GENOME NOTE

Inference of a genome-wide protein-coding gene set of the inshore hagfish *Eptatretus burgeri* [version 1; peer review: 2 approved with reservations]

Osamu Nishimura[1*], Kazuaki Yamaguchi[1*], Yuichiro Hara[1,2*], Kaori Tatsumi[1*], Jeramiah J Smith[3], Mitsutaka Kadota[1], Shigehiro Kuraku[1,4,5]

[1]Laboratory for Phyloinformatics, RIKEN Biosystems Dynamics Research, Kobe, Hyogo, 650-0047, Japan
[2]Research Center for Genome & Medical Sciences, Tokyo Metropolitan Institute of Medical Science (TMiMS), Tokyo, Japan
[3]Department of Biology, University of Kentucky, Lexington, KY, 40506, USA
[4]Molecular Life History Laboratory, Department of Genomics and Evolutionary Biology, National Institute of Genetics, Mishima, Shizuoka, 411-8540, Japan
[5]Department of Genetics, Sokendai (Graduate University for Advanced Studies), Mishima, Shizuoka, 411-8540, Japan

* Equal contributors

v1 **First published:** 08 Nov 2022, **11**:1270
https://doi.org/10.12688/f1000research.124719.1
Latest published: 08 Nov 2022, **11**:1270
https://doi.org/10.12688/f1000research.124719.1

Abstract

The hagfishes (Myxiniformes) arose from agnathan (jawless vertebrate) lineages and they are one of only two extant cyclostome taxa, together with lampreys (Petromyzontiformes). Even though whole genome sequencing has been achieved for diverse vertebrate taxa, genome-wide sequence information has been highly limited for cyclostomes. Here we sequenced the genome of the inshore hagfish *Eptatretus burgeri* using DNA extracted from the testis, with a short-read sequencing platform, aiming to reconstruct a high-coverage protein-coding gene catalogue. The obtained genome assembly, scaffolded with mate-pair reads and paired RNA-seq reads, exhibited an N50 scaffold length of 293 Kbp, which allowed the genome-wide prediction of coding genes. This computation resulted in the gene models whose completeness was estimated at the complete coverage of more than 83 % and the partial coverage of more than 93 % by referring to evolutionarily conserved single-copy orthologs. The high contiguity of the assembly and completeness of the gene models promise a high utility in various comparative analyses including phylogenomics and phylome exploration.

Keywords

hagfish, cyclostome, whole genome assembly, gene prediction

Open Peer Review

Approval Status ? ?

	1	2
version 1 08 Nov 2022	? view	? view

1. **Vincent Laudet**, Okinawa Institute of Science and Technology, Okinawa, Japan

2. **Daniel Ocampo Daza**, CTC Clinical Trial Consultants AB, Uppsala, Sweden

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

This article is included in the Genomics and Genetics gateway.

Corresponding author: Shigehiro Kuraku (skuraku@nig.ac.jp)

Author roles: Nishimura O: Data Curation, Investigation, Methodology, Resources, Validation, Writing – Review & Editing; **Yamaguchi K**: Data Curation, Methodology, Resources, Writing – Review & Editing; **Hara Y**: Data Curation, Formal Analysis, Investigation, Methodology, Resources, Writing – Review & Editing; **Tatsumi K**: Methodology, Resources, Writing – Review & Editing; **Smith JJ**: Methodology, Resources, Writing – Review & Editing; **Kadota M**: Data Curation, Methodology, Resources, Writing – Review & Editing; **Kuraku S**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This study was supported by RIKEN and JSPS KAKENHI Grant Numbers 17K07426 and 20H03269 to S.K. and an NSF Grant Number MCB-1818012 to J.J.S.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Nishimura O, Yamaguchi K, Hara Y *et al.* **Inference of a genome-wide protein-coding gene set of the inshore hagfish *Eptatretus burgeri* [version 1; peer review: 2 approved with reservations]** F1000Research 2022, **11**:1270 https://doi.org/10.12688/f1000research.124719.1

First published: 08 Nov 2022, **11**:1270 https://doi.org/10.12688/f1000research.124719.1

Introduction

Extant jawless fishes (cyclostomes) are divided into two groups, hagfishes (Myxiniformes) and lampreys (Petromyzontiformes).[1] They have been studied from various viewpoints mainly because they occupy an irreplaceable phylogenetic position among the extant vertebrates, having diverged from all other vertebrates during the early Cambrian period. Even after massive efforts of whole genome sequencing for invertebrate deuterostomes,[2,3] genome-wide sequence information for species in this irreplaceable taxon was limited until the genome analyses for two lamprey species, the sea lamprey *Petromyzon marinus* and the Arctic lamprey *Lethenteron camtschaticum*, which were published in 2013.[4,5]

In parallel, biological studies involving individual genes have been conducted for both lampreys and hagfishes. Developmental biologists, in particular, have largely relied on lampreys whose embryonic materials are accessible through artificial fertilization,[6] whereas studies on hagfishes have been limited to non-embryonic materials, with a few notable exceptions.[7–9] This type of molecular biological study is expected to be more thoroughly performed if a comprehensive catalogue of genes is available. For lampreys, derivation of a reliable comprehensive gene catalogue was long hindered by the peculiar nature of protein-coding sequences, which are characterized by high GC-content, codon usage bias, and biased amino acid compositions.[5,10,11] To reinforce existing resources for lampreys, we previously performed a dedicated gene prediction for *L. camtschaticum*[12] and provided a gene catalogue with comparable or superior completeness to other equivalent resources.[4,13]

As of July 2022, no whole genome sequence information is available for hagfishes except for the one at Ensembl[13] that remains unpublished, a fact that hinders the comprehensive characterization of gene repertoires and their expression patterns. Currently, some efforts for genome sequencing and analysis are ongoing that aim to resolve large-scale evolutionary and epigenomic signatures,[9] inspired partly by the relevance of hagfish to understanding patterns of whole genome duplications[14–19] and chromosome elimination.[20–22] In contrast to those efforts, which are necessarily targeting reconstruction of the genome at chromosome scale, in this study we aimed at providing a data set covering as many full-length protein-coding genes as possible, to enable gene-level analysis on molecular function and evolution of hagfishes, an indispensable component of the vertebrate diversity.

Methods

Genome sequencing

A 48cm-long adult male individual of *Eptatretus burgeri* caught at the Misaki Marine Station in June 2013 was used for the study. After anesthetization in 1% tricaine and decapitation, the testis was sampled from which genomic DNA was extracted with the conventional phenol/chloroform extraction method,[23] and the genome sequencing was performed as outlined in Figure 1. The study was conducted with all efforts to ameliorate any suffering of animals, in accordance with the institutional guideline Regulations for the Animal Experiments by the Institutional Animal Care and Use Committee (IACUC) of the RIKEN Kobe Branch. The extracted genomic DNA was sheared using an S220 Focused-ultrasonicator (Covaris), which allowed us to retrieve DNA fragments of variable length distributions. Table 1 includes detailed information on amounts of starting DNA as well as conditions for shearing. The sheared DNA was subjected to paired-end library preparation using the KAPA LTP Library Preparation Kit (KAPA Biosystems). The optimal number of PCR cycles for library amplification was determined by quantitative PCR based on SYBR Green, using the KAPA Real-Time Library Amplification Kit (KAPA Biosystems) with Illumina library compatible primers (5′-AATGATACGGCGACC ACCGA-3′ and 5′-CAAGCAGAAGACGGCATACGA-3′), and an aliquot of adaptor-ligated DNA,[24] at 98°C for 45 sec followed by 25 cycles of amplification at 98°C for 15 seconds, 60°C for 30 seconds, and 72°C for 30 seconds, in ABI 7900HT Real Time PCR system (Thermo Fisher Scientific). The optimal Ct value was determined in SDS 2.4 software (Thermo Fisher Scientific) as cycles that reaches FS1 (Fluorescent Standard 1) but does not exceed FS2 (Fluorescent Standard 2). Libraries were further size selected using Agencourt AMPure XP (Beckman Coulter). Mate-pair libraries were prepared using the Nextera Mate Pair Sample Prep Kit (Illumina), employing our customized iMate protocol.[25] Detailed information of the paired-end and mate-pair libraries are described in Table 1. Libraries were quantified using the KAPA Library Quantification Kit (KAPA Biosystems) and sequenced on HiSeq 1500 (Illumina) operated by HiSeq Control Software v2.0.12.0 using HiSeq SR Rapid Cluster Kit v2 (Illumina) and HiSeq Rapid SBS Kit v2 (Illumina), or on HiSeq X (Illumina) operated by HiSeq Control Software v3.3.76, or on MiSeq operated by MiSeq Control Software v2.3.0.3 using the MiSeq Reagent Kit v3 (600 Cycles) (Illumina). Read lengths were 127 or 251 nt on HiSeq 1500, 151 nt on HiSeq X, and 251 nt on MiSeq. Base calling and generation of fastq files were performed with RTA v1.17.21.3 (Illumina, RRID:SCR_014332) and bcl2fastq v1.8.4 (Illumina, RRID:SCR_015058) for the sequencing data of HiSeq 1500 and MiSeq, or by RTA v2.7.6 and bcl2fastq v2.15.0 for the sequencing data of HiSeq X. Illumina adaptor sequences and low-quality bases were removed from the paired-end sequencing reads by Trim Galore v0.3.3 (RRID:SCR_011847) with the '--stringency 2 --quality 30 --length 25 --paired --retain_unpaired' options. Mate-pair reads were processed to identify the junction adaptor by NextClip v1.1[26] (RRID:SCR_005465) with the default parameters.

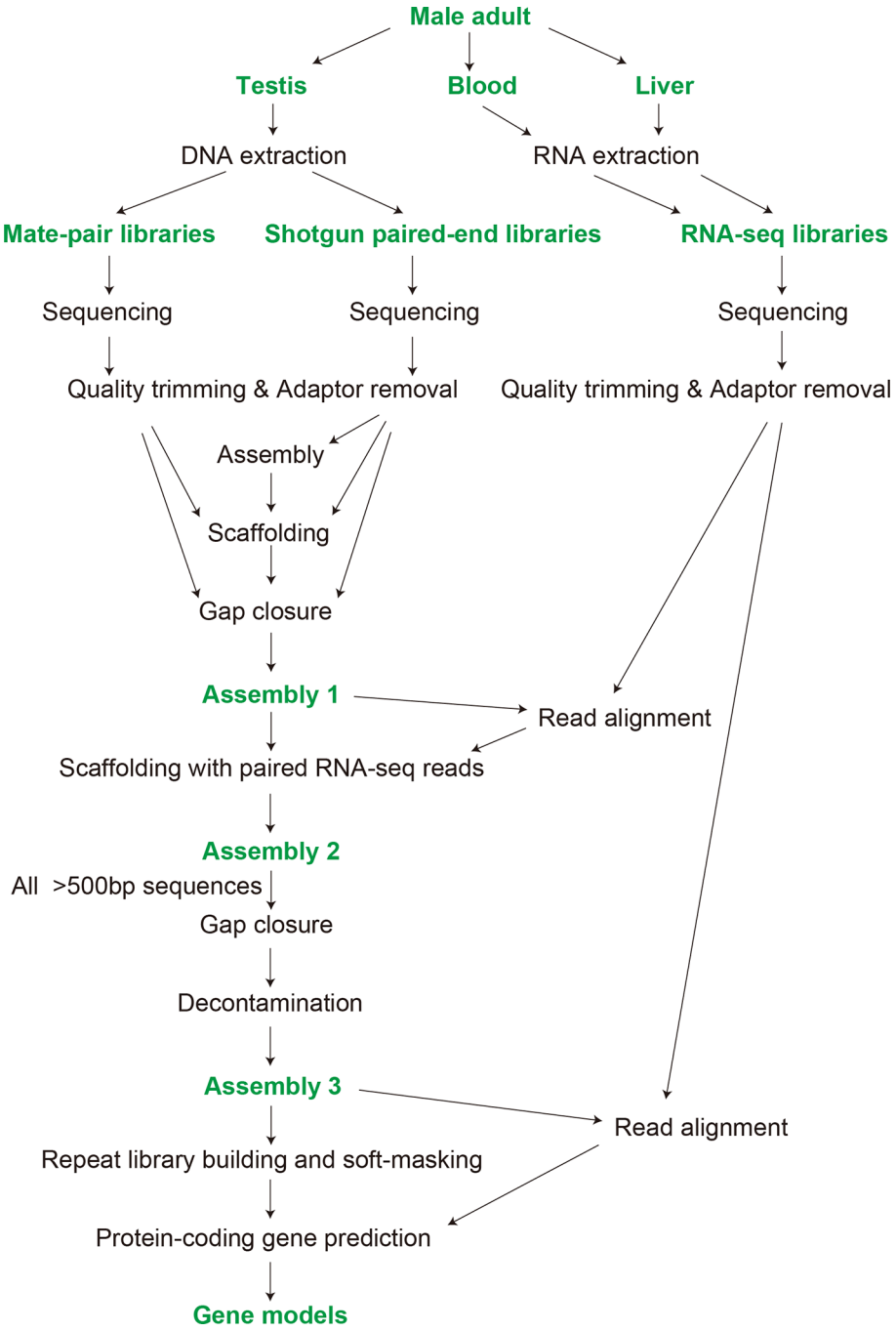

Figure 1. Data production workflow. Samples, raw data, and products are indicated with green letters, while computational steps are labelled in black. See Methods for the details including the choice of the programs used in individual computational steps.

RNA-seq and transcriptome data processing

Total RNAs were extracted from the liver tissue and the blood of the above-mentioned adult individual with Trizol reagent (Thermo Fisher Scientific) following the manufacturer's instruction. The RNA was treated with DNase I to digest genomic DNA. Quality control was performed with Bioanalyzer 2100 (Agilent Technologies), which yielded the RIN values of 8.7 and 9.1 for the respective tissues. Libraries were prepared with TruSeq Stranded mRNA LT Sample Prep Kit (Illumina).[27] The amount of total RNA used for library preparation and the number of PCR cycles applied for library amplification are described in Table 1 and Figure 2. Removal of Illumina adaptor sequences and low-quality bases was performed with Trim Galore v0.3.3 as outlined above. Alignment of the RNA-seq reads to the genome assembly was performed by HISAT2 v2.2.1[28] (RRID:SCR_015530) with the options '-k 3 -p 20 --pen-noncansplice 1000000'.

Table 1. Properties of prepared sequencing libraries.

A. Paired-end genome shotgun libraries					
Accession ID	**Library ID**	**Average insert size (bp)**	**Amount of DNA used (μg)**	**# PCR cycles**	**# read pairs**
DRX218807, DRX218808, DRX218809	P167_02_1	420	0.05	3	185,747,472
DRX218810, DRX218811, DRX218812, DRX218813	P167_02_2	690	0.05	5	174,756,277
DRX218814, DRX218815	P167_12_1	644	3	0	127,124,057
DRX218816, DRX218817, DRX218818	P167_12_2	873	3	4	278,906,224
DRX218819, DRX218820, DRX218821	P167_12_4	381	3	0	329,285,268
DRX218822, DRX218823, DRX218824	P167_12_5	418	3	2	129,764,303
B. Mate-pair genome libraries					
Accession ID	**Library ID**	**Mate distance (Kb)**	**Amount of DNA used (μg)**	**# PCR cycles**	**# read pairs**
DRX218825	P167_02_5	6-10	4	10	9,632,719
DRX218826	P167_02_6	12-18	4	13	10,230,697
DRX218827, DRX218828	P167_02_7	6-10	4	10	219,246,424
DRX218829, DRX218830	P167_02_8	12-18	4	13	139,814,746
C. RNA-seq libraries					
Accession ID	**Library ID**	**Tissue**	**Amount of total RNA used (μg)**	**# PCR cycles**	**# read pairs**
DRX218831	P238_01_1	Liver	1	6	55,675,220
DRX218832	P238_02_1	Blood	1	5	57,600,815

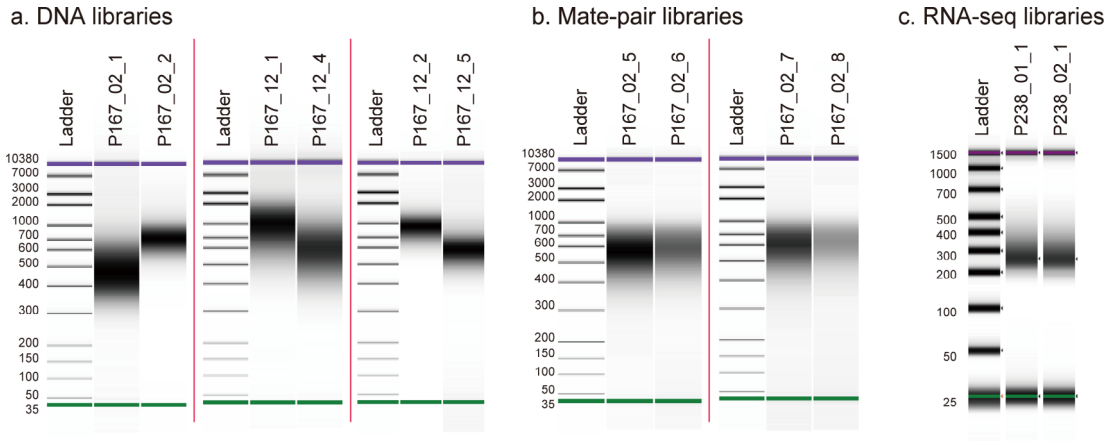

Figure 2. Size distribution of the sequencing libraries. a, Shotgun DNA libraries analyzed by Bioanalyzer High Sensitivity DNA Kit (Agilent). b, Mate-pair libraries analyzed by Bioanalyzer High Sensitivity DNA Kit. c, RNA-seq libraries analyzed by TapeStation High Sensitivity D1000 ScreenTape Assay Kit (Agilent).

Genome assembly

De novo genome assembly and scaffolding of Illumina short reads processed as described above were performed by the program PLATANUS v1.2.4[29] (RRID:SCR_015531) with its default parameters. The assembly employed paired-end reads and single reads whose pairs had been removed for quality filtering, and the scaffolding employed paired-end and mate-pair reads. The gap closure employed all of the single, paired-end, and mate-pair reads after processing. The

obtained sequences were further scaffolded with paired-end RNA-seq reads with the program P_RNA_Scaffolder[30] (commit 7941e0f on May 30, 2019, at GitHub) with the options '-s yes -b yes -p 0.90 -t 20 -e 100000 -n 100', followed by another gap closure run with PLATANUS 'gap_closure' using the same set of reads used in the above-mentioned gap closure run. The resultant genomic sequences were further screened for the species' own mitochondrial DNA fragments, contaminating organismal sequences, PhiX sequences loaded as a control in the Illumina sequencing system, and sequences shorter than 500 bp, as performed previously.[31]

Repeat detection and masking

To obtain species-specific repeat libraries, RepeatModeler v1.0.8[32] (RRID:SCR_015027) was executed with its default parameters. Repeat element detection in the genome sequence was performed by RepeatMasker v4.0.5[33] (RRID: SCR_012954), which employs the National Center for Biotechnology Information (NCBI) RMBlast v2.2.27[34] (RRID:SCR_022710), using the custom repeat library obtained above by RepeatModeler. Genomic regions detected as repeats were soft-masked by RepeatMasker with the '-nolow -xsmall' options.

Construction of gene models

Construction of gene models was performed by employing the gene prediction pipeline BRAKER v2.1.4[35] (RRID: SCR_018964) with the options '--min_contig=500 --prg=gth --softmasking --UTR=off' (Figure 1). This computation employed RNA-seq read alignments in BAM files onto the genome assembly and a set of peptide sequences prepared as follows. The set of peptide sequences used as homolog hints included the predicted proteins of the Arctic lamprey (34,362 sequences, previously designated as GRAS-LJ[12]), which were aligned to the soft-masked genome assembly.

Results

Genome assembly

Our technical procedure employing the genome assembly program PLATANUS[29] that previously produced genome assemblies for multiple shark species with modest investment[36] yielded genome sequences consisting of 4,519,897 scaffolds (Assembly 1 in Figure 1) with an N50 length of 238 Kbp (length cutoff=500 bp). To improve the continuity of fragmentary sequences that were derived from transcribed regions but were separated from exons, the sequences in Assembly 1 were further scaffolded with paired-end RNA-seq reads, which resulted in 4,505,643 sequences (Assembly 2) with an N50 length of 264 Kbp (length cutoff=500 bp). These sequences were filtered for the length of >500 bp, processed again for gap closure with the program PLATANUS, and scanned for contaminants of microbes and artificial oligos used for sequencing. Through this procedure, we have obtained 114,941 sequences with the minimum and maximum lengths of 500 bp and 2.064 Mbp, respectively, marking the N50 scaffolding length of 293 Kbp (Assembly 3).

Gene models

Using the resultant genome sequences (Assembly 3), genome-wide prediction of protein-coding sequences were performed with the program pipeline BRAKER.[35] After preliminary runs with variable parameters and input data sets, we conducted a prediction run with transcript evidence and peptide hints, which resulted in a set of 46,295 genes, with the maximum length of the putative peptides of 19,580 amino acids. These sequences have systematic identifiers Eptbu0000001–Eptbu0046295 with suffixes '.t1'–'.t6' depending on the multiplicity of predicted peptide variants derived from alternative splicing. These sequences are available under https://figshare.com/projects/eburgeri-genome/77052.[37–41]

Mapping RNA-seq reads to the genome assembly

To confirm the coverage of the genome assembly, paired-end RNA-seq reads were aligned to the genome sequence (Assembly 3) with splicing-aware read mapping program HISAT2 as described in the Methods section. This computation resulted in mapping of the reads to the nuclear and the mitochondrial genome sequences of *E. burgeri* at high proportion, at 91.64% and 5.17% respectively.

Gene space completeness assessment of genome assembly and gene models

It has been previously shown that completeness scores of cyclostome genomes tend to be underestimated, when their rapid-evolving nature and phylogenetic position is not taken into consideration.[27] In this study, completeness of the genome assemblies was assessed with CEGMA v2.5[42] (RRID:SCR_015055) and BUSCO v2.0.1[43] (RRID: SCR_015008). For both CEGMA and BUSCO, we employed not only the reference gene sets provided with these pipelines but also the core vertebrate genes (CVG) that was developed specifically for vertebrates from isolated lineages such as elasmobranchs and cyclostomes.[27] The completeness assessments executed using CEGMA and CVG on the gVolante webserver[44,45] returned percentages of single-copy orthologs detected as 'complete' of 65%, and 'complete or partial/fragmented' of 91%. Use of BUSCO v2.0.1[43] with CVG resulted in the detection of 'complete' single-copy orthologs of 83.7%, and 'complete or partial/fragmented' single-copy orthologs of 93.6% (Table 2). The difference of the

Table 2. Statistics of the newly produced gene models compared with published cyclostome gene models.

Species	Source	# Genes (# Peptides)	Maximum peptide length (amino acids)	Completeness score[b] (%)	
				Only 'Complete'	Including 'Fragmented'
Eptatretus burgeri	This study	46295 (50127)	19580	83.7	93.6
Lentheteron camtschaticum	GRAS-LJ[12,a]	34435	19612	90.1	98.7
Petromyzon marinus	PMZ_v3.0[19]	20940 (20950)	18818	57.1	89.3
Petromyzon marinus	Ensembl gene build[13]	10415 (11442)	18900	84.1	94.9
Petromyzon marinus	PMZ1.0[5]	24132 (24271)	17467	63.5	89.3

[a]The construction of this gene model was performed without predicting alternative splice variants, and the number of peptides is thus not included in the relevant cell.
[b]The completeness was scored by the use of the pipeline BUSCO v2 with the one-to-one ortholog set CVG (see Methods).

completeness scores between the assessments of the genome assembly and the gene models might be explained by decreased sensitivity of detecting divergent multi-exon genes in the genome. Altogether, the resultant set of gene models is expected to encompass more than 90% of the protein-coding genes in the *E. burgeri* genome.

Notes for data usage

This data set is oriented towards gene-level analysis including phylogenomic analysis and phylome exploration aiming at studying gene family evolution, rather than the analysis of complete genome structure. Importantly, the total length of the genome sequences obtained in this study amounts only to approximately 1.7 Gbp which is smaller by more than 1 Gbp than the genome size estimate based on flow cytometry of nuclear DNA content[21] (2.91 Gbp). For investigating the structural evolution of the whole genome, such as chromosome elimination or large-scale synteny conservation, it may be advisable to wait for other resources to be released without embargo.

The obtained gene models sometimes include multiple transcripts and their deduced amino acid sequences per gene, because of predicted alternative splice variants. For use in phylogenomics and ortholog clustering, a set of amino acid sequences without splice variants (doi: 10.6084/m9.figshare.11971932)[37] has also been made available. These sequence data are available for BLAST searches on the Squalomix project site (https://transcriptome.riken.jp/squalomix/).

Data availability

Underlying data

Figshare: Underlying data for 'Inference of a genome-wide protein-coding gene set of the inshore hagfish *Eptatretus burgeri*' (https://figshare.com/projects/eburgeri-genome/77052).

This project contains the following underlying data:

- Data file 1: gene coding nucleotide sequences, Eburgeri_v1.gene.fna.gz (https://doi.org/10.6084/m9.figshare.11967795.v2)[37]
- Data file 2: genes' peptide sequences, Eburgeri_v1.gene.faa.gz (https://doi.org/10.6084/m9.figshare.11968119.v2)[38]
- Data file 3: genes' peptide sequences without alternative splicing variants, Eburgeri_v1.gene-noisoform.faa.gz (https://doi.org/10.6084/m9.figshare.11971932.v2)[39]
- Data file 4: Inshore hagfish genome assembly, Eburgeri_v1.genome.fna.gz (https://doi.org/10.6084/m9.figshare.11967789.v3)[40]
- Data file 5: gene model, Eburgeri_v1.gene-model.gff3.gz (https://doi.org/10.6084/m9.figshare.11967474.v2)[41]

Accession numbers

NCBI Protein: Whole genome shotgun sequencing project [*Eptatretus burgeri* (Inshore hagfish)]. Accession number BROF01000000, https://identifiers.org/ncbiprotein:BROF01000000

DDBJ SRA Submission: Sequence data [*Eptatretus burgeri* (Inshore hagfish)]. Accession number DRA010216, https://ddbj.nig.ac.jp/resource/sra-submission/DRA010216

Acknowledgements

We thank Masumi Nozaki for assistance in sampling. The authors acknowledge Kazu Tanimoto, Kaori Tanaka, and Chiharu Tanegashima at Laboratory for Phyloinformatics in RIKEN Center for Biosystems Dynamics Research (BDR) for technical assistance.

References

1. Kuraku S, Ota KG, Kuratani S: *Timetree of life.* Kumar S, Hedges B, editors. 2009.
2. Dehal P, *et al.*: **The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins.** *Science.* 2002; **298**: 2157–2167. PubMed Abstract | Publisher Full Text
3. Putnam NH, *et al.*: **The amphioxus genome and the evolution of the chordate karyotype.** *Nature.* 2008; **453**: 1064–1071. PubMed Abstract | Publisher Full Text
4. Mehta TK, *et al.*: **Evidence for at least six Hox clusters in the Japanese lamprey (Lethenteron japonicum).** *Proc. Natl. Acad. Sci. U. S. A.* 2013; **110**: 16044–16049. PubMed Abstract | Publisher Full Text
5. Smith JJ, *et al.*: **Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution.** *Nat. Genet.* 2013; **45**: 415–421. 421e411-412. PubMed Abstract | Publisher Full Text
6. Nikitina N, Bronner-Fraser M, Sauka-Spengler T: **The sea lamprey *Petromyzon marinus*: a model for evolutionary and developmental biology.** *Cold Spring Harb. Protoc.* 2009; **2009**: pdb emo113. PubMed Abstract | Publisher Full Text
7. Oisi Y, Ota KG, Kuraku S, *et al.*: **Craniofacial development of hagfishes and the evolution of vertebrates.** *Nature.* 2013; **493**: 175–180. PubMed Abstract | Publisher Full Text
8. Ota KG, Kuraku S, Kuratani S: **Hagfish embryology with reference to the evolution of the neural crest.** *Nature.* 2007; **446**: 672–675. PubMed Abstract | Publisher Full Text
9. Pascual-Anaya J, *et al.*: **Hagfish and lamprey Hox genes reveal conservation of temporal colinearity in vertebrates.** *Nat. Ecol. Evol.* 2018; **2**: 859–866. PubMed Abstract | Publisher Full Text
10. Manousaki T, *et al.*: *Jawless Fishes of the World.* Orlov AM, Beamish RJ, editors. Cambridge Scholars Publishing; 2016; Vol. 1: pp. 2–16.
11. Qiu H, Hildebrand F, Kuraku S, *et al.*: **Unresolved orthology and peculiar coding sequence properties of lamprey genes: the KCNA gene family as test case.** *BMC Genomics.* 2011; **12**: 325. Publisher Full Text
12. Kadota M, *et al.*: **CTCF binding landscape in jawless fish with reference to Hox cluster evolution.** *Sci. Rep.* 2017; **7**: 4957. PubMed Abstract | Publisher Full Text
13. Yates AD, *et al.*: **Ensembl 2020.** *Nucleic Acids Res.* 2020; **48**: D682–D688. PubMed Abstract | Publisher Full Text
14. Kuraku S: **Insights into cyclostome phylogenomics: pre-2R or post-2R.** *Zool. Sci.* 2008; **25**: 960–968. PubMed Abstract | Publisher Full Text
15. Sacerdot C, Louis A, Bon C, *et al.*: **Chromosome evolution at the origin of the ancestral vertebrate genome.** *Genome Biol.* 2018; **19**: 166. PubMed Abstract | Publisher Full Text
16. Escriva H, Manzon L, Youson J, *et al.*: **Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution.** *Mol. Biol. Evol.* 2002; **19**: 1440–1450. PubMed Abstract | Publisher Full Text
17. Simakov O, *et al.*: **Deeply conserved synteny resolves early events in vertebrate evolution.** *Nat. Ecol. Evol.* 2020; **4**: 820–830. PubMed Abstract | Publisher Full Text
18. Smith JJ, Keinath MC: **The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications.** *Genome Res.* 2015; **25**: 1081–1090. PubMed Abstract | Publisher Full Text
19. Smith JJ, *et al.*: **The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution.** *Nat. Genet.* 2018; **50**: 270–277. PubMed Abstract | Publisher Full Text
20. Kojima NF, *et al.*: **Whole chromosome elimination and chromosome terminus elimination both contribute to somatic differentiation in Taiwanese hagfish *Paramyxine sheni.*** *Chromosom. Res.* 2010; **18**: 383–400. PubMed Abstract | Publisher Full Text
21. Nakai Y, *et al.*: **Chromosome elimination in three Baltic, south Pacific and north-east Pacific hagfish species.** *Chromosom. Res.* 1995; **3**: 321–330. PubMed Abstract | Publisher Full Text
22. Nakai Y, Kubota S, Kohno S: **Chromatin diminution and chromosome elimination in four Japanese hagfish species.** *Cytogenet. Cell Genet.* 1991; **56**: 196–198. PubMed Abstract | Publisher Full Text
23. Kuraku S, Qiu H, Meyer A: **Horizontal transfers of Tc1 elements between teleost fishes and their vertebrate parasites, lampreys.** *Genome Biol. Evol.* 2012; **4**: 929–936. PubMed Abstract | Publisher Full Text
24. Tanegashima C, *et al.*: **Embryonic transcriptome sequencing of the ocellate spot skate Okamejei kenojei.** *Sci Data.* 2018; **5**: 180200. Publisher Full Text
25. Tatsumi K, Nishimura O, Itomi K, *et al.*: **Optimization and cost-saving in tagmentation-based mate-pair library preparation and sequencing.** *Biotechniques.* 2015; **58**: 253–257. Publisher Full Text
26. Leggett RM, Clavijo BJ, Clissold L, *et al.*: **NextClip: an analysis and read preparation tool for Nextera long mate pair libraries.** *Bioinformatics.* 2014; **30**: 566–568. PubMed Abstract | Publisher Full Text
27. Hara Y, *et al.*: **Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation.** *BMC Genomics.* 2015; **16**: 977. PubMed Abstract | Publisher Full Text
28. Kim D, Paggi JM, Park C, *et al.*: **Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.** *Nat. Biotechnol.* 2019; **37**: 907–915. PubMed Abstract | Publisher Full Text
29. Kajitani R, *et al.*: **Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads.** *Genome Res.* 2014; **24**: 1384–1395. Publisher Full Text
30. Zhu BH, *et al.*: **P_RNA_scaffolder: a fast and accurate genome scaffolder using paired-end RNA-sequencing reads.**

BMC Genomics. 2018; **19**: 175.
PubMed Abstract | Publisher Full Text

31. Hara Y, *et al.*: **Madagascar ground gecko genome analysis characterizes asymmetric fates of duplicated genes.** *BMC Biol.* 2018; **16**: 40.
PubMed Abstract | Publisher Full Text

32. Smit AFA, Hubley R: *RepeatModeler Open-1.0.* (2008-2010).
Reference Source

33. Smit AFA, Hubley R, Green P: *RepeatMasker Open-4.0.* (2013-2015).
Reference Source

34. Altschul SF, *et al.*: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res.* 1997; **25**: 3389–3402.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Hoff KJ, Lomsadze A, Borodovsky M, *et al.*: **Whole-Genome Annotation with BRAKER.** *Methods Mol. Biol.* 2019; **1962**: 65–95.
Publisher Full Text

36. Hara Y, *et al.*: **Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates.** *Nat. Ecol. Evol.* 2018; **2**: 1761–1771.
PubMed Abstract | Publisher Full Text

37. RIKEN Kobe PL: Inshore hagfish gene coding nucleotide sequence. figshare. Dataset. 2022.
Publisher Full Text

38. RIKEN Kobe PL: Inshore hagfish genes' peptide sequences. figshare. Dataset. 2022.
Publisher Full Text

39. RIKEN Kobe PL: Inshore hagfish genes' peptide sequences without alternative splicing variants. figshare. Dataset. 2022.
Publisher Full Text

40. RIKEN Kobe PL: Inshore hagfish genome assembly. figshare. Dataset. 2022.
Publisher Full Text

41. RIKEN Kobe PL: Inshore hagfish gene model. figshare. Dataset. 2022.
Publisher Full Text

42. Parra G, Bradnam K, Ning Z, *et al.*: **Assessing the gene space in draft genomes.** *Nucleic Acids Res.* 2009; **37**: 289–297.
PubMed Abstract | Publisher Full Text

43. Simao FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**: 3210–3212.
PubMed Abstract | Publisher Full Text

44. Nishimura O, Hara Y, Kuraku S: **Evaluating genome assemblies and gene models using gVolante.** *Methods Mol. Biol.* 1962; **1962**: 2019.
PubMed Abstract | Publisher Full Text

45. Nishimura O, Hara Y, Kuraku S: **gVolante for standardizing completeness assessment of genome and transcriptome assemblies.** *Bioinformatics.* 2017; **33**: 3635–3637.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 22 December 2022

https://doi.org/10.5256/f1000research.136943.r155853

?

Daniel Ocampo Daza
CTC Clinical Trial Consultants AB, Uppsala, Sweden

The authors have provided an important resource for the bioinformatic study of phylogenomics and gene family evolution representing an under-studied yet crucial branch of vertebrate evolution, the agnathans. I look forward to start using this resource in my own research. The way the authors have shared all their data openly on the figshare platform is ideal.

In my estimation, the procedures undertaken to isolate, sequence, assemble and evaluate this genome-wide protein coding gene set from the inshore hagfish *Eptatretus burgeri* constitute the current standard practices, even the state of the art. To accompany the genomic sequencing with concurrent transcriptomic sequencing for the identification of transcripts is excellent. Of course, the identification of transcripts from more tissues is always desirable, but the tissues selected here (blood and liver) are appropriate and I doubt that the inclusion of more tissues would have raised the completeness statistics achieved in the study in a major way.

The study is only limited by avoiding a full-scale whole-genome sequencing project with the associated whole-genome analyses, however considering the substantial known challenges for such a project in an agnathan species, some of which the authors mention in this manuscript, it is understandable and acceptable that the publication of a coding gene set is a worthwhile goal in and of itself. Like I mentioned above, it will be a valuable resource.

I have only the following minor comments pertaining to the text of the manuscript itself or some general points on vertebrate evolution. I have used the pagination of the provided PDF as a guide to where in the manuscript I direct my comments.

The authors have used "testis", singular, consistently throughout the manuscript rather than the plural "testes". Perhaps it is worth shortly mentioning that hagfishes only have one testis? This could be done simply on page 3 by writing "the *single* testis was sampled..." The word "sampled" also suggests that only part of the testis was used. Was the whole organ used for DNA extraction?

In the abstract the authors write that hagfishes "arose from agnathan (...) lineages". This is a

confusing statement. Hagfishes in themselves are an *extant* agnathan lineage that arose from an ancestral, *extinct* agnathan lineage. Perhaps this statement can be rewritten more clearly.

The word choice "*irreplaceable*" in the first paragraph on page 3 is strange and I don't think it applies. The authors have similarly used *"indispensable"* further down in the text, at the end of the 3rd paragraph on page 3. I suggest using *"a scientifically valuable phylogenetic position among extant vertebrates"* (remove *"the"* before *"extant"*) in the first instance and perhaps *"an important and under-studied component of vertebrate diversity"* (remove *"the"* before *"vertebrate diversity"*).

It is an overstatement to write that agnathans diverged from the vertebrate stem during the early Cambrian. What is this based on? The best fossil evidence indicates that agnathans were present in the late Silurian, at the earliest, and were abundant in the Ordovician. *Haikouichthys*, an early Cambrian vertebrate is sometimes called an agnathan, but it is not at all clear that it belongs to the stem agnathan lineage that gave rise to lampreys and hagfishes. It is likely more closely related to the stem craniate lineage. Is the overstatement based on molecular time-estimations? These often overshoot time estimations in the far past, whereby it is important to also consider the fossil evidence.

In paragraph 2 on page 3, I suggest writing "the peculiar nature of *their* protein-coding sequences".

Regarding the first sentence on paragraph 3, page 3: I don't think this conveys the situation accurately. I suggest something like this instead: "As of July 2022, there is only one whole genome resource for hagfishes, a genome assembly from *Eptatretus burgeri* with corresponding gene models available in Ensembl. A description and analysis of this whole genome sequence has not yet been published." This "unpublished" genome has also been a valuable resource, especially because it also allows for the study of conserved synteny, so it should not be downplayed. I also don't entirely agree that because a description and analysis has not yet been published, this "hinders the comprehensive characterization of gene repertoires and their expression patterns." There are predicted gene models/annotations available in the Ensembl database. The gene set presented in this manuscript obviously provide better evidence for protein-coding gene sequence, but the Ensembl genome assembly and gene models are not entirely useless. The authors should present the advantages of their approach and the resulting gene set without understating the usefulness of the previously available genome assembly.

In line 3 of paragraph 3 on page 3 I suggest: *"Currently, some efforts to sequence and analyse hagfish genomes are ongoing..."* The original phrasing does not mention hagfishes specifically.

How was the species of the sampled hagfish determined? The manuscript only describes that it was caught at the Misaki Marine Station in 2013. Do other hagfish species inhabit the location where the sampled hagfish was caught? The brown hagfish *Eptatretus atami*, for example, also occurs in the sea around Japan. Was the location or habitat where it was caught used to determine the species, or were physical/anatomical characteristics used? Or indeed both?

In line 3 of the 4th paragraph on page 3, i suggest "*and genome sequencing was performed...*" removing *"the"* before *"genome sequencing".*

In the last line of the 1st paragraph on page 6, *"as performed previously"* seems to suggest

something previously described in the present manuscript, not a previously published study. Please make this clearer.

The first sentence on paragraph 4 of page 6 (*"Our technical procedure..."*) is very long and difficult to follow. Please break this up and clarify. I the same sentence, what does *"with modest investment"* mean?

In the next to last line of paragraph 4 on page 6 I suggest *"we obtained"* rather than *"we have obtained"*.

In the second line of paragraph 7 on page 6 I suggest *"In this study, **the** completeness..."*

In line 5 of the same paragraph I suggest *"that **were** developed specifically for vertebrates..."*

In line 8 of the same paragraph I suggest *"**The** use of BUSCO..."*

In line 3 of the 2nd paragraph on page 7 I suggest *"which is more than 1 Gbp smaller"* rather than *"which is smaller by more than 1 Gbp"*.

Are the rationale for sequencing the genome and the species significance clearly described?
Yes

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

Are sufficient details of the sequencing and extraction, software used, and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a usable and accessible format, and the assembly and annotation available in an appropriate subject-specific repository?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Vertebrate evolution, comparative genomics, gene family evolution

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 30 November 2022

https://doi.org/10.5256/f1000research.136943.r155282

?

Vincent Laudet
Marine Eco-Evo-Devo Unit, Okinawa Institute of Science and Technology, Okinawa, Japan

In this manuscript the authors present a genome of a hagfish (*Eptatretus burgeri*) determined from testis DNA. This is completed by transcriptome data from two tissues. The importance of hagfish for vertebrate evolution is obvious and this sequence is of course a very useful resource for the community. However I found the paper a bit disappointing for three main reasons:

1. In lamprey there is a fascinating phenomenon called "programmed genome rearrangement" in which we see the loss of ca 10-15% of the DNA in somatic cells during early development, producing a somatic genome that is different from the germline genome. This process is known to exist in hagfish. However, the author does not mention this in the introduction whereas of course this explains why they choose to sequence testis DNA. This should be added of course. No analysis of this phenomenon is done. For example, one could wonder what is the extent of this genome completeness when compared to other cyclostome genomes available.

2. I wonder why the authors have only determind transcriptomes of two tissues, the liver and the blood. I think brain and testis (especially because of the programmed genome rearrangement present in lamprey) would have been also particularly interesting.

3. I found the Busco value for single copy gene (83%) quite low, indicating that many genes are only partial. But even the value for partial genes (93.6%) is relatively low. Given the importance of hagfish in vertebrate genome evolution I wonder why the authors have not put more efforts in reaching better levels. This is particularly vexing given the programmed genome rearrangement phenomenon discussed above.

Minor point

Material and Methods. The sentence "*The study was conducted with all efforts to ameliorate any suffering of animals, in accordance with the institutional guideline Regulations for the Animal Experiments by the Institutional Animal Care and Use Committee (IACUC) of the RIKEN Kobe Branch.*" is bizarre. I think it would be better to replace "*ameliorate*" by "*avoid*".

Are the rationale for sequencing the genome and the species significance clearly described?
No

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

Are sufficient details of the sequencing and extraction, software used, and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a usable and accessible format, and the assembly and

annotation available in an appropriate subject-specific repository?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Eco-Evo-Devo

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.