Journal of Heredity, 2023, **114**, 189–194
https://doi.org/10.1093/jhered/esad003
Advance access publication 20 January 2023
Genome Resources

American Genetic Association
OXFORD

Genome Resources

A chromosome-scale reference genome assembly of the great sand eel, *Hyperoplus lanceolatus*

Sven Winter[1,2,3], Jordi de Raad[2,4], Magnus Wolf[1,2], Raphael T.F. Coimbra[1,2], Menno J. de Jong[2], Yannis Schöneberg[1,2], Maria Christoph[1], Hagen von Klopotek[1], Katharina Bach[1], BehgolPashm Foroush [1], Wiebke Hanack[1], Aaron Hagen Kauffeldt[1], Tim Milz[1], Emmanuel Kipruto Ngetich[1], Christian Wenz[1], Moritz Sonnewald[5], Maria Anna Nilsson[2,4], Axel Janke[1,2,4]

[1]Institute for Ecology, Evolution, and Diversity, Goethe University, Frankfurt am Main, Germany,
[2]Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany,
[3]Research Institute of Wildlife Ecology, University of Veterinary Medicine, Vienna, Austria,
[4]LOEWE-Centre for Translational Biodiversity Genomics, Frankfurt am Main, Germany,
[5]Senckenberg Research Institute, Department of Marine Zoology, Section Ichthyology, Frankfurt am Main, Germany
Address correspondence to S. Winter at the address above, or e-mail: sven.winter@senckenberg.de.

Corresponding Editor: Alexander Suh

Abstract

Despite increasing sequencing efforts, numerous fish families still lack a reference genome, which complicates genetic research. One such understudied family is the sand lances (Ammodytidae, literally: “sand burrower”), a globally distributed clade of over 30 fish species that tend to avoid tidal currents by burrowing into the sand. Here, we present the first annotated chromosome-level genome assembly of the great sand eel (*Hyperoplus lanceolatus*). The genome assembly was generated using Oxford Nanopore Technologies long sequencing reads and Illumina short reads for polishing. The final assembly has a total length of 808.5 Mbp, of which 97.1% were anchored into 24 chromosome-scale scaffolds using proximity-ligation scaffolding. It is highly contiguous with a scaffold and contig N50 of 33.7 and 31.3 Mbp, respectively, and has a BUSCO completeness score of 96.9%. The presented genome assembly is a valuable resource for future studies of sand lances, as this family is of great ecological and commercial importance and may also contribute to studies aiming to resolve the suprafamiliar taxonomy of bony fishes.

Key words: Ammodytidae, Eupercaria, Omni-C, Oxford Nanopore, proximity-ligation scaffolding

Introduction

The great sand eel *Hyperoplus lanceolatus* (Le Sauvage, 1824) (Fig. 1) is a coastal species that is distributed in the northeastern Atlantic, more particularly on the European continental shelves between Portugal and Murmansk and the Baltic Sea, at a maximum depth of 60 m (Rutkowicz 1982). The species occurs along the continental coastline and around islands, most notably Iceland, Svalbard, and the British Isles (Rutkowicz 1982; Nadolna-Ałtyn *et al.* 2017).

Sand eels are commercially and ecologically important due to their high abundance and high-fat content. Natural predators include sea mammals, piscivorous birds, and predatory fish. Industrial fisheries target the species for fish meal and oil production, while small-scale fisheries aim for human consumption and fishing bait (Frimodt 1995). Therefore, concerns have been raised about the potentially detrimental effects of sand eel stock depletion on the marine food web (Dunn 2021).

The great sand eel is included in the family of sand lances (Ammodytidae), which contains 33 species in 7 genera (Fricke *et al.* 2022). Sand lances feed primarily on small crustaceans and small fishes and are characterized by an elongated body with long dorsal fins, reduced or missing pelvic fins, and the absence of a swim bladder (Muus and Nielsen 1999). The latter trait is likely an adaptation to a burrowing lifestyle (Muus and Nielsen 1999).

Besides the great sand eel, the genus *Hyperoplus* includes one other species, the less common Corbin’s sand eel (*H. immaculatus*), which also occurs in the northeastern Atlantic. These 2 species can be distinguished from other sand lances by their 2 sharply pointed vomerine teeth and by the relatively short pectoral fins, which do not extend to the base of the dorsal fin. Within the genus itself, the great sand eel can be distinguished from the Corbin’s sand eel by its larger size (up to 20 to 40 cm length) and a species-specific dark spot on either side of the snout below the anterior nostril (Reay 1986). The great sand eel is also more piscivorous, sometimes even feeding on other sand lances (Frimodt 1995).

Received October 24, 2022; Accepted January 18, 2023

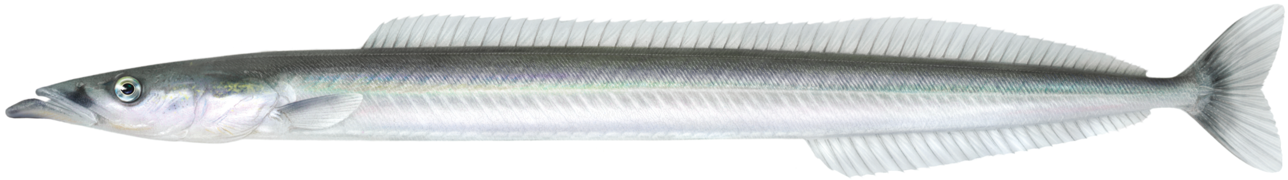

Fig. 1. *Hyperoplus lanceolatus*. Painting by Jan Fekjan. https://artsdatabanken.no/taxon/Hyperoplus%20lanceolatus/43111.

The sand lance family was originally classified as part of the large order Perciformes but has recently been moved into other orders, either Trachiniformes or Uranoscopiformes (Nelson *et al.* 2016; Betancur-R *et al.* 2017). These taxonomic revisions illustrate the uncertainty surrounding the phylogenetic relationships within the series Eupercaria as a whole, many of which are still unresolved and in need of clarification through genetic studies (Betancur-R *et al.* 2017).

To date, the only genetic data available for the great sand eel are mitochondrial gene sequences. As part of a master's course at the Goethe University in Frankfurt am Main, Germany (Prost *et al.* 2020), we generated a de novo, chromosome-level genome assembly of *H. lanceolatus*. This genome has been assembled from Nanopore long reads, polished with Illumina short reads, and scaffolded into chromosomes with Omni-C proximity-ligation data. The genome assembly represents the first in the genus *Hyperoplus* and the second within the family of sand lances after the recently published *Ammodytes dubius* assembly (Jones *et al.* 2023) and may facilitate future studies which aim to resolve the suprafamiliar taxonomy of sand lances or to evaluate fisheries' effects on individual species.

Materials and methods

Sampling, DNA extraction, and sequencing

Two adult *H. lanceolatus* individuals were collected in the North Sea during a regular monitoring expedition to the Dogger Bank (Hlan001: N 54°59′37.0608, E 2°56′26.9772; Hlan002: N 55°1′30.054, E 1°34′57.2952) with the permission of the Maritime Policy Unit of the UK Foreign and Commonwealth Office in 2020. One specimen (Hlan001) was initially frozen at −20 °C on the ship and later stored at −80 °C until further processing. High molecular weight genomic DNA of this individual was extracted from muscle tissue using the protocol of Mayjonade *et al.* (2016) with the addition of Proteinase K during lysis. We used the Genomic DNA ScreenTape on the Agilent 2200 TapeStation system (Agilent Technologies) to evaluate DNA quantity and quality. In addition, we dissected the second specimen (Hlan002) during the expedition and preserved tissues from different inner organs (brain, heart, gills, muscle, liver, gonads, and pyloric gland) in RNALater for RNA extraction. These tissue samples, along with a DNA sample from the first individual, were sent to Novogene (UK) Company Limited for RNA extraction and sequencing. A standard 150 base pair (bp) paired-end whole-genome sequencing library from genomic DNA was prepared using the NEBNext Ultra II library preparation kit for Illumina sequencing (New England Biolabs Inc., Ipswich, USA) and sequenced on a Novaseq 6000 Illumina platform (Illumina, Inc., San Diego, California, USA). In addition, short-read paired-end RNA-Seq libraries for each of the RNA extracts from the different tissue types were prepared and sequenced on the same Illumina platform.

Furthermore, we prepared five long-read libraries for sequencing on the Oxford Nanopore Technologies (ONT, Oxford, UK) MinION v.Mk1B sequencer following the protocol of ONTs Rapid Sequencing Kit (SQK-RAD004). Each library was sequenced on an individual flow cell (FLO-MIN106 v.9.41).

Lastly, we prepared a proximity-ligation library from muscle tissue using the Dovetail Omni-C Kit (Dovetail Genomics, Santa Cruz, California, USA). The library was sent to Novogene (UK) for sequencing on the Illumina Novaseq 6000.

Genome size estimation

The genome size of *H. lanceolatus* was estimated by k-mer frequencies. The frequencies of k-mers with k = 21 were calculated using Jellyfish v.2.2.10 (RRID: SCR_005491) (Marçais and Kingsford 2011) from the short-read Illumina data. The genome size and heterozygosity were then estimated with GenomeScope v.2.0 (RRID: SCR_017014) (Vurture *et al.* 2017; Ranallo-Benavidez *et al.* 2020).

Genome assembly and polishing

The raw Nanopore sequencing signal data (fast5) was base called with Guppy v.4.0.14 (Oxford Nanopore Technologies Ltd.) on the high-accuracy setting. Adapter sequences were removed with Porechop v.0.2.4 (RRID: SCR_016967) (Wick *et al.* 2017). The read length distribution and base quality of the 5 sequencing runs were analyzed both independently with Nanocomp v.1.0.0 (De Coster *et al.* 2018) and combined using Nanoplot v.1.0.0 (De Coster *et al.* 2018). The resulting long-read dataset was used to assemble the genome of *H. lanceolatus* with WTDBG2 v.2.5 (RRID: SCR_01722) (Ruan and Li 2019) using the preset for ONT reads (flag "*-x ont*"). The accuracy of the resulting assembly was further improved by a 3-step polishing approach. First, 3 iterations of long-read polishing were performed with racon v.1.4.3 (RRID: SCR_017642) (Vaser *et al.* 2017), followed by 1 iteration of long-read polishing with Medaka v.0.11.5 (Oxford Nanopore Technologies Ltd. 2018) to correct for errors typical for nanopore sequencing (homopolymers and repeat errors). Finally, we used Illumina short reads to correct for single-base errors and short indels in the assembly using 3 iterations of pilon v.1.23 (RRID: SCR_014731) (Walker *et al.* 2014). We also assembled the mitochondrial genome of *H. lanceolatus* from the short-read Illumina data using MitoZ v.2.4 (Meng *et al.* 2019). The resulting circular genome was then annotated using MitoAnnotator v.3.75 (Iwasaki *et al.* 2013).

Scaffolding and quality assessment

To anchor the contigs into chromosome-scale scaffolds, we used the Dovetail Genomics scaffolding service. For that, we sent the Omni-C data, generated in this study, and the polished assembly to Dovetail Genomics as input for the HiRise pipeline (Putnam *et al.* 2016). Afterwards, gaps in the scaffolded assembly were filled using TGS-GapCloser v.1.1.1 (RRID: SCR_017633) (Xu *et al.* 2020) using the same long reads used for the initial assembly. Finally, haplotypic duplications (haplotigs) were identified and removed using purge_dups v.1.2.5 (RRID: SCR_021173) (Guan *et al.* 2020) adjusting the command for minimap2 in the pipeline to the preset for ONT reads.

The contiguity and completeness of the final assembly were evaluated using Quast v.5.0.2 (RRID: SCR_001228) (Mikheenko *et al.* 2018) and BUSCO v.5.3.1 (RRID: SCR_015008) (Seppey *et al.* 2019) with the Actinopterygii specific orthologous gene set (actinopterygii_odb10). The completeness, base-level error rate, and consensus quality value (QV) of the assembly were evaluated with Merqury v.1.1 (Rhie *et al.* 2020). Furthermore, we mapped the short and long reads onto the assembly using BWA-mem v.0.7.17-r1188 (RRID: SCR_010910) (Li 2013) and Minimap2 v.2.17-r941 (RRID: SCR_018550) (Li 2018), respectively, to analyze mapping quality and rate with Qualimap v.2.2.1 (RRID: SCR_001209) (Okonechnikov *et al.* 2016). To assess potential contamination of the assembly, we used the previously generated mapping files and a BLASTN v.2.11.0+ (RRID: SCR_001598) (Zhang *et al.* 2000) search of the assembly against the NCBI nucleotide database to generate a BlobPlot with Blobtoolkit v.3.5.0 (RRID: SCR_017618) (Laetsch and Blaxter 2017).

Transcriptome assembly and quality assessments

In addition to the genome assembly, we assembled the transcriptome of *H. lanceolatus*. The RNA-seq data for the 7 different tissue types were combined into a single dataset and used to assemble the transcriptome with Trinity v2.9.0 (RRID: SCR_013048) (Grabherr *et al.* 2011; Haas *et al.* 2013) following the step-by-step protocol of (Freedman and Weeks 2020). The completeness of the transcriptome assembly and assembly statistics were assessed with BUSCO v5.3.1 (RRID: SCR_015008) (Seppey *et al.* 2019) and Quast v.5.0.2 (RRID: SCR_001228) (Mikheenko *et al.* 2018) using the same settings as described previously.

Genome annotation

Repeat annotation

To annotate repeats in the assembly, we first generated a de novo repeat library with RepeatModeler v.2.0.1 (RRID: SCR_015027) (Flynn *et al.* 2020), which was combined with an Actinopterygii database, derived from the RepeatMasker v.4.1.0 (http://www.repeatmasker.org/RepeatMasker/; RRID: SCR_012954) Repeat Sequence Database using the utility script "queryRepeatDatabase.pl," to a custom repeat library. Then RepeatMasker was used with the custom library to annotate and mask the repeats in the assembly. We hard-masked interspersed repeats and soft-masked simple repeats to increase the accuracy of the subsequent gene annotation.

Gene annotation

Homology-based gene prediction was performed with the GeMoMa pipeline v.1.7.1 (RRID: SCR_017646) (Keilwagen et al. 2016, 2018) using MMseqs2 v. 13.45111 (Steinegger and Söding 2017) as the alignment tool. The following 5 genomes and annotation files (GFF format) were used as references: *Acanthochromis polyacanthus* (GCA_002109545.1), *Perca fluviatilis* (GCA_010015445.1), *Gasterosteus aculeatus* (GCA_016920845.1), *Betta splendens* (GCA_900634795.3), and *Acanthopagrus latus* (GCA_904848185.1). In addition, corrected and trimmed RNA-seq reads generated during the Trinity transcriptome assembly process (Freedman and Weeks 2020) were mapped against the genome assembly with STAR v.2.7.9a (RRID: SCR_004463) (Dobin *et al.* 2013) and used as RNA-seq evidence during the annotation. The genes predicted by GeMoMa were further annotated by a search against the Swiss-Prot database (RRID: SCR_002380; release 2022-02) using BLASTP v.2.11.0+ (RRID: SCR_001010) (Zhang *et al.* 2000), applying an *e*-value cutoff of 10^{-6}. Further annotation of Gene Ontology (GO) terms, domains, and motifs was conducted with InterProScan v. 5.50.84 (RRID: SCR_005829) (Quevillon *et al.* 2005; Jones *et al.* 2014).

Results and discussion

Genome sequencing and assembly

The 5 sequencing runs on the ONT MinION generated a total of 32 Gbp or an approximate 40-fold coverage of long-read data with a mean read length of 4.56 kbp and a mean read quality of 12 (Supplementary Table 1A, Supplementary Fig. 1). Illumina whole-genome and Omni-C sequencing generated 42.3 and 43.1 Gbp of short-read and proximity-ligation data, respectively (Supplementary Table 1B).

The final chromosome-scale scaffolded, gap-closed, and haplotig-purged de novo genome assembly of *H. lanceolatus* has 965 scaffolds (incl. mitochondrial genome), a length of 808.5 Mbp, and 6 gaps of 100 N's each resulting in a scaffold/contig N50 of 33.7 and 31.3 Mbp, respectively (Table 1). Proximity-ligation scaffolding resulted in 97.1% of the total assembly length being anchored into 24 chromosome-scale scaffolds larger than 15 Mbp (Fig. 2A and B), which is the expected haploid number of exclusively acrocentric chromosomes (2*n* = 48) described for the species (Ocalewicz *et al.* 2019). The

Table 1. Assembly statistics for the *Hyperoplus lanceolatus* genome and transcriptome assemblies.

	Scaffold-level assembly	Contig-level assembly[a]	Transcriptome
No. of scaffolds/ contigs	965	970	118,200
No of scaffolds/ contigs (>1 kbp)	879	884	69,268
Total length (bp)	808,503,945	808,503,345	223,612,403
Largest scaffold/ contig (bp)	43,769,752	43,769,752	31,570
N50 (bp)	33,676,134	31,308,663	2,816
L50	11	12	23,663
GC%	41.09	41.09	46.28
No. of N's	600	0	0
No. of N's per 100 kbp	0.07	0	0

[a] Statistics in this column are based on contigs as the assembly was broken into contigs at gaps with a length ≥10 bp. Statistics for the remaining columns are based on scaffolds.

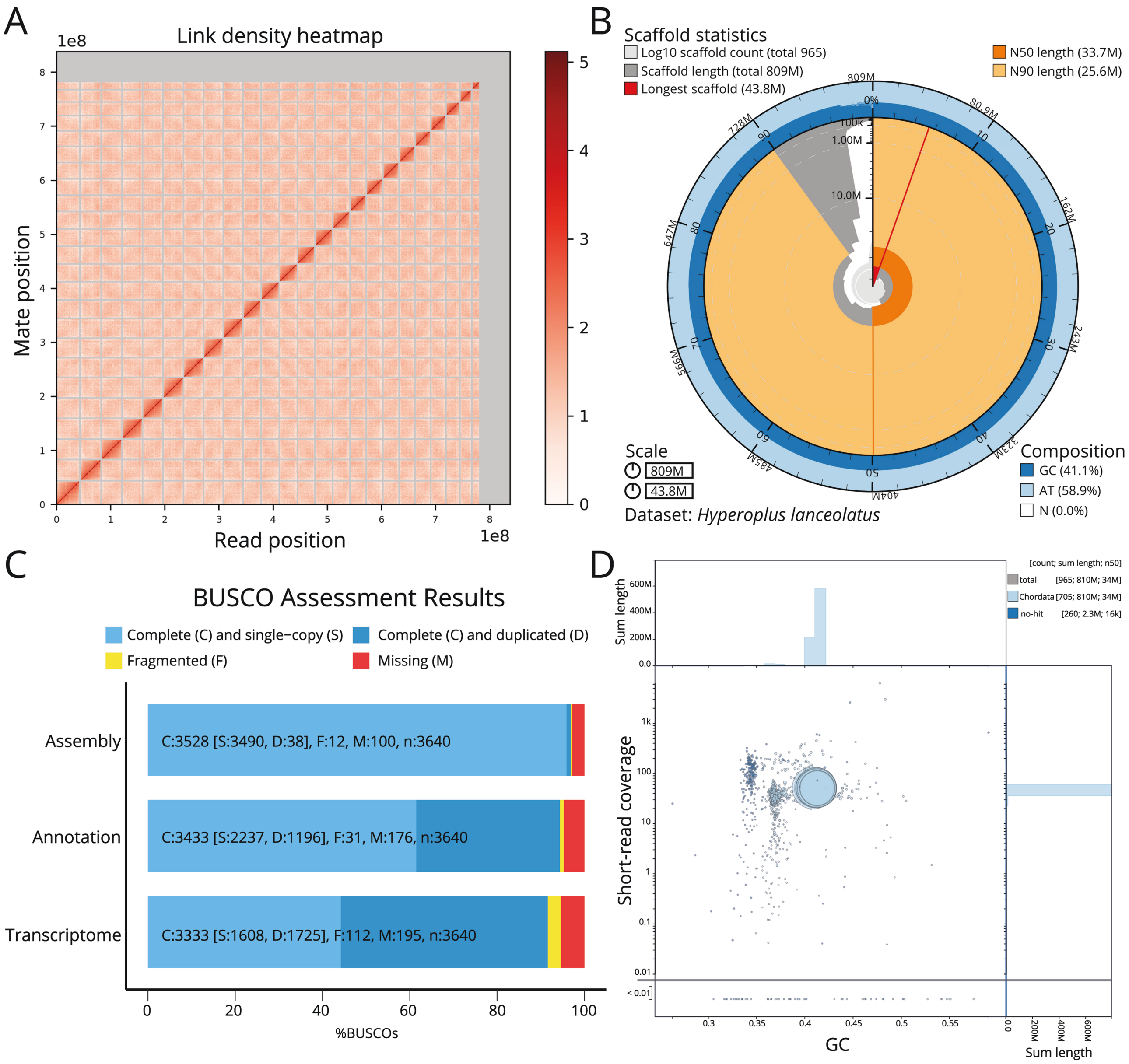

Fig. 2. Quality assessment of the *Hyperoplus lanceolatus* genome assembly. An Omni-C contact density map depicting the 24 distinct chromosome-level scaffolds (A). SnailPlot summarizing assembly statistics (B). Gene set completeness analyses for the assembly, annotation, and transcriptome (C). BlobPlot analysis comparing GC content (*x* axis), sequencing depth of short reads (*y* axis), and taxonomic assignment of contigs (colors) show no evidence of contamination (D).

remaining 2.9% are comprised of scaffolds/contigs smaller than 400 kbp. The separately conducted mitochondrial genome assembly resulted in a circular mitochondrial sequence with a length of 16,509 bp, which conforms to the standard vertebrate gene organization (Supplementary Fig. 2).

Genome completeness and quality assessment

The heterozygosity and haploid genome size of *H. lanceolatus* was estimated by GenomeScope as 0.48% and 695 Mbp, respectively, which is about 113 Mbp shorter than the length of the haplotig-free assembly.

A high percentage of identified complete BUSCO genes (96.9%) (Fig. 2C) of the Actinopterygii dataset and the k-mer completeness of 91.5% calculated by Merqury suggest an overall high completeness of the assembly. In addition, Merqury also suggests a low base-level error rate of 0.04% and a corresponding QV of 33.6. Furthermore, both long and short reads mapped to the assembly with a high mapping rate of 94.8% and 98.9%, respectively, and the BlobPlot generated with Blobtoolkit shows no clear evidence for contamination (Fig. 2D). Yet, a congregation of “no-hit” and “Chordata” scaffolds with a lower GC content (~34%) compared with the chromosome-scale scaffolds (41%) might be a sign of contamination of unknown origin due to a lack of sequences in the nucleotide database or simply scaffolds containing AT-rich repeats that were not placed into the chromosome-scale scaffolds.

Transcriptome assembly

The final transcriptome assembly is based on 50.9 Gbp of short-read RNA-seq data (Supplementary Table 1B) and has a total length of 223.6 Mbp (Table 1). BUSCO analyses found 91.6% of Actinopterygii orthologous genes in the transcriptome, indicating high transcriptome completeness (Fig. 2C).

Annotation

Repeat annotation

The de novo repeat library generated by RepeatModeler2 was comprised of 2,515 sequences (for details, see Supplementary Table 2). The annotation of repetitive elements in the genome identified 44.37% of the genome assembly of *H. lanceolatus* (359 Mbp) as repeats (Supplementary Table 3). DNA transposons were found to be the most common repeat elements spanning 16.6% of the genome, followed by Long Interspersed Nuclear Elements (LINEs) with 6.1% and simple repeats with 4.5%. However, a large percentage of repeats, spanning 13.6% of the genome, could not be classified.

Gene annotation

The homology-based gene prediction with GeMoMa identified 22,274 genes with a median length of 6,597 bp spanning 294.8 Mbp of the assembly. BUSCO analysis found 94.4% complete orthologous of the Actinopterygii dataset indicating high completeness of the predicted genes (Fig. 2C). Furthermore, InterProScan functionally annotated 50,694 (99.5%) of the 50,935 predicted proteins and assigned at least 1 GO term to 39,171 (76,9%) of the proteins. In addition, 43,600 proteins (95,2%) were assigned to entries within the Swiss-Prot database.

Conclusion

The chromosome-level reference assembly of *H. lanceolatus* presented here is not only the second genome assembly for the family Ammodytidae but, in fact, also the second of the order *Uranoscopiformes* that contains approximately 174 recognized species (Encyclopedia of Life, http://eol.org, 2018). It will be an invaluable resource for future phylogenomic and population genomic studies of sand lances and bony fishes in general, as the systematics of Eupercaria is not fully resolved yet (Betancur-R *et al.* 2017). In addition, it is an important reference for genomic assessments of fisheries stocks, as sand lances are a valuable resource and play an irreplaceable role in the survival and breeding success of many seabirds (Frimodt 1995; Dunn 2021).

Supplementary material

Supplementary material is available at *Journal of Heredity* online.

Funding

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

Acknowledgments

We thank Carola Greve and Damian Baranski from the TBG laboratory center for laboratory support.

Conflict of interest

None declared.

Data availability

All underlying read data, the assembly, and the transcriptome are available at GenBank under BioProject PRJNA835307. SRA accession numbers for each sequencing dataset are listed in Supplementary Table 1. The annotation, assembly, transcriptome, and all commands used in preparing the data are also available at Dryad (https://doi.org/10.5061/dryad.7pvmcvdxv).

References

Betancur-R R, Wiley EO, Arratia G, Acero A, Bailly N, Miya M, Lecointre G, Ortí G. Phylogenetic classification of bony fishes. *BMC Evol Biol*. 2017;17(1):162.

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*. 2018;34(15):2666–2669.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.

Dunn E. Revive our Seas: the case for stronger regulation of sandeel fisheries in UK waters. The Royal Society for the Protection of Birds (RSPB); 2021 [accessed 2022 Jun 1]. https://www.rspb.org.uk/globalassets/downloads/documents/campaigning-for-nature/rspb2021_the-case-for-stronger-regulation-of-sandeel-fisheries-in-uk-waters.pdf

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 USA*. 2020;117(17):9451–9457.

Freedman A, Weeks N. Best practices for de novo transcriptome assembly with Trinity. Harvard FAS Informatics; 2020 [accessed 2022 Jul 1]. https://informatics.fas.harvard.edu/./best-practices-for-de-novo-transcriptome-assembly-with-trinity.html

Fricke R, Eschmeyer WN, Van der Laan R. Eschmeyer's catalog of fishes: genera, species, references. 2022 [accessed 2022 Jun 1]. https://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp

Frimodt C. *Multilingual illustrated guide to the world's commercial coldwater fish*. Oxford, UK: Fishing News Books Ltd; 1995.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644–652.

Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 2020:36(9):2896–2898.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 2013;8(8):1494–1512.

Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, Sado T, Mabuchi K, Takeshima H, Miya M, et al. MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Mol Biol Evol*. 2013;30(11):2531–2540.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014;30(9):1236–1240.

Jones LF, Lou RN, Murray CS, Robert D, Bourne CM, Bouchard C, Kučka M, Chan YF, Carlon DB, Wiley DN, et al. Two distinct population clusters of northern sand lance (*Ammodytes dubius*) on the northwest Atlantic shelf revealed by whole genome sequencing. *ICES J Mar Sci*. 2023;80(1):122–132. doi:10.1093/icesjms/fsac217

Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics*. 2018:19(1). doi:10.1186/s12859-018-2203-5

Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res*. 2016;44(9):e89.

Laetsch DR, Blaxter ML. BlobTools: interrogation of genome assemblies. *F1000Research*. 2017:6:1287.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv, arXiv:1303.3997, 2013. doi:10.48550/arXiv.1303.3997.

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094–3100.

Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;27(6):764–770.

Mayjonade B, Gouzy J, Donnadieu C, Pouilly N, Marande W, Callot C, Langlade N, Muños S. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *Biotechniques*. 2016;61(4):203–205.

Meng G, Li Y, Yang C, Liu S. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Res*. 2019;47(11):e63.

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*. 2018;34(13):i142–i150.

Muus BJ, Nielsen JG. 1999. Sea fish. Scandinavian Fishing Year Book, *Hedehusene, Denmark*. p. 340.

Nadolna-Ałtyn K, Podolska M, Szostakowska B. Great sandeel (*Hyperoplus lanceolatus*) as a putative transmitter of parasite *Contracaecum osculatum* (Nematoda: Anisakidae). *Parasitol Res*. 2017;116(7):1931–1936.

Nelson JS, Grande TC, Wilson MVH. *Fishes of the world*. Hoboken, New Jersey, USA: John Wiley & Sons; 2016.

Ocalewicz K, Kirtiklis L, Mojsa K, Sapota M, Kwiatkowski M. First description of karyotypes and localization of ribosomal genes in two sand lances (Uranoscopiformes: Ammodytidae); small sand-eel (Ammodytes tobianus Linnaeus, 1758) and great sand-eel (*Hyperoplus lanceolatus* Le Sauvage, 1824). *Mar Biol Res*. 2019;15(8–9):523–529.

Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*. 2016;32(2):292–294.

Oxford Nanopore Technologies Ltd. Medaka [Python]. Oxford Nanopore Technologies; 2018 [accessed 2022 Jul 6]. https://github.com/nanoporetech/medaka

Prost S, Winter S, De Raad J, Coimbra RTF, Wolf M, Nilsson MA, Petersen M, Gupta DK, Schell T, Lammers F, et al. Education in the genomics era: generating high-quality genome assemblies in university courses. *GigaScience*. 2020;9(6). doi:10.1093/gigascience/giaa058

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res*. 2016;26(3):342–350.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. InterProScan: protein domains identifier. *Nucleic Acids Res*. 2005;33(suppl 2):W116–W120.

Ranallo-Benavidez TR, Jaron, KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 2020;11(1). doi:10.1038/s41467-020-14998-3

Reay PJ. Ammodytidae. In: Whitehead PJP, Beauchot ML, Hureau JC, Nielsen J, Tortonese E, editors. *Fishes of the Northeastern Atlantic and Mediterranean*. Paris, France: UNESCO; 1986. p. 945–950.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;21(1):245.

Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods*. 2019;17(2):155–158.

Rutkowicz S. *Encyklopedia ryb morskich*. Gdańsk, Poland; Morskie Gdańsk; 1982.

Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol*. 2019;1962:227–245.

Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. 2017;35(11):1026–1028.

Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 2017;27(5):737–746.

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(14):2202–2204.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 2014;9(11):e112963.

Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genomics*. 2017;3(10):e000132.

Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, Fan G, Liu X, Xu X, Deng L, et al. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience*. 2020;9(giaa094). doi:10.1093/gigascience/giaa094

Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol*. 2000;7(1–2):203–214.