Near chromosome-level and highly repetitive genome assembly of the snake pipefish *Entelurus aequoreus* (Syngnathiformes: Syngnathidae)

Magnus Wolf[1,2,3], Bruno Lopes da Silva Ferrette[1], Raphael T. F. Coimbra[1,2], Menno de Jong[1], Marcel Nebenfuehr[1,2], David Prochotta[1,2], Yannis Schöneberg[1,2], Konstantin Zapf[1,2], Jessica Rosenbaum[2], Hannah A. Mc Intyre[2], Julia Maier[2], Clara C.S. de Souza[2], Lucas M. Gehlhaar[2], Melina J. Werner[2], Henrik Oechler[2], Marie Wittekind[2], Moritz Sonnewald[4], Maria A. Nilsson[1,5], Axel Janke[1,2,5], Sven Winter[1,2,6]

[1] Senckenberg Biodiversity and Climate Research Centre (BiK-F), Frankfurt am Main, Germany

[2] Institute for Ecology, Evolution, and Diversity, Goethe University, Frankfurt am Main, Germany

[3] Institute for Evolution and Biodiversity, University of Münster, Münster, Germany

[4] Senckenberg Research Institute, Department of Marine Zoology, Section Ichthyology, Frankfurt am Main, Germany

[5] LOEWE-Centre for Translational Biodiversity Genomics (TBG), Frankfurt am Main, Germany

[6] Research Institute of Wildlife Ecology, University of Veterinary Medicine, Vienna, Austria

Authors for correspondence:
Sven Winter | Email: Sven.Winter@senckenberg.de
Magnus Wolf | Email: Magnus.Wolf@senckenberg.de

Name	e-mail	ORCID
Magnus Wolf (MW)	magnus.wolf@senckenberg.de	0000-0001-9212-9861
Bruno L. S. Ferrette (BF)	bruno.ferrette@senckenberg.de	0000-0002-3108-9867
Raphael T. F. Coimbra (RC)	raphael.t.f.coimbra@gmail.com	0000-0002-6075-7203
Menno de Jong (MDJ)	menno.de-jong@senckenberg.de	0000-0003-2131-9048
Marcel Nebenfuehr (MN)	marcel.nebenfuehr@senckenberg.de	0000-0001-8802-2105
David Prochotta (DP)	david.prochotta@senckenberg.de	0009-0000-6275-7752
Yannis Schöneberg (YS)	yannis.schoeneberg@gmx.de	0000-0003-1113-973X
Konstantin Zapf (KZ)	konstantin.zapf@senckenberg.de	
Jessica Rosenbaum (JR)	s5932510@stud.uni-frankfurt.de	0009-0008-2306-9015
Hannah A. Mc Intyre (HMI)	s3630184@stud.uni-frankfurt.de	0009-0002-3275-5048
Julia Maier (JM)	juliaomaier@googlemail.com	
Clara C.S. de Souza (CDS)	s6795932@stud.uni-frankfurt.de	0009-0000-9560-8905
Lucas M. Gehlhaar (LG)	s6244190@stud.uni-frankfurt.de	
Melina J. Werner (MJW)	s5033619@stud.uni-frankfurt.de	0009-0007-1081-009X

Henrik Oechler (HO)	henrik.oechler@uni-bayreuth.de	0009-0001-0413-8731
Marie Wittekind (MWI)	s0097351@stud.uni-frankfurt.de	0009-0001-2443-7552
Moritz Sonnewald (MS)	moritz.sonnewald@senckenberg.de	0000-0003-3042-8107
Maria A. Nilsson (MAN)	maria.nilsson-janke@senckenberg.de	0000-0002-8136-7263
Axel Janke (AJ)	axel.janke@senckenberg.de	0000-0002-9394-1904
Sven Winter (SW)	sven.winter@vetmeduni.ac.at	0000-0002-1890-0977

Abstract

The snake pipefish, *Entelurus aequoreus* (Linnaeus, 1758), is a slender, up to 60 cm long, northern Atlantic fish that dwells in open seagrass habitats and has recently expanded its distribution range. The snake pipefish is part of the family Syngnathidae (seahorses and pipefish) that has undergone several characteristic morphological changes, such as loss of pelvic fins and elongated snout. Here, we present a highly contiguous, near chromosome-scale genome of the snake pipefish assembled as part of a university master's course. The final assembly has a length of 1.6 Gbp in 7,391 scaffolds, a scaffold and contig N50 of 62.3 Mbp and 45.0 Mbp and L50 of 12 and 14, respectively. The largest 28 scaffolds (>21 Mbp) span 89.7% of the assembly length. A BUSCO completeness score of 94.1% and a mapping rate above 98% suggest a high assembly completeness. Repetitive elements cover 74.93% of the genome, one of the highest proportions so far identified in vertebrate genomes. Demographic modeling using the PSMC framework indicates a peak in effective population size (50 – 100 kya) during the last interglacial period and suggests that the species might largely benefit from warmer water conditions, as seen today. Our updated snake pipefish assembly forms an important foundation for further analysis of the morphological and molecular changes unique to the family Syngnathidae.

Keywords

long reads, proximity-ligation scaffolding, genome annotation, demographic history, repetitive elements

Introduction

The snake pipefish *Entelurus aequoreus* (Linnaeus 1758) is a member of the family Syngnathidae, which currently includes over 300 species of seahorses and pipefishes [1]. The

56 species shares typical features with other pipefishes such as a unique, elongated body plan and
57 fused jaws [2]. However, unlike most pipefish, which are found in benthic habitats, the snake
58 pipefish inhabits more open and deeper seagrass environments and occurs even in pelagic
59 waters [2]. They are ambush predators on small crustaceans and other invertebrates, thereby
60 indirectly contributing to the overall biodiversity and stability of these fragile habitats [3].
61 Adult snake pipefish are poor swimmers with small fins and rely on their elongated, thin
62 bodies for crypsis in eelgrass habitats [4–6].

63 The snake pipefish historically ranged from the waters of Azores northwards to the
64 waters of Norway and Iceland, and eastward to the Baltic [7]. However, since 2003, the
65 species has expanded its distribution [8] into the arctic waters of Spitsbergen [9], the Barents
66 Sea, and the Greenland Sea [10]. Simultaneously, population sizes seem to increase within its
67 former range, as indicated by substantially increased catch rates [11, 12]. Several factors have
68 been proposed to cause this expansion and population growth, including rising sea
69 temperatures, an increased potential for long-distance dispersal of juveniles via ocean currents
70 [4, 7] and an increased reproductive success facilitated by the dispersal of invasive seaweeds
71 [6, 8–10, 13]. The latter explanation has been confirmed in local field experiments in the
72 northern Wadden Sea, suggesting a mutual co-occurrence of the invasive Japanese seaweed
73 (*Sargassum muticum*) and the snake pipefish [5]. Studies based on mtDNA marker regions
74 did not discern any population structure thus far and suggest a previous population expansion
75 in the Pleistocene ca. 50–100 kya [6]. Yet, a comprehensive analysis of demographic events is
76 better studied from genomic data, requiring a high-quality reference genome of ideally the
77 same species or at least a closely related one.

78 Previously, genomes of Syngnathidae have been used to study the evolution of highly
79 specialized morphologies and life-history traits unique to pipefishes and seahorses [14–16].
80 The transition to male pregnancy was associated with major genomic restructuring events and
81 parallel modifications of the adaptive immune system. There is a remarkable variability in

genome sizes within the family, with estimates ranging from 350 Mbp to 1.8 Gbp [14]. The major shifts in body shape are assumed to be related to gene-family loss and expansion events and higher rates of protein and nucleotide evolution [16]. Genomic data using a direct sequencing approach of ultra-conserved elements (UCEs) improved the understanding of the phylogeny of pipefishes [15] and identified a likely radiation of the group in the waters of the modern Indo-Pacific Ocean. Nevertheless, high-quality genomes of Syngnathidae are only available for a few species, and according to the NCBI genome database, only 7% of the known species diversity has genome sequences available.

A draft genome of the snake pipefish was previously assembled using a combination of paired-end and mate-pair sequencing techniques, yielding an assembly with low continuity (N50 3.5 kbp, BUSCO C: 21%) and a large difference between the estimated and assembled genome sizes (1.8 Gbp vs. 557 Mbp) [14]. To obtain a higher quality, near chromosome-scale genome assembly for the snake pipefish for future population genomic, conservation, and evolutionary studies of fish, we used long-read sequencing technologies. This allows us to gain insight into the genetic properties of the species and to perform demographic analysis based on the PSMC framework [17]. The data generation and analyses presented here were conducted during a six-week master course in 2021 at the Goethe University, Frankfurt am Main, Germany. The concept of high-quality genome sequencing in a course setting has so far yielded three reference-quality genomes of fish and has proven to be a successful approach to introduce the technology to a new generation of scientists [18–21].

Results and Discussion

Genome sequencing and assembly

PacBio's continuous long reads (CLR) technology generated 401 Gbp of long-read data in ~60 million reads with an N50 of 7.9 kb (Table 1). Illumina sequencing yielded 38 Gbp of standard short-read data in ~257 million reads with a mean length of 148 bp after filtering. Sequencing of the Omni-C library generated 54.7 Gbp of raw short-read data.

The snake pipefish's genome was assembled *de novo* to a total size of 1.7 Gbp. It consisted of 2,204 scaffolds, with a scaffold N50 of 62 Mbp and an L50 of 11 (Table 1, Fig. 1A). The finalized assembly has 1.0 N's per 100 kbp and a GC content of 38.84%. A BUSCO completeness assessment resulted in 94.1% complete core genes, given the *actinopterygii_obd10* set, indicating high completeness of the assembly. Both long- and short-read data mapped onto the assembly with high mapping rates of 98.6% and 99.5%, respectively. HI-C mapping resulted in 28 larger scaffolds (Fig. 1B), indicating the near-chromosome level of the *de novo* assembly as past karyotype estimations of other pipefish and seahorses predicted 22 and 22-24 chromosomes, respectively [22–24]. The rest of the genome comprises only smaller scaffolds and contigs, which may result from the high amounts of repetitive regions described in the following section. Our Blobtools analysis of both long- and short-read data (Fig. 1C+D) found no apparent signs of contamination, although background noise of unknown origin was detected and removed in both datasets.

Variant calling resulted in ~301 million sites (including monomorphic sites), of which ~1.3 million were found to be biallelic. Genome-wide heterozygosity was determined to be 0.387%, which is in line with other fish species [25, 26]. The GenomeScope results based on short reads suggested a haploid genome size of 1.15 Gbp and an expected genome-wide heterozygosity of 1%, around 362 Mbp shorter and 0.57% more heterozygous when compared to the final assembly. This, again, might be explained by the high repeat content in the genome.

Annotation

In total, 0.9 Gbp, or 74.93%, of the entire assembly, were identified as repetitive during our *de novo* repeat-modeling and repeat-masking (Fig. 2). This high repeat content contrasts that of other fish genomes [27], but is similar, although at a smaller scale, to the closest relative *Nerophis ophidion* (65.7%) [14] and other genomes of syngnathid fish like e.g., seadragons [28]. The first draft assembly of the snake pipefish had a repeat content of 57.2% [14] and our improved long-read assembly identified 17.7% additional repeats that were missing from the previous assembly [14]. So far, among vertebrates, only the lungfish *Neoceratodus forsteri* [29] has more transposable elements (TEs) than the snake pipefish.

The annotation of the genome featuring *de novo* and homology-based identification approaches resulted in 33,202 genes with an average length of 13,828 bp. Each gene had on average 7.32 exons and 6.25 introns with average lengths of 188 bp and 2,240 bp, respectively. In total, we identified 243,038 exons and 207,467 introns within our annotation. The total number of genes is ~30% larger compared to other annotated genomes in the order of Syngnathiformes like, e.g., 23,458 for the tiger tail seahorse (*Hippocampus comes*) [16] or 24,927 for the greater pipefish (*Syngnathus acus*) [30] made by the NCBI Eukaryotic Genome Annotation pipeline. Given that these two genomes are also considerably smaller, 492 Mbp and 324 Mbp, respectively, it can be assumed that the large-scale genome increase in this species also included many coding sequences. A high content of repetitive regions as well as a lack of transcriptomic data might also have increased the number of false positive gene-calls; however, a BUSCO completeness analysis of the predicted proteins resulted in 82.6% complete sequences, of which only 6.8% were duplicated. 5.3% of the coding sequences appeared fragmented, and 12.1% were missing from the *actinopterygii_obd10* OrthoDB set. A functional annotation resulted in hits for 89% of the predicted proteins.

Demographic inference

The demographic inference analysis of the snake pipefish genome using the PSMC framework [17] traced population changes over the past 1 million years. Given the chosen substitution rate and generation time, there was a steady increase in the effective population size (N_e), starting at 15 thousand individuals 1 Mya, which peaked at an N_e of 250 thousand individuals at 100 kya. Thereafter, N_e decreased until reaching 30 thousand individuals at 10 kya and stagnated until the end of the model. The previously suggested population expansion during the Pleistocene (50 – 100 kya) was therefore confirmed with this model but was followed by another population decline that wasn't resolved by Braga Goncalves et al. [6]. This result may point to a different conclusion as drawn by the authors, because the snake pipefish might have resided in a comparable small population size during the Holocene and only recently expanded its distribution, resulting in a large population with a high degree of homogenization as observed by Braga Goncalves and colleagues [6]. Given that the presented peak in population size parallels with the last interglacial period between the Penultimate Glacial Period (135 – 192 kya [31]) and the last glacial period (present – 20 kya [32]), we assume that the snake pipefish largely benefitted from the warmer water conditions during the interglacial period as seen in the present range expansion.

Material & Methods

Sampling, DNA extraction, and sequencing

A single individual of *Entelurus aequoreus* (Linnaeus 1758) was caught by trawling during an annual monitoring expedition to the Dogger Bank in the North Sea in July 2021 (trawl start coordinates 54.993633, 2.940833; end coordinates 55.0077, 2.929867) with the permission of the Maritime Policy Unit of the UK Foreign and Commonwealth Office. The study was conducted in compliance with the 'Nagoya Protocol on Access to Genetic Resources and the

Fair and Equitable Sharing of Benefits Arising from Their Utilization'. The sample was initially frozen at -20 °C and later stored at -80 °C.

High molecular weight genomic DNA was extracted from muscle tissue, following the protocol by Mayjonade et al. [33] with the addition of Proteinase K. We evaluated the quantity and quality of the DNA with the Genomic DNA ScreenTape on the Agilent 2200 TapeStation system (Agilent Technologies), as well as with the Qubit® dsDNA BR Assay Kit.

For long-read sequencing, a PacBio SMRT Bell continuous long read (CLR) library was prepared using the SMRTbell Express Prep kit v3.0 kit (Pacific Biosciences – PacBio, Menlo Park, CA, USA) and sequenced on the PacBio Sequel IIe platform. A proximity-ligation library was compiled with muscle tissue following the Dovetail™ Omni-C protocol (Dovetail Genomics, Santa Cruz, California, USA). In addition, a standard whole-genome 150 base pair (bp) paired-end Illumina library was prepared using the NEBNext Ultra II library preparation kit (New England Biolabs Inc., Ipswich, USA). Finally, the proximity ligation and the paired-end library were shipped to Novogene (UK) for sequencing on the Illumina NovaSeq 6000 platform.

Pre-processing & Genome size estimation

The PacBio subreads were converted from BAM into FASTQ format using the PacBio Secondary Analysis Tool BAM2fastx v.1.3.0 (https://github.com/PacificBiosciences/pbbioconda). Quality control, trimming, and filtering of the Illumina reads were performed using fastp v0.23.1 [34] with the settings "*-g -3 -l 40 -y -Y 30 -q 15 -u 40 -c -p -j -h -R -w N*". To estimate the genome size of the snake pipefish, we performed *k*-mer profiling using the standard short-read Illumina data. We first ran Jellyfish v2.3.0 [35] to generate a histogram of *k*-mers with a length of 21 bp. Subsequently, we used this data to obtain a genome profile

using GenomeScope v2.0 [36]. We further tested alternative k-mer lengths between 17- and 25-mers which resulted in no meaningful differences of the estimated genome size except for the 17-mer, which resulted in a smaller genome size estimation of ~500 Mbp.

Genome Assembly & polishing

We assembled the genome from the PacBio long-read data using WTDBG v.2.5 [37]. The resulting assembly was first polished using the PacBio data with Flye v.2.9 [38], using Minimap v.2.17 [39] for mapping, followed by two rounds of short-read polishing by mapping reads onto the assembly with BWA-MEM v.0.7.17 [40] and error correction with Pilon v1.23 [41].

Assembly QC & Scaffolding

The polished assembly contigs were anchored into chromosome-scale scaffolds utilizing the generated proximity-ligation Omni-C data. First, the data were mapped and filtered to the assembly following the Arima Hi-C mapping pipeline used by the Vertebrate Genome Project (https://github.com/VGP/vgpassembly/blob/master/pipeline/salsa/arima_mapping_pipeline.sh). In brief, reads were mapped using BWA-MEM v.0.7.17 [40], mapped reads were filtered with samtools v.1.14 [42], and duplicated reads were removed with “MarkDuplicates” in Picard v.2.26.10 (Broad Institute, 2019). The filtered mapped reads were then used for proximity-ligation scaffolding in YaHs v.1.1 [43]. Gaps in the scaffolded assembly were closed with TGS-GapCloser v.1.1.1 [44] using a subset (25%) of the PacBio subreads due to computational constraints. To further improve the assembly's contiguity, scaffolding and gap-closing were performed a second time using a different subset of PacBio reads for gap-closing. The PacBio read subsets were generated with seqtk v.1.3 (https://github.com/lh3/seqtk) using the random number generator seeds 11 and 18. Gene set

completeness was analyzed with BUSCO v.5.4.7 [45] using the Actinopterygii set of core genes (*actinopterygii_odb10*). Assembly continuity was evaluated using QUAST v5.0.2 [46], and mapping rates were assessed by Qualimap v2.2.1 [47]. BlobToolsKit v.4.0.6 [48] was used to perform contamination screening.

Repeat landscape analysis & genome annotation

The TE annotation was done in three steps. First, we used RepeatMasker v4.1.5 [49] to annotate, and hard-mask known Actinopterygii repeats from RepBase, which comprises a database of eukaryotic repetitive DNA element sequences [50]. Secondly, a *de novo* library of transposable elements was created from the hard-masked genome assembly using RepeatModeler v2.0.4 [51] which includes RECON v1.08 [52], RepeatScout v1.0.6 [53], and LTRharvest/LTR_retriever [54, 55]. Finally, predicted repeats were annotated with a second run of RepeatMasker on the hard-masked assembly obtained in the first run. The results from both RepeatMasker runs were then combined. A summary of transposable elements and the relative abundance of repeat classes in the genome are shown in Table 2 and Fig. 2.

The genome was annotated using the BRAKER3 pipeline [56–61], combining a *de novo* gene calling and a homology-based gene annotation. For protein references, we combined the vertebrate-specific protein collection from OrthoDB and the protein collection of the greater pipefish (*Syngnathus acus*) genome [30] made by the NCBI (see: GCF_901709675.1, last accessed 12th Oct. 2023). To further filter genes based on the support of introns by extrinsic homology evidence, we used TSEBRA [62] with an “intron_support=0.1”. The resulting set of proteins was tested for completeness using BUSCO v.5.4.7 [45] in “protein mode” and run against the Actinopterygii-specific set of core genes. Functional annotation was done using InterProScan v5 [63].

Variant calling & demographic inference

The preprocessed short reads were mapped to the final assembly using BWA-MEM v. 0.7.17 [40] followed by removal of duplicate reads with "MarkDuplicates" in Picard v.2.26.10 (Broad Institute, 2019) and evaluation of mapping quality using Qualimap v2.2.1 [47] . Indels in the BAM files were first identified and then realigned with "RealignerTargetCreator" and "IndelRealigner" as part of the Genome Analysis Toolkit (GATK) v3.8-1 (https://gatk.broadinstitute.org/). Subsequently, samtools v.1.14 [42] was used to check and remove unmapped, secondary, QC failed, duplicated, and supplementary reads keeping only reads mapped in proper pairs in non-repetitive regions of the 28 chromosome-scale scaffolds.

Sambamba v 1.0.0 [64] was used to estimate site depth statistics. Minimum and maximum thresholds for the global site depth were set to d ± (5 × MAD), where d is the global site depth distribution median and MAD is the median absolute deviation. Variant calling was performed using the bcftools v1.17 [65] commands "mpileup" and "call" [-m]. Variants were then filtered with bcftools "filter" [-e "DP< d – (5 × MAD) || DP> d + (5 × MAD) || QUAL<30"] removing sites with low quality and out of range depth. Finally, bcftools was used to estimate the genome-wide heterozygosity as the proportion of heterozygous sites using the "stats" command.

Long-term changes in effective population size (N_e) over time were estimated with the Pairwise Sequentially Markovian Coalescent (PSMC) model [17] based on the diploid consensus genome sequences generated by bcftools v1.17 [65] with the script "*vcfutils.pl*" from the processed BAM files, as described above. Sites with read-depth up to a third of the average depth or above twice each sample's median depth and with a consensus base quality < 30 were removed. PSMC was executed using 25 iterations with a maximum $2N_0$-scaled coalescent time of 15, an initial θ/ρ ratio of 5, and 64 atomic time intervals (4 + 25 × 2 + 4 + 6) to infer the scaled mutation rate, the recombination rate, and the free population size parameters, respectively. We performed 100 bootstrap replicates by randomly sampling with

replacement of 1 Mb blocks from the consensus sequence for all individuals. A mutation rate (μ) of 1.7 x 10^{-9} per site per generation [66] and a generation length of 2.5 years [67] were employed for plotting.

Availability of Supporting Data

The *de novo* genome and all underlying raw data were uploaded to NCBI under the BioProject PRJNA1005573, BioSample SAMN36988691, genome assembly JAVRRV000000000. All other data, including the repeat and gene annotation, was uploaded to the GigaDB repository: DOI:XXXXX. [Rawdata available for review at https://dataview.ncbi.nlm.nih.gov/object/PRJNA1005573?reviewer=2i5vm98fdb4r9j0asoff8msn3]

Author Contributions

MW, BF, MS, AJ, and SW designed the study. SW, JR, HMI, JM, CDS, LG, MJW, HO, and MWI performed laboratory procedures and sequencing. MW, BF, RC, MDJ, MN, DP, YS, KZ, JR, HMI, JM, CDS, LG, MJW, HO, MWI, MAN, and SW conducted bioinformatic processing and analyses. All authors contributed to writing this manuscript.

List of Abbreviations

bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; CLR: continuous long reads; Gbp: Gigabase pairs; kbp: kilobase pairs; kya: thousand years ago; Mbp: megabase pairs; Mya: million years ago; N_e: effective population size; PacBio: Pacific Biosciences; PSMC: Pairwise Sequentially Markovian Coalescent; TEs: transposable elements; UCEs: ultra-conserved elements.

Conflict of Interest

The authors declare that they have no competing interests.

Ethics Approval and Consent for Publication

Not Applicable.

Acknowledgments

The present study is a result of the Centre for Translational Biodiversity Genomics (LOEWE-TBG) and was supported through the program 'LOEWE-Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz' of Hesse's Ministry of Higher Education, Research, and the Arts.

References

1. Froese R, Pauly D. FishBase. 2023. www.fishbase.org. Accessed 9 Aug 2023.
2. Dawson C. Syngnathidae. In: Smith M, Heemstra P, editors. Smiths' sea fishes. Berlin: Springer-Verlag; 1986. p. 445–458.
3. O'Gorman EJ. Multitrophic diversity sustains ecological complexity by dampening top-down control of a shallow marine benthic food web. Ecology. 2021;102:e03274. doi:10.1002/ecy.3274.
4. Vincent ACJ, Berglund A, Ahnesj I. Reproductive ecology of five pipefish species in one eelgrass meadow. Environ Biol Fish. 1995;44:347–61. doi:10.1007/BF00008250.
5. Polte P, Buschbaum C. Native pipefish Entelurus aequoreus are promoted by the introduced seaweed Sargassum muticum in the northern Wadden Sea, North Sea. Aquat. Biol. 2008;3:11–8. doi:10.3354/ab00071.
6. Braga Goncalves I, Cornetti L, Couperus AS, van Damme CJG, Mobley KB. Phylogeography of the snake pipefish, Entelurus aequoreus (Family: Syngnathidae) in the northeastern Atlantic Ocean. Biological Journal of the Linnean Society. 2017;122:787–800. doi:10.1093/biolinnean/blx112.
7. Wheeler A. Key to the Fishes of Northern Europe: A guide to the identification of more than 350 species. London: Frederick Warne & Co. Ltd; 1978.
8. Harris MP, Beare D, Toresen R, Nøttestad L, Kloppmann M, Dörner H, et al. A major increase in snake pipefish (Entelurus aequoreus) in northern European seas since 2003: potential implications for seabird breeding success. Mar Biol. 2007;151:973–83. doi:10.1007/s00227-006-0534-7.
9. Fleischer D, Schaber M, Piepenburg D. Atlantic snake pipefish (Entelurus aequoreus) extends its northward distribution range to Svalbard (Arctic Ocean). Polar Biol. 2007;30:1359–62. doi:10.1007/s00300-007-0322-y.
10. Rusyaev SM, Dolgov AV, Karamushko OV. Captures of snake pipefish Entelurus aequoreus in the Barents and Greenland Seas. J. Ichthyol. 2007;47:544–6. doi:10.1134/S0032945207070090.
11. Kloppmann MHF, Ulleweit J. Off-shelf distribution of pelagic snake pipefish, Entelurus aequoreus (Linnaeus, 1758), west of the British Isles. Mar Biol. 2007;151:271–5. doi:10.1007/s00227-006-0480-4.
12. van Damme CJ, Couperus AS. Mass occurrence of snake pipefish in the Northeast Atlantic: Result of a change in climate? Journal of Sea Research. 2008;60:117–25. doi:10.1016/j.seares.2008.02.009.
13. Lindley J, Kirby R, Johns D, Reid C. Exceptional abundance of the snake pipefish (Entelurus aequoreus) in the north-eastern North Atlantic Ocean. ICES Document. 2006.
14. Roth O, Solbakken MH, Tørresen OK, Bayer T, Matschiner M, Baalsrud HT, et al. Evolution of male pregnancy associated with remodeling of canonical vertebrate immunity in seahorses and pipefishes. Proc Natl Acad Sci U S A. 2020;117:9431–9. doi:10.1073/pnas.1916251117.
15. Stiller J, Short G, Hamilton H, Saarman N, Longo S, Wainwright P, et al. Phylogenomic analysis of Syngnathidae reveals novel relationships, origins of endemic diversity and variable diversification rates. BMC Biol. 2022;20:75. doi:10.1186/s12915-022-01271-w.
16. Lin Q, Fan S, Zhang Y, Xu M, Zhang H, Yang Y, et al. The seahorse genome and the evolution of its specialized morphology. Nature. 2016;540:395–9. doi:10.1038/nature20595.

17. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–6. doi:10.1038/nature10231.
18. Prost S, Winter S, Raad J de, Coimbra RTF, Wolf M, Nilsson MA, et al. Education in the genomics era: Generating high-quality genome assemblies in university courses. Gigascience 2020. doi:10.1093/gigascience/giaa058.
19. Prost S, Petersen M, Grethlein M, Hahn SJ, Kuschik-Maczollek N, Olesiuk ME, et al. Improving the Chromosome-Level Genome Assembly of the Siamese Fighting Fish (Betta splendens) in a University Master's Course. G3 (Bethesda). 2020;10:2179–83. doi:10.1534/g3.120.401205.
20. Winter S, Prost S, Raad J de, Coimbra RTF, Wolf M, Nebenführ M, et al. Chromosome-level genome assembly of a benthic associated Syngnathiformes species: the common dragonet, Callionymus lyra. GigaByte. 2020;2020:gigabyte6. doi:10.46471/gigabyte.6.
21. Winter S, Raad J de, Wolf M, Coimbra RTF, Jong MJ de, Schöneberg Y, et al. A chromosome-scale reference genome assembly of the great sand eel, Hyperoplus lanceolatus. J Hered. 2023;114:189–94. doi:10.1093/jhered/esad003.
22. Vitturi R, Catalano E. Karyotypes in two species of the genusHippocampus (Pisces: Syngnatiformes). Mar Biol. 1988;99:119–21. doi:10.1007/BF00644985.
23. Vitturi R, Libertini A, Campolmi M, Calderazzo F, Mazzola A. Conventional karyotype, nucleolar organizer regions and genome size in five Mediterranean species of Syngnathidae (Pisces, Syngnathiformes). Journal of Fish Biology. 1998;52:677–87. doi:10.1111/j.1095-8649.1998.tb00812.x.
24. Small CM, Bassham S, Catchen J, Amores A, Fuiten AM, Brown RS, et al. The genome of the Gulf pipefish enables understanding of evolutionary innovations. Genome Biol. 2016;17:258. doi:10.1186/s13059-016-1126-6.
25. Tigano A, Jacobs A, Wilder AP, Nand A, Zhan Y, Dekker J, Therkildsen NO. Chromosome-Level Assembly of the Atlantic Silverside Genome Reveals Extreme Levels of Sequence Diversity and Structural Genetic Variation. Genome Biol Evol 2021. doi:10.1093/gbe/evab098.
26. Barry P, Broquet T, Gagnaire P-A. Age-specific survivorship and fecundity shape genetic diversity in marine fishes. Evol Lett. 2022;6:46–62. doi:10.1002/evl3.265.
27. Shao F, Han M, Peng Z. Evolution and diversity of transposable elements in fish genomes. Sci Rep. 2019;9:15399. doi:10.1038/s41598-019-51888-1.
28. Small CM, Healey HM, Currey MC, Beck EA, Catchen J, Lin ASP, et al. Leafy and weedy seadragon genomes connect genic and repetitive DNA features to the extravagant biology of syngnathid fishes. Proc Natl Acad Sci U S A. 2022;119:e2119602119. doi:10.1073/pnas.2119602119.
29. Meyer A, Schloissnig S, Franchini P, Du K, Woltering JM, Irisarri I, et al. Giant lungfish genome elucidates the conquest of land by vertebrates. Nature. 2021;590:284–9. doi:10.1038/s41586-021-03198-8.
30. Scott-Somme K, McTierney S, Brittain R, Perry F, Brenen M. The genome sequence of the greater pipefish, Syngnathus acus (Linnaeus, 1758). Wellcome Open Res. 2023;8:274. doi:10.12688/wellcomeopenres.19528.1.
31. Obrochta SP, Crowley TJ, Channell JE, Hodell DA, Baker PA, Seki A, Yokoyama Y. Climate variability and ice-sheet dynamics during the last three glaciations. Earth and Planetary Science Letters. 2014;406:198–212. doi:10.1016/j.epsl.2014.09.004.
32. Armstrong E, Hopcroft PO, Valdes PJ. A simulated Northern Hemisphere terrestrial climate dataset for the past 60,000 years. Sci Data. 2019;6:265. doi:10.1038/s41597-019-0277-1.
33. Mayjonade B, Gouzy J, Donnadieu C, Pouilly N, Marande W, Callot C, et al. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. Biotechniques. 2016;61:203–5. doi:10.2144/000114460.
34. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884-i890. doi:10.1093/bioinformatics/bty560.
35. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70. doi:10.1093/bioinformatics/btr011.
36. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33:2202–4. doi:10.1093/bioinformatics/btx153.
37. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17:155–8. doi:10.1038/s41592-019-0669-3.
38. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6. doi:10.1038/s41587-019-0072-8.
39. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100. doi:10.1093/bioinformatics/bty191.
40. Li H. Aligning seuquence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013.

41. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9:e112963. doi:10.1371/journal.pone.0112963.
42. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.
43. Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. Bioinformatics 2023. doi:10.1093/bioinformatics/btac808.
44. Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, et al. TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. Gigascience 2020. doi:10.1093/gigascience/giaa094.
45. Manni M, Berkeley MR, Seppey M, Zdobnov EM. BUSCO: Assessing Genomic Data Quality and Beyond. Curr Protoc. 2021;1:e323. doi:10.1002/cpz1.323.
46. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34:i142-i150. doi:10.1093/bioinformatics/bty266.
47. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32:292–4. doi:10.1093/bioinformatics/btv566.
48. Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit – Interactive quality assessment of genome assemblies; 2019.
49. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013. http://www.repeatmasker.org.
50. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11. doi:10.1186/s13100-015-0041-9.
51. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117:9451–7. doi:10.1073/pnas.1921046117.
52. Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12:1269–76. doi:10.1101/gr.88502.
53. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21 Suppl 1:i351-8. doi:10.1093/bioinformatics/bti1018.
54. Ou S, Jiang N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol. 2018;176:1410–22. doi:10.1104/pp.17.01310.
55. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008;9:18. doi:10.1186/1471-2105-9-18.
56. Bruna T, Lomsadze A, Borodovsky M. GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistency with Extrinsic Data; 2023.
57. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 2021;3:lqaa108. doi:10.1093/nargab/lqaa108.
58. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20:278. doi:10.1186/s13059-019-1910-1.
59. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32:767–9. doi:10.1093/bioinformatics/btv661.
60. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60. doi:10.1038/nmeth.3176.
61. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, Stanke M. BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. bioRxiv 2023. doi:10.1101/2023.06.10.544449.
62. Gabriel L, Hoff KJ, Brůna T, Borodovsky M, Stanke M. TSEBRA: transcript selector for BRAKER. BMC Bioinformatics. 2021;22:566. doi:10.1186/s12859-021-04482-0.
63. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40. doi:10.1093/bioinformatics/btu031.
64. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4. doi:10.1093/bioinformatics/btv098.
65. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience 2021. doi:10.1093/gigascience/giab008.
66. He L, Long X, Qi J, Wang Z, Huang Z, Wu S, et al. Genome and gene evolution of seahorse species revealed by the chromosome-level genome of Hippocampus abdominalis. Mol Ecol Resour. 2022;22:1465–77. doi:10.1111/1755-0998.13541.
67. Schultz J. Entelurus aequoreus: IUCN Red List of Threatened Species, e.T18258072A44775951; 2014.

Figures

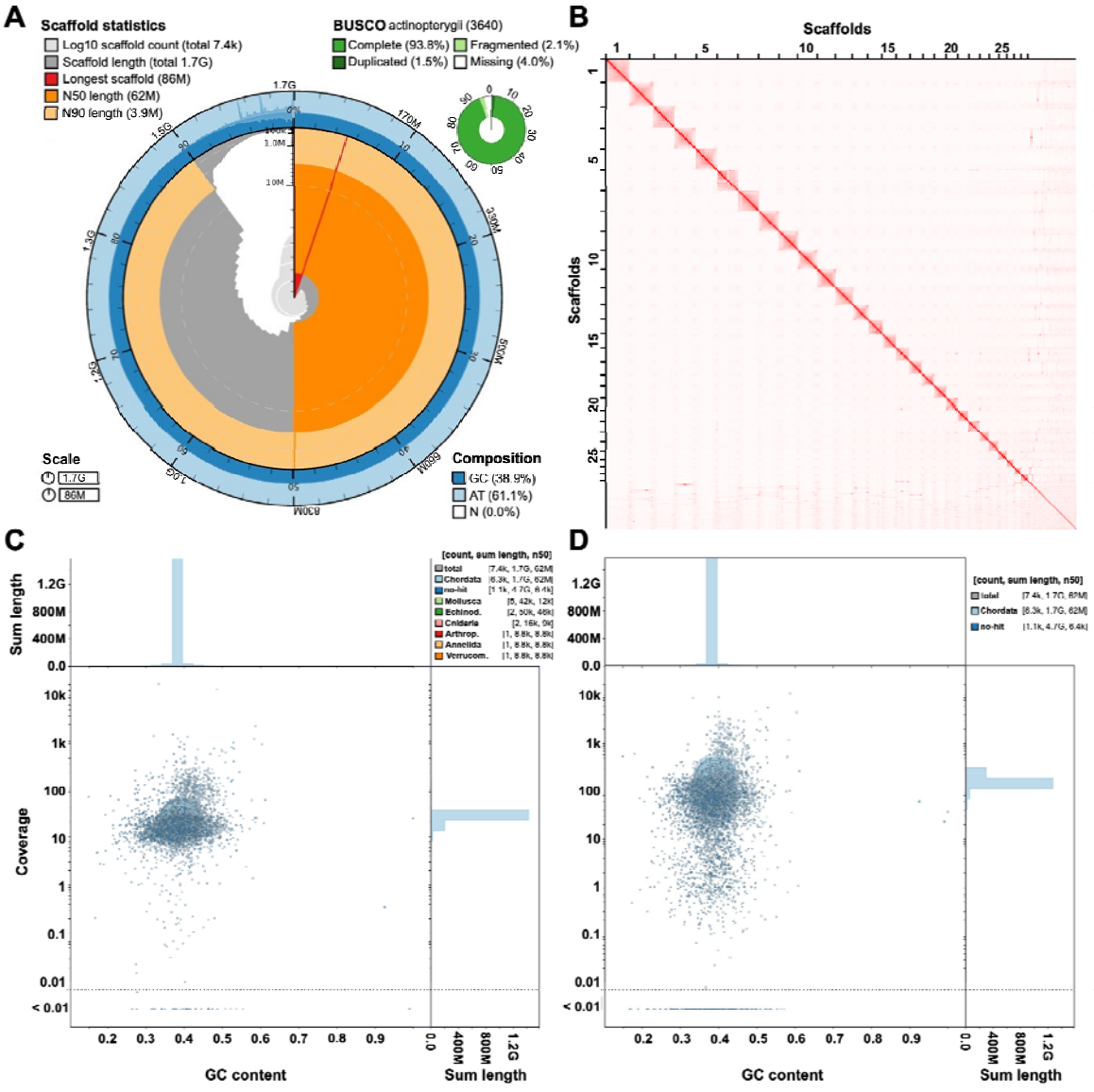
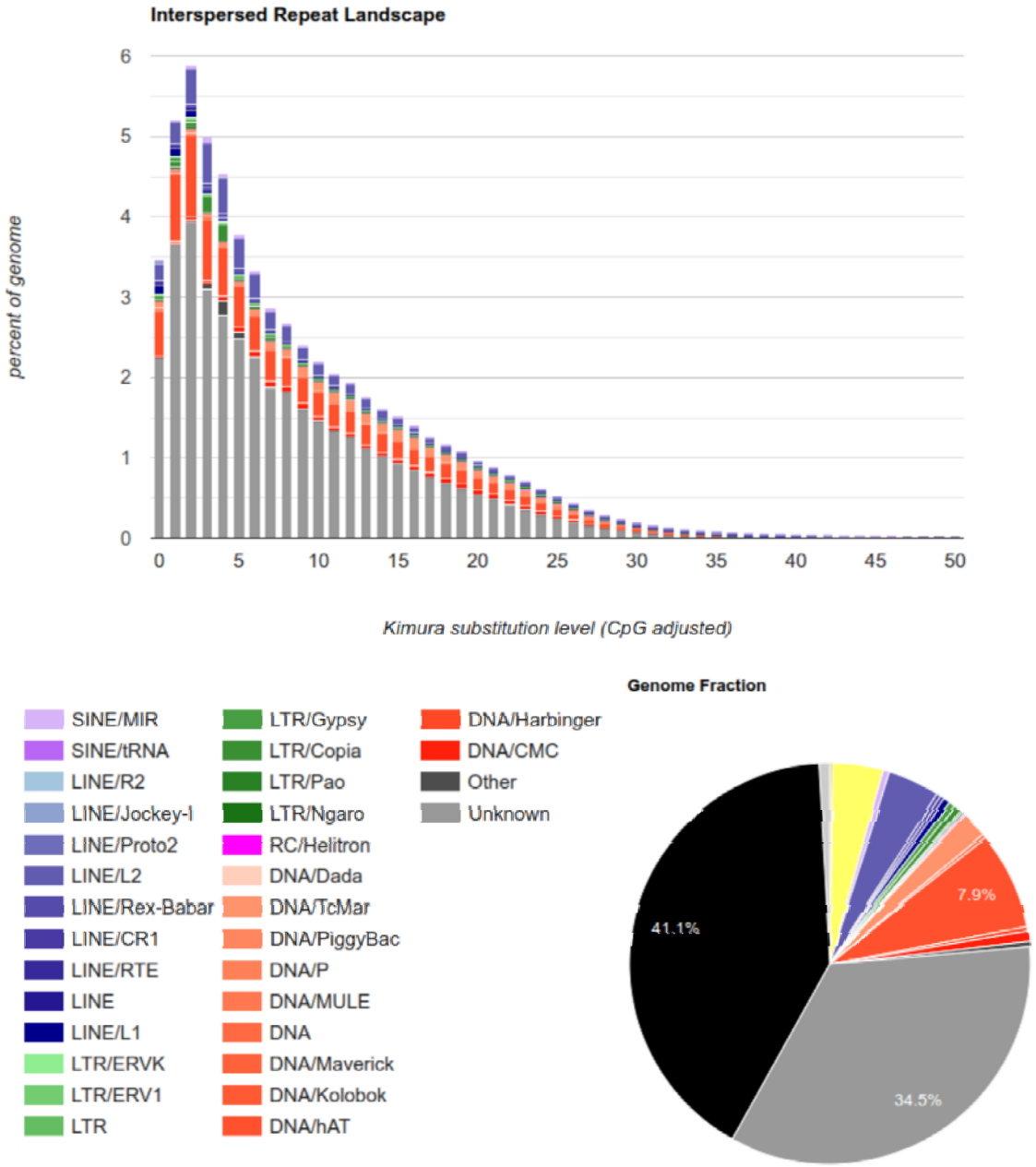


Figure 1 Assembly characteristics and quality assessments of the *de novo Entelurus aequoreus* genome. **A** The snail plot summarizes different assembly properties. Scaffold statistics are depicted in the innermost circle and the colors red to orange represent the longest scaffold N50 and N90, respectively. GC composition is shown in the outer blue circle. BUSCO completeness statistics are depicted in the small green circle. **B** Omni-C contact density map indicating 28 larger scaffolds and the near-chromosome level of the assembly. **C-D** The BlobPlot analysis compares GC content (x-axis), assembly coverage (y-axis) and taxonomic BLAST assignments of contigs (color) for both the Omni-C short reads (C) and PacBio long reads (D).

476



477

478 **Figure 2** Repeat landscape of the *de novo Entelurus aequoreus* genome. Colors represent repetitive element
479 types, gray areas indicate unclassified types of repetitive regions.

480

481

482

483

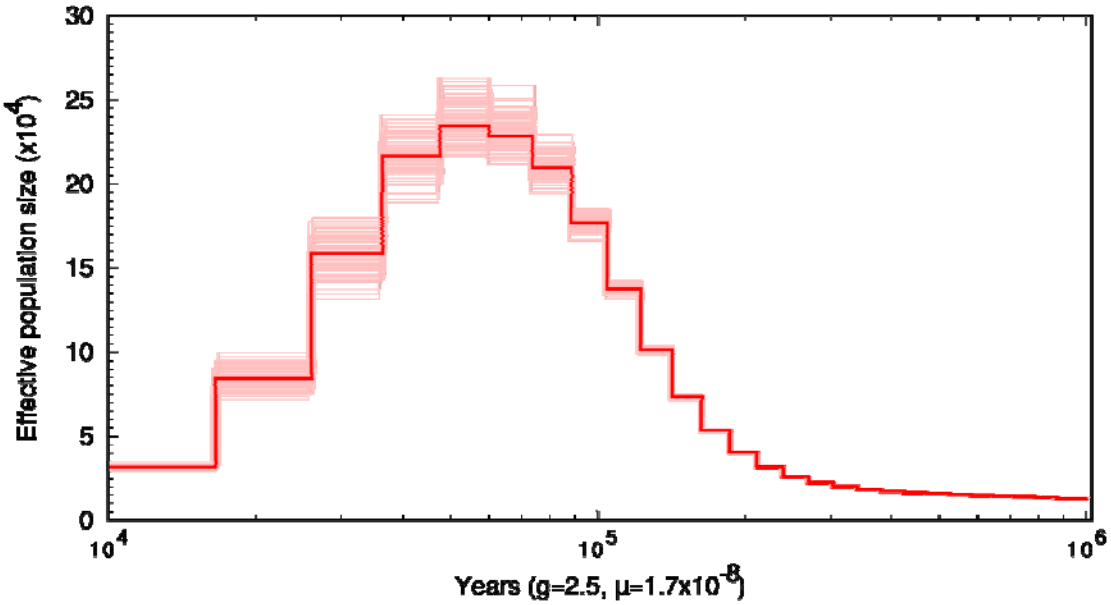


Figure 3 Demographic history of the snake pipefish estimated using the PSMC framework. Using a generation time of 2.5 years [67] and a substitution rate of 1.7×10^{-8} per site per generation [66] a model was created covering the last 10 kya to 1 Mya. The x-axis represents time in number of years ago and the y-axis shows the effective population (N_e) size in tens of thousands of individuals. The model indicates a peak in N_e of 250 thousand individuals during the Pleistocene at around 100 thousand years ago.

Tables

Table 1 Summary statistics of the snake pipefish reference genome. The table includes information for **A** the raw read sequencing, and **B** the scaffold- and contig-level *de novo* assembly and **C** the BUSCO completeness statistics.

(A)	**Raw read statistics**	
No. short reads	**264,111,731**	
Mapped short reads (%)	**99.53**	
Mean short read coverage (x)	**23**	
No. long reads	**130,590,372**	
Mapped long reads (%)	**98.61**	
Mean long read coverage (x)	**205.2**	
(B)	**Assembly statistics (**scaffold/contig)	
No. scaffolds/contigs	**7,387**	**7,473**
No. scaffolds/contigs (>50 kbp)	**466**	**526**
scaffold/contig L50	**12**	**14**
scaffold/contig N50 (bp)	**62,341,166**	**45,010,074**
Total length (bp)	**1,662,053,046**	**1,662,035,846**
GC (%)	**38.87**	**38.87**
No. of N's per 100 kb	**1.03**	**0.0**
heterozygosity (%)	**0.387**	
Total interspersed repeats (bp)	**1,237,929,559 (74.93 %)**	
(C)	**BUSCO completeness**	
Clade: *Actinopterygii*	**C:94.1%[S:92.6%, D:1.5%]**	
	F:2.0%, M:3.9%	
	n:3640	

BUSCO: Benchmarking Universal Single Copy Orthologs (*65*); C, complete; S, single copy; D, duplicated; F, fragmented; M, missing.

Table 2 Repeat content of the genome assembly. Class, class of the repetitive regions. Count, number of occuences of the repetitive region. bpMasked, number of base pairs masked; %Masked, percentage of base pairs masked. LINE, Long Interspersed Nuclear Elements (include retroposons); LTR, Long Terminal Repeat elements (including retroposons); SINE, Short Interspersed Nuclear Elements; RC, Rolling Circle.

Class	Count	bpMasked	%masked
ARTEFACT	4	84	0.00%
DNA	2765297	372407739	22.40%
LINE	850222	167337419	10.06%
LTR	177214	55439687	3.33%
PLE	1	0	0.00%
RC	32348	3385084	0.20%
SINE	435464	32709572	1.95%
Unknown	3628328	534216084	32.14%
Low complexity	127733	3095322	0.19%
Satellite	21221	7145469	0.43%
Simple repeat	1437090	61077339	3.67%
rRNA	4394	534599	0.03%
scRNA	5	504	0.00%
snRNA	695	46845	0.00%
tRNA	6029	533812	0.03%
Total	9486045	1237929559	74.93%