MitoChime: A Machine-Learning Pipeline for Detecting PCR-Induced Chimeras in Mitochondrial Illumina Reads

A Special Project Proposal

Presented to

the Faculty of the Division of Physical Sciences and Mathematics

College of Arts and Sciences

University of the Philippines Visayas

Miag-ao, Iloilo

In Partial Fulfillment

of the Requirements for the Degree of

Bachelor of Science in Computer Science

by

DURAN, Duranne

LIN, Yvonne

PAILDEN, Daniella

Adviser

Francis DIMZON

October 23, 2025

Contents

List of Figures

List of Tables

Chapter 1

0.1 Introduction

0.1.1 Overview

The rapid advancement of next-generation sequencing (NGS) technologies has transformed genomic research by enabling high-throughput and cost-effective DNA analysisMetzker (2010). Among current platforms, Illumina sequencing remains the most widely adopted, capable of producing millions of short reads that can be assembled into reference genomes or analyzed for genetic variationBentley et al. (2008); Glenn (2011). Despite its high base-calling accuracy, Illumina sequencing is prone to artifacts introduced during library preparation, particularly polymerase chain reaction (PCR)-induced chimeras, which are artificial hybrid sequences that do not exist in the true genomeJudo, Wedel, and Wilson (1998).

PCR chimeras form when incomplete extension products from one template anneal to an unrelated DNA fragment and are extended, creating recombinant readsQiu et al. (2001). In mitochondrial genome assembly, such artifacts are especially problematic because the mitochondrial genome is small, circular, and often repetitiveBoore (1999); Cameron (2014). Even a small number of chimeric or mis-joined reads can reduce assembly contiguity and introduce false junctions during organelle genome reconstructionDierckxsens, Mardulyn, and Smits (2017); Hahn, Bachmann, and Chevreux (2013); Jin et al. (2020). Existing assembly tools such as GetOrganelle and MITObim assume that input reads are largely free of such artifactsHahn et al. (2013); Jin et al. (2020). Consequently, unde-

tected chimeras may produce fragmented assemblies or misidentified organellar boundaries. To ensure accurate reconstruction of mitochondrial genomes, a reliable and automated method for detecting and filtering PCR-induced chimeras before assembly is essential.

This study focuses on mitochondrial sequencing data from the genus *Sardinella*, a group of small pelagic fishes widely distributed in Philippine waters. Among them, *Sardinella lemuru* (Bali sardinella) is one of the country's most abundant and economically important species, providing protein and livelihood to coastal communitiesLabrador, Agmata, Palermo, Ravago-Gotanco, and Pante (2021); Willette, Bognot, Mutia, and Santos (2011). Accurate mitochondrial assemblies are critical for understanding its population genetics, stock structure, and evolutionary history. However, assembly pipelines often encounter errors or fail to complete due to undetected chimeric reads. To address this gap, this research introduces **MitoChime**, a machine-learning pipeline designed to detect and filter PCR-induced chimeric reads using both alignment- and sequence-derived statistical features. The tool aims to provide bioinformatics laboratories, particularly the Philippine Genome Center Visayas, with an efficient, interpretable, and resource-optimized solution for improving mitochondrial genome reconstruction.

0.1.2 Problem Statement

While NGS technologies have revolutionized genomic data acquisition, the accuracy of mitochondrial genome assembly remains limited by artifacts produced during PCR amplification. These chimeric reads can distort assembly graphs and cause misassemblies, with especially severe effects in small, circular mitochon-

drial genomesBoore (1999); Cameron (2014). Existing assembly pipelines such as GetOrganelle, MITObim, and NOVOPlasty assume that sequencing reads are free of such artifactsDierckxsens et al. (2017); Hahn et al. (2013); Jin et al. (2020). At the Philippine Genome Center Visayas, several mitochondrial assemblies have failed or yielded incomplete contigs despite sufficient coverage, suggesting that undetected chimeric reads compromise assembly reliability. Meanwhile, existing chimera-detection tools such as UCHIME and VSEARCH were developed primarily for amplicon-based microbial community analysis and rely heavily on reference or taxonomic comparisonsEdgar, Haas, Clemente, Quince, and Knight (2011); Rognes, Flouri, Nichols, Quince, and Mahé (2016). These approaches are unsuitable for single-species organellar data, where complete reference genomes are often unavailable. Therefore, there is a pressing need for a reference-independent, data-driven tool capable of automatically detecting and filtering PCR-induced chimeras in mitochondrial sequencing datasets.

0.1.3 Research Objectives

General Objective

To develop and evaluate a machine-learning-based pipeline (MitoChime) capable of detecting PCR-induced chimeric reads in *Sardinella* mitochondrial sequencing data to improve the accuracy of mitochondrial genome assembly.

Specific Objectives

Specifically, the researchers aim to:

1. Construct simulated and empirical *Sardinella* Illumina paired-end datasets containing both clean and PCR-induced chimeric reads.

2. Extract alignment- and sequence-based features (e.g., k-mer composition, junction complexity, split-alignment counts) from both clean and chimeric reads.

3. Train, validate, and compare supervised machine-learning models (e.g., Random Forest, XGBoost) for classifying reads as clean or chimeric.

4. Determine feature importance and identify the most informative indicators of PCR-induced chimerism.

5. Integrate the optimized classifier into a modular and interpretable pipeline deployable on standard computing environments at PGC Visayas.

0.1.4 Scope and Limitations of the Research

This study focuses on detecting PCR-induced chimeric reads in Illumina paired-end mitochondrial sequencing data from *Sardinella* species. The work emphasizes `wgsim` simulations and selected empirical data obtained from open-access genomic repositories such as the National Center for Biotechnology Information (NCBI). The study excludes naturally occurring chimeras, nuclear mitochondrial pseudogenes (NUMTs), and large-scale structural rearrangements in nuclear genomes. Feature extraction prioritizes interpretable, shallow statistics and alignment metrics rather than deep-learning embeddings to ensure transparency and computational efficiency. Testing on long-read platforms (e.g., Nanopore, PacBio) and other taxa lies beyond the project's scope. The resulting pipeline will serve as a

foundation for future, broader chimera-detection frameworks applicable to diverse organellar genomes.

0.1.5 Significance of the Research

This research provides both methodological and practical contributions to mitochondrial genomics and bioinformatics. First, MitoChime enhances assembly accuracy by filtering PCR-induced chimeras prior to genome assembly, thereby improving the contiguity and correctness of *Sardinella* mitochondrial genomes. Second, it promotes automation and reproducibility by replacing subjective manual curation with a data-driven, machine-learning-based workflow. Third, the pipeline demonstrates computational efficiency through its design, enabling implementation on modest computing infrastructures commonly available in regional laboratories. Beyond technical improvements, MitoChime contributes to local capacity building by strengthening expertise in bioinformatics and machine-learning integration, aligning with the mission of the Philippine Genome Center Visayas. Finally, accurate mitochondrial assemblies are vital for fisheries management, population genetics, and biodiversity conservation, providing reliable genomic resources for species such as *Sardinella*. Through these contributions, MitoChime advances the reliability of mitochondrial genome reconstruction and supports sustainable, data-driven research in Philippine genomics.

References

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., ... Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. doi: 10.1038/nature07517

Boore, J. L. (1999). Animal mitochondrial genomes. *Nucleic Acids Research*, *27*(8), 1767–1780. doi: 10.1093/nar/27.8.1767

Cameron, S. L. (2014). Insect mitochondrial genomics: Implications for evolution and phylogeny. *Annual Review of Entomology*, *59*, 95–117. doi: 10.1146/annurev-ento-011613-162007

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). Novoplasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, *45*(4), e18. doi: 10.1093/nar/gkw955

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*(16), 2194–2200. doi: 10.1093/bioinformatics/btr381

Glenn, T. C. (2011). Field guide to next-generation dna sequencers. *Molecular Ecology Resources*, *11*(5), 759–769. doi: 10.1111/j.1755-0998.2011.03024.x

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial

genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research*, *41*(13), e129. doi: 10.1093/nar/gkt371

Jin, J.-J., Yu, W.-B., Yang, J., Song, Y., dePamphilis, C. W., Yi, T.-S., & Li, D.-Z. (2020). Getorganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*(1), 241. doi: 10.1186/s13059-020-02154-5

Judo, M. S. B., Wedel, W. R., & Wilson, B. H. (1998). Stimulation and suppression of pcr-mediated recombination. *Nucleic Acids Research*, *26*(7), 1819–1825. doi: 10.1093/nar/26.7.1819

Labrador, K., Agmata, A., Palermo, J. D., Ravago-Gotanco, R., & Pante, M. J. (2021). Mitochondrial dna reveals genetically structured haplogroups of bali sardinella (sardinella lemuru) in philippine waters. *Regional Studies in Marine Science*, *41*, 101588. doi: 10.1016/j.rsma.2020.101588

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. doi: 10.1038/nrg2626

Qiu, X., Wu, L., Huang, H., McDonel, P. E., Palumbo, A. V., Tiedje, J. M., & Zhou, J. (2001). Evaluation of pcr-generated chimeras, mutations, and heteroduplexes with 16s rrna gene-based cloning. *Applied and Environmental Microbiology*, *67*(2), 880–887. doi: 10.1128/AEM.67.2.880-887.2001

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). Vsearch: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. doi: 10.7717/peerj.2584

Willette, D., Bognot, E., Mutia, M. T., & Santos, M. (2011). *Biology and ecology of sardines in the philippines: A review* (Vol. 13; Tech. Rep. No. 1). NFRDI Technical Paper Series. Retrieved from `https://nfrdi.da.gov.ph/tpjf/`

etc/Willette%20et%20al.%20Sardines%20Review.pdf