

# Machine Learning Pipeline for Detecting PCR-Induced Chimeric Reads

MitoChime: Organellar Chimera Detection from Per-Read Features

Duran, Lin, Pailden

University of the Philippines Visayas  
Philippine Genome Center Visayas

December 8, 2025

# Outline

1 Objectives

2 Scope and Limitations

3 Methodology

# General Objective

- Develop and evaluate a machine-learning pipeline (MitoChime) to detect PCR-induced chimeric reads in *S. lemuru* mitochondrial sequencing data to improve downstream assembly quality.

# Specific Objectives

- ① Construct simulated *Sardinella lemuru* Illumina paired-end datasets containing both clean and PCR-induced chimeric reads.
- ② Extract alignment-based and sequence-based features such as k-mer composition, junction complexity, and split-alignment counts from both clean and chimeric reads
- ③ Train, validate, and compare supervised machine learning models for classifying reads as clean or chimeric.
- ④ Determine feature importance and identify indicators of PCR-induced chimerism.
- ⑤ Integrate the optimized classifier into a modular and interpretable pipeline deployable on standard computing environments at PGC Visayas.

# Scope of the Study

- Focuses on PCR-induced chimeric reads in *Sardinella lemuru* mitochondrial sequencing data to:
  - 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
  - to align the analysis with relevant *S. lemuru* sequencing projects at PGC Visayas
  - 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
  - to develop a tool that directly supports local studies on *S. lemuru* population structure and fisheries management produce tools applicable to local population and fisheries studies

# Scope of the Study

- Uses wgsim-based simulations and selected empirical mitochondrial datasets
- Analysis targets low-dimensional alignment and sequence features (k-mers, GC content, clipping, split alignments) to maintain interpretability and computational accessibility
- Long-read platforms (Nanopore, PacBio) and other taxa are not included

# Key Exclusions

- Naturally occurring chimeras
- NUMTs
- Large-scale nuclear genome rearrangements
- High-dimensional deep learning embeddings

## Other Limitations

- No simulations with variable sequencing error rates
- No testing of alternative parameter settings (k-mer length, microhomology windows)
- Reliance on supervised machine learning may limit detection of novel/unknown chimeric patterns

# Methodology

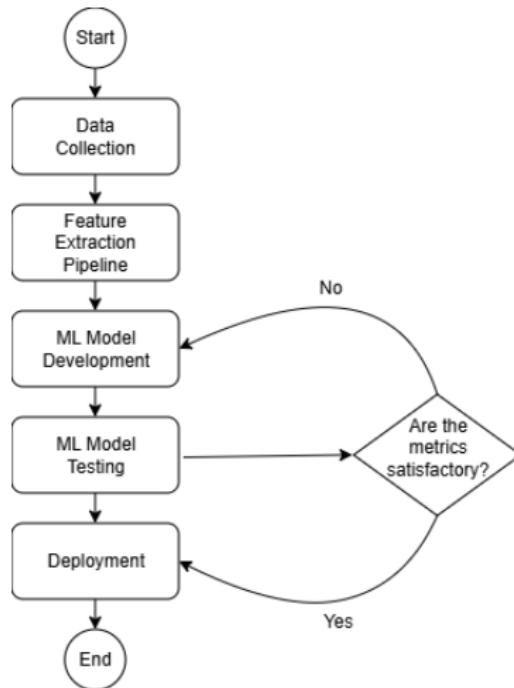


Figure: Process Diagram of the Special Project

# Data Collection

The *S. lemuru* mitochondrial reference genome (NCBI: NC\_039553.1) was downloaded in FASTA format and used as the basis for generating simulated reads.

# Data Preprocessing

- A Python script was used to generate the reads.
- Clean reads were produced with wgsim from the reference genome.
- A chimeric reference was created by creating a custom script to combine non-adjacent segments with microhomology
- Chimeric reads were simulated with wgsim.
- All reads were mapped with minimap2 to extract alignment information.
- SAM/BAM files were converted, sorted, and indexed with samtools.

# Data Preprocessing

- Final dataset: 40k reads, roughly balanced between clean and chimeric (19,984 clean reads and 20,000 chimeric).
- Some of the clean reads failed to align due to the set error rate.

# Data Preprocessing

NC_039553_1_3_540_8:0:0 6:0:0 ef2	163	NC_039553.1	3	60	150M	=	391	538	
TTGTAGCTTAAACAAAGCATAACACTGAAAGATGGTCGATAAGCCCCAACGCACTGAAAGTTGGCTCTGGCTTATTATCAGCTTACCGGAATTACACCAGCGAGGCCCTCCGCGGCCGTGAGGATGCCCTCA									NM:i:8 ms:i:220
AS:i:220	nn:i:8 tp:A:P cm:i:8 si:i:164	s2:i:0	def:f:0.0533	r1:i:0					
NC_039553_1_4_430_13:0:0 11:0:0 _243d	163	NC_039553.1	4	60	150M	=	281	427	
GGTGTAGCTTAAACAAAGCATAACACTGCAAGATGATCCGCTGGCCGTGATAAGCCGCAAGCAGGAGTGAAAGTTGGCCAGGCTTATTATCAGCTTACCCAAATTACACATGCGAGCCTCCGCGGCCGTGAGGATGCCCTCAG									NM:i:13 ms:i:170
AS:i:170	nn:i:0 tp:A:P cm:i:9 si:i:135	s2:i:0	def:f:0.0867	r1:i:0					
NC_039553_1_5_495_6:0:0 11:0:0 _1d49	163	NC_039553.1	5	60	150M	=	346	491	
GTGTAGCTTACACAAAGCATAACACTGAAAGATGTTAAGATGGCCGTGATCAGCCCCAACAGCACTGAAAGTTGGCTCTGGCTTATTATCAGTTTCCCCAATTACACATGCGAGCCTCCGCGGCCGTGAGGATGCCCTCAGC									NM:i:6 ms:i:170
AS:i:240	nn:i:0 tp:A:P cm:i:12 si:i:148	s2:i:0	def:f:0.04	r1:i:0					
NC_039553_1_6_523_6:0:0 9:0:0 _82c	163	NC_039553.1	6	60	150M	=	374	518	
TGTAGCTTAAACAAAGCATAACACTGAAAGATGTTAAGATGGCCGTGATAAGCCCCAACAGCACTGAAAGTTGGCTGGCTATTACAGCTTACCCAAATTACACATGCGCCCTCCGCGGCCGTGAGGATGCCCTCAGCC									NM:i:6 ms:i:240
AS:i:240	nn:i:0 tp:A:P cm:i:10 si:i:157	s2:i:0	def:f:0.04	r1:i:0					
NC_039553_1_9_574_7:0:0 7:0:0 _181b	163	NC_039553.1	9	60	150M	=	425	566	
AGCTTAAACAAAGCATAACACTGCAAGATGTTAAGCTGGCCGTGATAAGCCCCAACAGCACTGAAAGTTGGCTGGCTTATTATCAGCTTACCCAAATTACACATGCGAGCCTCCGCGGCCGTGAGGCTGCCCTCCGCTCC									NM:i:7 ms:i:230
AS:i:230	nn:i:0 tp:A:P cm:i:12 si:i:176	s2:i:0	def:f:0.0467	r1:i:0					
NC_039553_1_10_391_9:0:0 8:0:0 _256b	99	NC_039553.1	10	60	150M	=	242	382	
GCTTAAACAAAGCCAACACTGAAAGATGTTAAGATGGCCGTGATAAGCCCCAACAGCACAGAACAGAAAGTTGGCTCTGGCTTATTAAACAGCTTACCCAAATTAGACATGCGAGCCTCCGCGGCCGTGATGCTGCCCTCAGCTCC									NM:i:9 ms:i:210
AS:i:210	nn:i:0 tp:A:P cm:i:5 si:i:156	s2:i:0	def:f:0.06	r1:i:0					
NC_039553_1_11_509_6:0:0 11:0:0 _a19	99	NC_039553.1	11	60	150M	=	360	499	
CTTCACAAAGCATAACACTGAAAGATGTTAAGATGGCCGTATAAGCCCACAGCACTGAAAGTTAGCTGGCTTATTATGAGCTTACCCAAATTACACATGCGATCTCCGCGGCCGTGAGGATGCCCTCAGCTCCG									NM:i:6 ms:i:242
AS:i:242	nn:i:0 tp:A:P cm:i:10 si:i:150	s2:i:0	def:f:0.04	r1:i:0					
NC_039553_1_12_427_9:0:0 9:0:0 _157	163	NC_039553.1	12	60	150M	=	278	416	
TTAACAAAGCATAACACTGAAAGATTCAGATGGCCGTAAAGCCCCAACAGCACTGAAAGTTGGCTCTGGCTTATTATCAGCTTACCCAAATTACACATGCGAGCCTCGGGGGGCCGTGAGGATGCCCTCCGCTCCGT									NM:i:9 ms:i:210
AS:i:210	nn:i:0 tp:A:P cm:i:8 si:i:158	s2:i:0	def:f:0.06	r1:i:0					

Figure: SAM File of Clean Reads

# Data Preprocessing

```
AS:i:240      nn:i:0  tp:A:P  cm:1:19  s1:i:109      s2:i:0  def:f:0  SA:Z:NC_039553.1,2062,+,34M116S,1,0;  rl:i:0
chimera_1_A9831-10051_B14983-15061_MH0_40985_41514_0:0:0  161  NC_039553.1  89  60  415109M = 7383  7444
CTTAATTATAGGAGGTCCCGCCCTGCTGACCAAAAAGTTTATTATCAGCTTACCCAAATTACACATGGAGGCCCTCGGGCCCCCGTGAGGATGCCCTCAGCCTCCGGAGATGAGGAGCCGGCATCAGGCACGATGTCG
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| NM:i:0 ms:i:218
AS:i:218      nn:i:0  tp:A:P  cm:1:16  s1:i:194  s2:i:0  def:f:0  SA:Z:NC_039553.1,2051,+,45M105S,15,0;  rl:i:0
chimera_1_A9831-10051_B14983-15061_MH0_105028_105471_0:0:0  81  NC_039553.1  89  60  96M54S = 13068  12885
TTTATTATCAGCTTACCCCAAATTACACATGGAGGCCCTCGGGCCCCGGTAGGGATGCCCTAGCCCTCCGGAGATGAGGAGCCGGCATCACCACTTGACAAGGCCACCGGCCTGTACAATTGCGTTACAGCTCTAGCACTCA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| NM:i:0 ms:i:192
AS:i:192      nn:i:0  tp:A:P  cm:1:15  s1:i:87  s2:i:0  def:f:0  SA:Z:NC_039553.1,4313,-,92558M,48,0;  rl:i:0
chimera_1_A9831-10051_B14983-15061_MH0_40665_41142_0:0:0  2371  81  NC_039553.1  89  60  335117M = 3362  3158
-TATAGGAGGTCCGCCCTGCCCTGACCAAAAAGTTTATTATCAGCTTACCCAAATTACACATGGAGGCCCTCGGGCCCCGGTAGGGATGCCCTAGCCTCCGGAGATGAGGAGCCGGCATCAGGCACGATGTCGCCCATGA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| NM:i:0 ms:i:234
AS:i:234      nn:i:0  tp:A:P  cm:1:19  s1:i:109      s2:i:0  def:f:0  SA:Z:NC_039553.1,2059,-,37M113S,1,0;  rl:i:0
chimera_1_A9831-10051_B14983-15061_MH0_41927_41581_0:0:0  97  NC_039553.1  90  60  150M = 7450  7467
TTTATTATCAGCTTACCCCAAATTACACATGGAGGCCCTCGGGCCCCGGTAGGGATGCCCTAGCAGGAGATGAGGAGCCGGCATCACGCACGATGTTGGCCCATGACGCCCTGTTAGCCACACCCCCAAGGGAAATTCA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| NM:i:0 ms:i:300
AS:i:1300     nn:i:0  tp:A:P  cm:1:25  s1:i:139      s2:i:0  def:f:0  rl:i:0
chimera_1_A9831-10051_B14983-15061_MH0_5784_6251_0:0:0  145  NC_039553.1  90  60  150M = 6133  5895
TTTATTATCAGCTTACCCCAAATTACACATGGAGGCCCTCGGGCCCCGGTAGGGATGCCCTAGCAGGAGATGAGGAGCCGGCATCAGGCACGATGTTGGCCCATGACGCCCTGTTAGCCACACCCCCAAGGGAAATTCA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| NM:i:0 ms:i:300
AS:i:1300     nn:i:0  tp:A:P  cm:1:25  s1:i:139      s2:i:0  def:f:0  rl:i:0
chimera_1_A9831-10051_B14983-15061_MH0_32227_32777_0:0:0  161  NC_039553.1  91  60  150M = 6793  6812
NTTATTACAGCTTACCCCAAATTACACATGGAGGCCCTCGGGCCCCGGTAGGGATGCCCTAGCAGGAGATGAGGAGCCGGCATCAGGCACGATGTTGGCCCATGACGCCCTGTTAGCCACACCCCCAAGGGAAATTCA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| NM:i:2 ms:i:296
AS:i:296     nn:i:2  tp:A:P  cm:1:25  s1:i:139      s2:i:0  def:f:0:0.0133  rl:i:0
```

Figure: SAM File of Chimeric Reads

# Dataset construction and split

- Simulated feature tables:
  - Clean reads (label 0)
  - PCR-induced chimeras (label 1)
- `build_datasets.py`:
  - Concatenate tables
  - Shuffle rows (avoid file-order artefacts)
- 80/20 **stratified** train–test split
- Test set held out and used **only once** at the end

# Validation strategy

- Layer 1: 80/20 stratified train–test split
- Layer 2: 5-fold stratified cross-validation on training set
  - Train on 4 folds, validate on 1
  - Rotate so each fold is validation once
- Layer 3: Final evaluation on held-out test set
- Hyperparameter tuning:
  - RandomizedSearchCV inside CV for top models
- Goal: stable estimates and **unbiased** test performance

# Model zoo and preprocessing pipeline

- **Baseline:** dummy majority-class classifier
- **Linear models:** logistic regression, calibrated linear SVM
- **Tree ensembles:**
  - Random Forest, Extra Trees
  - Gradient Boosting, XGBoost, LightGBM, CatBoost
- **Others:** bagging trees, k-NN, Gaussian NB, shallow MLP
- Common scikit-learn pipeline:
  - Median imputation (numeric missing values)
  - Standardisation (zero mean, unit variance)
- Ensures a **fair comparison** across models

# Test-set performance (F1, tuned models)

Figure: Test F1 for tuned models (chimera class).

- Positive class: **chimeric reads**
- Dummy baseline:
  - $F1 \approx 0.67$
- Linear models:
  - $F1 \approx 0.74$  (logreg, linear SVM)
- Best ensembles:
  - CatBoost:  $F1 \approx 0.769$
  - Gradient Boosting:  $F1 \approx 0.767$
  - LightGBM:  $F1 \approx 0.766$
- k-NN, MLP: good but slightly below ensembles

# Effect of hyperparameter tuning

Figure: Test F1: baseline vs tuned.

Figure: Test ROC-AUC: baseline vs tuned.

- Tuning done with `RandomizedSearchCV` on training set
- Small but consistent gains ( $F1, AUC \approx 0.001\text{--}0.01$ )
- Top-ranked models remain the same (CatBoost, Gradient Boosting, LightGBM)

# ROC and precision–recall curves

**Figure:** ROC (left) and PR (right) curves for CatBoost, Gradient Boosting, Random Forest, and logistic regression.

- Ensembles: ROC–AUC  $\approx 0.84$ ; logreg:  $\approx 0.82$
- Average precision  $\approx 0.88$  for ensembles
- Precision  $> 0.9$  up to recall  $\approx 0.5\text{--}0.6$
- Good trade-off across thresholds

# Confusion matrix: CatBoost (test set)

Figure: Confusion matrix heatmap for CatBoost.

- Clean reads:
  - Recall  $\approx 0.95$  (3782 / 3997)
- Chimeric reads:
  - Precision  $\approx 0.92$
  - Recall  $\approx 0.66$  (2631 / 4000)
- Behaviour at default threshold:
  - **Conservative chimera filter**
  - Protects clean reads, misses some subtle chimeras

# Top features for CatBoost

**Figure:** Permutation importance (F1) for CatBoost.

- Strongest signals:
  - kmer\_js\_divergence
  - total\_clipped\_bases
  - kmer\_cosine\_diff
- Also important:
  - Left/right soft-clipping
  - Mapping quality (MAPQ)
  - SA count (supplementary alignments)
- Consistent with PCR chimera junctions

# Feature family importance

Figure: Aggregated feature families for CatBoost.

- Aggregated permutation importance:
  - **Clipping** features dominate
  - **K-mer jump** features also strong
- Smaller contributions:
  - SA structure
  - Micro-homology
  - Other alignment context
- Same pattern for Gradient Boosting and Random Forest

# ML component: summary and implications

- Per-read classifier for clean vs PCR chimeric reads
- Evaluation:
  - Stratified 80/20 split, 5-fold CV, held-out test set
- Best models: tree-based ensembles
  - CatBoost, Gradient Boosting, LightGBM
  - Test F1  $\approx 0.77$ , ROC-AUC  $\approx 0.84$
- Default threshold:
  - Conservative chimera filter (high clean recall, high chimera precision)
  - Removes  $\sim 2/3$  of chimeras
- Features match PCR chimera biology
- Practical, interpretable pre-filter before mitochondrial assembly