Assembly and variation analyses of *Clarias batrachus* mitogenome retrieved from WGS data and its phylogenetic relationship with other catfishes

Basdeo Kushwaha a, Ravindra Kumar a,⁎ Suyash Agarwal a, Manmohan Pandey a, N.S. Nagpure a, Mahender Singh a, Shreya Srivastava a, C.G. Joshi b, P. Das c, L. Sahoo c, P. Jayasankar c, P.K. Meher c, T.M. Shah b, A.B. Patel b, Namrata Patel b, P. Koringa b, Sofia Priyadarsani Das c, Siddhi Patnaik c, Amrita Bit c, Sarika d, M.A. Iquebal d, Dinesh Kumar d, J.K. Jena a

**A R T I C L E   I N F O**

Article history:
Received 24 February 2015
Revised 30 May 2015
Accepted 7 June 2015
Available online 17 June 2015

Keywords:
*Clarias batrachus*
Indian catfish
Mitogenome
Phylogenetics
WGS data

**A B S T R A C T**

Whole genome sequencing (WGS) using next generation sequencing technologies paves the way to sequence the mitochondrial genomes with greater ease and lesser time. Here, we used the WGS data of *Clarias batrachus*, generated from Roche 454 and Ion Torrent sequencing platforms, to assemble the complete mitogenome using both de novo and reference based approaches. Both the methods yielded almost similar results and the best assembled mitogenome was of 16,510 bp size (GenBank Acc. No. KM259918). The mitogenome annotation resulted in 13 coding genes, 22 tRNA genes, 2 rRNA genes and one control region, and the gene order was found to be identical with other catfishes. Variation analyses between assembled and the reference (GenBank Acc. No. NC_023923)) mitogenome revealed 51 variations. The phylogenetic analysis of coding DNA sequences resulted in 13 coding genes, 22 tRNA genes, 2 rRNA genes and one control region, and the gene order was found to be identical with other catfishes. Variation analyses between assembled and the reference (GenBank Acc. No. NC_023923) mitogenome revealed 51 variations. The phylogenetic analysis of coding DNA sequences resulted in 13 coding genes, 22 tRNA genes, 2 rRNA genes and one control region, and the gene order was found to be identical with other catfishes. Variation analyses between assembled and the reference (GenBank Acc. No. NC_023923) mitogenome revealed 51 variations. The phylogenetic analysis of coding DNA sequences resulted in 13 coding genes, 22 tRNA genes, 2 rRNA genes and one control region, and the gene order was found to be identical with other catfishes. Variation analyses between assembled and the reference (GenBank Acc. No. NC_023923) mitogenome revealed 51 variations. The phylogenetic analysis of coding DNA sequences resulted in 13 coding genes, 22 tRNA genes, 2 rRNA genes and one control region, and the gene order was found to be identical with other catfishes. Variation analyses between assembled and the reference (GenBank Acc. No. NC_023923) mitogenome revealed 51 variations. The phylogenetic analysis of coding DNA sequences resulted in 13 coding genes, 22 tRNA genes, 2 rRNA genes and one control region, and the gene order was found to be identical with other catfishes. Variation analyses between assembled and the reference (GenBank Acc. No. NC_023923) mitogenome revealed 51 variations. The phylogenetic analysis of coding DNA sequences resulted in 13 coding genes, 22 tRNA genes, 2 rRNA genes and one control region, and the gene order was found to be identical with other catfishes. Variation analyses between assembled and the reference (GenBank Acc. No. NC_023923) mitogenome revealed 51 variations. The phylogenetic analysis of coding DNA sequences resulted in 13 coding genes, 22 tRNA genes, 2 rRNA genes and one control region, and the gene order was found to be identical with other catfishes. Variation analyses between assembled and the reference (GenBank Acc. No. NC_023923) mitogenome revealed 51 variations. The phylogenetic analysis of coding DNA sequences resulted in 13 coding genes, 22 tRNA genes, 2 rRNA genes and one control region, and the gene order was found to be identical with other catfishes. Variation analyses between assembled and the reference (GenBank Acc. No. NC_023923) mitogenome revealed 51 variations. The phylogenetic analysis of coding DNA sequences resulted in 13 coding genes, 22 tRNA genes, 2 rRNA genes and one control region, and the gene order was found to be identical with other catfishes. Variation analyses between assembled and the reference (GenBank Acc. No. NC_023923) mitogenome revealed 51 variations. The phylogenetic analysis of coding DNA sequences resulted in 13 coding genes, 22 tRNA genes, 2 rRNA genes and one control region, and the gene order was found to be identical with other catfishes.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

**Introduction**

The *Clarias batrachus* is one of the most popular catfish owing to its characteristics such as taste, protein flavor, low fat content, medicinal value etc. The fish can thrive well in low oxygen conditions and its culture is economically viable (Debnath, 2011; Goswami, 2007; Hossain and Parween, 2006; Singh and Hughes, 1971). The occurrence of this species in natural water bodies has greatly been reduced due to over-exploitation and disease outbreaks (Ahmad et al., 2012; Binoy, 2010). The *C. batrachus* available in India have recently been renamed as *Clarias magur*. As per the IUCN Red List of Threatened Species (www.iucnredlist.org; Version 2014.3; downloaded on 20 November, 2014), this species has been listed under endangered (A3cde + 4acde) categories. Further, the lack of genomic resources for this species is one of the major constraints for its conservation, management and genetic improvement programs.

Mitogenome is a circularized DNA having the size ranges from 16–18 kb in fishes. It comprises of 37 genes, viz. 13 protein-coding genes (PCGs), 2 ribosomal RNAs (rRNAs), 22 transfer RNAs (tRNAs) and a variable control or D-loop region. Since the mitochondrial DNA (mtDNA) is maternally inherited and lacks recombination; hence comparatively more conserved during the course of evolution and, therefore, may be used as molecular markers for studying the taxonomy, phylogenetics, phylogeography, population and conservation genetics (Avise, 2000; Hillis et al., 1996; Santini et al., 2013). Till date, ~4636 complete metazoan mitogenomes have been sequenced and deposited in the NCBI organellar genome database (http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?txid=2759&hopt=stat&opt=organelle; November 20, 2014).
Conventional techniques for mitogenome sequencing (long range PCR with subsequent primer walking) are often cumbersome and time consuming (Cameron et al., 2009; Miller et al., 2004; Poulsen et al., 2013; Proscodimi et al., 2012). With the help of next generation sequencing (NGS) technologies, millions of short sequencing reads can be produced in short time with greater economy. The mtDNA, usually present in high copy number in eukaryotic cells, can be assembled independently from the raw reads of whole genome sequencing (WGS) data and the complete mitogenome sequence can be recovered easily with a few giga bases of NGS data (Cui et al., 2009; Gan et al., 2014; Groenenberg et al., 2012; Hahn et al., 2013; Iorizzo et al., 2012; Jex et al., 2013; Prosdocimi et al., 2012). With the help of next generation sequencing, millions of short sequencing reads can be produced in short time with greater economy. The mtDNA, usually present in high copy number in eukaryotic cells, can be assembled independently from the raw reads of whole genome sequencing (WGS) data and the complete mitogenome sequence can be recovered easily with a few giga bases of NGS data (Cui et al., 2009; Gan et al., 2014; Groenenberg et al., 2012; Hahn et al., 2013; Iorizzo et al., 2012; Jex et al., 2013; Prosdocimi et al., 2012).

In this study, the data obtained from two runs, each of Roche 454 and Ion Torrent, were de novo assembled, annotated and analyzed to extract the complete mitogenome of *C. batrachus*. The *C. batrachus* mitogenome from the GenBank sequence database was used for reference guided assembly and its results were compared with that of *de novo* assembled results. We applied different data pooling strategies to obtain the best assembly with the maximum read depth and high coverage.

As the reference mitogenome and the *de novo* assembled mitogenome belong to fish specimens of different geographical locations, the variation analyses were done to study the phylogeographic differentiation between them. Phylogenetic analysis, SSR prediction, analysis of conserved sequences, tRNA and gene clusters among the 24 catfish mitogenomes were carried out to get insight into the intergenic variation and evolutionary relationships among the catfishes.

Materials and methods

DNA isolation, library preparation and sequencing

A healthy farm reared specimen of *C. batrachus* was collected from experimental pond of ICAR-Central Institute of Freshwater Aquaculture, Bhubaneswar, Odisha, India. The specimen was properly anesthetized using conventional protocol as described by Sambrook et al. (1989), with minor modifications. The whole genome sequence data was generated on Roche 454 and Ion Torrent NGS platforms using the sequencing protocols as described by Groenenberg et al. (2012), Miller et al., 2013; Prosdocimi et al., 2012). With the help of next generation sequencing (NGS) technologies, millions of short sequencing reads can be produced in short time with greater economy. The mtDNA, usually present in high copy number in eukaryotic cells, can be assembled independently from the raw reads of whole genome sequencing (WGS) data and the complete mitogenome sequence can be recovered easily with a few giga bases of NGS data (Cui et al., 2009; Gan et al., 2014; Groenenberg et al., 2012; Hahn et al., 2013; Iorizzo et al., 2012; Jex et al., 2013; Prosdocimi et al., 2012).

In this study, the data obtained from two runs, each of Roche 454 and Ion Torrent, were de novo assembled, annotated and analyzed to extract the complete mitogenome of *C. batrachus*. The *C. batrachus* mitogenome from the GenBank sequence database was used for reference guided assembly and its results were compared with that of *de novo* assembled results. We applied different data pooling strategies to obtain the best assembly with the maximum read depth and high coverage.

As the reference mitogenome and the *de novo* assembled mitogenome belong to fish specimens of different geographical locations, the variation analyses were done to study the phylogeographic differentiation between them. Phylogenetic analysis, SSR prediction, analysis of conserved sequences, tRNA and gene clusters among the 24 catfish mitogenomes were carried out to get insight into the intergenic variation and evolutionary relationships among the catfishes.

Materials and methods

DNA isolation, library preparation and sequencing

A healthy farm reared specimen of *C. batrachus* was collected from experimental pond of ICAR-Central Institute of Freshwater Aquaculture, Bhubaneswar, Odisha, India. The specimen was properly anesthetized and tissue samples were taken from different organs, including muscles, and preserved in liquid nitrogen. The genomic DNA was isolated from liver, muscle and testis using conventional protocol as described by Sambrook et al. (1989), with minor modifications. The whole genome sequence data was generated on Roche 454 and Ion Torrent NGS platforms using the sequencing protocols as described in the kits provided.

Raw data processing, assembly and annotation

The quality control (QC) of the raw data obtained from Roche 454 and Ion Torrent was done using CLC Genomics Workbench (version 7.0.4) (Knudsen and Knudsen, 2013). The low quality reads were trimmed/filtered out and the reads with high quality (Phred 20–40) only were retained for the downstream analyses. The *de novo* assembly of Roche 454 data, Ion Torrent data and pooled data (Roche 454 + Ion Torrent) were carried out using CLC Genomics Workbench and Newbler (version 2.9) (Margulies et al., 2005). The *C. batrachus* mitogenome reference was downloaded from GenBank (NC_023923.1) and was used for the reference guided assembly using CLC Genomics Workbench and Newbler. The annotation of the best assembled mitogenome was carried out using Mitoannotator (Iwaski et al., 2013) and further verified by NCBI ORF finder (http://www.ncbi.nlm.nih.gov/projects/gorf/). Mitoannotator predicted the tRNA and rRNA genes of the *C. batrachus* mitogenome. The tRNA and rRNA structures of these predicted genes were further retrieved using Mitos Server (Bernt et al., 2013).

Variation analysis

Furthermore, the trimmed and processed reads from all the 4 runs were aligned with the reference mitogenome of *C. batrachus* using CLC Workbench and Newbler to elucidate intra-specific variations. The variations common to all the 4 runs and to both the assemblers were retrieved and mapped onto the assembled mitogenome of *C. batrachus* using an in-house Perl script. Venn diagram for the depiction of variations in each run was computed using VENNY tool (Oliveros, 2007). The nucleotide composition, AT and GC skews of all the 37 mitogenes was calculated using an in-house Perl script.

Phylogenetic analysis

Furthermore, the mitogenomes of all catfishes (24 species) available in public databases along with six other fishes as outgroups were downloaded from the NCBI. Two different data sets, viz. only CDS and CDS + tRNA, were prepared using a in-house Perl script to carry out the phylogenetic analysis. All the sequences were aligned using MUSCLE (Edgar, 2004) and the phylogenetic tree was constructed using the PhyML (Anisimova and Gascuel, 2006) option in Sea-View software (Gouy et al., 2010), with the following parameters: Model = GTR, Branch support = aLRT (SH like) and the tree searching operation = best of NNI and SPR. Further, refinement of the phylogenetic tree was done using FigTree (Morariu et al., 2009).

The calculation of codon usage, transition/transversion ratio and variation analyses of coding genes of all the catfishes was done using MEGA (Tamura et al., 2013). Heatmap of the codon usage was generated using R Script.

Relative rate of gene evolution

The pairwise distance and regression-based approach was applied to deduce the relative rate of gene evolution for all mitochondrial genes, among the catfishes. The p-distances for all genes were calculated against the respective consensus sequences using cons and distmat from the EMBOSS package. In similar fashion, the p-distances were calculated for all the protein coding genes considering 1st, 2nd, 3rd and 4-fold codon positions. The regression analysis was applied on the

---

**Table 1**

| Platform/Run       | Number of reads | Avg. length | Number of reads after trimming | Percentage of high quality reads |
|--------------------|-----------------|-------------|--------------------------------|---------------------------------|
| Run1, 454         | 14,88,526       | 350.2       | 14,45,418                      | 97.1                            |
| Run2, 454         | 15,39,353       | 372.7       | 15,03,731                      | 97.69                           |
| Run1_Ion_Torrent  | 27,56,593       | 333.2       | 27,26,188                      | 98.9                            |
| Run2_Ion_Torrent  | 33,94,980       | 300.6       | 32,96,863                      | 97.11                           |

**Table 2**

| Assembler   | Data set | De novo assembled mitogenome | Reference guided assembled mitogenome |
|-------------|----------|------------------------------|---------------------------------------|
|             |          | Mitogenome length | No. of reads | Query coverage (%) | Percent similarity (%) | Average coverage | Mitogenome length | No. of reads | Query coverage (%) | Percent similarity (%) |
| CLC         | Pooled 454 | 16,509 | 7655 | 100 | 99.4 | 198.77 | 16,474 | 7712 | 100 | 99.4 |
| CLC         | Pooled Ion Torrent | 16,512 | 4492 | 100 | 99.4 | 87.93 | 16,493 | 4506 | 100 | 99.4 |
| CLC         | Pooled all | 16,509 | 12,160 | 100 | 99.4 | 284.67 | 16,476 | 12,218 | 99 | 99.4 |
| Newbler     | Pooled 454 | 16,509 | 7636 | 100 | 99.4 | 198.3 | 16,509 | 7511 | 100 | 99.4 |
| Newbler     | Pooled Ion Torrent | 16,512 | 4469 | 100 | 99.4 | 87.2 | 16,331 | 4464 | 98 | 99.4 |
| Newbler     | Pooled all | 16,510 | 12,094 | 100 | 99.5 | 284.5 | 16,509 | 11,975 | 98 | 99.4 |
pairwise distances of all genes against the respective distances of the 12S rRNAs as reference. The regression coefficients of correlated distance values represented the relative rates of gene evolution.

Results and discussion

Raw data processing

Raw reads of good quality were obtained from the two runs of each Roche 454 and Ion Torrent with the insert sizes ranging from 500 to 1000 bp and 670 to 1070 bp, respectively. In case of Roche 454, the total reads obtained were 3.01 million (~2.7 GB) with a mean length of 361 bp. Similarly, 6.57 million (~4 GB) reads with a mean length of 316 bp were obtained from the Ion Torrent. After the QC, 97% raw
reads were found to be of high quality, which were retained for the downstream assembly process (Table 1).

De novo assembly

Three different data sets, viz. pooled Roche 454, pooled Ion Torrent and pooled all, were de novo assembled using CLC Genomics Workbench and Newbler. The longest contig obtained from the three data sets were almost identical and ~16 kb in length. The blast result of the longest contig showed the best hit with the *C. batrachus* mitogenome (NC_023923.1) (Mohindra et al., 2013) with 100% query coverage and ~99% identity. The longest mitogenome contig obtained from CLC Genomics Workbench and Newbler were highly similar in terms of percentage identity and sequence length (Table 2). The average coverage of the mitogenome was quite high (~284×) as compared to other contigs due to the presence of high copy number of mitochondria in eukaryotic cells and, thus, indirectly confirms it to be of the mitochondrial origin (Supplementary Fig. 1).

Reference guided assembly

The reference guided assembly of the three datasets using CLC genomic workbench and Newbler resulted in consensus sequence with almost identical assembly to the mitogenome as obtained from the de novo assembly. Based on the results obtained, it can be concluded that CLC Genomics Workbench yielded better results for Ion Torrent data and Newbler for Roche 454 data (Table 2).

Although the mitogenomes obtained from de novo and reference based assemblies showed good query coverage with the reference mitogenome (NC_023923.1), but they differed slightly in length. Out of the three mitogenome assemblies, the mitogenome assembly comprising 16,510 bp long contig was considered to be best one on the basis of the maximum percentage identity and minimum number of gaps and mismatches.

Gene contents and organization

The assembled mitogenome differed by just one base pair from the *C. batrachus*, sampled from India, mitogenome already submitted in GenBank. Gene prediction of the assembled mitogenome revealed 37 mitochondrial genes (13 PCGs, 22 tRNAs, and 2 rRNAs) and a control/D-loop region (Fig. 1). The gene order and orientation were found to be similar with the gene order of vertebrates. Coding genes spanned the maximum mitogenome of over 69%, followed by tRNA of ~9.4%, rRNA ~15.9%, D-loop ~5.2% and intergenic region ~0.3% (Fig. 2).

The nucleotide composition of the assembled mitogenome has a strong bias towards A + T richness (~57%) with overall content of ~32%A, ~25%T, ~15%G and ~28%C (Supplementary Table 1). The maximum AT content (62.76%) was observed in the D-loop region, followed by the PCGs. The unbalanced nucleotide frequencies were not only
dependent on AT content variation, but also on strand bias. The overall GC and AT skews of the *C. batrachus* mitogenome is found to be $-0.29$ and $0.13$, respectively. The GC skew of the PCGs ranged from $-0.47$ (ATPase 8) to $0.59$ (ND6), where only ND6 showed a positive value (Supplementary Table 1). This observation mainly confirms the origin of ND6 from the light strand (Fischer et al., 2013; Zhuang et al., 2013). The AT skew of the PCGs was ranged from $-0.52$ (ND6) to $0.24$ (ND2). The genes with a negative GC skew are coded on the heavy strand and vice versa. Thus, GC and AT skews are mainly responsible for characterizing the differences between the two strands of a mitochondrial genome, with one strand favoring CT over CA (Min and Hickey, 2007). This may be due to an asymmetry in replication process of mitochondrial genomes. Due to negative GC skew and positive AT skew, it is expected that the *C. batrachus* mitogenome encodes proteins that are relatively low in the proportion of cysteine, valine, phenylalanine, glycine and tryptophan, all of which are encoded by the GT rich codons.

Most of the PCGs had ATG as their start codon and TAA as the stop codon. A total of 11 noncoding intergenic regions were observed in the mitogenome of *C. batrachus*, the longest being 32 bp in length, which was found between tRNA-Asn and tRNA-Cys (Supplementary Table 2). This was also reported by Broughton et al. (2001) in zebrafish and it was similar to the origin of light strand replication (O$_s$) in other vertebrates (Wong and Clayton, 1985). The O$_s$ region has potential to fold into a stable stem loop secondary structure (Fig. 3), with a stem formed by 8-paired nucleotide and a loop of 13 nucleotides. A total of 13 overlapping regions were also reported with the maximum overlap of 38 bp that was observed between Cyt b and tRNA-Thr. The overlaps observed between ATPase 8 and ATPase 6, ND4L and ND4, ND5 and ND6 correlated with that of Black carp, Grass carp, Nile tilapia and Blue tilapia (He et al., 2011; Wang et al., 2008). Four consecutive genes, viz. ATPase 8, ATPase 6, COX3 and tRNA-Gly, were found to be completely overlapped (Supplementary Table 2). The protein coding genes, which were immediately followed by the tRNA genes on the same strand, do not overlap. On the other hand, adjacent protein coding genes always overlap if no tRNA is present between them. This observation strongly supported the idea that the tRNA genes, located between peptide genes, function as signals for the processing of transcripts (Ojala et al., 1981). The two bases of the tRNA-Trp were used as stop codons for ND2 gene.

A total of 22 tRNA genes, identified in the *C. batrachus* mitogenome, had lengths that ranged from 67 nt to 75 nt with $-57\%$ AT content and $-43\%$ GC content. Fourteen tRNAs were coded by the H-strand and 8 by the L-strand. All the tRNA genes were found to have a typical clover leaf structure, with the exception of tRNA-Ser that lacks the paired DHU arm (Supplementary Fig. 2).

The $\delta$-loop region of the mtDNA was found to be highly variable across the taxonomic groups and even in the closely related species. It regulates replication and transcription in mtDNA (Wong and Clayton, 1985). The identified $\delta$-loop region was 871 nt long, which spanned $-5\%$ of the mitogenome and was less similar to other fishes than the PCGs. A conserved element TAS (termination-associated sequence) was reported that was located at the 5‘ end of the control region. This TAS motif (TACATATTGTA) act as a signal for the termination of D-loop strand (7S DNA) synthesis (Faber and Stepien, 1998; Liu et al., 1998; Liu et al., 2000; Shi et al., 2013). Three conserved sequence blocks (CSBs) were identified (Fig. 4) using MEME tool, which were found to be highly conserved among the catfishes and also reported in zebrafish (Broughton et al., 2001) and *Odontamblyopus rubicundus* (Liu et al., 2013). These CSBs are involved in the positioning of RNA polymerase for transcription as well as for priming replication (Clayton, 1991; Shadel and Clayton, 1997)

**Average coverage analysis**

The average coverage of the mitogenome obtained from pooled 454 data, pooled Ion Torrent data and pooled all data were $-198 \times$, $-87 \times$ and $-284 \times$, respectively. Gene-wise coverage was computed for all the 3 data sets and the coding as well as the rRNA genes was found to assemble at higher coverage than the tRNA genes because of the large length of PCGs and rRNA (Fig. 5).
Fig. 8. A — CDS regions based phylogenetic relationships among the catfishes. B — CDS + tRNA regions based phylogenetic relationships among the catfishes.
Fig. 9. Heatmap of codon usage.
Variation analysis

Individual variations were observed between the *de novo* assembled and the reference mitogenomes. These variations may be a part of the sequencing error, so in order to remove the false positives from the list; we individually aligned the reads from 4 different data sets (two 454 and two Ion Torrent) onto the mitogenome reference using CLC Genomics Workbench and Newbler. A total of 51 variations (Supplementary Table 3) were found to be common among the 4 datasets (Fig. 6), of which 47 were single nucleotide variations (SNVs) and 4 multiple nucleotide variations (MNVs). All the reported variations were homozygous in nature, and of which 50 variations were seen in the coding region and 1 SNV in the longest intergenic spacer (32 bp in length) between tRNA-Asn and tRNA-Cys. Out of the 47 SNVs, 38 showed transition and 9 transversion. Maximum number of variations were observed in ND5 region (8), followed by tRNA-Ser (7) (Fig. 7). A total of 9 non-synonymous mutations were reported in PCGs, out of which 3 were in ND4 gene. The functional implication of the observed non-synonymous mutations remains to be elucidated experimentally.

Phylogenetic analysis

Phylogenetic analysis of *C. batrachus* was carried out with 24 other catfishes and 6 outgroup species to deduce the evolutionary relationship between the catfishes and to infer the genetic distance from other outgroups. The phylogenetic analysis was performed from two data sets, one pertaining to the concatenated PCGs and other to the concatenated PCGs + tRNAs. The phylogenetic relationship obtained from both the data sets yielded almost similar results and were mainly consistent to the previously reported phylogeny (Fig. 8). The topology of the tree showed higher aLRT-SH score for most of the branches; thus, supporting the authenticity of the phylogeny. The outgroups used in the phylogeny formed a separate clade and were consistent with the study reported by Betancur-R. et al. (2013). Among the catfishes, *C. batrachus* and *Heteropneustes fossilis* formed the monophyletic clade, as also reported by Mayden et al. (2008). As per the phylogenetic tree, *Palteobagrus*, *Pseudobagrus* and *Leiocassis* bagrids were non-monophyletic and were mainly divided into two lineages (referred to as Lineage I and Lineage II). Lineage I comprised of *Leiocassis longirostris*, *Palteobagrus nitidus*, *Pseudobagrus ussuriensis* and *Palteobagrus vachelli*, while Lineage II comprised of *Palteobagrus fulvidraco*, *Pseudobagrus tokiensis* and *Palteobagrus eurypogon* with high confidence value. Thus, the present finding correlated with that of Mo (1991) and Ku et al. (2007), which stated that genera *Leiocassis* and *Palteobagrus* should be considered as the members of *Pseudobagrus*.

Comparative analyses among catfishes

The mitogenomes of 24 catfishes were compared to study the variable sites between them. At the nucleotide level, the maximum variable sites were observed in δ-loop region, followed by ND2 gene and the minimum were seen in COX2 (Supplementary Fig. 3a). The size of δ-loop region also varied among the catfishes, ranging from 70 nt to 910 nt (Supplementary Table 4). At the protein level, the maximum variable sites were observed in ND2 and the minimum in COX1 genes (Supplementary Fig. 3b). The variations at protein level were much less than that at nucleotide level because of the degeneracy of the genetic code. This variation is extensive, and for both protein and tRNA genes it reflects considerable gene and lineage-specific variation in rates of gene loss. The transition-transversion ratio (Ts/Tv) was maximum for ND6 gene and minimum for δ-loop region (Supplementary Fig. 3c). The heatmap analysis of the codon usage co-related with the phylogenetic analyses (Fig. 9). Due to negative GC skew and positive AT skew, it was expected that the *C. batrachus* mitogenome encodes proteins that are relatively low in the proportion of cystein, valine, phenylalanine, glycine and tryptophan. The same inference can be drawn by looking at the heatmap of the codon usage.

Simple sequence repeat (SSR) analysis

The SSR prediction was carried out using MISA tool for all the 28 catfish mitogenomes available in NCBI database and our assembled mitogenome of *C. batrachus* (Supplementary Table 5). Out of 28 catfish species with 4,62,545 bp sequence size examined, we observed SSRs in 11 species. A total of 11 SSRs were present in δ-loop region in these 11 catfish species, whereas 1 SSR was present additionally in ND2 region of *C. batrachus* which was a striking inference drawn and may be used as a potential biomarker for *C. batrachus* identification. The SSR results of our assembled mitogenome were similar to that of the already submitted *C. batrachus* mitogenome.

Relative rate of gene evolution

The results of pairwise distance and regression-based approach (Supplementary Table 6) for relative rate of gene evolution showed that protein coding mitochondrial genes evolve at a much faster rate than the δ-loop, which was followed by the rRNAs; the tRNAs shows wide variability in the rate of sequence evolution, and on average evolve the slowest (Fig. 10). Among the coding genes, ND2 evolves most rapidly. The δ-loop region contains several conserved sequence blocks, whereas the remaining region is highly variable. Many insertions and deletions are seen when aligning the δ-loop regions of different catfishes; these
highly variable regions were removed using MEGA (β-σ2 loop Gb) in order to provide a comparative view on the effect of large scale variation. According to the neutral theory of molecular evolution, synonymous sites in protein-coding genes evolve faster than non-synonymous sites due to the difference in selection pressure. Our results correlate with this theory, since any mutation at the 2nd codon position is non-synonymous, whereas most mutations at 3rd position and some mutations at 1st position are synonymous. The rate of false positives in the detection of positively selected genes and sites increases at the levels of constraint at 4-fold degenerate sites.

Conclusion

The retrieval of complete mitogenome using WGS data is much appropriate and less time consuming than the Sanger sequencing method because of the presence of very high number of mitochondria in cells. The outcome of the present study was a well assembled mitogenome with very high coverage (~284×). The gene order of C. batrachus as well as the conservation of some intergenic sequences and genic contents suggested that evolution of mitogenome is consistent with fish taxonomy. Most tRNA can be folded as classical clover leaf structure, with the exception of tRNA-Ser that lacks the paired DHU arm. The functional implication of the observed non-synonymous mutation can further be elucidated experimentally. The TAS, CBS and O2 regions were found exclusively for the first time in C. batrachus mitogenome. The unique SSR observed in ND2 region can be a potential molecular marker for the characterization of Clarias spp. and can further be validated experimentally.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.mgene.2015.06.004.

Acknowledgments

The authors are thankful to the Director, NBFCGR, Lucknow; the Vice-Chancellor, AAU, Anand and the Director, CIFA, Bhubaneswar, India for providing necessary laboratory facilities. The financial support provided by Department of Biotechnology, Ministry of Science and Technology, Gov. of India, New Delhi, India vide Sanction Grant No. BT/PR3688/AEE/3/571/2011 dated 10.09.2013 for the present research works are also duly acknowledged.

References

Ahmad, R., Pandey, R.B., Arif, S.H., Nabi, N., Jabeen, N., Hasnain, A., 2012. Polymorphic β and γ lens crystallins demonstrate latitudinal distribution of threatened walking catfish Clarias batrachus (Linn.) populations in North-western India. J. Biol. Sci. 12, 98–104.

Anisimova, M., Gascuel, O., 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. 55, 539–552.

Avise, J.C., 2000. Phylogeography: the history and formation of species. Harvard University Press.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., Stadler, P.F., 2013. MITOS: improved de novo metazoon mitochondrial genome annotation. Mol. Phylogenet. Evol. 69, 313–319.

Betancur-R., R., Broughton, R.E., Wiley, E.O., Carpenter, K., López, J.A., Li, C., Holcroft, N.J., Arcila, D., Sanciangco, M., Cureton II, J.C., 2013. The tree of life and a new classification of the ray-finned fishes (Pisces, Osteichthyes) based on 1740 nuclear loci. PLoS ONE 8, e57048.

Binoy, V., 2010. Catfish Clarias is vanishing from the waters of Kerala.Curr. Sci. 99, 714–719.

Biollaz, J., Brändlein, F., Loidl, A., Sammarco, S., Squivido, G., Tschudi, U., 2013. O-Fold: a robust non-homology modeling method. Bioinformatics 29, 1373–1380.

Bouvier, D., 2007. New insights into the Placodermi (pre-chordate vertebrates) from the late Silurian of southern France. Palaeontology 50, 1013–1039.

Braun, C., Rietsch, L., Wissinger, B., Ganten, D., 2006. Molecular characterization of the Danio rerio mitochondrial genome. Mol. Biol. Evol. 23, 2436–2442.

Brosius, J., 1981. The mitochondrial DNA of Neurospora crassa: Molecular biology of a eukaryotic genome. Annu. Rev. Genet. 15, 113–139.

Cairns, N.J., 1989. The evolutionary significance of mitochondrial DNA at the species level. Annu. Rev. Genet. 23, 211–236.

Cai, Y., Wang, L., Zhou, S., Shi, J., 2012. Characterization of the mitochondrial genome of the large yellow croaker, Clarias batrachus, a potential molecular marker for identification of hatchery fish. Aquaculture 369–370, 74–81.

Cai, Y., Zou, J., 2013. Analysis of clade specific mitochondrial DNA control regions among different species of greenland halibut, Hippoglossus hippoglossus (Pisces, Gadidae). J. Mol. Evol. 77, 219–233.

Canas, S., 2001. Mitochondrial DNA diversity in Mediterranean fish species as a tool to assess genetic diversity. J. Appl. Ichthyol. 17, 279–287.

Carr, A.R., 2005. The RNA polymerase II transcriptome of the Drosophila melanogaster mitochon-
Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. Molecular cloning. Cold Spring Harbor Laboratory Press, New York.
Santini, F., Sorenson, L., Alfaro, M.E., 2013. A new multi-locus timescale reveals the evolutionary basis of diversity patterns in trigger fishes and filefishes (Balistidae, Monacanthidae; Tetraodontiformes). Mol. Phylogenet. Evol. 69, 165–176.
Shadel, G.S., Clayton, D.A., 1997. Mitochondrial DNA maintenance in vertebrates. Annu. Rev. Biochem. 66, 409–435.
Shi, W., Dong, X.-L, Wang, Z.-M., Miao, X.-G., Wang, S.-Y., Kong, X.-Y., 2013. Complete mitogenome sequences of four flatfishes (Pleuronectiformes) reveal a novel gene arrangement of L-strand coding genes. BMC Evol. Biol. 13, 1–9.
Singh, B., Hughes, G., 1971. Respiration of an air-breathing catfish Clarias batrachus (Linn.). J. Exp. Biol. 55, 421–434.
Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729.
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 phylogenetic position within cyprinidae. Gene 424, 96–101.
Wong, T.W., Clayton, D.A., 1985. In vitro replication of human mitochondrial DNA: accurate initiation at the origin of light-strand synthesis. Cell 42, 951–958.
Zhuang, X., Qu, M., Zhang, X., Ding, S., 2013. A comprehensive description and evolutionary analysis of 22 grouper (Perciformes, Epinephelidae) mitochondrial genomes with emphasis on two novel genome organizations. PLoS ONE 8 (8), e73561.