

EDM Bioinformatic SOP

U.S. Fish and Wildlife Service

Reference Database Curation

The following steps were taken to create the reference databases used to assign taxonomy to the genetic sequences :

- 1) **Created a target species list.** Created a target species list of native and established freshwater species for the Great Lakes using a combination of the GAPeDNA dataset (<https://shiny.cefe.cnrs.fr/GAPeDNA/> ;for the Saint Laurent basin) and the freshwater Aquatic Invasive Species (AIS) watchlist and a list of nonindigenous/range expanders (fish only) from the Great Lakes Aquatic Nonindigenous Species Information System (<https://www.glerl.noaa.gov/glansis/> ; GLANSIS; All Great Lakes Drainages).
- 2) **Pulled reference sequences.** Mitochondrial sequences from fish species were downloaded from the National Center for Biotechnology Information (NCBI). In addition to fish sequences, human and bovine mitochondrial sequences were downloaded from NCBI. These additional sequences provide insight into base levels of background eDNA (non-target DNA that is found in the environment including the field and lab). Reference sequences with known issues (i.e. wrong taxonomic assignment, high error sequences, etc.) were removed. Further, the software 'CRABS' v1.14.0 was used to perform in-silico PCR and isolate amplicon regions using the MiFish primers (Miya et al. 2020), dereplicate identical sequences, keep only sequences that were at least 150 base pairs, and remove sequences with more than 5 Ns.
- 3) **Created target reference database.** The fish sequences pulled from the NCBI were filtered to only those that were in the "target species list". These sequences were added to the human sequences from NCBI and the synthetic sequences that are used as an internal lab control (gBlock). These compiled sequences made up what is referred to as the "Target Reference Database". A list of species that made up the reference database was created. By using a curated reference database first, we can reduce the number of misassigned sequencing reads (Mugnai et al., 2023).
- 4) **Created 'unclassified' reference database.** All remaining reference sequences from the NCBI dataset were kept in a separate, alternative reference database. Any sequences that did not align to the target reference database will be queried against this 'unclassified' reference database.

Preparing Sequence Data and Conducting Quality Filtering

Sequences were demultiplexed via their unique index sequence combinations via the onboard software on the sequencer (NextSeq- DRAGEN). Each sample has an R1 and an R2 file that contains the raw sequence data. Each sequencing read went through multiple stages of quality filtering using the following steps:

- 1) **Fraction sequenced reads using primer tags.** Each of these sequences contains an in-line tag added to the MiFish primers. These primer tags were used to improve our ability to detect and remove contaminated sequences. The software cutadapt v5.2 was used to remove sequences that did not have the corresponding matching primer tags, allowing for 1 base-pair of variation per primer tag, before taxonomic classification (Martin, 2011). Further, we used cutadapt to remove primer sequence using an error rate of 0.2. Removing both the in-line tag and primer reduces the chance of the synthetic DNA constructs reducing the accuracy of the taxonomic assignments.
- 2) **Quality filtering sequence reads.** The software fastp v1.0.1 was used to 1) trim low quality base-pairs from the beginning and end of the sequence using a sliding window (the cut window was 4 base pairs, and reads were trimmed if the average quality phred score was <20 within the cut window), 2) remove sequences with an excess of Ns (occurs when sequencer software cannot confidently identify a nucleotide base; maximum of 5 Ns per read were allowed), and 3) remove sequences that were less than 150 base pairs long after quality filtering was applied.

Assigning Taxonomy to Sequences

Raw genetic sequences from each sample were aligned to the reference databases using the following steps:

- 1) **Aligned genetic sequences to the target reference database.** Used the software Centrifuge v1.0.4.2 to match filtered sequence data to the target reference database. An alignment had to have a hit length of ≥ 150 bp to be considered a match (Kim et al., 2016). The species with the highest match (in terms of total length and % base pair similarity) was selected as the species ID for that sequence. In instances where multiple species from the reference database had a high match to the raw sequence, the taxonomy was collapsed to the genus level. Raw sequences that did not have a high match to any species in the target reference database were put in a separate category titled 'unclassified' for further analyses (described in step 2 below).
- 2) **Aligned 'unclassified' reads to the 'unclassified' reference database.** All raw sequence reads in the 'unclassified' category from step 1 were further processed. The alignment methodology outlined in step 1 was repeated for 'unclassified' sequence reads except the 'unclassified' reference database was used instead of the target reference database. This allowed for novel detections and potentially catching species that were missed or not considered when creating the target reference database (this database encompasses a greater geographic representation of fish taxa).
- 3) **Created contingency tables.** For each sample, the number of sequences that matched each species in the target reference database were summed in a contingency table. Each row of the contingency table contained the total number of assigned reads for each species while each column represented a unique eDNA sample.
- 4) **Assigned common names to taxa in the contingency tables.** Lists of all Great Lakes fish species found between EDM programs are maintained and used to assign common names to species. If a species is detected in metabarcoding and not on the list, the R package 'Taxize' v0.10.1 was used to assign a common name. Species assignments were evaluated for taxonomic accuracy. For instance, species with identical sequences at the MiFish site

were checked to ensure that they were elevated to genus level taxonomic assignment when species resolution is not possible. Taxonomic assignments were reviewed and any species identified that is not commonly detected in the Great Lakes basin (based on data collected by the USFWS using traditional netting and electrofishing) were evaluated by examining the specific reference database sequence that matched the collected eDNA sample. Examinations included but were not limited to verifying correct taxonomic assignment of specific reference sequence, general reference sequence quality, and similarity of reference sequence to other taxa for the 12S MiFish gene region.

Performance Verification of Control Samples

Each sequencing run included a variety of control samples to ensure the integrity and reliability of the results. These controls comprised field blanks, extraction negatives, PCR negatives, and PCR positives (synthetic gblock sequences).

- 1) **Evaluation of PCR Positive Controls.** The PCR positive controls were examined to ensure the presence of gblock sequences. This assessment validated that the PCR reactions effectively amplified the target DNA, confirming the accuracy and efficiency of the amplification step.
- 2) **Assessment of Negative Controls.** Field blanks, extraction negatives, and PCR negatives were checked to determine if any sequencing reads had been assigned to fish taxa. Any detection of such reads in these negative controls indicated potential contamination, and samples with these results were flagged for downstream decontamination processing as outlined in the decontamination section.

Identify and Remove Failed Samples

After sequence reads went through quality control filters and taxonomic assignment, the PCR negative controls were used to identify samples that failed to amplify or sequence. If an environmental sample had less than the minimum number of reads across PCR negative controls (while containing at least 1,000 reads), the sample was dropped before decontamination and denoising. If a PCR negative had more than 10 reads assigned to any taxon, indicating contamination, it was not used for calculating the threshold to define failed samples.

Bioinformatic Decontamination of eDNA Samples

Metabarcoding is an extremely sensitive method for detecting eDNA, so even small amounts of foreign/non-target DNA can generate sequence reads. The field and laboratory process included multiple controls in order to detect contamination as well as determine when contamination occurred (field vs laboratory). We used these negative controls as well as minimum quality thresholds to account for contamination in the metabarcoding results. This bioinformatic decontamination included multiple steps—first to account for internal contamination that can occur when filters come into contact with a higher concentration of DNA from field or laboratory practices (results in a high sequence read count in negative controls) and then to account for base-level contamination that commonly results from free-floating DNA in the air and surroundings as well as sequence tag-jumping (usually resulting in low read counts in sample data; Schnell et al. 2015).

- 1) **Account for negative control contamination.** Next, we used negative controls (field, extraction, and PCR) to 1) identify contaminant species, 2) determine how widespread the contamination is, and 3) remove contamination by performing a common practice referred to as “read subtraction” (McKnight et al. 2019). To avoid overcorrecting for contamination, only negative controls with more than 10 reads and had a relative read frequency greater than 0.1% from a fish taxon were processed for decontamination (this low number of reads could be due to sequencing error or tag-jumping). Further, environmental samples only underwent bioinformatic decontamination if 2 or more field controls were contaminated (more than 10 reads and relative read frequency greater than 0.1%) by the same species for a given sampling event (unique sample location and timeframe). We used these guidelines to decide which samples underwent read subtraction:
 - A. **PCR Negative Contamination.** If sequencing reads were assigned to a PCR negative, the PCR plate was flagged for review by metabarcoding leads for further action (i.e. rerun the plate or perform bioinformatic decontamination).
 - B. **Extraction Negative Contamination.** If sequencing reads were assigned to a fish species in a laboratory extraction negative, read subtraction for the contaminant taxon was conducted across all samples processed in that extraction batch.
 - C. **Field Blank Contamination.** If sequencing reads were assigned to the same fish taxon in two or more field blanks for the same collection event, but the lab controls did not have any reads assigned to the same fish species, read subtraction was conducted for all samples collected before and after the contaminated field blanks until the next uncontaminated (or clean) field blank for that sampling event.
 - D. **Contamination of Multiple Negative Controls.** If sequencing reads from the same taxon were assigned to multiple negative controls, the maximum relative read abundance for that taxon within the negative controls was used for read subtraction across all impacted groups of samples.
- 2) **Remove background sequencing noise and account for tag-jumping.** To account for background contamination and tag-jumping (process that results from unique sample identifiers moving between samples in the library prep process), we used two filter thresholds. We used an absolute read abundance (total sequencing read count) to remove any species within a sample that had 10 or fewer reads. Then any sequencing reads that were below a minimum relative abundance of 0.1% were removed. This dual-filter process avoids over-filtering low sequencing depth samples and under-filtering high sequencing depth samples.

Upload data to ArcGIS Online “USFWS Database”

Finalized results were appended to the ID table within the [AGOL USFWS eDNA Projects Database](#) by a data scientist.

Interactive Report Generation

After the dataset was cleaned and appended, the data scientist will create an interactive report utilizing a html-based template that is stored in a shared GitHub repository.

FWCO Review of Results

The data scientist who created the interactive report will deliver the product to the appropriate FWCO. The FWCO has five business days to review the report and notify the data scientist of any errors or concerns. Specifically, staff should:

- Review classifications
- Review species status (e.g., native status, AIS determination)
- Assess the unexpected species and flag results of interest
- Confirm the AIS Watchlist table

When completed, the FWCO will notify the data scientist of any concerns, interesting results, or if the report looks OK to deliver via email.

Results Report sent to EDM Coordinator

A data scientist will deliver the results package to the EDM Coordinator. The EDM coordinator is responsible for verifying receiving entities, reviewing the species watchlist table, and noting any flagged species.

Great Lakes EDM Communications Plan Initiated

The EDM Coordinator will communicate results internally to USFWS leadership, followed by external communication to jurisdictional entities, as outlined in the *USFWS Great Lakes Early Detection and Monitoring (EDM) Program Communications Plan*. The EDM Coordinator will fill out the planning survey with the date sent to the entities, which establishes the date the data becomes publicly available. If an entity requests a delay to public posting or plans to issue a press release, the EDM Coordinator will check the adjust posting date box and enter the date the data will be posted.

References

Davis NM, Proctor DM, Holmes SP, et al. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, 6: 226.
<https://doi.org/10.1186/s40168-018-0605-2>

Hülpüsch C, Rauer L, Nussbaumer T, Schwierzeck V, Bhattacharyya M, Erhart V, Traidl-Hoffmann C, Reiger M & Neumann AU (2021). MicroblEM: A user-friendly tool for quality control and interactive analysis of microbiome data. Available at: <https://github.com/LuiseRauer/MicroblEM>

Karstens L, Asquith M, Davin S, Fair D, Gregory WT, Wolfe AJ, et al. (2019). Controlling for contaminants in low-biomass 16S rRNA gene sequencing experiments. *mSystems*, 4(4): e00290-19.
<https://doi.org/10.1128/mSystems.00290-19>

Kim D, Song L, Breitwieser FP & Salzberg SL (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*. <https://doi.org/10.1101/gr.210641.116>

Martin M (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1): 10–12. <https://doi.org/10.14806/ej.17.1.200>

McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA & Zenger KR (2019). microDeCon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. *Environmental DNA*, 1(1): 14–25. <https://doi.org/10.1002/edn3.11>

Miya M, Gotoh RO & Sado T (2020). MiFish metabarcoding: A high-throughput approach for simultaneous detection of multiple fish species from environmental DNA and other samples. *Fisheries Science*, 86(6): 939–970. <https://doi.org/10.1007/s12562-020-01461-x>

Mugnai F, Costantini F, Chenuil A, Leduc M, Gutiérrez Ortega JM & Meglécz E (2023). Be positive: Customized reference databases and new, local barcodes balance false taxonomic assignments in metabarcoding studies. *PeerJ*, 11: e14616. <https://doi.org/10.7717/peerj.14616>

Sato Y, Miya M, Fukunaga T, Sado T & Iwasaki W (2018). MitoFish and MiFish pipeline: A mitochondrial genome database of fish with an analysis pipeline for environmental DNA metabarcoding. *Molecular Biology and Evolution*, 35: 1553–1555.

Schnell IB, Bohmann K & Gilbert MTP (2015). Tag jumps illuminated: Reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, 15(6): 1289–1303.
<https://doi.org/10.1111/1755-0998.12402>