Check for updates

DATA NOTE

ERGA-BGE Genome of the viperine water snake (*Natrix maura*): a key species for evolutionary insights of freshwater snakes

[version 1; peer review: 2 approved]

Salvador Carranza[1], Daniel Fernández-Guiberteau[2], Laura Blasón[2], Sergi Tulloch[1], Rita Monteiro[3], Astrid Böhne[3], Rosa Fernández[1], Nuria Escudero[1], Laura Aguilera[4,5], Marta Gut[4,5], Tyler S Alioto[4,5], Francisco Câmara Ferreira[4,5], Jèssica Gómez-Garrido[4,5], Fernando Cruz[4,5], Swati Sinha[6], Leanne Haggerty[6], Fergal Martin[6], Tom Brown[7,8]

[1]Institute of Evolutionary Biology (IBE, CSIC-Universitat Pompeu Fabra), Barcelona, Spain
[2]CREAC, Centre de Recerca i Educació Ambiental de Calafell (GRENP-Ajuntament de Calafell), Tarragona, Spain
[3]Leibniz Institute for the Analysis of Biodiversity Change, Museum Koenig Bonn, Bonn, Germany
[4]Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain
[5]Universitat de Barcelona (UB), Barcelona, Spain
[6]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK
[7]Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany
[8]Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Berlin, Germany

v1 **First published:** 04 Aug 2025, **5**:215
https://doi.org/10.12688/openreseurope.20863.1
Latest published: 04 Aug 2025, **5**:215
https://doi.org/10.12688/openreseurope.20863.1

Abstract

The reference genome of the freshwater snake, *Natrix maura,* offers a crucial resource for uncovering the genetic basis of adaptability to freshwater environments, while providing insights into hybridization and gene flow within the genus *Natrix*. A total of 18 contiguous chromosomal pseudomolecules were assembled from the genome sequence, corresponding to the 16 autosomes and 2 sex chromosomes. This chromosome-level assembly encompasses 1.7 Gb, composed of 284 contigs and 209 scaffolds, with contig and scaffold N50 values of 36.5 Mb and 174.4 Mb, respectively.

Keywords

Natrix maura, genome assembly, European Reference Genome Atlas, Biodiversity Genomics Europe, Earth Biogenome Project, viperine water snake

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 04 Aug 2025	✓ view	✓ view

1. **Yannis Schöneberg**, Universitat Trier Biogeographie, Trier, Germany

2. **Zexian Zhu**, Zhejiang University School of Medicine, Hangzhou, China

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

This article is included in the Horizon Europe gateway.

This article is included in the Genome Reports from the Biodiversity Genomics Europe Project collection.

Corresponding author: Tom Brown (brown@izw-berlin.de)

Author roles: Carranza S: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Fernández-Guiberteau D**: Investigation, Resources; **Blasón L**: Investigation, Resources, Writing – Review & Editing; **Tulloch S**: Investigation, Resources, Writing – Review & Editing; **Monteiro R**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Böhne A**: Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing; **Fernández R**: Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing; **Escudero N**: Methodology, Project Administration, Supervision, Writing – Review & Editing; **Aguilera L**: Investigation, Writing – Review & Editing; **Gut M**: Investigation, Supervision, Writing – Review & Editing; **Alioto TS**: Data Curation, Formal Analysis, Supervision, Writing – Review & Editing; **Câmara Ferreira F**: Data Curation, Formal Analysis, Writing – Review & Editing; **Gómez-Garrido J**: Data Curation, Formal Analysis, Writing – Review & Editing; **Cruz F**: Data Curation, Formal Analysis, Writing – Review & Editing; **Sinha S**: Data Curation, Formal Analysis, Writing – Review & Editing; **Haggerty L**: Data Curation, Formal Analysis, Writing – Review & Editing; **Martin F**: Data Curation, Formal Analysis, Supervision, Writing – Review & Editing; **Brown T**: Visualization, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: Biodiversity Genomics Europe (Grant no.101059492) is funded by Horizon Europe under the Biodiversity, Circular Economy and Environment call (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 research was also made possible by funding from grant PID2021-128901NB-100 MCIN/AEI/10.13039/501100011033 and by ERDF, a way of making Europe. ST was funded by an AGAUR-FI Joan Oró grant from the Department of Research and Universities of the Government of Catalonia and the European Social Fund Plus (2024_FI-1_01035). Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III and the 2014–2020 Smart Growth Operating Program, co-financed with the European Regional Development Fund (MINECO/FEDER, BIO2015-71792-P).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2025 Carranza S *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Carranza S, Fernández-Guiberteau D, Blasón L *et al.* **ERGA-BGE Genome of the viperine water snake (*Natrix maura*): a key species for evolutionary insights of freshwater snakes [version 1; peer review: 2 approved]** Open Research Europe 2025, **5**:215 https://doi.org/10.12688/openreseurope.20863.1

First published: 04 Aug 2025, **5**:215 https://doi.org/10.12688/openreseurope.20863.1

Introduction

Natrix maura, a member of the Colubridae family, is a medium-sized, moderately robust freshwater snake that typically grows to just under one meter in length. It has a well-defined triangular head, a relatively thick body, and a short tail. The snake often features a distinctive dorsal zigzag stripe running along its body to the tip of the tail, which, along with its morphological and behavioral traits, can lead to confusion with viper species found in the western Mediterranean (Speybroeck *et al.*, 2016). This is an example of Batesian mimicry, where a harmless species imitates the appearance and actions of a dangerous one to avoid predation (Santos, 2021). *N. maura* is widely distributed across the western Mediterranean region, ranging from Liguria in Italy, through central and southern France, across the entire Iberian Peninsula, and into the Mediterranean-climate zones of northwestern Africa, including Morocco, Tunisia, and Algeria. Introduced to Sardinia, Menorca, Mallorca and Corsica (Speybroeck *et al.*, 2016).

N. maura is currently classified as 'Least Concern' on the IUCN Red List (Mateo Miras *et al.*, 2009) in view of its wide distribution, presumed large population, and because it is unlikely to be declining fast enough to qualify for listing in a more threatened category.

N. maura thrives in aquatic environments such as rivers, lakes, and wetlands, particularly in warm, Mediterranean habitats (Speybroeck *et al.*, 2016). These environments provide ideal conditions for the snake, where it hunts for prey such as amphibians and small fish. Additionally, *N. maura* plays an important ecological role as prey for larger predators, supporting the overall biodiversity and health of these ecosystems, contributing to the balance of species within its natural habitats (Santos, 2021).

Developing a high-quality reference genome for *N. maura* is essential for advancing our understanding of its unique genetic adaptations to freshwater environments. This genomic resource will not only enhance our knowledge of its physiological and ecological traits but also provide valuable insights into the evolutionary origins of venom in snakes. Additionally, it will offer broader implications for research on snake evolution and adaptation, contributing to our understanding of biodiversity and their ecological roles (Schöneberg *et al.*, 2023).

N. maura may host parasites like trematodes, cestodes, and nematodes. These parasites are primarily found in the intestines but can also affect other parts of the digestive and respiratory systems, such as the pharynx, cloaca, trachea, and lungs (Santos, 2021). The snake's varied diet and habitat contribute to its high parasite diversity, with some parasites being acquired from common prey like amphibians (Santos, 2021).

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). This species falls within the regional reach of the Catalan Initiative for the Earth BioGenome Project (CBP), which is linked to ERGA (Corominas *et al.*, 2024).

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

On 13th October 2023, an adult, female specimen of *Natrix maura* (Figure 1) was sampled by Salvador Carranza and Daniel Fernández-Guiberteau. The species identification via morphology based on Santos 2021 and barcoding was confirmed by Salvador Carranza, Daniel Fernández-Guiberteau and Laura Blasón. The specimen was collected by hand from the Siurana River, Poboleda, Tarragona, Spain. Sampling was performed under permit number SF/0179/23 issued to Salvador Carranza by the Generalitat de Catalunya. Euthanasia was performed by first administering anesthetic agents (alfaxalone [Alfaxan, 10 mg/kg intramuscular]), and once the animal was fully anesthetized, the spinal cord was severed at the neck to ensure death. The specimen's tissues were snap-frozen in liquid nitrogen immediately after harvesting and stored at -80°C until DNA extraction.

Vouchering information

Physical reference materials for the here sequenced specimen have been deposited in Museo Nacional de Ciencias Naturales (MNCN) https://www.mncn.csic.es, under the accession numbers MNCN-ADN-151759 & MNCN-ADN-151760.

Frozen reference tissue material of muscle tissue is available from the same individual at the MNCN under the voucher ID MNCN52208.



Figure 1. Example image of *Natrix Maura*. Image of *Natrix maura* taken by Daniel Fernández-Guiberteau. Further voucher images of the here sequenced specimen are available at the BioImage Archive under accession number SAMEA114541135: https://www.ebi.ac.uk/biostudies/BioImages/studies/S-BIAD1012

Images of the here sequenced specimen are available at the BioImage Archive under the accession SAMEA114541135: https://www.ebi.ac.uk/biostudies/BioImages/studies/S-BIAD1012

Data availability

Genetic information

The estimated genome size, based on direct measurements retrieved from Genomes on a Tree (Challis *et al.*, 2023), is 2.3 Gb, while the estimation based on reads kmer profiling is 1.6 Gb. This is a diploid genome with a haploid number of 17 chromosomes (2n=34), including ZW sex chromosomes in female individuals (De Smet, 1978).

DNA/RNA processing

DNA was extracted from kidney tissue using the Blood & Cell Culture DNA Midi Kit (Qiagen) following the manufacturer's instructions. DNA quantification was performed using a Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific), and DNA integrity was assessed using a Genomic DNA 165 Kb Kit (Agilent) on the Femto Pulse system (Agilent). The DNA was stored at +4°C until used.

RNA was extracted using an RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. RNA was extracted from four different specimen parts: ovary, liver, pancreas and lung. RNA quantification was performed using the Qubit RNA BR kit and RNA integrity was assessed using a Bioanalyzer 2100 system RNA 6000 Nano Kit (Agilent). The RNA was stored at -80°C until used.

Library preparation and sequencing

For long-read whole genome sequencing, a library was prepared using the SQK-LSK114 Kit (Oxford Nanopore Technologies, ONT), which was then sequenced on a PromethION 24 A Series instrument (ONT). A short-read whole genome sequencing library was prepared using the KAPA Hyper Prep Kit (Roche). A Hi-C library was prepared from liver and lung using the ARIMA High Coverage Hi-C Kit (ARIMA), followed by the KAPA Hyper Prep Kit for Illumina sequencing (Roche). The RNA libraries were prepared using the KAPA mRNA Hyper prep kit (Roche). All the short-read libraries were sequenced on a NovaSeq 6000 instrument (Illumina).

In total, 52x Oxford Nanopore, 88x Illumina WGS shotgun, and 163x HiC data were sequenced to generate the assembly.

Genome assembly methods

The genome was assembled using the CNAG CLAWS pipeline v2.2.0 (Gomez-Garrido, 2024). Briefly, reads were preprocessed for quality and length using Trim Galore v0.6.7 and Filtlong v0.2.1, and initial contigs were assembled using NextDenovo v2.5.0, followed by polishing of the assembled contigs using HyPo v1.0.3, removal of retained haplotigs using purge-dups v1.2.6 and scaffolding with YaHS v1.2a. Finally, assembled scaffolds were curated via manual inspection using Pretext v0.2.5 with the Rapid Curation Toolkit (https://gitlab.com/wtsi-grit/rapid-curation) to remove any false joins and incorporate any sequences not automatically scaffolded into their respective locations in the chromosomal pseudomolecules (or super-scaffolds). The blobtoolkit nextflow pipeline v0.6.0 (https://pipelines.tol.sanger.ac.uk/blobtoolkit/0.6.0/usage) confirmed the absence of contaminants. Finally, the mitochondrial genome was assembled as a single circular contig of 21,237 bp using the FOAM pipeline v0.5 (https://github.com/cnag-aat/FOAM) and included in the released assembly (GCA_964659585.1). Summary analysis of the released assembly was performed using the ERGA-BGE Genome Report ASM Galaxy workflow (De Panis, 2024b).

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 BioSamples: SAMEA114541140, SAMEA114541142, SAMEA114541145, and SAMEA114541148 to the genome. Gaps in the annotation were filled via protein-to-genome alignments of a select set of vertebrate proteins from UniProt (The UniProt Consortium, 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. 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 performed using the ERGA-BGE Genome Report ANNOT Galaxy workflow (De Panis, 2024a)

Results

Genome assembly

The genome assembly has a total length of 1,738,090,269 bp in 210 scaffolds including the mitogenome (Figure 2 and

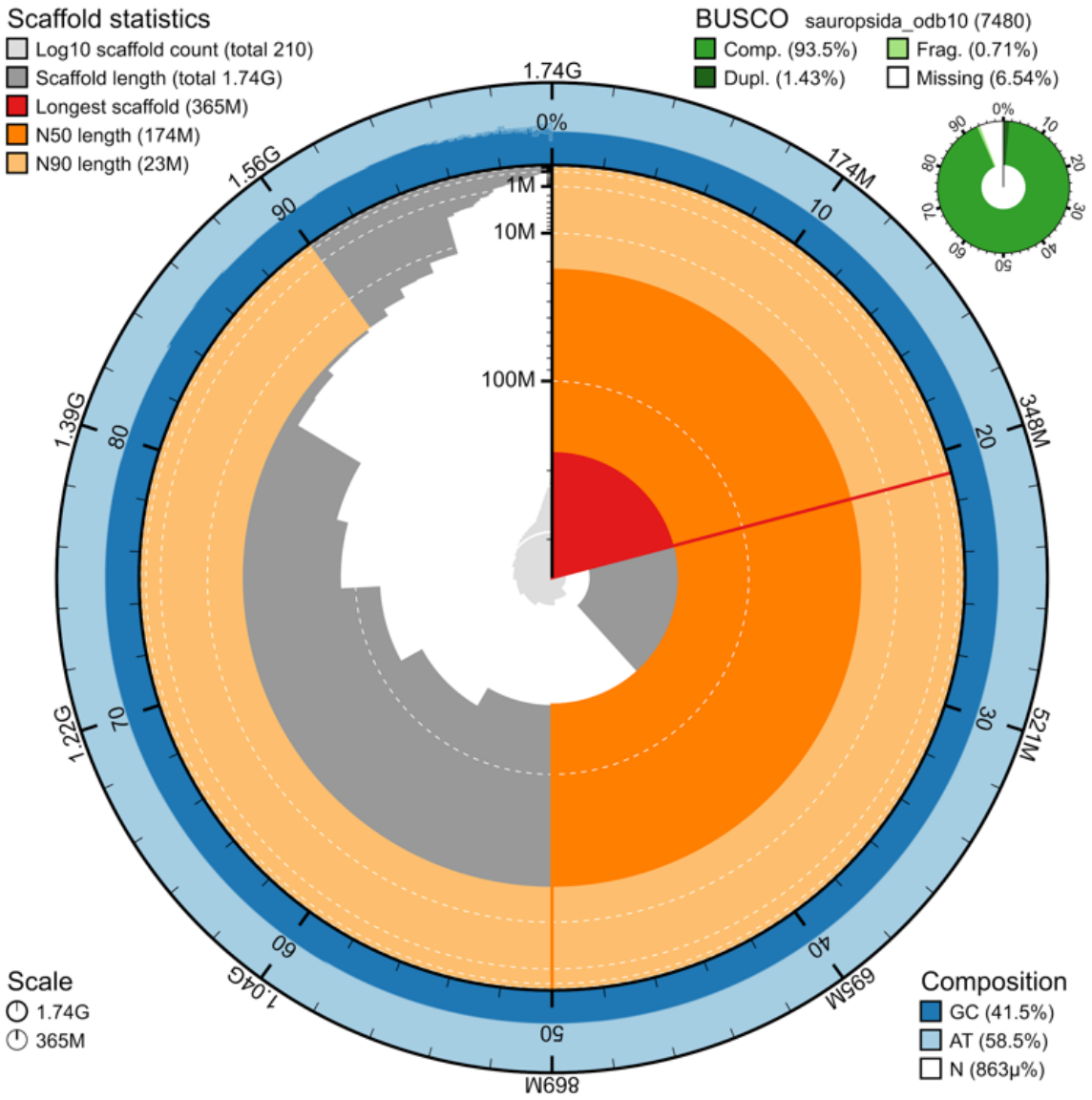

Figure 2. Snail plot summary of assembly statistics. The main plot is divided into 1,000 size-ordered bins around the circumference, with each bin representing 0.1% of the 1,738,090,269 bp assembly including the mitochondrial genome. The distribution of sequence lengths is shown in dark grey, with the plot radius scaled to the longest sequence present in the assembly (364,621,453 bp, shown in red). Orange and pale-orange arcs show the scaffold N50 and N90 sequence lengths (174,378,910 and 23,029,681 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 sauropsida database (odb10) is shown on the top right.

Figure 3), with a GC content of 41.45%. It features a contig N50 of 36,526,383 bp (L50=13) and a scaffold N50 of 174,378,910 bp (L50=4). There are 75 gaps, totaling 15 kb in cumulative size. The single-copy gene content analysis using the sauropsida_odb10 database with BUSCO v5.5.0 resulted in 93.4% completeness (92.0% single and 1.4% duplicated). 91.35% of reads k-mers were present in the assembly and the assembly has a base accuracy Quality Value (QV) of 46.28 as calculated by Merqury.

Genome annotation

The genome annotation consists of 17,402 protein-coding genes with an associated 51,854 transcripts, in addition to 27 pseudogenes and 1,491 non-coding RNA genes of various types (Table 1). Using the longest isoform per transcript, the single-copy gene content analysis using the sauropsida_odb10 database with BUSCO v5.5.0 resulted in 93.4% completeness. Using the OMAmer Episquamata database for OMArk resulted in 95.85% completeness and 97.59% consistency (Table 2).

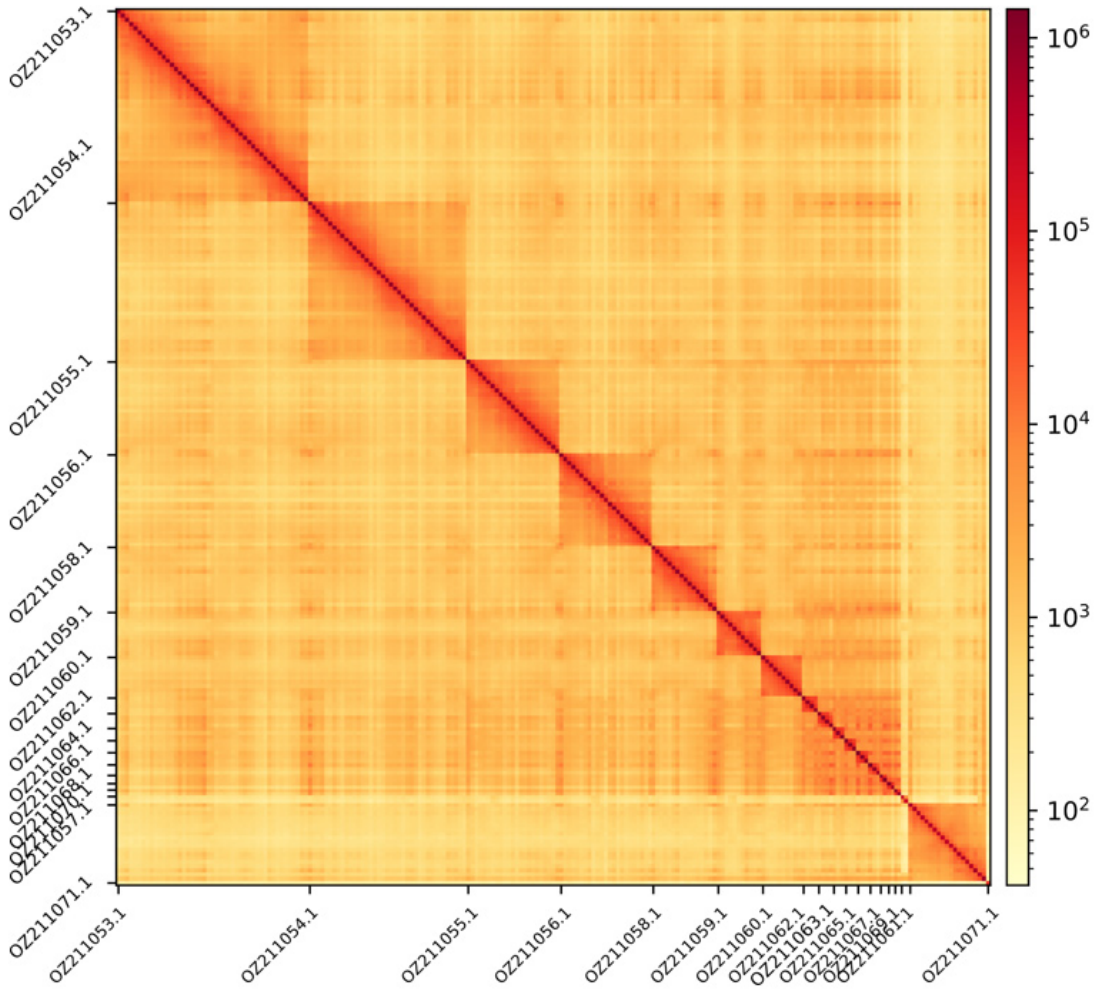

Figure 3. Hi-C contact map showing spatial interactions between regions of the genome. The diagonal corresponds to intra-chromosomal contacts, depicting chromosome boundaries. The frequency of contacts is shown on a logarithmic heatmap scale. Hi-C matrix bins were merged into a 150 kb bin size for plotting. Due to space constraints on the axes, only the GenBank names of the chromosomes are shown, with alternating names displayed for the micro-chromosomes on the x- and y-axes.

Table 1. Statistics from assembled gene models.

	No. genes	No. transcripts	Mean gene length (bp)	No. single-exon genes	Mean exons per transcript
Protein-coding	17,402	51,854	41,684	491	15.2
Pseudogenes	27	27	7,337	2	6.7
lncRNA	981	1,663	21,796	88	3.6
snRNA	162	162	124	162	1.0
snoRNA	197	197	116	197	1.0
rRNA	73	73	610	73	1.0
miRNA	52	52	86	52	1.0
scRNA	15	15	194	15	1.0
Other non-coding	11	11	183	11	1.0

*Combined categories show the range of the mean values

Table 2. Annotation completeness and consistency scores calculated by BUSCO run in protein mode (Sauropsida_odb10) and OMArk (Episquamata).

	Complete	Singular	Duplicated	Fragmented	Missing
BUSCO	93.4%	91.7%	1.7%	0.9%	5.7%
OMArk	95.85%	93.78%	2.06%	-	4.15%
	Consistent	**Inconsistent**	**Contaminants**	**Unknown**	
OMArk	97.59%	1.57%	0.00%	0.84%	

Data availability

Natrix maura and the related genomic study were assigned to Tree of Life ID (ToLID) 'rNatMau1' and all sample, sequence, and assembly information are available under the umbrella BioProject PRJEB77787. The sample information is available at the following BioSample accessions: SAMEA114541135, SAMEA114541139, SAMEA114541140, SAMEA114541142, SAMEA114541148. The genome assembly is accessible from ENA under accession number GCA_964659585.1 and the annotated genome is available through the Ensembl ERGA-BGE project page (https://projects.ensembl.org/erga-bge/) and Ensembl Genome data & annotation site (https://beta.ensembl.org/). Sequencing data produced as part of this project are available from ENA at the following BioProject: PRJEB77785. 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/Natrix_maura/rNatMau1. Further details and data about the project are hosted on the ERGA portal at https://portal.erga-biodiversity.eu/data_portal/8585.

Author contributions

SC & DF-G collected the species, SC, DF-G, ST & LB identified the species, SC, ST & DF-G sampled and preserved biological material and provided metadata, RM, AB, RF & NE provided sampling and metadata support and management, LA and MG extracted DNA, prepared libraries, and performed sequencing, FCF, JG-G & FC performed genome assembly and curation under the supervision of TSA, LH and FM performed genome annotation, 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.

We are deeply saddened by the passing of Daniel Fernández-Guiberteau, who died in May 2025 after a courageous battle with cancer. Daniel was a passionate biologist whose work made a lasting impact on the conservation of reptiles and amphibians, particularly in Catalonia. His dedication, knowledge, and commitment to fieldwork and outreach were truly inspiring. This manuscript is a small tribute to his memory and to the legacy he leaves in herpetological research and conservation.

Acknowledgements

We would like to express our gratitude to Jhulyana López Caro and Eduardo Cáceres for their invaluable assistance in the field and laboratory. We 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. We would like to acknowledge the assembly reviewer, Lola Demirdjian from Genoscope. We also acknowledge the support of the Generalitat de Catalunya through the Departament de Salut and the Departament d'Empresa i Coneixement.

References

Aken BL, Ayling S, Barrell D, *et al.*: **The ensembl gene annotation system.** *Database (Oxford).* 2016; **2016**: baw093.
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**

Corominas M, Marquès-Bonet T, Arnedo MA, *et al.*: **The catalan initiative for the earth biogenome project: contributing local data to global biodiversity genomics.** *NAR Genom Bioinform.* 2024; **6**(3): lqae075.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

De Panis D: **ERGA-BGE genome report ANNOT analyses**. WorkflowHub, 2024a.
Publisher Full Text

De Panis D: **ERGA-BGE genome report ASM analyses (one-asm WGS Illumina PE + HiC)**. WorkflowHub, 2024b.
Publisher Full Text

De Smet WHO: **The chromosomes of 23 species of snakes.** *Acta Zoologica et Pathologica Antverpiensia.* 1978; **80**: 85–118.
Reference Source

Gomez-Garrido J: **CLAWS (CNAG's long-read assembly workflow in Snakemake)**. WorkflowHub, 2024.
Publisher Full Text

Gruber AR, Lorenz R, Bernhart SH, *et al.*: **The Vienna RNA websuite.** *Nucleic Acids Res.* 2008; **36**(suppl_2): W70–W74.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Kalvari I, Nawrocki EP, Argasinska J, *et al.*: **Non-Coding RNA analysis using the rfam database.** *Curr Protoc Bioinformatics.* 2018; **62**(1): e51.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Kozomara A, Birgaoanu M, Griffiths-Jones S: **miRBase: from microRNA sequences to function.** *Nucleic Acids Res.* 2019; **47**(D1): D155–D162.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Mateo Miras JA, Cheylan M, Nouira MS, *et al.*: **Natrix maura: the IUCN red list of threatened species 2009**. (e. T61538A12510365) [Dataset], 2009.
http://www.doi.org/10.2305/IUCN.UK.2009.RLTS.T61538A12510365.en

Mazzoni CJ, Ciofi C, Waterhouse RM: **Biodiversity: an Atlas of European reference genomes.** *Nature.* 2023; **619**(7969): 252.
PubMed Abstract | **Publisher Full Text**

Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches.** *Bioinformatics.* 2013; **29**(22): 2933–2935.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Santos X: **Culebra viperina – Natrix maura. En: enciclopedia virtual de los vertebrados españoles**. (A Salvador & A Marco, Eds.). Museo Nacional de Ciencias Naturales, 2021.
Reference Source

Schöneberg Y, Winter S, Arribas O, *et al.*: **Genomics reveals broad hybridization in deeply divergent palearctic grass and water snakes (*Natrix* spp.).** *Mol Phylogenet Evol.* 2023; **184**: 107787.
PubMed Abstract | **Publisher Full Text**

Speybroeck J, Beukema W, Bok B, *et al.*: **Field guide to the amphibians and reptiles of Britain and Europe. British wildlife publishing field guides**. Bloomsbury Publishing, 2016.
Reference Source

UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–D515.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 18 September 2025

https://doi.org/10.21956/openreseurope.22571.r59863

© **2025 Zhu Z.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Zexian Zhu**
Zhejiang University School of Medicine, Hangzhou, China

General comment:
The authors present a high-quality, chromosome-level reference genome of Natrix maura, accompanied by a robust annotation. This genome will serve as an important resource for evolutionary, ecological, and conservation studies. The sequencing and assembly strategies follow standard protocols, and the results demonstrate good contiguity and completeness. I commend the authors for generating such a valuable dataset and genome report.
A few aspects would benefit from clarification and additional detail to improve reproducibility and reader understanding.

Detailed comments

1. The introduction emphasizes mimicry and distribution but less on the phylogenomic and comparative significance of this genome. It would be useful to explicitly connect N. maura with ongoing debates in snake evolution (e.g., sex chromosome evolution, repeated hybridization in Natrix, or genome size variation in colubrids)
2. The assembly metrics are excellent. However, the presentation is quite heavy. A concise comparative table (with other snake genomes' statistics) would make it easier to read.
3. Figures 2 and 3 (snail plot and Hi-C map) are clear, but additional annotation of chromosome identity (e.g., the sex chromosomes) would add value. Even if preliminary, indicating which scaffolds correspond to sex chromosomes would enhance interpretability.
4. Assembly quality control. BUSCO, Merqury QV, and k-mer completeness are reported in the results, but the corresponding methods are missing. Please add details (e.g., software versions, parameters) in the Methods.

Minor suggestions:

1. Missing noun after '52x Oxford Nanopore'; also 'data were sequenced" is inaccurate.' It could be like 'In total, sequencing yielded 52× coverage from Oxford Nanopore, 88× from Illumina WGS, and 163× from Hi-C libraries for the assembly.'
2. Words like 'Likelihood' in '...To distinguish true isoforms from fragments, the likelihood of each open reading frame (ORF) was evaluated against known vertebrate proteins...evidence of fragmented ORFs, were removed.' is vague. It could be like '...To distinguish..., each ORF

was compared against known..., and those with low-confidence (specify the threshold here, like frameshift or a premature stop codon?) were removed.
3. Missing a comma before the second 'did' in '...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...'

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

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomic

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 02 September 2025

https://doi.org/10.21956/openreseurope.22571.r58763

© 2025 Schöneberg Y. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Yannis Schöneberg
Biogeography, Universitat Trier Biogeographie, Trier, Rhineland-Palatinate, Germany

The aim of this study was to generate a reference genome for Natrix maura. The authors clearly show that they succeeded in generating a high quality, chromosome level genome and a high quality annotation. Therefore, this genome will be a valuable resource for future studies on this species, including but not limited to evolution and conservation of this species. The methods are state of the art and mostly described in high detail. However, I found a few sections where clarity could be improved and details are missing.
I congratulate the authors for generating such a valuable high quality genome resource and genome report.
Best whishes
Yannis Schöneberg

#Detailed Comments:
Unfortunately, it was not able to get a version with line numbers, even after inquiry with the editor. Therefore, I try to describe which text sections I mean.

Comment 1
"The species identification via morphology based on Santos 2021 and barcoding was confirmed by Salvador Carranza, Daniel Fernández-Guiberteau and Laura Blasón."

I think the citation was not properly inserted.

Comment 2
"The species identification via morphology based on Santos 2021 and barcoding was confirmed by Salvador Carranza, Daniel Fernández-Guiberteau and Laura Blasón."

It would be good to give a bit more details on how the barcoding was performed. Which marker sequence was used? Which Primers were used and how was it sequenced and how was the species identity confirmed?

Comment 3
"Data availability" (below Vouchering information)

This header seems misplaced in the methods section.

Comment 4
"In total, 52x Oxford Nanopore, 88x Illumina WGS shotgun, and 163x HiC data were sequenced to generate the assembly"

It would be good to speak of expected base coverage here, otherwise there might be room for misunderstandings.

Comment 5
Paragraph "Genome assembly methods"

In this section, many citations for the programs are missing (e.g. TrimGalore or YaHS) and some existing ones are inadequate (e.g. the one of blobtoolkit). This should be revised.

Comment 6
"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."

To me it's not entirely clear whether this is still part of the Ensemble Gene Annotation System, or if this was done by the authors manually. It would be good to have more information on how this was done.

Comment 7
"The genome assembly has a total length of 1,738,090,269 bp in 210 scaffolds including the

mitogenome (Figure 2 and Figure 3), with a GC content of 41.45%. It features a contig N50 of 36,526,383 bp (L50=13) and a scaffold N50 of 174,378,910 bp (L50=4). There are 75 gaps, totaling 15 kb in cumulative size. The single-copy gene content analysis using the sauropsida_odb10 database with BUSCO v5.5.0 resulted in 93.4% completeness (92.0% single and 1.4% duplicated). 91.35% of reads k-mers were present in the assembly and the assembly has a base accuracy Quality Value (QV) of 46.28 as calculated by Merqury."

The quality control methods are missing from the method section. They should be added e.g. in the "Genome assembly methods" section.

Comment 8
The genome annotation consists of 17,402 protein-coding genes with an associated 51,854 transcripts, in addition to 27 pseudogenes and 1,491 non-coding RNA genes of various types (Table 1). Using the longest isoform per transcript, the single-copy gene content analysis using the sauropsida_odb10 database with BUSCO v5.5.0 resulted in 93.4% completeness. Using the OMAmer Episquamata database for OMArk resulted in 95.85% completeness and 97.59% consistency (Table 2).

The quality control methods are missing from the method section. They should be added e.g. in the "Genome annoation methods" section.

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

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genome Assembly, Comparative Genomics, Phylogenetics

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.