Norén & Kullander, *Cogent Biology* (2018), 4: 1525857
https://doi.org/10.1080/23312025.2018.1525857

cogent biology

CELL, MOLECULAR & DEVELOPMENTAL BIOLOGY | RESEARCH ARTICLE

The enigmatic *Betadevario ramachandrani* (Teleostei: Cyprinidae: Danioninae): phylogenetic position resolved by mitogenome analysis, with remarks on the prevalence of chimeric mitogenomes in GenBank

Michael Norén[1]* and Sven Kullander[1]

Received: 11 June 2018
Accepted: 14 September 2018
First Published: 20 September 2018

*Corresponding author: Michael Norén, Swedish Museum of Natural History, P.O. Box 50007, SE-10405, Stockholm, Sweden
E-mail: michael.noren@nrm.se

Reviewing editor:
Jason Abernathy, USDA-ARS Southeast Area, USA

Additional information is available at the end of the article

Abstract: We present the complete mitochondrial genome and a phylogenetic analysis of the danionine cyprinid *Betadevario ramachandrani*, endemic to the Western Ghats in India. Bayesian phylogenetic analysis of all available mitochondrial genomes of Danionina show that *B. ramachandrani* is the most basal member of a clade also containing *Devario, Microdevario* and *Microrasbora*, and this clade is the sister group of *Danio*. Seven of 20 mitochondrial genomes downloaded from GenBank for phylogenetic analysis were found to be chimeric, including five curated reference genomes, and this did affect our phylogenetic analysis. At least three of these erroneous sequences have been used in other studies. There is reason to suspect that there are numerous chimeric mitogenomes in GenBank.

Subjects: Ichthyology; Phylogenetic analysis; Conservation Biology

Keywords: DNA barcoding; next generation sequencing; phylogeny

1. Introduction

Betadevario ramachandrani Pramod, Fang, Rema Devi, Liao, Indra, Jameela Beevi & Kullander, 2010 is a small (up to 61 mm SL) danionine cyprinid fish which combines morphological characters

ABOUT THE AUTHOR

Dr. Michael Norén and Professor Sven Kullander are members of the ichthyology team at the Swedish Museum of Natural History (NRM) in Stockholm, Sweden. Professor Kullander (ORCID: 0000-0001-6075-0266) is the Scientific Curator for the ichthyological and herpetological collections of the NRM, team leader for FishBase Sweden, and an expert on fish taxonomy, specializing in Cichlidae and Cyprinidae. Dr. Norén (ORCID: 0000-0003-2561-6760) is Curator of FishBase Sweden, and specializes in molecular systematics.

PUBLIC INTEREST STATEMENT

This study reports on the first sequencing of the whole mitochondrial genome of the rare carpfish *Betadevario ramachandrani*, and performs an analysis which confirms that it is a primitive relative of the genus *Devario*. During analysis, it was found that 7 out of 20 mitochondrial genomes downloaded from GenBank for inclusion in this study were amalgams ("chimeras") from several different species; in one case, the genome is comprised of sequences from three different species from three different subfamilies of Cyprinidae. The erroneous genomes affected the outcome of the analysis, at least three of the erroneous genomes have been used in other studies, and further analysis suggests that the problem of chimeric genomes in GenBank may be widespread.

typical of the genus *Danio*, such as two pairs of long barbels, with characters typical of the genus *Devario*, such as a prominent dark spot immediately posterior to the gill opening, and is the only species in the genus *Betadevario*. It is restricted to a single high-altitude stream, 2.27 m wide and at most 0.3 m deep, in the upper Sita River drainage in India (Pramod et al., 2010). Its limited known distribution suggests that it is vulnerable to human activities such as fishing, pollution, logging, or damming, and there are no records of *B. ramachandrani* after its description. Pramod et al. (2010) analysed the morphology of *B. ramachandrani* and mitochondrial *CYTB* and nuclear *RHO* data from DNA samples stored at the Swedish Museum of Natural History and concluded that *B. ramachandrani* "is a basal species in the clade containing *Devario* in both morphological and molecular analyses". In 2017, it was found that despite being stored in alcohol at −80°C, the DNA of the samples was degraded, with no remaining fragments longer than 450 base pairs (bp), and a decision was made to use high-throughput sequencing technology to sequence the whole mitochondrial genome.

2. Materials and methods

2.1. DNA extraction and sequencing

DNA was extracted from material stored at the Swedish Museum of Natural History (voucher ID: NRM 57,780, capture locality: 13°29′22.8″N 75°03′53.5″E, close to Barkana falls, Karnataka, India) using the GeneMole automated DNA extraction system (Mole Genetics, Lysaker, Norway) with recommended protocol.

Twenty microlitres of DNA extract (sample concentration 14 ng DNA per µL) was sent to Macrogen Inc. (Seoul, Republic of Korea) for shotgun sequencing (HiSeq X sequencer with TruSeq DNA nano kit (Illumina, San Diego, USA)), producing 523 million paired reads, 145 bp, with 281 bp insert. The reads have been deposited at the NCBI Sequence Read Archive (SRA), accession number SRP157898. Assembly of the mitochondrial genome was performed using the computer software Geneious v.10 (Biomatters, Auckland, New Zealand) (Kearse et al., 2012). The paired reads were merged, and unmerged reads and merged reads shorter than 50 bp were deleted, leaving a total of 25 million merged reads. The reads were quality-trimmed to remove positions with >2% probability of error, and mapped to a published reference mitochondrial genome (*Danio dangila*, Genbank accession number NC_015525). A total of 35,430 reads mapped to the reference sequence, producing a minimum coverage of 131. The 85% majority rule consensus was extracted, and annotated by transferring annotations from published reference sequences (*Danio dangila* NC_015525, *Rasbora daniconius* NC_015527, *Microdevario nanus* NC_015546), with manual adjustment.

2.2. Phylogenetic analysis

There is at present no consensus on the taxonomy of Cypriniformes, with several novel and partly conflicting classifications having been proposed in the last few years (Nelson, Grande, & Wilson, 2016; Stout, Tan, Lemmon, Lemmon, & Armbruster, 2016; Van Der Laan, Eschmeyer, & Fricke, 2014). We use Nelson et al. (2016) as main taxonomic reference, but our concept of subtribe Danionina is from Liao, Kullander, and Fang (2011). Up to two published complete or nearly complete mitochondrial genomes per species of the subtribe Danionina were downloaded from GenBank. Seven of the 20 downloaded genomes (GenBank accession numbers AB937094, KP407138, NC_015528, NC_026122, NC_027688, NC_028526, NC_029771) were chimeric and were removed from analysis. A total of 13 mitogenomes, representing 12 species and all genera of Danionina except *Chela* and *Laubuka*, were aligned using the MAFFT (Katoh, Misawa, Kuma, & Miyata, 2002) plug-in for Geneious. All included species share the same gene order. The control region could not be confidently aligned and was deleted. One unique insertion in the *Microdevario* and two in the *Danionella* genomes were deleted. A distantly related potential danionine, *Sundadanio rubellus* (in GenBank under the trade name *Sundadanio axelrodi* "RED", with accession number AP011401), served as outgroup. The final alignment was 15,872 bp. For the phylogenetic analysis, the data were partitioned by protein coding or not protein coding, and coding data further

partitioned based on codon position, for a total of four partitions. Phylogenetic analysis was performed using the parallel-computing version of the computer software MrBayes v3.2 (http://mrbayes.sourceforge.net) (Ronquist et al., 2012), with General Time Reversible model, assuming a Γ distribution of rates, and estimating the proportion of invariant sites (GTR + Γ + I) for all partitions, as recommended by the computer software PartitionFinder2 (http://www.robertlanfear.com/partitionfinder) (Lanfear, Frandsen, Wright, Senfeld, & Calcott, 2017) with Akaike information criterion and restricting to models supported by MrBayes. The analysis was run for 5 million generations, sampled every 1,000 generations, the first 25% of samples discarded as burn-in, and convergence was checked with Tracer v1.4 (http://tree.bio.ed.ac.uk/software/tracer) (Rambaut & Drummond, 2007). The alignment, with MrBayes data block with data partition scheme and model, is available in Nexus format from the authors.

Two alternative data partitioning schemes were tried: (1) no partitioning, and (2) partitioning the dataset by gene, with protein coding genes further partitioned by codon position, resulting in a total of 37 initial partitions, then analysing with the data partitioning (22 partitions; PartitionFinder2 merges partitions it determines to have similar characteristics) and model for each partition recommended by PartitionFinder2. All analyses produced trees with differing branch lengths but identical topology and posterior probability (Bayesian posterior probability = 1 for all nodes in all trees).

3. Results

3.1. The mitochondrial genome of Betadevario ramachandrani

The genome sequence of *B. ramachandrani* is 16,932 bp and comprises 13 protein coding genes, 22 tRNA genes, 2 rRNA genes and D-loop region (control region). The control region is 1,322 bp, from position 15,611 to 16,932, and contains 2 repetitious regions. The first, from position 15,633 to 16,034, consists of one 69 bp tandem repeat and one 21 bp tandem repeat; the number of repetitions is uncertain, as it varied between 3 and 4 for the 69 bp repeat, and 4 and 6 for the 21 bp repeat, depending on assembly parameters. The second repetitious region, from position 16,799 to 16,828, is a dinucleotide microsatellite AT repeat. All protein coding genes start with ATG (Met), and end with TAA as stop codon, except *ND2, ATP8, ND3, ND4* and *ND6* which end with TAG. The *12S* (small) ribosomal RNA gene is 956 bp, and the *16S* (large) ribosomal RNA gene is 1668 bp. The mitogenome has a base composition of A (32.5%), C (25.3%), G (15.5%) and T (26.7%).

The complete annotated mitochondrial genome, with GenBank accession number MH817023, is illustrated in Figure 1.

3.2. Phylogenetic analysis

The result of the phylogenetic analysis is summarized in Figure 2. *Esomus metallicus* is the most basal species of Danionina. *Danionella* is monophyletic, and the sister group of all remaining Danionina. *Danio* is monophyletic, and the sister group of a clade comprising *Microdevario nanus, Betadevario ramachandrani, Microrasbora rubescens* and *Devario devario. Betadevario ramachandrani* is the sister taxon of *M. rubescens* + *D. devario*.

3.3. Chimeric mitogenomes

To test for chimeric sequences, individual genes from the mitochondrial genomes downloaded from GenBank were BLAST-searched against the *nr* database of GenBank (17 May 2018). To minimize the risk of false positives, only genes longer than 600 bp (*12S, 16S, ND1, ND2, COX1, COX2, ATP6, COX3, ND4, ND5, CYTB* and D-loop region) were BLAST-searched. Seven genomes were found to be chimeric: NC_026122, NC_028526, NC_029771, KP407138, AB937094, NC_027688 and NC_015528. Below is a list of the sequences and genes which appear to be at least partly chimeric, with GenBank accession number of the top BLAST hit. Genes which appear to be from the target organism are not listed. BLAST search hits from other suspected chimeric sequences are not reported.

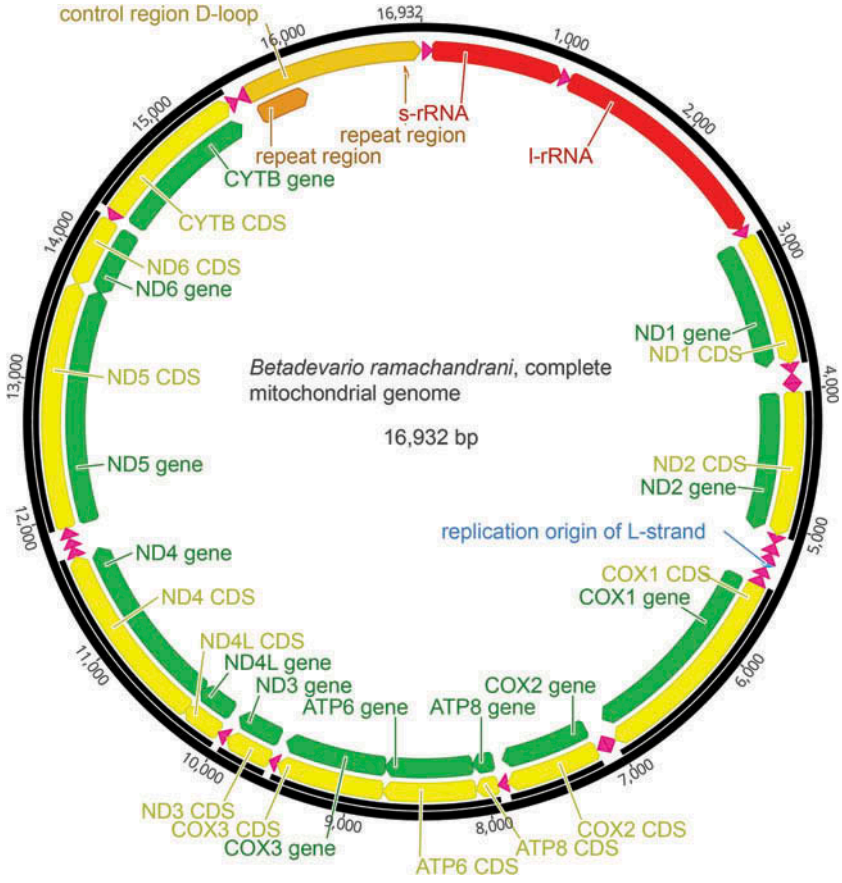

Figure 1. Annotated schematic representation of the mitochondrial genome of *Betadevario ramachandrani*. GenBank accession number MH817023.

NC_026122 "*Devario laoensis*": *COX1, COX2, ND2, ND4* and *ATP6* genes are from a barbine cyprinid of the genus *Schizothorax*. *COX1* = 100% similar to HQ235962 *S. malacanthus; COX2* = 98% similar to KT833107 *S. oconnori; ND2* = 91% similar to KC51374 *S. waltoni; ND4* = 89% similar to HQ235822 *S. malacanthus* (note: none of the top 100 search hits were danionines); *ATP6* = 95% similar to KT833107 *S. oconnori*. The *CYTB* and *16S* rRNA genes are from two other species of *Devario* (*16S* = 99% similar to GQ406286 *D. apogon; CYTB* = 99% similar to EU241433 *D. chrysotaeniatus*. Note, however, that *D. laoensis* does not occur in China (Fang, 2000), and the source specimen for NC_026122 is likely a misidentified *D. chrysotaeniatus*.

NC_028526 "*Danio margaritatus*": *ND1, ND2, COX1, COX2, ATP6, ND4* and *ND6* are from the labeonine cyprinid *Discogobio tetrabarbatus* (BLAST similarity to KJ669372 91%, 95%, 90%, 95%, 94%, 90% and 99%, respectively), whereas *ND5* and probably also the control region is from the squaliobarbine cyprinid *Ctenopharyngodon idella* (*ND5* = 95% similar to KM401549 *C. idella* x *Elopichthys bambusa*; control region = 82% similar to KT894100 *C. idella*; no danionines among the 100 most similar sequences).

NC_029771 "*Danio albolineatus*": *COX1, COX2, ND1, ND2, ND4, ND5* and control region are from a squaliobarbine cyprinid, probably *C. idella* (respectively, 91%, 91%, 84%, 88%, 88%, 91% and 82% similar to KM401549 *C. idella* x *E. bambusa* or KT894100 *C. idella*).

KP407138 "*Devario chrysotaeniatus*": *ND1, ND2, COX1, ATP6, COX3, ND4, ND5* and *CYTB* are 99–100% similar to MG570437 *C. idella*. *COX2* and *ATP6* are 95% similar to MG570437 *C. idella*. Aligning MG570437 to KP407138 reveals that the two sequences are nearly identical, differing mainly in one 1998 bp region from the 3′ half of *12S* to near the 3′ end of *16S*, and one 2061 bp

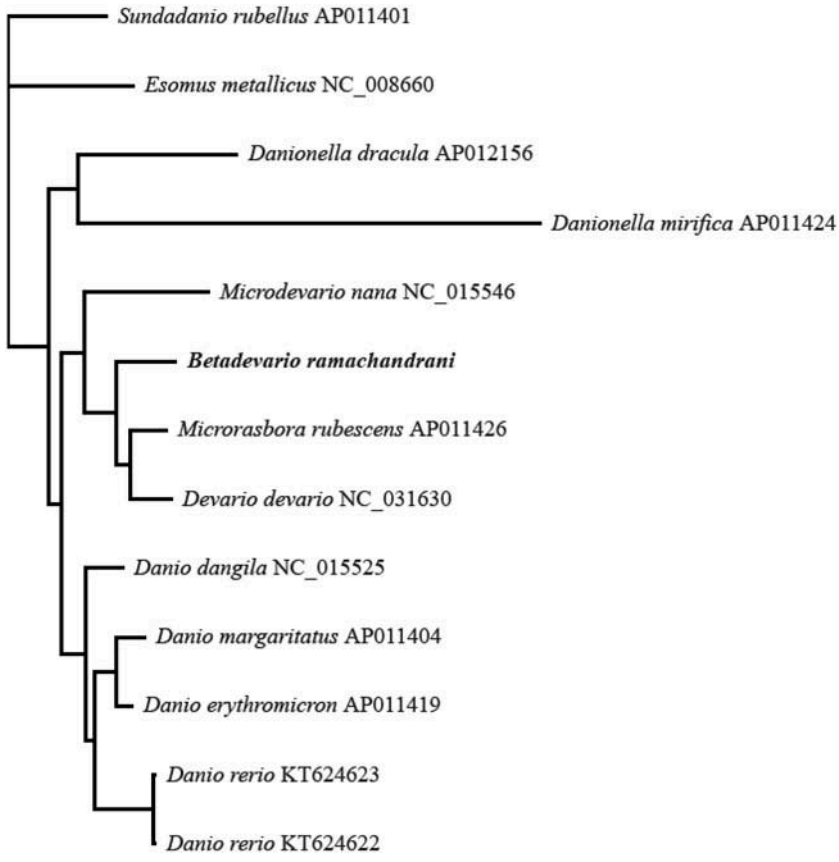

Figure 2. Phylogram of all species of subtribe Danionina represented on GenBank on the 17 May 2018, based on Bayesian analysis of nearly complete mitochondial genome sequences, with *Sundadanio rubellus* as outgroup. All nodes have Bayesian posterior probability = 1. Terminal labels end with GenBank accession number. Branch lengths are proportional to number of expected substitutions per site.

region from the 3′ end of *COXI* to the 3′ end of *ATP6*. Submitting these two regions to a BLAST search reveals that first is 97% similar to *Devario laoensis* (KP115291), whereas the second is 99% similar to the cyprinine cyprinid *Procypris mera* (KM461699).

AB937094 "*Microdevario kubotai*": *CYTB*, control region, and *12S* are 94%, 99% and 94% similar to AB924546 from the rasborine cyprinid *Rasbora borapetensis*, respectively.

NC_027688 "*Danio nigrofasciatus*": *ND4, ND5, ND6* and *CYTB* are 100% similar to KT624623 *Danio rerio*.

NC_015528 "*Leptobarbus hoevenii*": leptobarbine cyprinid downloaded for use as outgroup. *12S* = 94% similar to KJ679504 from the xenocypridine cyprinid *Hypophthalmichthys nobilis; 16S* = 94% similar to KT894100 from squaliobarbine cyprinid *C. idella*.

4. Discussion

4.1. Phylogenetic position of B. ramachandrani

The result of the phylogenetic analysis is summarized in Figure 2. Our result is, despite different taxon sampling and an order of magnitude more molecular data, compatible with the conclusions of Pramod et al. (2010), which were based on mitochrondrial cytochrome *b* and nuclear rhodopsin data, and morphology. Both studies support the view that *B. ramachandrani* is the most basal member of a clade containing *Devario* + *Microrasbora*, and that (*Microdevario* +*Betadevario* +*Devario* +*Microrasbora*) are the sister group of *Danio*.

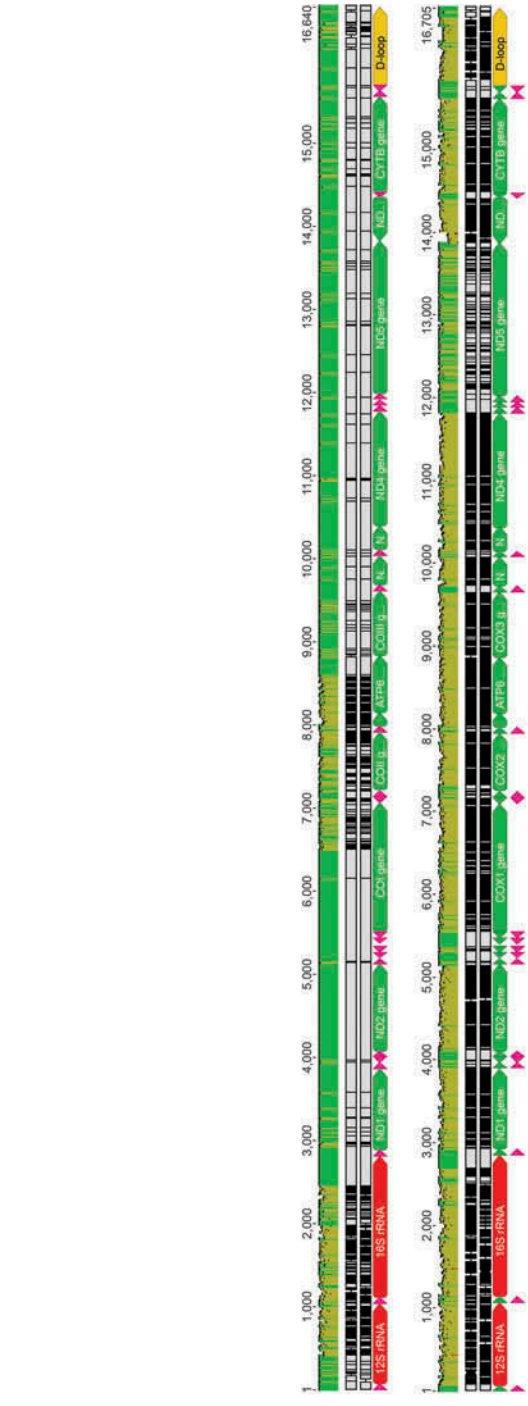

Figure 3. Two example chimeric genomes aligned to a genome of the main contaminant (*Ctenopharyngodon idella*). The numbers at top indicate position in the aligned genomes. The top bar indicates identity: if both sequences are identical, the bar is tall and bright green. As the aligned sequences come from distantly related taxa, they should not have significant identical regions, so bright green effectively indicates contamination. The grey bars are the aligned sequences, with dissimilar bases highlighted in black. The bottom bar indicates gene location, with arrows pointing in read direction. (A): KP407138 ("*Devario chrysotaeniatus*") is nearly identical to *C. idella*, with the exception of one region spanning parts of the *COI*, *COII* and *ATP6* genes, and one region spanning parts of the *12S* and *16S* rRNA genes; the first region is 99% similar to *Procypris mera*, a contamination, and the second is 97% similar to *Devario laoensis*, and possibly correct. (B): Contamination from *C. idella* in NC_028526 ("*Danio margaritatus*") is evident in the ND5 gene, and in short regions flanking protein coding genes.

4.2. Chimeric mitochondrial genomes

Our initial phylogenetic analysis produced surprising results, with three danionine sequences grouping distantly from all other Danioninae: NC_029771 "*Danio albolineatus*" and KP407138 "*Devario chrysotaeniatus*" were recovered as members of Xenocypridinae, while NC_028526 "*Danio margaritatus*" was recovered as a member of Labeoninae. These species are morphologically highly distinct, and misidentification seemed unlikely. A BLAST search of the individual genes of the 20 mitogenomes downloaded from GenBank revealed that 7 mitogenomes are partly chimeric. Six of the chimeric genomes were produced in China, by four different institutions, and one in Japan, so it is not an issue restricted to one institution. At least three of the chimeric mitogenomes have been used by other studies, and five are curated GenBank reference genomes (RefSeq sequences). We have notified GenBank of our findings. There is reason to think that the problem is not restricted to the subfamily Danioninae, as the contaminant sequences frequently are from species of different subfamilies of the Cyprinidae.

Analysis suggests the contamination may be of two different types (Figure 3). Some, often spanning several 1,000 bp, appear to be a result of the person assembling the genome failing to notice that some reads are from a non-target organism. Other consist of short regions flanking protein coding genes; we speculate that these regions correspond to primer bind positions with low or zero read coverage, and that the reference genome used to assemble the reads was not removed before calculating the consensus sequence.

Performing a BLAST search of GenBank's *nr* database (11 September 2018) with one of these short flanking regions (a 211 bp fragment corresponding to the *tRNA-Ile, tRNA-Gln* and *tRNA-Met* genes, from position 3,843 to 4,053 in the NC_029771 genome) found genomes from seven different subfamilies of Cyprinidae with corresponding regions which were 98–100 percent similar. Randomly selecting and investigating one of the suspected chimeric genomes, KJ801524, ostensibly from *Chanodichthys erythropterus* (in GenBank as *Culter erythropterus*), revealed that its *12S* and *16S* rRNA genes are 98 % similar to the corresponding genes of KJ756343 *Hypophthalmichthys nobilis*, and almost certainly chimeric.

That a third of the genomes downloaded for this study were chimeric suggests that the problem may be widespread, and sequence identity an issue for any study which uses data from downloaded whole mitochondrial genomes. We urge researchers to test downloaded mitogenomes by doing a GenBank BLAST search for each gene prior to use, to report erroneous sequences they detect to GenBank, and to check their own sequences before submitting new genomes to GenBank.

Acknowledgements
Te Yu Liao (National Sun Yat-sen University, Taiwan) made the extraction used in this study, and Mazen Sarhan (Macrogen Europe, Netherlands) provided help with high-throughput sequencing.

Funding
The authors received no direct funding for this research.

Competing Interest
The authors report no conflicts of interest.

Author details
Michael Norén[1]
E-mail: michael.noren@nrm.se
ORCID ID: http://orcid.org/0000-0003-2561-6760
Sven Kullander[1]
E-mail: sven.kullander@nrm.se
ORCID ID: http://orcid.org/0000-0001-6075-0266
[1] Department of Zoology, Swedish Museum of Natural History, Stockholm, Sweden.

Citation information
Cite this article as: The enigmatic *Betadevario ramachandrani* (Teleostei: Cyprinidae: Danioninae): phylogenetic position resolved by mitogenome analysis, with remarks on the prevalence of chimeric mitogenomes in GenBank, Michael Norén & Sven Kullander, *Cogent Biology* (2018), 4: 1525857.

References

Fang, F. (2000). A review of Chinese *Danio* species. *Acta Zootaxonomica Sinica*, *25*, 214–227.
Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.
Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … Drummond, A. (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649.
Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., & Calcott, B. (2017). PartitionFinder 2: New methods

for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, *34*(3), 772–773.

Liao, T. Y., Kullander, S. O., & Fang, F. (2011). Phylogenetic position of rasborin cyprinids and monophyly of major lineages among the Danioninae, based on morphological characters (Cypriniformes: Cyprinidae). *Journal of Zoological Systematics and Evolutionary Research*, *49*(3), 224–232.

Nelson, J. S., Grande, T., & Wilson, M. V. H. (2016). *Fishes of the world* (5th ed). Hoboken: John Wiley & Sons.

Pramod, P. K., Fang, F., Devi, K. R., Liao, T. Y., Indra, T. J., Beevi, K. S., & Kullander, S. O. (2010). *Betadevario ramachandrani*, a new danionine genus and species from the Western Ghats of India (Teleostei: Cyprinidae: Danioninae). *Zootaxa*, *2519*(1), 31–47.

Rambaut, A., & Drummond, A. J. (2007). Tracer version 1.4. Computer software and documentation distributed by the author. Retrieved from http://tree.bio.ed.ac.uk/software/tracer

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., … Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–542.

Stout, C. C., Tan, M., Lemmon, A. R., Lemmon, E. M., & Armbruster, J. W. (2016). Resolving Cypriniformes relationships using an anchored enrichment approach. *BMC Evolutionary Biology*, *16*(1), 244.

Van Der Laan, R., Eschmeyer, W. N., & Fricke, R. (2014). Family-group names of recent fishes. *Zootaxa*, *3882*(1), 1–230.

Cogent Biology **(ISSN: 2331-2025) is published by Cogent OA, part of Taylor & Francis Group.**