RESEARCH ARTICLE

First draft genome assembly and identification of SNPs from hilsa shad (*Tenualosa ilisha*) of the Bay of Bengal [version 1; peer review: 1 approved, 2 approved with reservations]

Md. Bazlur Rahman Mollah [1], Mohd Golam Quader Khan[2], Md Shahidul Islam[3], Md Samsul Alam [2]

[1]Poultry Biotechnology and Genomics Laboratory, Department of Poultry Science, Bangladesh Agricultural University, Mymensingh, 2202, Bangladesh
[2]Department of Fisheries Biology and Genetics, Bangladesh Agricultural University, Mymensingh, 2202, Bangladesh
[3]Department of Biotechnology, Bangladesh Agricultural University, Mymensingh, 2202, Bangladesh

v1 **First published:** 22 Mar 2019, **8**:320 (https://doi.org/10.12688/f1000research.18325.1)
Latest published: 22 Mar 2019, **8**:320 (https://doi.org/10.12688/f1000research.18325.1)

Abstract

Background: Hilsa shad (*Tenualosa ilisha*), a widely distributed migratory fish, contributes substantially to the economy of Bangladesh. The harvest of hilsa from inland waters has been fluctuating due to anthropological and climate change-induced degradation of the riverine habitats. The whole genome sequence of this valuable fish could provide genomic tools for sustainable harvest, conservation and productivity cycle maintenance. Here, we report the first draft genome of *T. ilisha* from the Bay of Bengal, the largest reservoir of the migratory fish.
Methods: A live specimen of *T. ilisha* was collected from the Bay of Bengal. The whole genome sequencing was performed by the Illumina HiSeqX platform (2 × 150 paired end configuration). We assembled the short reads using SOAPdenovo2 genome assembler and predicted protein coding genes by AUGUSTUS. The completeness of the *T. ilisha* genome assembly was evaluated by BUSCO (Benchmarking Universal Single Copy Orthologs). We identified single nucleotide polymorphisms (SNPs) by calling them directly from unassembled sequence reads using discoSnp++.
Results: We assembled the draft genome of 710.28 Mb having an N50 scaffold length of 64157 bp and GC content of 42.95%. A total of 37,450 protein coding genes were predicted of which 29,339 (78.34%) were annotated with other vertebrate genomes. We also identified 792,939 isolated SNPs with transversion:transition ratio of 1:1.8. The BUSCO evaluation showed 78.1% completeness of this genome.
Conclusions: The genomic data generated in this study could be used as a reference to identify genes associated with physiological and ecological adaptations, population connectivity, and migration behaviour of this biologically and economically important anadromous fish species of the Clupeidae family.

Keywords

Hilsa, anadromous, Bay of Bengal, whole genome, SNP

Open Peer Review

Reviewer Status ✔ ? ?

	Invited Reviewers		
	1	**2**	**3**
version 1 published 22 Mar 2019	✔ report	? report	? report

1 **Joel A. Malek**, Weill Cornell Medicine–Qatar, Doha, Qatar

2 **Yuzine B. Esa**, University Putra Malaysia (UPM), Serdang, Malaysia

3 **Shotaro Hirase**, University of Tokyo, Hamamatsu, Japan

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

This article is included in the Draft Genomes collection.

Corresponding authors: Md. Bazlur Rahman Mollah (mbrmollah.ps@bau.edu.bd), Md Samsul Alam (samsul.alam@bau.edu.bd)

Author roles: Mollah MBR: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Khan MGQ**: Conceptualization, Funding Acquisition, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Islam MS**: Conceptualization, Funding Acquisition, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Alam MS**: Conceptualization, Funding Acquisition, Investigation, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Mollah MBR, Khan MGQ, Islam MS and Alam MS. **First draft genome assembly and identification of SNPs from hilsa shad (*Tenualosa ilisha*) of the Bay of Bengal [version 1; peer review: 1 approved, 2 approved with reservations]** F1000Research 2019, **8**:320 (https://doi.org/10.12688/f1000research.18325.1)

First published: 22 Mar 2019, **8**:320 (https://doi.org/10.12688/f1000research.18325.1)

Introduction

Hilsa shad (*Tenualosa ilisha*) is a migratory fish of the Clupeidae family. It is distributed from the South China Sea and through the Bay of Bengal to the Persian Gulf. The riverine habitats of this fish include the Padma, Jamuna, Meghna, Karnaphuli and the coastal rivers of Bangladesh, the Tigris and Euphrates of Iran and Iraq, the Indus of Pakistan, the rivers of Eastern and Western India and the Irrawaddy of Myanmar (Freyhof, 2014; Pillay & Rosa, 1963). Ecologically, three different types of hilsa shad have been recognized in Bangladesh waters such as anadromous, potamodromous and marine (Milton, 2010).

Hilsa is the most popular and economically important food fish in Bangladesh, contributing 12% of the total fish production and 1.15% of GDP. Of its world catch, 60% amounting to 0.5 million metric tons, comes from Bangladesh (DoF, 2018). Though the overall production of hilsa increased over the years, gross decline in productions were evident from inland waters. A number of factors such as overexploitation, siltation in river beds, decrease in water flow from upstream and fragmentation of the rivers are attributed to this fluctuation in productivity (Ahsan *et al.,* 2014). To enhance hilsa production, programs like the establishment of sanctuaries, restrictions on the use of fishing equipment and a ban on fishing in certain periods of the year to protect parent and juvenile fish have been initiated. It is, however, very important that the management activities be matched with the biological features of the fish for their effectiveness. Inconclusive information about the management units and the level of connectivity amongst them is considered as the major constraint in formulating appropriate hilsa management plans.

There are controversies regarding the number of hilsa stocks in Bangladesh waters. Studies involving morphological and genetic analyses using allozyme, Random Amplification of Polymorphic DNA (RAPD) and mtDNA-restriction fragment length polymorphism (RFLP) markers proved to be insufficient in resolving the stock disputes of this species (Ahmed *et al.,* 2004; Mazumder & Alam, 2009; Salini *et al.,* 2004). DNA markers derived from whole genome sequencing are more efficient to define management units, quantify the extent of adaptive divergence and connectivity among stocks, and to perform mixed-stock analysis (Martinez Barrio *et al.,* 2016; Figueras *et al.,* 2016; Machado *et al.,* 2018). Single nucleotide polymorphic (SNP) markers allow whole genome coverage and high levels of automation. Conventionally, SNP markers are developed by comparing nucleotide sequences with a reference genome. Recent advancement in generating SNPs from reference-free whole genome sequences accelerated identification of SNPs from non-model organisms. Although hilsa is a very important fish biologically and economically, it lacks a reference genome and genomic resources, imposing a severe bottleneck to understand its physiological and ecological requirements. Therefore, we performed whole genome sequencing (Mollah *et al.,* 2018), constructed a draft genome assembly and identified SNPs from *T. ilisha* of the Bay of Bengal.

Methods

Sample collection and genomic DNA extraction

We captured ten *T. ilisha* specimens from the seashore of the Bay of Bengal (21.981753 N 90.305556 E) (Figure 1). All efforts were made to ameliorate harm to the fish by using a seine net of appropriate mesh size (20 mm) to avoid any physical injuries and suffocation. Due to the nature of this species, the fish died immediately after taking them out of the water. Dorsal and caudal fin tissues were immediately clipped on board from a dead female fish (560.42g) and preserved in 96% ethanol. The fish were handled according to the guidelines of the Animal Welfare and Ethical Committee (AWEC) of Bangladesh Agricultural University. The genomic DNA was isolated using the standard phenol:chlorofom:isoamyl alcohol method (Sambrook & Russell, 2001). DNA purity was evaluated using a NanoDrop 2000 Spectrophotometer (ThermoFisher Scientific, cat # ND-2000) and 0.8% agarose gel electrophoresis. DNA was quantified using Qubit 2.0 Fluorometer and Qubit dsDNA HS Assay Kit (ThermoFisher Scientific, Cat. # Q32851) and used to systematically generate the whole genome sequence data (Figure 2).

PCR-free DNA library preparation and sequencing

For sequencing, a PCR-free DNA library was prepared using the Illumina TruSeq DNA PCR-free Library Preparation Kit (Cat. # 20015963), following the manufacturer's recommendations (Illumina, CA, USA). The library was fragmented, size was selected following the 350 bp insert size scheme and validated using TapeStation (2100 Bioanalyzer,Agilent Technologies, CA, USA). The DNA library was quantified using a Qubit 2.0 Fluorometer as well as by real time PCR (ABI 7500, Applied Biosystems, CA, USA) using the KAPA Library Quantification Kit (Cat. # KK4824) following the manufacturer's standard protocol with the primer pair Primer 1: 5'-AAT GAT ACG GCG ACC ACC GA-3' Primer 2: 5'-CAA GCA GAA GAC GGC ATA CGA-3'. The PCR condition was followed as initial denaturation at 95°C for 5 min followed by 35 cycles (denaturation at 95°C for 30 sec, annealing/extension/data acquisition at 60°C for 45 sec) and melt curve analysis at 65 – 95°C. Sequencing was performed on the Illumina HiSeqX instrument according to the manufacturer's instructions. The library was sequenced using a 2× 150 paired-end (PE) configuration (GENEWIZ, LLC. South Plainfield, NJ, USA).

DNA sequence processing and genome size estimation

The raw reads were filtered based on quality and length using Trimmomatic-0.32 (Bolger *et al.,* 2014) after evaluating with FastQC v. 0.11.8 (Andrews, 2018) as follows: i) removal of adaptor sequences; ii) removal of read pairs from either ends when the base quality was <20; iii) trimming low quality fragments at both ends of the reads within a window size of 4 bp and an average quality threshold of 15; iv) removal of read pairs having <75 nucleotides. Jellyfish v. 2.2.6 (Marçais & Kingsford, 2011) was used to obtain a separate frequency distribution of three different high occurring kmers (21, 31 and 33) in the raw HiSeq sequence reads, and the histograms were uploaded to GenomeScope for estimating genome size, repeat content,

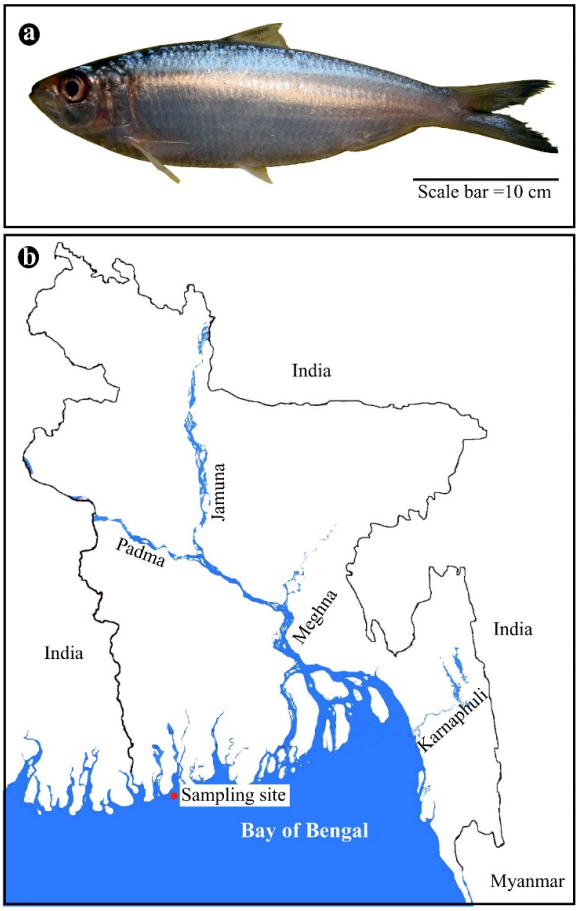

Figure 1. The experimental fish and its collection site. Photograph of a *T. ilisha* specimen (**a**) and a map of Bangladesh showing the sampling site (21.981753 N 90.305556 E) (**b**).

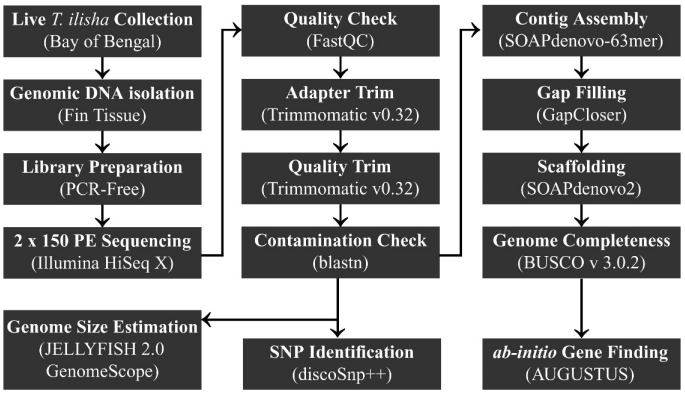

Figure 2. Methodology outline. Schematic diagram illustrating the methodology of whole genome sequencing, *de novo* assembly and identification of SNPs in *T. ilisha* from the Bay of Bengal.

repeat length, unique length and heterozygosity following kmer-based statistical approaches (Vurture *et al.*, 2017).

Genome assembly, genome quality evaluation and annotation

We assembled the short reads using SOAPdenovo2 genome assembler (Luo *et al.*, 2012), developed specifically for use with next-generation short-read sequences. SOAPdenovo2 uses the de Bruijn graph algorithm. We tested several kmers to assemble the *T. ilisha* genome and finally selected the assembly with a kmer of 33. The completeness of the *T. ilisha* genome assembly was evaluated by BUSCO (Benchmarking Universal Single Copy Orthologs) (Simão *et al.*, 2015). For BUSCO analysis (-m geno

–sp zebrafish settings), the genome was searched against the Actinopterygii database (actinopterygii_odb9), which was constructed from 20 fish species consisting of 4584 orthologs. AUGUSTUS *ab initio* gene prediction was performed to predict protein-coding genes. The protein sequences of fish species and other vertebrates, including *Rhincodon typus*, *Cyprinus carpio*, *Takifugu rubripes*, *Salmo salar*, *Mus musculus* and *Homo sapiens*, were downloaded from the NCBI non-redundant protein sequences (nr) database (Table 3) and aligned against the *T. ilisha* genome using BLASTP (Altschul *et al.,* 1997).

Reference-free detection of isolated SNPs

We used discoSnp++ v2.2.10 (Uricaru *et al.,* 2015) with default parameters for reference-free detection of isolated SNPs (SNPs not flanked by other SNPs, Indels or structural variants) by calling SNPs directly from sequence reads without a reference genome. This method identifies isolated SNPs from the sequences of two homologous chromosomes within a single individual.

Results and discussion

The estimated haploid genome size of *T. ilisha* ranged from 649.48 to 660.73 Mb. We observed heterozygosity and a repeat peak (Figure 3), with an estimated heterozygosity of 0.579 to 0.660% and repeats of 8.30 to 13.57% (Table 1). We assembled the draft genome of 710.28 Mb, having an N50 scaffold length of 64157 bp and a GC content of ~43% (Table 2). The whole genome assembly of a notable Clupeid fish, the Atlantic herring, based on short reads (170 bp to 20 kb inserts) was 808 Mb with a scaffold N50 of 1.84 Mb and GC content of 44%, with repetitive elements making up 31% of the assembly (Martinez Barrio *et al.,* 2016). The genome size of another important Clupeid fish, the European sardine (*Sardina pilchardus*), was estimated to be 655-850 Mb (Machado *et al.,* 2018) and 935-950Mb (Louro *et al.,* 2018).

The assembled *T. ilisha* genome was searched for BUSCO analysis against the Actinopterygii database, consisting of 4,584 orthologs constructed from 20 fish species. We found

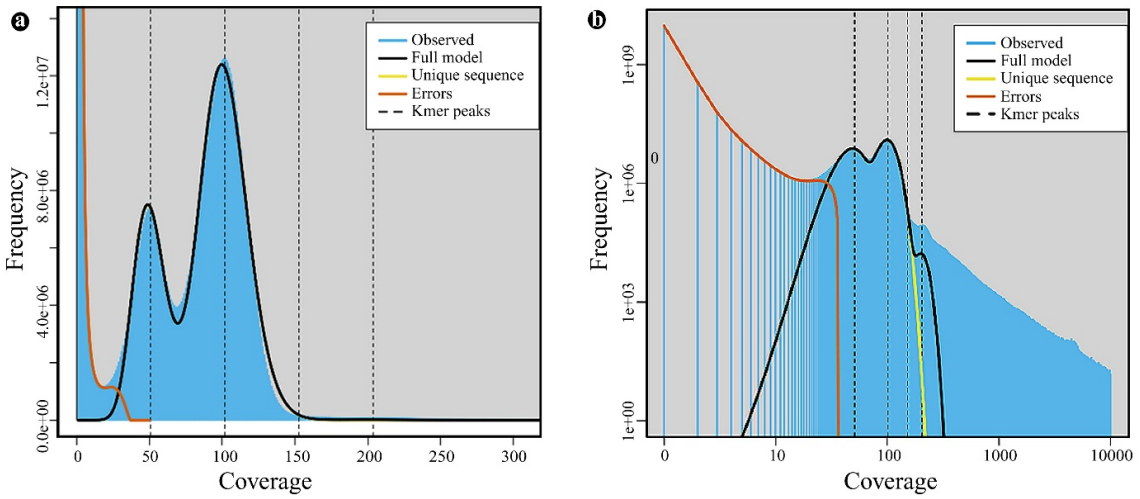

Figure 3. GenomeScope kmer profile plot of the *T. ilisha*. Dataset show the fit of the GenomeScope model (black) based on 33-kmers in Illumina HiSeq sequence reads, max kmer coverage at 300× (**a**) and 10000× coverage (**b**).

Table 1. Properties of *T. ilisha* genome estimated at three different kmers[1].

Properties	Kmer 21		Kmer 31		Kmer 33	
	min	max	min	max	min	max
Genome Haploid Length (bp)	649,475,766	649,949,877	659,441,333	659,890,585	660,289,984	660,728,342
Heterozygosity (%)	0.654	0.660	0.590	0.594	0.579	0.583
Genome Repeat Length (bp)	88,141,621	88,205,963	57,904,063	57,943,510	54,791,371	54,827,747
Genome Unique Length (bp)	561,334,145	561,743,914	601,537,270	601,947,074	605,498,612	605,900,595
Read Error Rate (%)	0.527		0.477		0.468	

[1]Kmers are unique subsequences of a sequence of length k. The estimated genome size varies according to kmer value. The estimated haploid genome lengths obtained from kmer 31 and kmer 33 are very close.

Table 2. Contig and scaffold properties of *T. ilisha* genome.

Contig			Scaffold	
Parameters	**Value**	**%**	**Parameters**	**Value**
Read pairs	769,262,291	-	Scaffold Number	100181
Contig Number	1724390	-	Mean Scaffold size	7090
Mean Contig Size	378	-	Longest Scaffold	832708
Median Contig Size	209	-	Shortest Scaffold	200
Longest Contig	27277	-	N10	254367
Shortest Contig	100	-	N30	118059
Contig >100bp	1704389	98.84	N50	64157
Contig >500bp	335422	19.45	N70	26438
Contig >1K	121379	7.04	N90	4991
Contig >10K	58	0.00	N count	96164104
Contig N50 (bp)	594	-	Assembled Genome Size (bp)	710279582
G+C content %	-	43.01	G+C content (%)	42.95

Table 3. Significant matches (blastp e ≤ 0.001) of *T. ilisha* genes with other vertebrates.

Species	No. of matches	% match
Rhincodon typus	27062	72.26
Cyprinus carpio	28373	75.76
Takifugu rubripes	28325	75.63
Salmo salar	29339	78.34
Mus musculus	26480	70.71
Homo sapiens	27497	73.42

3,578 complete (C: 78.1%), 3,456 complete and single-copy (S: 75.4%), 122 complete and duplicated (D: 2.7%), 351 fragmented (F: 7.7%) and 655 missing BUSCOs (M: 14.2%). These results suggest higher completeness of the *T. ilisha* genome assembly of the Bay of Bengal. The BUSCO analysis of its closely related species, the European sardine, showed 84% genome completeness (Louro *et al.,* 2018). The completeness of genome assembly may depend on the sequencing platform used. For example, 92.3% BUSCO completeness was obtained using only the Illumina reads compared to 94.2% completeness from the Illumina + Nanopore reads in the Murray cod (*Maccullochella peelii*) (Austin *et al.,* 2017).

AUGUSTUS *ab initio* gene prediction was performed to predict protein-coding genes. We found 37,450 protein coding genes from the assembled *T. ilisha* genome (Mollah *et al.,* 2019a). To annotate the proteins, predicted amino acid sequences of *T. ilisha* were aligned against the NCBI non-redundant protein sequences (nr) database of other vertebrates (Table 3) using BLASTP. Among the five vertebrate genomes compared, a minimum of 70.71% genes of *T. ilisha* was annotated by *Mus musculus* and a maximum of 78.34% by *Salmo salar* (Table 3). The numbers of predicted protein coding genes in two other important Clupeids, the Atlantic herring and the European sardine, were 23,336 and 29,408, respectively (Martinez Barrio *et al.,* 2016; Louro *et al.,* 2018).

We identified a total of 792,939 isolated SNPs in *T. ilisha* genome, of which 510,251 were transitions and 282,688 were transversions (Table 4) (Mollah *et al.,* 2019b). We also detected 155,574 indels ranging in sizes from 1 to 60 nucleotides (Figure 4). A total of 5.3 million raw SNPs in the Atlantic stock and 5.2 million SNPs in the Baltic stock of the Atlantic herring were detected by Feng *et al.* (2017). In contrast, Louro *et al.* (2018) identified a total of 2.3 million filtered heterozygous SNPs in the European sardine. Since there is no high quality reference genome available in *T. ilisha*, we used discoSNP++ because of its nobility and efficiency in detection of SNPs from unassembled genome sequences. Uricaru *et al.* (2015) genotyped 384 SNPs out of a total of 312,088 discoSNP++

Table 4. Single nucleotide polymorphism (SNP) and Indels in the *T. ilisha* genome.

SNPs/indels	Type	Number	%
Total SNPs	-	792939	100
Transitions	A>G	256209	32.31
	C>T	254042	32.04
Transversions	A>C	75912	9.57
	A>T	78469	9.90
	C>G	55810	7.04
	G>T	72497	9.14
Transition : Transversion	1.8 : 1		
Number of Indels	155574		

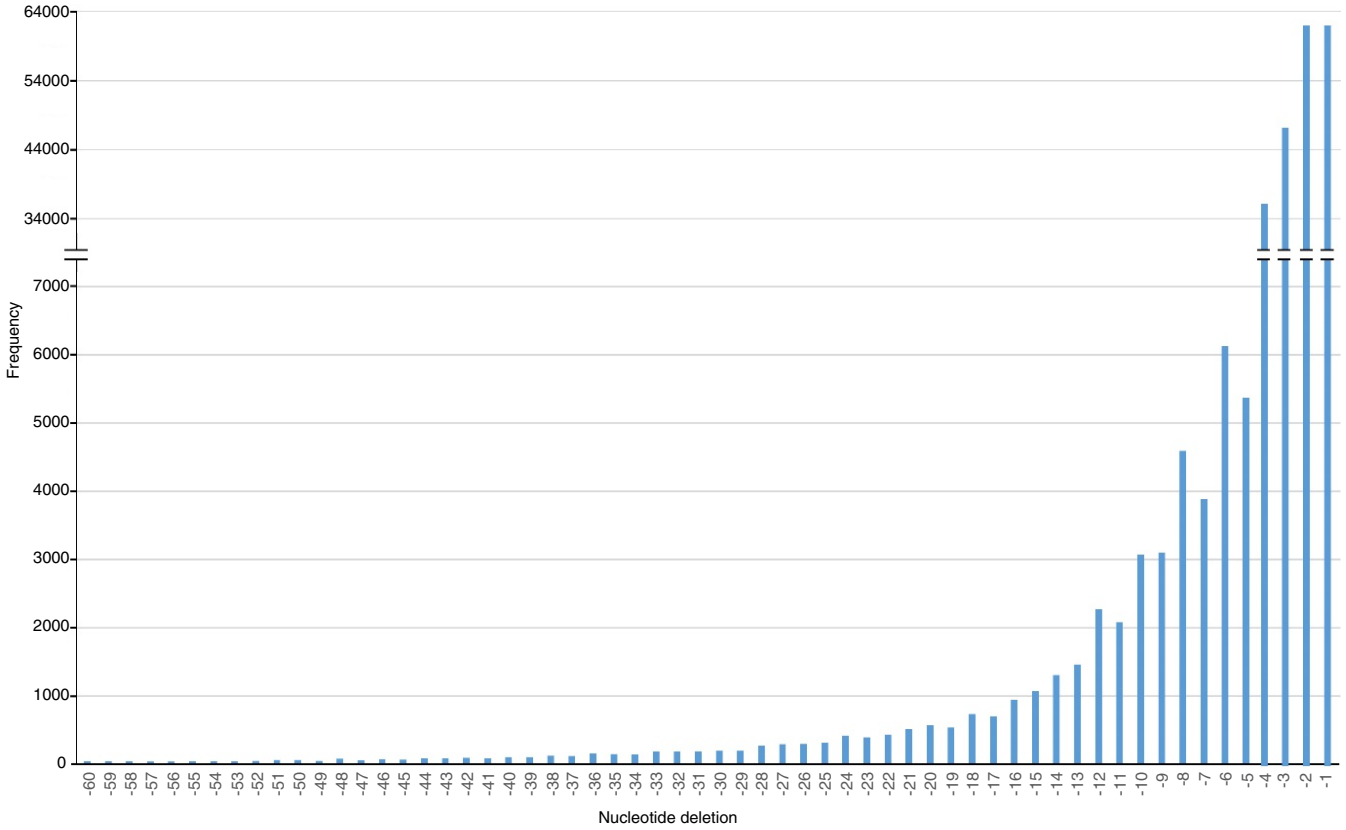

Figure 4. Distribution of Indels in the *T. ilisha* genome. The indel values ranged from 1 to 60 nucleotides. It shows that the frequency of indels decreased with the increase in size.

predicted SNPs, of which 368 (95.8%) were accurately validated in the tick *Ixodes ricinus*.

Conclusions

We report here the first *de novo* genome of *T. ilisha* from the Bay of Bengal. The assembled genome can be used as a reference for genetic studies of *T. ilisha* and related species. The SNPs generated could provide a valuable resource for resolving stock disputes and phylogenetic or adaptation investigation of the Clupeidae family.

Data availability

Underlying data

Tenualosa ilisha collected from Bay of Bengal, Accession number SAMN07556897: https://www.ncbi.nlm.nih.gov/biosample/?term=SAMN07556897

Tenualosa ilisha whole genome sequencing and assembly, Accession number SRP116260: https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP116260

Tenualosa ilisha whole genome shotgun sequencing project, Accession number SCED00000000: https://www.ncbi.nlm.nih.gov/nuccore/SCED00000000

Zenodo: Amino acid sequences of the proteins predicted from the whole genome of hilsa shad (Tenualosa ilisha) of the Bay of Bengal. https://doi.org/10.5281/ZENODO.2539223 (Mollah *et al.,* 2019a)

Zenodo: Single Nucleotide Polymorphisms (SNPs) identified from the whole genome sequences of hilsa shad (Tenualosa ilisha) of the Bay of Bengal. https://doi.org/10.5281/ZENODO.2538155 (Mollah *et al.,* 2019b)

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgements

We thank Bangladesh Fisheries Research Institute (BFRI) for assistance in sample collection. We also acknowledge Bangladesh Agricultural University Research System (BAURES) for institutional support and BdREN (Bangladesh Research and Education Network), University Grants Commission of Bangladesh, for sharing high performance computing infrastructure.

References

Ahmed A, Islam M, Azam M, *et al.*: **RFLP analysis of the mtDNA D-loop region in Hilsa shad (*Tenualosa ilisha*) population from Bangladesh.** *Indian J Fish.* 2004; **51**(1): 25–31.
Reference Source

Ahsan DA, Naser MN, Bhaumik U, *et al.*: **Migration, spawning patterns and conservation of *Hilsa* Shad (*Tenualosa ilisha*) in Bangladesh and India.** 1st ed. Academic Foundation, New Delhi, India, 2014.
Reference Source

Altschul SF, Madden TL, Schäffer AA, *et al.*: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res.* 1997; **25**(17): 3389–3402.
PubMed Abstract | Publisher Full Text | Free Full Text

Andrews S: **FastQC -A quality control tool for high throughput sequence data.** [WWW Document]. 2018.
Reference Source

Austin CM, Tan MH, Harrisson KA, *et al.*: ***De novo* genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (*Maccullochella peelii*), from Illumina and Nanopore sequencing read.** *GigaScience.* 2017; **6**(8): 1–6.
PubMed Abstract | Publisher Full Text | Free Full Text

Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics.* 2014; **30**(15): 2114–2120.
PubMed Abstract | Publisher Full Text | Free Full Text

DoF: **National Fish-Week 18-24 July, 2018 Compendium.** 2018.
Reference Source

Feng C, Pettersson M, Lamichhaney S, *et al.*: **Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate.** *eLife.* 2017; **6**: pii: e23907.
PubMed Abstract | Publisher Full Text | Free Full Text

Figueras A, Robledo D, Corvelo A, *et al.*: **Whole genome sequencing of turbot (*Scophthalmus maximus*; Pleuronectiformes): a fish adapted to demersal life.** *DNA Res.* 2016; **23**(3): 181–192.
PubMed Abstract | Publisher Full Text | Free Full Text

Freyhof J: **The IUCN Red List of Threatened Species 2014.** International Union for Conservation of Nature - IUCN. 2014.
Reference Source

Louro B, De Moro G, Garcia CM, *et al.*: **A haplotype-resolved draft genome of the European sardine (*Sardina pilchardus*).** *bioRxiv.* 2018; 441774.
Publisher Full Text

Luo R, Liu B, Xie Y, *et al.*: **SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler.** *GigaScience.* 2012; **1**(1): 18.
PubMed Abstract | Publisher Full Text | Free Full Text

Machado AM, Felício M, Fonseca E, *et al.*: **A resource for sustainable management: *De novo* assembly and annotation of the liver transcriptome of the Atlantic chub mackerel, *Scomber colias*.** *Data Brief.* 2018; **18**: 276–284.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Martinez Barrio A, Lamichhaney S, Fan G, *et al.*: **The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing.** *eLife.* 2016; **5**: pii: e12081.
PubMed Abstract | Publisher Full Text | Free Full Text

Mazumder SK, Alam MS: **High levels of genetic variability and differentiation in hilsa shad, *Tenualosa ilisha* (Clupeidae, Clupeiformes) populations revealed by PCR-RFLP analysis of the mitochondrial DNA D-loop region.** *Genet Mol Biol.* 2009; **32**(1): 190–196.
PubMed Abstract | Publisher Full Text | Free Full Text

Milton DA: **Status of hilsa (*Tenualosa ilisha*) management in the Bay of Bengal: an assessment of population risk and data gaps for more effective regional management.** Bay of Bengal Large Marine Ecosystem Project (BOBLME), Phuket, Thailand. 2010.
Reference Source

Mollah MB, Khan MG, Islam MS, *et al.*: **First Draft Genome Sequence of Anadromous Hilsa Shad (*Tenualosa ilisha*) and Development of Genomic Resources for Conservation.** In: *Plant and Animal Genome Conference XXVI*. San Diego, 2018; P0330.
Reference Source

Mollah MBR, Khan MGQ, Islam MS, *et al.*: **Amino acid sequences of the proteins predicted from the whole genome of hilsa shad (*Tenualosa ilisha*) of the Bay of Bengal.** 2019a.
http://www.doi.org/10.5281/ZENODO.2539223

Mollah MBR, Khan MGQ, Islam MS, *et al.*: **Single Nucleotide Polymorphisms (SNPs) identified from the whole genome sequences of hilsa shad (*Tenualosa ilisha*) of the Bay of Bengal.** 2019b.
http://www.doi.org/10.5281/ZENODO.2538155

Pillay SR, Rosa JH: **Synopsis of biological data on hilsa, *Hilsa ilisha* (Hamilton) 1882.** FAO fisheries biology Synopsis 25. 1963.

Salini JP, Milton DA, Rahman MJ, *et al.*: **Allozyme and morphological variation throughout the geographic range of the tropical shad, hilsa *Tenualosa ilisha*.** *Fish Res.* 2004; **66**(1): 53–69.
Publisher Full Text

Sambrook JF, Russell DW, (Eds.): **Molecular Cloning: A Laboratory Manual, 3rd ed.** Cold Spring Harbor Laboratory Press, New York. 2001.
Reference Source

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–3212.
PubMed Abstract | Publisher Full Text

Uricaru R, Rizk G, Lacroix V, *et al.*: **Reference-free detection of isolated SNPs.** *Nucleic Acids Res.* 2015; **43**(2): e11.
PubMed Abstract | Publisher Full Text | Free Full Text

Vurture GW, Sedlazeck FJ, Nattestad M, *et al.*: **GenomeScope: fast reference-free genome profiling from short reads.** *Bioinformatics.* 2017; **33**(14): 2202–2204.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ? ?

Version 1

Reviewer Report 01 October 2019

https://doi.org/10.5256/f1000research.20044.r53826

?

Shotaro Hirase
Fisheries Laboratory, Graduate School of Agricultural and Life Sciences, University of Tokyo, Hamamatsu, Japan

Authors constructed a draft genome assembly by whole genome sequencing and identified SNPs from *Tenualosa ilisha* of the Bay of Bengal. This paper is well written, and that the analysis methods are generally appropriate. I have several suggestions to improve this manuscript as follows:

1. Authors described that “We captured ten *T. ilisha* specimens from the seashore of the Bay of Bengal” in the Method section. But it is confusing because one individual was used finally in this study. Please revise this point.

2. Authors said that three different ecological types of hilsa shad have been recognized in Bangladesh. What kind of ecotype is the individual used for genome assembly expected to have? This information may be important for future studies focusing genomic bases of the ecotype divergence.

3. Authors described that “These results suggest higher completeness of the *T. ilisha* genome assembly of the Bay of Bengal.” It is a difficult problem whether it is higher completeness or not, but it seems that there are too many missing genes to say higher completeness.

4. SNP typing was performed with a reference-free method. However, I think that it would be better to detect SNP using the reference constructed newly in this study (mapping and SNP calling by samtools etc). The SNP information (e.g. vcf file) based on the constructed reference genome should be useful for readers.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

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

Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Population genomics.

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 21 May 2019

https://doi.org/10.5256/f1000research.20044.r47105

?

Yuzine B. Esa
Department Of Aquaculture, Faculty of Agriculture, University Putra Malaysia (UPM), Serdang, Malaysia

Overall, the manuscript was well written and contains all relevant information in accordance with its scope.

Probably, it would be better if Table 3 is being presented in a figure form (e.g. Venn diagram), showing overlapping regions of genes between species.

A linkage map (e.g using zebrafish as reference genome) and gene synteny analyses would be another interesting aspect that can be included.

It would be more interesting if the genome of male and female of this species can be compared and analysed, since the fish is known as protandrous hermaphrodite.

References

1. Wang Y, Lu Y, Zhang Y, Ning Z, Li Y, Zhao Q, Lu H, Huang R, Xia X, Feng Q, Liang X, Liu K, Zhang L, Lu T, Huang T, Fan D, Weng Q, Zhu C, Lu Y, Li W, Wen Z, Zhou C, Tian Q, Kang X, Shi M, Zhang W, Jang S, Du F, He S, Liao L, Li Y, Gui B, He H, Ning Z, Yang C, He L, Luo L, Yang R, Luo Q, Liu X, Li S, Huang W, Xiao L, Lin H, Han B, Zhu Z: The draft genome of the grass carp (Ctenopharyngodon idellus)

provides insights into its evolution and vegetarian adaptation.*Nat Genet*. 2015; **47** (6): 625-31 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Fish genetics.

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 29 April 2019

https://doi.org/10.5256/f1000research.20044.r47164

Joel A. Malek
Genomics Core, Department of Genetic Medicine, Genomics Laboratory, Weill Cornell Medicine–Qatar, Doha, Qatar

The authors provide a new reference genome for the hilsa shad fish of significant commercial importance in the Bay of Bengal. The methods they use are standard and the resulting completeness analysis by BUSCO showed approximately 78% of core genes present. While it would be ideal to have added libraries for scaffolding the assembly (reads from 5-20kb inserts, or 10X Genomics reads, or Pacific Bioscience), this initial assembly will certainly allow some general comparisons and the building of genomic resources to better study the fish.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, genome sequencing, assembly and annotation.

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.

Comments on this article

Version 1

Reader Comment 20 Apr 2019

Chungkeun Lee, Inha university, Incheon, South Korea

This study aim to genome assembly of hilsa shad. these sequences are assembled by SOAPdenovo2. But, the progress of sequences assembly doesn't represent.
I would like to see more information on the series of the assembly procedure on adequate explanation or explicit mean.

Competing Interests: No competing interests were disclosed.