PROJECT COMPLETION REPORT

Whole genome sequencing of Indian Major Carps (*Labeo rohita*, *Catla catla/Gibelion catla*, *Cirrhinus cirrhosus*, *Labeo calbasu*, *Platanista gangetica*) and genome annotation to explore genetic variations.

JULY
2021

Implemented jointly by

Halda River Research Laboratory
Department of Zoology
University of Chittagong (CU)
Hathazari, Chittagong- 4331, Bangladesh

Genomics Research Group
Chittagong Veterinary and Animal Sciences University (CVASU)
Khulshi, Chittagong- 4202, Bangladesh

Sponsored by
Palli Karma-Sahayak Foundation (PKSF)
Integrated Development Foundation (IDF)

PROJECT COMPLETION REPORT

Submitted by: Prof. AMAM Zonaed Siddiki and Professor MM Kibria

Project Title

Whole genome sequencing of Indian Major Carps (*Labeo rohita*, *Catla catla/Gibelion catla, Cirrhinus cirrhosus, Labeo calbasu, Platanista gangetica*) and genome annotation to explore genetic variations

Funded by	:	Integrated Development Foundation (IDF)
Implementing organization	:	Department of Zoology, University of Chittagong & Genomics Research Group, CVASU, Chittagong
International collaborator	:	Dr. Abdul Baten, AgResearch, New Zealand Mr Masum Billah, Northwest A & F University, China
Time duration	:	January-December, 2020

Project Personnel

Principal investigator:

Professor Dr. AMAM Zonaed Siddiki
Team Lead, Genomics Research Group
Chittagong Veterinary and Animal Sciences University
Khulshi, Chittagong-4202 Bangladesh
Tel. 0088 0171 771 8884
Email: zsiddiki@gmail.com

Associate Investigator:

Dr. Md. Manzoorul Kibria
Professor and Coordinator,
Halda River Research Laboratory,
Department of Zoology,
University of Chittagong, Bangladesh
Tel. 01711164408
Email. mnzoorul@yahoo.com

Project Scientists: (Research Fellow)
Ms. Naznin Islam, MS Fellow, Department of Zoology, CU
Ms. Azlina Kabir, MS Fellow, Department of Zoology, CU
Mr. Suma Akter, MS Fellow, Department of Zoology, CU
Ms. Syeda Sabiha Mostafa, MS Fellow, Department of Zoology, CU
Mr. Abdullah Al Asek, Graduate Student, Department of Zoology, CU
Mr. Md. Anamul Bhuiyan, Bioinformatician
Mr. Shamsur Rahaman, Bioinformatician
Mr. Masum Billah, Bioinformatician

Project Staff:
Mr. Salahuddin, Accountant
Mr. Mohammed Hossain, Computer operator

Preamble

In Bangladesh, the Halda River is one of the most significant rivers due to natural breeding ground and significant seedling sources of freshwater fishes. Very recently it has been declared as **"Bangabandhu Fisheries Heritage"** to commemorate the **"Mujib Barsha"**. This river provides favorable physicochemical factors which creates a congenial environment during monsoon between April and June. Therefore, due to this river's unique feature, the Indian major carps spawn naturally which makes it the most economic heritage of this country as well as in South Asia.

Whole genome sequence (WGS) approach has become very useful in recent time as relevant technologies are readily available to the scientific community. It is crucial to explore the WGS of different carp species as well as that of River Dolphin to reveal novel and accurate ways to explore the heterozygous nature of the carp genome and to foster further research for their conservation and branding.Genetic variation is essential in maintaining the evolutionary potential and robustness of a population. Preservation of genetic diversity is also crucial for both wild and cultivated species to preserve the fitness of a population. Accessibility to high-quality genomics information of these species will aid us in interpreting its evolutionary aspects along with aptitude in adapting to different territories as well as environmental fluctuations, through comparative studies with other subspecies.

We are grateful to our sponsors for all out support to conduct this study. All the fish and dolphin specimens were kindly provided by the local authorities including Upazila Fisheries Office, Hathazari and UNO, Hathazari. We are grateful to the Vice Chancellor of CVASU to provide laboratory support for the molecular and bioinformatics analyses. All researchers, avademicians, students and support staff of the Halda River Research Lab (HRRL) deserve special thanks for kind support during the project period. We are especially grateful to the members of the Halda Research Lab management Committee for all out support and cooperation in conducting this research.

We hope and believe that the findings from this study will provide substantial genomics and proteomics datasets and the researcher community were highly benefitted to accomplish further relevant research on this unique resource of Bangladesh.

Professor Dr. Md. Manzoorul Kibria
Coordinator,
Halda River Research Laboratory
Department of Zoology
University of Chittagong
Bangladesh

Professor Dr. AMAM Zonaed Siddiki
Team Lead, Genome Research Group
Chittagong Veterinary and Animal Sciences
University (CVASU) Khulshi, Chittagong-4202
Bangladesh

Acknowledgements

We are grateful to our collaborators and partners for extraordinary support towards successful implementation of the project activities.

Index

Abbreviations

bp: Base pair

BUSCO: Benchmarking Universal Single-Copy Orthologs

DNA: Deoxyribonucleic acid.

EDTA: Ethylene_diamine-tetraacetic acid

Gbp: Giga base pair

GO: Gene ontology

IMC: Indian Major Carp

Kb: Kilo base pair

Mb: Mega base pair

qPCR: Quantitative Polymerase Chain Reaction

PCR: Polymerase Chain Reaction

SRA Sequence Read Archive

WGS: Whole Genome Sequencing

BLAST: Basic Local Alignment Search Tool

BWA: Burrows-Wheeler Aligner

CVASU: Chittagong Veterinary and Animal Sciences University

CU: University of Chittagong

HRRL: Halda River Research Lab

Software/ Tools used for the Genome Sequencing

Software name	URL/ website	Function
BWA	http://bio-bwa.sourceforge.net	Mapping
SAMTOOLS	http://samtools.sourceforge.net	Mapping
GATK	https://gatk.broadinstitute.org/hc/en-us	Mapping
NovoPlasty	https://github.com/ndierckx/NOVOPlasty	Assembly
MITOS	http://mitos.bioinf.uni-leipzig.de/index.py	Annotation
GeSeq	https://chlorobox.mpimp-golm.mpg.de/geseq.html	Annotation
CLC main workbench	https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-clc-main-workbench/	Phylogenetic tree
MEGA X	https://www.megasoftware.net	Phylogenetic tree
OGDRAW	https://chlorobox.mpimp-golm.mpg.de/OGDraw.html	Visualization
Software	URL/ Website	Function
FastQC	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/	Quality Check
ABYSS	https://github.com/bcgsc/abyss	Assembly
Platanus	http://platanus.bio.titech.ac.jp	Assembly
Repeatmasker	http://www.repeatmasker.org	Repeat content
BUSCO	https://busco.ezlab.org	To assess genome assembly completeness
MAKER	https://www.yandell-lab.org/software/maker.html	Annotation
INTERPROSCAN	http://www.ebi.ac.uk/InterProScan/	Protein Analysis/ Functional annotation
OrthoVenn2	https://orthovenn2.bioinfotoolkits.net/home	To show orthologous gene cluster

Project background

Aquaculture in Bangladesh had experienced massive growth during the last twenty-five years and currently, 50% of total fish production comes from aquaculture (FRSS, 2012) which was only 20% in the 1980s (Dey *et al*., 2008).

The Halda River is one of the most significant rivers of Bangladesh due to natural breeding ground and significant seedling sources of freshwater fishes. This river provides favorable physicochemical factors which creates a congenial environment during monsoon between April and June (Kibria *et al*., 2009). Therefore, due to this river's unique feature, the Indian major carps spawn naturally which makes it the most economic heritage of this country as well as in South Asia (Akter and Ali, 2012; Kabir *et al*., 2013). The three Indian major carp species such as catla (*Catla catla*), rohu (*Labeo rohita*), and mrigal (*Cirrhinus cirrhosus*) together constitute a traditional and popular aquaculture practice termed as carp polyculture system in Bangladesh. *C*. *catla* is an endemic species in Bangladesh, India, Pakistan and Myanmar (Basak *et al*., 2014). Indian major carps group of the family Cyprinidae is naturally distributed in the rivers of Bangladesh, India, Pakistan and Myanmar (Talwar and Jhingran, 1991). The column feeder rohu (*L. rohita*), along with the surface feeder catla, *Catla catl*a (Hamilton) and bottom feeder mrigal, *Cirrhinus cirrhosus* (Hamilton) constitute an essential and popular aquaculture practice in the region named carp polyculture system.

Genetic variation is essential in maintaining the evolutionary potential and robustness of a population (Vandewoestijne *et al*., 2008). Preservation of genetic diversity is also crucial for both wild and cultivated species to preserve the fitness of a population (Vandewoestijne *et al*., 2008). Accessibility to high-quality genomics information of these species will aid us in interpreting its evolutionary aspects along with aptitude in adapting to different territories as well as environmental fluctuations, through comparative studies with other subspecies.

Whole genome sequence (WGS) approach has become very useful in recent time as relevant technologies are readily available to the scientific community. Researchers already conducted a de novo assembly of *Labeo rohita* (Breed: Jayanti) whole-genome sequence to reveal novel and accurate ways to explore the heterozygous nature of the carp genome. Several studies were conducted to investigate genetic differentiation of wild & hatchery populations of Indian major carps Such as *Labeo rohita, Catla catla, Cirrhinus cirrhosis* (Ullah *et al*., 2012). With few preliminary studies, this is high time for genome-level investigation of these major Indian carps to explore their vulnerability towards environmental alterations, susceptibility to illness and disease as well as other crucial "biological phenomena."

Several studies were conducted to investigate genetic differentiation of wild & hatchery populations of Indian major carps Such as *Labeo rohita, Catla catla, Cirrhinus cirrhosis* (Ullah *et al*., 2012). With few preliminary studies, this is high time for genome-level investigation of these major Indian carps to explore their vulnerability towards environmental alterations, susceptibility to illness and disease as well as other crucial "biological phenomena."

The Ganges river dolphin, commonly known as as "shusuk" in Bangla *Platanista gangetica* (Roxburgh, 1801) is a freshwater dolphin distributed throughout the Ganges–Brahmaputra–Meghna and Karnaphuli–Sangu river systems of Nepal, India, Bangladesh, and potentially Bhutan (Mohan et al. 1997; Sinha et al. 2000; Smith et al. 2001). Its geographical range has been reduced, and its abundance has declined in many areas where the animals still appear (Reeves, et al, 1995, Smith et, al, 2012, paudel, et al, 2015). The International Union for Conservation of Nature (IUCN) revised the threatened status of the Ganges river dolphin from vulnerable (Klinowska, 1991) to endangered as per IUCN Global Red List guidelines (Smith et, al, 2012). At present, it is thought that there are about 3500 individuals of this species throughout its distribution range (Sinha et al, 2014). This species is placed on the First Schedule of Bangladesh Wildlife (Conservation & Security) Act, 2012, and it is classified as endangered because of the numerous anthropogenic and natural constraints to its survival, for example, water pollution, boat traffic, and reduction of water flow during the dry season (IUCN Bangladesh, 2000); however, the population of river dolphins has declined rapidly in recent years with much of their habitat already degraded (Smith, et al, 2012).

There are nearly 200 species of aquatic vertebrates, almost exclusively fishes, in the Brahmaputra River System. This faunal composition includes not only a wide variety of food-fishes, but also about 50 varieties of aquarium fishes. The most spectacular animal in the Brahmaputra is undoubtedly the river dolphin, *Platanista gangetica*. The population of many species, particularly of the dolphin, is in steady decline. River dolphins, mostly juveniles, often entangle themselves in gill and drift nets, while feeding on trapped fishes. Although vulnerable due to respective extinction risks, river dolphin is the most inhabited dolphin- subspecies containing the most genetically diverse and has the strongest possibility of survival in the freshwater. Accessibility to high quality genomics information of this animal will assist us in interpreting its evolutionary aspects along with capability in adapting to different habitats as well as environmental changes, through comparative studies with other subspecies.

Whole genome sequence (WGS) approach has become very useful in recent time as relevant technologies are readily available to the scientific community. Recently,

reseahers conducted a *De Novo* assembly of a bottlenose dolphin whole-genome sequence to reveal novel and accurate ways to explore the heterozygous nature of the dolphin genome (Martinex et al, 2018). Likeewise, it is essential to analyse the whole genome sequece of river dolphin to explore their vulnerability towards environmental alterations, susceptibility to illness and disease as well as other crucial factors associated with its conservation.

Significance of the project

The Halda river occupies a prominent position as a natural breeding ground of Indian major carps among existing open inland water bodies of Bangladesh. This river is unique and distinct due to genetically purity of fertilized spawn of *Labeo rohita* and other carps. Simultaneously this river has also a great significant role in the livelihood of a considerable number of people who are egg collectors and Hatcher and supply freshwater fishes.

Though about 50 years ago, 40-50 thousand kgs of eggs could be collected from Halda river; in the last year, only about 12 kgs of egg were collected from it. This is because the natural breeding has become constrained by the degradation of habitats as a result of environmental modifications and anthropological intervention (Akhtar *et al.,* 2017). The density of the species has declined significantly because of overfishing, constructions of sluice gates in the branch rivers and rubber dam and hunting mother fishes, the quality of eggs released in the river by mother fishes are also gradually reducing to an alarming rate. On the contrary, the demand for fish fry has increased several folds through aquaculture expansion. However, little attention has been given to explore the unique genetic features of these fish species which has considerable economic significance. Thus it is essential to understand the genetic composition of Indian major carps for the management of their natural population in the river. **Thereby the branding of the Halda carps can be ensured** which were invaluable for commercialization and entrepreneurship of its resources.

Traditionally, it is recommended that the genetic structure of the base population (e.g., Brood bank collections) be analyzed so that this information can be used to assess the quality of the hatchery stocks in future. Whole genome sequencing (WGS) is one of the modern technique to achieve this objective. As entire genome map of Jayanti Rui is available, **now it is possible to establish the draft genome of these four carps, fully characterize the gene for any specific trait.** It will increase our understanding of the

transcriptome and thereby select particular gene groups unique to halda river population.

The genome sequencing of Indian major carps will provide valuable information on genome organization, evolutionary divergence, conservation and overall endemic diversity. It's also important to identify some genes related to a particular trait such as those associated with adaptation, evolution. **Comparative genome analysis will also reveal more information on fishes which have desired characteristics.** The genetic tools and experiences from this current project may offer a new frontier to disease management, the ability to better support towards their conservation.

Fresh water mammals from different mammalian orders share several phenotypic traits adapted to the aquatic environment and therefore represent a classic example of convergent evolution. The population of many species, particularly of the dolphin, is in steady decline. River dolphins, mostly juveniles, often entangle themselves in gill and drift nets, while feeding on trapped fishes. This species is placed on the First Schedule of Bangladesh Wildlife (Conservation & Security) Act, 2012, and it is classified as endangered because of the numerous anthropogenic and natural constraints to its survival, for example, water pollution, boat traffic, and reduction of water flow during the dry season; however, the population of river dolphins has declined rapidly in recent years with much of their habitat already degraded. So, it is necessary to develop modern and fast monitoring tools for endangered biodiversity. To do this whole genome sequencing is important. As whole genome map of bottlenose dolphin is available , now it is possible to establish a draft genome of river dolphin, fully characterize the genes and their for any specific trait. It will increase our understanding about the transcriptome and thereby select particular gene groups unique to river dolphin population.

It is expected that, this study will provide a useful platform for the functional genome and **conservation research of Halda river carps and other aquatic animals including Dolphins** in the future. The knowledge will help us develop a better policy for their breeding, behavioral pattern analyses with a view towards their in situ conservation. The Government of Bangladesh, as well as international communities, has a particular interest in Halda river population conservation. **The present proposal will certainly help our effort towards their effective conservation.**

Aims and Objectives

The proposed research activities were focused on following specific objectives:-

- To develop a draft whole genome and mitogemome of Indian Major Carps (*Labeo rohita*, *Catla catla/Gibelion catla*, *Cirrhinus cirrhosus*, *Labeo calbasu*) with high sequence coverage (50X) using state-of-the-art Next-generation sequencing (NGS) approach and other modern computational biology tools

- To determine the genetic characterization & genetic diversity of selected Indian major carps, as well as to detect distinct features of Halda river carps with a view to branding of the Halda carps

- To develop a draft genome and mitogenome of *Platanista gangetica* (South Asian River Dolphin) with high sequence coverage using state-of-the-art next generation sequencing (NGS) approach and other modern computational biology tools

- To strengthen national and international collaboration towards capacity building and human resource development to promote River Dolphin conservation

Methodology of the Study

Sampling and DNA isolation:

The fish specimens provided by local Fisheries Department and preserved at the -20 degree freezer of Halda River Research Lab (HRRL) of Department of Zoology, University of Chittagong were used for WGS study. All fishes were collected from river Halda (Longitude/Latitude-22°28'56.09"N & 91°54'07.62"E) and during 2019-20. The muscle tissue was dissected and stored with 95% ethanol, later a high molecular weight genomic DNA was extracted using the AddPrep Genomic DNA extraction kit (AddPrep Inc., Korea). Samples were further tested for qualitative and quantitative evaluation of the DNA through Fluorumetry. All the methods had been performed following the "Regulations for Animal Experiments in Chittagong Veterinary and Animal Sciences University" as required.

Library preparation

The extracted DNA was cleaned up using commercial kits and sent for subsequent library preparation and whole-genome sequencing (WGS) at the BGI genomics, China. Using Next-generation sequencing (NGS) technology on an IlluminaNovaSeq 6000 platform a total of 45 Gigabase pair (Gb) of subread bases with a read length of 150 bp were induced. After sequencing quality of primary sequence reads and trimmed sequencing reads were investigated using FastQC version 0.11.8. The quality control of the reads was done including removing adaptor sequences, low-quality reads and contamination from raw reads using BFC. A total of 55,264,452,615 clean reads were included in the assembly with 48X coverage.

Genome assembly:

To assemble the whole genome ABySS ver. 1.9.0 program were used along with Platanus ver.1.2.4 assembler. Since there is currently no de novo assembler assured to outperform others and as assemblers overall performance can differ relying on the dataset, two unique assemblers were used and to determine the best assembler an assembly evolution was subsequently performed. Both assemblers follow the classic De Bruijn graph illustration even though the assembly algorithm differs across methods. Finally, BUSCO v.4.1.2 was operated to assess the quality of the assembly in respect of gene completeness.

WGS data processing and filtering

The raw sequence data quality were assessed using FASTX- Toolkit (Pearson et al. 1997; doi: 10.1006/geno.1997.4995), Trimmomatic (http://www.usadellab.org/cms/?page =trimmomatic) and Trim Galore (http://www. bioinformatics. babraham.ac.uk/projects/trim_galore/). Low quality reads were filtered and potential sequencing errors were removed. We defined the following types of reads as low quality reads:

1. Reads with Ns more than 10% of their length;
2. Reads with low quality base more than 50% of their length (quality score ≤ 5);
3. Reads contain more than 10 bp adapter sequences (allowing ≤ 2 bp mismatches);
4. Small insert size paired-end reads that were overlapped (≥ 10bp);

Read1 and read2 of paired-end reads are completely identical. These paired-end reads would be considered as artefacts of PCR experiment.We also implemented an error correction procedure on the remaining high quality reads, withthe same method described by Li, et al.2010. All the reads were then checked using FASTQC (https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/fastqc.html)

Scaffold construction

The high quality reads after the above filtering and correction steps were exploited to generate scaffolds using SOAPdenovo2 software (version 2.04 41,https://github.com/aquaskyline/SOAPdenovo2). All the high quality reads were loaded into computer memory and de Bruijn graph data structure were used to represent the overlap among the reads. The graph were then be simplified by removing erroneous connections and solving tiny repeats by read path. Then all the high quality reads were realigned onto the contigs and aligned PEs (Pair-End sequences) were obtained. We calculated the PE relationships between each pair of contigs and then construct the scaffolds step by step; from short insert size tolong insert size PEs. Finally, the gaps in scaffolds were filled, which are most likely be caused by repeats, using the high quality PE reads. ABySS 2 (https://bio.tools/ABySS), MaSuRCA (www.genome.umd.edu/masurca.html) were also be exploited to assemble river genome scaffolds.

Assembly visualization

A number of web based open source software are available for WGS analyses. Bandage (https://rrwick.github.io/Bandage/), Cortex (https://bio.tools/cortex) were used for visualizing de novo assembly graphs. QUAST (quast.sourceforge.net) were also be used for genome assemblies evaluation and comparison.

Validation of assembly and Gap resolution

Once the assembly were achieved, further verification were required. Sequence gap locations were determined. In order to identify identical gap regions on assemblies, we will use simple alignment tools like BLAST. If fragments aligned to two separate scaffolds or chromosomes, then the region were considered a trans-scaffold break. If one or both fragments surrounding a gap do not align, or if there are two or more ambiguous bases between aligned fragments, the gap is to be considered open. Gaps were confirmed by checking Illumina WGS read alignments from the sequenced animal to the gap region using SAMtools depth version 1.3. If one or more bases in the filled region have a read depth <5, the gap were considered unresolved. GMcloser (https://omictools.com/gmcloser-tool) and GapFiller (https://omictools.com/gapfiller-tool) were also used.

Final evaluation of the assembly

To further evaluate the assembly, base qualities of the generated contigs and also the N50 metric were utilized. Base qualities, as reported by BWA, are based on the PHRED quality scale (Ewing and Green 1998), a de facto standard for genome assemblies. The N50 metric is the size of the smallest contig that has to be considered to cover at least50% of a sequence (in our case, each chromosome). High values of N50 indicate better assemblies. MUMmer v.4 (https://bio.tools/mummer) can also be used.

Gene prediction and functional annotation:

The first step in genome annotation in a given the genomic sequence is gene structure prediction. Gene prediction was conducted using MAKER ver. 3.01.03 which defines probability distributions for the different sections of the genomic sequence. Gene prediction was performed ab initio with using both given and default parameters. Functional annotation was obtained by InterProScan ver. 5.46-81.0. The functional annotation report has been stored at Figshare database.

Annotation of protein-coding genes

To predict protein-coding genes, information were integrated from different methods, specifically, ab initio prediction, homology-based annotation and. For ab initio prediction, GENSCAN (http://genes.mit.edu/GENSCANinfo.html, version 1.0) and GlimmerHMM

(http://www.cbcb.umd.edu/software/glimmerhmm, version 3.02,) software were used to predict tiger genes with parameters trained with the PanTig1.0 genome.. GeneWise (unreleased, compiled on Jul-25-2007) software were used to predict gene structure contained in each protein region. Pseudogenes were filtered out of the homology-based results. ORF Finder(http://www.geneinfinity.org/sms/sms_orffinder.html) is also used for gene prediction. BreakDancer (gmt.genome.wustl.edu/packages/breakdancer), VarScan2 (https://omictools.com/varscan-tool) were used to detect structural variations.

A high-confidence homology-based gene set were also be constructed based on the syntenic relationship between tiger and lion/human/cattle sequence using LASTZ (version1.01.50, built 20090316) software. Given that real orthologs of the reference protein are likely within a syntenic region, genes which are not in these regions were filtered. Homologous genes which are shorter than the reference genes by two or more exons will also be filtered out. The homology gene set derived from sheep reference proteins and cattle reference proteins were then be merged. For each gene locus, the record with the longest coding regions and/or highest genewise score were retained. The GLEAN (http://sourceforge.net/projects/glean-gene) software were used to integrate data derived from different methods into a GLEAN-derived gene set. Short genes (CDSlength < 150 bp) and low-quality genes (gaps in more than 10 percent of the coding region) were filtered. Finally all protein-coding gene models were annotated.

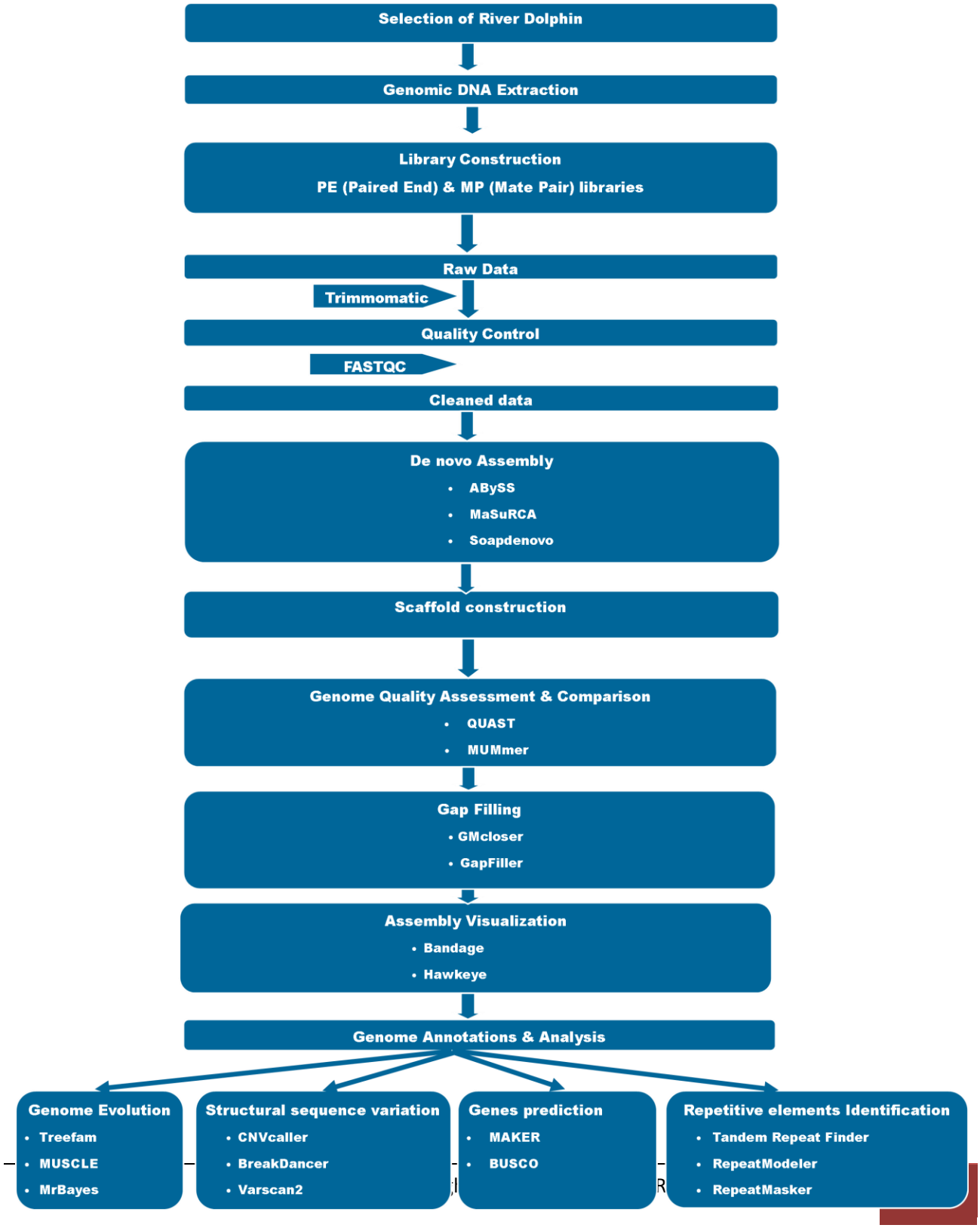
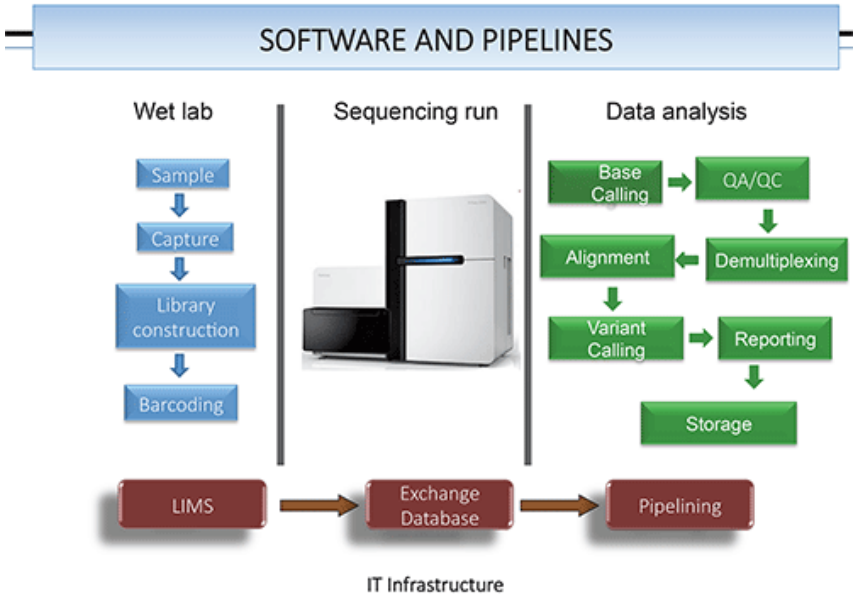

Selection of River Dolphin
Genomic DNA Extraction
Library Construction
PE (Paired End) & MP (Mate Pair) libraries
Raw Data
Trimmomatic
Quality Control
FASTQC
Cleaned data
De novo Assembly
ABySS
MaSuRCA
Soapdenovo
Scaffold construction
Genome Quality Assessment & Comparison
QUAST
MUMmer
Gap Filling
GMcloser
GapFiller
Assembly Visualization
Bandage
Hawkeye
Genome Annotations & Analysis
Genome Evolution
Treefam
MUSCLE
MrBayes
Structural sequence variation
CNVcaller
BreakDancer
Varscan2
Genes prediction
MAKER
BUSCO
Repetitive elements Identification
Tandem Repeat Finder
RepeatModeler
RepeatMasker

Fig: Steps (Pipeline) followed during this WGS Bioinformatics analyses

Computational infrastructures and data formats

Dedicated Server used for Whole genome sequencing project

Dedicated server Configuration	
Processor	AMD Ryzen Threadripper 1950X 16-core/32-thread
Motherboard	ASUS Prime X399-A . TR-4
RAM	(16*8) 128 GB DDR-4
HDD	10 TB SATA HDD
CPU Cooler	ThermalTake Liquid CPU Water Cooler
Power Supply	Tough Power Guard 1050W Fully Modular Ring Full Rang



Research objective 1

Draft genome sequencing and assembly of *Catla catla* [Hamilton, 1822] from the Halda river of Bangladesh

Abstract

Catla catla (Catla) is a South Asian popular carp belonging to the Cyprinidae family. It is inherent to rivers and lakes in India, Bangladesh, Myanmar, Nepal and Pakistan but has also been proposed elsewhere in South Asia for cultivation. This surface dweller fish has a low demand for feed and has great socio-economic significance in Bangladesh. Here, we report a scaffold-level reference genome of *C. catla* generated by employing the Illumina Hiseq technology. A healthy male adult *C. catla* belonging to river Halda of Chittagong, Bangladesh was captured and used for reference based assembly. The whole genome sequences were assembled in 5,345 contigs with a total length of 1,233,067,729 bp and a contig N50 length of 723,178 bp. Genome annotations identified 24,571 gene models using MAKER gene annotation tool. The Benchmarking Universal Single-Copy Orthologs (BUSCO) tools assessed 97.6% completeness of the assembled genome. The assembly provides an important resource for further comparative genomics analyses to explore the unique biology of *C. catla* fish variants.

Introduction

C. catla (Hamilton, 1822) is a fast-growing freshwater carp fish species under the family Cyprinidae, a very substantial group of fish primarily originating from Europe and Asia. The wide range of exciting morphological functions of *C. catla* offers an interesting possibility to hyperlink genotype to phenotype and to apprehend the dynamics of genome evolution and speciation. It is the sixth most important finfish aquaculture species, with approximately 2.8×10^6 tons produced globally in 2015 [1]. *C. catla* is considered one of the "Four famous Indian carps" of Halda River which has extensive demand for its enriched nutritive properties, high productivity rate, delicate flavour and helpful habit of increasing water quality [2]. The Halda river is considered as one of the top rivers in the world, which can sustain the spawning process of the four major carps of the subcontinent naturally. It is geographically situated in South-East region of Bangladesh which is a major branch of the River Karnaphuli, known as fish mine of Bangladesh, in Chittagong district. The origin of Halda is at the hill ranges of Patachara, Ramgarh near in the Khagrachari hill specifically speaking from the hilly fountain called Haldachora [3,4].

As a principal species, *C. catla* alone contributed about 3,650,000 ton to the global aquaculture production in 2010 [5] and 6.40% of annual fish production in inland water

bodies was reported from solely *C. catla* in 2017-2018 FY in Bangladesh. But to meet the developing seed demand, hatchery operators did not consider the genetic quality issue which has led to inbreeding a common scenario in Bangladeshi hatcheries [6,7]. Thus *C. catla* may be susceptible to drop of genetic diversity and variability in the wild populations because of inbreeding depression [8]. Hence, it is very urgent to annotate genetic variability and unique traits of river populations of *C. catla* to take initiatives for improvement of its reduced genetic properties that have great importance for the sustainable aquaculture practices. Already to some extent, the gene pools of our indigenous varieties of carps viz: *Labeo rohita* (Rohu), *C. Catla* (Catla) & *Cirrhinus cirrhosis* (Mrigal) have been contaminated [9]. As a result, soon, it's feared that pure seeds of these carps, endemic to this region will gradually disappear from the culture system. To avoid such genetic contaminations of the seed production, the genetic variability within a population is extremely useful to gather the information on individual identity, breeding pattern, degree of relatedness and distribution of genetic variation among them along with evolutionary and adaptive behaviour. The present investigation was intended at developing the whole genome sequences of Halda riverine wild populations of *C. catla*, to explore fish genomics by conducting transgenesis, mutation or chromosomal manipulation. These will eventually facilitate conservation of genetic materials, crossbreeding along with avoiding intergeneric hybridization among the wild species and to encourage pure gene strain of this indigenous species for Bangladesh.

Methodology of the Study:

Sampling and DNA isolation:

A fresh sample of adult male *C. catla* was collected from river Halda (Longitude/Latitude-22°28'56.09"N & 91°54'07.62"E) and conveyed to the laboratory to preserve at -70° C for further analysis during September 2019. The muscle tissue was dissected and stored with 95% ethanol, later a high molecular weight genomic DNA was extracted using the AddPrep Genomic DNA extraction kit (AddPrep Inc., Korea). Samples were further tested for qualitative and quantitative evaluation of the DNA through Fluorumetry. All the methods had been performed following the "Regulations for Animal Experiments in Chittagong Veterinary and Animal Sciences University" as required.

Library preparation

The extracted DNA was cleaned and sent for both library preparation and whole-genome sequencing (WGS) at the BGI genomics, China. Using Next-generation sequencing (NGS) technology on an IlluminaNovaSeq 6000 platform [10] a total of 45 Gigabase pair (Gb) of subread bases with a read length of 150 bp were induced. After sequencing quality of

primary sequence reads and trimmed sequencing reads were investigated using FastQC version 0.11.8 [11]. The quality control of the reads was done including removing adaptor sequences, low-quality reads and contamination from raw reads using BFC [12]. A total of 55,264,452,615 clean reads were included in the assembly with 48X coverage.

Genome assembly:

To assemble the *C. catla* genome we used ABySS ver. 1.9.0 [13] and Platanus ver.1.2.4 [14] assembler. Since there is currently no de novo assembler assured to outperform others and as assemblers overall performance can differ relying on the dataset, two unique assemblers were used and to determine the best assembler an assembly evolution was subsequently performed. Both assemblers follow the classic De Bruijn graph illustration even though the assembly algorithm differs across methods. Finally, BUSCO v.4.1.2 [15] was operated to assess the quality of the assembly in respect of gene completeness.

Gene prediction and functional annotation:

The first step in genome annotation in a given the genomic sequence is gene structure prediction. Gene prediction was conducted using MAKER ver. 3.01.03 [16] which defines probability distributions for the different sections of the genomic sequence. Gene prediction was performed ab initio with using both given and default parameters. Functional annotation was obtained by InterProScan ver. 5.46-81.0 [17]. The functional annotation report has been stored at Figshare database [18].

Result and discussion:

A total of 1.2 Gbp reads were generated and subsequently, both assemblers generated 5344 scaffolds along with 702,160 contigs in Platanus and 823,556 contigs in Abyss. Standard assembly metrics are detailed in Table: 2.Between the assemblers, Platanus surpassed ABySS as it generated the fewest but longest sequence with an N50 scaffold measure. BUSCO analysis suggested that Platanus is the best assembler in terms of genome completeness. The BUSCO analysis on Platanus assembly revealed 97.6% completeness, as well as a significantly lower number of scaffolds and considerably better N50 indicates the genome to be of high-quality. In the NCBI GeneBank under the Accession numbers JACDQS000000000 the genome sequence data has been deposited.

Table: 1. Overview of data files/data sets

Label	Name of data file/data set	File types [file extension]	Data repository and identifier [DOI or accession number]
Data file 1	Site of sample collection	.docx	https://doi.org/10.6084/m9.figshare.12980375
Data file 2	Whole genome assembly data	FASTA	NCBI GeneBank Assembly [Accession number: GCA_014525385.1] [https://www.ncbi.nlm.nih.gov/assembly/GCA_014525385.1]
Data file 3	Whole genome sequence	FASTA	NCBI GeneBank [Accession numbers :JACDQS010000001-JACDQS010005344] [https://www.ncbi.nlm.nih.gov/Traces/wgs/JACDQS01?display=contigs]
Data file 4	Annotation files	.tsv	https://doi.org/10.6084/m9.figshare.12948266

Table: 2. Genome assembly parameters

Total sequence length	**1,232,079,142**
Total Ungapped length	**1,130,442,603**
Gaps between scaffolds	**0**
Number of Scaffolds	**5,344**
Scaffold N50	**723,740**
Scaffold L50	**412**
No. of contigs	**702,160**
Contigs N50	**4,548**
Contigs L50	**62,790**
Total number of chromosomes and plasmids	**-**
Number of component sequences [WGS or clone]	**5,344**

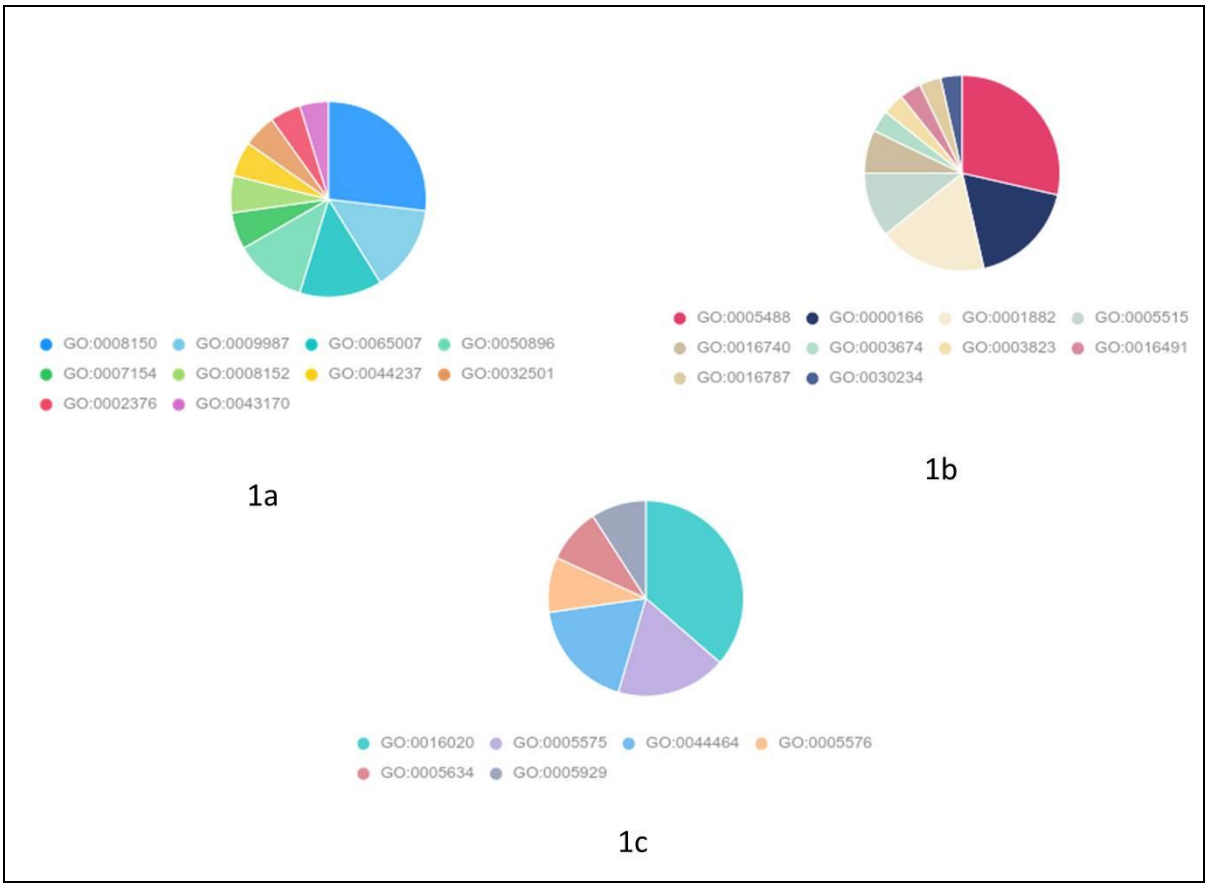


Figure: 1(a-c). Gene Ontology (GO) functional annotations of predicted genes represented by three categories- 443 genes were assigned with GO terms in Biological process (a), 27 genes in Molecular Function (b) and 12 genes in Cellular component (c) Each category is sub-categorized in different GO terms, represented on percentage in the pie charts.

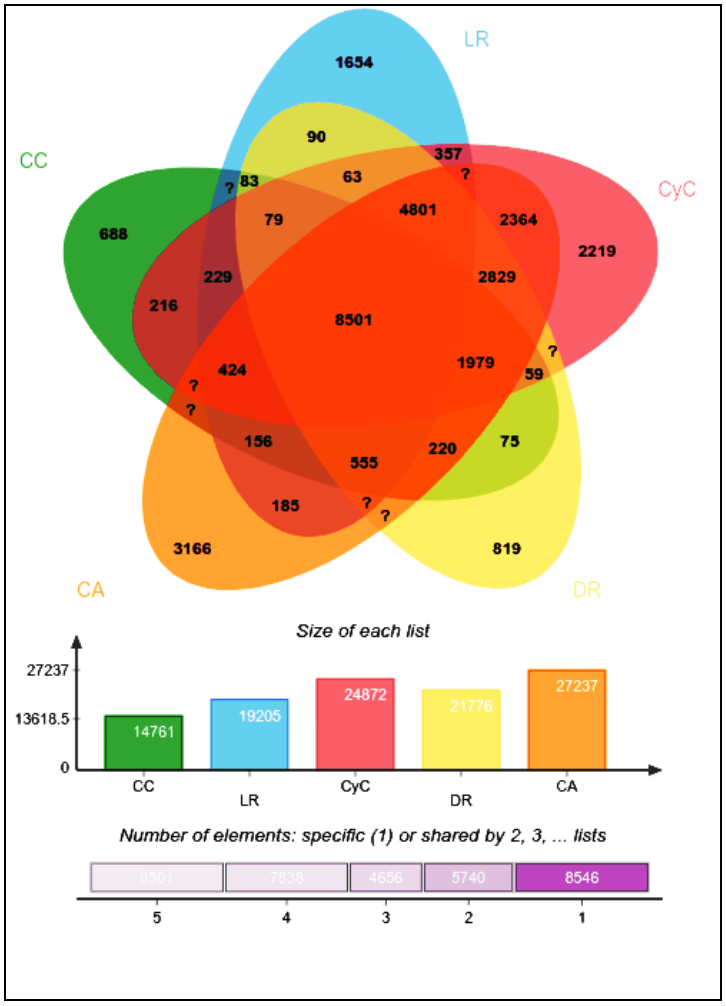


Figure: 2. Comparative genomics of *C. catla*: Venn diagram showing orthologous gene clusters among five diploid cyprinid species, *C. catla* (CC), *C. auratus* (CA), *C. carpio* (CyC), *L. rohita* (LR) and *D. rerio* (DR).

MAKER ver. 3.01.03 [16] pipeline was used for structural annotation. GC content of the genome was determined to be 37.03%. RepeatMasker ver. 4.1.0 [19] and Repeatmodeler using the latest version of the Repbase database [20] identified 24.16% repeat contents. Overall, 24,571 gene models were predicted using the MAKER gene annotation pipeline based on both de novo and reference-based predictions using

genes and proteins from other fish species (Goldfish, Common Carp, Zebrafish). 22,616 genes were located in a total of 5,344 scaffolds with an average of 4.23 genes per scaffold. Out of the 24,571 genes, 12,429 were identified as GO terms using InterProScan ver. 5.46-81.0 [17]. Based on functional annotation, 482 GO terms were found (Figure: 1) to be associated with *C. catla* genome of which highest number of gene (443) were under Biological process, followed by 27 genes of Molecular functions and 12 genes of Cellular components. Among these, the highest number of genes was under Biological process, 69 genes (GO: 0008150).

To elucidate orthologous relationships for the *C. catla* (CC) annotated genes, we compared them employing OrthoVenn [21] with four other diploid cyprinid species, *Carassius auratus* (CA), *Cyprinus carpio* (CyC), *Labeo rohita* (LR) and Danio *rerio* (DR). Orthologous genes shared among these species were delineated through a Venn diagram. The orthologous gene family analysis in diploid cyprinids resulted in total 32,111 clusters (*C. Catla*, 12,576; *C. auratus*, 22,014; *C. carpio*, 21,901; *L. rohita* , 15,523;and *D. rerio,* 19,244 orthologous clusters and 8,546 single-copy gene clusters) (Fig. 2). An add up of 8,501 orthologs were shared by all five species, along with 688 species-specific gene clusters of *C. catla*.

The genome sequencing and assembly of the *C. catla* provides a valuable tool for future population genetics and fish genomics studies, which will allow for targeting specific genes and particularly interesting regions of the *C. catla* genome. For instance, the species could be a model organism in which to study inbreeding and aquaculture of major carps as it is the highest common IMC species. Additionally, the species seem to be greatly demonstrated in which to study natural breeding adaptation and potential genomic specializations that allow for adaptation to pollution in otophysine fish species, by comparing the *C. catla* genome with that of closely-related species like the *Labeo rohita* (Rohu) with a more south-Asian geographic distribution. As shown in the recent special issue of Science, sequencing new fish genomes can lead to a better understanding of crucial aspects of the biology and ethology of fish. In the case of the *C. catla* fish, it could gather new information on the genetic diversity of the species and infer the compelling population size of each of the putative populations.

Finally, the availability of the *C. catla* genome can offer assistance to gather the statistic history of the species, i.e. how high and low have the population sizes been within the past. Besides the estimation of genetic variability, this is typically a critical point because species that have experienced low population sizes in the past might be more vulnerable to human threats and more inclined to extinction.

Data availability

The Illumina raw reads have been deposited in the SRA [Project ID: PRJNA623322] under the Accession numbers SRR12102514. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession SRX8626849 and the assembled genome at GCA_014525385.1. The version described in this paper is version JACDQS010000000.

Reference

1. FAO. Nuestro futuro: Un mundo sin hambre. FAO Publ [Internet]. 2015;83. Available from: www.fao.org/publications
2. M. Shafi and MAAQ. Bangladesher Matshya Sampad [in Bengali]. Bangla Acad Dhaka. 1982;314–9.
3. Akter A, Ali MH. Débits environnementaux requis: évaluation pour la rivière Halda, Bangladesh. Hydrol Sci J. 2012;57[2]:326–43.
4. Kabir M, Kibria M, Hossain M. Indirect and Non-use Values of Halda River- A Unique Natural Breeding Ground of Indian Carps in Bangladesh. J Environ Sci Nat Resour. 2015;6[2]:31–6.
5. FAO. The state of world fisheries and aquaculture. 2010;[disponível em http://www.fao.org/docrep/013/i1820e/i1820e.pdf]:218.
6. Zakiur Rahman SM, Khan MMR, Islam S, Alam S. Genetic variation of wild and hatchery populations of the catla Indian major carp [Catla catla Hamilton 1822: Cypriniformes, Cyprinidae] revealed by RAPD markers. Genet Mol Biol. 2009;32[1]:197–201.
7. Simonsen V, Hansen MM, Mensberg KLD, Sarder RI, Alam S. Widespread hybridization among species of Indian major carps in hatcheries, but not in the wild. J Fish Biol. 2005;67[3]:794–808.
8. Ali MR, Rahi ML, Islam SS, Shah MS, Shams FI. Genetic Variability Assay of Different Strains of Catla catla. Int J Life Sci. 2015;9[1]:37–42.
9. Khatun N, Islam MT, Sultana N, Mrong S, Huq MA. Present status of carp hatchery and breeding operations in Bangladesh: A review. Res Agric Livest Fish. 2017;4[2]:123–9.
10. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. 2010;5[6].
11. S. A. . FastQC: a quality control tool for high throughput sequence data.
12. Li H. BFC: correcting Illumina sequencing errors. Available from: http://bit.ly/biobin
13. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, et al. ABySS 2 . 0 : Resource-Efficient Assembly of Large Genomes using a Bloom Filter Effect of Bloom Filter False Positive Rate. Genome Res. 2017;27:768–77.
14. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 2014;24[8]:1384–95.
15. Simã FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Available from: https://academic.oup.com/bioinformatics/article/31/19/3210/211866
16. Campbell MS, Holt C, Moore B, Yandell M. Genome Annotation and Curation Using MAKER and MAKER-P. Available from: www.yandell-lab.org.
17. Jones P, Binns D, Chang H-Y, Fraser M, Li W, Mcanulla C, et al. InterProScan 5: genome-scale protein function classification. 2014;30[9]:1236–40. catla.out.all.maker.proteins.fasta.tsv [Internet]. [cited 2020 Oct 2]. Available from: https://figshare.com/articles/dataset/catla_out_all_maker_proteins_fasta_tsv/12948266

20. Jurka J, Kapitonov V V., Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110[1–4]:462–7.
21. Wang Y, Coleman-Derr D, Chen G, Gu YQ. OrthoVenn: A web server for genome wide comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res. 2015;43[W1]:W78–84.

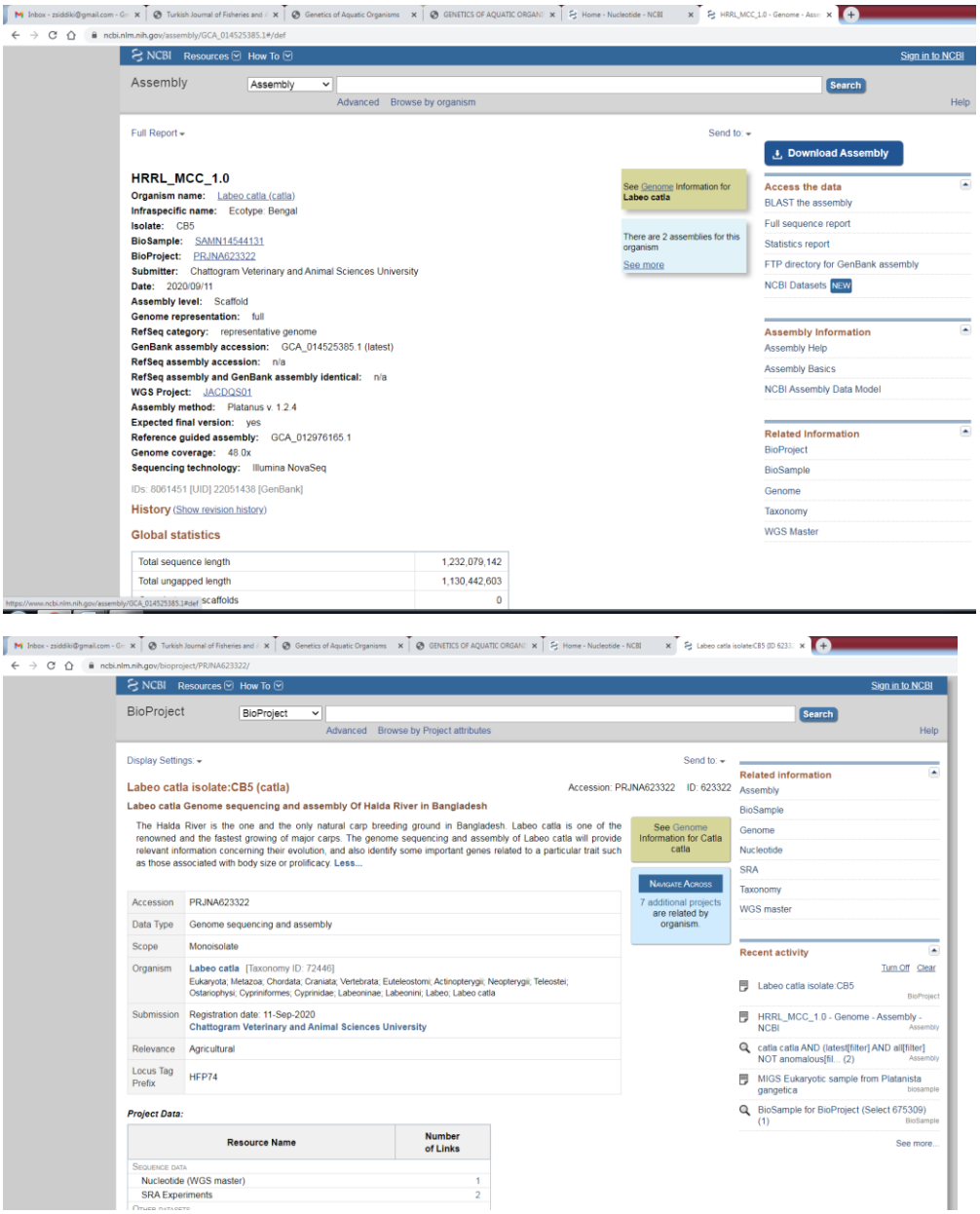


Fig. Screenshot of NCBI SRA datasets of Catla genome

Research objective 2

Complete mitochondrial genome sequence of ***Gibelion catla*** (Hamilton, 1822) from the Halda river of Bangladesh

ABSTRACT

Catla (***Gibelion catla***) is one of the fastest-growing major carp found in South Asia as well as Bangladesh. ***Gibelion catla*** is abundant inBangladesh and one of the most important aquacultured freshwater species with economic impact. In this study, we disclosed the complete mitochondrial genome sequence of Bangladeshi *Catla* fish from Halda river located in Chittagong. The circular mitogenome of *Gibelion catla* is 16,597 bp in length and nucleotide composition is AT-based (72%), contains 37 genes including 13 protein-coding genes, 22 tRNA genes, 2 rRNA genes and a D-loop control region.

INTRODUCTION

Gibelion catla is a member of the Cyprinidae family, which is endemic to the perennial river network of northern India, the Indus plain and adjacent hills of Pakistan, Bangladesh, Nepal and Myanmar (Reddy, 1999). It has become one of the most well-established fish populations of all the rivers, lakes and reservoirs where they have been introduced. The Halda River is located in South-East region of Bangladesh which is a major tributary of the river Karnaphuli in Chittagong district originated from the hilly Haldachora fountain at the Patachara hill ranges of Ramgarh in the Khagrachari hill and renowned for being the only pure Indian carp breeding ground in Bangladesh where Catla is very common (Akter and Ali, 2012; Kabir et al., 2015; Tsai et al., 1981). A major portionof the country's pond carp culture is dependent on these wild seed that has an important and potential contribution in the agro-based economic development, poverty alleviation, employment and supplying of animal protein and earning the foreign currency for the national sector (Azadi, 1979, DoF, 2005). *G. Catla* is oneof the "Four famous Indian carp" of Halda river which has extensive demand for it's higher nutrition content, productivity rate, delicate flavour and helpful habit of increasing water quality (Shafi and Quddus, 1982). For being small in size, high evolutionary rate, and maternal inheritance mood, the complete mitochondrial genome sequences provide insight into the assessment of wide variation in animals and the comparison of sequence data contribute to the exploration of improved markers for population ecological studies (Avise, 1995; Zhou et al., 2009). Here we reported the entire mtDNA sequences of *Gibelion catla* from the Halda river.

The specimen was collected from Halda river, Chattogram (geographic coordinate: 22°33'34.7" N 91°50'41.8"E). Fresh tissue (from muscle) sample was stored at −20 °C until used to isolate genomic DNA using commercial DNA extraction kit (AddBio, Korea)

and the total DNA was stored with a voucher number (DPP/CVASU/2019-12-44). Purified DNA was sent for library preparation and sequencing through commercial suppliers. DNA was sequenced using Illumina NovaSeq 6000 platform from BGI, China. The mitochondrial genome reads were separated from the whole genome sequence by mapping it against the reference Catla mitochondrial datasets using SAMTOOLS. The organelle assembler NOVOPlasty V.2.7.2 (Dierckxsens, Mardulyn & Smits, 2017) was used to assemble the clean reads. Web based tools like MITOS (Bernt et al., 2013) and GeSeq (Tillich et al., 2017) were applied to perform structural and functional annotation. Another tool, OGDRAW was used to construct the circular representation of the entire mitogenome (Greiner et al., 2019). Finally, mtDNA sequences were aligned and a phylogenetic tree was constructed by using MEGA X (Kumar et al., 2018)

The complete mitogenome of *Gibelion catla* (NCBI accession number **MT303069**) is 16,597 bp in length and consists of 13 protein-coding genes, two ribosomal RNA genes (rRNA), 22 transfer RNA (tRNA) genes, and a putative control region (D-loop). The structural organization and location of the different features of these mito genomes were consistent with the common vertebrate mtgenome model (Liu & Cui, 2009). The relative order of nucleotide composition corresponds to the nucleotide pattern of other fish mitogenome A>C>T>G (Wang et al., 2008). The mitochondrial genome of *Gibelion catla* contains an A + T bias with an overall nucleotide composition of A = 5383 (32.43%), T =4087(24.62%), C=4580 (27.60%), and G =2547(15.35%). The GC content of the mitogenome is 42.94 %. Furthermore, the AT-skew is positive which is 0.13 and GC-skew is observed negative which is -0.28.

Most of the protein-coding genes (PCGs) have been encoded on the H-strand of mtDNA. Only one PCG (*nad6*) and 8 transfer RNA genes (*trnA, trnC, trnE, trnN, trnP, trnS2, trnY*) were encoded in the L-strand of mtDNA. Most of the PCG starts with a standard ATG start codon, whereby *nad2, nad1, nad5,* starts with ATA and *Cox2* starts with AAT. The length of the 12S rRNA and 16S rRNA genes were 952 bp and 1685 bp respectively. The tRNA genes encoded in the genome ranged from 60 to 75 bp. The control region is between *trnaP* and *tnaF* and has a size of 930 bp. The phylogenetic relationship were estimated using Neighbor Joining method implemented in CLC main workbench (Fig. 1). Two different closely related species, *Gibelion catla* and *Labeo rohita* were placed in the sister clade. All other *Labeo* were also placed in different sister clades.

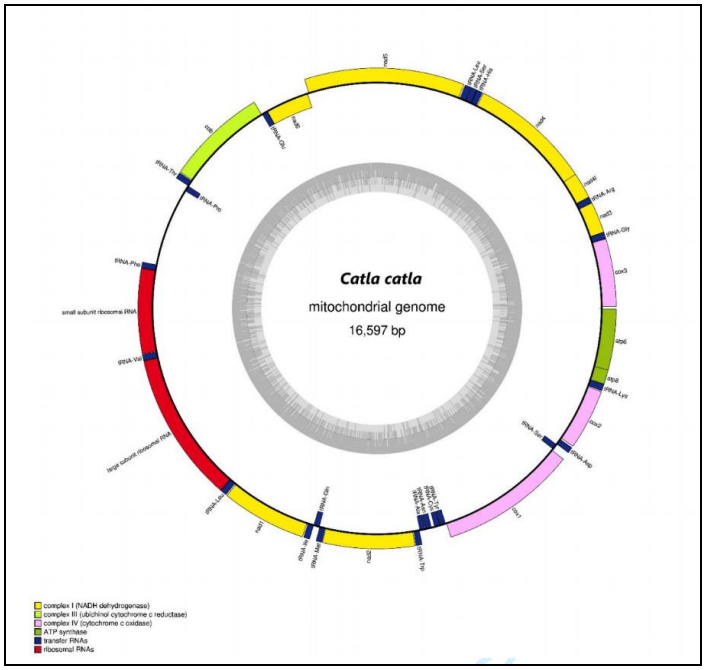


Figure 4. Circular mitogenome representation of Catla catla. The map is annotated and shows 13 for protein coding genes (PCGs), 2 genes for ribosomal RNA (rrrnS [12S ribosomal RNA] and rrnL [16S ribosomal RNA]), 22 genes for RNA transfer (tRNA) and the putative control region. The inner circle represents the content of GC alongside the genome. There is no annotated putative D-Loop / Control region

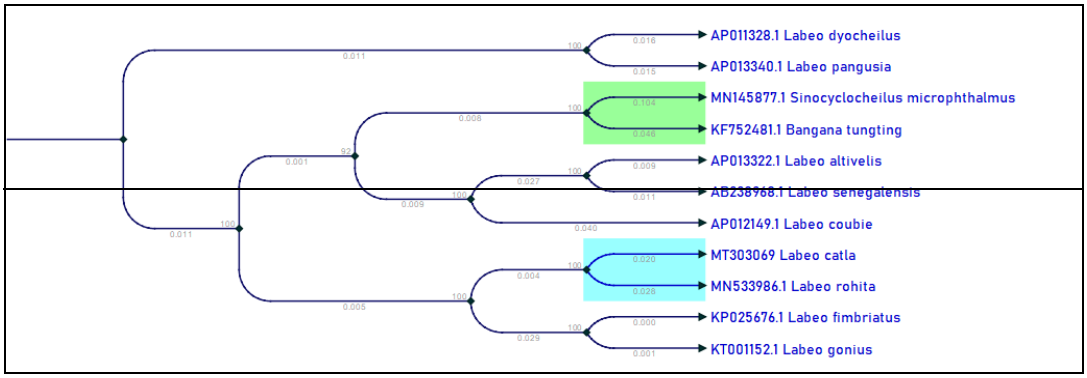


Figure 1.The Neighbor Joiningtree of Gibelion catla and 9 Labeo based on complete mitochondrial genome. Numbers above the branches indicate the bootstrap support values, and values lower than 50 are not shown.

To sum up, this study provides the information of *Gibelion catla* mitogenome collected from the Halda river of Bangladesh. The data will be useful for future research by the fish geneticists and evolutionary biologists. This study will also provide crucial information for further taxonomic and phylogenetic analyses among closely related species and implementation of the effective conservation strategy of this unique resource of Bangladesh.

References

Akter, A., & Ali, M.H. (2012). Environmental flow requirements assessment in the Halda River, Bangladesh. Hydrological Sciences Journal, 57(2), 326–343.

Avise, J. C. (1995). Mitochondrial DNA polymorphism and a connection between genetics and demography of relevance to conservation. *Conservation Biology, 9*(3), 686-690.

Azadi, M. A. 1979. Studies on the limnology of theRiver Halda with special reference to thespawning of major carps. p232.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., . . . Stadler, P. F. (2013). MITOS: improved de novo metazoan mitochondrial genome annotation. *Molecular phylogenetics and evolution, 69*(2), 313-319.

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic acids research, 45*(4), e18-e18.

DoF, 2005. Fishery Statistical Yearbook of Bangladesh 2003–2004. Fisheries Resources Survey System, Department of Fisheries, Ministry of Fisheries and Livestock, Matshya Bhaban, Dhaka, Government of Bangladesh publication. p46.

Greiner, S., Lehwark, P., & Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic acids research, 47*(W1), W59-W64.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution, 35*(6), 1547-1549.

Liu, Y., & Cui, Z. (2009). The complete mitochondrial genome sequence of the cutlassfish Trichiurus japonicus (Perciformes: Trichiuridae): Genome characterization and phylogenetic considerations. *Marine genomics, 2*(2), 133-142.

Reddy, P. (1999). Genetic resources of Indian major carps. *FAO fisheries technical paper*(387).

M. Shafi and M. A. A. Quddus (1982). “Bangladesher Matshya Sampad (in Bengali),”Bangla Academy of Dhaka, Bangladesh, pp314-319.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq–versatile and accurate annotation of organelle genomes. *Nucleic acids research, 45*(W1), W6-W11.

Tsai, Chu-fa, Islam, M. N., Karim, R., & Rahman, K. S. (1981). Spawning of major carps in the lower Halda River, Bangladesh. *Estuaries, 4*(2), 127-138.

Wang, C., Chen, Q., Lu, G., Xu, J., Yang, Q., & Li, S. (2008). Complete mitochondrial genome of the grass carp (Ctenopharyngodon idella, Teleostei): insight into its phylogenic position within Cyprinidae. *Gene, 424*(1-2), 96-101.

Zhou, Y., Zhang, J.-Y., Zheng, R.-Q., Yu, B.-G., & Yang, G. (2009). Complete nucleotide sequence and gene organization of the mitochondrial genome of Paa spinosa (Anura: Ranoidae). *Gene, 447*(2), 86-96.

Research Objective 3

Whole Genome Sequence of *Labeo rohita* (Hamilton, 1822) from Halda river of Bangladesh

Introduction:

Labeo rohita commonly known as 'Rohu' belongs to the family Cyprinidae. It's good taste and high market price makes it one of the favorite freshwater aquaculture fish in Bangladesh. In Bangladesh among the three Indian major carp species, (*Labeo rohita*, *Catla catla* and *Cirrhinus cirrhosus*) Rohu ensures the top position of consumer preference due to its good test, quality, reasonable price and availability and also in the national economy, the contribution of Rohu is beyond question. *L. rohita* is widely distributed in the country. The natural populations show a declining trend due to habitat loss. The fish is extensively cultured in the country and is used in floodplain programmes. The threats to the fish are in general. Hence, the fish is assessed as Least Concern (IUCN., 2015).

Morphological features of *Labeo rohita*:

The back of the Rohu's body is brownish and on the sides and beneath its silvery. Fins are greyish or dark and dusky pectoral fins are observed. Both lips are covered by cartilaginous covering. One pair of short maxillary barbells is present. Dorsal and abdominal profiles are convex and the caudal peduncle is short. The complete lateral line is observed. Fin formula: D. 3/12-13, P1.16-17, P2. 9. A. ⅖ (Rahman, 2005). Labeo is column feeder at mid-water and prefers to feed on plant matters including decaying vegetation (Jhingram and Pullin, 1985). Its food comprises crustaceous and insect larvae in the early stages (Mookerjee et al., 1946).

Distribution:

Rohu is a natural inhabitant of the freshwater section of Bangladesh, India, Pakistan, Burma and Terai region of Nepal and also has been transplanted into Sri Lanka, Japan, Philippines, Malaysia and some countries of Africa (Jhingram and Pullin, 1985). In Bangladesh, Rohu is mostly found in the Padma-Brahmaputra, their tributaries and Halda river system (Alam et al., 2009), though Halda is the only river in Bangladesh from where fertilized eggs of major carps are collected (Tsai et al., 1981; Patra and Azadi, 1985).

Importance of the genomic study:

Halda river is considered to be the natural breeding ground for Rohu and other major carps and the only river from where fertilized eggs of Rohu are collected. Quality of

eggs, fry and fingerling collected from this natural source are far greater than those of induced breeding as induced breeding can lead to genetic degradation. To meet up the protein requirement of our increasing population aquaculture is expanding in our country as a result seed demand is increasing in hatcheries. Poor management of broodstock and lack of scientific approach in those fish farms can lead to gene loss which is a common scenario in Bangladesh (Simonsen, et al., 2005; Rahman, et al., 2009). The gene pools of our indigenous varieties of carps viz: Rohu, Catla & Mrigal have been contaminated (Khatun et al., 2017). On the contrary, the environmental condition of the Halda river is degrading due to anthropogenic activities, as a result, the natural population of major carps is decreasing gradually. So also in natural habitat, this species is at risk. Rohu contributed 10.50% of the total annual fish production of inland water bodies, which is far greater than the other three major carps of Bangladesh (DoF, 2018). But still, we lack sufficient genome label study records for this economically important fish of Halda River in Bangladesh. As we already stated that Halda river is the only river in Bangladesh from where fertilized eggs of major carps are collected so it is very urgent to annotate genetic variability and unique traits of river populations of *L. rohita.* Here we present the first draft genome of pure wild Rohu of river Halda to compliment the ongoing increasing aquaculture sector of our country through providing genetic resources, which in turn will allow researchers to carry out further studies of Rohu evolution and resistance to an extreme environment. The genomic information can also be used to avoid loss of genetic diversity, inbreeding depression in the hatchery population and a proper management program can be taken to conserve the pure wild stock of *L. rohita* of river Halda. This information will aid the evolutionary and biological study of Rohu.

The methodology of the study:

A healthy male *L. rohita* belonging to river Halda of Chittagong, Bangladesh was collected, later on, sequenced its de novo assembly. Illumina platform was employed for sequencing. Fresh tissue (from muscle, liver) samples were stored at −20 °C until used to isolate genomic DNA using a commercial DNA extraction kit (AddBio, Korea) and the total DNA was stored. Purified genomic DNA was sent for library preparation and whole-genome sequencing (WGS) at BGI Group (Shenzhen, Guangdong, China). A total of 60.2 Gb (Gigabase pair) of subread bases were generated using Next-generation sequencing (NGS) technology on an Illumina NovaSeq 6000 platform. After sequencing, quality of raw sequence reads and trimmed sequencing reads were inspected using FastQC version 0.11.8 (Andrews, 2010). Reads were quality controlled including removing adaptor sequences, contamination and low-quality read from raw reads using BFC (Li,H., 2015). For de novo assembly we used ABySS v. 1.9.0 (Shaun D Jackman et al., 2017) assembler to assemble the Rohu genome. This assembler follows the classic De Bruijin graph illustration. We have submitted our raw data in NCBI under the Biosample SAMN15846808 (SRR12474558).

Table1: Genome assembly parameters

Total sequence length	1.427 Gb
Number of scaffolds	13661
Scaffold N50	2006958
Scaffold L50	182
Contig Size	1.4 Gb
Number of contigs	1553666
Contig L50	13194
Contig N50	18563

Reference:

Alam, M.S., Jahan, M., Hossain, M.M. and Islam, M.S., 2009. Population genetic structure of three major river populations of rohu, *Labeo rohita* (Cyprinidae: Cypriniformes) using microsatellite DNA markers. *Genes & Genomics*, *31*(1), pp.43-51.

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data.

Department of Fisheries, 2018. Yearbook of fisheries statistics of Bangladesh 2017–18.

Jackman, S.D., Vandervalk, B.P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S.A., Jahesh, G., Khan, H., Coombe, L., Warren, R.L. and Birol, I., 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome research*, *27*(5), pp.768-777.

Jhingran, V.G. and Pullin, R.S., 1985. *A hatchery manual for the common, Chinese, and Indian major carps* (No. 252). WorldFish.

Khatun, N., Islam, M.T., Sultana, N., Mrong, S. and Huq, M.A., 2017. Present status of carp hatchery and breeding operations in Bangladesh: A review. *Research in Agriculture Livestock and Fisheries*, *4*(2), pp.123-129.

Li, H., 2015. BFC: correcting Illumina sequencing errors. *Bioinformatics*, *31*(17), pp.2885-2887.

Mookerjee. H.K. Gupta, S.M., Chaudhury, P.K.R. 1946. Food and its percentage composition of the common adult fishes of Bengal. Sci. Cult. 12: 247-249.

Naser, M. N., 2015. *Labeo rohita.* In: IUCN Bangladesh. Red List of Bangladesh Volume 5: Freshwater Fishes. IUCN, International Union for Conservation of Nature, Bangladesh Country Office, Dhaka, Bangladesh. pp. 190.

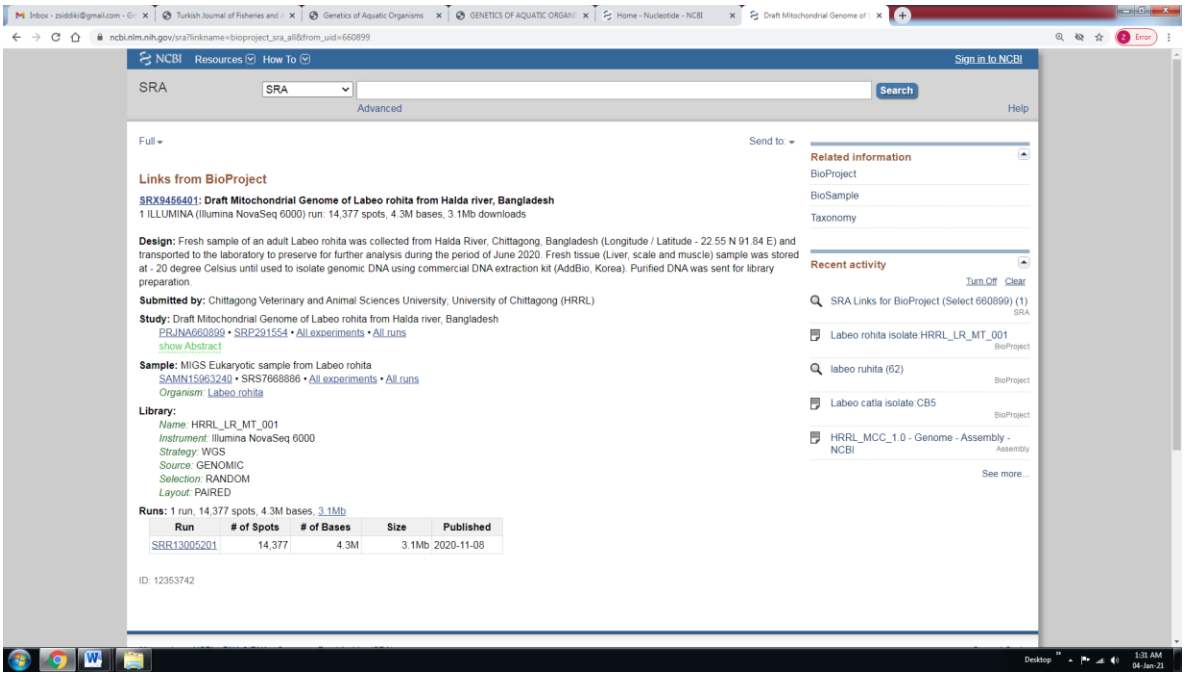
Patra, R.W. and Azadi, M.A., 1985. Hydrological conditions influencing the spawning of major carps in the Halda River, Chittagong, Bangladesh. *Bangladesh Journal of Zoology*, *13*(1), pp.63-72.

Rahman, A.A., 2005. *Freshwater fishes of Bangladesh*. Zoological society of Bangladesh.

Rahman, S.M., Khan, M.R., Islam, S. and Alam, S., 2009. Genetic variation of wild and hatchery populations of the catla Indian major carp (Catla catla Hamilton 1822: Cypriniformes, Cyprinidae) revealed by RAPD markers. *Genetics and Molecular Biology*, *32*(1), pp.197-201.

Simonsen V, Hansen MM, Mensberg KLD, Sarder RI and Alam S. 2005. Wide spread hybridization among species of Indian major carps in hatcheries, but not in the wild. Journal of Fish Biology, 67:794-808.

Tsai, C.F., Islam, M.N., Karim, R. and Rahman, K.S., 1981. Spawning of major carps in the lower Halda River, Bangladesh. *Estuaries*, *4*(2), pp.127-138.



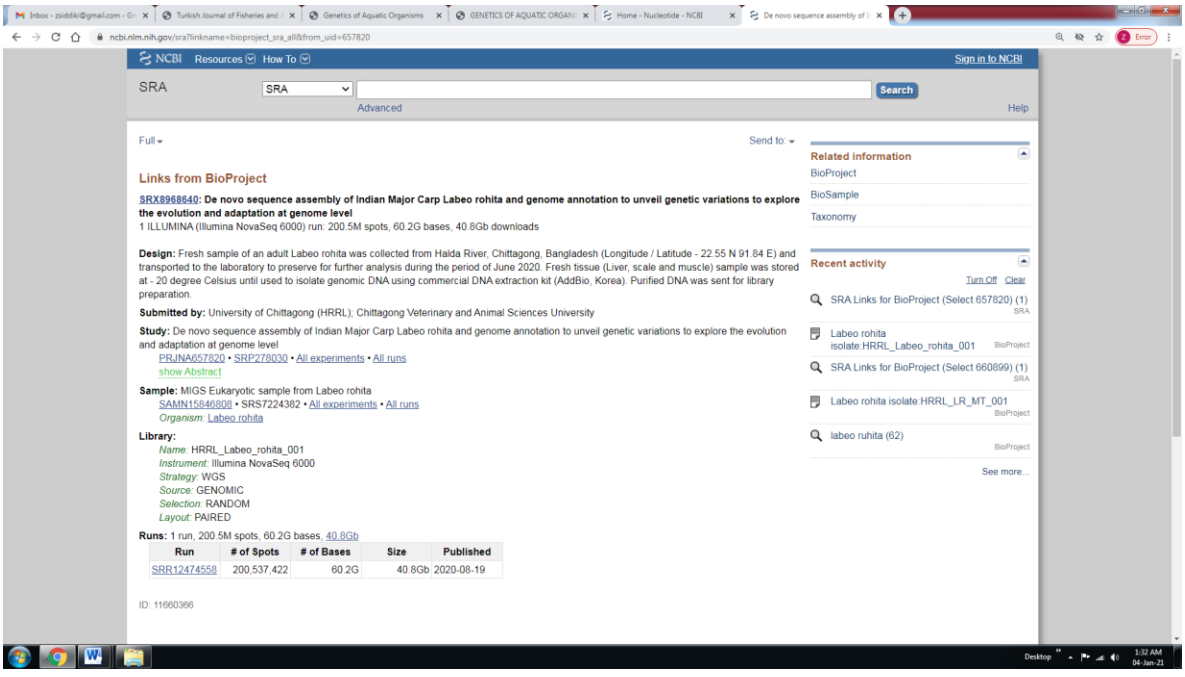


Fig. Screenshot of NCBI SRA datasets of *Labeo ruhita* genome

Research Objective 4

Complete Mitochondrial Genome Sequence of pure wild stock of *Labeo rohita* (Hamilton, 1822) from Halda river of Bangladesh

Abstract

Labeo rohita commonly known as Rohu, is an important Indian major carp that thrives as a prime species in aquaculture practice among the other four major carps available in Bangladesh. Halda river serves as a major reservoir of wild pure Rohu stock from where fry and fingerlings are supplied all over Bangladesh for aquaculture purposes. Due to anthropogenic activities, the ecological condition of this river is degrading, which negatively influences the breeding capacity of these major carps. Regardless of Rohu's economic potential, still we lack genetic information regarding this species of Halda river. Wild stock *Labeo rohita* of Halda river, Bangladesh was collected and processed to reveal the feature of its mitochondrial genome. Total length of the mitochondrial genome was 16,609 base pairs (bp), containing 13 protein-coding genes, two ribosomal RNAs, 22 transfer RNAs, and one control region. Control region was located between tRNA proline and tRNA phenylalanine which was 885 bp in length. The overall base composition of the mtDNA was found to be 24.32% of T (4039), 27.84% of C (4624), 32.62% of A (5418), and 15.22% of G (2528). The entire mtDNA of Rohu showed a slight AT rich bias (56.94%) with positive A-T skew (0.15) and negative G-C skew (-0.29). This data will provide us with an insight into the phylogeny of *Labeo rohita* and also for further analysis in future, this study can aid researchers to understand evolutionary biology, population genetics of Indian major carps and provide a way to conserve this pure wild stock Rohu of Halda river.

Introduction

Labeo rohita belongs to the family Cyprinidae, is a natural inhabitant of the freshwater section of Bangladesh, India, Pakistan, Burma and Terai region of Nepal and also has been transplanted into Sri Lanka, Japan, Philippines, Malaysia and some countries of Africa (Jhingram and Pullin, 1985). In Bangladesh, Rohu is mostly found in the Padma-Brahmaputra, their tributaries and Halda river system (Alam et al., 2009), though Halda is the only river in Bangladesh from where fertilized eggs of major carps are collected (Tsai et al., 1981; Patra and Azadi, 1985). Halda river is of special interest as it is the only tidal freshwater river in the world that serves as a natural spawning ground for Indian major carps and is the only of its kind in the world from where fishermen collect fertilized eggs directly (Kabir et al., 2013; Kibria, 2009). In Bangladesh natural populations of *Labeo rohita* are showing declining trends but still, it is available in good numbers throughout its habitat range; hence the fish is assessed as Least Concern (IUCN, 2015). It ensures the top position of consumer preference due to its good test, quality, reasonable price and availability and also in the national economy, the contribution of Rohu is beyond question. Rohu contributed 10.50% of the total annual

fish production of inland water bodies, which is far greater than the other three major carps of Bangladesh (DoF, 2018). In spite of such importance, there are no records of phylogenetic study for this economically valuable pure wild Rohu of Bangladesh. Genetic information provides a foundation for the protection and management of biological diversity which enables researchers to decipher the evolutionary histories of diverse biological species and in this respect, the mitochondrial genome is regarded as the marker of choice for the reconstruction of phylogenetic relationships at several taxonomic levels, from population to phyla, and has been widely used for the resolution of taxonomic controversies (Iwasaki et al., 2013; Gissi et al., 2008). In general, mtDNA is described as a closed circular, double-stranded molecule, smaller in size (15-17 kb) in comparison with the nuclear genome and have some distinctive features like relatively stable and compact gene organisation, faster replication, maternal inheritance, lack of recombination and presence of an orthologous gene (Wu et al., 2003; Cao et al., 2006; Saccone et al., 1999). These traits have made mtDNA extensively used for testing hypotheses of macroevolution, studying population structure, phylogeography, and phylogenetic relationships at various taxonomic levels (Saccone et al., 1999; Zhang et al., 2005; Cao et al., 2006). Thus mtDNA is considered as an effective tool for studying phylogenetic and population genetic analysis in vertebrates (Satoh et al., 2016). So, here we reported the entire mtDNA sequences of *L. rohita* from the Halda river revealing what's hidden in its mitochondrial genome structure.

Result and Discussion

Gene organization and base composition

The complete mitogenome of *L. rohita* (MT950724) was 16,609 bp in length containing 37 genes in total. In those 37 genes, 13 protein-coding genes (68.09%), two ribosomal RNA genes (15.91%), 22 transfer RNA (9.42%) genes were found (Table 1 & Table 2). The overall nucleotide composition of A = 5418 (32.62%), T= 4039 (24.32%), C= 4624 (27.84%), and G= 2528 (15.22%) were determined. From this analysis, it was clear that the relative order of nucleotide composition corresponds to the nucleotide pattern of other fish A>C>T>G (Wang et al., 2008). Out of 37 genes, 28 genes were found to be encoded in the F-strand except for nad6 and 8 tRNA (trnQ, trnA, trnN, trnC, trnY, trnS2, trnE, trnP) were encoded in the R-strand of the mitochondrial genome of *L. rohita*. The structural organization and location of the different features of these mito genomes were consistent with the common vertebrate mitogenome genome model (Liu & Cui, 2009).The whole mitochondrial genome showed a positive A-T skew (0.15) and a negative G-C skew (-0.29) (Figure 1). AT and GC content of the total mitogenome were observed to be 56.94% and 43.06% respectively, indicating that the overall nucleotide composition was biased toward adenine and thymine. For comparative analysis, five datasets of mtDNA of Rohu were taken which were available in the NCBI (JQ231111; KR185963; JN412817; AP011201 and MN533986) and determined the AT, GC content

as well as their A-T skew and G-C skew for both the mtDNAs and for their PCGs respectively (Figure 1).

Protein Coding Genes (PCGs) and Their Base Composition

13 protein coding genes constituted 11,310 bp in the whole mitogenome of *Labeo rohita* which accounted for 68.10% (Table 2) of the total mitogenome. Out of 13 PCGs, 12 of them were encoded in the F-strand (nad1, nad2, cox1, cox2, atp8, atp6, cox3, nad3, nad4l, nad4, nad5 and cob) where nad6 (13850-14368) was encoded in the R-strand of the mtDNA.

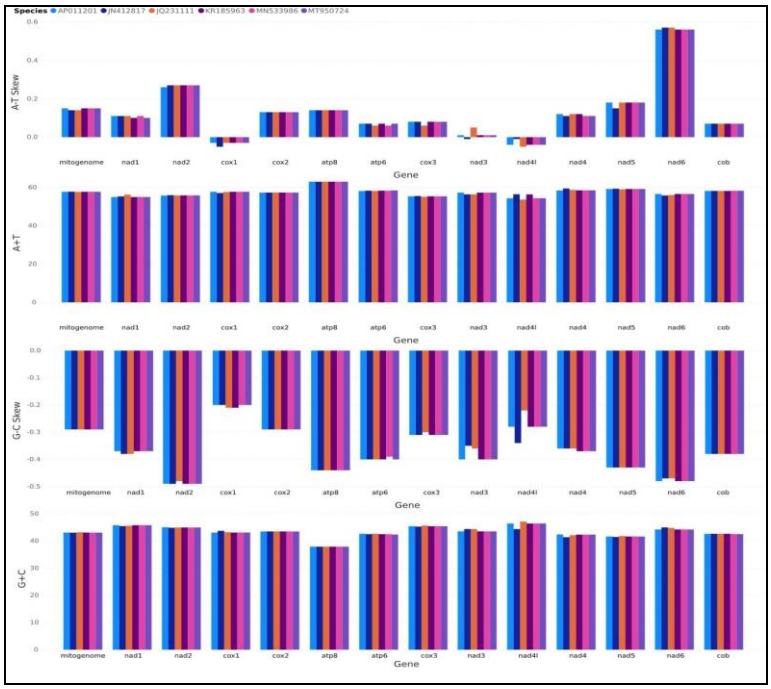


Figure 1: Comparative analysis of skewness, AT and GC percentage of two Bangladeshi Rohu (MN533986 and MT950724), three Indian Rohu (JQ231111; KR185963; JN412817) and one Japanese Rohu (AP011201).

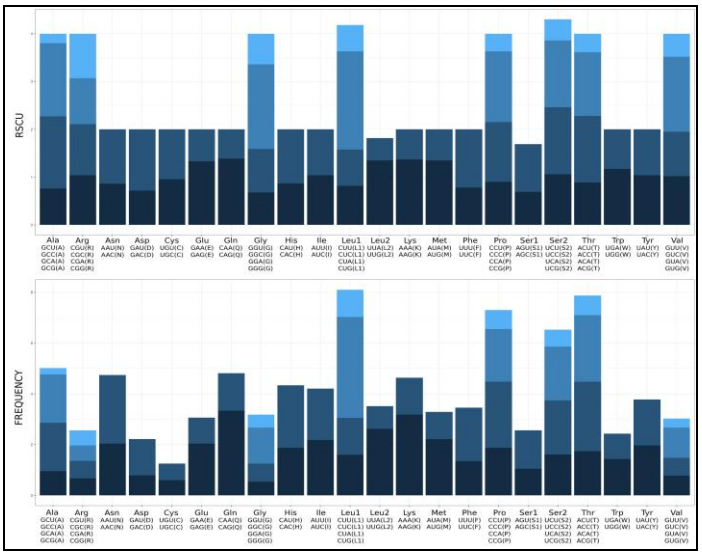


Figure 2: Relative synonymous codon usage (RSCU) and codon usage frequency in *Labeo rohita* (X axis represents the codon families and Y axis represents the RSCU and frequency respectively).

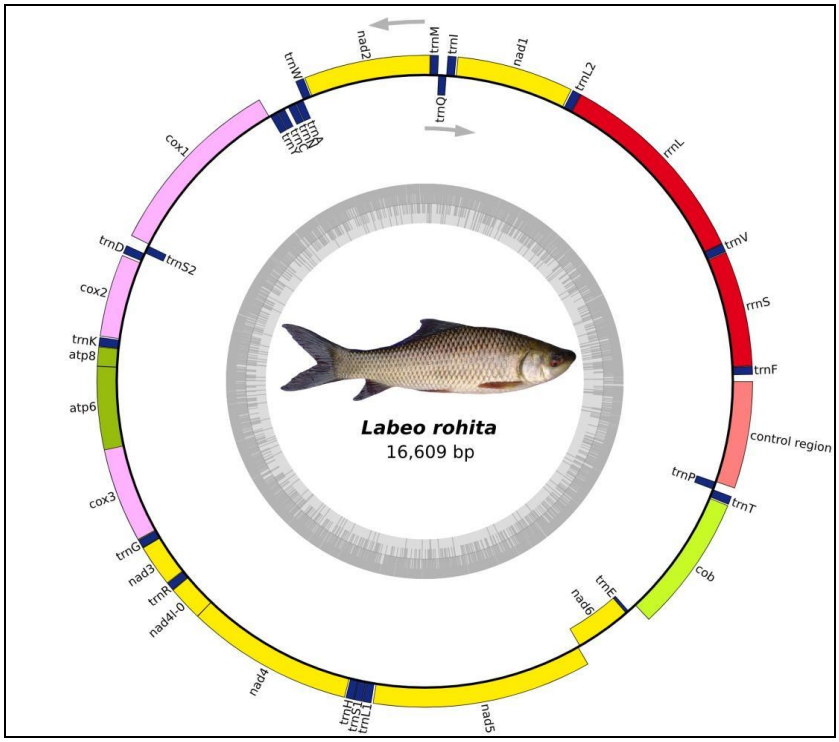


Figure 3: The circular representation of the whole mitochondrial genome of *Labeo rohita* (Hamilton 1822).

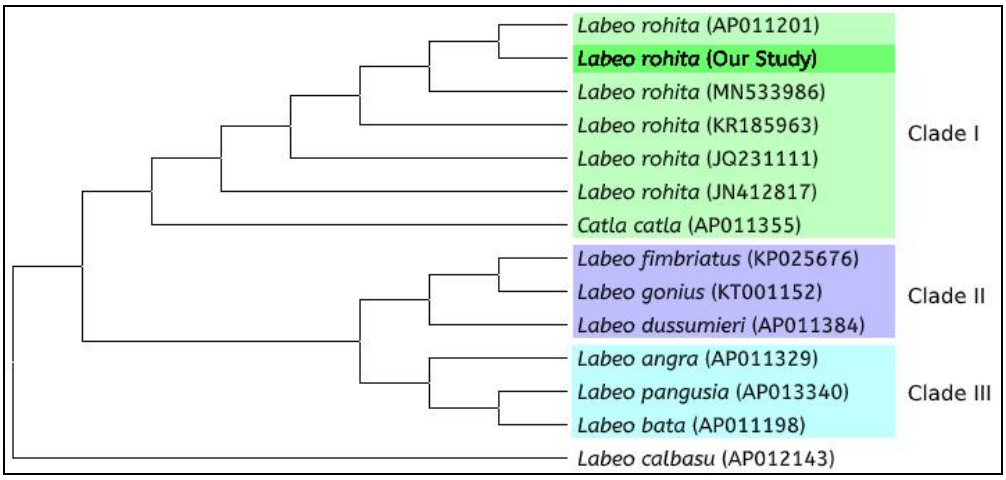


Figure 5: Phylogenetic analysis of *Labeo rohita* (Hamilton, 1822). The evolutionary history was inferred using the Maximum Parsimony method. The MP tree was obtained using the Subtree-Pruning-Regrafting (SPR) algorithm and branch lengths were calculated using the average pathway method (Nei and Kumar, 2000). This analysis involved 14 nucleotide sequences. Evolutionary analyses were conducted in MEGA X (Kumar, et al 2018).

The AT and GC content of the total PCGs was 56.72% and 43.28% respectively (Table 2). A-T and G-C skews of PCGs were 0.08 and -0.32 respectively reflecting the fact that adenine content is comparatively higher than thymine while cytosine content is higher than guanine which was also observed in the case of the whole mitochondrial genome (Table 2). In the individual analysis, out of 13 PCGs, 10 showed positive A-T skew except for the gene cox1 (-0.02), nad3 (-0.01) and nad4l (-0.04), whereas G-C skew was observed to be negative in all PCGs (Figure 1). The size of the PCGs was ranging from 162-1800 bp where nad5 (1800 bp) being the longest and atp8 (168 bp) being the shortest among all the PCGs (Table 1). Overlapping was observed between two adjacent pairs of PCGs (atp8-atp6, and nad4l-nad4) (Table 1). The overlap between atp8-atp6 and nad4l-nad4 genes were observed to be -4 respectively in each pair and in total -8 bp overlap was observed in our current study, this sort of overlap is common in most vertebrate mitochondrial genomes (Broughton et al. 2001). No overlapping had been observed between PCGs and other genes (tRNA and rRNA). In RSCU analysis CUA, GGA, GUA and GCA showed the highest utilization in the protein coding genes of *L. rohita* of The Halda river. From the frequency pattern, it was visible that Lucine, Threonine, Proline and Serine were the most common amino acids among the PCGs.

rRNAs, tRNAs and Their Base Composition

The total size of the rRNA was 2,643 bp and formed by two subunits that are 12S rRNA (954 bp) and 16S rRNA (1689 bp) which constituted 15.91% of the total mitochondrial

genome (Table 1 and Table 2). A-T and G-C skew of total rRNA were 0.29 and -0.10 respectively. AT and GC content of the total rRNA were 54.69 and 45.31 respectively. In individual analysis, both 12S rRNA and 16S rRNA showed biasness towards AT content. So it could be concluded that the occurrence of adenine and cytosine were higher than thymine and guanine in the rRNAs, as observed in the whole mitochondrial genome of *L. rohita*. No overlapping base pairs had been detected between the rRNAs with their adjacent tRNAs.

Phylogenetic Analysis

Phylogenetic tree had been constructed following the Maximum Parsimony (MP) method where the bootstrap value was 1000 (Figure 5). 14 mitogenome sequences of 9 species were taken into consideration while constructing this tree. Except for our own species, the rest of the 13 sequences of mitochondrial genome were taken from NCBI database. In our analysis, three distinctive clades were found to be originating from a common ancestral point. From the tree, it was visible that our Rohu from Halda river formed a sister group with another Rohu species (AP011201) as both of them shared the most recent common ancestor, while with another Rohu (MN533986), Halda's Rohu showed paraphyletic relationship. All these six Rohu from three different geographical regions belonged to a single clade (Clade I) forming a monophyletic group along with *Catla catla* (AP011355). In Clade I, the distant related species is *Catla catla* (AP011355), another Indian major carp, showing the continuity of evolutionary changes from the clustal node. Here in this phylogenetic analysis, another Indian major carp *Labeo calbasu* (AP012143) was included, which formed an outgroup with the rest of the three clade and their descendants, showing an early speciation and least relatedness.

References

Alam, M.S., Jahan, M., Hossain, M.M. and Islam, M.S., 2009. Population genetic structure of three major river populations of rohu, *Labeo rohita* (Cyprinidae: Cypriniformes) using microsatellite DNA markers. Genes & Genomics, 31(1), pp.43-51.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M. and Stadler, P.F., 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. Molecular phylogenetics and evolution, 69(2), pp.313-319.

Broughton, R.E., Milam, J.E. and Roe, B.A., 2001. The complete sequence of the zebrafish (*Danio rerio*) mitochondrial genome and evolutionary patterns in vertebrate mitochondrial DNA. Genome research, 11(11), pp.1958-1967.

Cao, S.Y., Wu, X.B., Yan, P., Hu, Y.L., Su, X. and Jiang, Z.G., 2006. Complete nucleotide sequences and gene organization of mitochondrial genome of *Bufo gargarizans*. Mitochondrion, 6(4), pp.186-193.

DoF. 2018. Yearbook of Fisheries Statistics of Bangladesh, 2017-18. Fisheries Resources Survey System (FRSS), Department of Fisheries. Bangladesh : Ministry of Fisheries, Vol (35), p. 23.

Dierckxsens, N., Mardulyn, P. and Smits, G., 2017. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic acids research, 45(4), pp.e18-e18.

Gissi, C., Iannelli, F. and Pesole, G., 2008. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. Heredity, 101(4), pp.301-320.

Greiner, S., Lehwark, P. and Bock, R., 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Research, 47(W1), pp.W59-W64.

Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T.P., Sado, T., Mabuchi, K., Takeshima, H., Miya, M. and Nishida, M., 2013. MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Molecular biology and evolution, 30(11), pp.2531-2540.

Jhingran, V.G. and Pullin, R.S., 1985. A hatchery manual for the common, Chinese, and Indian major carps. Asian Development Bank, Manila, Philippines and International Center for Living Aquatic Resources Management, Manila, Philippines, p. 252.

Kabir, M.H., Kibria, M.M. and Hossain, M.M., 2013. Indirect and non-use values of Halda River-a unique natural breeding ground of Indian carps in Bangladesh. Journal of Environmental Science and Natural Resources, 6(2), pp.31-36.

Kibria, M.M., Farid, I. and Ali, M., 2009. Halda Restoration Project: Peoples Expectation and Reality, A Review Report Based on the Peoples Opinion of the Project Area (In Bangla). Chittagong: Chattagram Nagorik Oddogh & Actionaid Bangladesh, p. 67

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K., 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution, 35(6), pp.1547-1549.

Naser, M. N., 2015. Labeo rohita. Red List of Bangladesh, Freshwater Fishes. IUCN, International Union for Conservation of Nature, Bangladesh Country Office, Dhaka, Bangladesh, Vol (5), p. 190.

Nei, M. and Kumar, S., 2000. Molecular evolution and phylogenetics. Oxford university press, p.126.

Patra, R.W. and Azadi, M.A., 1985. Hydrological conditions influencing the spawning of major carps in the Halda River, Chittagong, Bangladesh. Bangladesh Journal of Zoology, 13(1), pp.63-72.

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G. and Reyes, A., 1999. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. Gene, 238(1), pp.195-209.

Satoh, T.P., Miya, M., Mabuchi, K. and Nishida, M., 2016. Structure and variation of the mitochondrial genome of fishes. BMC genomics, 17(1), p.719.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R. and Greiner, S., 2017. GeSeq–versatile and accurate annotation of organelle genomes. Nucleic acids research, 45(W1), pp.W6-W11.

Tsai, C.F., Islam, M.N., Karim, R. and Rahman, K.S., 1981. Spawning of major carps in the lower Halda River, Bangladesh. Estuaries, 4(2), pp.127-138.

Wu, X., Wang, Y., Zhou, K., Zhu, W., Nie, J. and Wang, C., 2003. Complete mitochondrial DNA sequence of Chinese alligator, *Alligator sinensis*, and phylogeny of crocodiles. Chinese Science Bulletin, 48(19), pp.2050-2054.

Zhang, P., Zhou, H., Liang, D., Liu, Y.F., Chen, Y.Q. and Qu, L.H., 2005. The complete mitochondrial genome of a tree frog, *Polypedates megacephalus* (Amphibia: Anura: Rhacophoridae), and a novel gene organization in living amphibians. Gene, 346, pp.133-143.

Research Objective 5 and 6

Draft whole Genome and mitogenome of Mrigal (*Cirrhinus cirrhosus* Bloch, 1795) from the river Halda, Bangladesh

Abstract

Mrigal (*Cirrhinus cirrhosus*), freshwater fish and a species of ray-finned fish belonging to the Cyprinidae family, which is very common and widely distributed in the Southeast asian countries including Bangladesh. It is a bottom dweller as well as detritus feeder and stenophagic in feeding habit, which makes this fish important for the ecosystem. Due to fast growth rate and high consumer preference it has economical significance too. It is often used as game fish in Bangladesh. The world aquaculture production of *C. cirrhosus* was 378622 MT in 2010, which ranked 24th among all aquaculture species. Here, we report the first draft genome assembly (1,152,819,041 bp with 10154 bp N50) of *C. cirrhosus* of Halda River of Bangladesh. It reveals 61.7% completeness with actinopterygii and out of the 17018 predicted genes, 14745 were identified as GO terms. This study also provides a complete mitochondrial genome of 16,607 bases and it is featured by 13 protein-coding genes, 2rRNA genes, 22 tRNA genes and one non-coding control region (D loop). Most of the genes were encoded on the heavy strand, while the nad6 gene and 8 tRNA genes were on the light strand. The total nucleotide composition was A:32.6%, G:15.3%, C:27.4%, T:24.6% and the A+T content (57.2%) was higher than G+C content (42.8%).

Introduction

Cirrhinus cirrhosus (Bloch, 1795), (Cypriniformes: Cyprinidae) is a warm-water teleost, inhabitant of Indo-Gangetic riverine system spread across northern and central India, and the rivers of Pakistan, Bangladesh, Nepal and Myanmar (Reddy, 1999; Dahanukar, 2010). It is also called *Cirrhinus mrigala* at Fishbase, Morakhi, Moree, White carp and Mrigal carp fish which are generally a species of freshwater, but can also tolerate high levels of salinity and the only surviving wild population of this fish is in the Cauvery River of India (https://www.roysfarm.com/mrigal-fish/). *Cirrhinus cirrhosus* (Mrigal) is widely cultured under polyculture system along with Indian major carps as the third popular major carp species next to *Catla catla* and *Labeo rohita* inBangladesh with it's neighboring countries in South Asia. Besides Bangladesh it has become an important component in the fish culture systems of India, Pakistan, Myanmar, the Lao People's Democratic Republic, Thailand , Nepal and so many countries. It plays an impotrant role in the control of the algal population. The Richness of Cyprinids, indicates that these species from the carp family probably take benefit from eutrophication (Ådjers *et al*., 2006; Snickars *etal*., 2015). The growth rate of *C.cirrhosus* is fast, whose fry can grow to more than 400g in the first year and it's fertility is high, which increases with age, usually 100,000-150,000 eggs per kg of

body weight (Chondar, 1999). The species is of commercial significance due to its aquaculture potential and high consumer preference (Chauhan T. *et al*., 2007). This species is resistant to hypoxia and has strong disease resistance, it can grow normally in reservoirs, lakes, ponds and rivers (Chen *et al*., 2004). Halda river is one of the rivers which is unique and distinct due to genetically purity of fertilized spawn of *C. cirrhosus* and other carps which makes it the most economic heritage of this country as well as in South Asia (Akter and Ali, 2012). The spawn or post-larvae of Mrigal are called Renu or Dhani, which are collected during the early monsoon from the Halda river. After the Karnafuli and the Sangu, Halda is the third main river of southern Chattogram sub-region, is well-established of genetic resources. 60% of the country's pond carps culture on the fish fry naturally produced in the river and it also provides navigation, drinking water supply, sand querying, irrigation which makes the river a natural resource of immense economic value (Kabir *et al*., 2103). Halda river has become constrained by the degradation of habitats as a result of environmental modifications and anthropological intervention (Akhtar *et al*., 2017). So due to this degradation, as an element of Halda river it is obvious that the density of Mrigal species and eggs released by mother fishes are gradually reducing to an alarming rate but the demand

Fig. 1. Mrigel (*Cirrhinus cirrhosis) sampled during this WGS study*

for fish fry has increased several folds through aquaculture expansion. Thus it is essential to understand the genetic composition of Indian major carps for the management of their natural population in the river. Therefore, it is essential to examine the stock structure of *C. cirrhosus* to overcome this fall in catch of this important Indian major carp and to assess the quality of the hatchery stocks in future. Already to some extent, the gene pools of our indigenous varieties of carps viz: Rohu, Catla & Mrigal have been contaminated (Khatun *et al.*, 2017). Information on individual identity, breeding pattern, degree of pertinence and distribution of genetic variation

among them along with evolutionary and adaptive behavior should be known to get rid of such genetic contamination of the seed production. To do this whole genome sequencing is important. Although some study of mtgenome (Bej *et al*., 2012, Zhai *et al.*, 2020) and mitochondrial gene sequencing (Das *et al.,* 2013, Behera *et al*., 2015, Karim *et al*., 2018, Das *et al*., 2018), microsatellite DNA marker analysis of *C. Cirrhosus* (Lal *et al.,*2004, Chauhan T *et al*., 2007, Hasanat *et al*., 2015, Sharker and Siddik, 2015) and transferrin cDNA sequence (Sahoo *et al*., 2008) has been reported still there has been no records of full genomic information of this species. With the advancement of sequencing technologies, there has been a rapid increase in the number of genome assemblies for terrestrial species compared to aquatic species (including fish) in the last decade, with a very small (Kelley *et al*., 2016) percentage of fish genomes given the most numerous taxonomic group and huge diversity exhibited by teleosts (Ravi and Venkatesh, 2018). The current study *de novo* genome assembly and annotation of *Cirrhinus cirrohsus* using next generation sequencing will help us to develop a draft genome of *C. Cirrhosus* and allow us the knowledge of genetic variation to explore the evolution and adaptation at genome level. This study will provide a useful platform for the functional genome and conservation research of halda river mrigal carps in the future. The knowledge will help us develop a better policy for their breeding, behavioral pattern analyses with a view towards their in situ conservation.

Methodology

Sampling and DNA isolation:

Fresh sample of adult female *Cirrhinus cirrhosus* of length 81.28 cm (32 inch) and of width 27.94 cm (11 inch) was collected from river Halda (Longitude/Latitude-22°28'56.09"N & 91°54'07.62"E) and transported to the laboratory to preserve at -70° C for further analysis during the period of September, 2019. The muscle tissue was dissected and stored with 95% ethanol , later a high molecular weight gDNAs were isolated and purified using the AddPrep Genomic DNA extraction kit for future evaluation of the quality and quantity of the DNA. All the methods had been performed in accordance with the "Regulations for Animal Experiments in Chittagong Veterinary and Animal Sciences University's unique feature, the Indian and GOB ethical clearance" as required.

Genome sequencing and assembly

Library preparation:

Purified genomic DNA was sent for library preparation and whole genome sequencing (WGS) at BGI Group (Shenzhen, Guangdong, China). Almost 89.8 gb of high-quality reads were generated with a read length of 150 bp using Next- generation sequencing (NGS) technology on an Illumina Hiseq 4000 platform. After sequencing quality of raw sequence reads and trimmed sequencing reads were inspected using FastQC version 0.11.8 (Andrews, S., 2010). Then both the assemblies of mitogenome and whole genome were done step by step.

Mitochondrial Genome Sequence analysis:

The mitochondrial genome reads were separated from the whole genome by mapping it against the reference Mrigal mitochondrial genome (NC_033964) by cleaning up Human and bacterial gene contamination using BWA V0.7.17 and SAMTOOLS V0.1.19. To separate mitochondrial genome, firstly that reference mitogenome was downloaded from NCBI. Now BWA V0.7.17 was used to index the ref.fasta. After indexing, alignment of raw reads was done by BWA V0.7.17 and SAMTOOLS V0.1.19. Finally new mitogenome reads were mapped and then splitted using samtools and gatk. The organelle assembler NOVOPlasty V.2.7.2 (Dierckxsens, Mardulyn, & Smits, 2017) was used to assemble the clean reads. To verify assembly mummer was used. Web servers MITOS (Bernt et al., 2013), GeSeq (Tillich *et al*., 2017) and MitoAnnotator (Iwasaki et al., 2013) were applied to perform structural and functional annotation. For structural annotation, the fasta file which was generated from novoplasty, was uploaded into MITOS and MitoAnnotator with default parameters. Within one hour Mitos generated 9 types of statistic files and 6 from MitoAnnotator within few minutes. The assembled fasta file with required parameters, was submitted into GeSeq for functional annotation and verification. From GeSeq the locations and secondary structures of 22 tRNA genes were determined using tRNAscan-SE v2.0.6 (Chan *et al*., 2019) and ARWEN v1.2.3. (Laslett and Canback, 2008). Codon count values were generated by SMS which means Sequence Manipulation Suite (Stothard P, 2000) and MEGA X (Kumar et al., 2018). OGDRAW (Greiner, Lehwark, & amp; Bock, 2019) was used to construct the circular representation of the entire mitogenome which was also got from MitoAnnotator. Nucleotide composition, Relative Synonymous Codon Usage (RSCU) values and genetic divergence between species were calculated using the software MEGA. Finally, mtDNA sequences were aligned and a phylogenetic tree was constructed by using MEGA X (Kumar et al., 2018). The Maximum Likelihood method and Neighbor-Joining method to construct tree using the software MEGA with 5,00 bootstrap value.

Whole Genome assembly:

De novo genome assembly comprises step such as read preprocessing, contig assembly and refinement, scaffolding and gap filling. We used ABySS v. 1.9.0 (Shaun D Jackman et al., 2017), SOAPdenovo2 v. 2.04, Platanus v.1.2.4 (Kajitani R, 2014) and Megahit v.1.2.9 Li *et al*. 2015) assemblers to assemble the Mrigal genome. Since there is currently no de novo assembler assured to outperform others and as assemblers overall performance can differ relying on the dataset, four unique assemblers had been used and an assembly evolution was subsequently performed in order to choose the best assembler. All assemblers follow the classic De Bruijin graph illustration even though the assembly algorithm differs across methods. Finally BUSCO v.4.1.2 (Simão, F *et al*. , 2015) was used to assess the assembly quality in terms of gene completeness. In the NCBI GeneBank under the Accession numbers JACDQS000000000 the genome sequence data has been deposited.

Gene prediction and functional annotation:

The first step in genome annotation in a given genomic sequence is gene structure prediction. Gene prediction was conducted using MAKER ver 3.01.03 (Campbell *et al*., 2014) which defines probability distributions for the different sections of genomic sequence . Gene prediction was performed ab initio with using both given and default parameters. Functional annotation was obtained by InterProScan ver 5.46-81.0 (Jones et al., 2014) and Pannzer2 (Törönen, 2018).

Results

Whole Genome:

Rapid development of sequencing technologies enables highly accurate base calling of DNA. Here, Illumina Hiseq 4000 was used for a cost efficient high coverage short read sequencing of genomic data. High coverage data substantially decreases sequencing errors. Shotgun sequencing was mainly used to produce continuous consensus sequences (contigs). Building the longest possible consensus sequences is the eventual aim of genome assembly to illustrate individual chromosomes. A contig is a contiguous length of genomic sequence in which the order of bases is known to a high confidence level. Contigs, by chaining together and using additional information about the relative position and orientation of the contigs in the genome create scaffolds. Scaffolds are composed of contigs and gaps. Gaps occur when only the distance (not the nucleotides) between two contigs is known. Contigs in a scaffold are seperated by gaps which can be closed later in the gap-filling process (Nagarajan *et al*., 2010). So to

say Genome assembly functions to order and orient the contigs in the assembly of a draft genome into larger scaffolds (Shieh *et al*., 2020).

Whole genome sequencing was performed on a single individual female *Cirrhinus cirrhosus* obtained from Halda river. A total of 89.9 Gb (Gigabase pair) of subread bases with a read length of 150 bp were generated using Next- generation sequencing (NGS) technology on an Illumina Hiseq 4000 platform. After sequencing quality of raw sequence reads and trimmed sequencing reads were inspected using FastQC version 0.11.8 (Andrews, S., 2010). The FastQC result ensured that the raw data is too good which was being analysed by html report generated from FastQC. The html report evaluated that the results of the all modules are entirely normal. All the modules are described below:

Basic statistics:

The Basic Statistics module generates some simple composition statistics for the file analysed.

Measure	**Value**
Filename	FP200000269TL_L01_SP2012180236_1.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	72145277
Sequences flagged as poor quality	0
Sequence length	150
%GC	36

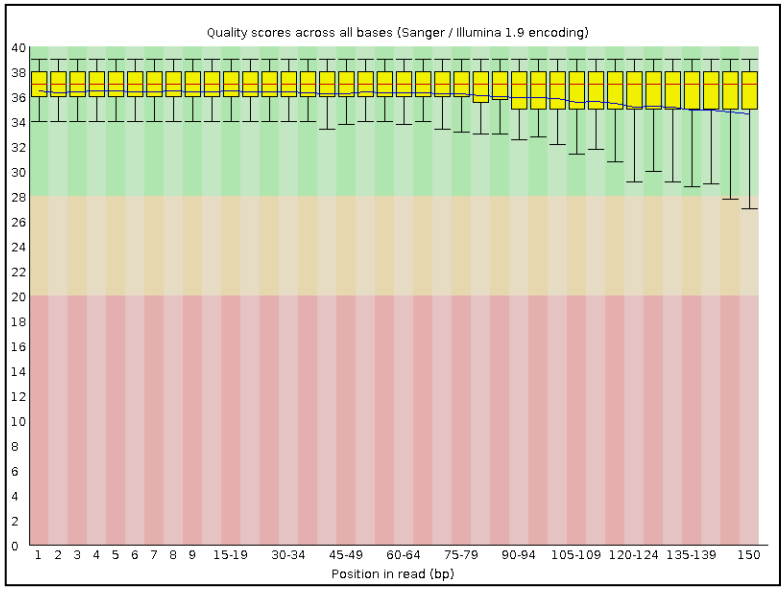


Fig.2 : FastQC quality result of the raw dataset

After correction of the high quality reads through BFC (Li,H. , 2015) these sequences were assembled to construct scaffold. The assembly resulted in 788002 contigs with 13.1 kb N50 value that is generated by Megahit. Before assembly genome size estimation was done by KmerGenie which estimates the best k-mer length for de novo assembly. The genome was observed to have an estimated genome size of 1.02 gb for the best *k*-mer value 85.

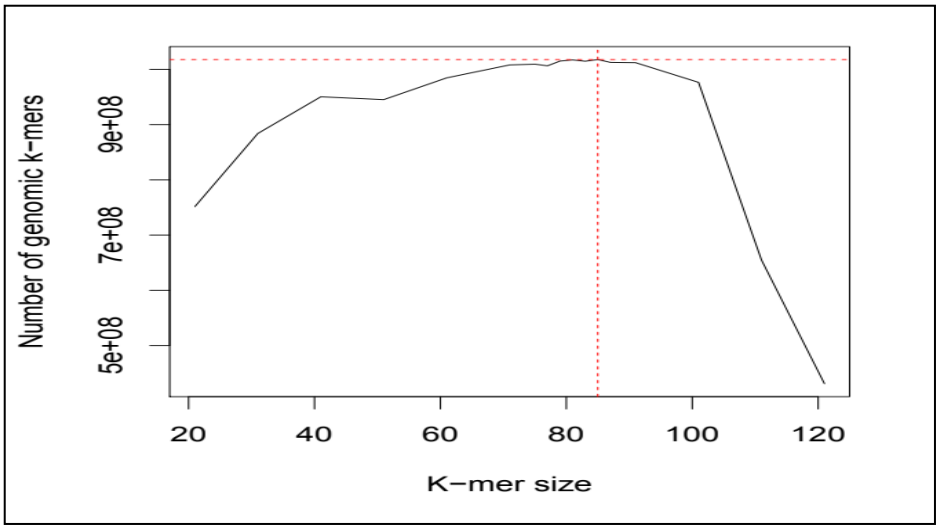


Fig.3 : KmerGenie result of the raw dataset

BUSCO analyses

The BUSCO analysis revealed 68.3% completeness with eukaryote and 61.7% with actinopterygii. The assembled genome size of *C. cirrhosus* is 1.19 gb accounting for 85% of the estimated Mrigal genome size of 1.02 gb. Including Megahit, assembly was done by four different assembly softwares. These assemblers generated 3617186 scaffolds along with 4994783 contigs in Platanus, 1905303 scaffold along with 1838589 contigs in Abyss and 1261374 scaffold along with 4692660 contigs in SOAPdenovo. These values are surpassed by Megahit as it generated the contig set with the fewest sequences, highest nucleotide content, highest mean sequence length and highest N50 value. In the NCBI GeneBank under the Accession numbers JACDQS000000000 the genome sequence data has been deposited

Abyss:

name	n	n:500	L50,	min,	N75,	N50,	N25,	E-size,	max,	sum
unitigs	2270343	520808	123118	500	1075	1876	3134	2411	22233	767.5e6
contigs	1905303	362040	63238	500	1897	3561	6411	4844	54598	795.8e6
scaffolds	1838589	317524	42596	500	2230	5145	9607	7011	86838	795.3e6

MEGAHIT:

name	n	n:500	L50	min	N75	N50	N25	E-size	max	sum
contgs	788002	209353	20438	500	5553	13155	24721	17591	122795	976e6

Platanus:

name	n	n:50	L:50	min	N75	N50	N25	E-size	max	sum
gapClosed	3617186	287587	68749	500	1195	2025	3369	2628	38954	462.2e6
contigs	4994783	445053	152788	500	642	844	1192	1004	8369	372.7e6
scaffolds	3617186	287586	69463	500	1154	1946	3240	2530	38514	448.5e6

SOAPdenovo:

name	n	n:50	L:50	min	N75	N50	N25	E-size	max	sum
contigs	4692660	506871	125767	500	1036	1746	2821	2221	21970	714e6
scaffolds	1261374	203850	26609	500	3900	8788	16603	12139	122260	853e6

Raw data statistics:

Nucleotide_A	326256582	30.73%
Nucleotide_C	184205451	17.35%
Nucleotide_G	183051471	17.24%
Nucleotide_T	322738550	30.40%
GapContent_N	45514982	4.29%
Non_ACGT_N	0	0.00%;
GC_Content	36.14%	

Genome Annotation:

MAKER ver. 3.01.03 pipeline was used for structural annotation. GC content of the genome was determined to be 37.03%. RepeatMasker ver. 4.1.0 has been run three times but every time it failed because of technological problem so we are still working on it. Overall, 17018 gene models were predicted using the MAKER gene annotation pipeline based on both de novo and reference-based predictions using genes and proteins from other fish species (Goldfish, Common Carp, Zebrafish). Out of the 17018 genes, 14745 were identified as GO terms and 6560 DE terms using Pannzer2 (Törönen, 2018).

Mitochondrial genome sequencing

To assemble the mitogenome, a total of 16,607 bp of assembly size was generated by NOVOplasty (Dierckxsens *et al*., 2017). Annotation was analysed by three methods; Geseq (Tillich *et al*., 2017), Mitos (Bernt *et al*., 2013) and MitoAnnotator (Iwasaki et al., 2013). These three methods resulted similar sizes for all tRNAs whereas differences in the PCGs and two rRNAs. For submission to the GenBank, the results obtained from MitoAnnotator was selected.

Genome structure, organization and composition:

Cirrhinus cirrhosus complete mitochondrial genome (NCBI accession number MW649087) is a closed circular and double-stranded molecule of 16,607 bp in length and included 37 genes: 13 protein-coding genes (cox1-3, nad1-6, nad4L, cob, atp6, and atp8), 2 rRNAs (large or 16S and small or 12S), 22 tRNAs (one for each amino acid and two each for leucine and serine) a Control region and a Light origin replicate as in other vertebrate (Cheng *et al*. 2010; Shi *et al*. 2012). The gene distributions are

similar to other teleost mitochondrial genomes and among 37 genes, except nad6 gene and 8 tRNA genes (trnA, trnC, trnE, trnN, trnY, trnQ, trnS2 and trnP) which were encoded on the light strand, other 28 were found to be encoded on the heavy strand. The overall base composition of the Mrigal mitogenome was A:32.6%, G:15.3%, C:27.4%, T:24.6% and the A+T content (57.2%) was higher than G+C content (42.8%) which corresponds to other teleosts (Tzeng *et al*. 1992; Wang *et al*. 2007, Yue *et al*., 2006, Cui *et al*., 2009). The relative order of nucleotide composition assimilates to the pattern of other cyprinids fishes: A>C>T>G (Wang *et al*., 2008). The highest AT content was found in control region (62.8%) followed by PCGs (56.3%), rRNAs (53.5%) and tRNAs (53.2%). The narration of overall patterns of nucleotide composition in DNA sequences remains vested in AT and GC skew, which were observed for *C. cirrhosus* is +0.14 and -0.28 respectively meaning that adenine and cytosine is of high content than thymine and guanine.

MitoAnnotator:

In total there are 21 overlapping nucleotides in the range from 1 to 7 bp, which are found at 5 distinct locations. The largest overlapping regions (7 bp) are observed in four protein coding genes- between 2 NADH dehydrogenase subunits (nad4L and nad4) and between 2 ATPase subunits (atp8 and atp6) and 4 bp between ND5 and ND6.. 2 bp overlapping regions were found between trnI and trnQ and 1 bp between trnT and trnP. Intergenic spacers were observed in 14 genes with a total length of 74 bp, with varying ranges from 1 to 33 bp. The most significant intergenic spacers (33 nucleotides) is located between trnN and trnC followed by trnD and Cox2 (15 nucleotides). The putative control region of the mitochondrial genome is located between the tRNA-P and tRNA-F consisting of 938 bp in length.

Fig: Complete automated annotated sequences of *C. cirrhosus* (16607) generated in MitoAnnotator.

Protein	Strand	Position		Size (bp)	Start codon	Stop codon	Intergenic nucleotide
		Start	Stop				
$tRNA^{Phe}$	+	1	69	69			
12S rRNA	+	70	1024	955			0
$tRNA^{Val}$	+	1025	1096	72			0
16S rRNA	+	1097	2784	1688			0

$tRNA^{Leu}$	+	2785	2860	76			0
ND1	+	2862	3836	975	ATG	TAA	1
$tRNA^{Ile}$	+	3841	3912	72			4
$tRNA^{Gln}$	-	3911	3981	71			-2
$tRNA^{Met}$	+	3984	4052	69			2
ND2	+	4053	5097	1045	ATG	T_{++}	0
$tRNA^{Trp}$	+	5098	5168	71			0
$tRNA^{Ala}$	-	5171	5239	69			2
$tRNA^{Asn}$	-	5242	5314	73			2
$tRNA^{Cys}$	-	5348	5414	67			33
$tRNA^{Tyr}$	-	5416	5486	71			1
COI	+	5488	7038	1551	GTG	TAA	1
$tRNA^{Ser}$	-	7039	7109	71			0
$tRNA^{Asp}$	+	7113	7184	72			3
COII	+	7200	7890	691	ATG	T_{++}	15
$tRNA^{Lys}$	+	7891	7966	76			0
ATPase 8	+	7968	8132	165	ATG	TAA	1
ATPase 6	+	8126	8808	683	ATG	T_{++}	-7
COIII	+	8809	9594	786	ATG	TAA	0
$tRNA^{Gly}$	+	9595	9666	72			0
ND3	+	9667	10015	349	ATG	T_{++}	0
$tRNA^{Arg}$	+	10016	10085	70			0
ND4L	+	10086	10382	297	ATG	TAA	0
ND4	+	10376	11756	1381	ATG	T_{++}	-7
$tRNA^{His}$	+	11757	11825	69			0
$tRNA^{Ser}$	+	11826	11894	69			0
$tRNA^{Leu}$	+	11896	11968	73			1
ND5	+	11972	13795	1824	ATG	TAA	3

ND6	-	13792	14313	522	ATG	TAA	-4
$tRNA^{Glu}$	-	14314	14382	69			0
Cyt b	+	14388	15528	1141	ATG	T_{++}	5
$tRNA^{Thr}$	+	15529	15600	72			0
$tRNA^{Pro}$	-	15600	15669	70			-1
CR		15670	16607	938			0

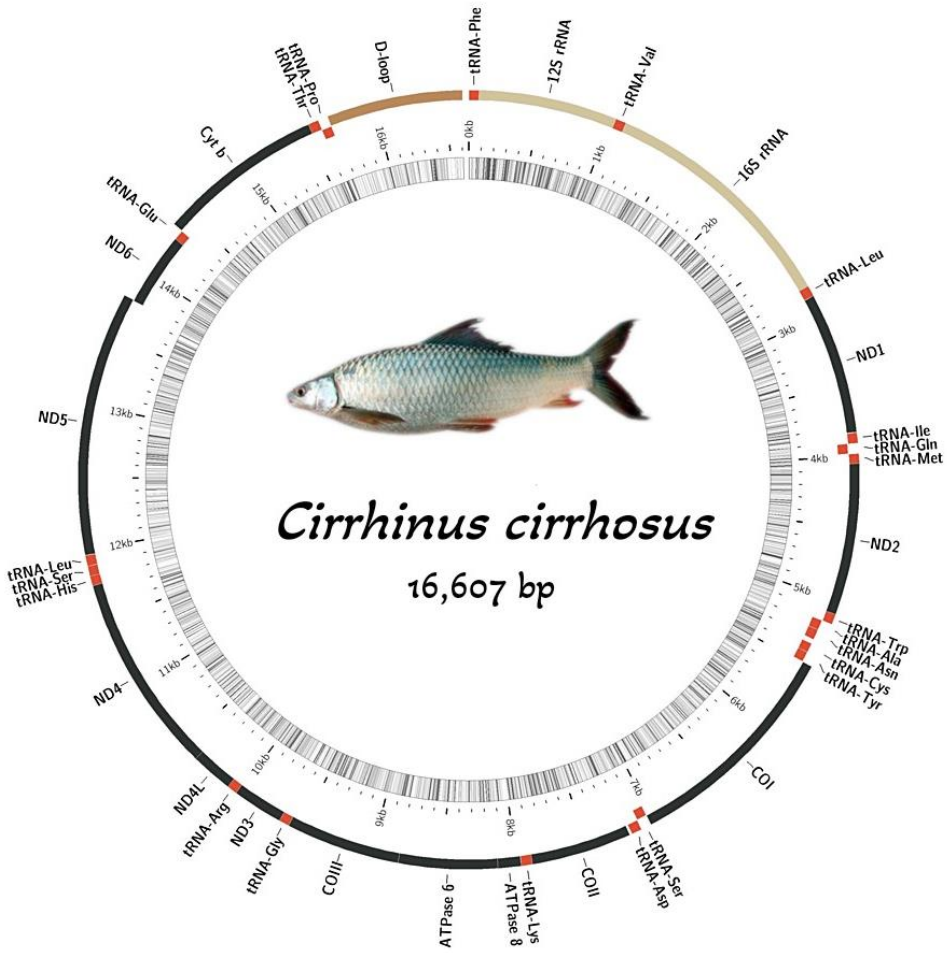


Fig 4: Fig 4: Mitigenome organization of *C. cirrhosus* generated by MitoAnnotator. The genes on the outer side of the circle are coded on the H-strand while those on the inner side are coded with L-strand.

Phylogenetic analysis of mitochondrial genome

Phylogenetic tree of *C. cirrhosus* is constructed (Fig. 5) using using both Maximum Likelihood method followed by Kimura 2-parameter model (Kimura M., 1980) and Neighbor-Joining method (Saitou N. and Nei M. 1987) by MEGA X (Kumar *et al*., 2018), where default bootstrap value was 500. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein J., 1985). Bootstrap value was raised from 500 to 1000. The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura *et al*., 2004) and are in the units of the number of base substitutions per site. This analysis involved 21 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. After removing All ambiguous positions through pairwise deletion option the aligned sequence was subjected to phylogenetic analysis and an unrooted tree was constructed by MEGA. The consensus tree seperates species into clear clades or tribes. According to both analysis, *Cirrhinus cirrhosus* from the Halda river shows maximum similarity with the *Labeo calbasu* (AP012143) and *Labeo chrysophekadion* (AP11199).

In order to investigate the evolutionary relationships between *Cirrhinus cirrhosus* and other 20 closely related species, based on the whole mitogenome sequences a phylogenetic tree was constructed. The data of 20 species were downloaded from NCBI and the accession numbers are shown on figure **6.** Among these 21 species the tree indicates that there are 19 groups which are branched from one common ancestor that means there are 19 monophyletic groups. In the case of the investigated species *Labeo calbasu* is grouped as monophyletic because *C. cirrhosus* and *L. calbasu* are seen to be descendants of common ancestors. So to say *Cirrhinus cirrhosus* and *L. calbasu are* sister taxa as they are being shown as closely related to each other. These two are sister taxa to another species of *Labeo* genus *L. chrysophekadion.* The another species from Japan (AP012150) and the present species *C. cirrhosus* of Halda river are seen as a polyphyletic group. Both method revealed the same evolutionary history among these 21 Cyprinids.

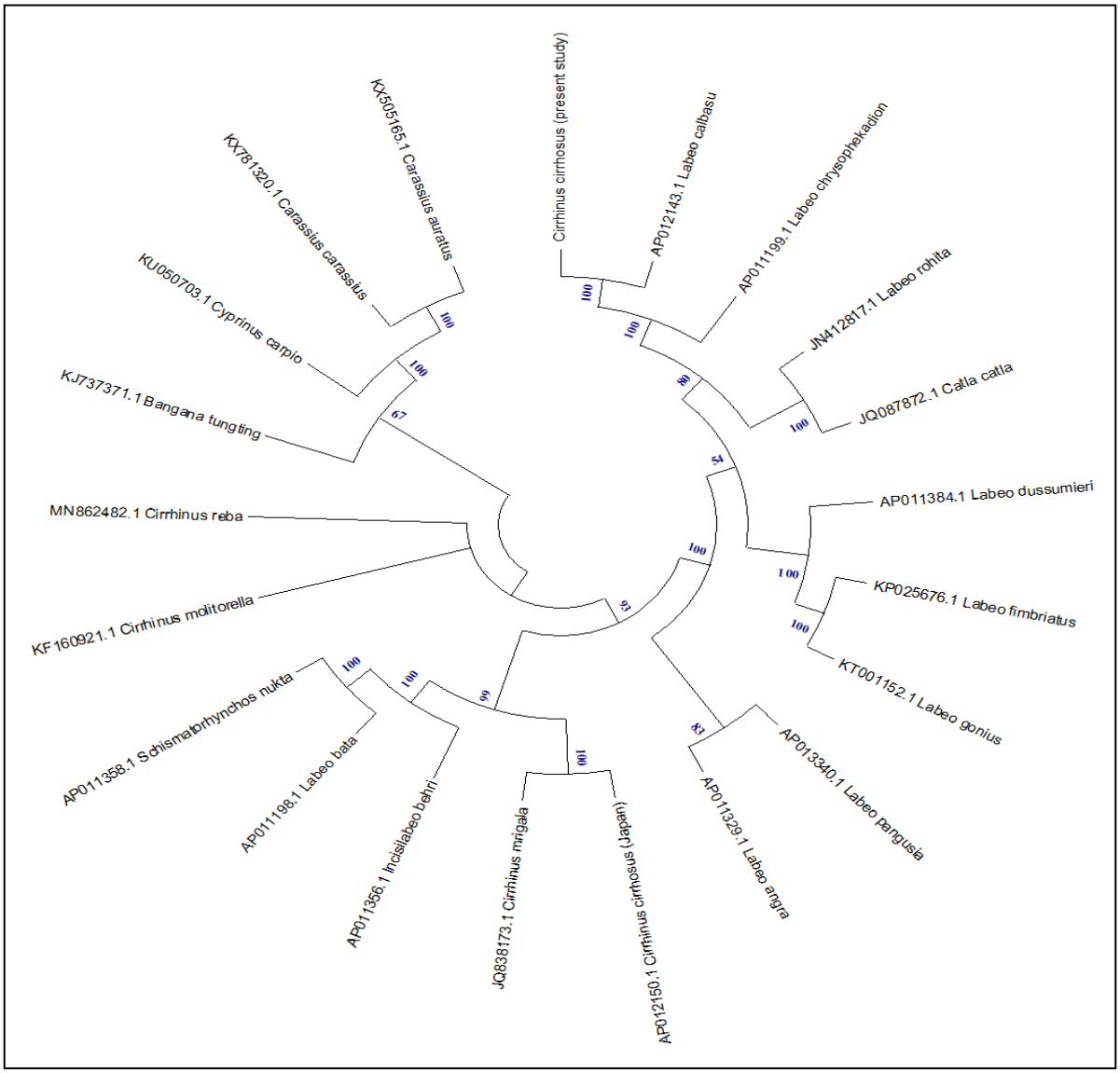


Fig. 5. The Phylogenetic circle tree of *Cirrhinus cirrhosus* and 20 Cyprinids based on whole mitochondrial genome using Maximum Likelihood method.

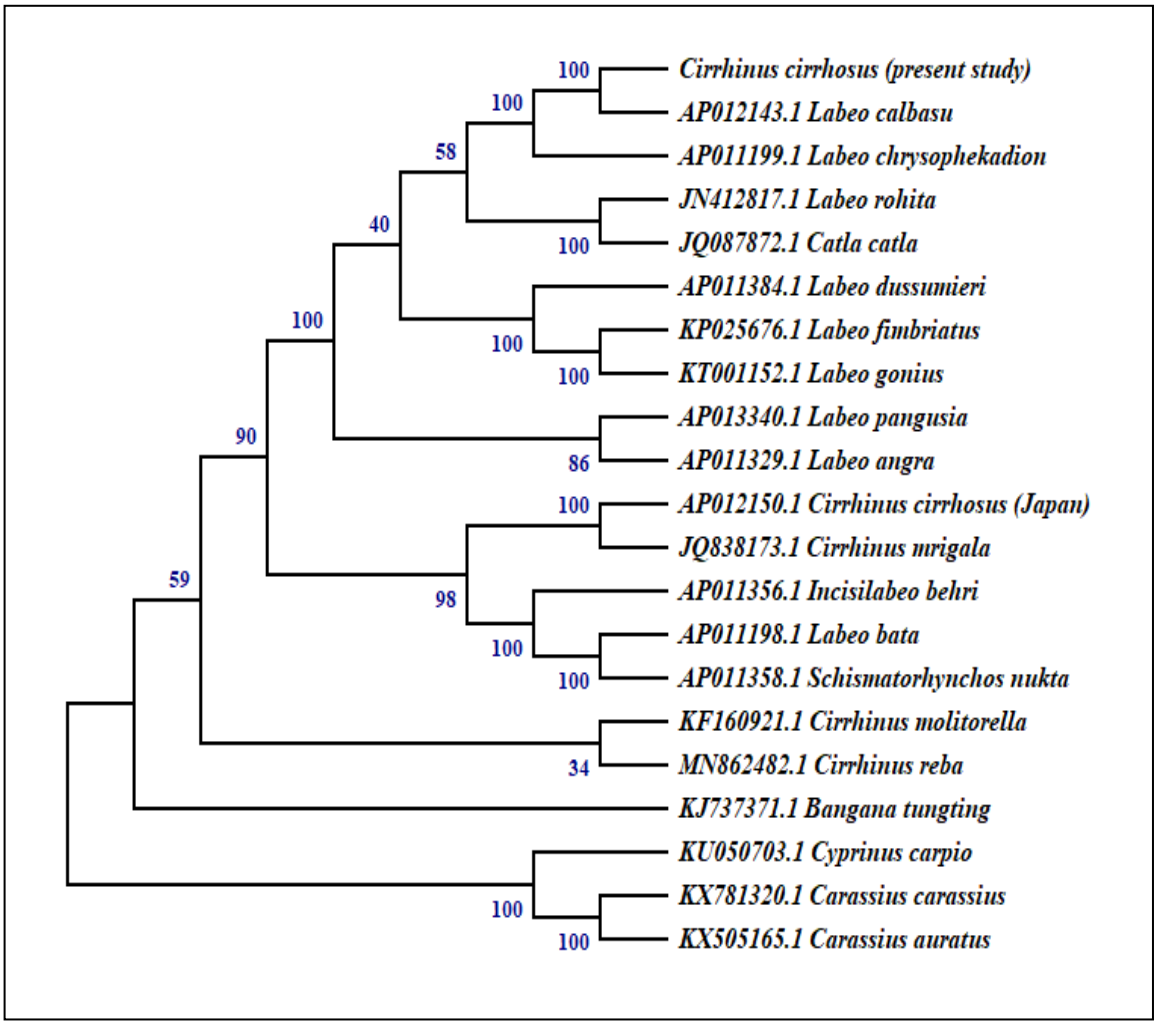


Fig.6. : The Phylogenetic traditional tree of *Cirrhinus cirrhosus* and 20 Cyprinids based on whole mitochondrial genome using Neighbor-Joining method.

Reference

Akter A. and Ali M.H., 2012. Environmental flow requirements assessment in the Halda River, Bangladesh. Hydrological sciences journal. Feb 1;57(2):326-43.

Akhtar A., Islam M.T., Islam M.S., Mia M.M., Bhuyan M.S., Kibria M.M., Sharif A.S., and Kamal A.H., 2017. Risk and coping mechanisms of the carp spawn fishing community of the Halda river, Bangladesh. Bangladesh Journal of Zoology. 45(1):85-96.

Alok, D., N.S. Haque and J.S. Killan. 1995. Transfer of human growth hormone gene into Indian major carp (*L. rohita*) Int.J.Anim.Sci. 10: 5–8.

Azadi, M.A., 1985. Spawning of commercial freshwater fish and brackish and marine water shrimps of Bangladesh. Bangladesh Fisheries Information Bulletin, 2 (2), 1 – 74.

Azadi M.A. and Alam M.A.U., 2013. Ichthyofauna of the river Halda, Chittagong, Bangladesh J.Zool., 41(2013), pp. 113-133.

Askokbhai B.N., Sharma B. and Shah T., 2016. Age, growth and harvestable size of *Cirrhinus mrigala* (HAM.) from the lake Picchola, Udaipur, India. J. Exp. Zool. India Vol. 19, No. 2, pp. 000-000.

Achakzai W.M., Mohammad S.W., Baloch W.A.,Shivastava N and Samroo A.N., 2015. Length-Weight relationship and Condition factor of C. mrigala from Manchar lake, Sindh, Pakistan. Proc.Zool.Soc.India. 14(1) :57-63.

Andrews S., 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data.

Aparicio S., Chapman J., Stupka E., Putnam N., Chia J.M., Dehal P., Christoffels A., Rash S., Hoon S., Smit A., Gelpke M.D.S., Roach J., Oh T., Ho I.Y., Wong M., Detter C., Verhoe F., Predki P., Tay A., Lucas S., Richardson P., Clark M.S., Edward Y.J.K., Doggett N., Zharkikh A., Tavtigian S.V., Pruss D., Barnstead M., Evans C., Powell J., Glusman G., Rowen L., Hood L., Tan Y.H., Elgar G., Hawkins T., Venkatesh B., Rokhsar D. and Brenner, S.(2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science, 297(5585), 1301–1310.

Austin Christopher M., Tan Mun Hua, Harrisson Katherine A., Lee Yin Peng, Croft Laurence J., Sunnucks Paul, Pavlova Alexandra and Gan Han Ming, 2017. De novo genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (Maccullochella peelii), from Illumina and Nanopore sequencing read.GigaScience, 6, 2017, 1–6.doi: 10.1093/gigascience/gix063

Avise, J.C., 1995. Mitochondrial DNA polymorphism and a connection between genetics and demography of relevance to conservation. Conserv Biol 9:686–690

Avise, J.C., 2000. Phylogeography – the History and Formation of Species. Harvard University Press, USA, 447

Avise J.C. 2009. Phylogeography: retrospect and prospect. Journal of Biogeography. 36: 3–15.

Bachtrog D., 2007. Reduced selection for codon usage bias in *Drosophila miranda*. J Mol Evol 64:586–590

Bej D., Sahoo L., Das S.P., Swain S., Jayasankar P. and Das P., 2012. Complete mitochondrial genome sequence of *Cirrhinus mrigala* (Hamilton, 1822). Mitochondr DNA 24:91--3

Behera B.K., Kunal S.P., Paria P., Das P., Meena D.K., Pakrashi S., Sahoo A.K., Panda D., Jena J. and Sharma A.P., 2015. Genetic differentiation in Indian Major Carp, Cirrhinus mrigala from Indian Rivers, as revealed by direct sequencing analysis of Mitochondrial Cytochrome b gene. Mitochondrial DNA 26 (3), 334-336, 2015.

Bernt M., Donath A., Jühling F., Externbrink F., Florentz C., Fritzsch G., Pütz J., Middendorf M. and Stadler P.F., 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. *Molecular phylogenetics and evolution* , 69 (2), pp.313-319.

Boore, J.L., 1999. Animal mitochondrial genome. Nucleic Acids Res 27:1767–1780.

Boore, J.L., 2006. The use of genome-level characters for phylogenetic reconstruction. Trends Ecol Evol.; 21: 439–446. doi: 10.1016/j.tree.2006.05.009 PMID: 16762445.

Burger G, Gray MW, Lang BF: Mitochondrial genomes: anything goes. Trends Genet 2003, 19:709–716.

Braasch, I., Gehrke, A. R., Smith, J. J., Kawasaki, K., Manousaki, T., Pasquier, J., Amores, A., Desvignes, T., Batzel, P., Catchen, J., Berlin, A. M., Campbell, M. S., Barrell, D., Martin, K. J., Mulley, J. F., Ravi, V., Lee, A. P., Nakamura, T., Chalopin, D., Fan, S., Wcisel, D., Cañestro, C., Sydes, J.,

Beaudry, F. E. G., Sun, Y., Hertel, J., Beam, M. J., Fasold, M., Ishiyama, M., Johnson, J., Kehr, S., Lara, M., Letaw, J. H., Litman, G. W., Litman, R. T., Mikami, M., Ota, T., Saha, N. R., Williams, L., Stadler, P. F., Wang, H., Taylor, J. S., Fontenot, Q., Ferrara, A., Searle, S. M. J., Aken, B., Yandell, M., Schneider, I., Yoder, J. A., Volff, J.-N., Meyer, A., Amemiya, C. T., Venkatesh, B., Holland, P. W. H., Guiguen, Y., Bobe, J., Shubin, N. H., Di Palma, F., Alföldi, J., Lindblad-Toh, K., and Postlethwait, J. H. (2016). The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nature genetics, 48:427–437.

Cameron, S.L., 2014. Insect mitochondrial genomics: implications for evolution and phylogeny. Annu Rev Entomol.; 59: 95–117. doi: 10.1146/annurev-ento-011613-162007 PMID: 24160435.

Campbell, M. S., Holt, C., Moore, B., & Yandell, M., 2014. Genome annotation and curation using MAKER and MAKER-P. Current protocols in bioinformatics, 48(1), 4-11.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A.S., Yandell, M., 2008. MAKER: an easy-to-use annotation pipeline designedfor emerging model organism genomes. Genome Res18:188–196

Chan, P.P., Lin, B.Y., Mak, A.J. and Lowe, T.M., 2019. tRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. *bioRxiv* 614032

Chauhan T., Lal K. K., Mohindra V., Singh R.K., Punia P., Gopalakrishnan A., Sharma P. C., and Lakra W. S., 2007. Evaluating genetic differentiation in the wild populations of the Indian major carp *Cirrhinus mrigala* (Hamilton-Buchanana, 1882): Evidence from allozyme and microsatellite markers. Aquaculture 269 (2007) 135 – 149.

Chen K, Xiao D, Wen Z, Lin S, Kuang G. 2004. A comparative study on the biological characteristics of Cirrhinus mrigala and Labco rohita. Inland Fisheries. (6):37–38. (In Chinese)

Cheng Y. Z., Xu T. J., Shi G. and Wang R. X., 2010. Complete mitochondrial genome of the miiuy croaker *Miichthys miiuy* (Perciformes, Sciaenidae) with phylogenetic consideration. Mar. Genomics 3, 201–209.

Chikhi R., Medvedev P., "Informed and automated k-mer size selection for genome assembly." Bioinformatics. 2014 Jan 1;30(1):31-7. 2013 Jun 3. https://doi.org/10.1093/bioinformatics/btt310

Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011;29: 987–91. doi:10.1038/nbt.2023

Curole, J.P., Kocher, T.D., 1999. Mitogenomics: digging deeper with complete mitochondrial genomes. Trends Ecol Evol 14:394–398

Das, A., Lanakiev, P., *et al*., 2018. Genome of *Tenualosa ilisha* from the river Padma, Bangladesh. BMC Res Notes 11:921. https://doi.org/10.1186/s13104-018-4028-8

Das P., Sahoo L., Das S.P., Bit A., Joshi C.G., Kushwaha B., et al., 2020. De novo assembly and genome-wide SNP discovery in rohu carp *Labeo rohita*. Front Genet. 11:386.

Das, S. P., Bej, D., Swain, S., Mishra, C. K., Sahoo, L., Jena, J., Jayasankar, P., and Das, P., 2013. Population divergence and structure of *Cirrhinus mrigala* from peninsular rivers of India, revealed by mitochondrial cytochrome b gene and truss morphometric analysis. Mitochondrial DNA, Early Online: 1–8. DOI: 10.3109/19401736.2013.792055.

Das S.P., Swain S., Jena J. and Das P., 2018. Gentic diversity and population structure of *Cirrhinus mrigala* revealed by mitochondrial ATPase 6 gene. Mitochondrial DNA Part A 29 (4), 495-500.

Das, S.P., Bej D., Swain S., Jena J.K. and Das P., 2012. Relative age and growth of Indian Major Carp, *Cirrhinus mrigala*, from peninsular rivers of India. J. Aqua., 20: 35-43.

Desai V. R. and Shrivastava N.P., 1990. Studies on Age, Growth and Gear Selectivity of *Cirrihinus Mrigala* (Hamilton) from Rihand Reservoir, Uttar Pradesh. Reservoir Fisheries Research Centre of CICFRI, Raipur - 492 007. Indian J. Fish., 37 (4): 305 - 311.

Dierckxsens N., Mardulyn P. and Smits G., 2017. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic acids research, 45 (4), pp.e18-e18.

Dierckxsens N, Mardulyn P, Smits G. Unraveling heteroplasmy patterns with NOVOPlasty *NAR Genomics and Bioinformatics, https://doi.org/10.1093/nargab/lqz011.*

Dowton M, Castro LR, Austin AD. Mitochondrial gene rearrangements as phylogenetic characters in the invertebrates: the examination of genome 'morphology'. Invertebr Syst. 2002; 16: 345–356.
Dwivedi A.K., Sarkar U.K., Mir J.I., Tomat P., and Vyas V., 2019. The Ganges basin fish *Cirrhinus mrigala* (Cypriniformes: Cyprinidae): detection of wild populations stock structure with landmark morphometry. Rev. Biol. Trop. (Int. J. Trop. Biol. ISSN-0034-7744) Vol. 67(3): 541-553, June.
Food and Agriculture Organization of the United Nations Fisheries and Aquaculture Department report for production of mrigal, http://www.fao.org/fishery/culturedspecies/Cirrhinusmrigala/en, 2015 (accessed 10.11.15).
FAO. FAO Yearbook, Fishery and Aquaculture Statistics 2010. Food and Agriculture Organization of the United Nations, Rome, Italy. 2012;78.
FAO, 2005. Aquaculture Production, 2004. Year book of Fishery Statistics - Vol. 96/2. FAO Rome, Italy, pp.1-7.
Felsenstein J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791.
Froese R. Length-weight relationships for 18 less studied fish species. J. Appl. Ichthyol. 1998; 14:117-118.

Gissi C, Iannelli F, Pesole G (2008) Evolution of the mitochon- drial genome of Metazoa as exemplified by comparison of congeneric species. Heredity 101:301–320.
Gurevich, A, Saveliev, V, Vyahhi, N, Tesler, G., QUAST: quality assessment toolfor genome assemblies.Bioinformatics. 2013; 29: 1072–1075.
Hasanat M.A., Mollah M.F.A., and Alam M.S., 2015. Microsatellite DNA Marker Analysis Revealed Low Levels of Genetic Variability in the Wild and Captive Populations of *Cirrhinus cirrhosus* (Hamilton) (Cyprinidae: Cypriniformes). BBJ, 5(4): 206-215, 2015;Article no.BBJ.2015.020.
Hora, S. L., and Pillay, T.V.R., 1962. Handbook on fish culture in the Indo-Pacific Region. FAO Fish Tech.Pap.14. pp.204.
Hossain, M.A.R., 2014. An overview of fisheries sector of Bangladesh. Res. Agric., Livest. Fish. 1(1); 109-126.
Hinaux H., Poulain J., Da Silva C., Noirot C., Jeffery W.R., *et al*. 2013. De Novo Sequencing of *Astyanax mexicanus* Surface Fish and Pachón Cavefish. Transcriptomes Reveals Enrichment of Mutations in Cavefish Putative Eye Genes. PLoS ONE 8(1): e53553. doi:10.1371/journal.pone.0053553
Hussain M.G. and Mazid M.A., 2001. Genetic improvement and conservation of carp species in Bangladesh. Bangladesh Fisheries Research Institute and International Centre for Living Aquatic Resources Management, 74pp.
Imran S.; Thakur S., Jha, D.N. and Dwivedi A.C., 2015. Size composition and exploitation pattern of *Labeo calbasu* (Hamilton 1822) from the lower stretch of the Yamuna river. Asian J. Bio. Sci., 10(2): 171-173. DOI : 10.15740/HAS/AJBS/ 10.2/171-173.
Iwasaki W., Fukunaga T., Isagozawa R., Yamada K., *et al*., 2013. MitoFish and MitoAnnotator: A Mitochondrial Genome Database of Fish with an Accurate and Automatic Annotation Pipeline. Mol Biol Evol. 30:2531-2540.
Johal, M. S. and Tandon, K. K. 1987. Age and growth of *Cirrhinus mrigala* (Pisces: Cyprinidae). Spolec Zoological., 51:252-280.
JENA, J. K., ARAVINDAK S., P.K. AND SINGH, W.J. 1998. Nursery rearing of Indian major carp fry under different stocking densities. Indian J. Fish., 45(2): 163- 168.
JHINGRAN, V.G. AND PULLIN, R.S.V., 1985. A Hatchery Manual of Chinese and Indian major carps. Asian Development Bank. Manila, Philippines. pp 1-18.
Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., ... & Pesseat, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics, 30(9), 1236-1240
Kabir, M.H., 2012. Economic valuation of Halda River in Chit- tagong. MS thesis, Institute of Forestry and Environmental Sciences, University of Chittagong, Bangladesh, 111.
Kabir, H., Kibria, M., Jashimuddin, M. and Hossain, M. M., 2015. Conservation of a river for biodiversity

and ecosystem services: the case of the Halda – the unique river of Chittagong, Bangladesh. International Journal of River Basin Management. http://dx.doi.org/10.1080/15715124.2015.1012514

Kabir, H., Kibria, M., Jashimuddin, M. and Hossain, M. M., 2013. Economic Valuation of Tangible Resources From Halda– The Carp Spawning Unique River Located at Southern Part of Bangladesh. International Journal of Water Research 2013; 1(2): 30-36

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, Kohara Y, Fujiyama A, Hayashi T, Itoh T, "Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads". Genome Res. 2014 Aug;24(8):1384-95. doi: 10.1101/gr.170720.113.

Karim, A., Saif R., Faisal, S.A., Gil, Z., Ali, W., 2018. Use of CO1 gene sequences for computing genetic diversity between *Cirrhinus mrigala* from two different habitats (Farm and River). JFLS | Vol 3(2) | Pp 54-57.

Kawaguchi A, Miya M, Nishida M (2001) Complete mitochondrial DNA sequence of *Aulopus japonicus* (Teleostei: Aulopiformes), a basal Eurypterygii: longer DNA sequences and higher-level relationships. Ichthyol Res 48:213–223

Kelley, J. L., Brown, A. P., Therkildsen, N. O., and Foote, A. D. (2016). The life aquatic: advances in marine vertebrate genomics. Nat. Rev. Genet. 17, 523–534.

Khan, H. A. and Jhingran, V. G. 1975. Synopsis of biological data on the Rohu, *Labeo rohita* (Hamilton, 1822). FAO Fish Synopsis, p. 72.

Khatun, N., Islam, M., Sultana, N., Mrong, S., and Huq, M. A., 2017. Present status of carp hatchery and breeding operations in Bangladesh: A review. *Research in Agriculture Livestock and Fisheries*, *4*(2), 123-129. https://doi.org/10.3329/ralf.v4i2.33724.

Kibria, M.M., Farid, I., and Ali, M., 2009. Halda restoration project: peoples expectation and reality. Nasirabad, Chittagong: Chittagram, 57.

Kimura M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*16:111-120.

Koutrakis E.T., Tsikliras A.C., Length-weight relationships of fishes from three northern Aegean estuarine systems (Greece). J. Appl. Ichthyol. 2003; 19:258-260.

Kumar S., Stecher G., Li M., Knyaz C., Tamura K., 2018 MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 35(6):1547-1549.

Lal K. K., Chauhan T., Mandal A., Singh R.K., Khulbe L., Ponniah A.G. and Mohindra V., 2004. Identification of microsatellite DNA markers for population structure analysis in Indian major carp, Cirrhinus mrigala (Hamilton-Buchanan, 1882). J. Appl. Ichthyol. 20, 87–91.

Laslett, D. and Canback, B., 2008. ARWEN: A program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics 24:172–5.

Le Cren E D (1951) The length-weight relationship and seasonal cycle in gonadal weight and condition in the Perch (Percafluviatilis). J. Animal Ecol. 20, 201-219.

Lee W. J. and Kocher T. D. 1995 Complete sequence of a sea lamprey (*Pettromyzon marinus*) mitochondrial genome: early establishment of the vertebrate genome organization. Genetics 139, 873–887.

Li, H., 2015. BFC: correcting Illumina sequencing errors. *Bioinformatics*, *31*(17), 2885-2887.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al.2012.SOAPdenovo2: an empirically improved memory-efficient short-readde novoassembler.Gigascience1:1–18.

Manna, G.K. and Prasad R., 1971. A new perspective in the mechanism of evolution of chromosomes in fishes. J. cytol. Genet. 239–240. Proc. 1st All India Cong. Cytol. & Genti.

Menon A.G.K., 2004.Threatened Fishes of India and Their Conservation. Zoological Survey of India, Kolkata, 170pp.

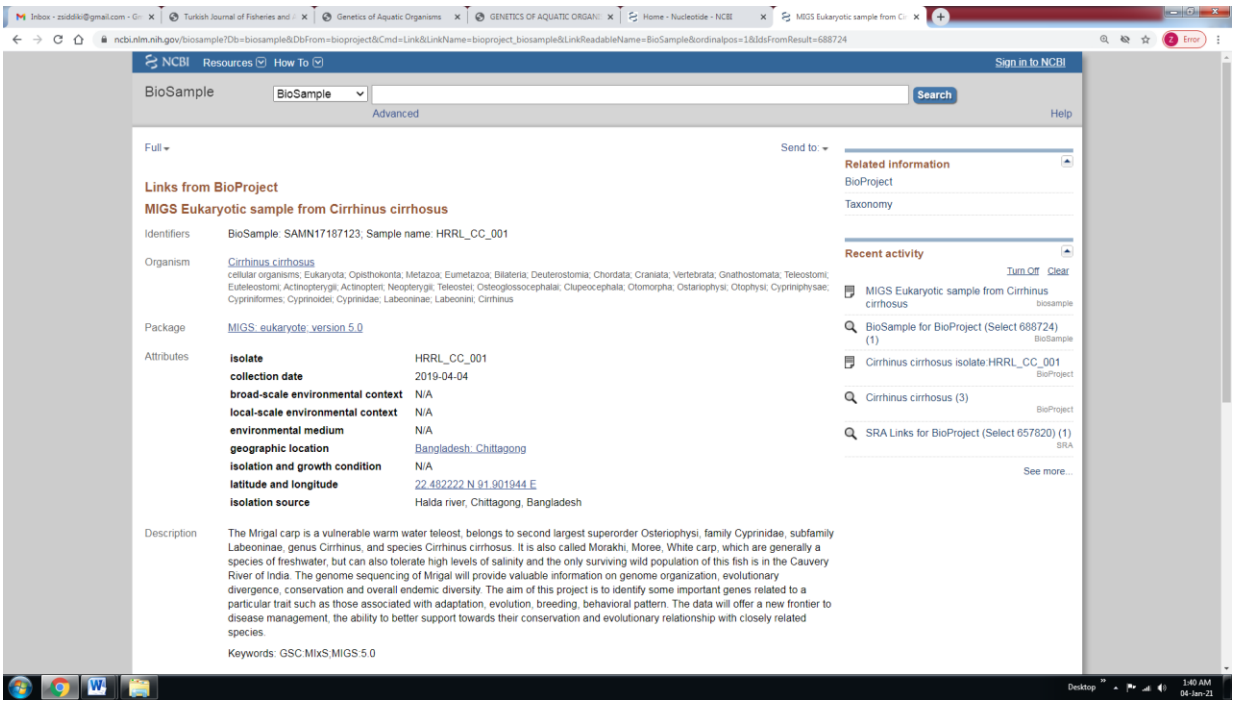
Majumdar K. and Chaudhuri S.P.R., 1976. Studies on the chromosomes of Indian major carps. Proc. Sym. "Modern treads in Zoological Researches in India" Calcutta, July 26–27. Abs. 68–69.

Majumdar K.C., Ravinder K. and Nasaruddin K., 1997. DNA fingerprinting in Indian major carps and

tilapia by Bkm 2 (8) and M 13 probes. *Aquaculture Research* 28: 129–138.
Martin-Smith, K.M., 1996. Length-weight relationships of fishes in a diverse tropical freshwater community, Sabah, Malaysia. J. Fish Biol. 1996; 49:731-734.
Masta SE. Mitochondrial rRNA secondary structures and genome arrangements distinguish chelicerates: comparisons with a harvestman (Arachnida: Opiliones: Phalangium opilio). Gene. 2010; 449: 921. doi: 10.1016/j.gene.2009.09.009 PMID: 19800399.
Mayank P., Tyagi R.K. and Dwivedi A.C., 2015. Studies on age, growth and age composition of commercially important fish species, *Cirrhinus mrigala* (Hamilton, 1822) from the tributary of the Ganga river, India. Euro. J. Exp. Bio., 2015, 5(2):16-21.
Mayank P., Tiwari A. and Dwivedi A.C., 2016. Reproductive profile of Cirrhinus mrigala and suggestion for restoration (Hamilton, 1822) from the Yamuna river, India. Bioved, 27(1) : 115–120.
Li, D.; Liu, C.M.; Luo, R.; Sadakane, K.; Lam, T.W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 2015, 31, 1674–1676.
Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., *et al*. InterPro, progress and status in 2005. Nucleic Acids Res. 2005;33:D201–D205.
Nagarajan N, Cook C, Di Bonaventura M, Ge H, Richards A, Bishop-Lilly KA, et al., 2010. Finishing genomes with limited resources: Lessons from an ensemble of microbial genomes. BMC Genomics. 242.
Nagarajan N, Pop M., 2013. Sequence assembly demystified. Nat Rev Genet. Nature Publishing Group;14: 157–167. doi:10.1038/nrg3367
Padhi, B.K. and R.K. Mandal. 1993. Rapid isolation of mitochondrial DNA from fish. *Indian Journal of Experimental Biology* Vol. 31: 790–792.
Padhi, B.K. and R.K. Mandal. 1995. Genetic resource mapping and fisheries management current science, Vol. 68 (5): 490–493.
Padhi B.K., Ghosh S.K. and Mandal R.K., 1998. Characterization of Mbol satellites in *Cirrhinus mrigala* and *Clarias batrachus* (Pisces). Genome, Vol. 41: 34–39.
Patra, R.W.R. and Azadi, M.A., 1985. Hydrological conditions influencing the spawning of major carps in the Halda River, Chittagong, Bangladesh. Bangladesh Journal of Zoology, 13 (1), 63– 72.
Patel A, Das P, Barat A and Sarangi N, 2009. Estimation of genome size in Indian major carps Labeo rohita (Hamilton), Catla catla (Hamilton), Cirrhinus mrigala (Hamilton) and Labeo calbasu (Hamilton) by Feulgen microdensitometry method. Indian J. Fish., 56(1) : 65-67.
Petrakis G, Stergiou KI. Weightlength relationships for33 fish species in Greek waters. Fish. Res. 1995; 21:465-469.
Powel, A. B., 1981. Annulus formation of otoliths and growth of young summer flounder from Pambier Sound, North Carolina. Trans. Am. Fish. Soc., 111: 688-693.
Pathak, R.K.; Gopesh, A.; Dwivedi, A.C. and Joshi, K.D., 2014. Age and growth of alien fish species, Cyprinus carpio var. communis (Common carp) in the lower stretch of the Yamuna river at Allahabad. Nat. Acad. Sci. Lett., 37(5): 419-422. DOI: 10.1007/s40009-014-0262-3.
Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. 2005. InterProScan: protein domains identifier. Nucleic acids research, 33(Web Server issue), W116–W120. https://doi.org/10.1093/nar/gki442
Rahman A.K.A., 2005. Freshwater fishes of Bangladesh. 2nd ed., Zool. Soc. Bangladesh, Dhaka, Bangladesh. 394 pp.
Rahman M.M. and Balcombe S.R., 2016. Competitive interactions under experimental conditions affect diel feeding of two common aquaculture fish species Labeo calbasu and Cirrhinus cirrhosis of Southern Asia. https://doi.org/10.1111/jai.13157.
Ravi, V., and Venkatesh, B. (2018).The divergent genomes of teleosts. Annu. Rev. Anim. Biosci. 6, 47–68. doi: 10.1146/annurev-animal-030117-014821
Rema Devi K.R. & Ali A., 2013. *Cirrhinus cirrhosus*. The IUCN Red List of Threatened Species 2013: e.T166531A6230103. http://dx.doi.org/10.2305/IUCN.UK.2011-1.RLTS.T166531A6230103.en

Sanger F, Thompson EOP. The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. Biochem J 1953;53:366–74.
Sanger F, Thompson EOP. The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem J 1953;53:353–66.
Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 1975;94:441–8.
Sahoo P.K., Mohanty B.R., Kumari J., Barat A. and Sarangi N., 2008. Cloning, nucleotide sequence and phylogenetic analyses, and tissue-specific expression of the transferrin gene in Cirrhinus mrigala infected with Aeromonas hydrophila. Comp. Immun. Microbiol. Infect. Dis. 32 527–537.
Sahoo L., Das P., Sahoo B. *et al.,* 2020. The draft genome of *Labeo catla*. *BMC Res Notes* 13, 411.
Saitou N. and Nei M., 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.
Sarang N., Sharma L.L., and Saini V.P., 2009. Age, growth and harvestable size of Cirrhinus mrigala from the Jawahar Sagar Dam, Rajasthan, India. Indian J. Fish., 56(3) : 215-218.
Sarder M.R.I., Rafiquzzaman S.M., Sultana R. and Islam M.F., 2009. Cryopreservation of spermatozoa of Mrigal, *Cirrhinus cirrhosus* with a view to minimize inbreeding and hybridization. Journal of the Bangladesh Agricultural University, 7(1), 211-218.
Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M., (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210-3212.a
Simakov, O., Marletaz, F., Cho, S.-J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo,D.-H., Larsson, T., Lv, J., Arendt, D., Savage, R., Osoegawa, K., de Jong, P., Grimwood,J., Chapman, J. A., Shapiro, H., Aerts, A., Otillar, R. P., Terry, A. Y., Boore, J. L., Grigoriev,I. V., Lindberg, D. R., Seaver, E. C., Weisblat, D. A., Putnam, N. H., and Rokhsar, D. S.(2013). Insights into bilaterian evolution from three spiralian genomes. Nature, 493:526–531.
Simon C, Buckley TR, Frati F, Stewart JB, Beckenbach AT. Incorporating molecular evolution into phylogenetic analysis, and a new compilation of conserved polymerase chain reaction primers for animal mitochondrial DNA. Annu Rev Ecol Evol Syst. 2006; 37: 545–579. doi: 10.1146/annurev.ecolsys.37.091305.110018
Shadel, G.S. and Clayton, D.A., 1997. Mitochondrial DNA maintenance in vertebrates. Annu Rev Biochem 66:409–435.
Sharker, R. and Siddik, M.A.B., 2015. Genetic differentiation of wild and hatchery populations of Indian major carp *Cirrhinus cirrhosus* in Bangladesh. Journal of Environmental Biology 36(5):1223-1227.
Shi G., Jin X. X., Zhao S. L., Xu T. J. and Wang R. X. 2012. Complete mitochondrial genome of *Trypauchen vagina* (Perciformes, Gobioidei). Mitochondrial DNA 23, 151–153.
Shieh Y.K., Liu S.C. and Lu L., 2020. Scaffolding Contigs Using Multiple Reference Genomes. DOI: 10.5772/intechopen.93456.
Simison WB, Boore JL: Molluscan evolutionary genomics. In Phylogeny and evolution of the mollusca. Edited by Ponder W, Lindberg DR. Berkeley: University of California Press; 2008:447–461.
Staden R., A strategy of DNA sequencing employing computer programs. Nucleic Acids Res 1979;6:2601–10.
Stothard P., 2000. The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28:1102-1104
Talwar P.K. and Jhingran A.G., 1991. Inland fisheries of India and adjacent countries, Vol. 2.
Tamura K., Nei M., and Kumar S., 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences* (*USA*) 101:11030-11035.
Tandon K. K., and Johal M.S., 1996. Age and Growth in Indian Freshwater Fishes. Narendra Publishing House, New Delhi, (Reprint 2006), pp 378.
Tzeng C. S., Hui C. F., Shen S. C. and Huang P. C., 1992. The complete nucleotide sequence of the

Crossostoma lacustre mitochondrial genome: conservation and variation among vertebrates. Nucleic Acids Res. 20, 4853–4858.
Tillich M., Lehwark P., Pellizzer T., Ulbricht-Jones E.S., Fischer A., Bock R., Greiner S. 2017. GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 45(W1):W6–W11.
Törönen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation web server. Nucleic Acids Res. 2018 Jul 2;46(W1):W84-W88. doi: 10.1093/nar/gky350. PMID: 29741643; PMCID: PMC6031051.
Tsai, C., Islam M.N., Karim M.R., and Rahman K.U.M.S., 1981. Spawning of major carps in the lower Halda River, Bangladesh. Estuaries, 4 (2), 127– 138.
Ujjania N.C. and Soni N., 2018. Harvestable Size of Indian Major Carp (Cirrhinus mrigala, Ham. 1822) in Vallabhsagar Reservoir, Gujarat (India). IJBS: 5(1): 71-73, June.
Valentini, A., Pompanon, F., Taberlet, P., 2009. DNA barcoding for ecologists. Trends Ecol Evol 24:110–117
Wang X. Z., Wang J., He S. P. and Mayden R. L. 2007 The complete mitochondrial genome of the Chinese hook snout carp *Opsariichthys bidens* (Actinopterygii: Cypriniformes) and an alternative pattern of mitogenomic evolution in vertebrate. Gene 399, 11–19.
Ward, R.D., Grewe, P.M., 1994. Appraisal of molecular genetic techniques in fisheries. In: Pitcher, T.J. (Ed.), Molecular Genetics in Fisheries. Chapman & Hall, UK, pp. 29–54.
Watson JD, Crick FHC. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 1953;171:737–8.
Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., and Zdobnov, E. M., 2017. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. 35(3):543–548. doi:10.1093/molbev/msx319
Wu P., Chen D., Guo X. and Chu W., 2016. complete Mitochondrial Genome Sequence of Cirrhinus mrigala (♀) × Labeo rohita (♂). Mitochondrial DNA Part a 27 (3).
Xia J., Xia K., Gong J. and Jiang S., 2007. Complete mitochondrial DNA sequence, gene organization and genetic variation of control regions in *Parargyrops edita*. Fish Sci. 73, 1042-1049.
Xu, P., Zhang, X., Wang, X. *et al.* Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nat Genet* 46, 1212–1219 (2014).
Yandell, M. and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics, 13(5):329–342.
Yuan ML, Wei DD, Wang BJ, Dou W, Wang JJ. The complete mitochondrial genome of the citrus redmite *Panonychus citri* (Acari: Tetranychidae): high genome rearrangement and extremely truncated tRNAs. BMC Genomics. 2010; 11: 597. doi: 10.1186/1471-2164-11-597 PMID: 20969792
Zdobnov, E.M., Apweiler, R., 2001. InterProScan—an integration platform for the signature recognition methods in InterPro. Bioinformatics. 2001;17:847–848.
Zhang, S.M. and P.V.G.K. Reddy. 1991. On the comparative karyomorphology of three Indian major carps, *Catla catla* (Ham.), *Labeo rohita* (Ham.) and *Cirrhinus mrigala* (Ham.) Aquaculture. 97: 7–12.
Zhang, D., Guo, H., Zhu, et al., 2013. The complete mitochondrial genome of mud carp *Cirrhinus molitorella* (Cypriniformes, Cyprinidae). Mitochondrial DNA, 2015; 26(1): 149–150. http://informahealthcare.com/mdn.
Zhai, D., Cha, J., Yu, J., Wang, Y., Chen, Y., Xia, M., Liu, H. and Xion, F., 2020. Complete mitochondrial genome characterization of *Cirrhinus. mrigala* from the Three Gorges Reservoir in China. Mitochondrial DNA part B 5(2), 1500-1501.
Zhong Q., Xu W., Wu Y. and Xu H., 2012. Patterns of synonymous codon usage on human metapneumovirus and its influencing factors. J Biomed Biotechnol. 2012;2012:460837. doi: 10.1155/2012/460837.
Zhou, Y., Zhang, J.Y., Zheng, R.Q., Yu, B.G., Yang, G., 2009. Complete nucleotide sequence and gene organization of the mitochondrial genome of *Paa spinosa* (Anura: Ranoidae). Gene 447:86–9



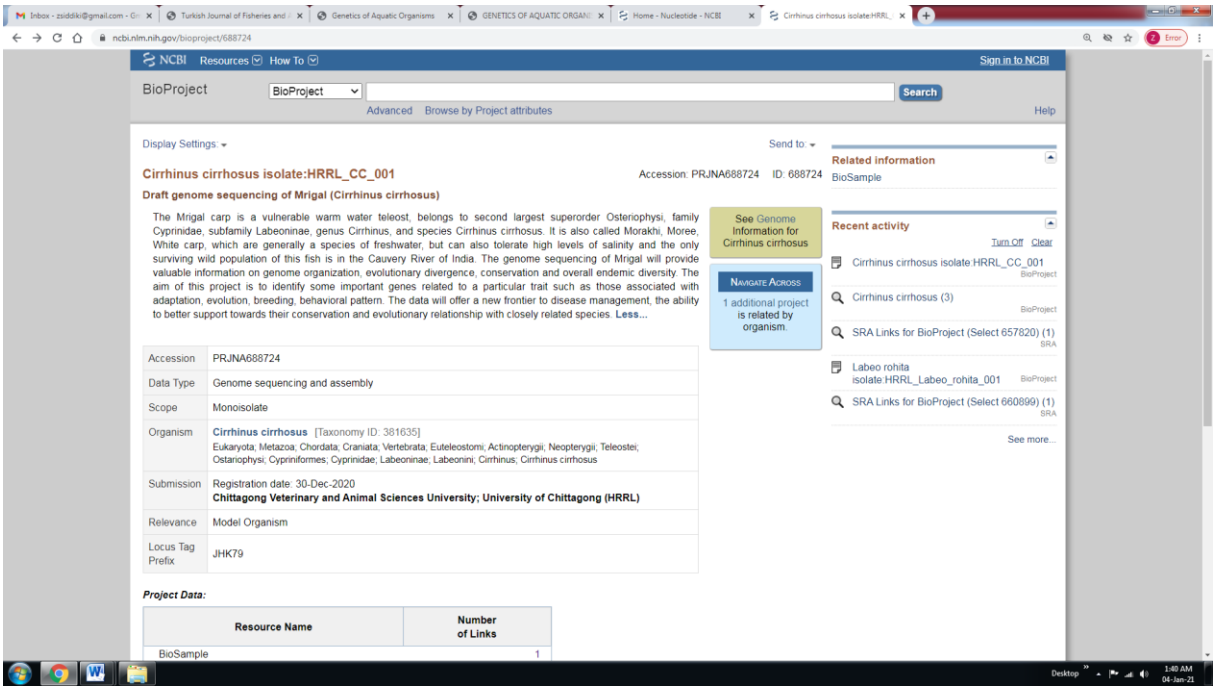


Fig. Screenshot of NCBI SRA datasets of Mrigel genome

Research Objective 7

Whole genome sequencing and assembly of *Labeo calbasu* [Hamilton, 1822] from the Halda river of Bangladesh

Abstract

Kalibaus- a small-sized carp species belonging to the family Cyprinidae, subfamily labeoninae, genus Labeo, and species Labeo calbasu is an important IMCs that Inhabitat naturally in the Halda river, Bangladesh. Here in this study genome sequencing and assembly of *L. calbasu* is done which will provide valuable information on genome organization, evolutionary divergence, conservation and overall endemic diversity. A healthy male adult L. calbasu belonging to the river Halda of Chittagong, Bangladesh was captured and used for reference-based assembly. The whole genome sequences were assembled in 538835 contigs with a total length of 932.2 Mb and N50 is 15758. Genome annotations identified 18273 gene models using AUGUSTUS gene annotation tool. The Benchmarking Universal Single-Copy Orthologs (BUSCO) tools assessed 76.9% completeness with the datasets of Eukaryota and 66.1% completeness with the datasets of Actinopterygii of the assembled genome.

Introduction

Halda is an indigenous river of Bangladesh which is important for its outstanding feature of being the natural breeding ground of carp fishes. It accommodates the Indian major carps (IMCs) *Labeo rohita*, *Catla catla*, *Cirrhinus cirrhosus* and *Labeo calbasu* that occur naturally here (Podder et al. 2017). Halda River is also important for its contribution to the national economy of Bangladesh. It contributes about US$ 20.5 million to the national economy by its tangible resources (Kabir et al. 2013 (a)). The total indirect and non-use values of this River were about Tk. 29.50 million (Kabir et al. 2013 (b)). Its Kalibaus is important because of the wild variety. *L. calbasu* is widely distributed in Asian countries including Bangladesh, India, Myanmar, Nepal, Pakistan, South China and Thailand (Talwar and Jhingran 1991). Once Kalibaus was found in all-natural water bodies (Rahman 2005). But later the population has reduced greatly from natural sources due to the Reduction of food availability (Rahman et al., 2008). Indiscriminate fishing, habitat modification and other ecological changes (Rahman et al. 2012) and its production dropped by about 20% during the period 2001 to 2009 in beel fishery (FRSS 2009). However, The fish is extensively used in stocking culture ponds. Hence, the species is assessed as Least Concern (IUCN, 2015). *L. calbasu* is a valuable food fish and also used as game fish in several places of the Indian subcontinent (Talwar and Jhingran, 1991, Rahman, 2005: Mishra and Saksena, 2012). Its liver oil contains Vitamin-A (Ghosh et al., 1993). Also, it provides 16.47% of protein and 2.65% of lipid (Ahmed et al.,

2012). So the demand for this fish in the local market is huge. In spite of such importance, there are no records of genomic information of this species. So, in this study, genome sequencing, assembly and annotation is done. The aim of this study is to develop draft information of *L. calbasu*'s genome that will help in future to identify some important genes related to a particular trait such as those associated with adaptation, muscle strength or prolificacy. The data will help to explore the evolutionary relationships with closely related species.

Methodology of the Study:

Sampling and DNA isolation:

An adult female of *Labeo calbasu* was collected from Halda river, Chattogram (geographic coordinate: 22.50018759362011 N 91.8654045869928 E). Blood was collected from the fresh specimen and preserved in an anticoagulant solution. Later the sample was sent to BGI, China where a high molecular weight genomic DNA was isolated and purified using the conventional blood DNA extraction kit for future evaluation of the quality and quantity of the DNA.

Table 1: Some information about the collected sample

Serial	Information Type	Information Values
1	collection date	07/03/2021
2	sex	Female
3	Age	1 year
4	Total Lenght	26.5 cm
5	Weight	230 gm

Library preparation

The extracted DNA was cleaned and sent for both library preparation and whole-genome sequencing (WGS) at the BGI genomics, China. Using Next-generation sequencing (NGS) technology on an IlluminaNovaSeq 6000 platform (Meyer et al. 2010) a total of 92.3 Gigabase pair (Gb) of subread bases with a read length of 150 bp were induced. After sequencing quality of primary sequence reads and trimmed sequencing reads were investigated using FastQC ver 0.11.9 (Andrews, S. 2010). The quality control of the reads

was done including removing adaptor sequences, low-quality reads and contamination from raw reads using BFC ver r181 (Li H. 2015).

Genome assembly:

To assemble the *Labeo calbasu* genome we used ABySS ver. 2.3.1 (Jackman et al. 2017), Platanus ver.1.2.4 (Kajitani et al. 2014), Soapdenovo2 ver. 2.40 (Luo et al. 2012) and MEGAHIT ver 1.2.9 (Li, D et al. 2015) assembler. Since there is currently no de novo assembler assured to outperform others and as assemblers overall performance can differ relying on the dataset, three unique assemblers were used and to determine the best assembler an assembly evolution was subsequently performed. All of the assemblers follow the classic De Bruijn graph illustration even though the assembly algorithm differs across methods. Finally, Busco ver.5.1.3 (Seppey et al. 2019) was operated to check the completeness of genes.

Genome Annotation:

AUGUSTUS ver. 3.4.0 (Stanke et al. 2005) was used for structural gene prediction. Web tools PANNZER2 (Törönen et al. 2018) was used to identify the protein annotation and Gene Ontology (GO) terms.

Result and discussion:

Fastqc Result:

The quality of the raw trim data was checked by FastQC ver 0.11.9 (Andrews, S. 2010). The report indicates the best quality to do the further assembly.

Measure	Value
Filename	r1.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	307620120
Sequences flagged as poor quality	0
Sequence length	150
%GC	37

Measure	Value
Filename	r2.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	307620120
Sequences flagged as poor quality	0
Sequence length	150
%GC	37

Fig.1 : Basic statistics of both raw trim data

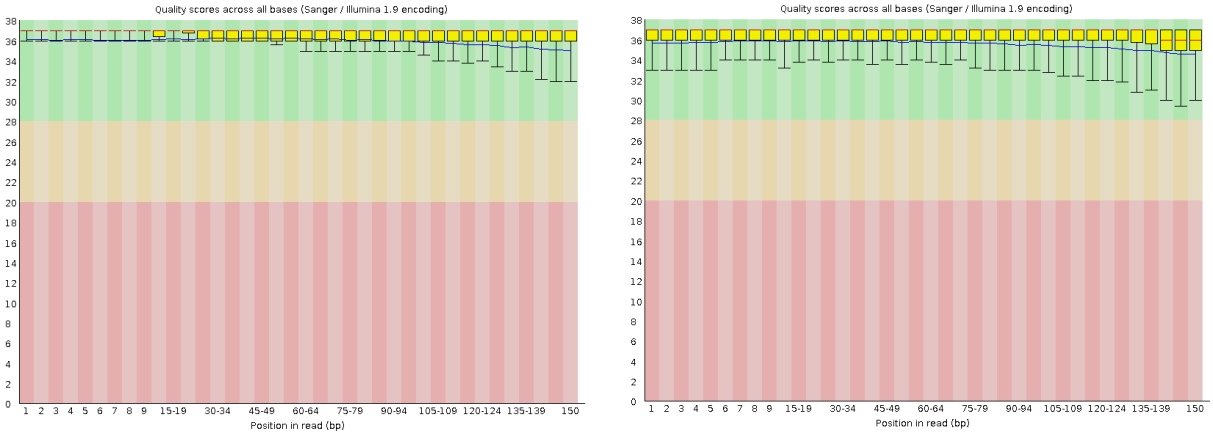


Figure 2: Per base sequence quality data of read 1 and read 2.

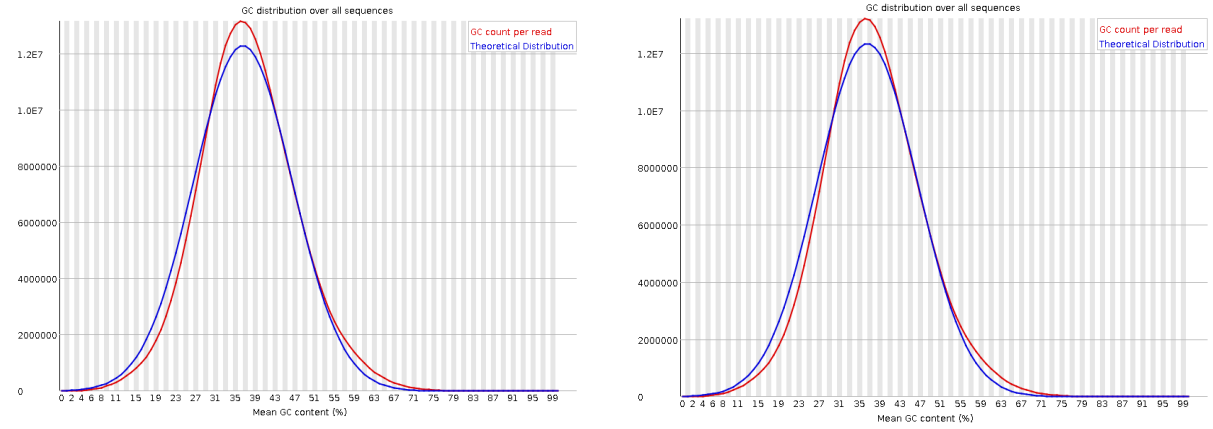


Figure 3: Per sequence GC content data of read 1 and read 2.

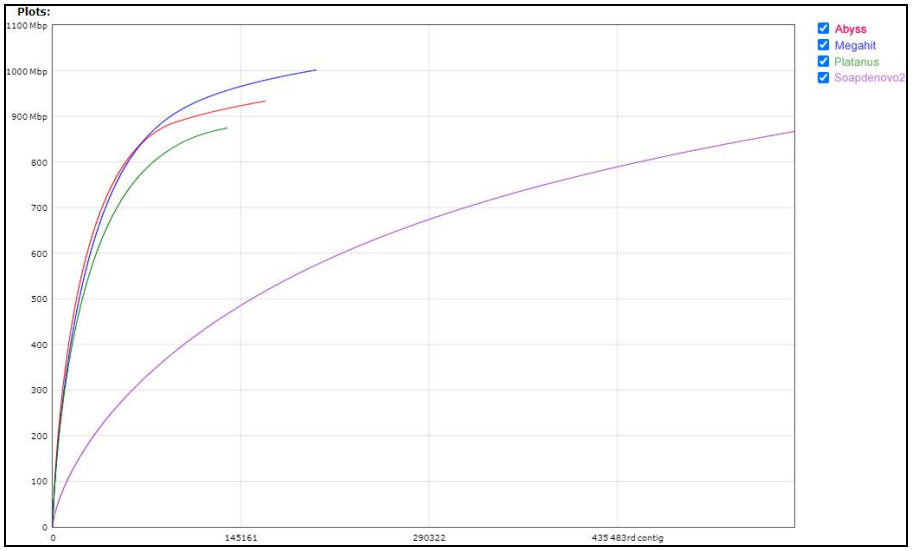
Assembler result:

A total of 932.2 Mb assembly is generated through ABySS ver. 2.3.1 (Jackman et al. 2017) assembler where 870.5 Mb, 866.8 and 1002 are from Platanus ver.1.2.4 (Kajitani et al. 2014), soapdenovo2 ver. 2.40 (Luo et al. 2012) and Megahit ver 1.2.9 (Li, D et al. 2015) respectively. The number of contigs is 538835 in Abyss, 2508213 in Platanus, 3702254 in Soapdenovo2 and 642264 in Megahit.

Table 2: Comparative assembly data

Values	Abyss	Platanus	Soapdenovo2	Megahit
Assembly Size (MB)	932.2	870.5	866.8	1002
Number of contigs (Number)	538835	2508213	3702254	642264
Min contig size (bp)	500	500	500	500
Max contig size (bp)	415821	271774	402511	521471
L50	16197	17746	117197	20567
N50	15758	13301	2007	13362

N50 of abyss data is 15758 in Abyss when 13301, 2007 and 13362 in Platanus, Soapdenovo2 and Megahit respectively. According to data, Abyss's result is the best one among.



Comparative Genome Assembly representation from Quast ver. 5.0.0 (Gurevich et al. 2013) Data. Contigs are ordered from largest (contig #1) to smallest.

Busco Result:

The Busco ver.5.1.3 (Seppey et al. 2019) analysis on Abyss assembly revealed 76.9% completeness with the datasets of Eukaryota and 66.1% completeness with the datasets of Actinopterygii (Table 3 and Table 4). A total of 255 groups of datasets were searched to find completeness with Eukaryota and 3640 groups of datasets were searched to find the completeness with Actinopterygii (Table 3 and Table 4).

Table 3: Results from generic domain eukaryota_odb10

Parameters	Number of Datasets	Percentages of Datasets
Complete BUSCOs (C)	196	76.9%
Complete and single-copy BUSCOs (S)	191	74.9%
Complete and duplicated BUSCOs (D)	5	2.0%
Fragmented BUSCOs (F)	52	20.4%
Missing BUSCOs (M)	7	2.7%
Total BUSCO groups searched	255	100%

Table 4: Results from dataset actinopterygii_odb10

Parameters	Number of Groups	Percentages of Groups
Complete BUSCOs (C)	2407	66.1%
Complete and single-copy BUSCOs (S)	2345	64.4%
Complete and duplicated BUSCOs (D)	62	1.7%
Fragmented BUSCOs (F)	457	12.6%
Missing BUSCOs (M)	776	21.3%
Total BUSCO groups searched	3640	100%

Annotation results:

GC content of the genome was determined to be 37.1%. Overall, 18273 gene models were predicted using the AUGUSTUS ver. 3.4.0 (Stanke et al. 2005) gene annotation pipeline based on both de novo and reference-based predictions using genes and proteins from Zebrafish (Danio ratio). Out of the 18273 genes, 15831 were identified as GO terms using PANNZER2 (Törönen et al. 2018).

Serial	Name of the dataset	Data type	Citation
1	Labeo calbasu (Kalibaus) from Halda river, Bangladesh AUGUTUS predicted genes in FASTA format	FASTA	Asek et al. 2021(a)
2	Labeo calbasu AUGUTUS predicted gene annotation in GFF format	GFF	Asek et al. 2021(b)
3	Labeo calbasu (Kalibaus) from Halda river, Bangladesh PANNZER2 annotation files in Excel (xlsx) format	XLSX	Asek et al. 2021(c)

Data availability

The Illumina raw reads have been deposited in the SRA [Project ID: PRJNA689123] under the Accession numbers SRR14651364.

Reference

Ahmed, M. S., 2015. *Labeo calbasu*. Red List of Bangladesh, Freshwater Fishes. IUCN, International Union for Conservation of Nature, Bangladesh Country Office, Dhaka, Bangladesh, Vol (5), p. 189.

Andrews, S., 2010. Babraham bioinformatics-FastQC a quality control tool for high throughput sequence data. URL: https://www. bioinformatics. babraham. ac. uk/projects/fastqc.

Asek, A. A., Siddiki, A. Z., Bhuiyan, M. A. B., Rahman, S. S., Akter, S. and Kibria, M. M. 2021(a). Labeo calbasu AUGUTUS predicted genes in FASTA format. figshare. Dataset. https://doi.org/10.6084/m9.figshare.14980713.v1

Asek, A. A., Bhuiyan, M. A. B.; Siddiki, A. Z., Rahman, S. S., Akter, S. and Kibria, M. M. 2021(b). Labeo calbasu AUGUTUS predicted gene annotation in GFF format. figshare. Dataset. https://doi.org/10.6084/m9.figshare.14980719.v1

Asek, A. A., Bhuiyan, M. A. B.; Siddiki, A. Z., Rahman, S. S., Akter, S. and Kibria, M. M. 2021(c). Labeo calbasu (Kalibaus) from Halda river, Bangladesh PANNZER2 annotation files in Excel (xlsx) format. figshare. Dataset. https://doi.org/10.6084/m9.figshare.14980722.v1

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), pp.1072-1075.

FRSS (Fisheries Resources Survey System).2009. Fisheries Statistical Yearbook of Bangladesh (July 2008-June 2009). Department of Fisheries, Bangladesh, Ministry of Fisheries and Livestock, 26(1), p. 41.

Jackman, S.D., Vandervalk, B.P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S.A., Jahesh, G., Khan, H., Coombe, L., Warren, R.L. and Birol, I., 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome research, 27(5), pp.768-777.

Kabir, M.H., Kibria, M.M. and Hossain, M.M., 2013. Indirect and non-use values of Halda River-a unique natural breeding ground of Indian carps in Bangladesh. Journal of Environmental Science and Natural Resources, 6(2), pp.31-36.

Kabir, M.H., Kibria, M.M., Jashimuddin, M. and Hossain, M.M., 2013. Economic valuation of tangible resources from Halda-the carp spawning unique river located at southern part of Bangladesh. Int. J. Water Res, 1, pp.30-36.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H. and Kohara, Y., 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome research, 24(8), pp.1384-1395.

Li, D., Liu, C.M., Luo, R., Sadakane, K. and Lam, T.W., 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, 31(10), pp.1674-1676.

Li, H., 2015. BFC: correcting Illumina sequencing errors. Bioinformatics, 31(17), pp.2885-2887.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y. and Tang, J., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience, 1(1), pp.2047-217X.

Mishra, S. and Saksena, D.N., 2012. Gonadosomatic index and fecundity of an Indian major carp *Labeo calbasu* in Gohad reservoir. The Bioscan, 7(1), pp.43-46.

Meyer, M. and Kircher, M., 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols, 2010(6), pp.pdb-prot5448.

Podder, M., Islam, M.S., Kibria, M.M., Bhuyan, M.S., 2017. Inventory of Watershed Area of the Halda River Basin for Ecosystem Health Management. Res. J. Environ. Sci, 11(4), pp. 170-176.

Rahman, A.K.A., 2005. Freshwater Fishes of Bangladesh. The Zoological Society of Bangladesh, Dhaka, pp. 115.

Rahman, M.M., Jo, Q., Gong, Y.G., Miller, S.A. and Hossain, M.Y., 2008. A comparative study of common carp (*Cyprinus carpio* L.) and calbasu (*Labeo calbasu* Hamilton) on bottom soil resuspension, water quality, nutrient accumulations, food intake and growth of fish in simulated rohu (Labeo rohita Hamilton) ponds. Aquaculture, 285(1-4), pp.78-83.

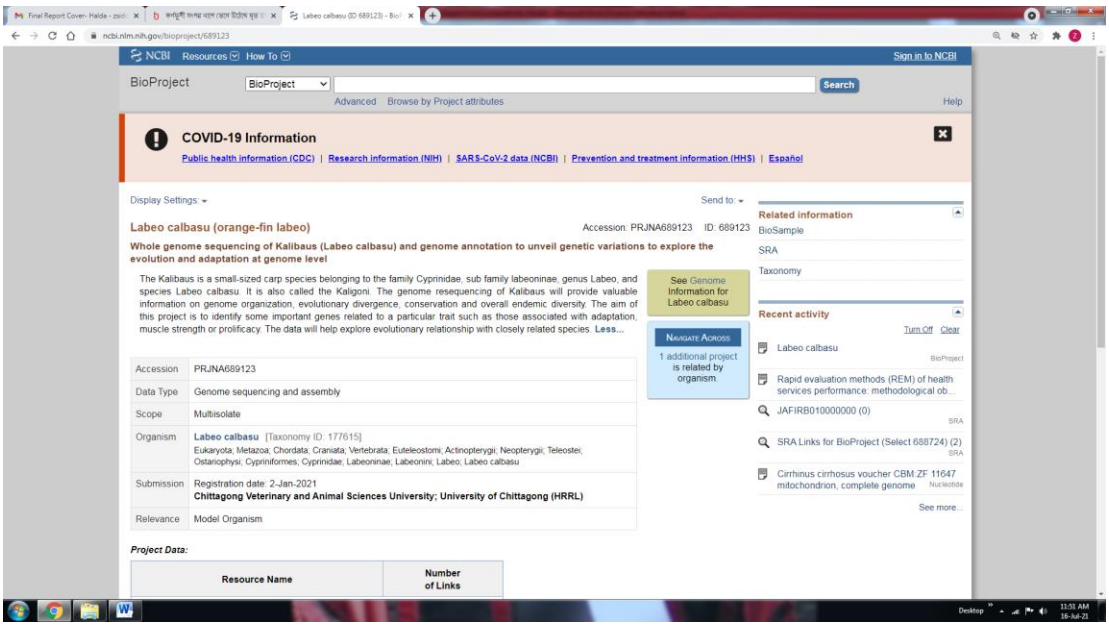
Rahman, M.M., Hossain, M.Y., Ahamed, F., Fatematuzzhura, S.B., Abdallah, E.M. and Ohtomi, J., 2012. Biodiversity in the Padma distributary of the Ganges River, Northwestern Bangladesh: Recommendations for conservation. World Journal of zoology, 7(4), pp.328-337.

Seppey, M., Manni, M. and Zdobnov, E.M., 2019. BUSCO: assessing genome assembly and annotation completeness. Methods in molecular biology (Clifton, NJ), 1962, pp.227-245.

Stanke, M. and Morgenstern, B., 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic acids research, 33(suppl_2), pp.W465-W467.

Talwar, P.K. and Jhingran, A.G., 1991. Inland fish
es of India and adjacent countries. vol. 1. A.A. Balkema, Rotterdam, pp. 203-304.

Törönen, P., Medlar, A. and Holm, L., 2018. PANNZER2: a rapid functional annotation web server. Nucleic acids research, 46(W1), pp.W84-W88.



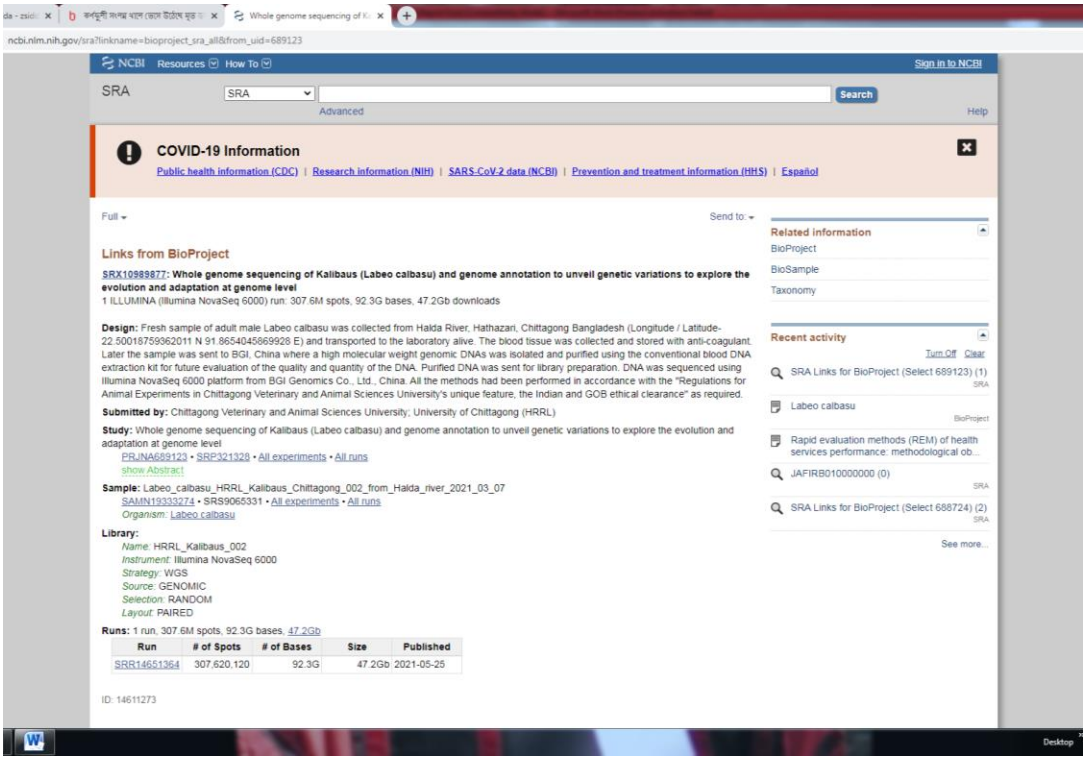


Fig. Screenshot of NCBI SRA datasets of L. calbasu genome

Research Objective 8

Complete Mitochondrial Genome Sequence of pure wild stock of *Labeo calbasu* (Hamilton, 1822) from Halda river of Bangladesh

Abstract

Labeo calbasu or Kalibaus, a freshwater species next to three IMC *Labeo rohita*, *Catla catla* and *Cirrhinus cirrhosus* naturally breeds in Halda river, Bangladesh of which pure wild stock is under threat due to reduction of food availability, indiscriminate fishing, habitat modification and other ecological changes. Yet, the genetic information was lacking. Here in this study wild stock *Labeo calbasu* of Halda river, Bangladesh was collected and processed to reveal the feature of its mitochondrial genome. The total length of the mitochondrial genome is 16,620 base pairs (bp), containing 13 protein-coding genes, two ribosomal RNAs, 22 transfer RNAs, and one control region. Control region is located between tRNA proline and tRNA phenylalanine which is 932 bp in length. The overall base composition of the mtDNA is found to be 24.59% of T (4087), 27.43% of C (4559), 32.65% of A (5427), and 15.32% of G (2547). The entire mtDNA of Kalibaus showed a slight AT rich bias (57.2%) with positive A-T skew (0.14) and negative G-C skew (-0.28). This data will provide us with a clear view into the phylogeny, evolutionary relationships and population genetics of *Labeo calbasu*.

Introduction

Labeo calbasu of the family Cyprinidae is widely distributed in Asian countries including Bangladesh, India, Myanmar, Nepal, Pakistan, South China and Thailand (Talwar and Jhingran 1991). Halda is the only tidal freshwater river in the world that is a natural spawning ground for IMPs (Including Kalibaus) from where fishermen collect fertilized eggs instead of larva or fry directly (Kabir et al., 2013; Kibria, 2009). Once Kalibaus was found in all-natural water bodies (Rahman 2005). But later the population has reduced greatly from natural sources due to the Reduction of food availability (Rahman et al., 2008). Indiscriminate fishing, habitat modification and other ecological changes (Rahman et al. 2012) and its production dropped by about 20% during the period 2001 to 2009 in beel fishery (FRSS 2009). However, The fish is extensively used in stocking culture ponds. Hence, the species is assessed as Least Concern (IUCN, 2015).

L. calbasu is a valuable food fish and also used as game fish in several places of the Indian subcontinent (Talwar and Jhingran, 1991, Rahman, 2005: Mishra and Saksena, 2012). Its liver oil contains Vitamin-A (Ghosh et al., 1993). Also, it provides 16.47% of protein and 2.65% of lipid (Ahmed et al., 2012). So the demand for this fish in the local market is huge. In spite of such importance, there are no records of phylogenetic activity and thus no information of evolutionary relationships of this fish. The mitochondrial genome is regarded as the marker of choice for the reconstruction of phylogenetic relationships at several taxonomic levels, from population to phyla, and has been widely used for the resolution of taxonomic controversies (Gissi et al., 2008). Mitochondrial

DNA has some distinctive features like relatively stable and compact gene organisation, faster replication, maternal inheritance, lack of recombination and presence of an orthologous gene (Wu et al., 2003; Cao et al., 2006; Saccone et al., 1999) which have made mitochondrial DNA extensively used for testing hypotheses of macroevolution, studying population structure, phylogeography, and phylogenetic relationships at various taxonomic levels (Saccone et al., 1999; Zhang et al., 2005; Cao et al., 2006). So, here we have done the mitochondrial DNA sequencing of wild *L. calbasu* to reveal the phylogenetic and evolutionary information.

Materials and methods

Collection of Sample and DNA Extraction

An adult female of *Labeo calbasu* was collected from Halda river, Chattogram (geographic coordinate: 22.50018759362011 N 91.8654045869928 E). Blood was collected from the fresh specimen and preserved in an anticoagulant solution. Later the sample was sent to BGI, China where a high molecular weight genomic DNA was isolated and purified using the conventional blood DNA extraction kit for future evaluation of the quality and quantity of the DNA.

Sequence Analysis

DNA was sequenced using Illumina NovaSeq 6000 platform from BGI Genomics Co., Ltd., China. All the methods had been performed in accordance with the "Regulations for Animal Experiments in Chittagong Veterinary and Animal Sciences University's unique feature, the Indian and GOB ethical clearance" as required. We used BWA V0.7.17 (Li et al. 2010) and SAMTOOLS V0.1.19 (Li et al. 2009) for separating the mitochondrial genome reads from the whole genome sequence by mapping it against the reference Kalibaus mitochondrial genome (NC_017614.1). The clean reads were assembled by using the organelle assembler NOVOPlasty V.4.0 (Dierckxsens et al., 2017). For functional and structural annotation web servers MITOS (Bernt et al., 2013) and GeSeq (Tillich et al., 2017) were used. The 22 tRNA genes' secondary structure and location were determined by using MITOS (Bernt et al. 2013). We used OGDRAW (Greiner et al., 2019) for constructing the circular representation of the entire mitochondrial genome. MEGA X (Kumar et al., 2018) was used to construct a phylogenetic tree following the Neighbor-Joining method (Saitou et al. 1987) where the bootstrap value was 500 (Felsenstein J. 1985).

Result and Discussion

Gene organization and base composition

The complete mitogenome of *L. calbasu* (From Halda) was 16,620 bp in length containing 37 genes in total. In those 37 genes, 13 protein-coding genes (68.66%), two ribosomal RNA genes (15.90%), 22 transfer RNA (9.44%) genes were found (Table 1 & Table 2). The overall nucleotide composition of A = 5427 (32.65%), T= 4087 (24.59%), C= 4559 (27.43%), and G= 2528 (15.32%) were determined. From this analysis, it was clear that the relative order of nucleotide composition corresponds to the nucleotide pattern of other fish A>C>T>G (Wang et al., 2008). Out of 37 genes, 28 genes were found to be encoded in the F-strand except for nad6 and 8 tRNA (trnQ, trnA, trnN, trnC, trnY, trnS2, trnE, trnP) were encoded in the R-strand of the mitochondrial genome of *L. calbasu*. The structural organization and location of the different features of these mitogenomes were consistent with the common vertebrate mitogenome genome model (Liu & Cui, 2009). The whole mitochondrial genome showed a positive A-T skew (0.14) and a negative G-C skew (-0.28). AT and GC content of the total mitogenome were observed to be 56.94% and 43.06% respectively, indicating that the overall nucleotide composition was biased toward adenine and thymine.

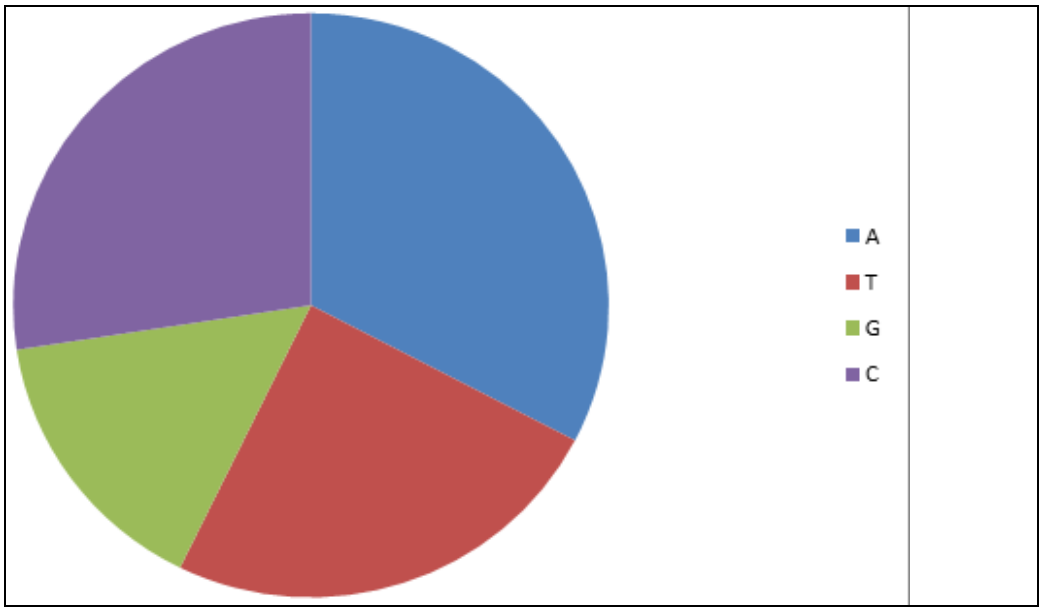


Figure 1: A, T, G, C Content of whole *Labeo calbasu* Mitochondrial Genome

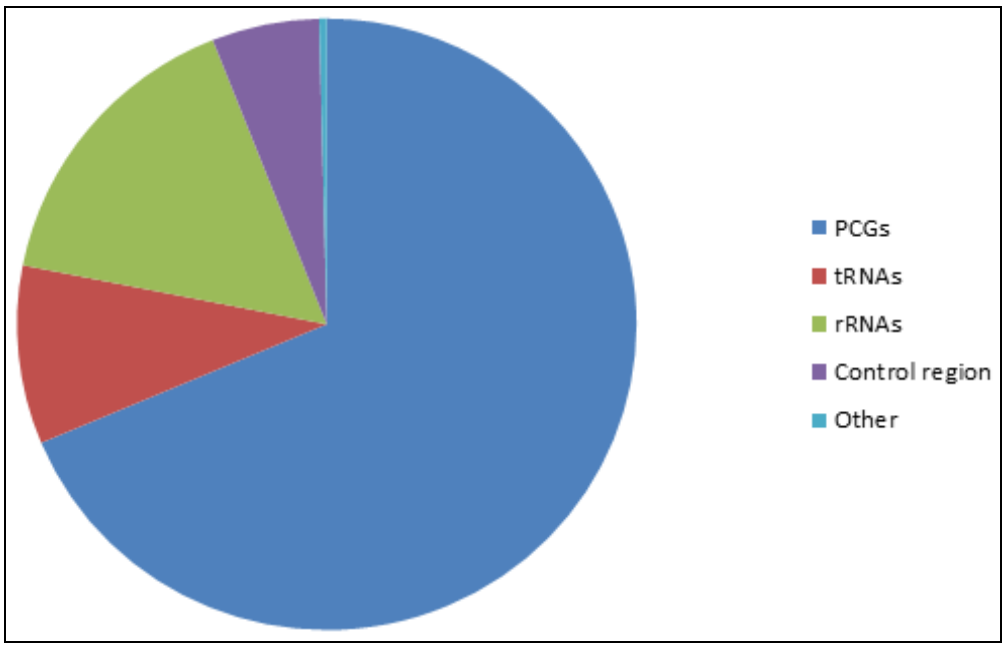


Figure 2: PCGs, tRNAs, rRNAs, Control Region and Other content of whole *Labeo calbasu* Mitochondrial Genome

Table 1: Mitochondrial genome organization in pure wild stock *Labeo calbasu*. Here 'F' and 'R' represent forward and reverse strands respectively.

Gene	**Direction**	**Location**	**Size**	**Anticodon**	**Intergenic nucleotides**
ATP8	F	1-165	165	-	-7
ATP6	F	159-842	684	-	-1
COX3	F	842-1627	786	-	0
tRNA-Gly	F	1628-1699	72	UCC	0
ND3	F	1700-2048	349	-	0
tRNA-Arg	F	2049-2118	70	UCG	0
ND4L	F	2119-2415	297	-	0
ND4	F	2416-3796	1,381	-	0
tRNA-His	F	3797-3865	69	GUG	0
tRNA-Ser 1	F	3866-3934	69	GCU	1
tRNA-Leu 1	F	3936-4008	73	UAG	3
ND5	F	4012-5835	1,824	-	-4
ND6	R	5832-6353	522	-	0

tRNA-Glu	R	6354-6422	69	UUC	5
CYTB	F	6428-7569	1,142	-	-1
tRNA-Thr	F	7569-7641	73	UGU	0
tRNA-Pro	R	7642-7711	70	UGG	2
tRNA-Phe	F	8649-8717	69	GAA	0
s-rRNA	F	8718-9673	956	-	-1
tRNA-Val	F	9673-9744	72	UAC	0
l-rRNA	F	9745-11431	1,687	-	1
tRNA-Leu 2	F	11433-11508	76	UAA	1
ND1	F	11510-12484	975	-	3
tRNA-Ile	F	12488-12560	73	GAU	0
tRNA-Gln	R	12561-12630	70	UUG	1
tRNA-Met	F	12633-12701	69	CAU	0
ND2	F	12702-13746	1,045	-	0
tRNA-Trp	F	13747-13819	73	UCA	1
tRNA-Ala	R	13821-13890	70	UGC	1
tRNA-Asn	R	13892-13967	76	GUU	32

tRNA-Cys	R	14000-14068	69	GCA	0
tRNA-Tyr	R	14069-14139	71	GUA	1
COX1	F	14141-15691	1,551	-	0
tRNA-Ser 2	R	15692-15762	71	UGA	3
tRNA-Asp	F	15766-15837	72	GUC	15
COX2	F	15853-16543	691	-	0
tRNA-Lys	F	16544-16619	76	UUU	1
D LOOF	F	7714-8645	932	-	3

Protein Coding Genes (PCGs) and Their Base Composition

13 protein coding genes constituted 11,412 bp in the whole mitogenome of *Labeo calbasu* which accounted for 68.66% of the total mitogenome. Out of 13 PCGs, 12 of them were encoded in the F-strand (nad1, nad2, cox1, cox2, atp8, atp6, cox3, nad3, nad4l, nad4, nad5 and cob) where nad6 (5832-6353) was encoded in the R-strand of the mtDNA (Table 1).

The AT and GC content of the total PCGs was 57.4% and 42.6% respectively (Table 2). A-T and G-C skews of PCGs were 0.12 and -0.36 respectively reflecting the fact that adenine content is comparatively higher than thymine while cytosine content is higher than guanine which was also observed in the case of the whole mitochondrial genome (Table 2). The size of the PCGs was ranging from 165-1824 bp where nad5 (1824 bp) being the longest and atp8 (165 bp) being the shortest among all the PCGs (Table 1). Overlapping was observed between three adjacent pairs of PCGs (atp8-atp6, atp6-cox3 and nad5-nad6) (Table 1). This sort of overlap is common in most vertebrate mitochondrial genomes (Broughton et al. 2001). No overlapping had been observed between PCGs and other genes (tRNA and rRNA) except Cytb and tRNA-Thr overlapping.

rRNAs, tRNAs and Their Base Composition

The total size of the rRNA was 2643 bp and formed by two subunits are 12S rRNA (956 bp) and 16S rRNA (1687 bp) which constituted 15.90% of the total mitochondrial genome (Table 1 and Table 2). A-T and G-C skew of total rRNA were 0.29 and -0.10 respectively. AT and GC content of the total rRNA were 54.6 and 45.4 respectively. In individual analysis, both 12S rRNA and 16S rRNA showed biasness towards AT content. So it could be concluded that the occurrence of adenine and cytosine were higher than thymine and guanine in the rRNAs, as observed in the whole mitochondrial genome of *L. calbasu*. Single overlapping of base pairs at 12S rRNA- tRNA-Val had been detected (Table 1). No other overlapping is detected between the rRNAs with their adjacent tRNAs.

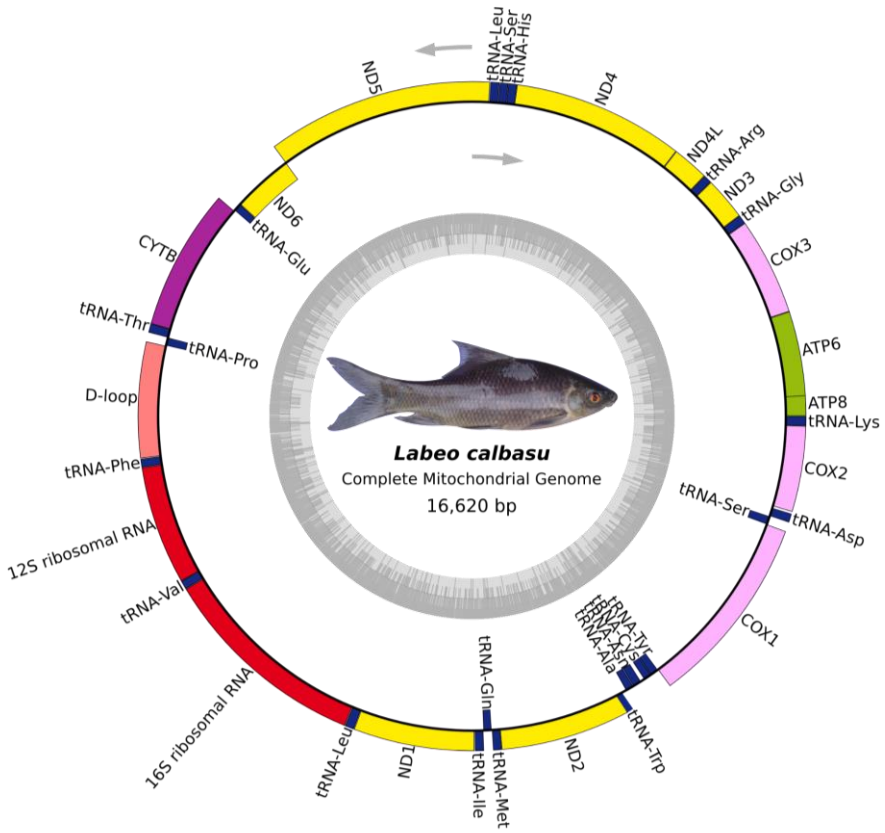
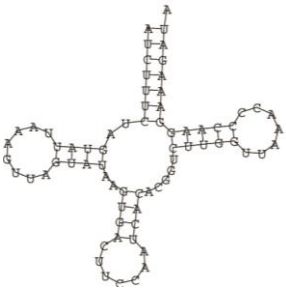


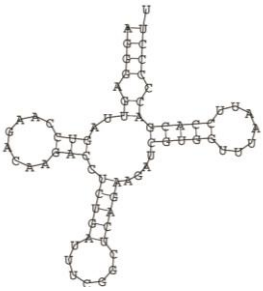
Figure 3: The circular representation of the whole mitochondrial genome of *Labeo calbasu* (Hamilton 1822).

In this study 22 tRNA genes were found in the mitochondrial genome which occupied 1569 bp and 9.44% of the total mtDNA. Out of 22 tRNA genes 14 tRNA genes were

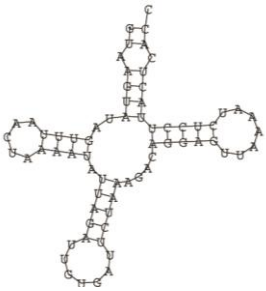
encoded by the F-strand (trnF, trnV, trnL2, trnI, trnM, trnW, trnD, trnK, trnG, trnR, trnH, trnS1, trnL1, trnT) while the rest were encoded by the R-strand (trnQ, trnA, trnN, trnC, trnY, trnS2, trnE and trnP). Size variation among the tRNA coding genes were ranging from 69-76 bp. Among the tRNAs, tRNA-His, tRNA-Ser 1, tRNA-Glu, tRNA-Phe, tRNA-Met and tRNA-Cys were found to have the shortest (69 bp) while tRNA-Leu 2 and tRNA-Lys exhibited the longest sequence (76 bp) in the mtDNA. The overall A-T and G-C skews of these 22 tRNA was 0.10 and -0.09. No overlapping is observed between tRNA. In secondary structure analysis, all of the 22 tRNAs exhibited cloverleaf structure (Figure 4).



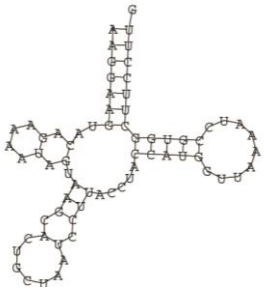
tRNA-Gly



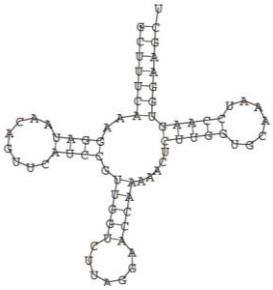
tRNA-Arg



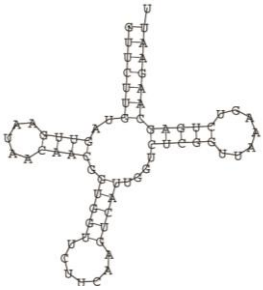
tRNA-His



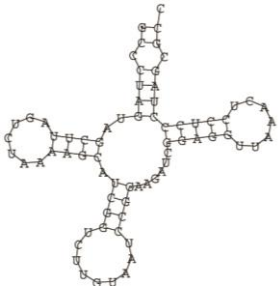
tRNA-Ser 1



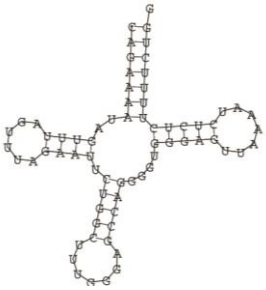
tRNA-Leu 1



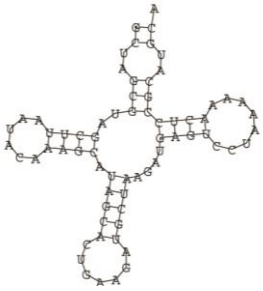
tRNA-Glu



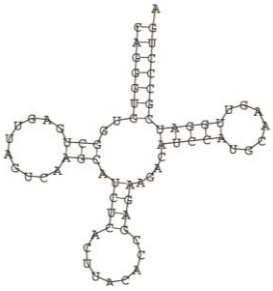
tRNA-Thr



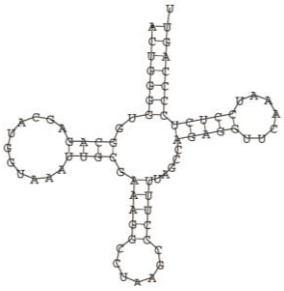
tRNA-Pro



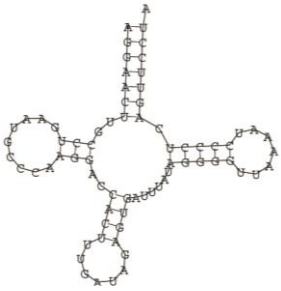
tRNA-Phe



tRNA-Val



tRNA-Leu 2



tRNA-Ile

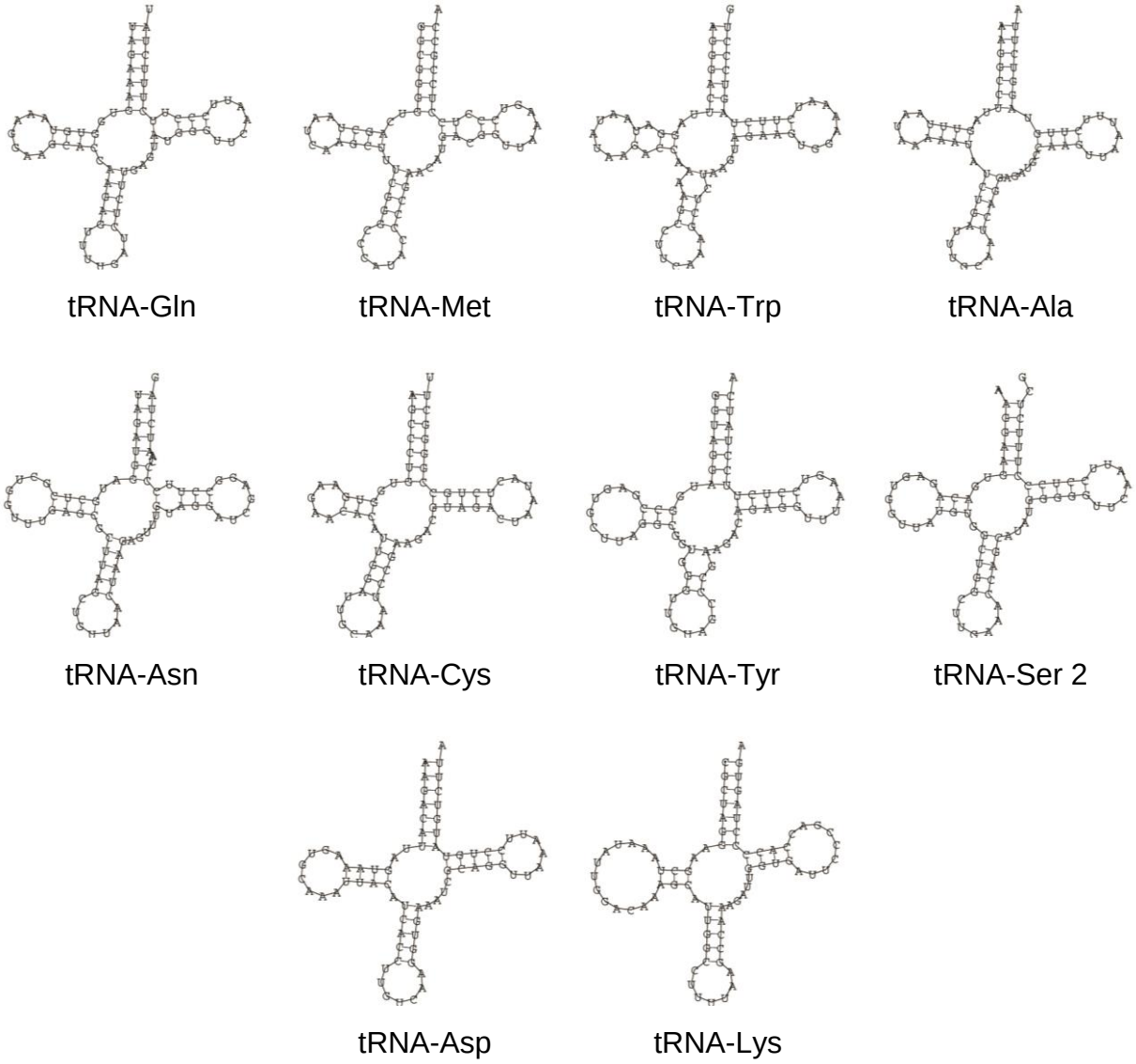


Figure 4: Typical tRNA secondary structure for 22 tRNA coding genes.

Control Region

A control region of a total of 932 bp was detected between proline (trnP) and phenylalanine (trnF), constituting 5.61% of the total mtDNA. The observed A-T skew of the CR was positive (0.06) which followed the whole mtDNA while the G-C skew was negative (-0.22) reflecting the whole mtDNA as well (Table 2). Again the control region exhibited bias toward AT content which was 65.6% and also the highest AT-rich region in the mitochondrial genome of *L. calbasu*. While GC content was found to be 34.4%, the lowest GC containing region in the whole mitochondrial genome of *L. calbasu*.

Table 2: Total percentage of PCGs, tRNAs, rRNAs and control region in the *Labeo calbasu* along with their AT, GC percentage and A-T, G-C skew.

Name	Length	A	T	G	C	A+T %	G+C %	A-T skew	G-C skew	% in the mtDNA
Whole mitochondrial genome	16620	5427	4087	2547	4559	57.2	42.8	0.14	-0.28	100
PCGs	11412	3672	2878	1559	3303	57.4	42.6	0.12	-0.36	68.66
tRNAs	1569	484	398	311	376	56.2	43.8	0.10	-0.09	9.44
rRNAs	2643	928	514	541	660	54.6	45.4	0.29	-0.10	15.90
Control region	932	323	288	125	196	65.6	34.4	0.06	-0.22	5.61
Other	64	20	9	11	24	-	-	-	-	0.39

Phylogenetic Analysis

The phylogenetic analysis was inferred using the Neighbor-Joining method (Saitou et al. 1987). The optimal tree is shown below (figure 5). The tree was made using bootstrap tests (500 replicates) are shown next to the branches (Felsenstein J. 1985). The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al. 2004). This analysis involved 11 nucleotide sequences and an out-group which is removed manually. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were a total of 17958 positions in the final dataset. Evolutionary analyses were conducted in MEGA X (Kumar et al. 2018). Except for our own species, the rest of the 11 sequences of the mitochondrial genome were taken from the NCBI database. In our analysis, *Labeo calbasu* of Halda river is paired with another *L. calbasu* that is taken from NCBI. Other species of the family Cyprinidae show a close relationship with our Kalibaus.

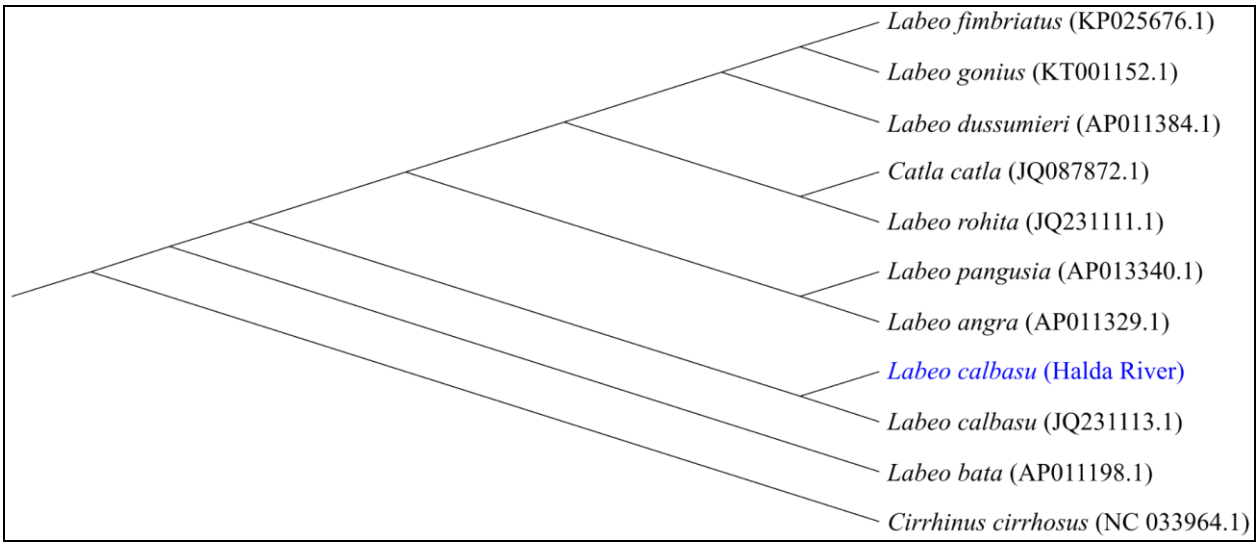


Figure 5: Phylogenetic analysis of *Labeo calbasu* (Hamilton, 1822)..

References

Ahmad, M.F. and Niazi, M.S., 1988. Important edible fishes of Pakistan. Zoological Survey Department, Government of Pakistan, p. 31.

Ahmed, M. S., 2015. *Labeo calbasu*. Red List of Bangladesh, Freshwater Fishes. IUCN, International Union for Conservation of Nature, Bangladesh Country Office, Dhaka, Bangladesh, Vol (5), p. 189.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M. and Stadler, P.F., 2013. MITOS: improved de novo metazoan mitochondrial genome annotation. Molecular phylogenetics and evolution, 69(2), pp.313-319.

Cao, S.Y., Wu, X.B., Yan, P., Hu, Y.L., Su, X. and Jiang, Z.G., 2006. Complete nucleotide sequences and gene organization of mitochondrial genome of *Bufo gargarizans*. Mitochondrion, 6(4), pp.186-193.

Dierckxsens, N., Mardulyn, P. and Smits, G., 2017. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic acids research, 45(4), pp.e18-e18.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. evolution, 39(4), pp.783-791.

FRSS (Fisheries Resources Survey System).2009. Fisheries Statistical Yearbook of Bangladesh (July 2008- June 2009). Department of Fisheries, Bangladesh, Ministry of Fisheries and Livestock, 26(1), p. 41.

Gissi, C., Iannelli, F. and Pesole, G., 2008. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. Heredity, 101(4), pp.301-320.

Ghosh, A.R., Chakravorty, P.N. and Guha, B.C., 1933. Further Observations on Vitamin A in Indian Fish-Liver Oils. Indian Journal of Medical Research, 21, pp.441-6.

Greiner, S., Lehwark, P. and Bock, R., 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Research, 47(W1), pp.W59-W64.

Kabir, M.H., Kibria, M.M. and Hossain, M.M., 2013. Indirect and non-use values of Halda River-a unique natural breeding ground of Indian carps in Bangladesh. Journal of Environmental Science and Natural Resources, 6(2), pp.31-36.

Kibria, M.M., Farid, I. and Ali, M., 2009. Halda Restoration Project: Peoples Expectation and Reality, A Review Report Based on the Peoples Opinion of the Project Area (In Bangla). Chittagong: Chattagram Nagorik Oddogh & Actionaid Bangladesh, p. 67

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K., 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution, 35(6), pp.1547-1549.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), pp.2078-2079.

Li, H. and Durbin, R., 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics, 26(5), pp.589-595.

Mishra, S. and Saksena, D.N., 2012. Gonadosomatic index and fecundity of an Indian major carp *Labeo calbasu* in Gohad reservoir. The Bioscan, 7(1), pp.43-46.

Rahman, A.K.A., 2005. Freshwater Fishes of Bangladesh. The Zoological Society of Bangladesh, Dhaka, pp. 115.

Rahman, M.M., Jo, Q., Gong, Y.G., Miller, S.A. and Hossain, M.Y., 2008. A comparative study of common carp (*Cyprinus carpio* L.) and calbasu (*Labeo calbasu* Hamilton) on bottom soil resuspension, water quality, nutrient accumulations, food intake and growth of fish in simulated rohu (*Labeo rohita* Hamilton) ponds. Aquaculture, 285(1-4), pp.78-83.

Rahman, M.M., Hossain, M.Y., Ahamed, F., Fatematuzzhura, S.B., Abdallah, E.M. and Ohtomi, J., 2012. Biodiversity in the Padma distributary of the Ganges River, Northwestern Bangladesh: Recommendations for conservation. World Journal of zoology, 7(4), pp.328-337.

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G. and Reyes, A., 1999. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. Gene, 238(1), pp.195-209.
Saitou, N. and Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution, 4(4), pp.406-425.
Talwar, P.K. and Jhingran, A.G., 1991. Inland fishes of India and adjacent countries. vol. 1. A.A. Balkema, Rotterdam, pp. 203-304.
Tamura, K., Nei, M. and Kumar, S., 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences, 101(30), pp.11030-11035.
Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R. and Greiner, S., 2017. GeSeq–versatile and accurate annotation of organelle genomes. Nucleic acids research, 45(W1), pp.W6-W11.
Wu, X., Wang, Y., Zhou, K., Zhu, W., Nie, J. and Wang, C., 2003. Complete mitochondrial DNA sequence of Chinese alligator, *Alligator sinensis*, and phylogeny of crocodiles. Chinese Science Bulletin, 48(19), pp.2050-2054.
Zhang, P., Zhou, H., Liang, D., Liu, Y.F., Chen, Y.Q. and Qu, L.H., 2005. The complete mitochondrial genome of a tree frog, *Polypedates megacephalus* (Amphibia: Anura: Rhacophoridae), and a novel gene organization in living amphibians. Gene, 346, pp.133-143.

Research Objective 9
Whole Genome Sequence of *Platanista gangetica* from Halda river of Bangladesh

Abstract

Objectives: The Ganges river dolphin, *Platanista gangetica gangetica* (Roxburgh 1801) , an essentially blind freshwater cetacean; endemic to the Ganges-Brahmaputra-Meghna and Karnaphuli- Sangu river systems in South Asian countries, has been classified as “Endangered” on the IUCN Red List of Threatened Species. Being a tertiary organism in the food chain, it is an important indicator species of river ecosystem. The species bear a special navigation mechanism through echolocation which helps in feeding, locomotion, and sensory perception. To better understand the bioscience of the Ganges dolphin, we sequenced the genome generated by employing the Illumina Novaseq 6000 technologies and produced an assembly that contains ~95% of the genes known to be highly conserved among eukaryotes. Here, we report the chromosome-level reference genome of a healthy male adult *Platanista gangetica*, assembled *de novo* from an individual originating in the river Halda of Chittagong,Bangladesh. River Halda, which is known as the national fish treasury of Bangladesh, is very intimately linked in the genesis of various carp species. In total, 22788799445 bases of raw reads were generated by whole-genome sequencing, using an Illumina NovaSeq platform ; a draft genome were assembled of 2.9 Gb, Genome annotation identified. Within the predicted genes we have confirmed the presence of >20 genes or gene families that have been associated with adaptive evolution in other cetaceans. Our genome assembly were an irreplaceable resource for further genetic research on adaptive ecology and this will enable comparative studies of natural selection in freshwater cetaceans. Overall, this genome assembly and draft annotation represent a crucial addition to the genomic resources currently available for the study of order Cetartiodactyla and Platanistidae evolution, phylogeny, and conservation biology .

Introduction

Technical advancements in recent years have reduced sequencing costs and improved access to sequencing data. Subsequent transformation in DNA extraction, preparation and assembly algorithms facilitates the low-cost accurate de novo genome assemblies. Such assemblies are indispensable for constructing haplotype diversity databases for comparative biology,breeding,medicine, and conservation planning (Martinez-Viaud et al., 2019). Without whole-genome sequencing, the genomic basis of key adaptations remains difficult to identify and besides genome-wide resources are no longer limited to a few model organisms. With these resourceful data, it is now accomplishable to detect more subtle differences among species and populations and to extend their investigation beyond nucleotide polymorphisms to larger structural variants, linkage groups, and highly divergent genes and regions on a genomic scale (e.g., Autenrieth et al., 2018 ,Gloss et al. 2017; Johannesson et al. 2017; Li et al. 2017; Willoughby et al. 2017). However, Genetic amenities available for cetaceans do not yet adequately cover

the diversity of this group (Foote et al. 2016; Keane et al. 2015; Nery et al. 2013; Sun et al. 2013; Yim et al. 2014; Zhou et al. 2013).Within the toothed whales (odontoceti) species from five of the ten families (Phocoenidae, Physeteroidae, Lipotidae, Monodontidae and Delphinidae) have been whole-genome sequenced to date (Autenrieth et al., 2018).

Ganges River dolphin, *Platanista gangetica gangetica*, is one of the three obligate freshwater dolphins in the world, are distributed throughout the Ganges–Brahmaputra–Meghna and Karnaphuli–Sangu river systems of India, Bangladesh,Nepal and possibly Bhutan (Mohan et al. 1997; Sinha et al. 2000; Sinha & Kannan, 2014;Smith et al. 2001). One of the noteworthy features is, this species moves and feeds in a murky riverine environment using echolocation. Similar to other freshwater cetaceans, who have evolved these traits convergently, Ganges dolphin rely very little on eyesight because of the muddy waters it inhabits, and as a result their vestigial effectively non-functional eyes (even lack lenses) are only capable of distinguishing light from dark, designating them as blind-river dolphin (Rice, 1998). This dolphin is almost identical to its closest relative subspecies, Indus River Dolphin (*Platanista gangetica minor*) (Grill, 2000). Both subspecies are assigned to a monotypic family, Platanistidae ; one of the most basal cetacean families which is very closely related to Kogiidae(dwarf and pygmy sperm whales) and Physeteridae (sperm whales) (McGowen et al.2009; Zhou et al.2011). This ancient cetacean family diverged approximately 29 Million Years (MY) ago, 22 MY before modern marine dolphins arose (Xiong et al., 2009). Modern genetic studies showed that Indus and Ganges dolphins diverged from each other approximately 0.5 MY ago and, if they are shown to have morphological differences, the two subspecies may be recognised as separate species in the future (Braulik et al.,2014b).Although several marine dolphin species are commonly found in rivers far upstream of freshwater ecosystems, Ganges river dolphins are morphologically and phylogenetically distinct from marine dolphins (Sinha et al., 2010).

According to the IUCN Red List (Smith et al. 2012, IUCN-Bangladesh 2015), Ganges dolphin is globally Endangered(EN) and Vulnerable(VU) in Bangladesh. The current range of Ganges subspecies has been estimated about 1,200-1,800 individuals, but the actual population is believed to be larger because some potentially crucial areas have not been surveyed and at least some of the counting and estimations were considered negatively biased (Khan, (2019),Smith et al. 2012, Braulik et al. 2012). As gangetic dolphin mainly live in the human-dominated floodplain rivers in South Asia, they faces threats on their survival from habitat loss and fragmentation due to damming of rivers for hydropower and irrigation (Baruah et al. 2012; Choudhary et al. 2012; Braulik et al. 2012), incidental by-catch in fishing gear (Mansur et al. 2008), intentional killing for their oil (Sinha, 2002), water pollution (Kannan et al. 1997) and population decline of their prey (Kelkar et al. 2010). Even so it is a flagship species for river conservation and their preservation can benefit other aquatic species to be conserved as well as wider local communities for their subsistence(Sinha et al., 2010). Furthermore, Since *Lipotes vexilifer* (Yangtze River dolphin) of China was declared extinct in 2006 this incidence

implies a severe warning to the Ganges dolphin because both species have similar environmental conditions (Turvey et al., 2007). Likewise, *P. gangetica* illustrates an ancient lineage of cetaceans thus its extinction may lead to “missing link” for whole mammalian class. The conservation demand of these charismatic aquatic mammal is therefore paramount. Considering the importance of the Ganges dolphin, Genomic information is imperative for understanding the evolution,adaptation and conservation of the species. Although the morphology of the Gangetic dolphin has been studied intensively because of its extraordinary features, the underlying genetics have received attention to a lesser extent.

We have developed a high quality genome assembly for the *Platanista gangetica;* from an individual originating in the river Halda of Chittagong,Bangladesh which river is home to approximately 160 Ganges dolphins (reference). Among 720 rivers in Bangladesh, the Halda is an incomparable river which acts as influential mother fisheries and natural breeding ground for pure south asian carp fishes. This tidal River serves as the pure natural gene bank of major Indian carps which is the only of its kind in the world from where fishermen can collect fertilized eggs directly (Kibria et al., 2009). The ox bow bends of this river, combined with various other chemical and biological features make it unequaled. During seven years of study period (from September 2004 to December 2011) a total of 83 finfish species under 13 orders and 35 families and a total of 10 shellfish (9 prawns and 1 crab) under 1 order and 3 families were identified from the river Halda (Alam et al., 2013). These varieties of fish and crustaceans make this river a good home ground for the Ganges River Dolphin that feeds on fish and crustaceans. However, the molecular data (nuclear and mitochondrial) of P. gangetica gangetica was lacking and availability of first hand data on such pristine species can substantially aid in resolving the taxonomic perplex and phylogenetic relationship of this clade within cetaceans (Braulik et al., 2014)

In this paper, we present the first de novo assembly of the full genome of *Platanista gangetica gangetica*, scaffolded and draft-annotated to predict its coding proteins and their functions (Deposited at NCBI as BioProject: PRJNA675309 with BioSample-ID: SAMN16703806). We demonstrate a high level of completeness of the assembly, showing that many of our scaffolds are near-chromosome-level, and identify a number of candidate genes for future evolutionary analyses on adaptation in cetaceans.
The development of informative genetic assays would be a boon to *P. gangetica* conservation, as the ability to identify individuals and family groups, delineate populations, and track patterns of genetic diversity over space and time would result in more informed management decisions. To usher this species into the era of genomics, a high-quality reference genome is essential. It provides structure to catalogue diversity within and between species at the whole-genome level.

Methodology of the Study:

Sampling and DNA isolation:

Sampling and DNA isolation: Fresh sample of skin, liver and muscle tissues were collected from an adult male *Platanista gangetica* which was found dead from Akbaria point of Halda River under Hathazari upazila (Longitude / Latitude- 22.446661N, 91.861295E) on Thursday, December 5, 2019. Then, transported to the laboratory to preserve at -70° C for further analysis during the period of December 2019. The liver and muscle tissue was dissected and stored with 95% ethanol. Later for Sample preparation, a high molecular weight genomic DNA isolation and purification were performed following the AddPrep Genomic DNA extraction kit (AddBio, Korea) for future evaluation of the quality as well as quantity of the DNA. All the methods had been performed in accordance with the "Regulations for Animal Experiments in Chittagong Veterinary and Animal Sciences University's unique feature, the Indian and GOB ethical clearance" as required.

Library preparation

Purified genomic DNA was sent for library preparation and whole genome sequencing (WGS) at BGI Group (Shenzhen, Guangdong, China). A total of **22.8 Gb** of subread bases with a read length of bp were generated using Next-generation sequencing (NGS) technology on an Illumina NovaSeq6000 platform. After sequencing quality of raw sequence reads and trimmed sequencing reads were inspected using FastQC version 0.11.8 (Andrews, S., 2010). Reads were quality controlled including removing adaptor sequences, contamination and low-quality reads from raw reads using BFC (Li,H. , 2015). A total of 20407880100 bp clean reads were included in the assembly with 48X coverage.

Genome assembly:

For de novo assembly we used ABySS v. 2.2.4 (Shaun D Jackman et al., 2017), SOAPdenovo2 (Luo, R et al., 2012) and MEGAHIT v1.2.9 (Li,D et al., 2015) assembler to assemble the *P. gangetica* genome. Since there is currently no de novo assembler assured to outperform others and as assemblers overall performance can differ relying on the dataset, three unique assemblers had been used and an assembly evolution was subsequently performed in order to choose the best assembler. All assemblers follow the classic De Bruijin graph illustration even though the assembly algorithm differs across methods. Finally BUSCO v.4.1.2 (Simão, F et al. , 2015) was used to assess the assembly quality in terms of gene completeness.

SOAPdenovo assembly:

Information for assembly Scaffold (Kmer 27)

Size include N	1821064417
Size without N	1820439352
Scaffold_Number	5308908
Mean_Size	343
Median_Size	212
Longest_Sequence	6498
Shortest_Sequence	100
Singleton_Number	5278877
Average_length_of_break(N)_in_scaffold	0

NAME	NUMBERS	PERCENTAGE
scaffolds>100	4549637	85.70%
scaffolds>500	1085559	20.45%
scaffolds>1K	282431	5.32%
scaffolds>10K	0	0.00%

Table: Nucleotide composition

NAME	NUMBER	PERCENTAGE
Nucleotide_A	546224654	29.99%
Nucleotide_C	376569646	20.68%
Nucleotide_G	371969539	20.43%
Nucleotide_T	525675513	28.87%
GapContent_N	625065	0.03%
Non_ACGTN	0	0.00%

GC_Content 41.12% (G+C)/(A+C+G+T)

NAME	NUMBERS	LENGTH
N10	1409	99942
N20	1044	252036
N30	823	449349
N40	661	696862
N50	528	1005575
N60	412	1396058
N70	306	1908590
N80	216	2618812
N90	142	3659324

Information for assembly Contig (Kmer 27)

Size_includeN	1820560946
Size_withoutN	1820560946
Contig_Number	5330112
Mean_Size	341
Median_Size	211
Longest_Sequence	6498
Shortest_Sequence	100

NAME	NUMBERS	PERCENTAGE
scaffolds>100	4570450	85.75%
scaffolds>500	1082708	20.31%
scaffolds>1K	279549	5.24%
scaffolds>10K	0	0.00%

Table: Nucleotide composition

NAME	NUMBER	PERCENTAGE
Nucleotide_A	546440167	30.01%
Nucleotide_C	376637156	20.69%
Nucleotide_G	371971312	20.43%
Nucleotide_T	525512311	28.87%
GapContent_N	0	0.00%
Non_ACGTN	0	0.00%

GC_Content 41.12% (G+C)/(A+C+G+T)

NAME	NUMBERS	LENGTH
N10	1399	100588
N20	1037	253650
N30	818	452159
N40	656	701161
N50	525	1011713
N60	409	1404529
N70	304	1920230
N80	214	2634519
N90	141	3679387

Number of contigs in scaffolds 51235
Number of contigs not in scaffolds(Singleton) 5278877
Average number of contigs per scaffold 1.7

AbySS assembly:

n	n:500	L50	min	N75	N50	N25	E-size	max	sum
14.42e6	797646	291600	500	617	776	1040	887	8112	621.5e6
	797284	291465				1041	889		621.7e6
	797262	291443				1041	889		621.7e6

MEGAHIT assembly: (scaffold)

n	N:500	L50	minimum	N75	N50	N25	E-size	max	sum
692208	472832	1	500	3592	1.267e9	1.267e9	742.5e6	1.267e9	2.162e9

Gene prediction and functional annotation :

The first step in genome annotation in a given genomic sequence is gene structure prediction. Gene prediction was conducted using MAKER ver 3.01.03 (Campbell et al., 2014) which defines probability distributions for the different sections of genomic sequence. Gene prediction was performed ab initio with using both given and default parameters. Functional annotation was obtained by InterProScan ver 5.46-81.0 (Jones et al., 2014).The functional annotation report has been deposited at Figshare database

Conclusion

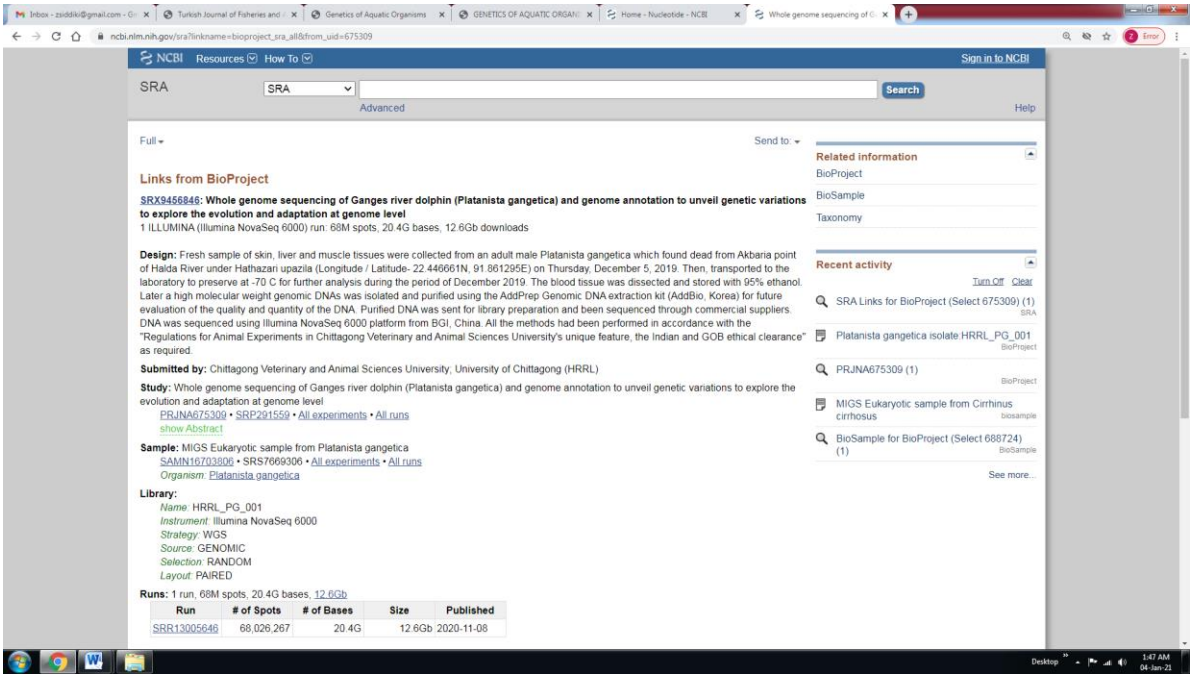
With a GC-content of 43.6% and 20.41 GBp bases in SRA reads, this sequencing data is comparable to other high quality cetacean genomes. In summary, we generated and analyzed the first whole genome assembly and annotation of *Platanista gangetica gangetica*. This genome adds to available Cetacean genomes by supplying essential resources for advance investigation within the Odontoceti and Cetacea.

The acquired data should facilitate further studies of the genetic basis of divergence between two subspecies Indus and Ganges dolphin and of the molecular differences between freshwater, marine, and terrestrial mammals. This data will also furnish invaluable information for further genetic studies of this species,both for whole-genome investigations into population structure and to denote key genes associated with local adaptation.

Data availability

The Illumina raw reads have been deposited in the SRA (Project ID : PRJNA675309) under the Accession numbers SRR13005646 . This Whole Genome Shotgun project

has been deposited at DDBJ/ENA/GenBank under the accession SRX9456846.



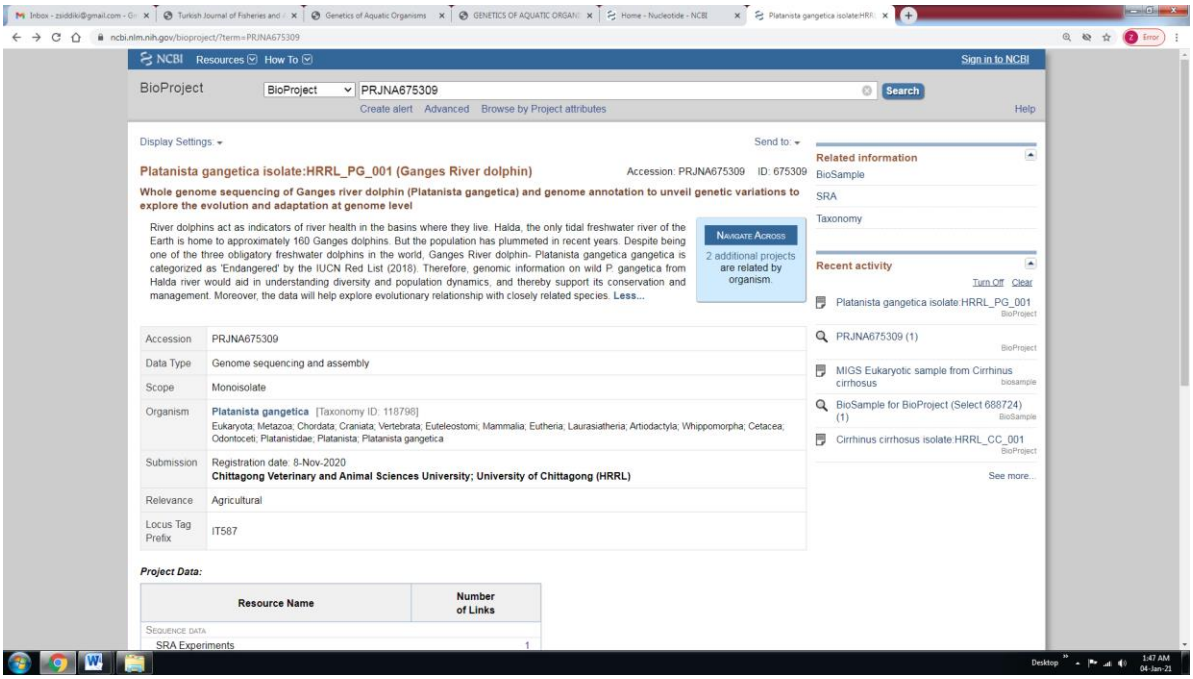


Fig. Screenshot of NCBI SRA datasets of *Platanista* genome

Reference

Alam, M. S., Hossain, M. S., Monwar, M. M., & Hoque, M. E. (2013). Assessment of fish distribution and biodiversity status in Upper Halda River, Chittagong, Bangladesh. *International Journal of Biodiversity and Conservation, 5(6), 349-357*.

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Baruah, D., L. P. Hazarika, B. Bakalial, S. Borah, R. Dutta, and S. P. Biswas. 2012. A grave danger for the Ganges dolphin (Platanista gangetica Roxburgh) in the Subansiri River due to a large hydroelectric project. Environmentalist 32:85–90.

Braulik, G. T., A. P. Reichert, T. Ehsan, S. Khan, S. P. Northridge, J. S. Alexander, and R. Garstang. 2012. Habitat use by a freshwater dolphin in the low-water season. Aquatic Conservation:Marine and Freshwater Ecosystems 22:533–546.

Braulik, G.T., Reichert, A.P., Ehsan, T., Khan, S., Northridge, S.P., Alexander, J.S., Garstang, R.,2012b.Habitat use by a freshwater dolphin in the low-water season. *Aquat. Conserv.Mar. Freshwat. Ecosyst.* 22, 533–546.

Braulik GT, Barnett R, Odon V, Islas-Villanueva V, Hoelzel AR, and Graves JA (2014)One Species or Two? Vicariance, Lineage Divergence and Low mtDNA Diversity in Geographically Isolated Populations of South Asian River Dolphin. *J.Mamm.Evol*.1–10.doi: http://dx.doi.org/10.1007/s10914-014-9265-6

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. Current protocols in bioinformatics, 48(1), 4-11.

Choudhary, S., S. Dey, S. Dey, V. Sagar, T. Nair, and N. Kelkar. 2012. River dolphin distribution inregulated river systems: implications for dry-season flow regimes in the Gangetic basin. AquaticConservation: Marine and Freshwater Ecosystems 22:11–25.

Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., et al. (2016). Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nature Communications*, 7(May), 11693. doi:10.1038/ncomms11693

Keane, M., Semeiks, J., Webb, A. E., Li, Y. I., Quesada, V., Craig, T., et al. (2015). Insights into theevolution of longevity from the bowhead whale genome. Cell Reports, 10(1), 112–122.doi:10.1016/j.celrep.2014.12.008

Li, H. (2015). BFC: correcting Illumina sequencing errors. *Bioinformatics* , 31 (17),2885-2887.

Mansur, E. F., B. D. Smith, R. M. Mowgli, and M. A. A. Diyan. 2008. Two Incidents of Fishing Gear Entanglement of Ganges River Dolphins (Platanista gangetica gangetica) in Waterways of theSundarbans Mangrove Forest, Bangladesh. Aquatic Mammals 34. Aquatic Mammals.

M. Autenrieth, S. Hartmann, L. Lah, A. Roos, A.B. Dennis, R. Tiedemann (2018). High-quality whole-genome sequence of an abundant Holarctic odontocete, the harbour porpoise (*Phocoena phocoena*) Mol. Ecol. Resour., 18 , pp. 1469-1481

Nery, M. F., Gonzalez, D. J., & Opazo, J. C. (2013). How to make a folphin: Molecular signature ofpositive selection in Cetacean genome. *PLoS ONE*, 8(6), 2–8. doi:10.1371/journal.pone.0065491

Gloss, A. D., Groen, Simon, C., & Whiteman, N. K. (2017). A genomic perspective on the generation and maintenance of genetic diversity in herbivorous insects. *Annual Review of Ecology, Evolution, and Systematics, November(47)*, 165–187. doi:10.1146/annurev-ecolsys-121415-032220

Grill, P., 2000. The Little Guide: Whales, Dolphins and Porpoises. Fog City Press, San Francisco.

IUCN-Bangladesh 2015. Red List of Bangladesh – A Brief on Assessment Result 2015. IUCN-Bangladesh, Dhaka. Bangladesh.

Johannesson, K., Butlin, R. K., Panova, M., & Westram, A. M. (2017). Mechanisms of adaptive di-vergence and speciation in *Littorina saxatilis*: Integrating knowledge from ecology and geneticswith new data emerging from genomic studies. In *Population Genomics* (pp. 1–25). Springer,Cham. doi:https://doi.org/10.1007/13836_2017_6

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C.,& Pesseat, S.(2014). InterProScan 5:

genome-scale protein function classification. Bioinformatics, 30(9),1236-1240.
Kannan, K., K. Senthilkumar, and R. K. Sinha. 1997. Sources and Accumulation of ButyltinCompounds in Ganges River Dolphin, *Platanista gangetica*. Applied Organometallic Chemistry11:223–230.
Kelkar, N., J. Krishnaswamy, S. Choudhary, and D. Sutaria. 2010. Coexistence of fisheries with river dolphin conservation. Conservation Biology 24:1130–1140.
Kibria, M. M., Farid, I., & Ali, M. (2009). Halda Restoration Project: Peoples Expectation and Reality, A Review Report Based on the Peoples Opinion of the Project Area (In Bangla). *Chittagong: Chattagram Nagorik Oddogh & Actionaid Bangladesh. 67p.*
Khan, M. M. H. (2019). MANAGEMENT PLAN FOR THE GANGES RIVER DOLPHIN IN HALDA RIVER OF BANGLADESH.
Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England)*, *31*(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033
Li, L. F., Li, Y. L., Jia, Y., Caicedo, A. L., & Olsen, K. M. (2017). Signatures of adaptation in the weedy rice genome. *Nature Genetics, 49(5)*, 811–814. doi:10.1038/ng.3825
Martinez-Viaud, K. A., Lawley, C. T., Vergara, M. M., Ben-Zvi, G., Biniashvili, T., Baruch, K., et al. (2019). New de novo assembly of the Atlantic bottlenose dolphin (*Tursiops truncatus*) improves genome completeness and provides haplotype phasing. *Gigascience* 8:giy168. doi:10.1093/gigascience/giy168
McGowen MR, Spaulding M, and Gatesy J (2009)Divergence date estimation and a comprehensive molecular tree of extant cetaceans. *Mol.Phylogenet.Evol*.53:891–906.doi: http://dx.doi.org/10.1016/j.ympev.2009.08.018
Mohan, R.S.L., S.C. Dey, S.P. Bairagi, and S. Roy. 1997. On a survey of the Ganges River dolphin *Platanista gangetica* of Brahmaputra River, Assam. *The Journal of the Bombay Natural History Society 94: 483–495.*
Rice, D.W., 1998. Marine mammals of the world. Marine Mammal Society, Lawrence, KS.
Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam, Jun Wang, SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler, *GigaScience*, Volume 1, Issue 1, December 2012, 2047–217X–1–18, https://doi.org/10.1186/2047-217X-1-18
Shaun D Jackman, Benjamin P Vandervalk, Hamid Mohamadi, Justin Chu, Sarah Yeo, SAustin Hammond, Golnaz Jahesh, Hamza Khan, Lauren Coombe, René L Warren, and InancBirol (2017). ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter.Genome research, 27(5), 768-777. doi:10.1101/gr.214346.116
Simpson, Jared T., Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven JMJones, and Inanc Birol (2009). ABySS: a parallel assembler for short read sequence data.Genome research, 19(6), 1117-1123. doi:10.1101/gr.089532.108
Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M.(2015). BUSCO: assessing genome assembly and annotation completeness with single-copyorthologs. Bioinformatics , 31 (19), 3210-3212.a
Sinha, R.K. 2000. Status of the Ganges River dolphin (*Platanista gangetica*) in the vicinity of Farakka Barrage, India. *In Biology and conservation of freshwater cetaceans in Asia*, ed. by R.R.
Reeves, B.D. Smith, T. Kasuya, Vol. 23, 42–48 pp. Occasional Gland, Switzerland: Paper of the IUCN Species Survival Commission.
Sinha, R. K. 2002. An alternative to dolphin oil as a fish attractant in the Ganges River system:Conservation of the Ganges River dolphin. Biological Conservation 107:253–257.
Sinha, K., Behera, S., Choudhary, B.C., 2010. The Conservation Action Plan for the Ganges River Dolphin 2010-2020.
Sinha, R. K., & Kannan, K. (2014). Ganges River dolphin: An overview of biology, ecology, and

conservation status in India. *Ambio*, 43, 1029–1046. https://doi.org/10.1007/s13280-014-0534-7
Smith, B.D., B. Ahmed, M.E. Ali, and G. Braulik. 2001. Status of the Ganges River dolphin or shushuk *Platanista gangetica* in Kaptai Lake and the southern rivers of Bangladesh. *Oryx 35:* 61–72.
Smith, B.D. and Braulik, G.T. 2012. *Platanista gangetica*. The IUCN Red List of ThreatenedSpecies 2012: e.T41758A17355810. http://dx.doi.org/10.2305/ IUCN.UK.2012.RLTS.T41758A17355810.en
Sun, Y. B., Zhou, W. P., Liu, H. Q., Irwin, D. M., Shen, Y. Y., & Zhang, Y. P. (2013). Genome-widescans for candidate genes involved in the aquatic adaptation of dolphins. *Genome Biology andEvolution*, 5(1), 130–139. doi:10.1093/gbe/evs123
Turvey, ST et al. 2007 First human-caused extinction of a cetacean species? Biology Letters 3: 537-540
Willoughby, J. R., Ivy, J. a., Lacy, R. C., & DeWoody, J. A. (2017). The effects of inbreeding andselection on genomic diversity in captive populations: implications for the conservation of endangered species. *PLoS ONE*, 12(4), 1–17. doi:10.1371/journal.pone.0175996
Xiong, Y., Brandley, M.C., Xu, S., Zhou, K., Yang, G., 2009. Seven new dolphin mitochondrialgenomes and a time-calibrated phylogeny of whales. *BMC Evol. Biol.* 9.http://dx.doi.org/10.1186/1471-2148-1189-1120
Yim, H.-S., Cho, Y. S., Guang, X., Jeong, J.-Y., Cha, S.-S., et al. (2014). Minke whalegenome ndtion in cetaceans. *Nature Genetics, 46(1), 88–92.*doi:10.1038/ng.2835
Zhou X, Xu S, Yang Y, Zhou K, and Yang G (2011)Phylogenomic analyses and improved resolution of Cetartiodactyla. *Mol.Phylogenet. Evol*.61(2):255–264.doi: http://dx.doi.org/10.1016/j.ympev.2011.02.009
Zhou, X., Sun, F., Xu, S., Fan, G., Zhu, K., Liu, X., et al. (2013). Baiji genomes reveal low genetic variability and new insights into secondary aquatic adaptations. *Nature Communications,* 4, 2708. doi:10.1038/ncomms3708

Research Objective 10

Complete Mitochondrial Genome Sequence of *Platanista gangetica* from Halda river of Bangladesh

Abstract :

Platanista gangetica (Roxburgh 1801) Ganges River dolphins are an evolutionarily ancient cetacean threatened with extinction in the near future due to the absence of necessary conservation action.This genus is the solitary member of its family, Plantanistidae. Yet the phylogenetic relationships of this peculiar and poorly known dolphin among extinct and extant cetaceans and river dolphins has been the subject of conjecture for long. Therefore it becomes an obligation to sequence the mitochondrial genome of Ganges River dolphin. Mitochondrial genome (mitogenome) plays some influential roles in evolutionary and ecological studies. It becomes routine to utilize multiple genes on mitogenome or the entire mitogenomes to investigate phylogeny and biodiversity of focal groups with the onset of High Throughput Sequencing (HTS) technologies. Here, we describe the complete mitogenome of *Platanista gangetica gangetica*, derived from an animal stranded on Halda river, Bangladesh. Using an Illumina platform, we shot-gun sequenced the complete mitochondrial genome of Ganges river dolphin to an average coverage of 152X. We performed a de novo assembly using Novoplasty and determined the total mitogenome length to be 16,347 bp. The nucleotide composition was asymmetric (33.3% A, 24.6% C, 12.6% G, 29.5% T) with an overall GC content of 37.2%. The gene organization was similar to that of other cetaceans with 13 protein-coding genes, 2 rRNAs (12S and 16S), 22 predicted tRNAs and 1 control region or D-loop. Among the 37 genes, 28 were positioned on the H-strand and 9 were positioned on the L-strand.The entire mtDNA showed a slight AT rich bias (56.94%) with positive A-T skew (0.15) and negative G-C skew (-0.29).The phylogenetic relationships of 15 Odontoceti species were reconstructed based on the 13 protein-coding genes using the Bayesian inference method.No evidence of heteroplasmy or nuclear copies of mitochondrial DNA were found in this individual. These data will provide an insight into the genetic structure; besides contribute to resolving the phylogeography and population ecology of this sole living representative of Platanistidae family.

Inroduction:

Ganges River dolphin , *Platanista gangetica gangetica*(Superfamily Platanistoidea, Infraorder: Cetacea), belonging the order Artiodactyla, is one of the most charismatic and iconic species; which is endemic to one of the world's most biodiverse river basins The Ganges–Brahmaputra–Meghna and Karnaphuli-Sangu (GBMK River Basin) in India, Bangladesh and Nepal. These river dolphins prefer areas that create eddy countercurrents, such as small islands, river bends, and convergent tributaries (Moreno,

2003).Unfortunately, various anthropogenic and natural factors jeopardize the future of this relict freshwater cetacean.Modification of river streams (particularly from dams), alterations in sediment and nutrient fluxes, habitat destruction, river water contamination from urbanization and agriculture, boat traffic, illicit intentional hunting, and over exploitative fisheries are known to threaten the habitats of this species in South Asia (Payne and Temple 1996, Bannerjee 1999, Dudgeon 2000a, 2000b, 2005; Manel et al. 2000, Gergel et al. 2002). Thus, *Platanista gangetica* is categorized as endangered species as per IUCN global redlist (Braulik et al. 2017), is on appendix I of the CITES (CITES 2020).

The *P. gangetica* has been classified into two subspecies by different researchers elucidating ambiguities in their taxonomy. According to Smith and Braulik (Smith and Braulik, 2012; Braulik et al. 2014), The South Asian river dolphins (Ganges river dolphin: *Platanista gangetica gangetica* and Indus river dolphin: *P. g.minor*) are two closely related but geographically isolated, endangered freshwater cetaceans, currently classified as subspecies in a monotypic family (Platanistidae). An assessment of divergence rates in low mtDNA Diversity of the two subspecies indicates the long-term absence of gene flow and clear genetic differentiation between them and also shows that they diverged from a common ancestor around 550,000 years ago(95 % posterior probability 0.13–1.05 million years ago), possibly when dolphins from the Ganges dispersed into the Indus during drainage capture. This ancestor is thought to have been a marine Platanistid occupying the epi-continental seas in Southern part of Asia amid the sea level rises in the middle Miocene (Braulik et al., 2014).

Although there is very less physical similarities, molecular analyses demonstrated that the Platanistidae family is most closely related to the Kogiidae (dwarf and pygmy sperm whales) and Physeteridae (sperm whales) (McGowen et al. 2009; Steeman et al. 2009; Zhou et al. 2011). Among the four river dolphin families, the Platanistidae was the earliest divergent clade, the Lipotidae was the next, and then the Iniidae and Pontoporiidae(Kaiya, 1999). Since no affinity was revealed between the Platanistidae and the other river dolphin families; so, the freshwater cetaceans are paraphyletic, and genus *Platanista* Wagler, 1830 as well as Platanistidae got place at superfamily level. Although, Lawrence, G. B. (2006) stated that,the concept of superfamily Platanistoidea has become dramatically emended presently. However,In spite of several endeavors of utilizing paleontological, morphological and molecular approaches for the task of exact taxonomic position of this dolphin clade this issue remained still slightly uncertain (Kasuya, 1973; Milinkovitch et al., 1994; Arnason and Gullberg, 1996; Messenger and Mcguire, 1998; Cassens et al., 2000; Nikaido et al.,2001; Hamilton et al., 2001).

In this study, we disclosed the complete mitochondrial genome sequence of a healthy male adult *Platanista gangetica gangetica,* an individual originating in the tidal river Halda of Chittagong,Bangladesh.This river is the claimant of the country's national fish breeding heritage for being one of the most important natural carp spawning grounds in Bangladesh and has long been the major source of naturally produced carp fry for pond

culture in much of the country. A total of 83 finfish species under 13 orders and 35 families and a total of 10 shellfish (9 prawns and 1 crab) under 1 order and 3 families were identified from the river Halda within study period September 2004 to December 2011 (Alam et al., 2013).This immensely biodiversed ecosystem make the river a good home ground for the Ganges River Dolphin that feeds on fish and crustaceans,where the population is estimated to be 130-160 in number (reference). As Ganges River Dolphin is the apex predator in the river system where it lives (Klinowska 1991, Culik 2003), their population remains low and can be a good indicator of the health of the aquatic ecosystem where it inhabits (Gómez-Salazar 2012).

Comparative mitogenomic information has revolutionised several concepts of molecular phylogeny and evolution across multiple taxonomic levels (Miya and Nishida,2015).Mt-DNA is the ideal molecular marker for phylogenetic studies due to its conserved gene content, maternal inheritance(shared by everyone in maternal lineage), Absence of recombination, multiple copies in single cell, high rate of nucleotide substitution (Brown et al., 1979; Curole and Kocher, 1999; Olivo et al., 1983). Genetic information coupled with biological and behavioural data is pivotal for the preservation and management of imperiled species just like the Ganges river dolphin. In order to provide a theoretical foundation for the conservation strategy of *P. gangetica* within Platanistidae and new sight for further studies of phylogenetically-informative sequence data; in the current study the complete mtDNA of *P. gangetica gangetica* was sequenced, assembled and annotated, afterwards compared with other freshwater odontocetes

Materials and methods

Collection of Sample and DNA Extraction

Sampling and DNA isolation: Fresh sample of skin, liver and muscle tissues were collected from an adult male *Platanista gangetica* which was found dead from Akbaria point of Halda River under Hathazari upazila (Longitude / Latitude- 22.446661N, 91.861295E) on Thursday, December 5, 2019. Then, transported to the laboratory to preserve at -70° C for further analysis during the period of December 2019. The liver and muscle tissue was dissected and stored with 95% ethanol. Later for Sample preparation, a high molecular weight genomic DNA isolation and purification were performed following the AddPrep Genomic DNA extraction kit (AddBio, Korea) for future evaluation of the quality as well as quantity of the DNA. All the methods had been performed in accordance with the "Regulations for Animal Experiments in Chittagong Veterinary and Animal Sciences University's unique feature, the Indian and GOB ethical clearance" as required.

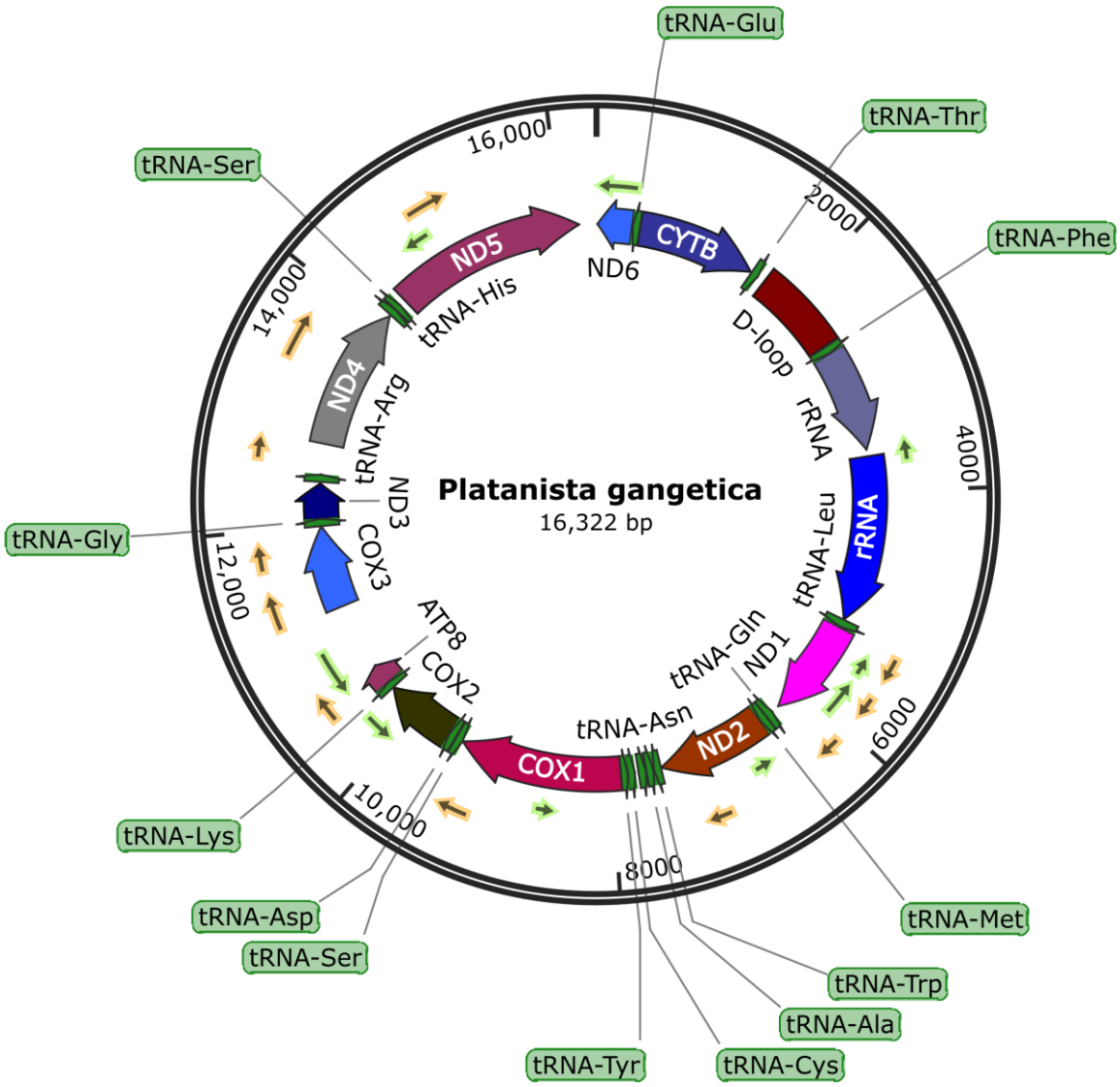


Figure 1. The circular mitochondrial gene map was drawn using SnapGene 5.2.2

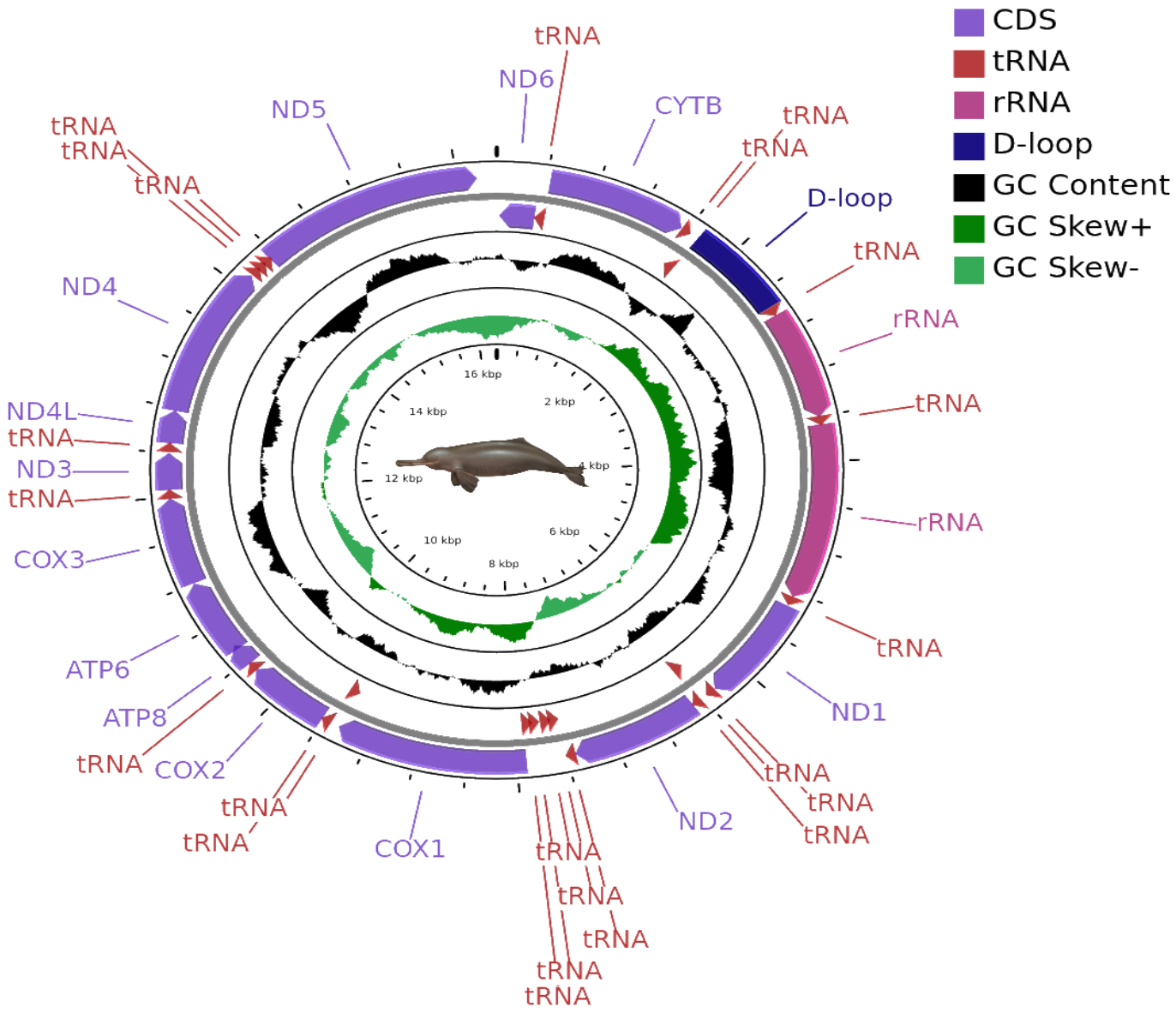


Figure 1. The circular representation of the complete mitochondrial genome of *P. gangetica.* Direction of gene *transcription is indicated by arrows in entire complete genome.The GC content is plotted using a black sliding window, as the deviation from the average GC content of the entire sequence. GC-skew is plotted using a colored sliding window (deep green and light green colour), as the deviation from the average GC skew of the entire sequence.*

Sequencing, assembly and annotation of mitochondrial genomes

Purified DNA was sequenced using Illumina NovaSeq 6000 platform from BGI, China. We used BWA V0.7.17 and SAMTOOLS V0.1.19 for separating the mitochondrial genome reads from the whole genome sequence by mapping it against the reference *Platanista gangetica* mitochondrial genbank accession (MF990206.1). The clean reads were assembled by using the organelle assembler NOVOPlasty V.4.0 (Dierckxsens et al., 2017). For functional and structural annotation web servers MITOS (Bernt et al., 2013) and GeSeq (Tillich et al., 2017) were used. The 22 tRNA genes secondary structure and location were determined by using MITOS and tRNA scan-SE. The circular image of mitochondrial genome was drawn by using online server CGView 29 (http://stothard.afns.ualberta.ca/cgview_server/). The length and locations of spacer regions (overlapping and intergenic) of *P.gangetica* mitochondrial genome were detected manually.. The nucleotide composition, codon usages, relative synonymous codon usage (RSCU) was done by MEGAX (Kumar, et al 2018).. To calculate the skewness, we used the formula: AT skew = (A − T)/(A + T) and GC skew = (G − C)/(G + C) (Perna et al., 1995).

Result and Discussion

Genome structure, organization and composition

P. g. gangetica complete mitochondrial genome (Submission 2405387) is 16,322 base pairs (bp) in length. It included 37 genes: 13 PCGs, large and small rRNAs, 22 tRNAs and one non-coding region (D-Loop) with the origin of light-strand replication (OL) (Fig. 1, Table 1). The majority strand contains 22 genes and minority with 15 genes (Table 1). The AT and GC content of nucleotide was 60% and 40%, respectively (Table S6), like other freshwater dolphin mitochondrial genomes. The highest AT content observed in tRNAs (69.68%,), followed by PCGs (68.72%), rRNAs (68.66%), and CR (67.67%). The mitochondrial genome showed positive AT (0.07) and negative GC (-0.36) skewness in contrast to other odontocetes mitochondrial genome.

To compare the mitogenomes of P. gangetica *and other related species, the orthologous average nucleotide identity (OrthoANI) value was analyzed using ANI calculator (http://www.ezbiocloud.net/tools/ani). The overall gene content and arrangement in the* P. gangetica *mitogenome were almost identical to those from three river dolphin species, and the highest OrthoANI value was obtained in the Indus river dolphin (*Platanista minor*, NC_005275.1, 99.57%),then Yangtze river dolphin (Lipotes vexilifer, NC_007629.1, 84.92%), La plata dolphin (*Pontoporia blainvillei, *NC_005277.1, 84.59%) whereas the Amazon river dolphin (*Inia geofrensis*, NC_005276.1) showed the lowest similarity (84.27%).*

Protein-coding genes

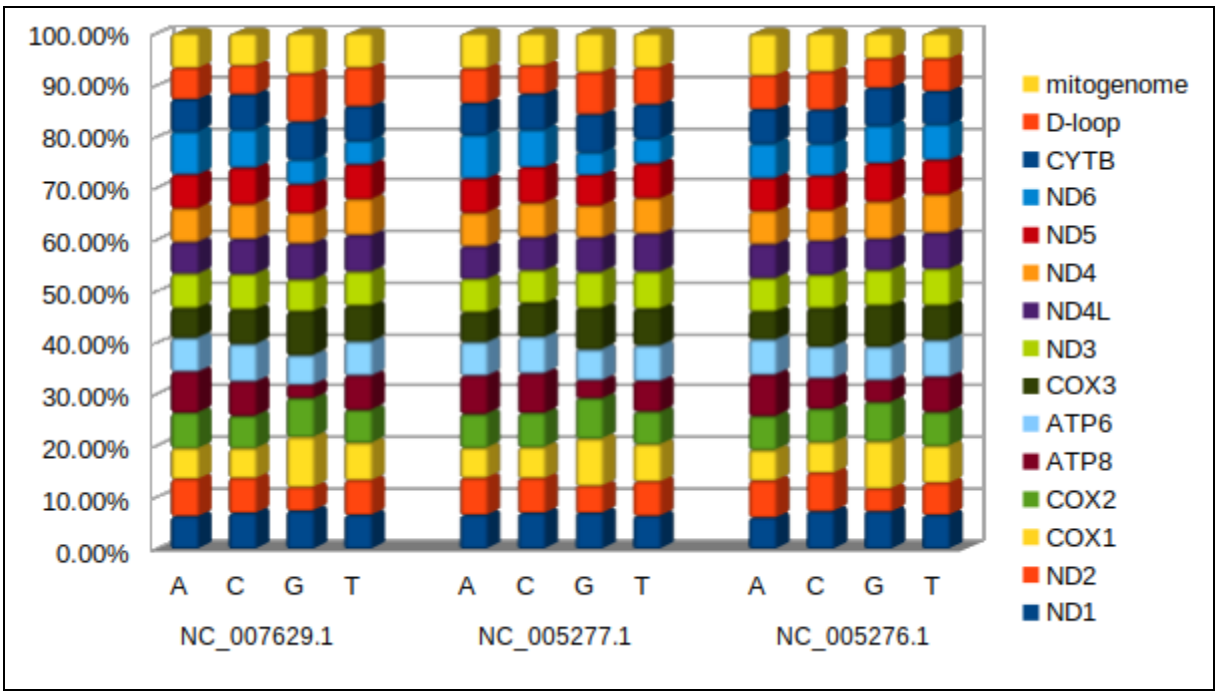
The total length of PCGs was 11406 bp in *P. gangetica* similar length as *Lipotes vexilifer* (NC_007629).All PCGs were encoded on the H-strand, except for ND6. Among the 13 PCGs, 12 genes had the typical ATG initiation codon, whereas the initiation codon GTG was found in ND2. Three types of stop codons were detected: AGA (CYTB), TAA (ND1, COI, COII, ATP6, COIII, ND3, ND4, ND4L, ND5, and ND6), and TAG (ND2, ATP8).

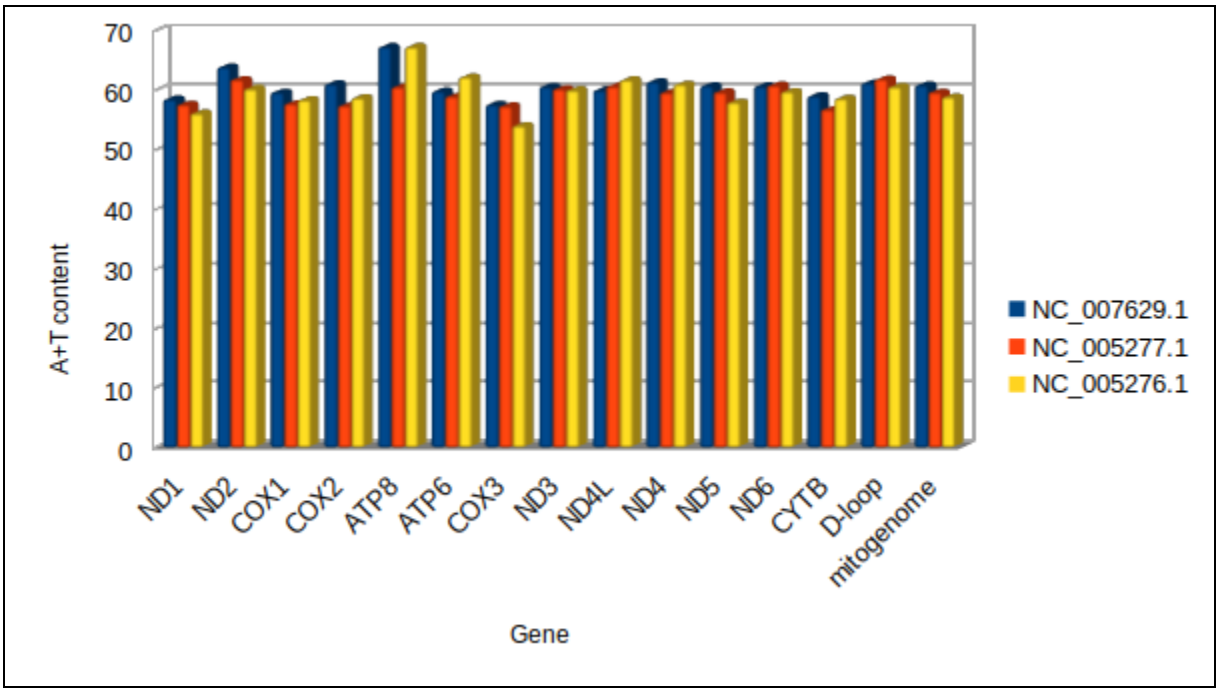
Codon usage bias and mutations.

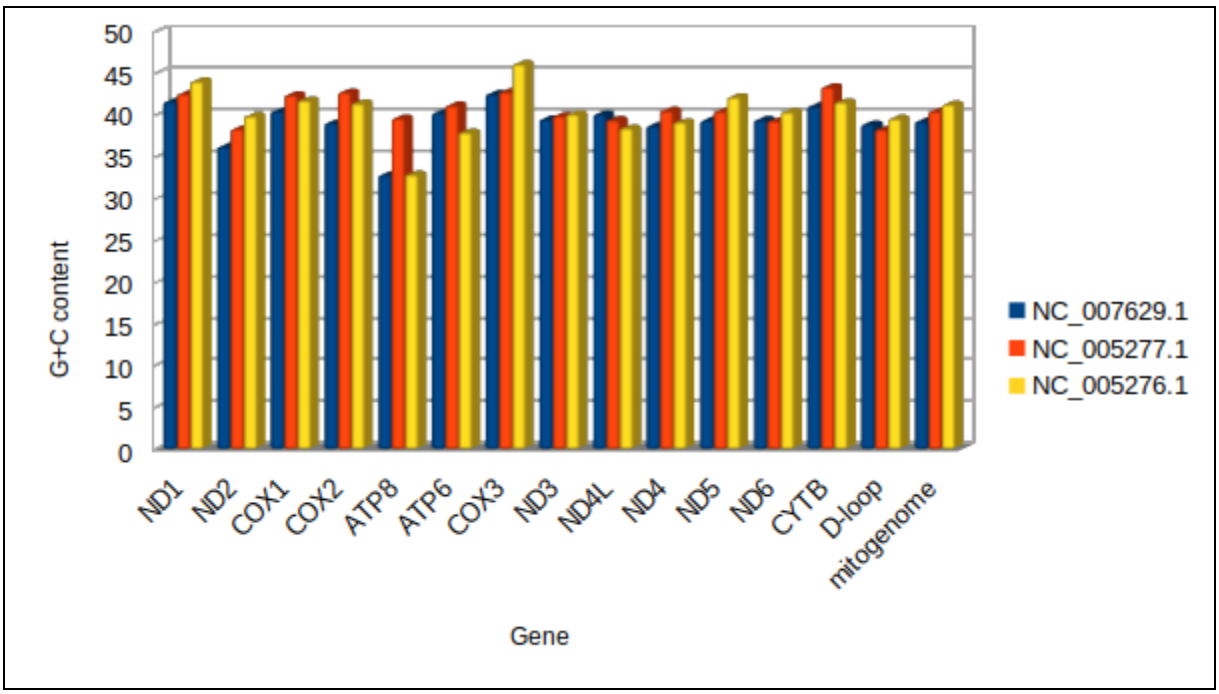
The use of codons or codon usage bias is a fundamental phenomenon in nature (Chakraborty et al., 2017;Whittle et al., 2016). The main influencing forces for codon usages are the mutation pressure and natural selection. Codon usage bias can be triggered by a number of other factors such as the content of nucleotides, gene length and their function, and the external environment (Whittle et al., 2016). We investigated the GC content to study the nucleotide distribution of all three codon positions of PCGs for Four river dolphin mitochondrial genomes (Fig. S1). The codon frequency ending with A/T is higher than G/C due to the AT rich segments which leads to the high codon bias (Mondal et al., 2017;Wei et al., 2014). The comparative study of river dolphin mitochondrial genomes revealed 21 codons (9 with A-ending, 12 with U-, and none with G- or C-ending) with high frequency (Table S8). This result suggested that compositional constrain may play an important role in the codon usage patterns in dolphin species.

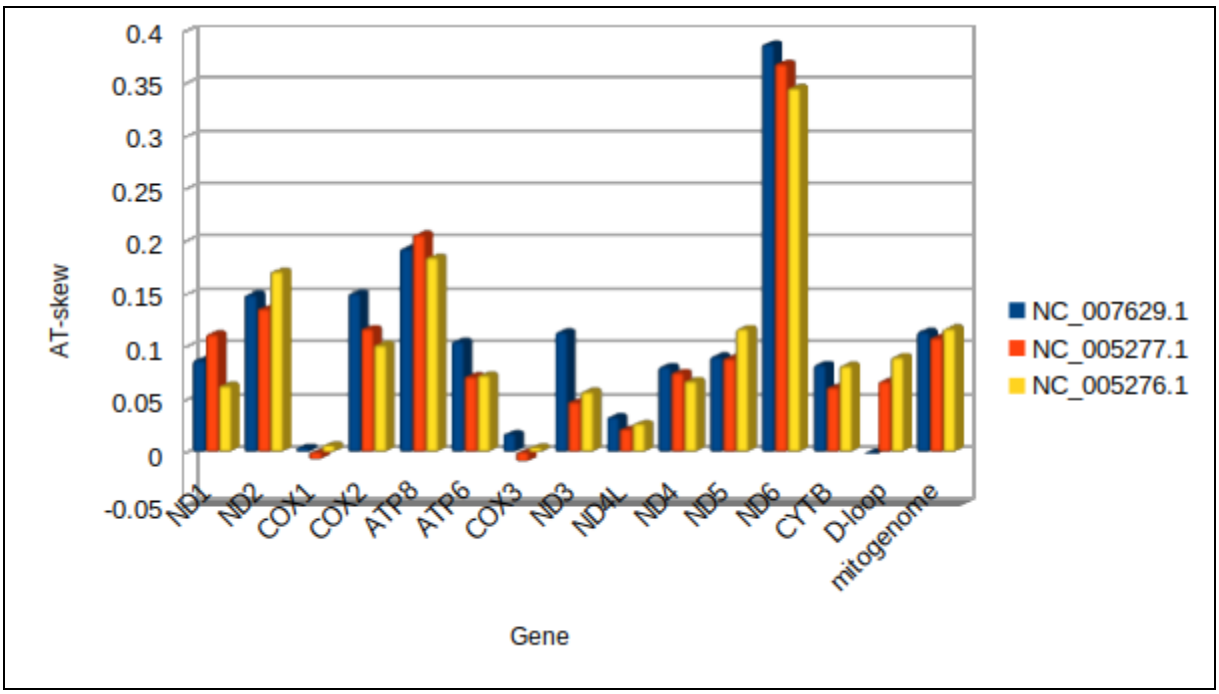
Ribosomal and transfer RNA genes.

Two rRNAs were observed in *P. g. gangetica* and other dolphin mitochondrial genome. The large ribosomal rrnL (16S RNA) placed between trnV and trnL1, was 1575 bp long; the small rrnS (12S RNA) between trnI and trnV, was 975 bp long (Table 1). *P. gangetica* has 22 tRNAs (total length 1519 bp), ranging from 60 bp (tRNA Ser(AGY) to 75 bp (tRNA Leu(UUR)) in length. (Table S5).


100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
A C G T
A C G T
A C G T
NC_007629.1
NC_005277.1
NC_005276.1
mitogenome
D-loop
CYTB
ND6
ND5
ND4
ND4L
ND3
COX3
ATP6
ATP8
COX2
COX1
ND2
ND1


70
60
50
40
30
20
10
0
A+T content
ND1
ND2
COX1
COX2
ATP8
ATP6
COX3
ND3
ND4L
ND4
ND5
ND6
CYTB
D-loop
mitogenome
Gene
NC_007629.1
NC_005277.1
NC_005276.1


G+C content
50
45
40
35
30
25
20
15
10
5
0
ND1
ND2
COX1
COX2
ATP8
ATP6
COX3
ND3
ND4L
ND4
ND5
ND6
CYTB
D-loop
mitogenome
Gene
NC_007629.1
NC_005277.1
NC_005276.1


AT-skew
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
ND1
ND2
COX1
COX2
ATP8
ATP6
COX3
ND3
ND4L
ND4
ND5
ND6
CYTB
D-loop
mitogenome
Gene
NC_007629.1
NC_005277.1
NC_005276.1

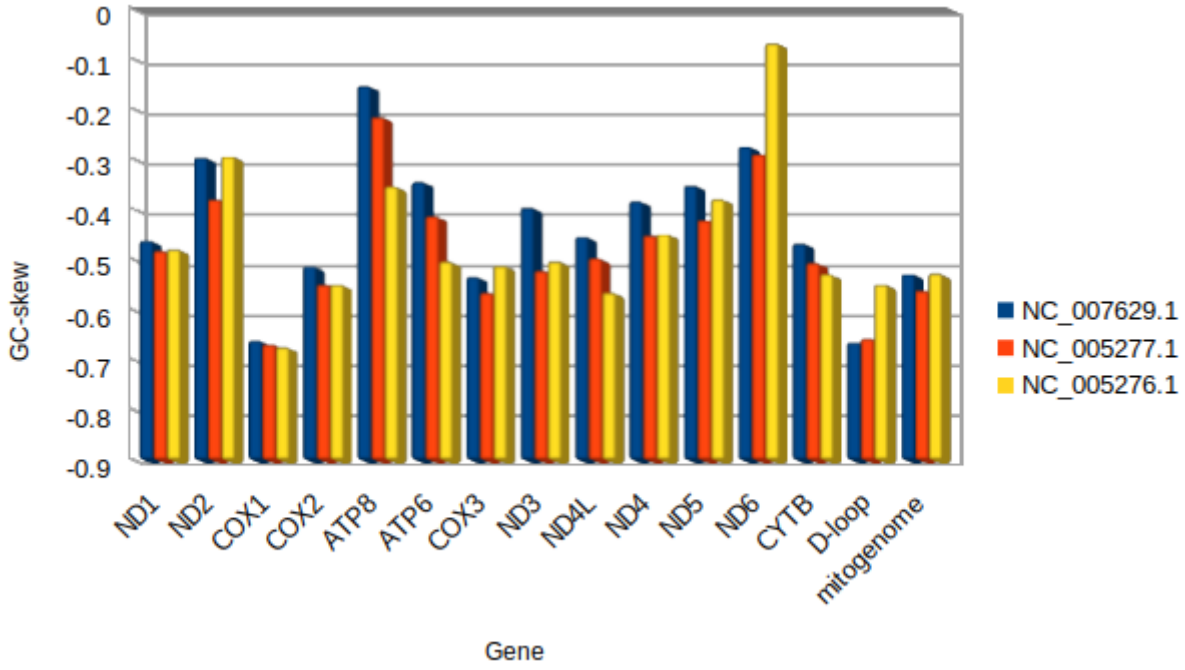
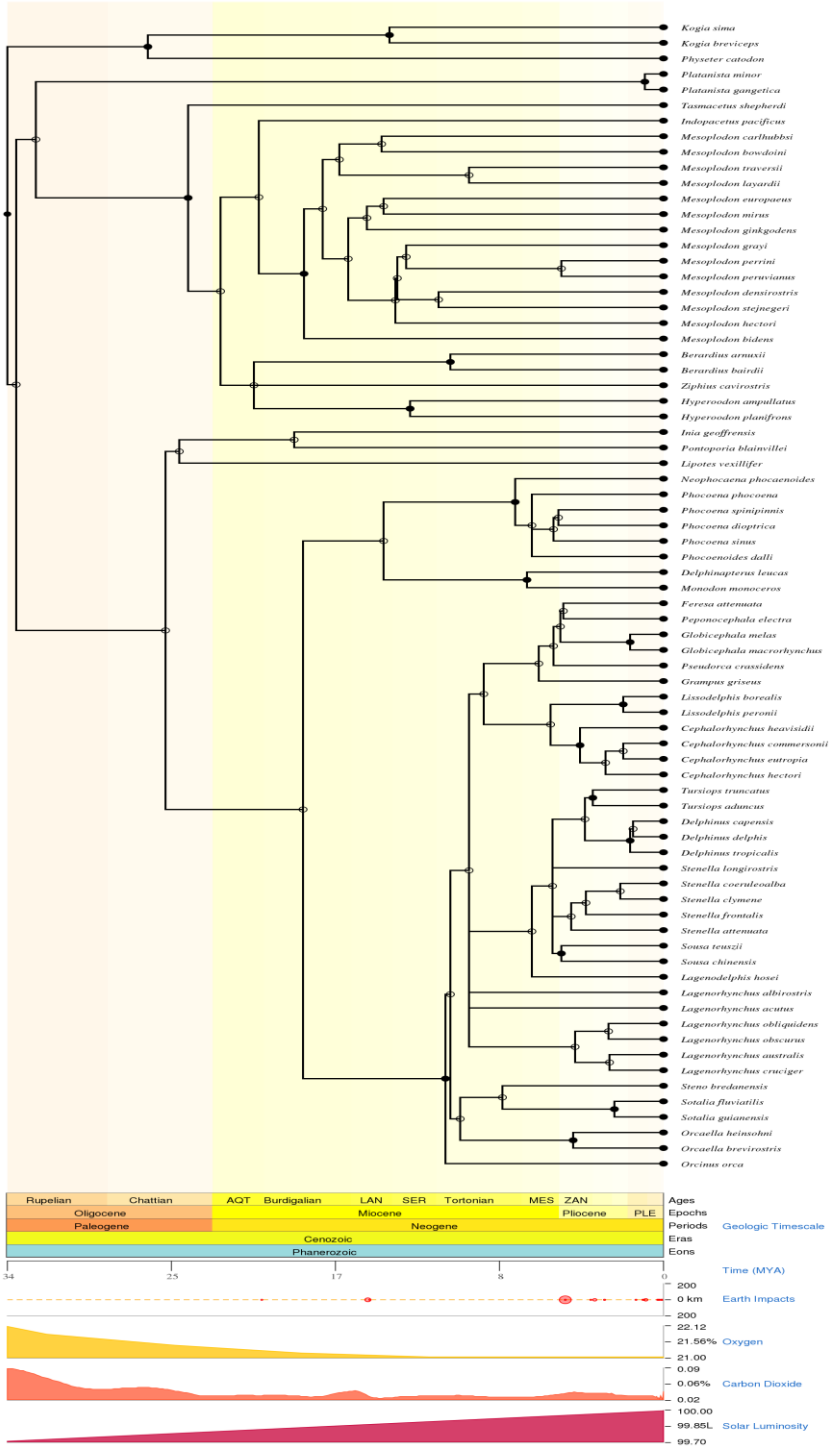


Figure 2: Comparative analysis of Nucleotide composition, AT and GC percentage and skewness of four river dolphin species Yangtze River dolphin (NC_007629.1), franciscana or la plata dolphin(NC_005277.1), Amazon river dolphin (NC_005276.1) and Ganges river dolphin)

The *P. g. gangetica* mitogenome (Submission 2405387) was 16,322 bp in length (60% A+T content), and consisted of the typical set of 37 genes (two rRNAs, 22 tRNAs, and 13 PCGs) which were very similar to those of other freshwater dolphin species in the Infraorder Odontoceti (Supplementary Fig. 1). The two rRNAs were 975 bp (12S rRNA) and 1575 bp (16S rRNA) in length, and separated by tRNA Val (67 bp) . The size of the 22 tRNAs varied from 60 bp (tRNA Ser(AGY)) to 75 bp (tRNA Leu(UUR)), with a total length of 1519 bp. Eleven intergenic spacers (1–7 bp), with a total length of 32 bp, and eight intergenic overlapping regions (1–31 bp), with a total length of 68 bp were detected (Table 1).


Kogia sima
Kogia breviceps
Physeter catodon
Platanista minor
Platanista gangetica
Tasmacetus shepherdi
Indopacetus pacificus
Mesoplodon carlhubbsi
Mesoplodon bowdoini
Mesoplodon traversii
Mesoplodon layardii
Mesoplodon europaeus
Mesoplodon mirus
Mesoplodon ginkgodens
Mesoplodon grayi
Mesoplodon perrini
Mesoplodon peruvianus
Mesoplodon densirostris
Mesoplodon stejnegeri
Mesoplodon hectori
Mesoplodon bidens
Berardius arnuxii
Berardius bairdii
Ziphius cavirostris
Hyperoodon ampullatus
Hyperoodon planifrons
Inia geoffrensis
Pontoporia blainvillei
Lipotes vexillifer
Neophocaena phocaenoides
Phocoena phocoena
Phocoena spinipinnis
Phocoena dioptrica
Phocoena sinus
Phocoenoides dalli
Delphinapterus leucas
Monodon monoceros
Feresa attenuata
Peponocephala electra
Globicephala melas
Globicephala macrorhynchus
Pseudorca crassidens
Grampus griseus
Lissodelphis borealis
Lissodelphis peronii
Cephalorhynchus heavisidii
Cephalorhynchus commersonii
Cephalorhynchus eutropia
Cephalorhynchus hectori
Tursiops truncatus
Tursiops aduncus
Delphinus capensis
Delphinus delphis
Delphinus tropicalis
Stenella longirostris
Stenella coeruleoalba
Stenella clymene
Stenella frontalis
Stenella attenuata
Sousa teuszii
Sousa chinensis
Lagenodelphis hosei
Lagenorhynchus albirostris
Lagenorhynchus acutus
Lagenorhynchus obliquidens
Lagenorhynchus obscurus
Lagenorhynchus australis
Lagenorhynchus cruciger
Steno bredanensis
Sotalia fluviatilis
Sotalia guianensis
Orcaella heinsohni
Orcaella brevirostris
Orcinus orca
Rupelian
Chattian
AQT
Burdigalian
LAN
SER
Tortonian
MES
ZAN
Oligocene
Miocene
Pliocene
PLE
Paleogene
Neogene
Cenozoic
Phanerozoic
Ages
Epochs
Periods
Eras
Eons
Geologic Timescale
34
25
17
8
0
Time (MYA)
200
0 km
200
Earth Impacts
22.12
21.56%
21.00
Oxygen
0.09
0.06%
0.02
Carbon Dioxide
100.00
99.85L
99.70
Solar Luminosity

Phylogenetic analyses

The phylogenetic analysis was performed with MegaX(Kumar, et al 2018) and edited on timetree server using mitochondrial genome among all odontocete species. The topological structure generated from both methods and analyses yielded similar results. The tree revealed that Platanista clustered with Ziphiidae clade. Within cetaceans, the tree topology wasIniidae,Pontoporiidae),(Delphinidaea,Phocoenidae,Monodontidae). ,Lipotidae),((Platanistidae,Ziphiidae),((Kogiidae,Physeteridae),Mysticeti). The two sub-species of *Platanista. gangetica* and *P.minor* clustered together as subspecies.

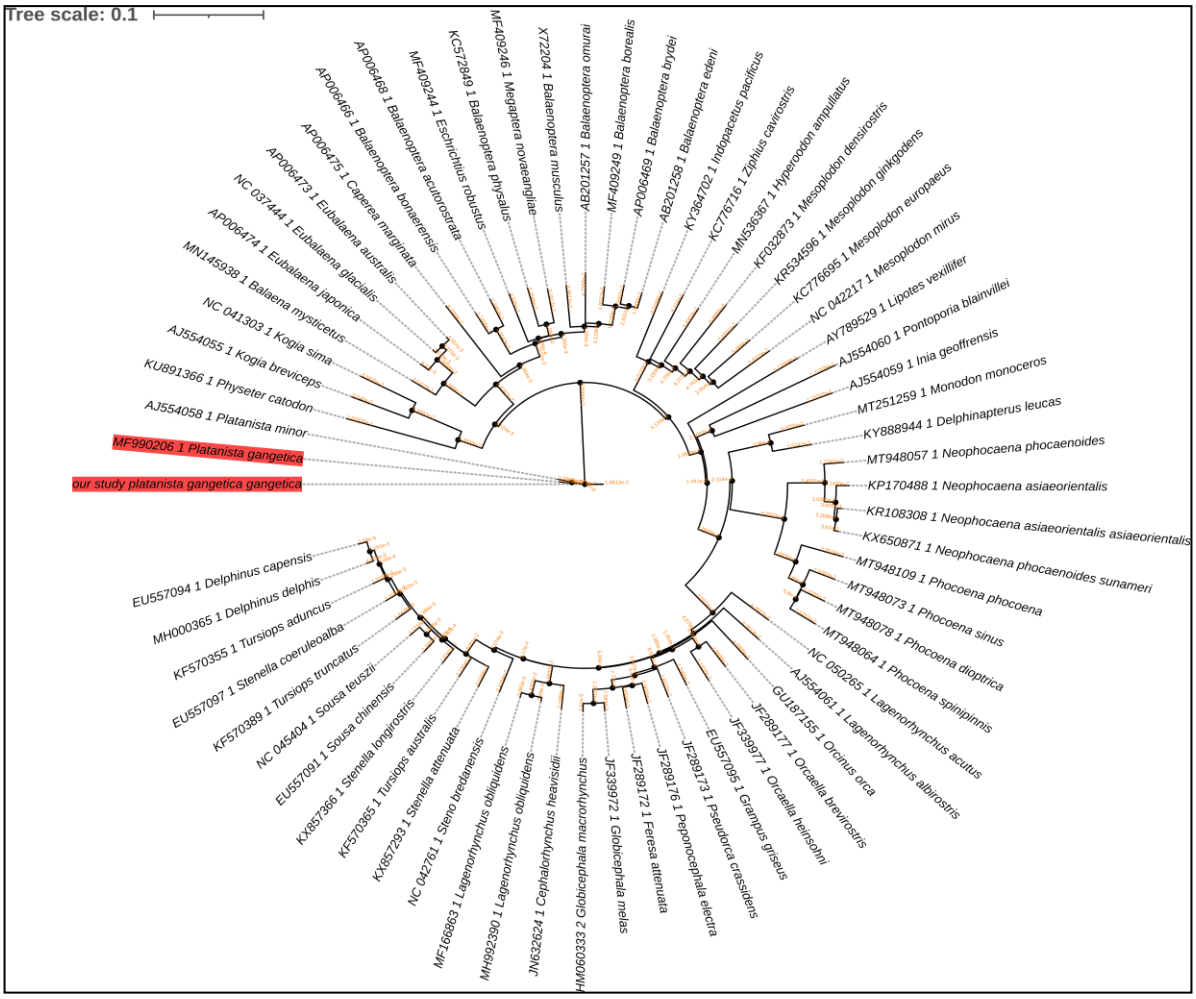


Figure : Phylogram of retained BIONJ Tree based on the available Odontoceti mitogenomes. The tree can be viewed interactively or downloaded from https://itol.embl.de/tree/1032301051213773160917786 4#

REFERENCES

Braulik, G. T.; Barnett, R.; Odon, V.; Islas-Villanueva, V.; Hoelzel, A. R.; Graves, J. A. (2014). "One Species or Two? Vicariance, Lineage Divergence and Low mtDNA Diversity in Geographically Isolated Populations of South Asian River Dolphin". *Journal of Mammalian Evolution*. **22** (1): 111–120. doi:10.1007/s10914-014-9265-6

Braulik, G.T. & Smith, B.D. 2017. Platanista gangetica. The IUCN Red List of Threatened Species 2017: e.T41758A50383612. http://dx.doi.org/10.2305/IUCN.UK.2017-3.RLTS.T41758A50383612.en

Chakraborty, S., Uddin, A. & Choudhury, M. N. Factors affecting the codon usage bias of SRY gene acrossmammals. Gene. 630,13–20 (2017).

CITES 2020. CITES Appendices I, II and III. <https://cites.org/sites/default/files/eng/app/2020/E-Appendices-2020-08-28.pdf> Downloaded on December 01, 2020.

CULIK, B. 2003. Review on small cetaceans: distribution, behaviour, migration andthreats. Compiled for CMS/UNEP. (available from CMS website).

GÓMEZ-SALAZAR, C. 2012. River Dolphins as Indicators of Ecosystem Degradation inLarge Tropical Rivers. PhD Thesis. Dalhousie University, Halifax, Nova Scotia,

Canada. http://dalspace.library.dal.ca:8080/bitstream/handle/10222/14446/GomezSalazar,%20Catalina,% 20PhD,%20BIOL,%20Mar%202012.pdf?sequence=1.

Kaiya, Y. G. Z. (1999). A STUDY ON THE MOLECULAR PHYLOGENY OF RIVER DOLPHINS [J]. *ACTA THERIOLOGICA SINICA*, *1*.

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K., 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution, 35(6), pp.1547-1549.

Miya, M. and Nishida, M. 2015.The mitogenomic contributions to molecular phylogenetics and evolution of fishes: a 15year retrospect. Ichthyol. Res, 62(2): 29-71. DOI: 10.1007/s10228-014-0440-9.

Moreno, P. 2003. Ganges and Indus Dolphins. Pp. 13-17 in M Hutchins, D Kleiman, V Geist, J Murphy, D Thoney, eds. *Grzimek's Animal Life Encyclopedia*, Vol. 15, 2 Edition. Farmington Hills: Gale Group.

Saito, S., Tamura, K. & Aotsuka, (2005). T. Replication origin of mitochondrial DNA in insects. Genetics. **171**, 1695–1705 .

Lawrence, G. B. (2006). A phylogenetic analysis of the superfamily Platanistoidea (Mammalia, Cetacea, Odontoceti).— Beitr. Palaont., 30:25-42, Wien.

Perna, N. T. & Kocher, T. D. (1995) . Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *J. Mol. Evol. **41**, 353–358 .*

Whittle, C. A. & Extavour, C. G. Expression-Linked Patterns of Codon Usage, Amino Acid Frequency, and Protein Length in theBasally Branching Arthropod Parasteatoda tepidariorum. Genome Biol. Evol. 8, 2722–2736 (2016).

Mondal, S. K., Kundu, S., Das, R. & Roy, S. J (2016). Analysis of phylogeny and codon usage bias and relationship of GC content, amino acidcomposition with expression of the structural nif genes. Biomol. Struct. Dyn. 34, 1649–66

Wei, L. et al (2014). Analysis of codon usage bias of mitochondrial genome in Bombyx mori and its relation to evolution. BMC Evol. Biol.14, 262

Zhang, D. X. & Hewitt, G. M (1997). Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary Studies. Biochem. Syst. Ecol. 25, 99–120

List of publications from the project

(Published and manuscript prepared)

1. MM Kibria, N. Islam, KSM Shawrob, M billah, MH Rumi, AMAMZ Siddiki **(**2020) "Complete mitochondrial genome sequence of *Catla catla* (Hamilton, 1822) from the Halda river of Bangladesh". Mitochondrial DNA Part B 5(3), 3233-3235. https://doi.org/10.1080/23802359.2020.1809542 (Published)

2. AMAMZ Siddiki, S Akter, AA Asek, SS Rahman, MAB Bhuiyan and MM Kibria (2021) **Complete Mitochondrial Genome Sequence of *Catla catla* (Hamilton, 1822) from Halda river of Bangladesh**. (Submitted)

3. AMAMZ Siddiki, S Akter, A Kabir, SS Rahman, MAB Bhuiyan and MM Kibria (2021) **Complete Mitochondrial Genome Sequence of *Labeo rohita* (Hamilton, 1822) from Halda river of Bangladesh**. (Submitted)

4. MM Kibria, S Akter, AA Asek, SS Rahman, MAB Bhuiyan and AMAMZ Siddiki (2021) **Whole Genome of *Labeo rohita* (Hamilton, 1822) from Halda river of Bangladesh.** (Submitted)

5. MM Kibria, SS Mostafa, N Islam, SS Rahman, MAB Bhuiyan and AMAMZ Siddiki (2021) **Complete mitochondrial genome sequence of *Cirrhinus cirrhosus* (Bloch, 1795) from Halda river of Bangladesh**. (Submitted)

6. AMAMZ Siddiki, SS Mostafa, N Islam, SS Rahman, MAB Bhuiyan and MM Kibria (2021) **Whole Genome of *Cirrhinus cirrhosus* (Bloch, 1795) from the river Halda, Bangladesh.** (Submitted)

7. AMAMZ Siddiki, AA Asek, SS Mostafa, SS Rahman, MAB Bhuiyan and MM Kibria (2021) **Complete mitochondrial genome sequence of *Labeo calbasu* (Hamilton, 1795) from the river Halda, Bangladesh.** (Submitted)

8. MM Kibria, AA Asek, S Akter, SS Rahman, MAB Bhuiyan and AMAMZ Siddiki (2021) **Whole genome sequence of *Labeo calbasu* (Hamilton, 1795) from Halda river of Bangladesh**. (Submitted)

9. MM Kibria, A Kabir, S Akter, SS Rahman, MAB Bhuiyan and AMAMZ Siddiki (2021) **Whole Genome Sequence of *Platanista gangetica* from Halda river of Bangladesh.** (Submitted)

10. AMAMZ Siddiki, A Kabir, AA Asek, SS Rahman, MAB Bhuiyan and MM Kibria (2021) **Complete mitochondrial genome sequence of *Platanista gangetica* from Halda river of Bangladesh**. (Submitted)

Project activities in pictures


Unusual death of Dolphin (*Platanista gangetica*)


Dolphin postmortem at Halda River Research Laboratory, University of Chittagong

Rescued brood fish from Halda River

Handover program of one brood fish to Halda River Research Laboratory by Raozan Upazila Administration (30th May, 2017)

Measurement of Rui (*Labeo rohita)*

Measurement of Catla (*Catla catla)*

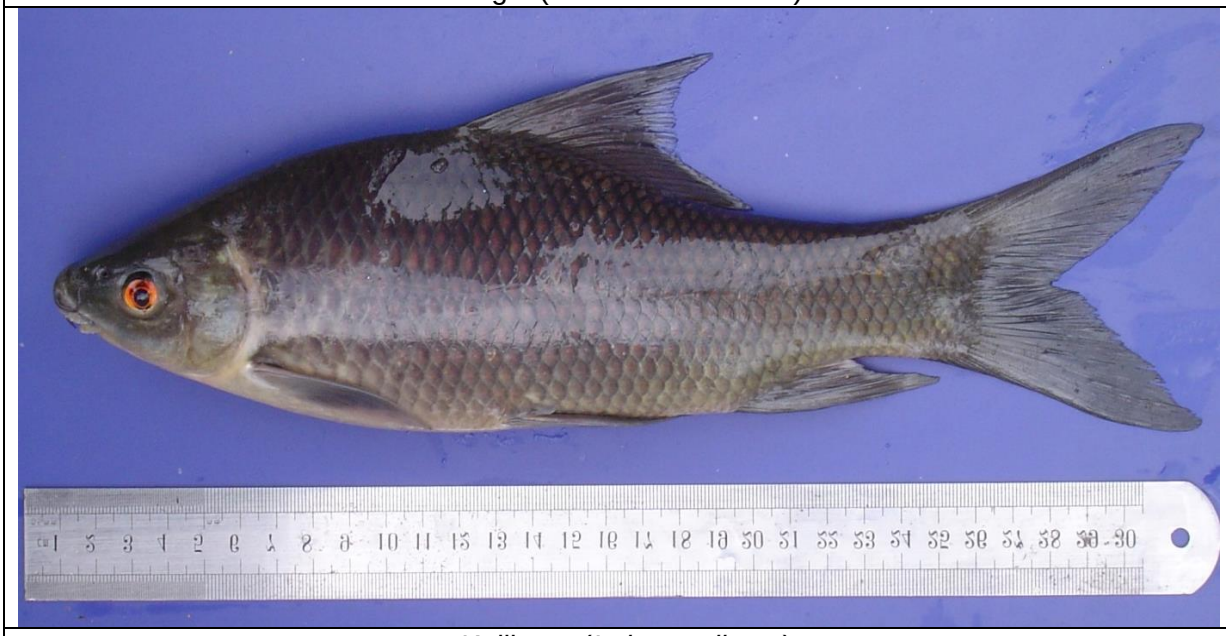
Measurement of Mrigel (*Cirrhinus cirrhosus)*

Measurement of Kalibaus (*Labeo calbasu)*

Rui (*Labeo rohita*)

Catla (*Catla catla*)

Mrigel (*Cirrhinus cirrhosus*)



Kalibaus (*Labeo calbasu*)

Dolphin (*Platanista gangetica*)

Tissue & Liver collection from Dolphin

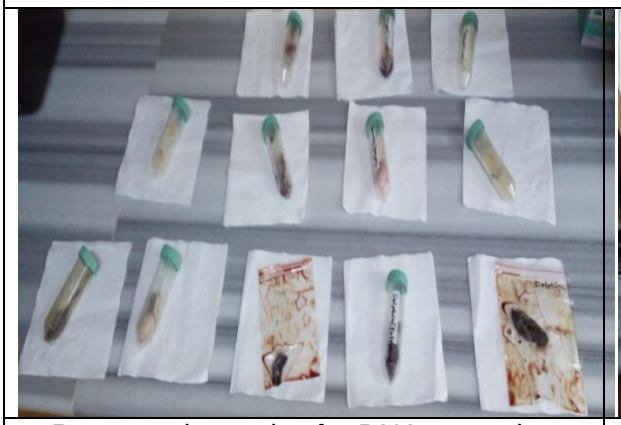
Tissue & Liver collection from Dolphin

Tissue, Liver, Scale collection from Carps

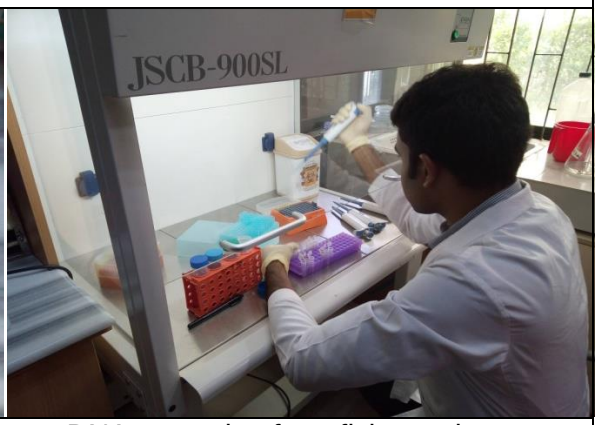
Tissue, Liver and Scale collection from Carps

Sample preparation for DNA extraction

Sample preparation for DNA extraction before sending DNA to whole genome sequencing



Processed samples for DNA extraction



DNA extraction from fish specimen

Training on Genome assembly

Bioinformatics research ongoing

Picture: The Halda Fish Genome Research Team with Team Leader Prof. M M Kibria and Prof. AMAM Zonaed Siddiki

Whole genome data resources in NCBI

Catla catla

Bioproject - https://www.ncbi.nlm.nih.gov/bioproject/623322
Mitogenome Data - https://www.ncbi.nlm.nih.gov/nuccore/MT303069
Whole Genome Sequence - https://www.ncbi.nlm.nih.gov/sra/SRX8626849[accn]
https://www.ncbi.nlm.nih.gov/sra/SRX8063618[accn]
Whole Genome Assembly - https://www.ncbi.nlm.nih.gov/assembly/GCA_014525385.1#/st

Labeo ruhita

Bioproject – https://www.ncbi.nlm.nih.gov/bioproject/657820
Mitogenome Data - https://www.ncbi.nlm.nih.gov/sra/SRX9456401%5Baccn%5D
https://www.ncbi.nlm.nih.gov/bioproject/660899
Whole Genome Sequence - https://www.ncbi.nlm.nih.gov/sra/SRX8968640%5Baccn%5D
Whole Genome Assembly: https://www.ncbi.nlm.nih.gov/assembly/GCA_017311145.1

Cirrhinus cirrhosus

Bioproject – https://www.ncbi.nlm.nih.gov/bioproject/688724
https://www.ncbi.nlm.nih.gov/bioproject/690704
Mitogenome Data - https://www.ncbi.nlm.nih.gov/bioproject/660899
Whole Genome Sequence:
https://www.ncbi.nlm.nih.gov/sra?LinkName=biosample_sra&from_uid=18021126
https://www.ncbi.nlm.nih.gov/sra?LinkName=biosample_sra&from_uid=17187123
Whole Genome Assembly: https://www.ncbi.nlm.nih.gov/assembly/GCA_019207145.1#/st

Labeo calbasu

Bioproject – https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA689123
Biosample - https://www.ncbi.nlm.nih.gov/biosample/?term=SAMN17199932
Mitogenome Data – Submitted at NCBI
Whole Genome Sequence:
https://www.ncbi.nlm.nih.gov/sra?LinkName=biosample_sra&from_uid=19333274
Whole Genome Assembly: Submitted at NCBI

Platanista gangetica

Bioproject – https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA675309
SRA Data - https://www.ncbi.nlm.nih.gov/sra/SRX9456846[accn]
Mitogenome Data - https://doi.org/10.6084/m9.figshare.13536764
Whole Genome Sequence –
https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=675309
Whole Genome Assembly: https://www.ncbi.nlm.nih.gov/assembly/GCA_017311385.1

Conclusion and recommendation

The genome sequencing and assembly of the Halda carps will provide a valuable tool for future population genetics and fish genomics studies, which will allow for targeting specific genes and particularly interesting regions of the specific genome. For instance, each species could be a model organism in which to study inbreeding and aquaculture of major carps as it is the highest common Indian major Carp (IMC) species. As shown in the recent special issue of Science, sequencing new fish genomes can lead to a better understanding of crucial aspects of the biology and ethology of fish. In the case of the different carp species, it could gather new information on the genetic diversity of the species and infer the compelling population size of each of the putative populations.

The availability of the draft genomes can offer assistance to gather the evolutionary history of the species, i.e. how high and low have the population sizes been within the past. Besides the estimation of genetic variability, this is typically a critical point because species that have experienced low population sizes in the past might be more vulnerable to human threats and more inclined to extinction.

At field level, to avoid genetic contaminations of the seed production, the genetic variability within a population is extremely useful to gather the information on individual identity, breeding pattern, degree of relatedness and distribution of genetic variation among them along with evolutionary and adaptive behaviour. The present research project was intended at developing the whole genome sequences of Halda riverine wild populations of several carp species to explore fish genomics with a view to generate baseline datasets. These will eventually facilitate conservation of genetic materials, crossbreeding along with avoiding intergeneric hybridization among the wild species and to encourage pure gene strain of this indigenous species for Bangladesh.

The *Platanista* (River Dolphin) sequencing data is comparable to other high quality cetacean genomes as reported elsewhere. Present study generated and analyzed the first whole genome assembly and annotation of *Platanista gangetica gangetica*. This genome adds to available Cetacean genomes by supplying essential resources for advance investigation within the Odontoceti and Cetacea. The acquired data should facilitate further studies of the genetic basis of divergence between two subspecies, *Indus* and *Ganges* dolphin and of the molecular differences between freshwater, marine, and terrestrial mammals. This data will also furnish invaluable information for further genetic studies of this species, both for whole-genome investigations into population structure and to denote key genes associated with local adaptation. Moreover, these data will provide an insight into the genetic structure along with resolving the phylogeography and population ecology of this sole living representative of Platanistidae family.

Future research areas (projected)

Fish, as a highly diversified group of the vertebrate family, shares a range of environmental conditions to which their physiologies, body shapes, and lifestyles have adapted over the years. In their aquatic habitat, water is in direct contact with several tissues and internal compartments of the animal, potentially inducing a high sensitivity to water-borne parameters such as temperature, oxygen levels, salinity, and sometimes toxic chemicals. This intimate relationship between an organism and a wide range of different environments has recently prompted the view that fish could be used as models for "environmental genomics", in other words the study of the interface between an organism and its environment using genomic approaches. **The quality of Halda river water and habitats can be verified through such novel research where fish genetics data can be employed.**

Up to now fish genomics can be used to exploit **the similarities and the differences between mammalian and fish genomes to gain profound insights into the evolution of vertebrate genomes in general, and into the function of individual genes** often associated with human disorders in particular. This is a new frontier of research and future studies will certainly based on such facts where each genome can provide complementary information.

The Government of Bangladesh, as well as international communities, has a particular interest in Halda river population conservation. Present study will provide a useful platform for the functional genome and conservation research of Halda river carps and other aquatic animals including Dolphins in the future. **The knowledge will help develop a better policy for their breeding, behavioral pattern analyses with a view towards their *in situ* conservation.**

In the coming decades, technologies like the applications of genomic techniques such as genome editing and genomic selection, along with the use of emerging intelligence systems, in aquaculture and fisheries will contribute significantly to genetic improvements of farmed fish and sustainable exploitation of fishery resources. **Present research will be invaluable in providing baseline molecular datasets (at genomic and proteomic level) for such future research initiatives.**

Molecular markers play an essential role in the selection and breeding programs in aquaculture and have been broadly used to construct the linkage maps of important economic phenotypic traits such as growth, sex determination, and pathogen resistance. Genomics has brought new tools that can help address fundamental questions in fisheries management such as stock identification, population structure, and adaptive response to environmental change. The identification of SNP markers through NGS has enhanced the ability to trace fisheries recourse or products to their original locations, allowing regulation enforcement in some commercially important fish species. **Therefore using population genomics, we can identify markers of Halda carps towards their branding in the near future.** This will also have commercial implications to characterize high quality produce from Halda at national and international level.

Appendix 1 (Published article)

MITOCHONDRIAL DNA PART B
2020, VOL. 5, NO. 3, 3215–3217
https://doi.org/10.1080/23802359.2020.1809542

Taylor & Francis
Taylor & Francis Group

MITOGENOME ANNOUNCEMENT

OPEN ACCESS Check for updates

Complete mitochondrial genome sequence of *Catla catla* (Hamilton, 1822) from the Halda river of Bangladesh

M. M. Kibria[a,b], N. Islam[a,b], M. Billah[c,d], K. S. M. Shawrob[c,e], M. H. Rumi[c] and AMAM Zonaed Siddiki[c]

[a]Department of Zoology, University of Chittagong, Chittagong, Bangladesh; [b]Halda River Research Laboratory, University of Chittagong, Chittagong,Bangladesh; [c]Genomics Research Group, Chittagong Veterinary and Animal Sciences University (CVASU), Chittagong, Bangladesh; [d]College of Animal Science and Technology, Northwest A&F University, Yangling, China; [e]Department of Biotechnology, Inland Norway University of Applied Sciences, Elverum, Norway

ABSTRACT

Catla (*Catla catla*) is one of the fastest-growing major carp found in South Asia as well as Bangladesh. *Catla catla* is the second most popular indigenous carp species in the freshwater aquaculture industry of Bangladesh due to its relatively good taste and high market price. In this study, we disclosed the complete mitochondrial genome sequence of Bangladeshi Catla fish from Halda river located in Chittagong. The circular mitogenome of *Catla catla* is 16,597 bp in length and nucleotide composition is AT-based (72%), contains 37 genes including 13 protein-coding genes, 22 tRNA genes, 2 rRNA genes and a D-loop (control region).

ARTICLE HISTORY
Received 8 July 2020
Accepted 1 August 2020

KEYWORDS
Catla catla; mitochondrial genome; protein-coding gene; tRNA; rRNA

Introduction

Catla catla is a member of the Cyprinidae family, which is endemic to the perennial river network of northern India, the Indus plain and adjacent hills of Pakistan, Bangladesh, Nepal and Myanmar (Reddy 1999). It has become one of the most well-established fish populations of all the rivers, lakes and reservoirs where they have been introduced. The Halda River is located in South-East region of Bangladesh which is a major tributary of the river Karnaphuli in Chittagong district originated from the hilly Haldachora fountain at the Patachara hill ranges of Ramgarh in the Khagrachari hill and renowned for being the only natural spawning ground of Indian major carp in Bangladesh (Tsai et al. 1981; Akter and Ali 2012; Kabir et al. 2015). A major portion of the country's pond carp culture is dependent on these wild seed that has an important and potential contribution in the agro-based economic development, poverty alleviation, employment, supplying of animal protein and earning the foreign currency for the national sector (Azadi 1979, DoF 2005). *C. Catla* is one of the "Four famous Indian carp" of Halda river which has extensive demand in carp polyculture system among the fish farmers due to it's higher productivity rate and compatibility with other major carps, specific surface feeding habit that help increase water quality, enriched protein and vitamin content with lower calories, delicate flavor and consumer preference (Shafi and Quddus, 1982). For being small in size, high evolutionary rate, and maternal inheritance mood, the complete mitochondrial genome sequences provide insight into the assessment of wide variation in animals and the comparison of sequence data contribute to the exploration of improved markers for population ecological studies (Avise 1995; Zhou et al. 2009). Here we reported the entire mtDNA sequences of Catla *catla* from the Halda river.

The specimen was collected from Halda river, Chittagong (geographic coordinate: 22°33′34.7″ N 91°50′41.8″E). Fresh tissue (from muscle) sample was stored at −20 °C until used to isolate genomic DNA using commercial DNA extraction kit (AddBio, Korea) and the total DNA was stored with a voucher number (DPP/CVASU/2019-12-44). Purified DNA was sent for library preparation and sequencing through commercial suppliers. DNA was sequenced using Illumina NovaSeq 6000 platform from BGI, China. The mitochondrial genome reads were separated from the whole genome sequence by mapping it against the reference Catla mitochondrial datasets (KY419138) using SAMTOOLS. The organelle assembler NOVOPlasty V.2.7.2 (Dierckxsens et al. 2017) was used to assemble the clean reads. Web-based tools like MITOS (Bernt et al. 2013) and GeSeq (Tillich et al. 2017) were applied to perform structural and functional annotation. Another tool, OGDRAW was used to construct the circular representation of the entire mitogenome (Greiner et al. 2019). Finally, mtDNA sequences were aligned and a phylogenetic tree was constructed by using CLC Main Workbench.

The complete mitogenome of *Catla catla* (NCBI accession number **MT303069**) is 16,597 bp in length and consists of 13 protein-coding genes, two ribosomal RNA genes (rRNA), 22 transfer RNA (tRNA) genes, and a putative control region

CONTACT Amamz Siddiki zsiddiki@gmail.com Genomics Research Group, Chittagong Veterinary and Animal Sciences University (CVASU), Chittagong, Bangladesh