released back into their natural habitat. The species identification and sample collection were conducted by Dimitra-Lida Rammou, Dionisios Youlatos, and Anastasia Diakou.

The specimen with ID ERGA_A_GR_012 (ToLID mSpeCit3) was used for PacBio HiFi sequencing, from which the genome was assembled. The specimen with ID ERGA_AT_GR_011, ToLID mSpeCit2) was used for Hi-C scaffolding of the assembly.

Table 2. Genome assembly data for *Spermophilus citellus*.

Genome assembly		
Assembly name	mSpeCit3.1	
Assembly accession	GCA_964194105.1	
Alternate haplotype accession	*GCA_964194095.1*	
Assembly level for primary assembly	chromosome	
Span (Mb)	3,090.03	
Number of contigs	1,933	
Number of scaffolds	497	
Longest scaffold (Mb)	288.97	
Assembly metric	**Measure**	***Benchmark***
Contig N50 length	3.45 Mb	*≥ 1 Mb*
Scaffold N50 length	155.99 Mb	*= chromosome N50*
Consensus quality (QV)	Primary: 60.8; alternate: 58.9; combined 59.9	*≥ 40*
k-mer completeness	Primary: 95.78%; alternate: 62.61%; combined: 99.41%	*≥ 95%*
BUSCO*	C:96.1%[S:92.2%,D:3.8%], F:0.7%,M:3.2%,n:13,798	*S > 90%; D < 5%*
Percentage of assembly mapped to chromosomes	95.48%	*≥ 90%*
Sex chromosomes	X	*localised homologous pairs*
Organelles	Mitochondrial genome: 16.45 kb	*complete single alleles*

* BUSCO scores based on the glires_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.

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 mSpeCit3 sample was prepared for DNA extraction on dry ice (Jay *et al.*, 2023). The blood sample was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a). HMW DNA was extracted using the Manual MagAttract v1 protocol (Strickland *et al.*, 2023b). DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Todorovic *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Strickland *et al.*, 2023a). 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 skin of the mSpeCit2 sample was processed for Hi-C sequencing at the WSI Scientific Operations core, using the Arima-HiC v2 kit. In brief, 20–50 mg of 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

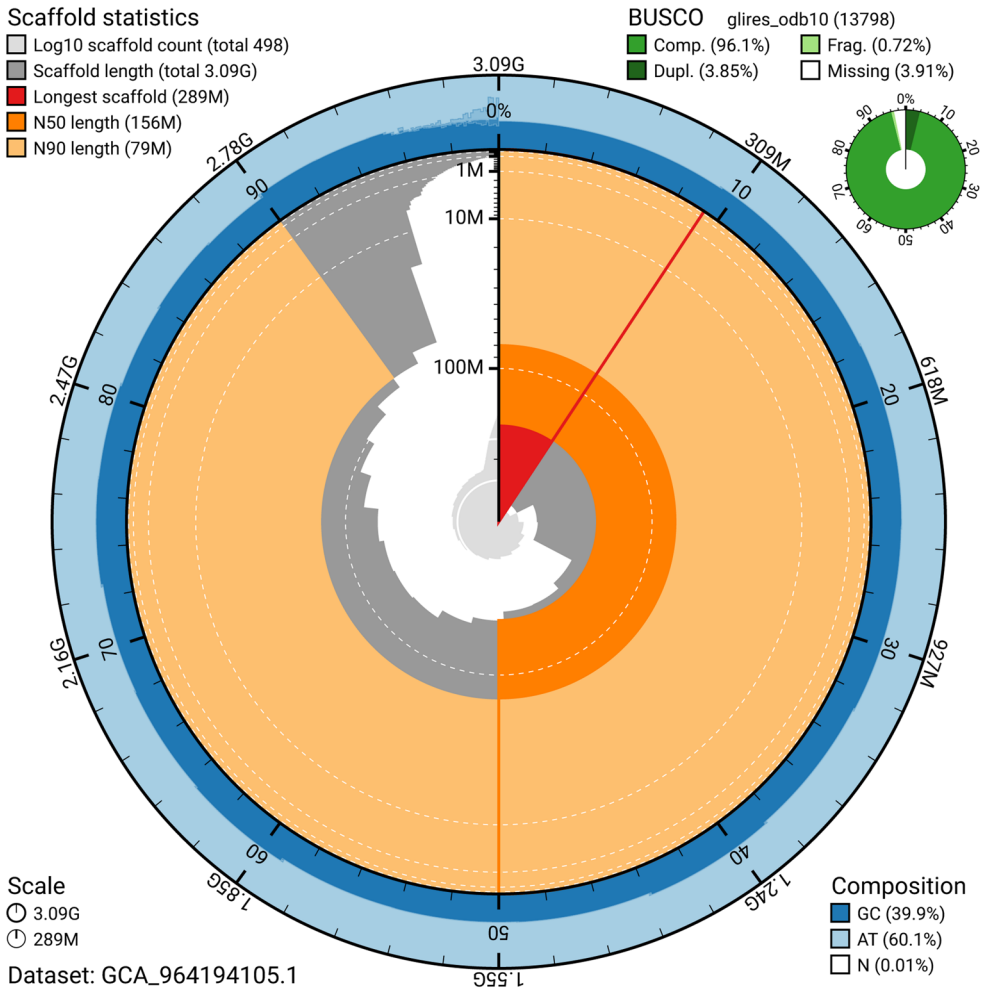

Figure 2. Genome assembly of *Spermophilus citellus*, mSpeCit3.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. 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 glires_odb10 set is presented at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_964194105.1/dataset/GCA_964194105.1/snail.

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

At a minimum, samples were required to have an average fragment size exceeding 8 kb and a total mass over 400 ng to proceed to the low input SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences, California, USA), depending on genome size and sequencing depth required. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA) as per the manufacturer's instructions. The kit includes the reagents required for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead cleanup, and nuclease treatment. Following the manufacturer's instructions, size selection and clean up was carried out using diluted AMPure PB beads (Pacific Biosciences, California, USA). DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) with Qubit 1X dsDNA HS assay kit and the final library fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and gDNA 55kb BAC analysis kit.

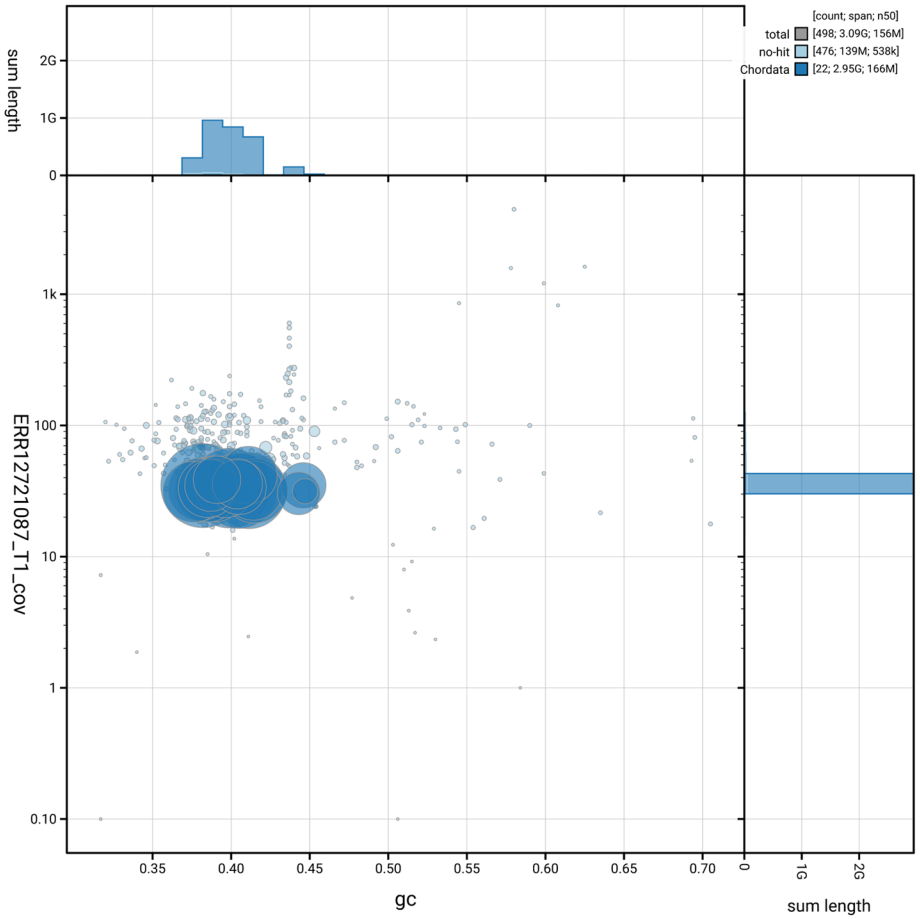

Figure 3. Genome assembly of *Spermophilus citellus*, mSpeCit3.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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_964194105.1/blob.

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 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.

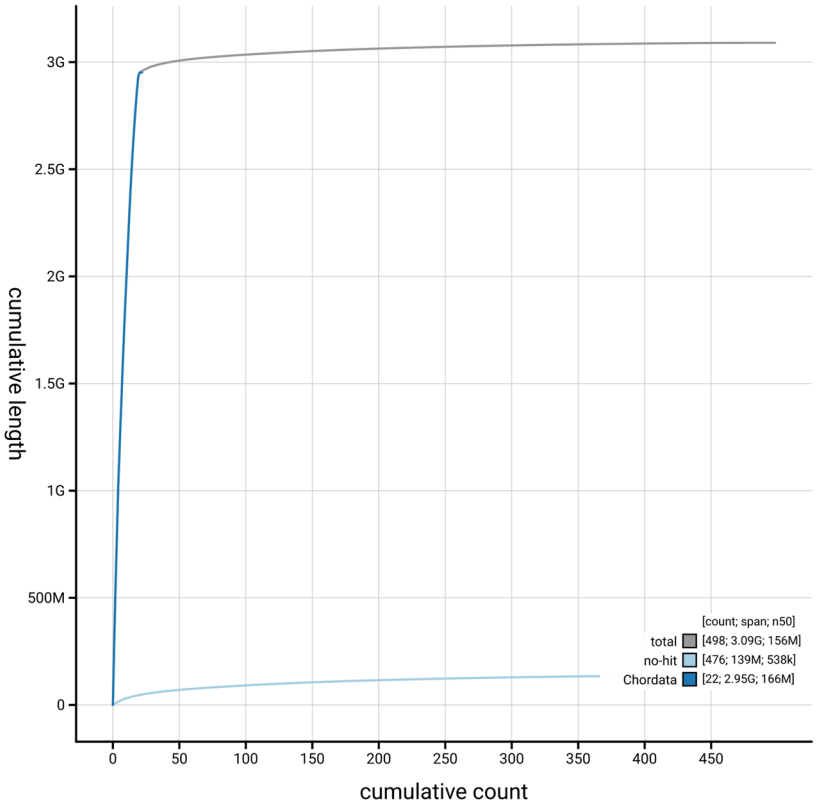

Figure 4. Genome assembly of *Spermophilus citellus*, mSpeCit3.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_964194105.1/dataset/GCA_964194105.1/cumulative.

The HiFi reads were first assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. 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. Flat files and maps used in curation were generated via the TreeVal pipeline (Pointon *et al.*, 2023). Manual curation was conducted primarily in PretextView (Harry, 2022) and HiGlass (Kerpedjiev *et al.*, 2018), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were amended, and duplicate sequences were tagged and removed. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation.

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 was 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

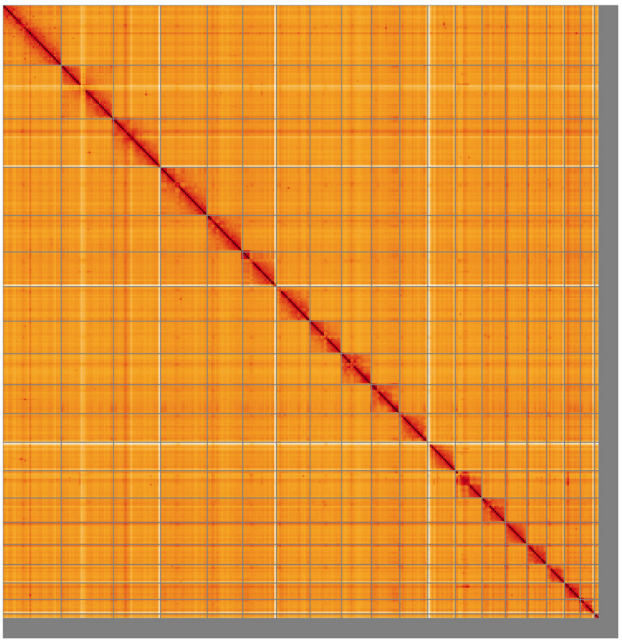

Figure 5. Genome assembly of *Spermophilus citellus*: Hi-C contact map of the mSpeCit3.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=c65KyKOKQG60ig0sTFmtoQ.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Spermophilus citellus*, mSpeCit3.

INSDC accession	Name	Length (Mb)	GC%
OZ077378.1	1	288.97	38
OZ077379.1	2	258.54	40
OZ077380.1	3	232.93	41
OZ077381.1	4	231.55	40.5
OZ077382.1	5	175.33	39
OZ077383.1	6	166.8	41.5
OZ077384.1	7	166.25	38.5
OZ077385.1	8	155.99	38
OZ077387.1	9	139.52	41.5
OZ077388.1	10	137.97	39.5
OZ077389.1	11	136.59	38.5
OZ077390.1	12	131.62	41
OZ077391.1	13	116.84	40.5
OZ077392.1	14	105.9	38.5
OZ077393.1	15	97.87	40.5
OZ077394.1	16	87.8	39
OZ077395.1	17	78.99	44.5
OZ077396.1	18	69.43	44.5
OZ077397.1	19	22.32	44.5
OZ077386.1	X	149.01	37.5
OZ077398.1	MT	0.02	35.5

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 soltions.

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 Tree of Life collaborator. 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

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	666652151335353eef2fcd58880bcef5bc2928e1	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	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b8 4aa44357826c0b6753eb28de	https://github.com/higlass/higlass
MerquryFK	d00d98157618f4e8d1a9190026b19b471055b 22e	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
Nextflow	23.10.0	https://github.com/nextflow-io/nextflow
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
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
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

to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible.

The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international)

Each transfer of samples is undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Tree of Life collaborator, Genome Research Limited (operating as the Wellcome Sanger Institute) and in some circumstances other Tree of Life collaborators.

Data availability

European Nucleotide Archive: Spermophilus citellus (European suslik). Accession number PRJEB73447; https://identifiers.org/ena.embl/PRJEB73447. The genome sequence is released openly for reuse. The *Spermophilus citellus* genome sequencing initiative is part of the European Reference Genome Atlas Pilot Project (https://www.erga-biodiversity.eu/pilot-project) and the Vertebrate Genomes Project (PRJNA489243). 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 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.

Acknowledgements

We extend our gratitude to Anastasia Diakou for her assistance with blood collection in the field and to Konstantinos Gagavouzis for his contributions to sample preparation and shipment. Field research complied with the laws of the Hellenic Ministry of Environment and Energy (research permit 178107/64/2019).

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

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

Ćosić N, Ćirović D, Fülöp T, *et al.*: ***Spermophilus citellus*.** *The IUCN Red List of Threatened Species.* 2024; e.T20472A221789466.
Publisher 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

Harry E: **PretextView (Paired Read Texture Viewer): a desktop application for viewing pretext contact maps.** 2022.
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

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

Kryštufek B, Bryja J, Bužan EV: **Mitochondrial phylogeography of the European ground squirrel, *Spermophilus citellus*, yields evidence on refugia for steppic taxa in the southern Balkans.** *Heredity (Edinb).* 2009; **103**(2): 129–35.
PubMed Abstract | Publisher 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

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

Matějů J: **Ecology and space use in a relict population of the European ground squirrel (*Spermophilus citellus*) at the north-western edge of its distribution range.** *Lynx.* 2008; **39**(2): 263–276.
Reference Source

Mead D, Fingland K, Cripps R, *et al.*: **The genome sequence of the eastern grey squirrel, *Sciurus carolinensis* Gmelin, 1788 [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2020a; **5**: 27.
PubMed Abstract | Publisher Full Text | Free Full Text

Mead D, Fingland K, Cripps R, *et al.*: **The genome sequence of the Eurasian red squirrel, *Sciurus vulgaris* Linnaeus 1758 [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2020b; **5**: 18.
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

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

Ramos-Lara N, Koprowski JL, Kryštufek B, *et al.*: ***Spermophilus citellus* (Rodentia: Sciuridae).** *Mammalian Species.* 2014; **46**(913): 71–87.
Publisher 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

Říčanová Š, Koshev Y, Říčan O, *et al.*: **Multilocus phylogeography of the European ground squirrel: cryptic interglacial Refugia of continental climate in Europe.** *Mol Ecol.* 2013; **22**(16): 4256–4269.
PubMed Abstract | Publisher Full Text

Ružić A: ***Citellus citellus* (Linnaeus, 1766) Der oder das Europäische Ziesel**. In: *Handbuch der Säugetiere Europas*. 1978; **1**: 123–144.

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

Strickland M, Moll R, Cornwell C, *et al.*: **Sanger Tree of Life HMW DNA extraction: manual MagAttract.** *protocols.io.* 2023b.
Publisher Full Text

Todorovic M, Sampaio F, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode megaruptor®3 for PacBio HiFi.** *protocols.io.* 2023.
Publisher 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 26 May 2025

https://doi.org/10.21956/wellcomeopenres.26450.r122130

✔ **Pawel Michalak**
[1] University of Haifa, Haifa, Israel
[2] Edward Via College of Osteopathic Medicine (Ringgold ID: 41066), Monroe, Louisiana, USA

This Data Note describes the chromosome-level genome assembly of the European ground squirrel, an endangered rodent species endemic to Central and Southeastern Europe. Using Pacific Biosciences HiFi long-read sequencing and Hi-C scaffolding, the authors report a high-quality genome assembly of 3.09 Gb, with 95.48% of the sequence scaffolded into 20 chromosomal pseudomolecules including the X chromosome. The mitochondrial genome was also assembled. The authors provide extensive details on sample collection, sequencing protocols, assembly pipelines, and quality assessment metrics, including QV scores, BUSCO completeness (96.1%), and k-mer completeness (99.41%). The study aims to contribute to the genomic resources of S. citellus to support research in evolutionary biology, population genetics, and conservation genomics, especially given the species' status as endangered and its ecological importance.

This is a well-executed and clearly presented genome assembly report. The data quality is high, the methodological framework is sound, and the dataset is well positioned to benefit researchers in genomics, ecology, and conservation biology. The only minor recommendation I would offer is to expand the background to reference existing studies on comparative genomics across rodents, chromosomal polymorphisms or karyotypic evolution in Spermophilus and related taxa, which would strengthen the evolutionary rationale for the genome assembly.

Incidentally, I couldn't help but notice that one of the reviewers expressed concern about the use of GenomeScope2 with PacBio data, suggesting the software is not appropriate for high-error-rate reads. However, the authors used PacBio HiFi sequencing, which achieves an accuracy of 99.9% (0.1% error rate), comparable to Illumina. GenomeScope2 is widely and successfully used with HiFi data, as demonstrated in recent studies (e.g., https://www.nature.com/articles/s41597-024-03808-w).

References

1. Castaño MI, Ye X, Uy FMK: First genome assembly of the order Strepsiptera using PacBio HiFi reads reveals a miniature genome.*Sci Data*. 2024; **11** (1): 934 PubMed Abstract | Publisher Full Text

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: Genomics, evolution, bioinformatics

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 24 April 2025

https://doi.org/10.21956/wellcomeopenres.26450.r122136

?

Gerrit Wehrenberg
University of Oulu, Oulu, Finland

I want to thank the authors for their work and the submission of this article.

The Data Note titled '*The genome sequence of the European ground squirrel, Spermophilus citellus (Linnaeus, 1766)*' by Dimitra-Lida Rammou, Dionisios Youlatos, and Alexandros Triantafyllidis provide and present the chromosome-level genome assembly and its methodological pipeline of two female European ground squirrels (*Spermophilus citellus*) from a population in Thessaloniki, Greece.

The authors provide a solid background on the focal species of the study and clearly explain the sequencing project within which the presented genome assembly was generated.

I think the Background section would be more valuable with mentioning the evolutionary research done in this clade, since a chromosome-level genome assembly could play an important role in this field. There are for example interesting evolutionary karyotype patterns in the Nearctic and Palearctic realms. Genomic data could shed a new light on this matter.
The validation of the karyotype in line with earlier studies would strengthen your findings as well.

Generally, the authors should point out more why this novel genome assembly is important overall and specifically for this species/higher taxon with referencing literature and maybe its conservation situation.

In the framework of a data note the workflow protocol was overall well provided.

Please consider to not utilise *GenomeScope2* for PacBio reads since this software is not fully appropriate for data with high sequencing error rates (e.g., PacBio reads). *GenomeScope2* is designed for accurate *k*-mer profiles, typically from Illumina reads. There are alternatives providing estimates of genome size, heterozygosity, and repeat content more suitable for your data.

Utilised field work, laboratory methodology and software (including versions) are provided. The provision of detailed information on the type locality and the studied population is particularly commendable, as such data are crucial for the prospective application of the genome assembly in population genetic research.

Please provide the information of the morphological sex identification method of the live specimens. Especially, since often very small Y-gonosomes can be hard to detect in Hi-C data even with a coverage of 30x. This would strengthen your molecular findings.

Specimen and sequencing data as well as assembly statistics are provided and the presented genome assembly is already assessable publicly.

I approve this article for publication to the editor, provided that the minor revisions mentioned above are addressed.

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

Yes

Are the protocols appropriate and is the work technically sound?

Partly

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: Wildlife and Conservation 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 18 April 2025

https://doi.org/10.21956/wellcomeopenres.26450.r122129

Camila do Nascimento Moreira
Universidade Federal de Mato Grosso Do Sul, Campo Grande, Brazil

The article brings new sequencing data about the European squirrel *Spermophilus citellus*, a rodent species classified as endangered. The sequencing process and assembly were so well-detailed. Also, all data produced and software used are available to be checked. However, I have some suggestions to be incorporated into the manuscript, as follows:

1. The species full name must be cited only once, upcoming citations must abbreviate the genus name;

2. Specify the meaning of the abbreviation 'EU' in the third sentence of the second paragraph of the section 'Background'. Also, the meaning of '.io' at the end of the second sentence of the 'Nucleic acid extraction' section;

3. I suggest changing the section from 'Nucleic acid extraction' to 'DNA extraction', since only DNA was sequenced in the work;

4. Finally, I suggest giving more attention to the chromosome composition of the species. Since details about the species taxonomy, morphology, ecology, and distribution were so well presented. What is the morphology of the chromosomes of this species? Is there any polymorphism reported for this species? At least for the X chromosome the answer is yes, see: Chassovnikarova T, et al., 2015 (Ref 1)

References

1. Chassovnikarova T, Rovatsos M, Atanasov N, Koshev Y: Sex chromosome variability of Spermophilus citellus (Linnaeus, 1766) in the Southeastern part of the Balkan Peninsula. *Mammalian Biology*. 2015; **80** (4): 365-371 Publisher Full Text

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: Genome evolution, chromosome rearrangements, repetitive DNA

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.