Inference of a genome-wide protein-coding gene set of the inshore hagfish *Eptatretus burgeri*

Kazuaki Yamaguchi[1,†], Yuichiro Hara[1,†,#], Kaori Tatsumi[1,†], Osamu Nishimura[1,†], Jeramiah J. Smith[2], Mitsutaka Kadota[1], Shigehiro Kuraku[1*]

Affiliations

1. Laboratory for Phyloinformatics, RIKEN Center for Biosystems Dynamics Research, Kobe, Japan
2. Department of Biology, University of Kentucky, Lexington, KY, USA

[#]Present address: Research Center for Genome & Medical Sciences, Tokyo Metropolitan Institute of Medical Science (TMiMS), Tokyo, Japan

[†]These authors contributed equally

[*]Corresponding author: Shigehiro Kuraku (shigehiro.kuraku@riken.jp)

Abstract

The group of hagfishes (Myxiniformes) arose from agnathan (jawless vertebrate) lineages and is one of the 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 at reconstructing a high-coverage 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 resulting gene models promises a high utility in various comparative analyses including phylogenomics and phylome exploration.

Background & Summary

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* 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 studies 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 June 2020, 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

Genomic DNA was extracted from the testis of a 48cm-long male individual of *Eptatretus burgeri* caught at the Misaki Marine Station in June 2013, with phenol/chloroform as previously described[23], and the genome sequencing was performed as outlined in Figure 1. The study was conducted 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 with an S220 Focused-ultrasonicator (Covaris) to retrieve DNA fragments of variable length distributions (see Table 1 for detailed amounts of starting DNA and conditions for shearing). The sheared DNA was used for paired-end library preparation with a KAPA LTP Library Preparation Kit (KAPA Biosystems). The optimal numbers of PCR cycles for individual libraries were determined with a Real-Time Library Amplification Kit (KAPA Biosystems) by preliminary qPCR-based quantification using an aliquot of adaptor-ligated DNAs as described previously[24]. Small molecules in the prepared libraries were removed by size selection using Agencourt AMPure XP (Beckman Coulter). The numbers of PCR cycles and conditions of size selection for individual libraries are included in Table 1. Mate-pair libraries were prepared using a Nextera Mate Pair Sample Prep Kit (Illumina), employing our customized iMate protocol[25] (http://www.clst.riken.jp/phylo/imate.html). The detailed conditions of mate-pair library preparation are included in Table 1. After size selection, the quantification of the prepared libraries was performed using a KAPA Library Quantification Kit (KAPA Biosystems). They were sequenced on a HiSeq 1500 (Illumina) operated by HiSeq Control Software v2.0.12.0 using a HiSeq SR Rapid Cluster Kit v2 (Illumina) and HiSeq Rapid SBS Kit v2 (Illumina), HiSeq X (Illumina) operated by HiSeq Control Software v3.3.76, and MiSeq operated by MiSeq Control Software v2.3.0.3 using 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 was performed with RTA v1.17.21.3, and the fastq files were generated by bcl2fastq v1.8.4 (Illumina) for HiSeq 1500 and MiSeq, while RTA v2.7.6 and bcl2fastq v2.15.0 (Illumina) were instead employed for HiSeq X. Removal of low-quality bases from paired-end reads was processed by TrimGalore v0.3.3 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with the options '--stringency 2 --quality 30 --length 25 --paired --retain_unpaired'. Mate-pair reads were processed by NextClip v1.1[26] with the default parameters.

RNA-seq and transcriptome data processing

Total RNAs were extracted from the liver and the blood of the above-mentioned adult individual with Trizol reagent (Thermo Fisher Scientific). Quality control of DNase I-treated RNA 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) as previously described[27]. The amount of starting total RNA and numbers of PCR cycles are included in Table 1. The obtained sequence reads were trimmed for removal of adaptor sequences and low-quality bases with TrimGalore v0.3.3 as outlined above. Alignment of the trimmed RNA-seq reads to the genome assembly employed HISAT2 v2.2.1[28] with the options of ‘-k 3 -p 20 --pen-noncansplice 1000000’.

Genome assembly

De novo genome assembly and scaffolding that employs the processed short reads were carried out by the program PLATANUS v1.2.4[29] with its default parameters. The assembly step employed paired-end reads and single reads whose pairs had been removed, and the scaffolding step employed paired-end and mate-pair reads. The gap closure step 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. Resultant genomic scaffold sequences were screened for own mitochondrial DNA fragments, contaminating organismal sequences, PhiX sequences loaded as a control, mitochondrial DNA sequences, and those shorter than 500▯bp, as performed previously[31].

Repeat detection and masking

To obtain species-specific repeat libraries, RepeatModeler v1.0.8[32] was run on the genome assemblies of the individual species with default parameters. Detection of repeat elements in the genomes was performed by RepeatMasker v4.0.5[33], which employs the National Center for Biotechnology Information (NCBI) RMBlast v2.2.27[34], using the custom repeat library obtained above. For gene prediction, the parts of genome sequences detected as repeats are soft-masked by RepeatMasker with the options ‘-nolow -xsmall’.

Construction of gene models

Construction of gene models was performed by employing the gene prediction pipeline Braker v2.1.4[35] with the options '--min_contig=500 --prg=gth --softmasking --UTR=off' (Figure 1). This computation employed RNA-seq read alignments in a .bam file onto the genome assembly and a set of peptide sequences prepared as following. 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.

Data Records

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 500bp and 2.064 Mbp, respectively, marking the N50 scaffolding length of 293 Kbp (Assembly 3). These sequences have systematic identifiers scf_eptbu00000001□scf_eptbu00114941 and are available under https://figshare.com/projects/eburgeri-genome/77052.

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.

Technical Validation

Mapping RNA-seq reads to the genome assembly

To confirm the coverage of the genome assembly, paired-end RNA-seq reads were aligned with splicing-aware read mapping program HISAT2 as described in Methods. This computation resulted in the high percentage of the paired reads mapped to the nuclear genome sequences of 91.64 % while the majority of the remaining reads (5.17%) were mapped to the mitochondrial genome sequence of *E. burgeri* itself.

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[37] and BUSCO v2.0.1[38]. 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 assessments were executed on the gVolante webserver[39,40]. The percentages of single-copy orthologs detected as 'complete' was approximately 65 %, with 91 % detected as either 'complete' or 'partial/fragmented', when CEGMA was used with CVG.

Similarly, we performed completeness of the gene models with BUSCO v2.0.1[38], again using CVG as a reference ortholog set. The assessment resulted in the complete coverage of 83.7 % and partial coverage of 93.6 % of CVGs (Table 2). The difference of the 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.

Data Citations

The sequence data are available in fastq files at DDBJ DRA under the accession ID DRA010216 and as multifasta files at https://figshare.com/projects/eburgeri-genome/77052, as well as formatted databases for BLAST searches at https://transcriptome.riken.jp/squalomix/.

Usage Notes

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 sequence files of the obtained gene models sometimes include multiple transcripts and its deduced amino acid sequences per gene, because of predicted alternative splice variants. To facilitate the use of the dataset without splice variants, a sequence file without splice variants (doi: 10.6084/m9.figshare.11971932) has also been made available.

Code Availability

No custom computer code was employed in this study.

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. This study was supported by RIKEN and JSPS KAKENHI Grant Number 17K07426 to S.K. and an NSF Grant Number MCB-1818012 to J.J.S..

Author contributions

S.K conceived the study. J.J.S. sampled the tissues. Y.H., K.Y., O.N., M.K, and S.K performed experiments. Y.H., K.Y., J.J.S. and S.K interpreted data. S.K drafted the manuscript. All authors contributed to the final manuscript editing.

Competing interests

The authors declare no conflict of interest.

Figures

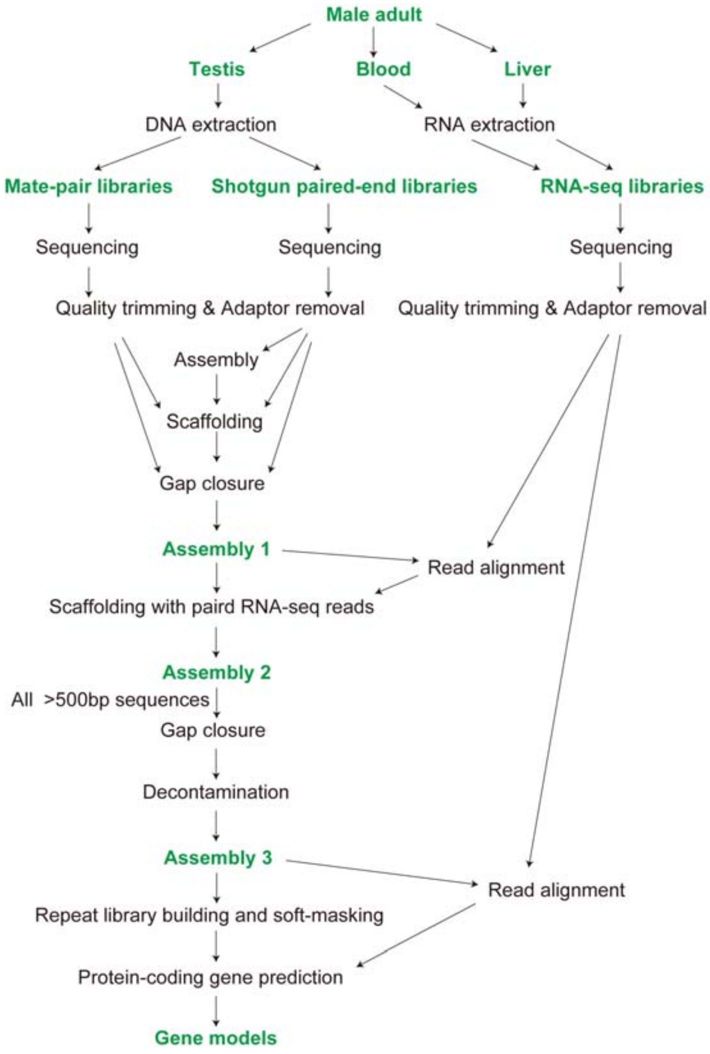

Figure 1. Data production workflow. Samples, raw data, and products are indicated with green letters, while computational steps are labelled in black. See Methods for the details including the choice of the programs used in individual computational steps.

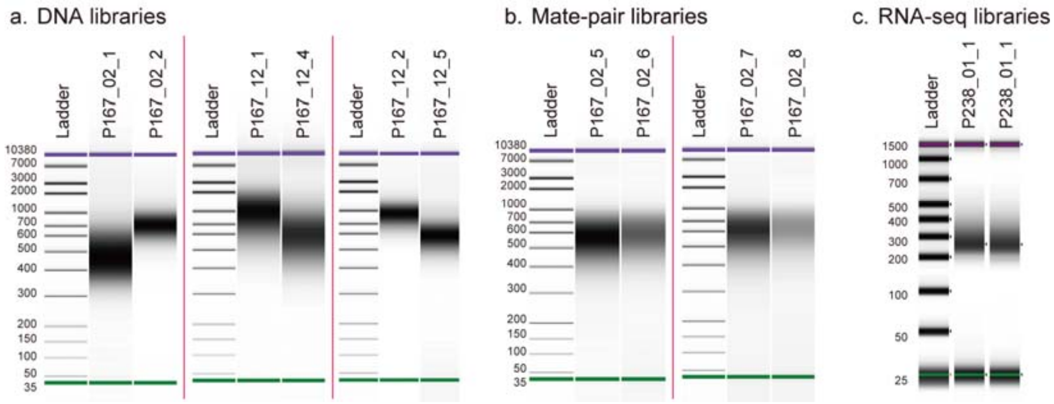

Figure 2. Length distributions of DNA molecules in sequencing libraries. a, Shotgun DNA libraries. b, Mate-pair libraries. c, RNA-seq libraries.

Tables

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

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).

References

1 Kuraku, S., Ota, K. G. & Kuratani, S. in *Timetree of life* (eds S. Kumar & B Hedges) (2009).

2 Dehal, P. *et al.* The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science* **298**, 2157-2167, doi:10.1126/science.1080049 (2002).

3 Putnam, N. H. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064-1071, doi:10.1038/nature06967 (2008).

4 Mehta, T. K. *et al.* Evidence for at least six Hox clusters in the Japanese lamprey (Lethenteron japonicum). *Proc Natl Acad Sci U S A* **110**, 16044-16049, doi:10.1073/pnas.1315760110 (2013).

5 Smith, J. J. *et al.* Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. *Nat Genet* **45**, 415-421, 421e411-412, doi:10.1038/ng.2568 (2013).

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**, pdb emo113, doi:10.1101/pdb.emo113 (2009).

7 Oisi, Y., Ota, K. G., Kuraku, S., Fujimoto, S. & Kuratani, S. Craniofacial development of hagfishes and the evolution of vertebrates. *Nature* **493**, 175-180, doi:10.1038/nature11794 (2013).

8 Ota, K. G., Kuraku, S. & Kuratani, S. Hagfish embryology with reference to the evolution of the neural crest. *Nature* **446**, 672-675, doi:10.1038/nature05633 (2007).

9 Pascual-Anaya, J. *et al.* Hagfish and lamprey Hox genes reveal conservation of temporal colinearity in vertebrates. *Nat Ecol Evol* **2**, 859-866, doi:10.1038/s41559-018-0526-2 (2018).

10 Manousaki, T. *et al.* in *Jawless Fishes of the World* Vol. 1 (ed A. & Beamish Orlov, R.) 2-16 (Cambridge Scholars Publishing, 2016).

11 Qiu, H., Hildebrand, F., Kuraku, S. & Meyer, A. Unresolved orthology and peculiar coding sequence properties of lamprey genes: the KCNA gene family as test case. *BMC Genomics* **12**, 325, doi:10.1186/1471-2164-12-325 (2011).

12 Kadota, M. *et al.* CTCF binding landscape in jawless fish with reference to Hox cluster evolution. *Sci Rep* **7**, 4957, doi:10.1038/s41598-017-04506-x (2017).

13 Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res* **48**, D682-D688, doi:10.1093/nar/gkz966 (2020).

14 Kuraku, S. Insights into cyclostome phylogenomics: pre-2R or post-2R. *Zoolog Sci* **25**, 960-968, doi:10.2108/zsj.25.960 (2008).

15 Sacerdot, C., Louis, A., Bon, C., Berthelot, C. & Roest Crollius, H. Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biol* **19**, 166, doi:10.1186/s13059-018-1559-1 (2018).

16 Escriva, H., Manzon, L., Youson, J. & Laudet, V. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. *Mol Biol Evol* **19**, 1440-1450, doi:10.1093/oxfordjournals.molbev.a004207 (2002).

17 Simakov, O. *et al.* Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol*, doi:10.1038/s41559-020-1156-z (2020).

18 Smith, J. J. & Keinath, M. C. The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. *Genome Res* **25**, 1081-1090, doi:10.1101/gr.184135.114 (2015).

19 Smith, J. J. *et al.* The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nat Genet* **50**, 270-277, doi:10.1038/s41588-017-0036-1 (2018).

20 Kojima, N. F. *et al.* Whole chromosome elimination and chromosome terminus elimination both contribute to somatic differentiation in Taiwanese hagfish Paramyxine sheni. *Chromosome Res* **18**, 383-400, doi:10.1007/s10577-010-9122-2 (2010).

21 Nakai, Y. *et al.* Chromosome elimination in three Baltic, south Pacific and north-east Pacific hagfish species. *Chromosome Res* **3**, 321-330, doi:10.1007/bf00713071 (1995).

22 Nakai, Y., Kubota, S. & Kohno, S. Chromatin diminution and chromosome elimination in four Japanese hagfish species. *Cytogenet Cell Genet* **56**, 196-198, doi:10.1159/000133087 (1991).

23 Kuraku, S., Qiu, H. & Meyer, A. Horizontal transfers of Tc1 elements between teleost fishes and their vertebrate parasites, lampreys. *Genome Biol Evol* **4**, 929-936, doi:10.1093/gbe/evs069 (2012).

24 Tanegashima, C. *et al.* Embryonic transcriptome sequencing of the ocellate spot skate Okamejei kenojei. *Sci Data* **5**, 180200, doi:10.1038/sdata.2018.200 (2018).

25 Tatsumi, K., Nishimura, O., Itomi, K., Tanegashima, C. & Kuraku, S. Optimization and cost-saving in tagmentation-based mate-pair library preparation and sequencing. *Biotechniques* **58**, 253-257, doi:10.2144/000114288 (2015).

26 Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D. & Caccamo, M. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics* **30**, 566-568, doi:10.1093/bioinformatics/btt702 (2014).

27 Hara, Y. *et al.* Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation. *BMC Genomics* **16**, 977, doi:10.1186/s12864-015-2007-1 (2015).

28 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915, doi:10.1038/s41587-019-0201-4 (2019).

29 Kajitani, R. *et al.* Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res* **24**, 1384-1395, doi:10.1101/gr.170720.113 (2014).

30 Zhu, B. H. *et al.* P_RNA_scaffolder: a fast and accurate genome scaffolder using paired-end RNA-sequencing reads. *BMC Genomics* **19**, 175, doi:10.1186/s12864-018-4567-3 (2018).

31 Hara, Y. *et al.* Madagascar ground gecko genome analysis characterizes asymmetric fates of duplicated genes. *BMC Biol* **16**, 40, doi:10.1186/s12915-018-0509-4 (2018).

32 Smit, A. F. A. & Hubley, R. *RepeatModeler Open-1.0.*, <http://www.repeatmasker.org> (2008-2010).

33 Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0.*, <http://www.repeatmasker.org> (2013-2015).

34 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

35 Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-Genome Annotation with BRAKER. *Methods Mol Biol* **1962**, 65-95, doi:10.1007/978-1-4939-9173-0_5 (2019).

36 Hara, Y. *et al.* Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates. *Nat Ecol Evol* **2**, 1761-1771, doi:10.1038/s41559-018-0673-5 (2018).

37 Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. *Nucleic Acids Res* **37**, 289-297, doi:10.1093/nar/gkn916 (2009).

38 Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212, doi:10.1093/bioinformatics/btv351 (2015).

39 Nishimura, O., Hara, Y. & Kuraku, S. Evaluating Genome Assemblies and Gene Models Using gVolante. *Methods Mol Biol* **1962**, 247-256, doi:10.1007/978-1-4939-9173-0_15 (2019).

40 Nishimura, O., Hara, Y. & Kuraku, S. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **33**, 3635-3637, doi:10.1093/bioinformatics/btx445 (2017).