

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

Duran, Lin, Pailden

University of the Philippines Visayas

December 14, 2025

# Outline

- 1 Introduction
- 2 Problem Statement & Proposed Solution
- 3 Objectives
- 4 Scope and Limitations
- 5 Methodology

## Next Generation Sequencing (NGS)

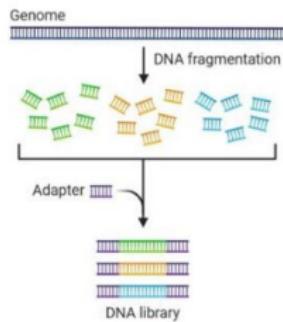


*Source: University of the Philippines  
Visayas, 2022*

## Illumina Seq Workflow

### Step 1. Library Preparation

#### ① Library preparation



Source: Microbe Notes, 2024

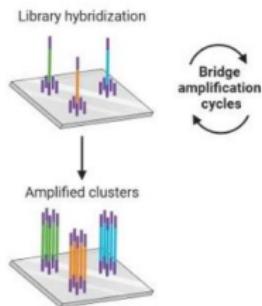


Source: Philippine Genome Center Visayas, 2025

## Illumina Seq Workflow

### Step 2. Library Bridge Amplification (PCR)

#### ② DNA library bridge amplification



Source: Microbe Notes, 2024

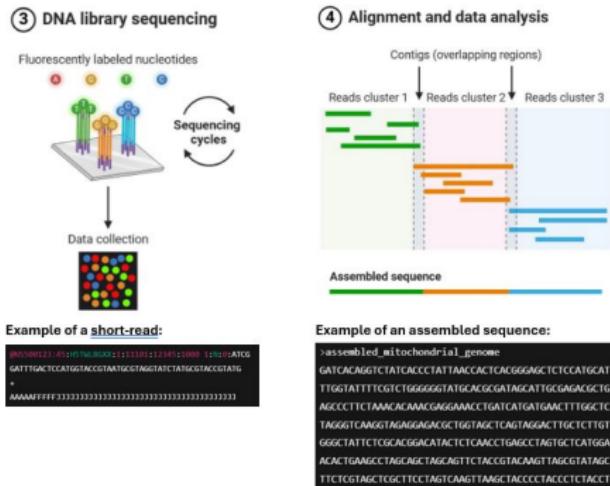


Source: Philippine Genome Center Visayas, 2025

## Real-World Problem

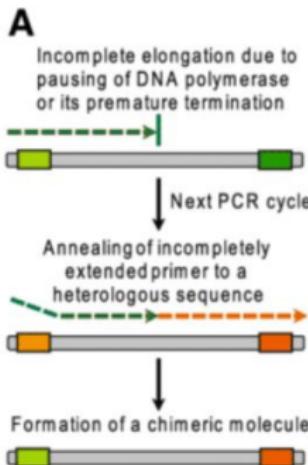
# Illumina Seq Workflow

### Step 3. Sequencing and Alignment



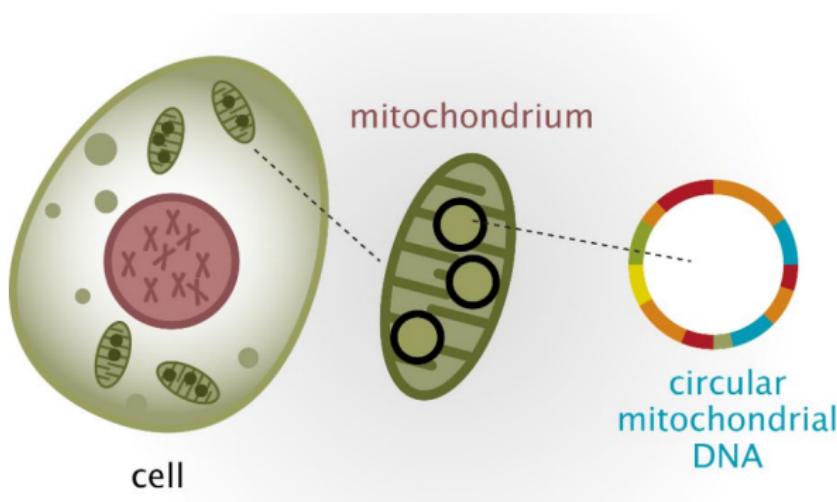
# Real-World Problem

## PCR-Chimera Formation



Source: Omelina et al., 2024

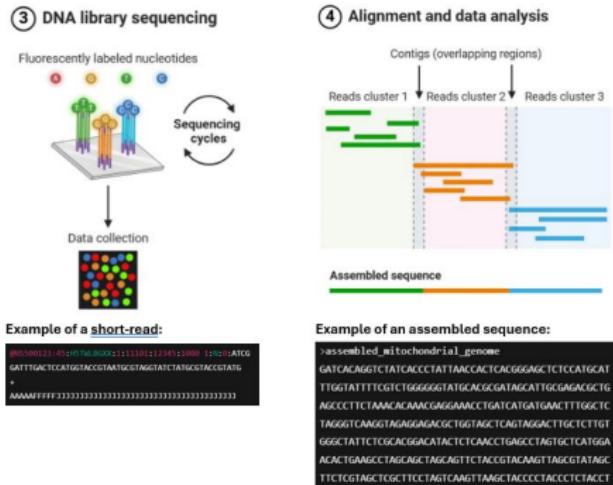
## The Mitochondrial Genome



Source: UZ Bruzel., 2020

## Real-World Problem

## Disrupts Genome Assembly



# Existing Approaches

Table 2.1: Comparison of Chimera Detection Approaches and Tools

Method / Tool	Core Approach	Key Limitations
Reference-based Detection	Compares each query sequence against curated databases of verified, non-chimeric sequences; evaluates segment similarity to identify mosaic patterns.	Accuracy depends on database completeness; performs poorly for novel taxa or missing parents; limited sensitivity for low-divergence chimeras.
De novo Detection	Identifies chimeras using only internal dataset structure; leverages abundance hierarchy and compositional similarity to infer whether low-abundance sequences can be reconstructed from abundant parents.	Assumes true sequences are more abundant; fails when amplification bias distorts abundances; struggles when parental sequences are similarly abundant or highly similar.
UCHIME	Alignment-based model that partitions the query into segments, identifies parent candidates, and computes a chimera score via a three-way alignment; supports reference and de novo modes.	Reduced accuracy for very closely related parents (<0.8% divergence); sensitive to incomplete databases; de novo mode fails if parents are absent or not sufficiently more abundant.
UCHIME2	Updated UCHIME with improved benchmarking (CHSIMA) and multiple sensitivity/specificity presets; better handles incomplete references and dataset variability.	"Fake models" limit theoretical accuracy; genuine variants may mimic chimeras; not recommended as a standalone step in OTU or denoising pipelines due to increased false positives/negatives.
CATCh	First ensemble ML model for 16S chimera detection; integrates outputs of UCHIME, ChimeraSlayer, DECIPHER, Pintail, and Perseus using an SVM to boost overall prediction accuracy.	Performance constrained by underlying tools; ML model cannot capture features not present in component algorithms; may misclassify in highly novel or low-coverage datasets.
ChimPipe	Pipeline for detecting biological chimeras in RNA-seq using discordant paired-end reads and split-read alignments; identifies isoforms and breakpoint coordinates.	Requires high-quality genome and annotation; tailored to RNA-seq rather than amplicons; computationally intensive; limited to organisms with available reference genomes.

# Problem Statement & Proposed Solution

- **Problem Statement:** Chimeric sequencing reads can disrupt mitochondrial genome assembly, but current assembly pipelines assume artifact-free input and existing chimera detection tools are not designed specifically for organellar, particularly mitochondrial datasets, leaving assemblies vulnerable to undetected artifacts.
- **Proposed Solution:** A machine-learning pipeline designed to detect PCR-induced chimeric reads using both alignment-based and sequence-derived features to improve the quality and reliability of downstream mitochondrial genome assemblies.

# 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, microhomology, 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 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

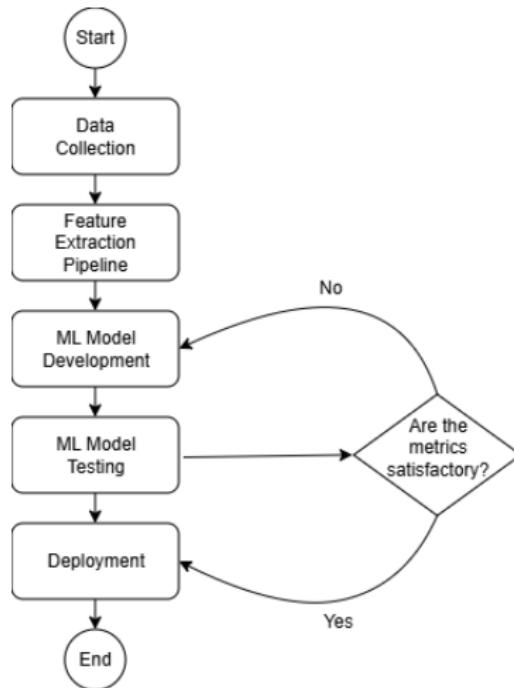


Figure: Process Diagram of the Special Project

# 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
    TGGTAGCTTAAACAAAGCATAACACTGAAAGATGGCTCGTATAAGCCCCAACGCACGTGAAAGTTTGCTGCTGCTTATTATCAGCTTACCGGAATTACACCAGCGAGGCCCTCCGCGGCCGTGAGGATGCCCTCA
    .....
AS:i:220    nn:i:0  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
    GGTTAGCTTAAACAAAGCATAACACTGCAAGATGATCCGCTGGCCGTGATAAGCCGAGCAGGAGTGAAGTTTGCTGAGGCTTATTATCAGCTTACCCAAATTACACATGCGAGCCTCCGCGGCCGTGAGGATGCCCTCAG
    .....
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
    GTGTAGCTTACACAAAGCATAACACTGAAAGATGTTAAGATGGCCGTGATCAGCCCCAACAGCACTGAAAGTTAGGTCTGGCTTATTATCAGTTTCCCCAATTACACATGCGAGCCTCCGCGGCCGTGAGGATGCCCTCAGC
    .....
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
    TGTAGCTTAAACAAAGCATAACACTGAAAGATGTTAAGATGGCCGTGATAAGCCCCAACAGCTGAAAGTTAGGTCTGGCTATTACAGCTTACCCAAATTACACATGCGCCCTCCGCGGCCGTGAGGATGCCCTCAGCC
    .....
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
    AGCTTAAACAAAGCATAACACTGCAAGATGTTAAGCTGGCCGTGATAAGCCCCAACAGCACTGAAAGTTAGGTCTGGCTTATTATCAGCTTACCCAAATTACACATGCGAGCCTCCGCGGCCGTGAGGCTGCCCTCCGCTCC
    .....
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
    GCTTAAACAAAGCTAACACTGAAAGATGTTAAGATGGCCGTGATAAGCCCCAACAGCACAGAAAGTTAGGTCTGGCTTATTACAGCTTACCCAAATTAGACATGCGAGCCTCCGCGGCCGTGATGCTGCCCTCAGCTCC
    .....
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
    CCTCACAAAGCATAACACTGAAAGATGTTAAGATGGCCGTATAAGCCCACAGCACTGAAAGTTAGGTCTGGCTTATTATGAGCTTACCCAAATTACACATGCGATCCCTCCGCGGCCGTGAGGATGCCCTCAGCTCCG
    .....
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
    TAAACAAAGCATAACACTGAAAGATTCAGATGGCCGTAAAGCCCCAACAGCACTGAAAGTTAGGTCTGGCTTATTATCAGCTTACCCAAATTACACATGCGAGCCTCGGGGGGCCGTGAGGATGCCCTCCGCTCCGT
    .....
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
CTTAATTATAGGAGGTCCCGCCCTGGCTTGACCAAAAAGTTTATTATCAGCTTACCCAAATTTCACACATGGAGGCCCTCGGGCCCCCGTGAGGATGCCCTCAGGCTCCCGTCCGGAGATGAGGAGCCGGCATCAGGCACGATGTCG
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| NM:i:0  ms:i:218
AS:i:218      nn:i:0  tp:A:P  cm:1:16  s1:i:94  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
TTTATTATCAGCTTACCCCAAATTTCACACATGGAGGCCCTCGGGCCCCGGTAGGGATGCCCTAGCCCTCCGTCGGAGATGAGGAGCCGGCATCACCACTTGACAAGGCCACCGGCCTGTACAATTGCGTTACAGCTCTAGCACTCA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| 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
-TATAGGAGGTCCGCCCTGCCCTTGACCAAAAAGTTTATTATCAGCTTACCCAAATTTCACACATGGAGGCCCTCGGGCCCCGGTAGGGATGCCCTCGAGGCTCCGTCAGGAGATGAGGAGCCGGCATCAGGCACGATGTCGCCCATGA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| 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  ae9  97  NC_039553.1  90  60  150M = 7450  7467
TTTATATCAGCTTACCCCAAATTTCACATGGAGGCCCTCGGGCCCCGGTAGGGATGCCCTCGGGAGATGAGGAGCCGGCATCAGGCACGATGTTCCGGCCATGAGCCCTGTAGCCACACCCCCAAGGGAAATTCA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| 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
TTTATATCAGCTTACCCCAAATTTCACATGGAGGCCCTCGGGCCCCGGTAGGGATGCCCTAGGGAGATGAGGAGCCGGCATCAGGCACGATGTTCCGGCCATGAGCCCTGTAGCCACACCCCCAAGGGAAATTCA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| 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  be3  161  NC_039553.1  91  60  150M = 6793  6812
NTTATCAGCTTACCCCAAATTTCACATGGAGGCCCTCGGGCCCCGGTAGGGATGCCCTAGGGATGCCCTCGGGAGATGAGGAGCCGGCATCAGGCACGATGTTCCGGCCATGAGCCCTGTAGCCACACCCCCAAGGGAAATTCA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| 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

# Feature Extraction Pipeline

- BAM files were processed with a Python script (`extract_features.py`) to build a TSV feature matrix.
- Used Pysam for parsing alignments and NumPy for computation.

# Feature Extraction Pipeline

- Focused on three features linked to PCR-induced chimeras:
  - ① **Supplementary Alignment (SA)**: Detects split alignments; counts and metrics extracted from SA tags
  - ② **K-mer Composition Difference**: Breakpoints inferred; left/right segments compared using cosine and JS metrics.
  - ③ **Microhomology**: Overlap at junction quantified (length + GC content) within a defined window.
- Pipeline design and outputs to be validated by experts.

# Feature Extraction Pipeline

read_id	label	read_length	mean_basenr	ref_start	l_strand	mappq	cigar	has_ss	sa_count	num_segs	sa_diff	co_sa_min	dk_sa_max	d_sa_mean	sa_same	sa_opp	st_sa_max	nr_sa_mean	sa_min	nr_sa_mean	softclip_le	softclip_le	n_total	clipp_breakpoint	kmer	cosi	kmer_js_d	microhom	microhom
NC_09950	0	150	13 NC_09950	3	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.9720	0.97143	1	0	
NC_09950	0	150	13 NC_09950	4	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98591	0.98571	1	0	
NC_09950	0	150	13 NC_09950	5	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.95887	0.95714	0	0	
NC_09950	0	150	13 NC_09950	6	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.97183	0.97143	1	1	
NC_09950	0	150	13 NC_09950	7	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98644	0.98571	1	0	
NC_09950	0	150	13 NC_09950	10	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.97296	0.97143	0	0	
NC_09950	0	150	13 NC_09950	11	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.97296	0.97143	1	0	
NC_09950	0	150	13 NC_09950	12	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	1	1	1	1	
NC_09950	0	150	13 NC_09950	12	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98648	0.98571	1	1	
NC_09950	0	150	13 NC_09950	12	0	24	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.95690	0.95714	1	1	
NC_09950	0	150	13 NC_09950	14	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	1	1	0	0	
NC_09950	0	150	13 NC_09950	15	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98649	0.98571	1	0	
NC_09950	0	150	13 NC_09950	17	0	60	146M4S	0	0	1	0	0	0	0	0	0	0	0	0	0	4	4	146	0	0.5	0	0		
NC_09950	0	150	13 NC_09950	18	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98649	0.98571	3	0	
NC_09950	0	150	13 NC_09950	18	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.97221	0.97143	3	0	
NC_09950	0	150	13 NC_09950	18	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98629	0.98571	3	0	
NC_09950	0	150	13 NC_09950	19	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.97221	0.97143	3	0	
NC_09950	0	150	13 NC_09950	20	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.97221	0.97143	0	0	
NC_09950	0	150	13 NC_09950	21	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98611	0.98571	1	0	
NC_09950	0	150	13 NC_09950	23	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98667	0.98571	0	0	
NC_09950	0	150	13 NC_09950	25	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98629	0.98571	0	0	
NC_09950	0	150	13 NC_09950	26	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.9863	0.98571	1	0	
NC_09950	0	150	13 NC_09950	32	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.97238	0.97143	2	1	
NC_09950	0	150	13 NC_09950	34	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	1	1	0	0	
NC_09950	0	150	13 NC_09950	34	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98611	0.98571	0	0	
NC_09950	0	150	13 NC_09950	35	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98611	0.98571	1	0	
NC_09950	0	150	13 NC_09950	36	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98648	0.98571	0	0	
NC_09950	0	150	13 NC_09950	38	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98611	0.98571	1	0	
NC_09950	0	150	13 NC_09950	39	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98684	0.98571	0	0	
NC_09950	0	150	13 NC_09950	41	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98296	0.97143	2	0.5	
NC_09950	0	150	13 NC_09950	43	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98611	0.98571	0	0	

Figure: TSV Dataset showing Clean Reads

# Feature Extraction Pipeline

**Figure:** TSV Dataset showing 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

# Effect of hyperparameter tuning

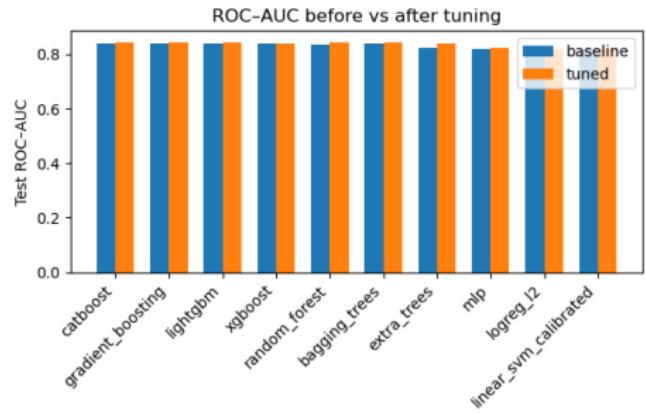
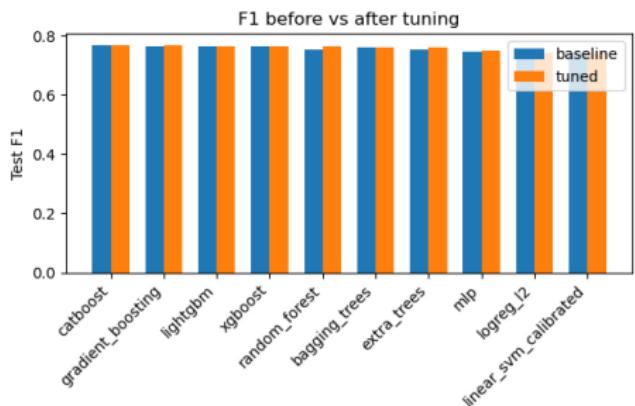


Figure: Test F1: baseline vs tuned.

Figure: Test ROC-AUC: baseline vs tuned.

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

# ROC and precision–recall curves

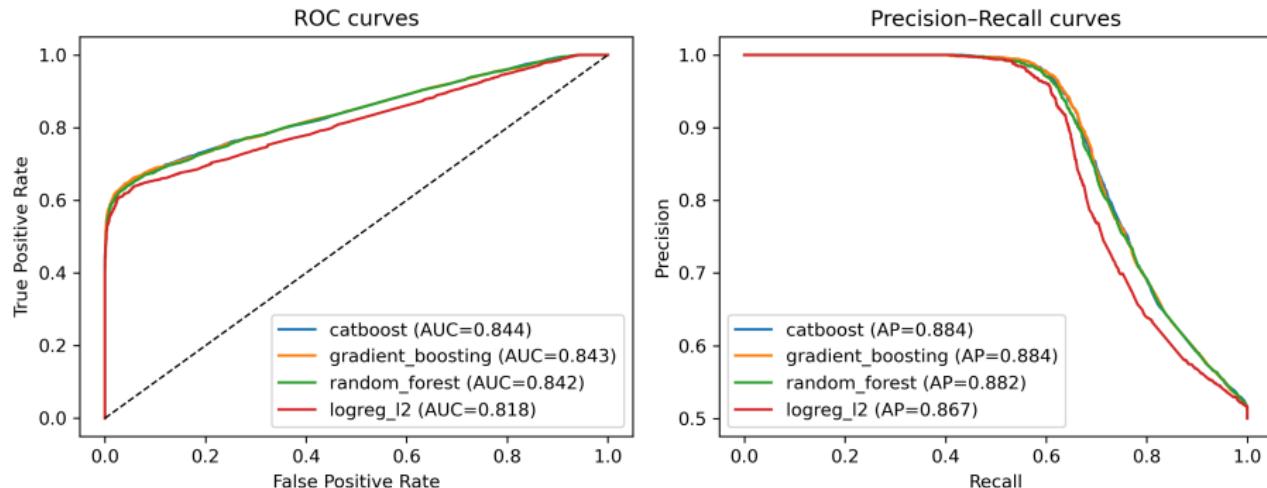


Figure: ROC (left) and PR (right) curves for CatBoost, Gradient Boosting, Random Forest, and logistic regression.

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

# Confusion matrix: CatBoost (test set)

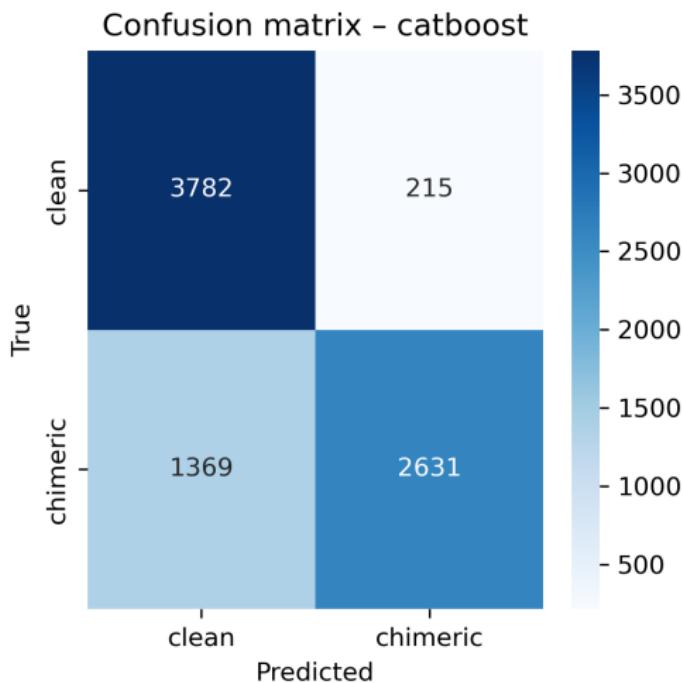
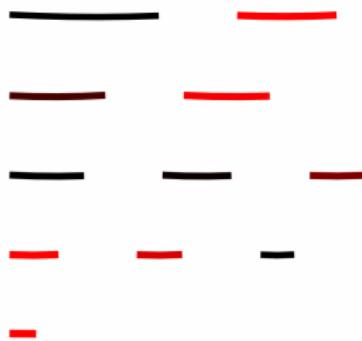
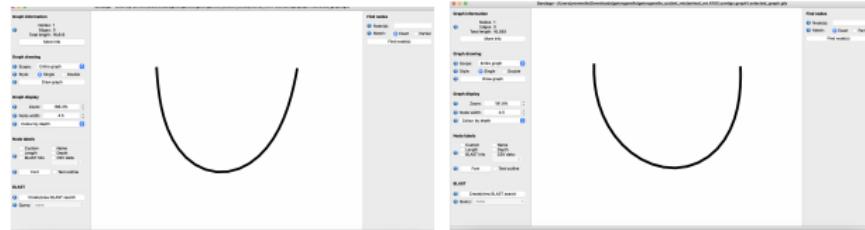


Figure: Confusion matrix heatmap for CatBoost.

- Clean reads (*negative class*):
  - Specificity  $\approx 0.95$  ( $TN = 3782 / 3997$ )
  - False positive rate  $\approx 0.05$  ( $FP = 215 / 3997$ )
- Chimeric reads (*positive class*):
  - Precision  $\approx 0.92$  ( $TP = 2631 / (2631 + 215)$ )
  - Recall  $\approx 0.66$  ( $TP = 2631 / 4000$ )
- Behaviour at default threshold:
  - **Conservative chimera filter:** low FP, higher FN
  - Misses  $\sim 34\%$  of chimeras ( $FN = 1369 / 4000$ )

# Downstream impact of residual chimeras (GetOrganelle)



- **Clean:** 1 contig (16,613 bp)
- **Mixed 50%:** 1 contig (16,593 bp)
- **Chimera-only:** 11 contigs (~39.7% mapped)
- **Implication:** missed chimeras can degrade assembly completeness/contiguity.

**Figure:** Clean, 50% mixed, and chimera-only assemblies.

# Top features for CatBoost

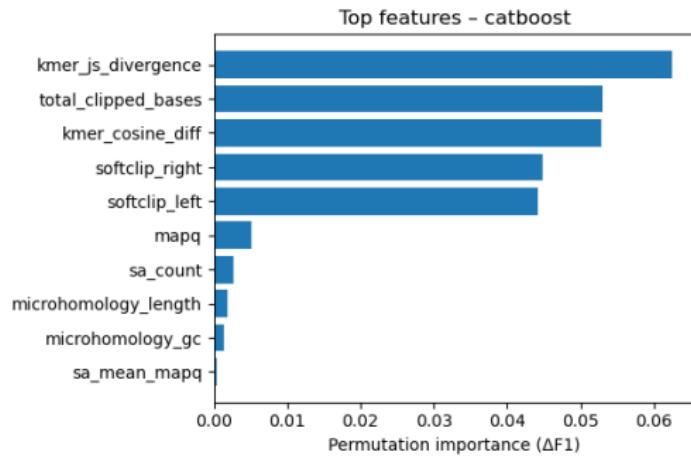
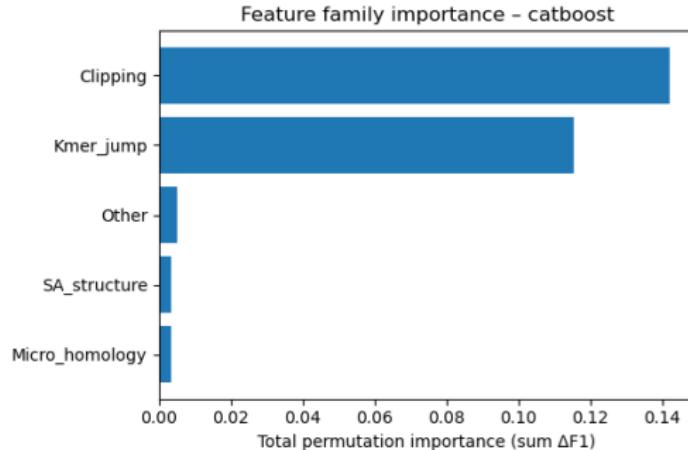


Figure: Permutation importance ( $\Delta F1$ ) for CatBoost.

- Strongest predictive signals:
  - kmer\_js\_divergence (within-read composition shift)
  - total\_clipped\_bases (junction-like clipping)
  - kmer\_cosine\_diff (windowed k-mer change)
- Also informative:
  - Left/right soft-clipping
  - Mapping quality (MAPQ)
  - SA count (supplementary alignments)
- Consistent with PCR chimera breakpoints and split mappings

# Feature family importance



**Figure:** Aggregated permutation importance by feature family (CatBoost).

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

# Summary of findings

- Tree-based ensembles (CatBoost, Gradient Boosting, LightGBM) outperform linear baselines.
- Best performance on held-out reads:
  - $F_1 \approx 0.77$
  - ROC-AUC  $\approx 0.84$
- Most predictive signals match chimera junction patterns:
  - within-read k-mer composition shifts (*k-mer jump*)
  - extensive soft-clipping / clipped bases
- At the default threshold, CatBoost is **conservative**:
  - specificity  $\approx 0.95$  (keeps clean reads)
  - recall  $\approx 0.66$  (misses some chimeras)

# Next steps

- **Error analysis:** characterize FP vs FN cases (focus on false negatives).
- **Calibration:** adjust threshold / use cost-sensitive objective to increase recall while controlling FP.
- **Biological validation:** compare assemblies before vs after filtering (contig count, length, coverage).
- **Exploratory extension:** sequence models (CNN / Transformer / RNN) for subtle breakpoint patterns.