











Data Paper (Biosciences)

Author-formatted document posted on 26/03/2025

Published in a RIO article collection by decision of the collection editors.

DOI: <https://doi.org/10.3897/arphapreprints.e153920>

ERGA-BGE genome of *Acomys minous* (Bate, 1906): the Crete spiny mouse, endemic to the island of Crete, Greece

 Petros Lymberakis,  Danae Karakasi, Manolis Papadimitrakis,  Rita Monteiro,  Astrid Böhne, Rosa Fernández, Nuria Escudero,  Jean-Marc Aury,  Alice Moussy, Corinne Cruaud, Karine Labadie, Sophie Mangenot, Caroline Belser, Lola Demirdjian,  Swati Sinha,  Leanne Haggerty, Fergal Martin, Patrick Wincker,  Pedro H. Oliveira,  Tom Brown

GENOME REPORT

ERGA-BGE genome of *Acomys minous* (Bate, 1906): the Crete spiny mouse, endemic to the island of Crete, Greece

Petros Lymberakis¹, Danae Karakasi^{1,2}, Manolis Papadimitrakis¹, Rita Monteiro³, Astrid Böhne³, Rosa Fernández⁴, Nuria Escudero⁴, Genoscope Sequencing Technical Team⁵, Alice Moussy⁵, Corinne Cruaud⁵, Karine Labadie⁵, Sophie Mangenot⁵, Caroline Belser⁶, Lola Demirdjian⁶, Swati Sinha⁷, Leanne Haggerty⁷, Fergal Martin⁷, Patrick Wincker⁶, Pedro H. Oliveira⁶, Jean-Marc Aury⁶, Tom Brown^{8,9*}

¹ Natural History Museum of Crete, School of Sciences and Engineering, University of Crete, Greece

² Department of Biology, School of Sciences and Engineering, University of Crete, Vassilika Vouton, Heraklion, GR-70013, Greece

³ Leibniz Institute for the Analysis of Biodiversity Change - Museum Koenig Bonn, Adenauerallee 127, 53113 Bonn, Germany

⁴ Metazoa Phylogenomics Lab, Institute for Evolutionary Biology (CSIC-UPF). Passeig marítim de la Barceloneta 37-49. 08003 Barcelona, Spain

⁵ Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, 91057, France

⁶ Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France

⁷ European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK

⁸ Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Straße 17, 10315 Berlin, Germany

⁹ Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Koenigin-Luise-Str 6-8, 14195 Berlin, Germany

* To whom correspondence should be addressed: brown@izw-berlin.de

Abstract

The *Acomys minous* reference genome offers a crucial resource for uncovering phylogenetic relationships within the genus and its complex phylogeographic history. The entirety of the genome sequence was assembled into 20 contiguous chromosomal pseudomolecules. This chromosome-level assembly encompasses 2.35 Gb, composed of 297 contigs and 113 scaffolds, with contig and scaffold N50 values of 29.3 Mb and 113 Mb, respectively.

Keywords

Acomys minous, genome assembly, European Reference Genome Atlas, Biodiversity Genomics Europe, Earth Biogenome Project, Muridae family, Crete spiny mouse, Agathopontikos

Introduction

Acomys minous is a species likely introduced to Crete during the Middle Pleistocene (Barome et al., 2001). There has been a long controversy on its taxonomic status, namely whether it should be considered as a distinct species or a subspecies (*A. cahirinus minous*) of the North-African *A. cahirinus* (Aghová et al., 2019; Giagia-Athanasopoulou et al., 2011; Kryštufek et al., 2009; Wilson et al., 2017).

The species has been recently assessed in Greece's regional Red Data list as Least Concern (Nikolaus, 2024).

Being an introduced species notwithstanding, *A. minous* has integrated into the environment of Crete. It is part of the diet of many carnivores, mainly avian, and it consumes both plants and invertebrate animals (Paragamian & Zivanovic, 1989; Renaud et al., 2020; Wilson et al., 2017).

Despite the importance of species from this genus alongside its large distribution and abundance in local communities, the phylogeny and the species limits in the genus are poorly resolved (Aghová et al., 2019). A high-quality reference genome for *A. minous* will enhance research for drawing a clearer picture on the phylogenetic and phylogeographic relationships within the genus *Acomys*.

The generation of this reference resource was coordinated by the European Reference Genome Atlas (ERGA) initiative's Biodiversity Genomics Europe (BGE) project, supporting ERGA's aims of promoting transnational cooperation to promote advances in the application of genomics technologies to protect and restore biodiversity (Mazzoni et al., 2023).

Materials & Methods

ERGA's sequencing strategy includes Oxford Nanopore Technology (ONT) and/or Pacific Biosciences (PacBio) for long-read sequencing, along with Hi-C sequencing for chromosomal architecture, Illumina Paired-End (PE) for polishing (i.e. recommended for ONT-only assemblies), and RNA sequencing for transcriptomic profiling, to facilitate genome assembly and annotation.

Sample and Sampling Information

Manolis Papadimitrakis sampled one specimen of female *Acomys minous*, identified by Petros Lymberakis, determined based on expert assessment, from Almyros Gorge, Irakleio, Crete, Greece on 19 May 2019. Sampling was performed under Presidential Decree 67/1981 issued by The Greek Government in Athens. Sampling was performed using traps. The specimen was euthanized by increasing concentration of CO₂ and subsequently frozen at -80C. Until DNA extraction, samples were preserved at -80C.

Vouchering information

Physical reference materials for the here sequenced specimen have been deposited in Natural History Museum of Crete, University of Crete (<https://www.nhmc.uoc.gr/en/departments/vertebrates>) under the accession number NHMC80.5.40.460.

Frozen reference tissue material of muscle is available from the same individual at the Biobank Natural History Museum of Crete, University of Crete (<https://www.nhmc.uoc.gr/en/departments/vertebrates>) under the voucher ID NHMC.80.5.40.460.

Data Availability

A. minous and the related genomic study were assigned to Tree of Life ID (ToLID) 'mAcoMin1'

and all sample, sequence, and assembly information are available under the umbrella BioProject PRJEB77214. The sample information is available at the following BioSample accessions: SAMEA112751364, SAMEA112751366, SAMEA112751371, SAMEA112751373 and SAMEA112751376. The genome assembly is accessible from ENA under accession number GCA_964271855.1. Sequencing data produced as part of this project are available from ENA at the following accessions: ERR12712697, ERR13351238, ERR13351239 and ERR13362573. Documentation related to the genome assembly and curation can be found in the ERGA Assembly Report (EAR) document available at https://github.com/ERGA-consortium/EARs/tree/main/Assembly_Reports/Acomys_minous/mAcoMin1. Further details and data about the project are hosted on the ERGA portal at https://portal.erga-biodiversity.eu/data_portal/Acomys%20minous.

Genetic Information

The estimated genome size, based on ancestral taxa, is 2.58 Gb. This is a diploid genome with a haploid number of 20 chromosomes ($2n=40$) and XY sex chromosomes. All information for this species was retrieved from Genomes on a Tree (Challis et al., 2023).

DNA/RNA processing

DNA was extracted from 200 mg of muscle tissue using a Genomic-tip 100/G kit (QIAGEN, MD, USA) following manufacturer's instructions. DNA fragment size selection was performed using Short Read Eliminator (PacBio, CA, USA). Quantification was performed using a Qubit dsDNA HS Assay kit (Thermo Fisher Scientific) and integrity was assessed in a FemtoPulse system (Agilent). DNA was stored at 4 °C until usage.

RNA was extracted from muscle (50 mg) using the RNeasy Plus Universal kit (Qiagen) following manufacturer instructions. Residual genomic DNA was removed with 6U of TURBO DNase (2 U/ μ L) (Thermo Fisher Scientific). Quantification was performed using a Qubit RNA HS Assay kit and

integrity was assessed in a Bioanalyzer system (Agilent). RNA was stored at -80 °C.

Library Preparation and Sequencing

Long-read DNA libraries were prepared with the SMRTbell prep kit 3.0 following manufacturers' instructions and sequenced on a Revio system (PacBio). Hi-C libraries were generated from muscle tissue using the Arima High Coverage HiC kit (following the Animal Tissues low input protocol v01) and sequenced on a NovaSeq 6000 instrument (Illumina) with 2x150 bp read length. Poly(A) RNA-Seq libraries were constructed using the Illumina Stranded mRNA Prep, Ligation kit (Illumina) and sequenced on a NovaSeq X+ instrument (Illumina). In total 77.3 Gb PacBio HiFi, 211.3 Gb Illumina WGS, and 43.4 Gb HiC data were sequenced to generate the assembly.

Genome Assembly Methods

The genome of *Acomys minous* was assembled using the Genoscope GALOP pipeline (<https://workflowhub.eu/workflows/1200>).

Briefly, raw PacBio HiFi reads were assembled using Hifiasm v0.19.5-r593. Retained haplotigs were removed using purge_dups v1.2.5 with default parameters and the proposed cutoffs. The purged assembly was scaffolded using YaHS v1.2 and assembled scaffolds were then curated through manual inspection using PretextView v0.2.5 to remove false joins and incorporate sequences not automatically scaffolded into their respective locations within the chromosomal pseudomolecules. The mitochondrial genome was assembled as a single circular contig using Oatk v1.0 and included in the released assembly. Summary analysis of the released assembly was performed using the ERGA-BGE Genome Report ASM Galaxy workflow (<https://doi.org/10.48546/workflowhub.workflow.1104.1>).

Genome Annotation Methods

A gene set was generated using the Ensembl Gene Annotation system (Aken et al., 2016), primarily by aligning publicly available short-read RNA-seq data from BioSample: SAMEA112751366 to the

genome. Gaps in the annotation were filled via protein-to-genome alignments of a select set of vertebrate proteins from UniProt (“UniProt: A Worldwide Hub of Protein Knowledge,” 2019), which had experimental evidence at the protein or transcript level. At each locus, data were aggregated and consolidated, prioritising models derived from RNA-seq data, resulting in a final set of gene models and associated non-redundant transcript sets. To distinguish true isoforms from fragments, the likelihood of each open reading frame (ORF) was evaluated against known vertebrate proteins. Low-quality transcript models, such as those showing evidence of fragmented ORFs, were removed (thresholds needed). In cases where RNA-seq data were fragmented or absent, homology data were prioritised, favouring longer transcripts with strong intron support from short-read data. The resulting gene models were classified into three categories: protein-coding, pseudogene, and long non-coding. Models with hits to known proteins and few structural abnormalities were classified as protein-coding. Models with hits to known proteins but displaying abnormalities, such as the absence of a start codon, non-canonical splicing, unusually small intron structures (<75 bp), or excessive repeat coverage, were reclassified as pseudogenes. Single-exon models with a corresponding multi-exon copy elsewhere in the genome were classified as processed (retrotransposed) pseudogenes. Models that did not fit any of the previously described categories did not overlap protein-coding genes, and were constructed from transcriptomic data were considered potential lncRNAs. Potential lncRNAs were further filtered to remove single-exon loci due to their unreliability. Putative miRNAs were predicted by performing a BLAST search of miRBase (Kozomara et al., 2019) against the genome, followed by RNAfold analysis (Gruber et al., 2008). Other small non-coding loci were identified by scanning the genome with Rfam (Kalvari et al., 2018) and passing the results through Infernal (Nawrocki & Eddy, 2013).

Summary analysis of the released annotation was carried out using the ERGA-BGE Genome Report ANNOT Galaxy workflow (<https://doi.org/10.48546/workflowhub.workflow.1096.1>).

Results

Genome Assembly

The genome assembly has a total length of 2,350,237,540 bp in 113 scaffolds including the mitogenome (Figures 1 & 2), with a GC content of 42.87%. The assembly has a contig N50 of 29,281,624 bp and L50 of 23 and a scaffold N50 of 122,851,966 bp and L50 of 9. The assembly has a total of 184 gaps, totaling 20.1 kb in cumulative size. The single-copy gene content analysis using the Glires database with BUSCO (Manni et al., 2021) resulted in 95.3% completeness (94.0% single and 1.3% duplicated). 92.44% of reads k-mers were present in the assembly and the assembly has a base accuracy Quality Value (QV) of 59.0 as calculated by Merqury (Rhie et al., 2020).

Genome Annotation

The genome annotation consists of 19,865 protein-coding genes with an associated 25,939 transcripts, in addition to 3,561 non-coding genes (Table 1). Using the longest isoform per transcript, the single-copy gene content analysis using the Glires database with BUSCO resulted in 97.9% completeness. Using the OMamer Myomorpha database for OMArk (Nevers et al., 2024) resulted in 97.93% completeness and 97.8% consistency (Table 2).

Table 1. Statistics from assembled gene models

	No. genes	No. transcripts	Mean gene length (bp)	No. single-exon genes	Mean exons per transcript
mRNA	19,865	25,939	40,353	2,501	10.4
pseudogene	416	416	11,537	18	9.7
snoRNA	1,455	1,455	118	1,455	1
lncRNA	111	121	2,456	81	1.6
miRNA	67	67	83	67	1
snRNA	1,235	1,235	118	1,235	1
rRNA	118	118	354	118	1
scRNA	44	44	160	44	1
Other ncRNA	115	115	106-271	115	1

Table 2. Annotation completeness and consistency scores calculated by BUSCO run in protein mode (glires_odb10) and OMArk (Myomorpha)

	Complete	Singular	Duplicated	Fragmented	Missing
BUSCO	13,514 (97.9%)	13,383 (97.0%)	131 (0.9%)	44 (0.3%)	240 (1.8%)
OMArk	15,571 (97.93%)	15,243 (95.87%)	328 (2.06%)	-	329 (2.07%)
	Consistent	Inconsistent	Contaminants	Unknown	
OMArk	19,653 (97.80%)	345 (1.72%)	0 (0.0%)	97 (0.48%)	

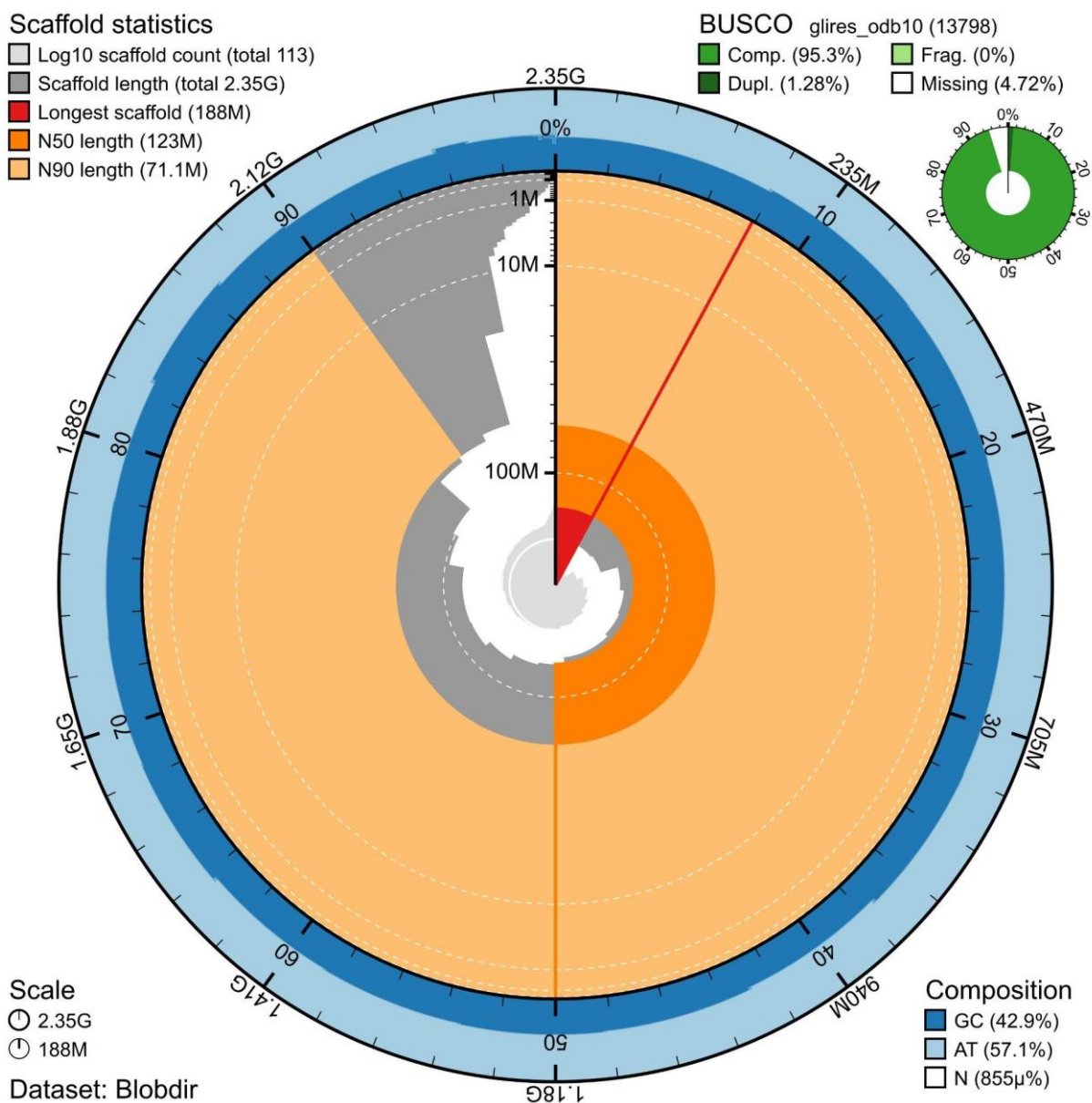


Figure 1. Snail plot summary of assembly statistics. The main plot is divided into 1,000 size-ordered bins around the circumference, with each bin representing 0.1% of the 2,350,237,540 bp assembly. The distribution of sequence lengths is shown in dark grey, with the plot radius scaled to the longest sequence present in the assembly (187,704,340 bp, shown in red). Orange and pale-orange arcs show the scaffold N50 and N90 sequence lengths (122,851,966 and 71,069,054 bp), respectively. The pale grey spiral shows the cumulative sequence count on a log-scale, with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT, and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated, and missing BUSCO genes found in the assembled genome from the Glires database (odb10) is shown in the top right.

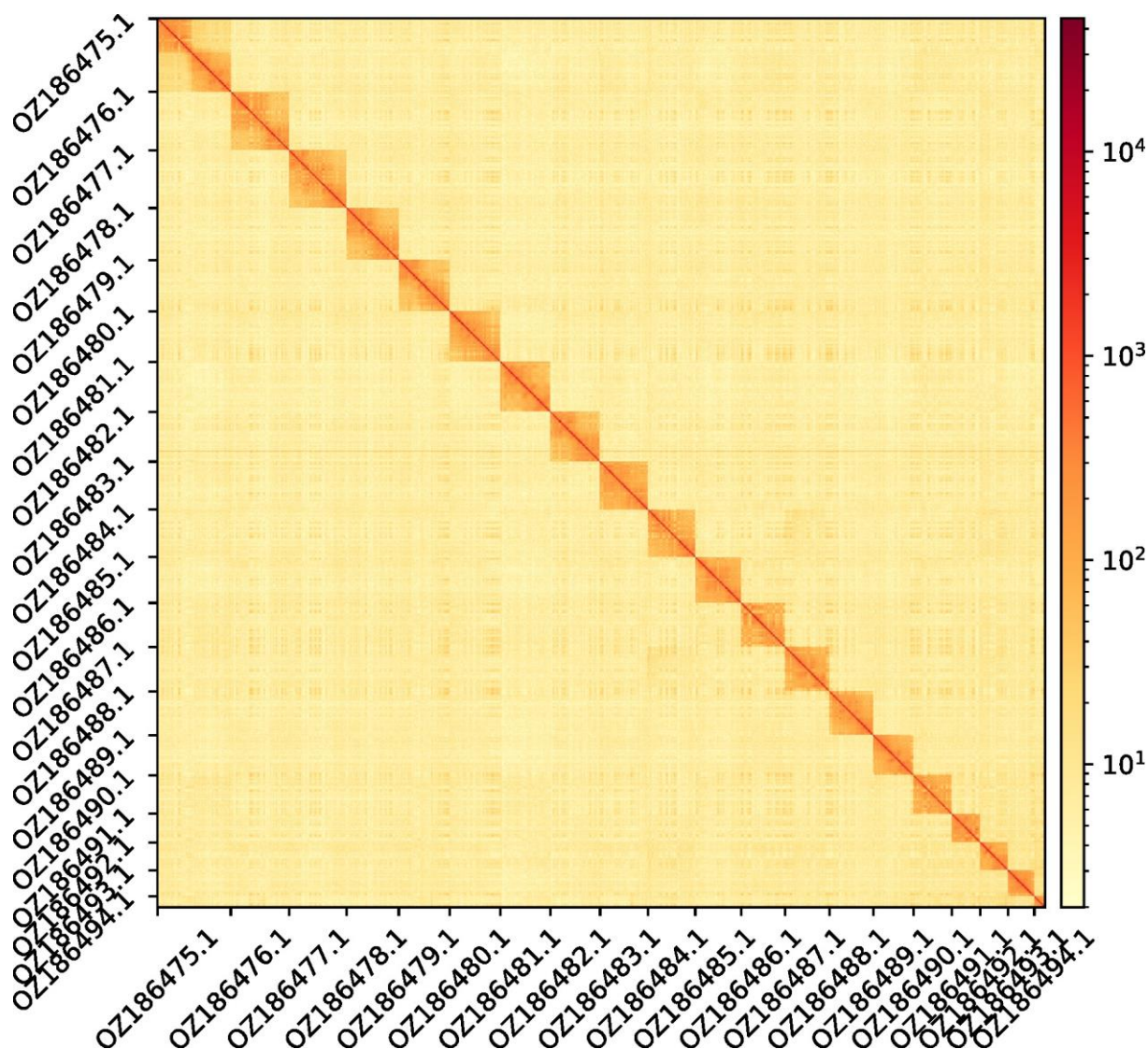


Figure 2. Hi-C contact map showing spatial interactions between regions of the genome. The diagonal corresponds to intra-chromosomal contacts, depicting chromosome boundaries. The frequency of contacts is shown on a logarithmic heatmap scale. Hi-C matrix bins were merged into a 25 kb bin size for plotting.

Acknowledgements

We would like to acknowledge the assembly reviewer, Tyler Alioto, from the Centro Nacional de Análisis Genómico (CNAG). The authors acknowledge the support of the Freiburg Galaxy Team: Saim Momin and Björn Grüning, Bioinformatics, University of Freiburg (Germany), funded by the German Federal Ministry of Education and Research BMBF grant 031 A538A de.NBI-RBC and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) within the framework of LIBIS/de.NBI Freiburg.

Conflict of Interest

The authors declare no conflict of interest related to this study. The funding sources had no involvement in the study design, collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to submit the article for publication. All authors have participated sufficiently in the work to take public responsibility for the content and agree to the submission of this manuscript.

Funder Information

This project received funding from Horizon Europe under the Biodiversity, Circular Economy and Environment (REA.B.3); co-funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract numbers 22.00173 and 24.00054; and by the UK Research and Innovation (UKRI) under the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme. This work was supported by the Genoscope, the Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA), France Génomique (ANR-10-INBS-09-08), and the exploratory research programme 'ATLASEa: Atlas of marine genomes' and its targeted project SEQ-Sea (ANR-22-EXAT-0003-SEQ-Sea)..

Author Contributions

MP and PL collected the species, PL identified the species, DK and MP sampled and preserved biological material and provided metadata, RM and AS provided sampling and metadata support and management, the Genomescope Sequencing Team extracted DNA, prepared libraries, and performed sequencing under the supervision of AM, CC, KL, PHO and PW, LD SM, CB and JMA performed genome assembly and curation, TB generated the analysis and report. All authors contributed to the writing, review, and editing of this genome note and read and approved the final version. This work is part of the species assigned to Genoscope, which was instrumental in the wet lab, sequencing, and assembly processes, and represents a key contribution to BGE's outputs

Author Information

Members of the Genoscope Sequencing Technical Team are listed here: <https://doi.org/10.5281/zenodo.14611490>

Literature Cited

- Aghová, T., Palupčíková, K., Šumbera, R., Frynta, D., Lavrenchenko, L. A., Meheretu, Y., Sádlová, J., Votýpka, J., Mbaou, J. S., Modrý, D., & Bryja, J. (2019). Multiple radiations of spiny mice (Rodentia: Acomys) in dry open habitats of Afro-Arabia: evidence from a multi-locus phylogeny. *BMC Evolutionary Biology*, *19*(1), 69. <https://doi.org/10.1186/s12862-019-1380-9>
- Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., Howe, K., Kähäri, A., Kokocinski, F., Martin, F. J., Murphy, D. N., Nag, R., Ruffier, M., Schuster, M., Tang, Y. A., ... Searle, S. M. J. (2016). The Ensembl gene annotation system. *Database*, *2016*, baw093. <https://doi.org/10.1093/database/baw093>
- Barome, P.-O., Lymberakis, P., Monnerot, M., & Gautun, J.-C. (2001). Cytochrome b Sequences Reveal *Acomys minous* (Rodentia, Muridae) Paraphyly and Answer the Question about the Ancestral Karyotype of *Acomys dimidiatus*. *Molecular Phylogenetics and Evolution*, *18*(1), 37–46. <https://doi.org/10.1006/mpev.2000.0859>
- Challis, R., Kumar, S., Sotero-Caio, C., Brown, M., & Blaxter, M. (2023). Genomes on a Tree (GoaT): A versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life. *Wellcome Open Research*, *8*, 24. <https://doi.org/10.12688/wellcomeopenres.18658.1>
- Giagia-Athanasopoulou, E. B., Rovatsos, M. T. H., Mitsainas, G. P., Martimianakis, S., Lymberakis, P., Angelou, L.-X. D., Marchal, J. A., & Sánchez, A. (2011). New data on the evolution of the Cretan spiny mouse, *Acomys minous* (Rodentia: Murinae), shed light on the phylogenetic relationships in the cahirinus group: NEW DATA ON THE EVOLUTION OF A. MINOUS. *Biological Journal of the Linnean Society*, *102*(3), 498–509. <https://doi.org/10.1111/j.1095-8312.2010.01592.x>
- Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., & Hofacker, I. L. (2008). The Vienna RNA Websuite. *Nucleic Acids Research*, *36*(suppl_2), W70–W74. <https://doi.org/10.1093/nar/gkn188>
- Kalvari, I., Nawrocki, E. P., Argasinska, J., Quinones-Olvera, N., Finn, R. D., Bateman, A., & Petrov, A. I. (2018). Non-Coding RNA Analysis Using the Rfam Database. *Current Protocols in Bioinformatics*, *62*(1), e51. <https://doi.org/10.1002/cpbi.51>

- Kiamos, N. (2024). *The Greek Red List of Threatened Species (2024)*. redlist.necca.gov.gr
- Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). miRBase: From microRNA sequences to function. *Nucleic Acids Research*, *47*(D1), D155–D162. <https://doi.org/10.1093/nar/gky1141>
- Kryštufek, B., Vohralík, V., Hošek, J., & Janžekovič, F. (with Grimmberger, E., & Spitzenberger, F.). (2009). *Mammals of Turkey and Cyprus. Rodentia II: Cricetinae, Muridae, Spalacidae, Calomyscidae, Capromyidae, Hystricidae, Castoridae*. Univerza na Primorskem.
- Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. <https://doi.org/10.1093/molbev/msab199>
- Mazzoni, C. J., Ciofi, C., & Waterhouse, R. M. (2023). Biodiversity: An atlas of European reference genomes. *Nature*, *619*(7969), 252–252. <https://doi.org/10.1038/d41586-023-02229-w>
- Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933–2935. <https://doi.org/10.1093/bioinformatics/btt509>
- Nikolaus, K. (2024). *The Greek Red List of Threatened Species (2024)*. redlist.necca.gov.gr
- Paragamian, K. V., & Zivanovic, S. (1989). Preliminary results of the examination of barn owl (*Tyto alba*) food pellets from two caves in central Crete, Greece. *Bulletin De La Société Spéléologique De Grèce*, *XX*, 95–97.
- Renaud, S., Hardouin, E. A., Chevret, P., Papayiannis, K., Lymberakis, P., Matur, F., Garcia-Rodriguez, O., Andreou, D., Çetintaş, O., Sözen, M., Hadjisterkotis, E., & Mitsainas, G. P. (2020). Morphometrics and genetics highlight the complex history of Eastern Mediterranean spiny mice. *Biological Journal of the Linnean Society*, *130*(3), 599–614. <https://doi.org/10.1093/biolinnean/blaa063>
- Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, *21*(1), 245. <https://doi.org/10.1186/s13059-020-02134-9>
- UniProt: A worldwide hub of protein knowledge. (2019). *Nucleic Acids Research*, *47*(D1), D506–D515. <https://doi.org/10.1093/nar/gky1049>

Wilson, D. E., Mittermeier, R. A., & Cavallini, P. (Eds.). (2017). *Handbook of the mammals of the world*.

Lynx Edicions : Conservation International : IUCN.

Scaffold statistics

- Log10 scaffold count (total 113)
- Scaffold length (total 2.35G)
- Longest scaffold (188M)
- N50 length (123M)
- N90 length (71.1M)

BUSCO glires_odb10 (13798)

- Comp. (95.3%)
- Frag. (0%)
- Dupl. (1.28%)
- Missing (4.72%)

