Fish Mitogenome Assembly Pipeline

Haley Atkins, Roza Gawin, Pranati Sukh, Vir Trivedi

Abstract

Across animal species, mitochondrial DNA (mtDNA) contains essential protein-coding rRNA and tRNA genes, alongside limited non-coding regions, characterized by maternal transmission, absence of recombination, and an elevated mutation rate relative to nuclear DNA. These properties make mtDNA an invaluable tool for reconstructing phylogenetic histories, allowing us to trace individual lineages and facilitate comparisons within populations and across distantly related species. Environmental DNA (eDNA) refers to the genetic material released by organisms into their environment, like bodies of water, where it can be sampled and analyzed to detect the presence of species without directly observing them. eDNA studies can provide valuable insights into the species' distribution, abundance, and habitat preferences, aiding in conservation efforts and management strategies. Mitochondrial reference databases can be useful tools for combining these two disciplines, however, there is no current consensus for mitogenome assembly for fish species. Here we show that GetOrganalle[2] can generate equally accurate mitogenomes as the SMART2 assembler (Shew and Lema[5]) while improving ease of use. For both assemblies, sequencing data from the skeletal muscles of the jack silverside were used with the jack silverside cytochrome B gene sequence as the seed sequence. The resulting mitogenome from this pipeline, when compared to the published assembly, returns 99.99% identity and 100% query coverage. The Mitofish annotation of the two also produces identical visualizations of the mitogenomes. These results support the use of this simplified pipeline for fish mitogenome assembly going forward. Ultimately this pipeline can be used to generate the mitogenomes for a fish mitogenome database for future eDNA studies.

Introduction

The biological problem addressed in this project is the need to assemble and characterize mitochondrial genomes of fish species. To test our methods, raw sequencing reads of the jack silverside (*Atherinopsis californiensis*), a fish species found in California, were used. Understanding the genetic makeup of species through mitogenome assembly is essential as it illuminates the species' evolutionary history, population structure, and ecological role within its habitat. As was performed in "Mitochondrial genome of the jack silverside, *Atherinopsis californiensis* (Atherinopsidae, Atheriniformes), a nearshore fish of the California Current Ecosystem," [5] SMART2[1] takes a partial sequence of the fish gene as the "seed sequence" and utilizes it for the assembly process. FastQC[8] is then employed to assess the quality of the raw sequencing reads. Fastp[6] is used to preprocess the raw sequencing reads and trim and filter out

low-quality reads. Bowtie2[3] is utilized to align the trimmed reads. Sequencher[4] and MitoFish[7] are employed to edit the sequence alignments and annotate the assembled mitochondrial genome.

We have developed a computational pipeline to simplify the workflow of assembling and analyzing marine genomic data. While the pipeline follows the general process explored in Shew and Lema's paper, we have opted for GetOrganelle[2] as a command line alternative for the web based SMART2.[5] The complete source code and documentation are available in the GitHub repository (https://github.com/rozgaw/MitogenomeAssemblyPipeline.git).

Implementation

The pipeline's input includes raw DNA sequencing data obtained from the skeletal muscle tissue of jack silverside fish (*Atherinopsis californiensis*). This data includes millions of short DNA sequence reads generated through high-throughput sequencing techniques. The pipeline's main output is the jack silverside's complete mitochondrial genome sequence.

The pipeline includes data preprocessing, genome assembly, and annotation beginning from raw sequence reads. These raw reads are first trimmed via fastp[6] to remove adaptors. The trimmed sequence reads are assembled into a complete mitochondrial genome sequence, and the assembly process involves aligning the reads with a reference sequence and reconstructing the complete genome. The assembled genome is annotated to identify protein-coding genes and other genetic elements using tools like MitoFish.[7] We utilized existing tools such as GetOrganelle[2], which requires dependencies such as Bowtie for indexing and aligning sequences and SPAdes for genome assembly. Additionally, we used MitoFish, a mitochondrial genome database for fish, to aid in annotating and comparing our assembled genomes.

The pipeline's accuracy in producing a complete and correct assembly is measured through ease of use and assembly alignment to published values. More specifically, we will be comparing the assembled genome to different reference sequences and validating the annotations against existing genomic databases. The complexity of the pipeline is evaluated based on factors such as efficiency and ease of use for researchers with different levels of bioinformatics exposures. To test the accuracy of all assemblies built through GetOrganelle[2], we ran the web megaBLAST tool to compare the generated assembly to the result achieved in the original paper.

Results and Discussion

The study done by Shew and Lema[5] involved a cut up pipeline utilizing both command line tools and web interface tools. In our workflow design, we wanted to create a more seamless pipeline that involved mostly command line tools. Our pipeline cuts the need for almost all web tools

besides the MitoFish tool MitoAnnotator[7] for the genome annotations and visuals.

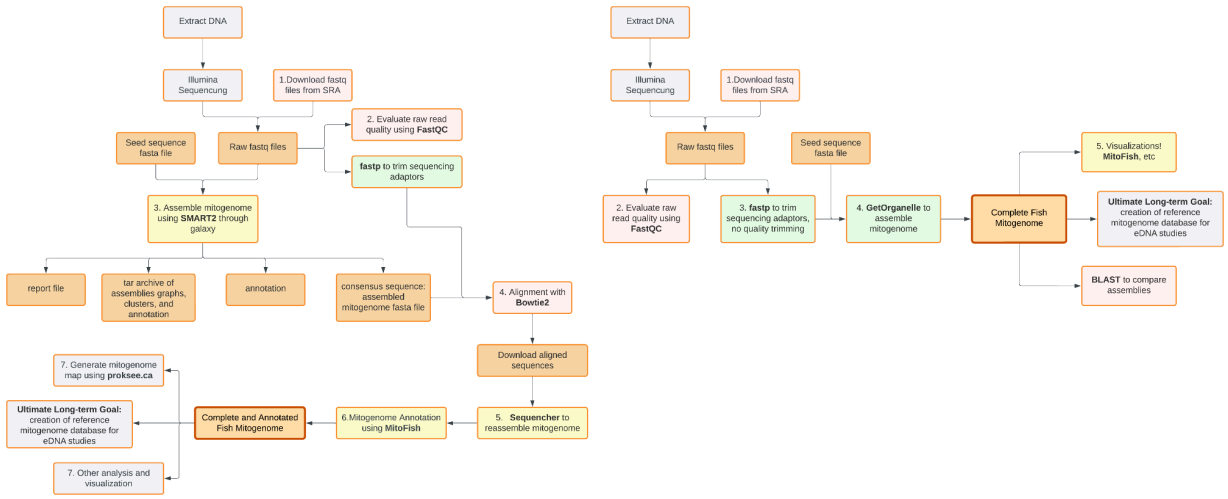

Figure 1. Comparison of "Mitochondrial genome of the jack silverside, Atherinopsis californiensis (Atherinopsidae, Atheriniformes), a nearshore fish of the California Current Ecosystem" workflow[5] (left) to mitogenome assembly with GetOrganelle (right). Each step is color coded as follows. Pink: tools which can be run using the command line and web interfaces or applications. Green: tools which can only be run using the command line (or python). Yellow: tools which can only be run through a web interface or application. Orange: specific input or output files. Grey: steps not included in pipeline. In the workflow on the right, all tools are run from the command line, except for MitoFish which is only available as a web interface.

Implementing GetOrganelle[2] for the assembly of the jack silverside mitogenome with the cytochrome B seed sequence (GenBank accession no. JQ282018) yielded an assembly with 100% query coverage and 99.99% identity to the assembly produced in Shew and Lema. This identity was determined using the web megaBLAST tool to align the generated assembly to the published assembly (see figure 2). The visualization using MitoFish[7] for both assemblies also produced the same mitogenome map (see figure 3).

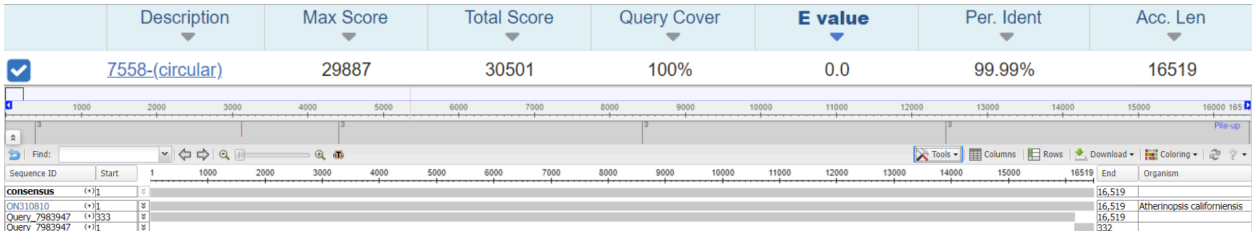

Figure 2. BLAST results comparing our assembly using the *A. californiensis* cytochrome B seed sequence to the published *A. californiensis* mitogenome assembly (GenBank accession no. ON310810).

Since this pipeline is developed specifically for the analysis of the mitochondrial genome of the jack silverside using the same sequencing reads and seed file as in Shew and Lima, sequences stemming from other species may not yield identical results. The assemblies may vary slightly

depending on what sort of seed sequences are used. In the GetOrganelle[2] documentation, it states that complete organelle genome sequences of a related species as the seed can help the assembly in many cases. This means that the seed sequence does not need to be a species specific mitochondrial gene such as we used in this assembly.

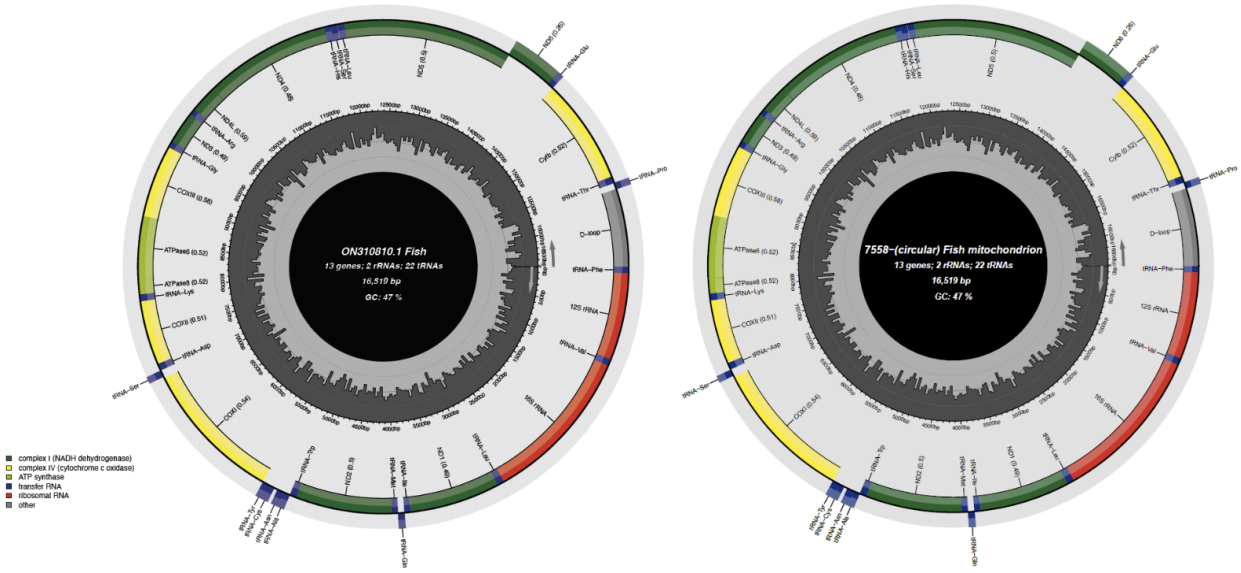

Figure 3. Comparison of the mitogenome visualizations produced by MitoFish of the GetOrganelle generated jack silverside mitogenome (right) and the published jack silverside mitogenome (GenBank accession no. ON310810) (left).

To further explore what sorts of seed sequences can be used for mitogenome assemblies with GetOrganelle, we ran an assembly of the Californian anchovy (*Engraulis mordax*) using our assembled jack silverside mitogenome as the seed. These two fish share the same taxonomic class, but are different orders, showing that they are related, but not very closely. This assembly took longer than when we had a species specific seed sequence, but still produced an assembled mitogenome with 100% query coverage and 99.75% identity with the published *E. mordax* mitogenome (NCBI accession: NC_041097.1) (see figure 4).

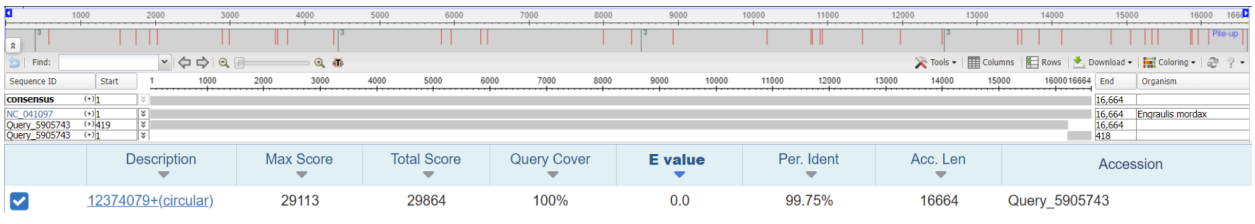

Figure 4. BLAST results comparing our assembled *E. mordax* mitogenome (seed sequence: assembled *A. californiensis* mitogenome) to the published *E. mordax* mitogenome (GenBank accession no. NC_041097.1).

We also compared both published *A. cali* and *E. mordax* mitogenomes to each other as well as both of our assemblies to see how similar the mitogenomes were overall (see figure 5). The query coverage for both of those megaBLASTs was 65% with percent identities at about 75% for both.

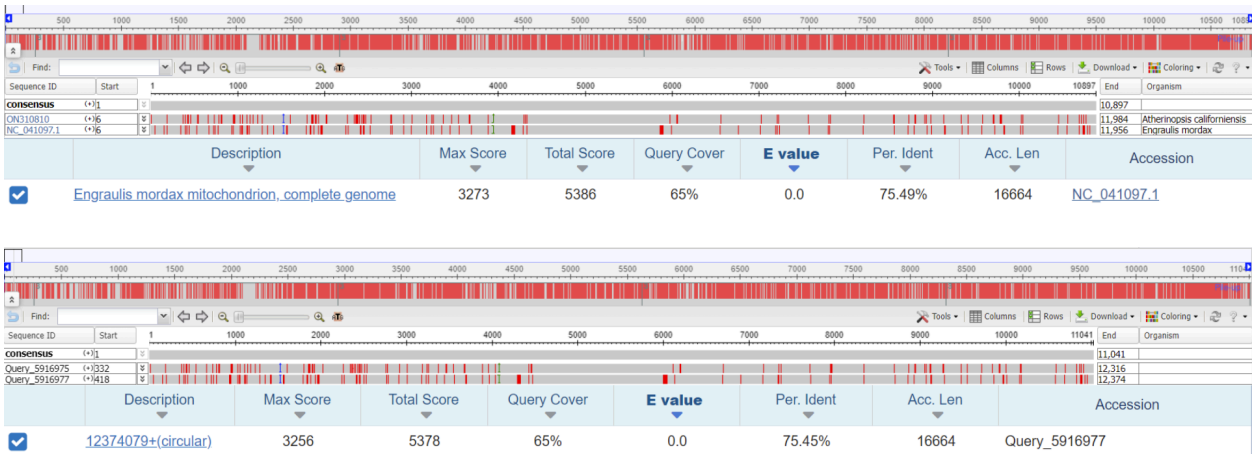

Figure 4. (Top) BLAST results comparing *A. californiensis* mitogenome (GenBank accession no. ON310810) to the *E. mordax* mitogenome (GenBank accession no. NC_041097.1). (Bottom) BLAST results comparing our assembled *A. californiensis* mitogenome (seed sequences: CytB GenBank accession no. JQ282018) to the assembled *E. mordax* mitogenome (seed sequence: assembled *A. californiensis* mitogenome)

These results help showcase how complete mitogenomes of more distantly related organisms can still be used as seed sequences for mitogenome assembly. All fish are grouped into three taxonomic classes, so theoretically even when assembling large amounts of fish mitogenomes, the same seed sequences should be able to be reused instead of having to find a new seed for each fish species. Ultimately, this will have to be explored further but it highlights the fact that mitogenome assemblies can still be produced by GetOrganelle using mitogenomes of species in different taxonomic orders.

References

[1] Alqahtani, Fahad and Măndoiu, Ion I (2020). Statistical mitogenome assembly with repeats. In *Journal of Computational Biology*

[8] Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data

[2] Jian-Jun Jin*, Wen-Bin Yu*, Jun-Bo Yang, Yu Song, Claude W. dePamphilis, Ting-Shuang Yi, De-Zhu Li. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology* 21, 241 (2020). https://doi.org/10.1186/s13059-020-02154-5

[3] Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, 9:357-359.

[4] Sequencher® version 5.4.6 DNA sequence analysis software, Gene Codes Corporation, Ann Arbor, MI USA http://www.genecodes.com

[5] Shew JE, Lema SC. Mitochondrial genome of the Jack Silverside, *atherinopsis californiensis* (Atherinopsidae, Atheriniformes), a nearshore fish of the California Current Ecosystem. Mitochondrial DNA Part B. 2023 Jan 1;8(1):13–7. doi:10.1080/23802359.2022.2158047

[6] Shifu Chen. 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta 2: e107. https://github.com/OpenGene/fastp

[7] Tao Zhu, Yukuto Sato, Tetsuya Sado, Masaki Miya, Wataru Iwasaki, MitoFish, MitoAnnotator, and MiFish Pipeline: Updates in 10 Years, *Molecular Biology and Evolution*, Volume 40, Issue 3, March 2023, msad035, https://doi.org/10.1093/molbev/msad035