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

Md. Bazlur Rahman Mollah¹, Mohd Golam Quader Khan², Md Shahidul Islam³, Md Samsul Alam ²

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

**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
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.

**Copyright:** © 2019 Mollah MBR et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**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: awaiting peer review] 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:chloroform: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 GCC 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 2x 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,
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
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

---

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

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

*1Kmers 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.*

---

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 300x (a) and 10000x coverage (b).
Table 2. Contig and scaffold properties of T. ilisha genome.

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

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

<table>
<thead>
<tr>
<th>Species</th>
<th>No. of matches</th>
<th>% match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhincodon typus</td>
<td>27062</td>
<td>72.26</td>
</tr>
<tr>
<td>Cyprinus carpio</td>
<td>28373</td>
<td>75.76</td>
</tr>
<tr>
<td>Takifugu rubripes</td>
<td>28325</td>
<td>75.63</td>
</tr>
<tr>
<td>Salmo salar</td>
<td>29339</td>
<td>78.34</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>26480</td>
<td>70.71</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>27497</td>
<td>73.42</td>
</tr>
</tbody>
</table>

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 peelli) (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.

<table>
<thead>
<tr>
<th>SNPs/indels</th>
<th>Type</th>
<th>Number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total SNPs</td>
<td>-</td>
<td>792939</td>
<td>100</td>
</tr>
<tr>
<td>Transitions</td>
<td>A&gt;G</td>
<td>256209</td>
<td>32.31</td>
</tr>
<tr>
<td></td>
<td>C&gt;T</td>
<td>254042</td>
<td>32.04</td>
</tr>
<tr>
<td>Transversions</td>
<td>A&gt;C</td>
<td>75912</td>
<td>9.57</td>
</tr>
<tr>
<td></td>
<td>A&gt;T</td>
<td>78469</td>
<td>9.90</td>
</tr>
<tr>
<td></td>
<td>C&gt;G</td>
<td>55810</td>
<td>7.04</td>
</tr>
<tr>
<td></td>
<td>G&gt;T</td>
<td>72497</td>
<td>9.14</td>
</tr>
<tr>
<td>Transition : Transversion</td>
<td>1.8 : 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Indels</td>
<td></td>
<td>155574</td>
<td></td>
</tr>
</tbody>
</table>

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

Tenuala osa ilisha whole genome shotgun sequencing project, Accession number SCED0000000: https://www.ncbi.nlm.nih.gov/nuccore/SCED00000000

Zenodo: Amino acid sequences of the proteins predicted from the whole genome of hilsa shad (Tenuala osa 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 (Tenuala osa ilisha) of the Bay of Bengal. https://doi.org/10.5281/ZENODO.2538155 (Mollah et al., 2019b)

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

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


Milton DA: Status of hilsa (Tenuala osa 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.


http://www.doi.org/10.5281/ZENODO.2539223

Mollah MBR, Khan MGO, Islam MS, et al.: Single Nucleotide Polymorphisms (SNPs) identified from the whole genome sequences of hilsa shad (Tenuala osa ilisha) of the Bay of Bengal. 2019b.

http://www.doi.org/10.5281/ZENODO.2538155


Publisher Full Text


Publisher Full Text
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com