

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

4 A Special Project Proposal
5 Presented to
6 the Faculty of the Division of Physical Sciences and Mathematics
7 College of Arts and Sciences
8 University of the Philippines Visayas
9 Miag-ao, Iloilo

10 In Partial Fulfillment
11 of the Requirements for the Degree of
12 Bachelor of Science in Computer Science

13 by

14 Duranne Duran
15 Yvonne Lin
16 Daniella Pailden

17 Adviser
18 Francis D. Dimzon, Ph.D.

19 December 1, 2025

Contents

21	1 Introduction	1
22	1.1 Overview	1
23	1.2 Problem Statement	3
24	1.3 Research Objectives	4
25	1.3.1 General Objective	4
26	1.3.2 Specific Objectives	4
27	1.4 Scope and Limitations of the Research	5
28	1.5 Significance of the Research	6
29	2 Review of Related Literature	7
30	2.1 The Mitochondrial Genome	7
31	2.1.1 Mitochondrial Genome Assembly	8

32	2.2	PCR Amplification and Chimera Formation	9
33	2.2.1	Effects of Chimeric Reads on Organelle Genome Assembly	10
34	2.3	Existing Traditional Approaches for Chimera Detection	11
35	2.3.1	UCHIME	12
36	2.3.2	UCHIME2	14
37	2.3.3	CATch	16
38	2.3.4	ChimPipe	17
39	2.4	Machine Learning Approaches for Chimera and Sequence Quality	
40		Detection	18
41	2.4.1	Feature-Based Representations of Genomic Sequences . . .	18
42	2.5	Synthesis of Chimera Detection Approaches	20
43	3	Research Methodology	23
44	3.1	Research Activities	23
45	3.1.1	Data Collection	24
46	3.1.2	Bioinformatics Tools Pipeline	28
47	3.1.3	Machine Learning Model Development	31
48	3.1.4	Validation and Testing	32

49	3.1.5 Documentation	32
50	3.2 Calendar of Activities	33

51 List of Figures

<small>52</small>	3.1 Process Diagram of Special Project	24
-------------------	--	----

53 List of Tables

<small>54</small>	2.1 Summary of Existing Methods and Research Gaps	21
<small>55</small>	3.1 Timetable of Activities	33

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 misjoined 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 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-based and sequence-derived statistical features. The tool aims to provide bioinformatics laboratories, partic-

ularly the Philippine Genome Center Visayas (PGC Visayas), with an efficient 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 particularly 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 PGC 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 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 detecting and filtering PCR-induced chimeras in mitochondrial sequencing datasets.

115 1.3 Research Objectives

116 1.3.1 General Objective

117 This study aims to develop and evaluate a machine learning-based pipeline (Mi-
118 toChime) that detects PCR-induced chimeric reads in *Sardinella lemuru* mito-
119 chondrial sequencing data in order to improve the quality and reliability of down-
120 stream mitochondrial genome assemblies.

121 1.3.2 Specific Objectives

122 Specifically, the study aims to:

- 123 1. construct simulated *Sardinella lemuru* Illumina paired-end datasets contain-
124 ing both clean and PCR-induced chimeric reads,
- 125 2. extract alignment-based and sequence-based features such as k-mer compo-
126 sition, junction complexity, and split-alignment counts from both clean and
127 chimeric reads,
- 128 3. train, validate, and compare supervised machine-learning models for classi-
129 fying reads as clean or chimeric,
- 130 4. determine feature importance and identify indicators of PCR-induced
131 chimerism,
- 132 5. integrate the optimized classifier into a modular and interpretable pipeline
133 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 *Sardinella lemuru*. The decision to restrict the taxonomic scope to a single species is based on four considerations: (1) to limit interspecific variation in mitochondrial genome size, GC content, and repetitive regions so that differences in read patterns can be attributed more directly to PCR-induced chimerism; (2) to align the analysis with relevant *S. lemuru* sequencing projects at PGC Visayas; (3) to take advantage of the availability of *S. lemuru* mitochondrial assemblies and raw datasets in public repositories such as the National Center for Biotechnology Information (NCBI), which facilitates reference selection and benchmarking; and (4) to develop a tool that directly supports local studies on *S. lemuru* population structure and fisheries management.

The study emphasizes `wgsim`-based simulations and selected empirical mitochondrial datasets from *S. lemuru*. It excludes naturally occurring chimeras, nuclear mitochondrial pseudogenes (NUMTs), and large-scale assembly rearrangements in nuclear genomes. Feature extraction is restricted to low-dimensional alignment and sequence statistics, such as k-mer frequency profiles, GC content, read length, soft and hard clipping metrics, split-alignment counts, and mapping quality, rather than high-dimensional deep learning embeddings. This design keeps model behaviour interpretable and ensures that the pipeline can be run on standard workstations at PGC Visayas. Testing on long-read platforms (e.g., Nanopore, PacBio) and other taxa is outside the scope of this project; the implemented pipeline is evaluated only on short-read *S. lemuru* datasets.

157 1.5 Significance of the Research

158 This research provides both methodological and practical contributions to mi-
159 tochondrial genomics and bioinformatics. First, MitoChime filters PCR-induced
160 chimeric reads prior to genome assembly, with the goal of improving the con-
161 tiguity and correctness of *Sardinella lemuru* mitochondrial assemblies. Second,
162 it replaces informal manual curation with a documented workflow, improving au-
163 tomation and reproducibility. Third, the pipeline is designed to run on computing
164 infrastructures commonly available in regional laboratories, enabling routine use
165 at facilities such as PGC Visayas. Finally, more reliable mitochondrial assemblies
166 for *S. lemuru* provide a stronger basis for downstream applications in the field of
167 fisheries and genomics.

168 Chapter 2

169 Review of Related Literature

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

176 2.1 The Mitochondrial Genome

177 Mitochondrial genome (mtDNA) is a small, typically circular molecule found in
178 most eukaryotes. It encodes essential genes involved in oxidative phosphorylation
179 and energy metabolism. Because of its conserved structure and maternal inher-
180 itance, mtDNA has become a valuable genetic marker for studies in evolution,
181 population genetics, and phylogenetics (Anderson et al., 1981; Boore, 1999). In

182 animal species, the mitochondrial genome ranges from 15–20 kilobase and contains
183 13 protein-coding genes, 22 tRNAs, and two rRNAs arranged compactly without
184 introns (Gray, 2012). In comparison to nuclear DNA the ratio of the number
185 of copies of mtDNA is higher and has relatively simple organization which make
186 it particularly suitable for genome sequencing and assembly studies (Dierckxsens
187 et al., 2017). Moreover, mitochondrial genomes provide crucial insights into evo-
188 lutionary relationships among species and are increasingly used for testing new
189 genomic assembly and analysis methods.

190 **2.1.1 Mitochondrial Genome Assembly**

191 Mitochondrial genome assembly refers to the reconstruction of the complete mito-
192 chondrial DNA (mtDNA) sequence from raw or fragmented sequencing reads. It is
193 conducted to obtain high-quality, continuous representations of the mitochondrial
194 genome that can be used for a wide range of analyses, including species identi-
195 fication, phylogenetic reconstruction, evolutionary studies, and investigations of
196 mitochondrial diseases. Because mtDNA evolves relatively rapidly and is mater-
197 nally inherited, its assembled sequence provides valuable insights into population
198 structure, lineage divergence, and adaptive evolution across taxa (Boore, 1999).
199 Compared to nuclear genome assembly, assembling the mitochondrial genome is
200 often considered more straightforward but still encounters distinct technical chal-
201 lenges such as sequencing errors, low coverage regions, and chimeric reads that can
202 distort the final assembly, leading to incomplete or misassembled genomes. These
203 errors can propagate into downstream analyses, emphasizing the need for robust
204 chimera detection and sequence validation methods in mitochondrial genome re-

205 search.

206 **2.2 PCR Amplification and Chimera Formation**

207 Polymerase Chain Reaction (PCR) plays an important role in next-generation
208 sequencing (NGS) library preparation, as it amplifies target DNA fragments for
209 downstream analysis. However, the amplification process can also introduce arti-
210 facts that affect data accuracy, one of them being the formation of chimeric se-
211 quences. Chimeras typically arise when incomplete extension occurs during a PCR
212 cycle. This causes the DNA polymerase to switch from one template to another
213 and generate hybrid recombinant molecules (Judo et al., 1998). Artificial chimeras
214 are produced through such amplification errors, whereas biological chimeras oc-
215 cur naturally through genomic rearrangements or transcriptional events. These
216 biological chimeras can have functional roles and may encode tissue-specific novel
217 proteins that link to cellular processes or diseases (Frenkel-Morgenstern et al.,
218 2012).

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

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

230 **2.2.1 Effects of Chimeric Reads on Organelle Genome As-** 231 **sembly**

232 In mitochondrial DNA (mtDNA) assembly workflows, PCR-induced chimeras pose
233 additional challenges. Assembly tools such as GetOrganelle and MitoBeam, which
234 operate under the assumption of organelle genome circularity, are vulnerable when
235 chimeric reads disrupt this circular structure. Such disruptions can lead to assem-
236 bly errors or misassemblies (Bi et al., 2024). These artificial sequences interfere
237 with the assembly graph, which makes it more difficult to accurately reconstruct
238 mitochondrial genomes. In addition, these artifacts propagate false variants and
239 erroneous annotations in genomic data. Hence, determining and minimizing PCR-
240 induced chimera formation is vital for improving the quality of mitochondrial
241 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

266 chimeric sequences form later resulting in the tendency to appear at lower relative
267 abundances compared to their true parental sequences. As such, the abundance
268 hierarchy is formed by treating the most abundant sequences as supposed parents
269 and testing whether less abundant sequences can be reconstructed as mosaics of
270 these dominant templates. In addition to abundance, de novo algorithms assess
271 compositional and structural similarity among sequences, examining whether cer-
272 tain regions of a candidate sequence align more closely with one high-abundance
273 sequence and other regions with a different one.

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

284 **2.3.1 UCHIME**

285 Developed by Edgar et al. (Edgar et al., 2011), UCHIME is one of the most widely
286 used computational tools for detecting chimeric sequences in amplicon sequencing
287 data. The UCHIME algorithm detects chimeras by evaluating how well a query
288 sequence (Q) can be explained as a mosaic of two parent sequences (A and B)

289 from a reference database. The query sequence is first divided into four non-
290 overlapping segments or chunks. Each chunk is independently searched against a
291 reference database that is assumed to be free of chimeras. The best matches to
292 each segment are collected, and from these results, two candidate parent sequences
293 are identified, typically the two sequences that best explain all chunks of the query.
294 Then a three-way alignment among the query (Q) and the two parent candidates
295 (A and B) is done. From this alignment, UCHIME attempts to find a chimeric
296 model (M) which is a hypothetical recombinant sequence formed by concatenating
297 fragments from A and B that best match the observed Q

298 **Chimeric Alignment and Scoring**

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

309 In de novo mode, UCHIME applies an abundance-driven strategy. Only se-
310 quences at least twice as abundant as the query are considered as potential parents.
311 Non-chimeric sequences identified at each step are added iteratively to a growing

312 internal database for subsequent queries.

313 **Limitations of UCHIME**

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

327 **2.3.2 UCHIME2**

328 To overcome the limitations of its predecessor, UCHIME2 (Edgar, 2016) intro-
329 duced several methodological and algorithmic refinements that significantly en-
330 hanced the accuracy and reliability of chimera detection. One major improve-
331 ment lies in its approach to uncertainty handling. In earlier versions, sequences
332 with limited reference support were often incorrectly classified as non-chimeric,

333 increasing the likelihood of false negatives. UCHIME2 addresses this issue by
334 designating such ambiguous sequences as “unknown,” thereby providing a more
335 conservative and reliable classification framework.

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

342 In comparative evaluations, UCHIME2 demonstrated superior detection per-
343 formance, achieving sensitivity levels between 93% and 99% and lower overall
344 error rates than earlier versions or other contemporary tools such as DECIPHER
345 and ChimeraSlayer. Despite these advances, the study also acknowledged a fun-
346 damental limitation in chimera detection: complete error-free identification is
347 theoretically unattainable. This is due to the presence of “perfect fake models,”
348 wherein genuine non-chimeric sequences can be perfectly reconstructed from other
349 reference fragments. This underscore the uncertainty in differentiating authentic
350 biological sequences from artificial recombinants based solely on sequence similar-
351 ity, emphasizing the need for continued methodological refinement and cautious
352 interpretation of results.

353 2.3.3 CATch

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

360 To address these inconsistencies, (Mysara, Saeys, Leys, Raes, & Monsieurs,
361 2015) developed the Classifier for Amplicon Tool Chimeras (CATCh), which rep-
362 resents the first ensemble machine learning system designed for chimera detection
363 in 16S rRNA amplicon sequencing. Rather than depending on a single detec-
364 tion strategy, CATCh integrates the outputs of several established tools, includ-
365 ing UCHIME, ChimeraSlayer, DECIPHER, Pintail, and Perseus. The individual
366 scores and binary decisions generated by these tools are used as input features for
367 a supervised learning model. The algorithm employs a Support Vector Machine
368 (SVM) with a Pearson VII Universal Kernel (PUK) to determine optimal weight-
369 ings among the input features and to assign each sequence a probability of being
370 chimeric.

371 Benchmarking in both reference-based and de novo modes demonstrated signif-
372 icant performance improvements. CATCh achieved sensitivities of approximately
373 85 percent in reference-based mode and 92 percent in de novo mode, with corre-
374 sponding specificities of approximately 96 percent and 95 percent. These results
375 indicate that CATCh detected 7 to 12 percent more chimeras than any individual
376 algorithm while maintaining high precision. Integration of CATCh into amplicon-

377 processing pipelines also reduced operational taxonomic unit (OTU) inflation by
378 23 to 35 percent, producing diversity estimates that more closely reflected true
379 community composition.

380 **2.3.4 ChimPipe**

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

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

396 ChimPipe’s modular design lets users adjust parameters to fit different se-
397 quencing protocols or organism characteristics. Experimental results have con-
398 firmed that many chimeric transcripts detected by the tool correspond to func-

399 tional fusion proteins, showing its value for understanding chimera biology and
400 its potential applications in disease research (Rodriguez-Martin et al., 2017).

401 **2.4 Machine Learning Approaches for Chimera** 402 **and Sequence Quality Detection**

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

414 **2.4.1 Feature-Based Representations of Genomic Se-** 415 **quences**

416 In genomic analysis, feature extraction converts DNA sequences into numerical
417 representations suitable for ML algorithms. A common approach is k-mer fre-
418 quency 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.

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

456 2.5 Synthesis of Chimera Detection Approaches

457 To provide an integrated overview of the literature discussed in this chapter, Ta-
458 ble 2.1 summarizes the major chimera detection studies, their methodological
459 approaches, and their known limitations. This consolidated comparison brings to-
460 gether reference-based approaches, de novo strategies, alignment-driven tools, en-
461 semble machine-learning systems, and general ML-based sequence-quality frame-
462 works. Presenting these methods side-by-side clarifies their performance bound-
463 aries and highlights the unresolved challenges that persist in mitochondrial genome
464 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.
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.
ChimPipe	Pipeline for detecting fusion genes and transcript-derived chimeras in	Designed for RNA-seq, not amplicons; needs high-quality genome

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

479 Chapter 3

480 Research Methodology

481 This chapter outlines the steps involved in completing the study, including data
482 gathering, generating simulated mitochondrial Illumina reads, preprocessing and
483 indexing the data, developing a bioinformatics pipeline to extract key features,
484 applying machine learning algorithms for chimera detection, and validating and
485 comparing model performance.

486 3.1 Research Activities

487 As illustrated in Figure 3.1, this study carried out a sequence of procedures to
488 detect PCR-induced chimeric reads in mitochondrial genomes. The process began
489 with collecting a mitochondrial reference sequence of *Sardinella lemuru* from the
490 National Center for Biotechnology Information (NCBI) database, which was used
491 as a reference for generating simulated clean and chimeric reads. These reads
492 were subsequently indexed and mapped. These datasets will go through a bioin-

493 formatics pipeline that includes k-mer extraction and homology-based filtering to
 494 prepare the data for model construction. The machine learning model will subse-
 495 quently be trained and tested using the processed datasets to assess its precision
 496 and accuracy. The model will undergo refinement and retraining until it meets the
 497 required performance threshold, after which it will proceed to validation, testing,
 498 and deployment.

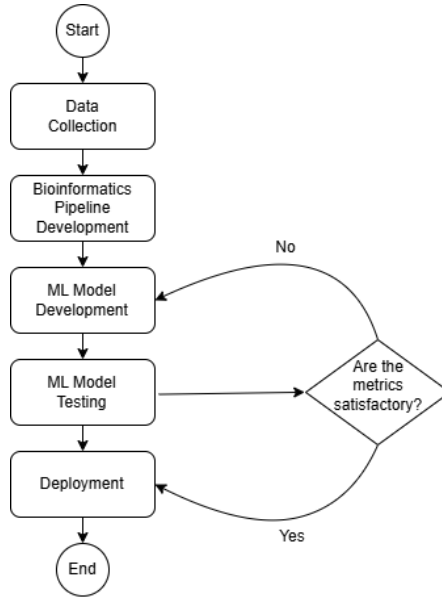


Figure 3.1: Process Diagram of Special Project

499 3.1.1 Data Collection

500 The mitochondrial genome reference sequence of *S. lemuru* was obtained from the
 501 NCBI database (accession number NC_039553.1) in FASTA format. This sequence
 502 served as the basis for generating simulated chimeric reads for model development.

503 This step is scheduled to begin in the first week of November 2025 and is
 504 expected to be completed by the last week of November 2025, with a total duration

505 of approximately one (1) month.

506 Data Preprocessing

507 To reduce manual repetition, all steps in the simulation and preprocessing pipeline
508 were executed using a custom script in Python (Version 3.11). The script runs
509 each stage, including read simulation, reference indexing, mapping, and alignment
510 processing, in a fixed sequence.

511 Sequencing data were simulated from the NCBI reference genome using `wgsim`
512 (Version 1.13). First, 10,000 paired-end reads (R1 and R2) were generated from
513 the original reference (`original_reference.fasta`) and designated as clean reads
514 using the command:

```
515 wgsim -1 150 -2 150 -r 0 -R 0 -X 0 -e 0.001 -N 10000 \  
516         original_reference.fasta ref1.fastq ref2.fastq
```

517 The command parameters are as follows:

- 518 • `-1` and `-2`: read lengths of 150 base pairs for each paired-end read.
- 519 • `-r`, `-R`, `-X`: mutation rate, fraction of indels, and indel extension probability,
520 all set to a default value of 0.
- 521 • `-e`: base error rate, set to 0.001 to simulate realistic sequencing errors.
- 522 • `-N`: number of read pairs, set to 10,000.

523 Chimeric sequences were then generated from the same NCBI reference using a
524 separate Python script. Two non-adjacent segments were randomly selected such
525 that their midpoint distances fell within specified minimum and maximum thresh-
526 olds. The script attempts to retain microhomology, or short identical sequences
527 at segment junctions, to mimic PCR-induced template switching. The resulting
528 chimeras were written to `chimera_reference.fasta`, with headers recording seg-
529 ment positions and microhomology length. The `chimera_reference.fasta` file
530 was subsequently processed with `wgsim` to simulate 10,000 paired-end chimeric
531 reads (`chimeric1.fastq` and `chimeric2.fastq`) using the same command for-
532 mat.

533 Next, a `minimap2` index of the reference genome was created using:

```
534 minimap2 -d ref.mmi original_reference.fasta
```

535 Minimap2 (Version 2.28) is a tool used to map reads to a reference genome.
536 The index `ref.mmi` of the original reference sequence is required by `minimap2` for
537 efficient read mapping. Mapping allows extraction of alignment features from each
538 read, which will be used as input for the machine learning model. The simulated
539 clean and chimeric reads were then mapped to the reference index as follows:

```
540 minimap2 -ax sr -t 8 ref.mmi ref1.fastq ref2.fastq > clean.sam
```

```
541 minimap2 -ax sr -t 8 ref.mmi \  
542 chimeric1.fastq chimeric2.fastq > chimeric.sam
```

543 Here, `-ax sr` specifies short-read alignment mode, and `-t 8` uses 8 CPU

544 threads. The resulting clean and chimeric SAM files contain the alignment posi-
545 tions of each read relative to the original reference genome.

546 The SAM files were then converted to BAM format, sorted, and indexed using
547 `samtools` (Version 1.20):

```
548 samtools view -bS clean.sam -o clean.bam
549 samtools view -bS chimeric.sam -o chimeric.bam
550
551 samtools sort clean.bam -o clean.sorted.bam
552 samtools index clean.sorted.bam
553
554 samtools sort chimeric.bam -o chimeric.sorted.bam
555 samtools index chimeric.sorted.bam
```

556 BAM files are the compressed binary version of SAM files, which enables faster
557 processing and reduced storage. Sorting will arrange reads by genomic coordi-
558 nates, and indexing will allow detection of supplementary alignments (SA) as a
559 feature for the machine learning model.

560 The results of this process will be used for feature extraction. Once the primary
561 features have been extracted from the reads, a custom script will be created to
562 merge them into a single dataset and assign class labels: clean reads as “0” and
563 chimeric reads as “1”. The final dataset will contain 10,000 clean reads and 10,000
564 chimeric reads (a total of 20,000 entries) to ensure equal representation of both
565 classes. The merged dataset will be saved in TSV (`.tsv`) format and subsequently
566 split so that 80% are used for training and 20% for testing.

567 This whole process is scheduled to start in the first week of November 2025
568 and is expected to be completed by the last week of November 2025, with a total
569 duration of approximately one (1) month.

570 **3.1.2 Bioinformatics Tools Pipeline**

571 A bioinformatics pipeline will be developed and implemented to extract the nec-
572 essary analytical features. This pipeline will serve as a reproducible and modular
573 workflow that accepts FASTQ and BAM inputs, processes these through a series
574 of analytical stages, and outputs tabular feature matrices (TSV) for downstream
575 machine learning. All scripts will be version-controlled through GitHub, and
576 computational environments will be standardized using Conda to ensure cross-
577 platform reproducibility. To promote transparency and replicability, the exact
578 software versions, parameters, and command-line arguments used in each stage
579 will be documented. To further ensure correctness and adherence to best practices,
580 bioinformatics experts at the Philippine Genome Center Visayas will be consulted
581 to validate the pipeline design, feature extraction logic, and overall data integrity.
582 This stage of the study is scheduled to begin in the last week of November 2025
583 and conclude by the last week of January 2026, with an estimated total duration
584 of approximately two (2) months.

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

590 Alignment and Supplementary Alignment Count

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

604 K-mer Composition Difference

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

613 5 or 6) for each segment. The resulting k-mer frequency vectors will be normalized
614 and compared using distance metrics such as cosine similarity or Jensen–Shannon
615 divergence to quantify compositional disparity between the two halves of the same
616 read. The resulting difference scores will be stored in a structured TSV file.

617 **Micro-homology Length**

618 The micro-homology length will be computed using a custom Python script that
619 detects the longest exact suffix–prefix overlap within ± 30 base pairs surround-
620 ing a candidate breakpoint. This analysis identifies the number of consecutive
621 bases shared between the end of one segment and the beginning of another. The
622 presence and length of such micro-homology are classic molecular signatures of
623 PCR-induced template switching, where short identical regions (typically 3–15
624 base pairs) promote premature termination and recombination of DNA synthesis
625 on a different template strand. Quantifying micro-homology allows assessment of
626 whether the suspected breakpoint reflects PCR artifacts or true biological variants.
627 Each read will therefore be annotated with its corresponding micro-homology
628 length, overlap sequence, and GC content.

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

635 3.1.3 Machine Learning Model Development

636 This study will explore multiple machine-learning approaches to detect PCR-
637 induced chimeras from mitochondrial Illumina reads: Support Vector Machines
638 (SVM) to separate reads with complex patterns, decision trees to capture hier-
639 archical interactions among SA count, k-mer composition, and micro-homology
640 length, logistic regression as a linear baseline, Random Forest (RF) to improve
641 stability and reduce variance, and gradient boosting (e.g., XGBoost) to model
642 non-linear relationships among the extracted features. Using these approaches
643 enables a balanced assessment of predictive performance and interpretability.

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

651 Model development and evaluation will be implemented in Python (v3.11)
652 using the `scikit-learn` and `xgboost` libraries. Standard metrics including ac-
653 curacy, precision, recall, F1-score, and area under the ROC curve (AUC) will be
654 computed to quantify predictive performance.

655 3.1.4 Validation and Testing

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

664 Comparative evaluation across all candidate algorithms, including SVM, de-
665 cision trees, logistic regression, Random Forest, gradient boosting, and others,
666 will determine which models demonstrate the highest predictive performance and
667 computational efficiency under identical data conditions. Their metrics will be
668 compared to identify the which algorithms are most suitable for further refine-
669 ment.

670 3.1.5 Documentation

671 Comprehensive documentation will be maintained throughout the study to en-
672 sure transparency and reproducibility. All stages of the research, including data
673 gathering, preprocessing, feature extraction, model training, and validation, will
674 be systematically recorded in a .README file in the GitHub repository. For each
675 analytical step, the corresponding parameters, software versions, and command
676 line scripts will be documented to enable exact replication of results.

677 The repository structure will follow standard research data management
678 practices, with clear directories for datasets and scripts. Computational
679 environments will be standardized using Conda, with an environment file
680 (`environment.arm.yml`) specifying dependencies and package versions to main-
681 tain consistency across systems.

682 For manuscript preparation and supplementary materials, Overleaf (L^AT_EX)
683 will be used to produce publication-quality formatting and consistent referencing.

684 3.2 Calendar of Activities

685 Table 3.1 presents the project timeline in the form of a Gantt chart, where each
686 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

732 Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial
733 genomes directly from genomic next-generation sequencing reads—a baiting
734 and iterative mapping approach. *Nucleic Acids Research*, *41*(13), e129. doi:
735 10.1093/nar/gkt371

736 Jin, J.-J., Yu, W.-B., Yang, J., Song, Y., dePamphilis, C. W., Yi, T.-S., & Li,
737 D.-Z. (2020). Getorganelle: a fast and versatile toolkit for accurate de
738 novo assembly of organelle genomes. *Genome Biology*, *21*(1), 241. doi:
739 10.1186/s13059-020-02154-5

740 Judo, M. S. B., Wedel, W. R., & Wilson, B. H. (1998). Stimulation and sup-
741 pression of pcr-mediated recombination. *Nucleic Acids Research*, *26*(7),
742 1819–1825. doi: 10.1093/nar/26.7.1819

743 Labrador, K., Agmata, A., Palermo, J. D., Ravago-Gotanco, R., & Pante, M. J.
744 (2021). Mitochondrial dna reveals genetically structured haplogroups of
745 bali sardinella (*sardinella lemuru*) in philippine waters. *Regional Studies in*
746 *Marine Science*, *41*, 101588. doi: 10.1016/j.rsma.2020.101588

747 Layer, R., Hall, I., & Quinlan, A. (2014, 10). Lumpy: A probabilistic framework
748 for structural variant discovery. *Genome Biology*, *15*. doi: 10.1186/gb-2014-
749 -15-6-r84

750 Li, H. (2018, 05). Minimap2: pairwise alignment for nucleotide sequences. *Bioin-*
751 *formatics*, *34*(18), 3094-3100. Retrieved from [https://doi.org/10.1093/](https://doi.org/10.1093/bioinformatics/bty191)
752 [bioinformatics/bty191](https://doi.org/10.1093/bioinformatics/bty191) doi: 10.1093/bioinformatics/bty191

753 Liang, Q., Bible, P. W., Liu, Y., Zou, B., & Wei, L. (2020, 02). Deepmi-
754 crobes: taxonomic classification for metagenomics with deep learning. *NAR*
755 *Genomics and Bioinformatics*, *2*(1), lqaa009. Retrieved from [https://](https://doi.org/10.1093/nargab/lqaa009)
756 doi.org/10.1093/nargab/lqaa009 doi: 10.1093/nargab/lqaa009

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

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

759 Mysara, M., Saeys, Y., Leys, N., Raes, J., & Monsieurs, P. (2015). Catch,
 760 an ensemble classifier for chimera detection in 16s rna sequencing stud-
 761 ies. *Applied and Environmental Microbiology*, 81(5), 1573-1584. Retrieved
 762 from <https://journals.asm.org/doi/abs/10.1128/aem.02896-14> doi:
 763 10.1128/AEM.02896-14

764 Qin, Y., Wu, L., Zhang, Q., Wen, C., Nostrand, J. D. V., Ning, D., ... Zhou, J.
 765 (2023). Effects of error, chimera, bias, and gc content on the accuracy of
 766 amplicon sequencing. *mSystems*, 8(6), e01025-23. Retrieved from [https://](https://journals.asm.org/doi/abs/10.1128/msystems.01025-23)
 767 journals.asm.org/doi/abs/10.1128/msystems.01025-23 doi: 10.1128/
 768 msystems.01025-23

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

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

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

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

783 Sedlazeck, F., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler,

784 A., & Schatz, M. (2018, 06). Accurate detection of complex structural
785 variations using single-molecule sequencing. *Nature Methods*, 15. doi: 10
786 .1038/s41592-018-0001-7

787 Sfeir, A., & Symington, L. S. (2015). Microhomology-mediated end joining: A
788 back-up survival mechanism or dedicated pathway? *Trends in Biochemical*
789 *Sciences*, 40(11), 701-714. Retrieved from [https://www.sciencedirect](https://www.sciencedirect.com/science/article/pii/S0968000415001589)
790 [.com/science/article/pii/S0968000415001589](https://www.sciencedirect.com/science/article/pii/S0968000415001589) doi: [https://doi.org/](https://doi.org/10.1016/j.tibs.2015.08.006)
791 [10.1016/j.tibs.2015.08.006](https://doi.org/10.1016/j.tibs.2015.08.006)

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

795 Willette, D., Bognot, E., Mutia, M. T., & Santos, M. (2011). *Biology and ecology*
796 *of sardines in the philippines: A review* (Vol. 13; Tech. Rep. No. 1). NFRDI
797 Technical Paper Series. Retrieved from [https://nfrdi.da.gov.ph/tpjf/](https://nfrdi.da.gov.ph/tpjf/etc/Willette%20et%20al.%20Sardines%20Review.pdf)
798 [etc/Willette%20et%20al.%20Sardines%20Review.pdf](https://nfrdi.da.gov.ph/tpjf/etc/Willette%20et%20al.%20Sardines%20Review.pdf)