

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 November 27, 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	25
44	3.1 Research Activities	25
45	3.1.1 Data Collection	26
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

52	3.1 Process Diagram of Special Project	26
----	--	----

53 List of Tables

54	2.1 Summary of Existing Methods and Research Gaps	21
55	3.1 Timetable of Activities	33

⁵⁶ **Chapter 1**

⁵⁷ **Introduction**

⁵⁸ **1.1 Overview**

⁵⁹ The rapid advancement of next-generation sequencing (NGS) technologies has
⁶⁰ transformed genomic research by enabling high-throughput and cost-effective
⁶¹ DNA analysis (Metzker, 2010). Among current platforms, Illumina sequencing
⁶² remains the most widely adopted, capable of producing millions of short reads
⁶³ that can be assembled into reference genomes or analyzed for genetic variation
⁶⁴ (Bentley et al., 2008; Glenn, 2011). Despite its high base-calling accuracy,
⁶⁵ Illumina sequencing is prone to artifacts introduced during library preparation,
⁶⁶ particularly polymerase chain reaction (PCR)-induced chimeras, which are ar-
⁶⁷ tificial hybrid sequences that do not exist in the true genome (Judo, Wedel, &
⁶⁸ Wilson, 1998).

⁶⁹ 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

₉₅ Philippine Genome Center Visayas, with an efficient, interpretable, and resource-
₉₆ optimized solution for improving mitochondrial genome reconstruction.

₉₇ 1.2 Problem Statement

₉₈ While NGS technologies have revolutionized genomic data acquisition, the ac-
₉₉ curacy of mitochondrial genome assembly remains limited by artifacts produced
₁₀₀ during PCR amplification. These chimeric reads can distort assembly graphs and
₁₀₁ cause misassemblies, with especially severe effects in small, circular mitochon-
₁₀₂ drial genomes (Boore, 1999; Cameron, 2014). Existing assembly pipelines such
₁₀₃ as GetOrganelle, MITObim, and NOVOPlasty assume that sequencing reads are
₁₀₄ free of such artifacts (Dierckxsens et al., 2017; Hahn et al., 2013; Jin et al., 2020).
₁₀₅ At the Philippine Genome Center Visayas, several mitochondrial assemblies have
₁₀₆ failed or yielded incomplete contigs despite sufficient coverage, suggesting that
₁₀₇ undetected chimeric reads compromise assembly reliability. Meanwhile, exist-
₁₀₈ ing chimera-detection tools such as UCHIME and VSEARCH were developed
₁₀₉ primarily for amplicon-based microbial community analysis and rely heavily on
₁₁₀ reference or taxonomic comparisons (Edgar, Haas, Clemente, Quince, & Knight,
₁₁₁ 2011; Rognes, Flouri, Nichols, Quince, & Mahé, 2016). These approaches are un-
₁₁₂ suitable for single-species organellar data, where complete reference genomes are
₁₁₃ often unavailable. Therefore, there is a pressing need for a reference-independent,
₁₁₄ data-driven tool capable of automatically detecting and filtering PCR-induced
₁₁₅ chimeras in mitochondrial sequencing datasets.

₁₁₆ **1.3 Research Objectives**

₁₁₇ **1.3.1 General Objective**

₁₁₈ This study aims to develop and evaluate a machine-learning-based pipeline (Mi-
₁₁₉ toChime) that detects PCR-induced chimeric reads in *Sardinella lemuru* mito-
₁₂₀ chondrial sequencing data in order to improve mitochondrial genome assembly
₁₂₁ accuracy.

₁₂₂ **1.3.2 Specific Objectives**

₁₂₃ Specifically, the study aims to:

- ₁₂₄ 1. construct empirical and simulated *Sardinella lemuru* Illumina paired-end
₁₂₅ datasets containing both clean and PCR-induced chimeric reads,
- ₁₂₆ 2. extract alignment-based and sequence-based features such as k-mer compo-
₁₂₇ sition, junction complexity, and split-alignment counts from both clean and
₁₂₈ chimeric reads,
- ₁₂₉ 3. train, validate, and compare supervised machine-learning models for classi-
₁₃₀ fying reads as clean or chimeric,
- ₁₃₁ 4. determine feature importance and identify the most informative indicators
₁₃₂ of PCR-induced chimerism,
- ₁₃₃ 5. integrate the optimized classifier into a modular and interpretable pipeline
₁₃₄ deployable on standard computing environments at PGC Visayas.

135 1.4 Scope and Limitations of the Research

136 This study focuses on detecting PCR-induced chimeric reads in Illumina paired-
137 end mitochondrial sequencing data from *Sardinella lemuru*. The decision to re-
138 strict the taxonomic scope to a single species is based on four considerations:
139 (1) to limit interspecific variation in mitochondrial genome size, GC content, and
140 repetitive regions so that differences in read patterns can be attributed more di-
141 rectly to PCR-induced chimerism; (2) to align the analysis with ongoing *S. lemuru*
142 sequencing projects at the Philippine Genome Center Visayas; (3) to make use of
143 existing *S. lemuru* mitochondrial assemblies and raw datasets available in public
144 repositories such as the National Center for Biotechnology Information (NCBI);
145 and (4) to develop a tool that directly supports local studies on *S. lemuru* popu-
146 lation structure and fisheries management.

147 The study emphasizes `wgsim`-based simulations and selected empirical mito-
148 chondrial datasets from *S. lemuru*. It excludes naturally occurring chimeras, nu-
149 clear mitochondrial pseudogenes (NUMTs), and large-scale structural rearrange-
150 ments in nuclear genomes. Feature extraction is restricted to low-dimensional,
151 hand-crafted alignment and sequence statistics—such as k-mer frequency profiles,
152 GC content, read length, soft-clipping and split-alignment counts, and mapping
153 quality—rather than high-dimensional deep-learning embeddings, so that model
154 behaviour remains interpretable and the pipeline can be executed on standard
155 workstations at PGC Visayas. Testing on long-read platforms (e.g., Nanopore,
156 PacBio) and on other taxa lies beyond the scope of this project; the implemented
157 pipeline is evaluated only on short-read *S. lemuru* datasets.

158 1.5 Significance of the Research

159 This research provides both methodological and practical contributions to mi-
160tochondrial genomics and bioinformatics. First, MitoChime filters PCR-induced
161chimeric reads prior to genome assembly, with the goal of improving the conti-
162guity and correctness of *Sardinella lemuru* mitochondrial assemblies. Second, it
163replaces ad hoc manual curation with a documented, data-driven workflow, im-
164proving automation and reproducibility. Third, the pipeline is designed to run
165on modest computing infrastructures commonly available in regional laborato-
166ries, enabling routine use at the Philippine Genome Center Visayas. In addition,
167MitoChime supports local capacity building by providing a concrete example of
168machine-learning integration into mitochondrial genome analysis, consistent with
169the mandate of the Philippine Genome Center Visayas. Finally, more reliable
170mitochondrial assemblies for *S. lemuru* provide a stronger basis for downstream
171applications such as fisheries management, population genetics, and biodiversity
172assessments.

¹⁷³ **Chapter 2**

¹⁷⁴ **Review of Related Literature**

¹⁷⁵ This chapter presents an overview of the literature relevant to the study. It
¹⁷⁶ discusses the biological and computational foundations underlying mitochondrial
¹⁷⁷ genome analysis and assembly, as well as existing tools, algorithms, and techniques
¹⁷⁸ related to chimera detection and genome quality assessment. The chapter aims to
¹⁷⁹ highlight the strengths, limitations, and research gaps in current approaches that
¹⁸⁰ motivate the development of the present study.

¹⁸¹ **2.1 The Mitochondrial Genome**

¹⁸² Mitochondrial genome (mtDNA) is a small, typically circular molecule found in
¹⁸³ most eukaryotes. It encodes essential genes involved in oxidative phosphorylation
¹⁸⁴ and energy metabolism. Because of its conserved structure and maternal inher-
¹⁸⁵ itance, mtDNA has become a valuable genetic marker for studies in evolution,
¹⁸⁶ population genetics, and phylogenetics (Anderson et al., 1981; Boore, 1999). In

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

195 **2.1.1 Mitochondrial Genome Assembly**

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

210 search.

211 2.2 PCR Amplification and Chimera Formation

212 Polymerase Chain Reaction (PCR) plays an important role in next-generation
213 sequencing (NGS) library preparation, as it amplifies target DNA fragments for
214 downstream analysis. However, the amplification process can also introduce arti-
215 facts that affect data accuracy, one of them being the formation of chimeric se-
216 quences. Chimeras typically arise when incomplete extension occurs during a PCR
217 cycle. This causes the DNA polymerase to switch from one template to another
218 and generate hybrid recombinant molecules (Judo et al., 1998). Artificial chimeras
219 are produced through such amplification errors, whereas biological chimeras oc-
220 cur naturally through genomic rearrangements or transcriptional events. These
221 biological chimeras can have functional roles and may encode tissue-specific novel
222 proteins that link to cellular processes or diseases (Frenkel-Morgenstern et al.,
223 2012).

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

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

235 **2.2.1 Effects of Chimeric Reads on Organelle Genome As-**
236 **sembly**

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

247 2.3 Existing Traditional Approaches for Chimera

248 Detection

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

259 On the other hand, the de novo chimera detection, also referred to as reference-
260 free detection, represents an alternative computational paradigm that identifies
261 chimeric sequences without reliance on external reference databases. Instead of
262 comparing each query sequence to a curated collection of known, non-chimeric
263 sequences, de novo methods infer chimeras based on internal relationships among
264 the sequences present within the dataset itself. This approach is particularly
265 advantageous in studies of novel, under explored, or taxonomically diverse mi-
266 crobial communities where comprehensive reference databases are unavailable or
267 incomplete (Edgar, 2016; Edgar et al., 2011). The underlying assumption on this
268 method operates on the key biological principle that true biological sequences are
269 generally more abundant than chimeric artifacts. During PCR amplification, au-
270 thentic sequences are amplified early and tend to dominate the read pool, while

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

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

289 **2.3.1 UCHIME**

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

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

303 Chimeric Alignment and Scoring

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

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

317 internal database for subsequent queries.

318 **Limitations of UCHIME**

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

332 **2.3.2 UCHIME2**

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

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

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

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

358 **2.3.3 CATch**

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

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

376 Benchmarking in both reference-based and de novo modes demonstrated signif-
377 icant performance improvements. CATCh achieved sensitivities of approximately
378 85 percent in reference-based mode and 92 percent in de novo mode, with corre-
379 sponding specificities of approximately 96 percent and 95 percent. These results
380 indicate that CATCh detected 7 to 12 percent more chimeras than any individual
381 algorithm while maintaining high precision. Integration of CATCh into amplicon-

382 processing pipelines also reduced operational taxonomic unit (OTU) inflation by
383 23 to 35 percent, producing diversity estimates that more closely reflected true
384 community composition.

385 2.3.4 ChimPipe

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

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

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

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

406 **2.4 Machine Learning Approaches for Chimera** 407 **and Sequence Quality Detection**

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

419 **2.4.1 Feature-Based Representations of Genomic Se-** 420 **quences**

421 In genomic analysis, feature extraction converts DNA sequences into numerical
422 representations suitable for ML algorithms. A common approach is k-mer fre-
423 quency analysis, where normalized k-mer counts form the feature vector (Vervier,

424 2015). These features effectively capture local compositional patterns that often
425 differ between authentic and chimeric reads. In particular, deviations in k-mer
426 profiles between adjacent read segments can serve as a compositional signature
427 of template-switching events. Additional descriptors such as GC content and
428 sequence entropy can further distinguish sequence types; in metagenomic classifi-
429 cation and virus detection, k-mer-based features have shown strong performance
430 and robustness to noise (Ren et al., 2020; Vervier, 2015). For chimera detection
431 specifically, abrupt shifts in GC or k-mer composition along a read can indicate
432 junctions between parental fragments. Windowed feature extraction enables mod-
433 els to capture these discontinuities that rule-based algorithms may overlook.

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

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

461 2.5 Synthesis of Chimera Detection Approaches

462 To provide an integrated overview of the literature discussed in this chapter, Ta-
463 ble 2.1 summarizes the major chimera detection studies, their methodological
464 approaches, and their known limitations. This consolidated comparison brings to-
465 gether reference-based approaches, de novo strategies, alignment-driven tools, en-
466 semble machine-learning systems, and general ML-based sequence-quality frame-
467 works. Presenting these methods side-by-side clarifies their performance bound-
468 aries and highlights the unresolved challenges that persist in mitochondrial genome
469 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.

470 Across existing studies, no single approach reliably detects all forms of chimeric
 471 sequences, particularly those generated by PCR-induced template switching in
 472 mitochondrial genomes. Reference-based tools perform poorly when parental se-
 473 quences are absent; de novo methods rely strongly on abundance assumptions;
 474 alignment-based systems show reduced sensitivity to low-divergence chimeras; and
 475 ensemble methods inherit the limitations of their component algorithms. RNA-
 476 seq-oriented pipelines likewise do not generalize well to organelle data. Although
 477 machine learning approaches offer promising feature-based detection, they are
 478 rarely applied to mitochondrial genomes and are not trained specifically on PCR-

479 induced organelle chimeras. These limitations indicate a clear research gap: the
480 need for a specialized, feature-driven classifier tailored to mitochondrial PCR-
481 induced chimeras that integrates k-mer composition, split-alignment signals, and
482 micro-homology features to achieve more accurate detection than current heuristic
483 or alignment-based tools.

⁴⁸⁴ Chapter 3

⁴⁸⁵ Research Methodology

⁴⁸⁶ This chapter outlines the steps involved in completing the project, including data
⁴⁸⁷ gathering through simulation of mitochondrial Illumina reads, preprocessing and
⁴⁸⁸ indexing the data with quality control, developing a bioinformatics pipeline to
⁴⁸⁹ extract key features, applying machine learning algorithms for chimera detection,
⁴⁹⁰ and validating and comparing model performance.

⁴⁹¹ 3.1 Research Activities

⁴⁹² As illustrated in Figure 3.1, this study will carry out a sequence of computa-
⁴⁹³ tional procedures designed to detect PCR-induced chimeric reads in mitochon-
⁴⁹⁴ drial genomes. The process begins with the collection of mitochondrial reference
⁴⁹⁵ sequences from the NCBI database, which will serve as the foundation for gener-
⁴⁹⁶ ating simulated chimeric reads. These datasets will then undergo bioinformatics
⁴⁹⁷ pipeline development, which includes alignment, k-mer extraction, and homology-

498 based filtering to prepare the data for model construction. The machine-learning
499 model will subsequently be trained and tested using the processed datasets to
500 assess its accuracy and reliability. Depending on the evaluation results, the model
501 will either be refined and retrained to improve performance or, if the metrics meet
502 the desired threshold, deployed for further validation and application.

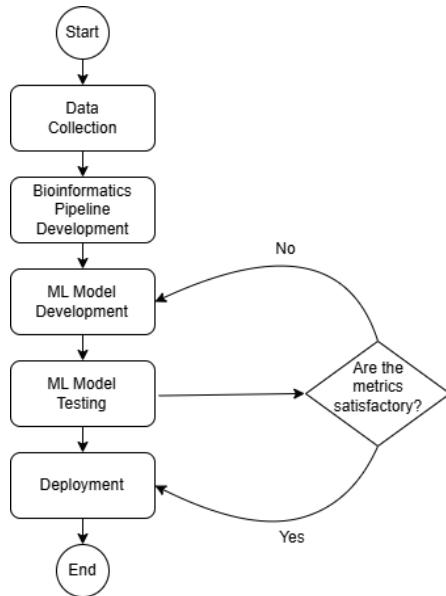


Figure 3.1: Process Diagram of Special Project

503 3.1.1 Data Collection

504 The mitochondrial genome reference sequences of *Sardinella lemuru* will be col-
505 lected from the National Center for Biotechnology Information (NCBI) database.
506 The downloaded files will be in FASTA format to ensure compatibility with bioin-
507 formatics tools and subsequent analysis. The gathered sequences will serve as the
508 basis for generating simulated chimeric reads to be used in model development.

509 The expected outcome of this process is a comprehensive dataset of *Sardinella*

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

514 **Data Preprocessing**

515 Sequencing data will be simulated using the reference sequences collected from
516 NCBI. Using `wgsim`, a total of 5,000 paired-end reads (R1 and R2) will be gener-
517 ated from the reference genome and designated as clean reads. These reads will
518 be saved in FASTQ (`.fastq`) format. From the same reference, a Bash script will
519 be created to deliberately cut and reconnect portions of the sequence, introducing
520 artificial junctions that mimic chimeric regions. The manipulated reference file,
521 saved in FASTA (`.fasta`) format, will then be processed in `wgsim` to simulate
522 an additional 5,000 paired-end chimeric reads, also stored in FASTQ (`.fastq`)
523 format. The resulting read files will be aligned to the original reference genome
524 using SAMtools, generating SAM (`.sam`) or BAM (`.bam`) alignment files. During
525 this alignment process, clean reads will be labeled as “0,” while chimeric reads will
526 be labeled as “1” in a corresponding CSV (`.tsv`) file. This results in a balanced
527 dataset with an equal number of clean and chimeric reads. This is important to
528 prevent model bias and ensure that the machine learning classifiers can learn to
529 detect chimeras accurately.

530 The expected outcome of this process is a complete set of clean and chimeric
531 paired-end reads prepared for subsequent analysis and model development. This
532 step is scheduled to start in the first week of November 2025 and is expected

533 to be completed by the last week of November 2025, with a total duration of
534 approximately one (1) month.

535 3.1.2 Bioinformatics Tools Pipeline

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

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

555 **Alignment and Supplementary Alignment Count**

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

569 **K-mer Composition Difference**

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

578 5 or 6) for each segment. The resulting k-mer frequency vectors will be normalized
579 and compared using distance metrics such as cosine similarity or Jensen–Shannon
580 divergence to quantify compositional disparity between the two halves of the same
581 read. The resulting difference scores will be stored in a structured TSV file.

582 **Micro-homology Length**

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

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

600 3.1.3 Machine-Learning Model Development

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

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

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

622 3.1.4 Validation and Testing

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

631 Comparative evaluation between the Random Forest and XGBoost classifiers
632 will establish which model achieves superior predictive accuracy and computa-
633 tional efficiency under identical data conditions.

634 3.1.5 Documentation

635 Comprehensive documentation will be maintained throughout the study to en-
636 sure transparency, reproducibility, and scientific integrity. All stages of the re-
637 search—including data acquisition, preprocessing, feature extraction, model train-
638 ing, and validation—will be systematically recorded. For each analytical step, the
639 corresponding parameters, software versions, and command-line scripts will be
640 documented to enable exact replication of results.

641 Version control and collaborative management will be implemented through
642 GitHub, which will serve as the central repository for all project files, including
643 Python scripts, configuration settings, and Jupyter notebooks. The repository

644 structure will follow standard research data management practices, with clear
645 directories for datasets, processed outputs, and analysis scripts. Changes will be
646 tracked through commit histories to ensure traceability and accountability.

647 Computational environments will be standardized using Conda, with environ-
648 ment files specifying dependencies and package versions to maintain consistency
649 across systems. Experimental workflows and exploratory analyses will be con-
650 ducted in Jupyter Notebooks, which facilitate real-time visualization, annotation,
651 and incremental testing of results.

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

658 3.2 Calendar of Activities

659 Table 3.1 presents the project timeline in the form of a Gantt chart, where each
660 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	• • •	• • •	• • •	• • •	• • •	• • •	• • •

661 References

- 662 Anderson, S., Bankier, A., Barrell, B., Bruijn, M., Coulson, A., Drouin, J., ...
663 Young, I. (1981, 04). Sequence and organization of the human mitochondrial
664 genome. *Nature*, 290, 457-465. doi: 10.1038/290457a0
- 665 Arango, G., Garner, E., Pruden, A., Heath, L., Vikesland, P., & Zhang, L. (2018,
666 02). Deeparg: A deep learning approach for predicting antibiotic resistance
667 genes from metagenomic data. *Microbiome*, 6. doi: 10.1186/s40168-018
668 -0401-z
- 669 Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J.,
670 Brown, C. G., ... Smith, A. J. (2008). Accurate whole human genome
671 sequencing using reversible terminator chemistry. *Nature*, 456(7218), 53–
672 59. doi: 10.1038/nature07517
- 673 Bi, C., Shen, F., Han, F., Qu, Y., Hou, J., Xu, K., ... Yin, T. (2024, 01).
674 Pmat: an efficient plant mitogenome assembly toolkit using low-coverage
675 hifi sequencing data. *Horticulture Research*, 11(3), uhae023. Retrieved
676 from <https://doi.org/10.1093/hr/uhae023> doi: 10.1093/hr/uhae023
- 677 Boore, J. L. (1999). Animal mitochondrial genomes. *Nucleic Acids Research*,
678 27(8), 1767–1780. doi: 10.1093/nar/27.8.1767
- 679 Cameron, S. L. (2014). Insect mitochondrial genomics: Implications for evolution

- 680 and phylogeny. *Annual Review of Entomology*, 59, 95–117. doi: 10.1146/
681 annurev-ento-011613-162007
- 682 Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). Novoplasty: de novo assembly
683 of organelle genomes from whole genome data. *Nucleic Acids Research*,
684 45(4), e18. doi: 10.1093/nar/gkw955
- 685 Edgar, R. C. (2016). Uchime2: improved chimera prediction for amplicon se-
686 quencing. *bioRxiv*. Retrieved from <https://api.semanticscholar.org/>
687 CorpusID:88955007
- 688 Edgar, R. C. (n.d.). Uchime in practice. Retrieved from https://www.drive5.com/usearch/manual7/uchime_practical.html
- 689
- 690 Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011).
691 Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*,
692 27(16), 2194–2200. doi: 10.1093/bioinformatics/btr381
- 693 Frenkel-Morgenstern, M., Lacroix, V., Ezkurdia, I., Levin, Y., Gabashvili, A.,
694 Prilusky, J., ... Valencia, A. (2012, 05). Chimeras taking shape: Potential
695 functions of proteins encoded by chimeric rna transcripts. *Genome research*,
696 22, 1231-42. doi: 10.1101/gr.130062.111
- 697 Glenn, T. C. (2011). Field guide to next-generation dna sequencers. *Molecular
698 Ecology Resources*, 11(5), 759–769. doi: 10.1111/j.1755-0998.2011.03024.x
- 699 Gonzalez, J. M., Zimmermann, J., & Saiz-Jimenez, C. (2004, 09). Evalu-
700 ating putative chimeric sequences from pcr-amplified products. *Bioin-
701 formatics*, 21(3), 333-337. Retrieved from <https://doi.org/10.1093/bioinformatics/bti008> doi: 10.1093/bioinformatics/bti008
- 702
- 703 Gray, M. W. (2012). Mitochondrial evolution. *Cold Spring Harbor perspectives
704 in biology*, 4. Retrieved from <https://doi.org/10.1101/cshperspect.a011403> doi: 10.1101/cshperspect.a011403
- 705

- 706 Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial
707 genomes directly from genomic next-generation sequencing reads—a baiting
708 and iterative mapping approach. *Nucleic Acids Research*, *41*(13), e129. doi:
709 10.1093/nar/gkt371
- 710 Jin, J.-J., Yu, W.-B., Yang, J., Song, Y., dePamphilis, C. W., Yi, T.-S., & Li,
711 D.-Z. (2020). Getorganelle: a fast and versatile toolkit for accurate de
712 novo assembly of organelle genomes. *Genome Biology*, *21*(1), 241. doi:
713 10.1186/s13059-020-02154-5
- 714 Judo, M. S. B., Wedel, W. R., & Wilson, B. H. (1998). Stimulation and sup-
715 pression of pcr-mediated recombination. *Nucleic Acids Research*, *26*(7),
716 1819–1825. doi: 10.1093/nar/26.7.1819
- 717 Labrador, K., Agmata, A., Palermo, J. D., Ravago-Gotanco, R., & Pante, M. J.
718 (2021). Mitochondrial dna reveals genetically structured haplogroups of
719 bali sardinella (sardinella lemuру) in philippine waters. *Regional Studies in*
720 *Marine Science*, *41*, 101588. doi: 10.1016/j.rsm.2020.101588
- 721 Layer, R., Hall, I., & Quinlan, A. (2014, 10). Lumpy: A probabilistic framework
722 for structural variant discovery. *Genome Biology*, *15*. doi: 10.1186/gb-2014
723 -15-6-r84
- 724 Li, H. (2018, 05). Minimap2: pairwise alignment for nucleotide sequences. *Bioin-*
725 *formatics*, *34*(18), 3094-3100. Retrieved from <https://doi.org/10.1093/bioinformatics/bty191>
726 doi: 10.1093/bioinformatics/bty191
- 727 Liang, Q., Bible, P. W., Liu, Y., Zou, B., & Wei, L. (2020, 02). Deepmi-
728 crobes: taxonomic classification for metagenomics with deep learning. *NAR*
729 *Genomics and Bioinformatics*, *2*(1), lqaa009. Retrieved from <https://doi.org/10.1093/nargab/lqaa009>
730 doi: 10.1093/nargab/lqaa009
- 731 Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature*

- 732 *Reviews Genetics*, 11(1), 31–46. doi: 10.1038/nrg2626
- 733 Mysara, M., Saeys, Y., Leys, N., Raes, J., & Monsieurs, P. (2015). Catch,
734 an ensemble classifier for chimera detection in 16s rrna sequencing stud-
735 ies. *Applied and Environmental Microbiology*, 81(5), 1573-1584. Retrieved
736 from <https://journals.asm.org/doi/abs/10.1128/aem.02896-14> doi:
737 10.1128/AEM.02896-14
- 738 Qin, Y., Wu, L., Zhang, Q., Wen, C., Nostrand, J. D. V., Ning, D., ... Zhou, J.
739 (2023). Effects of error, chimera, bias, and gc content on the accuracy of
740 amplicon sequencing. *mSystems*, 8(6), e01025-23. Retrieved from <https://journals.asm.org/doi/abs/10.1128/msystems.01025-23> doi: 10.1128/
741 msystems.01025-23
- 742 Qiu, X., Wu, L., Huang, H., McDonel, P. E., Palumbo, A. V., Tiedje, J. M., &
743 Zhou, J. (2001). Evaluation of pcr-generated chimeras, mutations, and het-
744 eroduplexes with 16s rrna gene-based cloning. *Applied and Environmental
745 Microbiology*, 67(2), 880–887. doi: 10.1128/AEM.67.2.880-887.2001
- 746 Ren, J., Song, K., Deng, C., Ahlgren, N., Fuhrman, J., Li, Y., ... Sun, F. (2020,
747 01). Identifying viruses from metagenomic data using deep learning. *Quan-
748 titative Biology*, 8. doi: 10.1007/s40484-019-0187-4
- 749 Rodriguez-Martin, B., Palumbo, E., Marco-Sola, S., Griebel, T., Ribeca, P.,
750 Alonso, G., ... Djebali, S. (2017, 01). Chimpipe: Accurate detection of
751 fusion genes and transcription-induced chimeras from rna-seq data. *BMC
752 Genomics*, 18. doi: 10.1186/s12864-016-3404-9
- 753 Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). Vsearch: a
754 versatile open source tool for metagenomics. *PeerJ*, 4, e2584. doi: 10.7717/
755 peerj.2584
- 756 Sedlazeck, F., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler,
757

- 758 A., & Schatz, M. (2018, 06). Accurate detection of complex structural
759 variations using single-molecule sequencing. *Nature Methods*, 15. doi: 10
760 .1038/s41592-018-0001-7
- 761 Sfeir, A., & Symington, L. S. (2015). Microhomology-mediated end joining: A
762 back-up survival mechanism or dedicated pathway? *Trends in Biochemical
763 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>
- 764 Vervier, M. P. T. M. V. J. B. . V. J. P., K. (2015). Large-scale machine learning
765 for metagenomics sequence classification. *Bioinformatics*, 32, 1023 - 1032.
766 Retrieved from <https://api.semanticscholar.org/CorpusID:9863600>
- 767 Willette, D., Bognot, E., Mutia, M. T., & Santos, M. (2011). *Biology and ecology
768 of sardines in the philippines: A review* (Vol. 13; Tech. Rep. No. 1). NFRDI
769 Technical Paper Series. Retrieved from <https://nfrdi.da.gov.ph/tpjf/etc/Willette%20et%20al.%20Sardines%20Review.pdf>
- 770
- 771
- 772