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

Duranne Duran
Yvonne Lin
Daniella Pailden

Adviser
Francis Dimzon

November 25, 2025

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The rapid advancement of next-generation sequencing (NGS) technologies has transformed genomic research by enabling high-throughput and cost-effective DNA analysis (Metzker, 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 variation (Bentley 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 genome (Judo, Wedel, & Wilson, 1998).

PCR chimeras form when incomplete extension products from one template anneal to an unrelated DNA fragment and are extended, creating recombinant reads (Qiu et al., 2001). In mitochondrial genome assembly, such artifacts are especially problematic because the mitochondrial genome is small, circular, and often repetitive (Boore, 1999; Cameron, 2014). Even a small number of chimeric or mis-joined reads can reduce assembly contiguity and introduce false junctions during organelle genome reconstruction (Dierckxsens, Mardulyn, & Smits, 2017; Hahn, Bachmann, & Chevreux, 2013; Jin et al., 2020). Existing assembly tools such as GetOrganelle and MITObim assume that input reads are largely free of such artifacts (Hahn et al., 2013; Jin et al., 2020). Consequently, undetected 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 communities (Labrador, Agmata, Palermo, Ravago-Gotanco, & Pante, 2021; Willette, Bognot, Mutia, & 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.

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 mitochondrial genomes (Boore, 1999; Cameron, 2014). Existing assembly pipelines such as GetOrganelle, MITObim, and NOVOPlasty assume that sequencing reads are free of such artifacts (Dierckxsens 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 comparisons (Edgar, Haas, Clemente, Quince, & Knight, 2011; Rognes, Flouri, Nichols, Quince, & 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.

1.3 Research Objectives

1.3.1 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.

1.3.2 Specific Objectives

Specifically, the researchers aim to:

1. Construct empirical as well as simulated *Sardinella* Illumina paired-end datasets containing both clean and PCR-induced chimeric reads.
2. Extract alignment and sequence-based features such as k-mer composition, junction complexity, split-alignment counts from both clean and chimeric reads.
3. Train, validate, and compare supervised machine-learning models 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.

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 the *Sardinella lemuru* species. The decision to limit the taxonomic scope is motivated by three factors: (1) To provide a biologically coherent system by eliminating interspecific variation, such as differences in mitochondrial genome size, GC content, and repetitive regions. Restricting the analysis to *S. lemuru* reduces biological noise and ensures that observed patterns reflect chimeric artifacts rather than taxonomic differences. (2) The selected species is directly relevant to ongoing research initiatives and sequencing efforts at the Philippine Genome Center Visayas, making it a strategically appropriate choice for developing and validating the analytical framework; and (3) *Sardinella lemuru* possesses a moderately complex but well-characterized mitochondrial genome, with clear gene boundaries and sufficient publicly available data from repositories such as the National Center for Biotechnology Information (NCBI).

The study emphasizes `wgsim`-based simulations and selected empirical mitochondrial datasets. It 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.

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.

Chapter 2

Review of Related Literature

This chapter presents an overview of the literature relevant to the study. It discusses the biological and computational foundations underlying mitochondrial genome analysis and assembly, as well as existing tools, algorithms, and techniques related to chimera detection and genome quality assessment. The chapter aims to highlight the strengths, limitations, and research gaps in current approaches that motivate the development of the present study.

2.1 The Mitochondrial Genome

Mitochondrial genome (mtDNA) is a small, typically circular molecule found in most eukaryotes. It encodes essential genes involved in oxidative phosphorylation and energy metabolism. Because of its conserved structure and maternal inheritance, mtDNA has become a valuable genetic marker for studies in evolution, population genetics, and phylogenetics (Anderson et al., 1981; Boore, 1999). In animal species, the mitochondrial genome ranges from 15–20 kilobase and contains 13 protein-coding genes, 22 tRNAs, and two rRNAs arranged compactly without introns (Gray, 2012). In comparison to nuclear DNA the ratio of the number of copies of mtDNA is higher and has relatively simple organization which make it particularly suitable for genome sequencing and assembly studies (Dierckxsens et al., 2017). Moreover, mitochondrial genomes provide crucial insights into evolutionary relationships among species and are increasingly used for testing new genomic assembly and analysis methods.

2.1.1 Mitochondrial Genome Assembly

Mitochondrial genome assembly refers to the reconstruction of the complete mitochondrial DNA (mtDNA) sequence from raw or fragmented sequencing reads. It is conducted to obtain high-quality, continuous representations of the mitochondrial genome that can be used for a wide range of analyses, including species identification, phylogenetic reconstruction, evolutionary studies, and investigations of mitochondrial diseases. Because mtDNA evolves relatively rapidly and is maternally inherited, its assembled sequence provides valuable insights into population structure, lineage divergence, and adaptive evolution across taxa (Boore, 1999). Compared to nuclear genome assembly, assembling the mitochondrial genome is often considered more straightforward but still encounters distinct technical challenges such as sequencing errors, low coverage regions, and chimeric reads that can distort the final assembly, leading to incomplete or misassembled genomes. These errors can propagate into downstream analyses, emphasizing the need for robust chimera detection and sequence validation methods in mitochondrial genome research.

2.2 PCR Amplification and Chimera Formation

Polymerase Chain Reaction (PCR) plays an important role in next-generation sequencing (NGS) library preparation, as it amplifies target DNA fragments for downstream analysis. However, the amplification process can also introduce artifacts that affect data accuracy, one of them being the formation of chimeric sequences. Chimeras typically arise when incomplete extension occurs during a PCR cycle. This causes the DNA polymerase to switch from one template to another and generate hybrid recombinant molecules (Judo et al., 1998). Artificial chimeras are produced through such amplification errors, whereas biological chimeras occur naturally through genomic rearrangements or transcriptional events. These biological chimeras can have functional roles and may encode tissue-specific novel proteins that link to cellular processes or diseases (Frenkel-Morgenstern et al., 2012).

In the context of amplicon-based sequencing, PCR-induced chimeras can significantly distort analytical outcomes. Their presence artificially inflates estimates of genetic or microbial diversity and may cause misassemblies during genome reconstruction. (Qin et al., 2023) has reported that chimeric sequences may account for more than 10% of raw reads in amplicon datasets. This artifact tends to be most prominent among rare operational taxonomic units (OTUs) or singletons, which are sometimes misinterpreted as novel diversity, which further causes the complication of microbial diversity analyses (Gonzalez, Zimmermann, & Saiz-

Jimenez, 2004). Moreover, the likelihood of chimera formation has been found to vary with the GC content of target sequences, with lower GC content generally associated with a reduced rate of chimera generation (Qin et al., 2023).

2.2.1 Effects of Chimeric Reads on Organelle Genome Assembly

In mitochondrial DNA (mtDNA) assembly workflows, PCR-induced chimeras pose additional challenges. Assembly tools such as GetOrganelle and MitoBeam, which operate under the assumption of organelle genome circularity, are vulnerable when chimeric reads disrupt this circular structure. Such disruptions can lead to assembly errors or misassemblies (Bi et al., 2024). These artificial sequences interfere with the assembly graph, which makes it more difficult to accurately reconstruct mitochondrial genomes. In addition, these artifacts propagate false variants and erroneous annotations in genomic data. Hence, determining and minimizing PCR-induced chimera formation is vital for improving the quality of mitochondrial genome assemblies, and ensuring the reliability of amplicon sequencing data.

2.3 Existing Traditional Approaches for Chimera Detection

Several computational tools have been developed to identify chimeric sequences in NGS datasets. These tools generally fall into two categories: reference-based and de novo approaches. Reference-based chimera detection, also known as database-dependent detection, is one of the earliest and most widely used computational strategies for identifying chimeric sequences in amplicon-based microbial community studies. These methods rely on the comparison of each query sequence against a curated, high-quality database of known, non-chimeric reference sequences to determine whether the query can be more plausibly explained as a composite or a mosaic of two or more reference sequences rather than as a genuine biological variant (Edgar et al., 2011).

On the other hand, the De novo chimera detection, also referred to as reference-free detection, represents an alternative computational paradigm that identifies chimeric sequences without reliance on external reference databases. Instead of comparing each query sequence to a curated collection of known, non-chimeric sequences, de novo methods infer chimeras based on internal relationships among the sequences present within the dataset itself. This approach is particularly advantageous in studies of novel, under explored, or taxonomically diverse microbial communities where comprehensive reference databases are unavailable or incomplete (Edgar, 2016; Edgar et al., 2011). The underlying assumption on this method operates on the key biological principle that true biological sequences are generally more abundant than chimeric artifacts. During PCR amplification, authentic sequences are amplified early and tend to dominate the read pool, while chimeric sequences form later resulting in the tendency to appear at lower relative abundances compared to their true parental sequences. As such, the abundance hierarchy is formed by treating the most abundant sequences as supposed parents and testing whether less abundant sequences can be reconstructed as mosaics of these dominant templates. In addition to abundance, de novo algorithms assess compositional and structural similarity among sequences, examining whether certain regions of a candidate sequence align more closely with one high-abundance sequence and other regions with a different one.

Both reference-based and de novo approaches are complementary rather than mutually exclusive. Reference-based methods provide stability and reproducibility when curated databases are available, whereas de novo methods offer flexibility and independence for novel or highly diverse communities. In practice, many modern bioinformatics pipelines combine both paradigms sequentially: an initial de novo step identifies dataset-specific chimeras, followed by a reference-based pass that removes remaining artifacts relative to established databases (Edgar, 2016). These two methods of detection form the foundation of tools such as UCHIME and later UCHIME2, exemplified by the dual capability of providing both modes within a unified computational framework.

2.3.1 UCHIME

Developed by Edgar et al. (Edgar et al., 2011), UCHIME is one of the most widely used computational tools for detecting chimeric sequences in amplicon sequencing data. The UCHIME algorithm detects chimeras by evaluating how well a query sequence (Q) can be explained as a mosaic of two parent sequences (A and B) from a reference database. The query sequence is first divided into four non-overlapping segments or chunks. Each chunk is independently searched against a reference database that is assumed to be free of chimeras. The best matches to each segment are collected, and from these results, two candidate parent sequences are identified, typically the two sequences that best explain all chunks of the query. Then a three-way alignment among the query (Q) and the two parent candidates (A and B) is done. From this alignment, UCHIME attempts to find a chimeric model (M) which is a hypothetical recombinant sequence formed by concatenating fragments from A and B that best match the observed Q

Chimeric Alignment and Scoring

To decide whether a query is chimeric, UCHIME computes several alignment-based metrics between Q, its top hit (T, the most similar known sequence), and the chimeric model (M). The key differences are measured as: dQT or the number of mismatches between the query and the top hit as well as dQM or the number of mismatches between the query and the chimeric model. From these, a chimera score is calculated to quantify how much better the chimeric model fits the query compared to a single parent. If the model's similarity to Q exceeds a defined threshold (typically $\geq$0.8% better identity), the sequence is reported as chimeric. A higher score indicates stronger evidence of chimerism, while lower scores suggest that the sequence is more likely to be authentic.

In de novo mode, UCHIME applies an abundance-driven strategy. Only sequences at least twice as abundant as the query are considered as potential parents. Non-chimeric sequences identified at each step are added iteratively to a growing internal database for subsequent queries.

Limitations of UCHIME

Although UCHIME was a significant advancement in chimera detection, it has notable limitations. According to (Edgar, 2016) and the UCHIME practical notes (Edgar, n.d), many of the accuracy results reported in the original 2011 paper were overly optimistic due to unrealistic benchmark designs that assumed complete reference coverage and perfect sequence quality. In practice, UCHIME's accuracy can decline when: (1) The reference database is incomplete or contains erroneous entries. (2) Low-divergence chimeras are present, as these closely resemble genuine biological variants. (3) Sequence datasets include residual sequencing errors, leading to spurious alignments or misidentification; and (4) The abundance ratio between parent and chimera is distorted by amplification bias. Additionally, UCHIME tends to misclassify sequences as non-chimeric when parent sequences are missing from the database. These limitations motivated the development of UCHIME2.

2.3.2 UCHIME2

To overcome the limitations of its predecessor, UCHIME2 (Edgar, 2016) introduced several methodological and algorithmic refinements that significantly enhanced the accuracy and reliability of chimera detection. One major improvement lies in its approach to uncertainty handling. In earlier versions, sequences with limited reference support were often incorrectly classified as non-chimeric, increasing the likelihood of false negatives. UCHIME2 addresses this issue by designating such ambiguous sequences as "unknown," thereby providing a more conservative and reliable classification framework.

Another notable advancement is the introduction of multiple application-specific modes that allow users to tailor the algorithm's performance to the characteristics of their datasets. The following parameter presets: denoised, balanced, sensitive, specific, and high-confidence, enable researchers to optimize the balance between sensitivity and specificity according to the goals of their analysis.

In comparative evaluations, UCHIME2 demonstrated superior detection performance, achieving sensitivity levels between 93% and 99% and lower overall error rates than earlier versions or other contemporary tools such as DECIPHER and ChimeraSlayer. Despite these advances, the study also acknowledged a fundamental limitation in chimera detection: complete error-free identification is theoretically unattainable. This is due to the presence of "perfect fake models," wherein genuine non-chimeric sequences can be perfectly reconstructed from other reference fragments. This underscore the uncertainty in differentiating authentic biological sequences from artificial recombinants based solely on sequence similarity, emphasizing the need for continued methodological refinement and cautious interpretation of results.

2.3.3 CATch

Early chimera detection programs such as UCHIME (Edgar et al., 2011) relied on alignment-based and abundance-based heuristics to identify hybrid sequences in amplicon data. However, researchers soon observed that different algorithms often produced inconsistent predictions. A sequence might be identified as chimeric by one tool but classified as non-chimeric by another, resulting in unreliable filtering outcomes across studies.

To address these inconsistencies, (Mysara, Saeys, Leys, Raes, & Monsieurs, 2015) developed the Classifier for Amplicon Tool Chimeras (CATCh), which represents the first ensemble machine learning system designed for chimera detection in 16S rRNA amplicon sequencing. Rather than depending on a single detection strategy, CATCh integrates the outputs of several established tools, including UCHIME, ChimeraSlayer, DECIPHER, Pintail, and Perseus. The individual scores and binary decisions generated by these tools are used as input features for a supervised learning model. The algorithm employs a Support Vector Machine (SVM) with a Pearson VII Universal Kernel (PUK) to determine optimal weightings among the input features and to assign each sequence a probability of being chimeric.

Benchmarking in both reference-based and de novo modes demonstrated significant performance improvements. CATCh achieved sensitivities of approximately 85 percent in reference-based mode and 92 percent in de novo mode, with corresponding specificities of approximately 96 percent and 95 percent. These results indicate that CATCh detected 7 to 12 percent more chimeras than any individual algorithm while maintaining high precision. Integration of CATCh into amplicon-processing pipelines also reduced operational taxonomic unit (OTU) inflation by 23 to 35 percent, producing diversity estimates that more closely reflected true community composition.

2.3.4 ChimPipe

Among the available tools for chimera detection, ChimPipe is a bioinformatics pipeline developed to identify chimeric sequences such as fusion genes and transcription-induced chimeras from paired-end RNA sequencing data. It uses both discordant paired-end reads and split-read alignments to improve the accuracy and sensitivity of detecting fusion genes, trans-splicing events, and read-through transcripts (Rodriguez-Martin et al., 2017). By combining these two sources of information, ChimPipe achieves better precision than methods that depend on a single type of signal.

The pipeline works with many eukaryotic species that have available genome and annotation data, making it a versatile tool for studying chimera evolution and transcriptome structure (Rodriguez-Martin et al., 2017). It can also predict multiple isoforms for each gene pair and identify breakpoint coordinates that are useful for reconstructing and verifying chimeric transcripts. Tests using both simulated and real datasets have shown that ChimPipe maintains high accuracy and reliable performance.

ChimPipe's modular design lets users adjust parameters to fit different sequencing protocols or organism characteristics. Experimental results have confirmed that many chimeric transcripts detected by the tool correspond to functional fusion proteins, showing its value for understanding chimera biology and its potential applications in disease research (Rodriguez-Martin et al., 2017).

2.4 Machine Learning Approaches for Chimera and Sequence Quality Detection

Traditional chimera detection tools rely primarily on heuristic or alignment-based rules. Recent advances in machine learning (ML) have demonstrated that models trained on sequence-derived features can effectively capture compositional and structural patterns in biological sequences. Although most existing ML systems such as those used for antibiotic resistance prediction, taxonomic classification, or viral identification are not specifically designed for chimera detection, they highlight how data-driven models can outperform similarity-based heuristics by learning intrinsic sequence signatures. In principle, ML frameworks can integrate diverse indicators such as k-mer frequencies, GC-content variation, and split-alignment metrics to identify subtle anomalies that may indicate a chimeric origin (Arango et al., 2018; Liang, Bible, Liu, Zou, & Wei, 2020; Ren et al., 2020).

2.4.1 Feature-Based Representations of Genomic Sequences

In genomic analysis, feature extraction converts DNA sequences into numerical representations suitable for ML algorithms. A common approach is k-mer frequency analysis, where normalized k-mer counts form the feature vector (Vervier, 2015). These features effectively capture local compositional patterns that often differ between authentic and chimeric reads. In particular, deviations in k-mer profiles between adjacent read segments can serve as a compositional signature of template-switching events. Additional descriptors such as GC content and sequence entropy can further distinguish sequence types; in metagenomic classification and virus detection, k-mer-based features have shown strong performance and robustness to noise (Ren et al., 2020; Vervier, 2015). For chimera detection specifically, abrupt shifts in GC or k-mer composition along a read can indicate junctions between parental fragments. Windowed feature extraction enables models to capture these discontinuities that rule-based algorithms may overlook.

Machine learning models can also leverage alignment-derived features such as the frequency of split alignments, variation in mapping quality, and local coverage irregularities. Split reads and discordant read pairs are classical signatures of genomic junctions and have been formalized in probabilistic frameworks for structural-variant discovery that integrate multiple evidence types (Layer, Hall, & Quinlan, 2014). Similarly, long-read tools such as Sniffles employ split-alignment and coverage anomalies to accurately localize breakpoints (Sedlazeck et al., 2018). Modern aligners such as Minimap2 (Li, 2018) output supplementary (SA tags) and secondary alignments as well as chaining and alignment-score statistics that can be summarized into quantitative predictors for machine-learning models. These alignment-signal features are particularly relevant to PCR-induced mitochondrial chimeras, where template-switching events produce reads partially matching distinct regions of the same or related genomes. Integrating such cues within a supervised-learning framework enables artifact detection even in datasets lacking complete or perfectly assembled references.

A further biologically grounded descriptor is micro-homology length at putative junctions. Micro-homology refers to short, shared sequences (often in the range of a few to tens of base pairs) that are near breakpoints and mediate non-canonical repair or template-switch mechanisms. Studies of double strand break repair and structural variation have demonstrated that the length of micro-homology correlates with the likelihood of micro-homology-mediated end joining (MMEJ) or fork-stalled template-switching pathways (Sfeir & Symington, 2015). In the context of PCR-induced chimeras, template switching during amplification often leaves short identical sequences at the junction of two concatenated fragments. Quantifying the longest exact suffix–prefix overlap at each candidate breakpoint thus provides a mechanistic signature of chimerism and complements both compositional (k-mer) and alignment (SA count) features.

2.5 Synthesis of Chimera Detection Approaches

To provide an integrated overview of the literature discussed in this chapter, Table 2.1 summarizes the major chimera detection studies, their methodological approaches, and their known limitations. This consolidated comparison brings together reference-based approaches, de novo strategies, alignment-driven tools, ensemble machine-learning systems, and general ML-based sequence-quality frameworks. Presenting these methods side-by-side clarifies their performance boundaries and highlights the unresolved challenges that persist in mitochondrial genome analysis and chimera detection.

Table 2.1: Summary of Existing Methods and Research Gaps

Method/Study	**Scope/Approach**	**Limitations**
Reference-based Chimera Detection	Compares query sequences against curated, non-chimeric reference databases; identifies mosaic sequences by evaluating similarity to known templates.	Depends heavily on completeness and quality of reference databases; often fails when novel taxa or missing parent sequences are present; reduced accuracy for low-divergence chimeras.
De novo Chimera Detection	Identifies chimeras using only internal dataset relationships; relies on abundance patterns and compositional similarity; reconstructs sequences as mosaics of high-abundance parents.	Assumes true sequences are more abundant—fails when amplification bias distorts abundance; struggles with evenly abundant parental sequences; can misclassify highly similar true variants.

Method/Study	**Scope/Approach**	**Limitations**
UCHIME	Alignment-based chimera detection; segments query sequence, identifies parent candidates, performs 3-way alignment, and computes chimera scores; supports both reference-based and de novo modes.	Accuracy inflated in original benchmarks; suffers under incomplete databases; poor performance on low-divergence chimeras; sensitive to sequencing errors; misclassifies when parents are missing.
UCHIME2	Improved uncertainty handling; classifies ambiguous sequences as unknown; offers multiple sensitivity/specificity modes; more robust with incomplete references; higher sensitivity (93–99%).	Cannot achieve perfect accuracy due to “perfect fake models”; genuine variants may be indistinguishable from artificial recombinants; theoretical detection limit remains.
CATCh	First ML ensemble tool for 16S chimera detection; integrates outputs of UCHIME, ChimeraSlayer, DECIPHER, Pintail, Perseus via SVM classifier; significantly improves sensitivity and specificity.	Depends on performance of underlying tools; ML model limited to features they output; ensemble can still misclassify in datasets with extreme novelty or low coverage.

Method/Study	Scope/Approach	Limitations
ChimPipe	Pipeline for detecting fusion genes and transcript-derived chimeras in RNA-seq; uses discordant paired-end reads and split-alignments; predicts isoforms and breakpoint coordinates.	Designed for RNA-seq, not amplicons; needs high-quality genome and annotation; computationally heavier; limited to organisms with reference genomes.
Machine-Learning Sequence Quality & Chimera Detection (general)	Uses k-mer profiles, GC content shifts, entropy, split-read statistics, mapping quality variation, and micro-homology signatures as predictive features; identifies subtle artifacts missed by heuristics.	Requires labeled training data; model performance depends on feature engineering; may capture dataset-specific biases; limited generalization if training data is narrow or unrepresentative.

Across existing studies, no single approach reliably detects all forms of chimeric sequences, particularly those generated by PCR-induced template switching in mitochondrial genomes. Reference-based tools perform poorly when parental sequences are absent; de novo methods rely strongly on abundance assumptions; alignment-based systems show reduced sensitivity to low-divergence chimeras; and ensemble methods inherit the limitations of their component algorithms. RNA-seq–oriented pipelines likewise do not generalize well to organelle data. Although machine learning approaches offer promising feature-based detection, they are rarely applied to mitochondrial genomes and are not trained specifically on PCR-induced organelle chimeras. These limitations indicate a clear research gap: the need for a specialized, feature-driven classifier tailored to mitochondrial PCR-induced chimeras that integrates k-mer composition, split-alignment signals, and micro-homology features to achieve more accurate detection than current heuristic or alignment-based tools.

Chapter 3

Research Methodology

This chapter outlines and explains the specific steps and activities to be carried out in completing the project.

3.1 Research Activities

As illustrated in Figure 3.1, the researchers will carry out a sequence of computational procedures designed to detect PCR-induced chimeric reads in mitochondrial genomes. The process begins with the collection of mitochondrial reference sequences from the NCBI database, which will serve as the foundation for generating simulated chimeric reads. These datasets will then undergo bioinformatics pipeline development, which includes alignment, k-mer extraction, and homology-based filtering to prepare the data for model construction. The machine-learning model will subsequently be trained and tested using the processed datasets to assess its accuracy and reliability. Depending on the evaluation results, the model will either be refined and retrained to improve performance or, if the metrics meet the desired threshold, deployed for further validation and application.

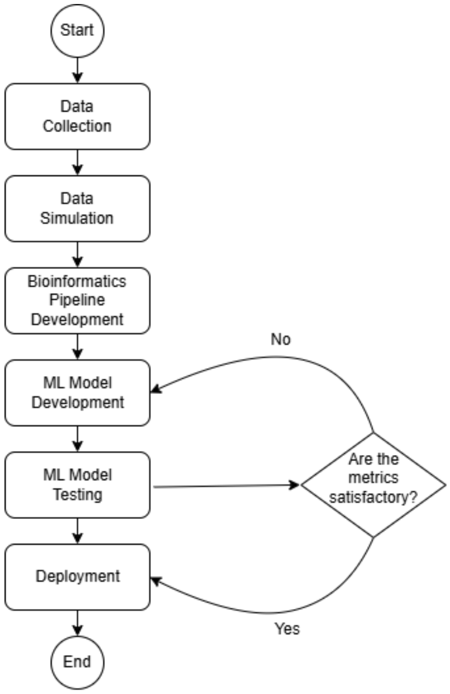

Figure 3.1: Process Diagram of Special Project

3.1.1 Data Collection

The researchers will collect mitochondrial genome reference sequences of *Sardinella lemuru* from the National Center for Biotechnology Information (NCBI) database. The downloaded files will be in FASTA format to ensure compatibility with bioinformatics tools and subsequent analysis. The gathered sequences will serve as the basis for generating simulated chimeric reads to be used in model development.

The expected outcome of this process is a comprehensive dataset of *Sardinella* *lemuru* mitochondrial reference sequences that will serve as the foundation for the succeeding stages of the study. This step is scheduled to start in the first week of November 2025 and is expected to be completed by the last week of November 2025, with a total duration of approximately one (1) month.

3.1.2 Data Simulation

The researchers will simulate sequencing data using the reference sequences collected from NCBI. Using `wgsim`, a total of 5,000 paired-end reads (R1 and R2) will be generated from the reference genome and designated as clean reads. These reads will be saved in FASTQ (`.fastq`) format. From the same reference, a Bash script will be created to deliberately cut and reconnect portions of the sequence, introducing artificial junctions that mimic chimeric regions. The manipulated reference file, saved in FASTA (`.fasta`) format, will then be processed in `wgsim` to simulate an additional 5,000 paired-end chimeric reads, also stored in FASTQ (`.fastq`) format. The resulting read files will be aligned to the original reference genome using SAMtools, generating SAM (`.sam`) or BAM (`.bam`) alignment files. During this alignment process, clean reads will be labeled as “0,” while chimeric reads will be labeled as “1” in a corresponding CSV (`.tsv`) file.

The expected outcome of this process is a complete set of clean and chimeric paired-end reads prepared for subsequent analysis and model development. This step is scheduled to start in the first week of November 2025 and is expected to be completed by the last week of November 2025, with a total duration of approximately one (1) month.

3.1.3 Bioinformatics Tools Pipeline

The researchers will obtain the necessary analytical features through the development and implementation of a bioinformatics pipeline. This pipeline will serve as a reproducible and modular workflow that accepts FASTQ and BAM inputs, processes these through a series of analytical stages, and outputs tabular feature matrices (TSV) for downstream machine learning. All scripts will be version-controlled through GitHub, and computational environments will be standardized using Conda to ensure cross-platform reproducibility. To promote transparency and replicability, the exact software versions, parameters, and command-line arguments used in each stage will be documented. To further ensure correctness and adherence to best practices, the researchers will consult with bioinformatics experts in Philippine Genome Center Visayas for validation of pipeline design, feature extraction logic, and overall data integrity. This stage of the study is scheduled to begin in the last week of November 2025 and conclude by the last week of January 2026, with an estimated total duration of approximately two (2) months.

The bioinformatics pipeline focuses on three principal features from the simulated and aligned sequencing data: (1) supplementary alignment count (SA count), (2) k-mer composition difference between read segments, and (3) microhomology length at potential junctions. Each of these features captures a distinct biological or computational signature associated with PCR-induced chimeras.

Alignment and Supplementary Alignment Count

This will be derived through sequence alignment using Minimap2, with subsequent processing performed using SAMtools and `pysam` in Python. Sequencing reads will be aligned to the *Sardinella lemuru* mitochondrial reference genome using Minimap2 with the `-ax sr` preset (optimized for short reads). The output will be converted and sorted using SAMtools, producing an indexed BAM file which will be parsed using `pysam` to count the number of supplementary alignments (SA tags) per read. Each read's mapping quality, number of split segments, and alignment characteristics will be recorded in a corresponding TSV file. The presence of multiple alignment loci within a single read, as reflected by a nonzero SA count, serves as direct computational evidence of chimerism. Reads that contain supplementary alignments or soft-clipped regions are strong candidates for chimeric artifacts arising from PCR template switching or improper assembly during sequencing.

K-mer Composition Difference

Chimeric reads often comprise fragments from distinct genomic regions, resulting in a compositional discontinuity between segments. Comparing k-mer frequency profiles between the left and right halves of a read allows detection of such abrupt compositional shifts, independent of alignment information. This will be obtained using Jellyfish, a fast k-mer counting software. For each read, the sequence will be divided into two segments, either at the midpoint or at empirically determined breakpoints inferred from supplementary alignment data, to generate left and right sequence segments. Jellyfish will then compute k-mer frequency profiles (with $k =$

5 or 6) for each segment. The resulting k-mer frequency vectors will be normalized and compared using distance metrics such as cosine similarity or Jensen–Shannon divergence to quantify compositional disparity between the two halves of the same read. The resulting difference scores will be stored in a structured TSV file.

Micro-homology Length

The micro-homology length will be computed using a custom Python script that detects the longest exact suffix–prefix overlap within ±30 base pairs surrounding a candidate breakpoint. This analysis identifies the number of consecutive bases shared between the end of one segment and the beginning of another. The presence and length of such micro-homology are classic molecular signatures of PCR-induced template switching, where short identical regions (typically 3–15 base pairs) promote premature termination and recombination of DNA synthesis on a different template strand. By quantifying micro-homology, the researchers can assess whether the suspected breakpoint exhibits characteristics consistent with PCR artifacts rather than true biological variants. Each read will therefore be annotated with its corresponding micro-homology length, overlap sequence, and GC content.

After extracting the three primary features, all resulting TSV files will be joined using the read identifier as a common key to generate a unified feature matrix. Additional read-level metadata such as read length, mean base quality, and number of clipped bases will also be included to provide contextual information. This consolidated dataset will serve as the input for subsequent machine-learning model development and evaluation.

3.1.4 Machine-Learning Model Development

The classification component of MitoChime will employ two ensemble algorithms—Random Forest (RF) and Extreme Gradient Boosting (XGBoost)—to evaluate complementary learning paradigms. Random Forest applies bootstrap aggregation (bagging) to reduce model variance and improve stability, whereas XGBoost implements gradient boosting to minimize bias and capture complex non-linear relationships among genomic features. Using both models enables a balanced assessment of predictive performance and interpretability.

The dataset will be divided into training (80%) and testing (20%) subsets. The training data will be used for model fitting and hyperparameter optimization through five-fold cross-validation, in which the data are partitioned into five folds; four folds are used for training and one for validation in each iteration. Performance metrics will be averaged across folds, and the optimal parameters will be selected based on mean cross-validation accuracy. The final models will then be evaluated on the held-out test set to obtain unbiased performance estimates.

Model development and evaluation will be implemented in Python (version 3.11) using the `scikit-learn` and `xgboost` libraries. Standard metrics including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) will be computed to quantify predictive performance. Feature-importance analyses will be performed to identify the most discriminative variables contributing to chimera detection.

3.1.5 Validation and Testing

Validation will involve both internal and external evaluations. Internal validation will be achieved through five-fold cross-validation on the training data to verify model generalization and reduce variance due to random sampling. External validation will be achieved through testing on the 20% hold-out dataset derived from the simulated reads, which will serve as an unbiased benchmark to evaluate how well the trained models generalize to unseen data. All feature extraction and preprocessing steps will be performed using the same bioinformatics pipeline to ensure consistency and comparability across validation stages.

Comparative evaluation between the Random Forest and XGBoost classifiers will establish which model achieves superior predictive accuracy and computational efficiency under identical data conditions.

3.1.6 Documentation

Comprehensive documentation will be maintained throughout the study to ensure transparency, reproducibility, and scientific integrity. All stages of the research—including data acquisition, preprocessing, feature extraction, model training, and validation—will be systematically recorded. For each analytical step, the corresponding parameters, software versions, and command-line scripts will be documented to enable exact replication of results.

Version control and collaborative management will be implemented through GitHub, which will serve as the central repository for all project files, including Python scripts, configuration settings, and Jupyter notebooks. The repository structure will follow standard research data management practices, with clear directories for datasets, processed outputs, and analysis scripts. Changes will be tracked through commit histories to ensure traceability and accountability.

Computational environments will be standardized using Conda, with environment files specifying dependencies and package versions to maintain consistency across systems. Experimental workflows and exploratory analyses will be conducted in Jupyter Notebooks, which facilitate real-time visualization, annotation, and incremental testing of results.

For the preparation of the final manuscript and supplementary materials, Overleaf (LaTeX) will be utilized to produce publication-quality formatting, consistent referencing, and reproducible document compilation. The documentation process will also include a project timeline outlining major milestones such as data collection, simulation, feature extraction, model evaluation, and reporting to ensure systematic progress and adherence to the research schedule.

3.2 Calendar of Activities

Table 3.1 presents the project timeline in the form of a Gantt chart, where each bullet point corresponds to approximately one week of planned activity.

Table 3.1: Timetable of Activities

Activities (2025)	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**
Data Collection and Simulation	••••						
Bioinformatics Tools Pipeline	••	••••	••••				
Machine Learning Development			••	••••	••••	••	
Testing and Validation						••	••••
Documentation	••••	••••	••••	••••	••••	••••	••••

References

Anderson, S., Bankier, A., Barrell, B., Bruijn, M., Coulson, A., Drouin, J., … Young, I. (1981, 04). Sequence and organization of the human mitochondrial genome. *Nature*, *290*, 457-465. doi: 10.1038/290457a0

Arango, G., Garner, E., Pruden, A., Heath, L., Vikesland, P., & Zhang, L. (2018, 02). Deeparg: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, *6*. doi: 10.1186/s40168-018-0401-z

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

Bi, C., Shen, F., Han, F., Qu, Y., Hou, J., Xu, K., … Yin, T. (2024, 01). Pmat: an efficient plant mitogenome assembly toolkit using low-coverage hifi sequencing data. *Horticulture Research*, *11*(3), uhae023. Retrieved from `https://doi.org/10.1093/hr/uhae023` doi: 10.1093/hr/uhae023

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. (2016). Uchime2: improved chimera prediction for amplicon sequencing. *bioRxiv*. Retrieved from `https://api.semanticscholar.org/CorpusID:88955007`

Edgar, R. C. (n.d). Uchime in practice. Retrieved from `https://www.drive5.com/usearch/manual7/uchime_practical.html`

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

Frenkel-Morgenstern, M., Lacroix, V., Ezkurdia, I., Levin, Y., Gabashvili, A., Prilusky, J., . . . Valencia, A. (2012, 05). Chimeras taking shape: Potential functions of proteins encoded by chimeric rna transcripts. *Genome research*, *22*, 1231-42. doi: 10.1101/gr.130062.111

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

Gonzalez, J. M., Zimmermann, J., & Saiz-Jimenez, C. (2004, 09). Evaluating putative chimeric sequences from pcr-amplified products. *Bioinformatics*, *21*(3), 333-337. Retrieved from `https://doi.org/10.1093/bioinformatics/bti008` doi: 10.1093/bioinformatics/bti008

Gray, M. W. (2012). Mitochondrial evolution. *Cold Spring Harbor perspectives in biology*, *4*. Retrieved from `https://doi.org/10.1101/cshperspect.a011403` doi: 10.1101/cshperspect.a011403

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

Layer, R., Hall, I., & Quinlan, A. (2014, 10). Lumpy: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*. doi: 10.1186/gb-2014-15-6-r84

Li, H. (2018, 05). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094-3100. Retrieved from `https://doi.org/10.1093/bioinformatics/bty191` doi: 10.1093/bioinformatics/bty191

Liang, Q., Bible, P. W., Liu, Y., Zou, B., & Wei, L. (2020, 02). Deepmicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genomics and Bioinformatics*, *2*(1), lqaa009. Retrieved from `https://doi.org/10.1093/nargab/lqaa009` doi: 10.1093/nargab/lqaa009

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature*

Reviews Genetics, *11*(1), 31–46. doi: 10.1038/nrg2626

Mysara, M., Saeys, Y., Leys, N., Raes, J., & Monsieurs, P. (2015). Catch, an ensemble classifier for chimera detection in 16s rrna sequencing studies. *Applied and Environmental Microbiology*, *81*(5), 1573-1584. Retrieved from `https://journals.asm.org/doi/abs/10.1128/aem.02896-14` doi: 10.1128/AEM.02896-14

Qin, Y., Wu, L., Zhang, Q., Wen, C., Nostrand, J. D. V., Ning, D., . . . Zhou, J. (2023). Effects of error, chimera, bias, and gc content on the accuracy of amplicon sequencing. *mSystems*, *8*(6), e01025-23. Retrieved from `https://journals.asm.org/doi/abs/10.1128/msystems.01025-23` doi: 10.1128/msystems.01025-23

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

Ren, J., Song, K., Deng, C., Ahlgren, N., Fuhrman, J., Li, Y., . . . Sun, F. (2020, 01). Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*, *8*. doi: 10.1007/s40484-019-0187-4

Rodriguez-Martin, B., Palumbo, E., Marco-Sola, S., Griebel, T., Ribeca, P., Alonso, G., . . . Djebali, S. (2017, 01). Chimpipe: Accurate detection of fusion genes and transcription-induced chimeras from rna-seq data. *BMC Genomics*, *18*. doi: 10.1186/s12864-016-3404-9

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

Sedlazeck, F., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. (2018, 06). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*. doi: 10.1038/s41592-018-0001-7

Sfeir, A., & Symington, L. S. (2015). Microhomology-mediated end joining: A back-up survival mechanism or dedicated pathway? *Trends in Biochemical Sciences*, *40*(11), 701-714. Retrieved from https://www.sciencedirect.com/science/article/pii/S0968000415001589 doi: https://doi.org/10.1016/j.tibs.2015.08.006

Vervier, M. P. T. M. V. J. B. . V. J. P., K. (2015). Large-scale machine learning for metagenomics sequence classification. *Bioinformatics*, *32*, 1023 - 1032. Retrieved from https://api.semanticscholar.org/CorpusID:9863600

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