MitoFish and MiFish Pipeline: A Mitochondrial Genome Database of Fish with an Analysis Pipeline for Environmental DNA Metabarcoding

Yukuto Sato,*[,1,2] Masaki Miya,[3] Tsukasa Fukunaga,[4] Tetsuya Sado,[3] and Wataru Iwasaki*[,4,5,6]

[1]Center for Strategic Research Project, Organization for Research Promotion, University of the Ryukyus, Okinawa, Japan

[2]Department of Integrative Genomics, Tohoku Medial Megabank Organization, Tohoku University, Miyagi, Japan

[3]Department of Ecology and Environmental Sciences, Natural History Museum and Institute, Chiba, Chiba, Japan

[4]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan

[5]Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan

[6]Center for Earth Surface System Dynamics, Atmosphere and Ocean Research Institute, The University of Tokyo, Chiba, Japan

***Corresponding authors:** E-mails: yuksato@lab.u-ryukyu.ac.jp; iwasaki@bs.s.u-tokyo.ac.jp.

Associate editor: Sudhir Kumar

Abstract

Fish mitochondrial genome (mitogenome) data form a fundamental basis for revealing vertebrate evolution and hydrosphere ecology. Here, we report recent functional updates of MitoFish, which is a database of fish mitogenomes with a precise annotation pipeline MitoAnnotator. Most importantly, we describe implementation of MiFish pipeline for metabarcoding analysis of fish mitochondrial environmental DNA, which is a fast-emerging and powerful technology in fish studies. MitoFish, MitoAnnotator, and MiFish pipeline constitute a key platform for studies of fish evolution, ecology, and conservation, and are freely available at http://mitofish.aori.u-tokyo.ac.jp/ (last accessed April 7th, 2018).

Key words: **database, fish, mitochondrial genome, metabarcoding, environmental DNA.**

Introduction

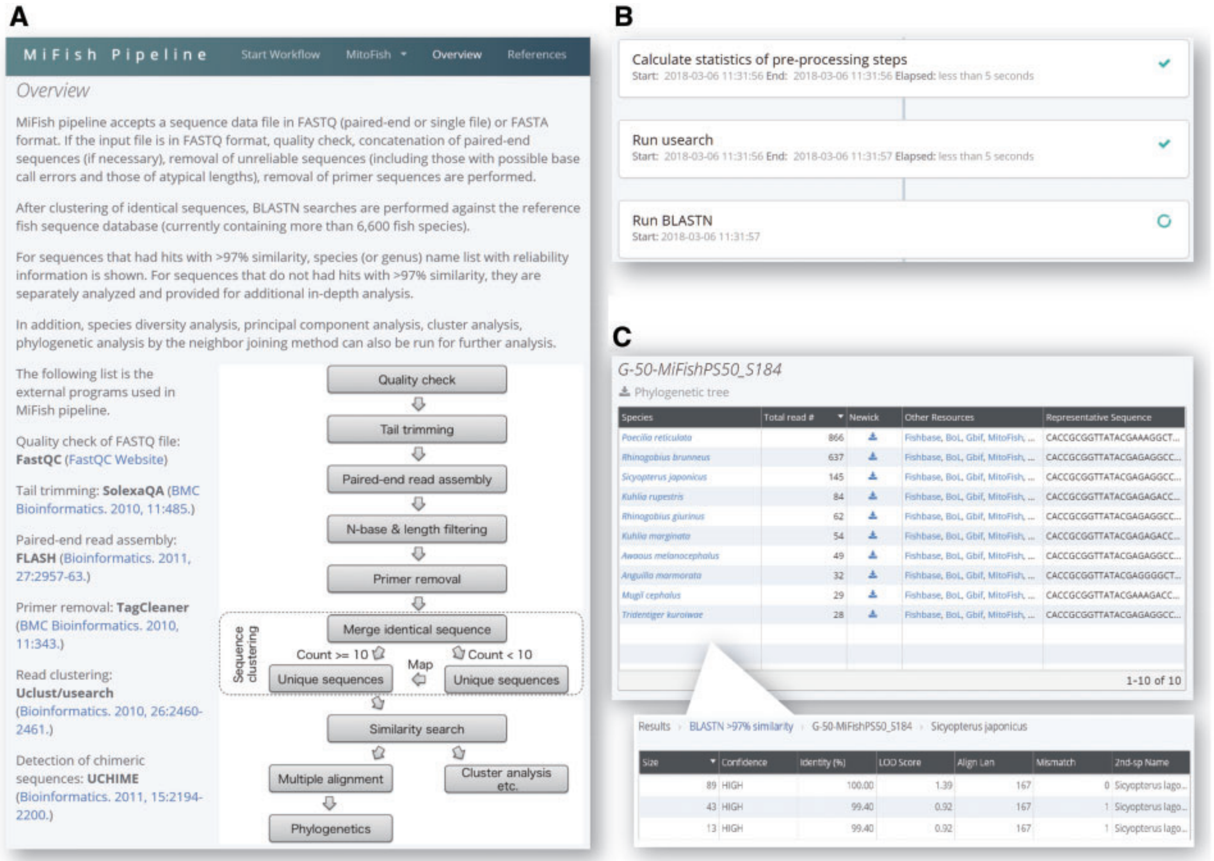
Fish occupy an important position in the vertebrate evolution and hydrosphere ecology, and genetic information from their mitochondrial genomes (mitogenomes) plays a key role in the investigation of their evolutionary histories and the protection and management of biological diversity. Mitofish is a database of fish mitogenomes with precise de novo annotations, and freely available at http://mitofish.aori.u-tokyo.ac.jp/. Since its major update in 2013 (Iwasaki et al. 2013), MitoFish has been actively and widely used by those in evolutionary science, ecology, ichthyology, fisheries, and conservation science from academia, government, and industry. MitoFish and its fish mitogenome annotation pipeline MitoAnnotator now receive >40,000 page views per year from around the world. In addition to its regular update of the data content, MitoFish has acquired two new functions since 2013. One is the multiple sequence annotation function, through which users can easily annotate many mitogenomic sequences for phylogeographic studies, for example.

The other, more important recent functional development in MitoFish is the implementation of MiFish pipeline (http://mitofish.aori.u-tokyo.ac.jp/mifish/). The recent advance in the high-throughput sequencing technology has enabled a new powerful approach in fish studies, that is, the metabarcoding analysis of environmental DNA (eDNA) (Deiner et al. 2017). It has been proved that fish (and tetrapod) mitochondrial DNA can be efficiently amplified by PCR from various environmental samples that include seawater, freshwater, sediment, and gut content (Miya et al. 2015; Ushio et al. 2017). eDNA analysis is a cost-effective and high-throughput approach to investigate species diversity in a noninvasive way, although several factors such as potential contaminations need to be taken cared of. MiFish is a set of universal PCR primers for effective metabarcoding of fish eDNA (Miya et al. 2015). As a powerful metabarcoding tool for biodiversity monitoring, MiFish primers were developed to target a hypervariable region within the fish mitochondrial 12S rRNA gene that is flanked by two highly conservative regions based on the MitoFish data.

MiFish pipeline on the MitoFish server is a user-friendly pipeline for analyzing fish metabarcoding data to estimate the species composition and ecological characteristics of natural environment. Whereas a number of computational tools are available for microbial metabarcoding analysis, there are few for the eDNA metabarcoding analysis of larger organisms. MiFish pipeline serves as a useful tool for those who are interested in diversity and ecological studies of fishes.

MiFish Pipeline

As a new function on the MitoFish server, MiFish pipeline accepts and analyzes fish mitogenomic metabarcoding data,



MiFish Pipeline   Start Workflow   MitoFish   Overview   References

Overview

MiFish pipeline accepts a sequence data file in FASTQ (paired-end or single file) or FASTA format. If the input file is in FASTQ format, quality check, concatenation of paired-end sequences (if necessary), removal of unreliable sequences (including those with possible base call errors and those of atypical lengths), removal of primer sequences are performed.

After clustering of identical sequences, BLASTN searches are performed against the reference fish sequence database (currently containing more than 6,600 fish species).

For sequences that had hits with >97% similarity, species (or genus) name list with reliability information is shown. For sequences that do not had hits with >97% similarity, they are separately analyzed and provided for additional in-depth analysis.

In addition, species diversity analysis, principal component analysis, cluster analysis, phylogenetic analysis by the neighbor joining method can also be run for further analysis.

The following list is the external programs used in MiFish pipeline.

Quality check of FASTQ file: **FastQC** (FastQC Website)

Tail trimming: **SolexaQA** (BMC Bioinformatics. 2010, 11:485.)

Paired-end read assembly: **FLASH** (Bioinformatics. 2011, 27:2957-63.)

Primer removal: **TagCleaner** (BMC Bioinformatics. 2010, 11:343.)

Read clustering: **Uclust/usearch** (Bioinformatics. 2010, 26:2460-2461.)

Detection of chimeric sequences: **UCHIME** (Bioinformatics. 2011, 15:2194-2200.)

Calculate statistics of pre-processing steps ✓
Start: 2018-03-06 11:31:56 **End:** 2018-03-06 11:31:56 **Elapsed:** less than 5 seconds

Run usearch ✓
Start: 2018-03-06 11:31:56 **End:** 2018-03-06 11:31:57 **Elapsed:** less than 5 seconds

Run BLASTN ○
Start: 2018-03-06 11:31:57

G-50-MiFishPS50_S184

Phylogenetic tree

Species	Total read #	Newick	Other Resources	Representative Sequence
Poecilia reticulata	866	⭳	Fishbase, BoL, Gbif, MitoFish, ...	CACCGCGGTTATACGAAAGGCT...
Rhinogobius brunneus	637	⭳	Fishbase, BoL, Gbif, MitoFish, ...	CACCGCGGTTATACGAGAGGCC...
Sicyopterus japonicus	145	⭳	Fishbase, BoL, Gbif, MitoFish, ...	CACCGCGGTTATACGAGAGGCC...
Kuhlia rupestris	84	⭳	Fishbase, BoL, Gbif, MitoFish, ...	CACCGCGGTTATACGAGAGACC...
Rhinogobius giurinus	62	⭳	Fishbase, BoL, Gbif, MitoFish, ...	CACCGCGGTTATACGAGAGGCC...
Kuhlia marginata	54	⭳	Fishbase, BoL, Gbif, MitoFish, ...	CACCGCGGTTATACGAGAGACC...
Awaous melanocephalus	49	⭳	Fishbase, BoL, Gbif, MitoFish, ...	CACCGCGGTTATACGAGAGGCC...
Anguilla marmorata	32	⭳	Fishbase, BoL, Gbif, MitoFish, ...	CACCGCGGTTATACGAGGGGCT...
Mugil cephalus	29	⭳	Fishbase, BoL, Gbif, MitoFish, ...	CACCGCGGTTATACGAAAGACC...
Tridentiger kuroiwae	28	⭳	Fishbase, BoL, Gbif, MitoFish, ...	CACCGCGGTTATACGAGAGGCC...

1-10 of 10

Results › BLASTN >97% similarity › G-50-MiFishPS50_S184 › Sicyopterus japonicus

Size	Confidence	Identity (%)	LOD Score	Align Len	Mismatch	2nd-sp Name
89	HIGH	100.00	1.39	167	0	Sicyopterus lago...
43	HIGH	99.40	0.92	167	1	Sicyopterus lago...
13	HIGH	99.40	0.92	167	1	Sicyopterus lago...

Fig. 1. Screenshots of MiFish pipeline. (*A*) An overview of the pipeline. (*B*) A view of a progress status shown during pipeline execution. (*C*) An HTML report that contains results of the species assignment, sequence counts, and web-page links. For each species, a detailed report with confidence scores is provided.

which include those produced using the MiFish primers (fig. 1*A*). The overall sequence quality is assessed by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and low-quality (Phred score < 10 by default) 3′-tails are trimmed by DynamicTirm.pl (Cox et al. 2010). The paired-end reads are merged by FLASH (Magoc and Salzberg 2011) and erroneous merged reads that contain N-nucleotides or do not have typical lengths are removed (229 ± 25 bp by default). The primer sequences are removed by TagCleaner (Schmieder et al. 2010) by allowing three-base mismatches at the maximum. Species-level taxonomic assignment is performed using Uclust (Edgar 2010) and NCBI Blast+ (Camacho et al. 2009; fig. 1*B*). Redundant sequences are merged into one sequence by keeping the count information. Then, low read-number sequences (<10 by default) are remapped onto high read-number sequences (≥10) at a given sequence-similarity threshold (99% by default), and the unmapped sequences are discarded. Blastn searches are conducted against MitoFish as a reference database with cutoff values of identity 97% and *e*-value 10^{-5}, and species names of the top-hit sequences are retrieved. If the second to fifth top-hit sequences of each Blast search contain those of different species, confidence scores of the species assignment are calculated using the following formula:

$$\ln \frac{\begin{array}{c}(\text{aligned length of the top-hit sequence})/\\(\text{mismatch numbers of the top-hit sequence}+1)\end{array}}{\begin{array}{c}(\text{aligned length of the second-hit sequence})/\\(\text{mismatch numbers of the second-hit sequence}+1)\end{array}}$$

All-species and within-species molecular phylogenetic trees are estimated for each environmental sample. Multiple sequence alignments are generated by MAFFT (Katoh and Standley 2013) and neighbor-joining phylogenetic trees are estimated by Morphy (Adachi and Hasegawa 1992). An HTML report is finally presented, which can also be used for calculating ecological indices such as alpha diversity, beta diversity, and correlation coefficients (fig. 1*C*). This report also contains links to major databases such as FishBase (Froese and Pauly 2017), Barcode of Life (Ratnasingham and Hebert 2007), Global Biodiversity Information Facility (Edwards 2004), and MitoFish.

Figure 2 shows an example of fish eDNA analysis results produced by MiFish pipeline. The water sample was taken at Uchidomari river in Okinawa Island, Japan (fig. 2*A*) and filtrated up to 960 ml using 0.45 μm Sterivex filter (Millipore) or 0.70 μm glass-fiber filter (Whatman GF/F). DNA was extracted using DNeasy PowerWater Sterivex and DNeasy Blood & Tissue kits (Qiagen) from the Sterivex and

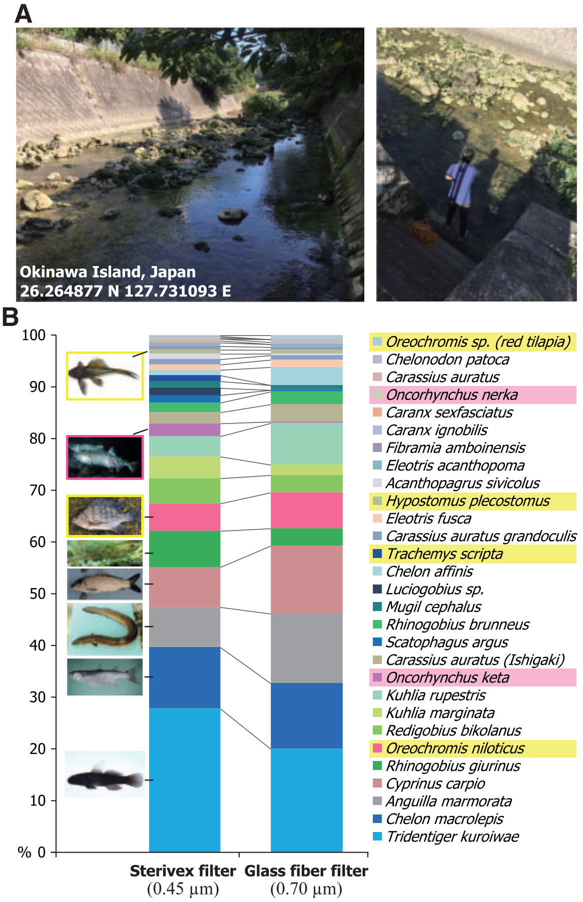


FIG. 2. Application of MiFish pipeline to a fish eDNA data set. (*A*) Photos of the sampling site in Okinawa Island, Japan. (*B*) A summary of the species assignment results produced by MiFish pipeline. The left and right bars indicate the data of eDNA extracted using the Sterivex and glass-fiber filters, respectively. Read numbers are expressed as percentages to the total read numbers. Nonnative species in Okinawa Island are highlighted (Yellow: Invasive species, Pink: Salmons). Fish pictures are from FishBase.

glass-fiber filters, respectively. MiFish primers (Miya et al. 2015) were used to amplify eDNA with the annealing temperature of 60°C. MiSeq with V2 chemistry (Illumina) was used for 150-bp paired-end sequencing.

The estimated species composition well represented the fish community in the rivers in Okinawa Island (fig. 2*B*). Dominant species included those typically observed in Okinawa Island rivers such as gobies (e.g., *Tridentiger kuroiwae*, *Rhinogobius giurinus*, and *Redigobius bikolanus*), mullets (e.g., *Chelon macrolepis* and *Chelon affinis*), and giant mottled eel (*Anguilla marmorata*), as well as invasive alien species such as nonnative tilapias (genus *Oreochromis*) and plecos (*Hypostomus plecostomus*). It may be noted that salmons (genus *Onchorhynchus*) also appeared in the list presumably because they are widely consumed as food in Okinawa and drainage likely contains their eDNA. These results would exemplify that MiFish pipeline is a useful tool for eDNA analysis of the endemic fish species composition, invasive species, and also impact of human activities, whereas the detection of salmons also suggests that eDNA analysis can be affected by unexpected environmental influences and/or contamination and need to be interpreted with caution. Taken together, MitoFish, MitoAnnotator, and MiFish pipeline constitute a key platform for studies of evolution, ecology, and conservation of fishes.

Acknowledgments

The authors thank Mutsumi Nishida, Koichiro Yamada, Takashi P. Satoh, Ryota Isagozawa, Yasunobu Maeda, Kohji Mabuchi, and Hirohiko Takeshima for their help in the development of MitoFish. This work was supported by the Japan Science and Technology Agency (CREST), the Japan Society for the Promotion of Science (Grant Numbers 15K18590, 16H06279, 17K19298, and 17HP8031), the University of the Ryukyus (the Spatiotemporal Genomics Project), and the Canon Foundation. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics. Sequence data are deposited in DDBJ Sequence Read Archive with an accession number DRA006629.

References

Adachi J, Hasegawa M. 1992. MOLPHY: programs for molecular phylogenetics, I. PROTML: maximum likelihood inference of protein phylogeny. Computer Science Monographs, No. 27. Tokyo: Institute of Statistical Mathematics.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11:485.

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, et al. 2017. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol Ecol*. 26(21):5872–5895.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460–2461.

Edwards JL. 2004. Research and societal benefits of the Global Biodiversity Information Facility. *Bioscience* 54:485–486.

Froese R, Pauly D. 2017. FishBase. Available from: www.fishbase.org, version October, 2017.

Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, Sado T, Mabuchi K, Takeshima H, Miya M, Nishida M. 2013. MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Mol Biol Evol*. 30(11):2531–2540.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.

Magoc T, Salzberg S. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27(21):2957–2963.

Miya M, Sato Y, Fukunaga T, Sado T, Poulsen JY, Sato K, Minamoto T, Yamamoto S, Yamanaka H, Araki H, et al. 2015. MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *R Soc Open Sci*. 2(7):150088.

Ratnasingham S, Hebert PD. 2007. bold: the Barcode of Life Data System (http://www.barcodinglife.org). *Mol Ecol Notes* 7(3):355–364.

Schmieder R, Lim YW, Rohwer F, Edwards R. 2010. TagCleaner: identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* 11:341.

Ushio M, Fukuda H, Inoue T, Makoto K, Kishida O, Sato K, Murata K, Nikaido M, Sado T, Sato Y, et al. 2017. Environmental DNA enables detection of terrestrial mammals from forest pond water. *Mol Ecol Resour*. 17(6):e63–e75.