

# 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)

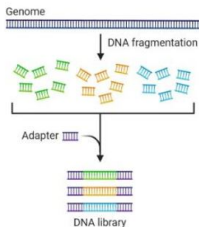


*Source: University of the Philippines  
Visayas, 2022*

## Illumina Seq Workflow

### Step 1. Library Preparation

#### ① Library preparation



Source: Microbe Notes, 2024

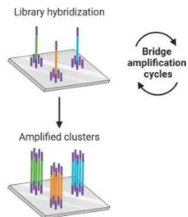


Source: Philippine Genome Center Visayas, 2025

## Illumina Seq Workflow

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

#### ② DNA library bridge amplification



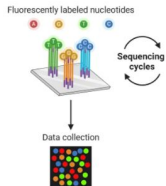
Source: *Microbe Notes*, 2024



Source: *Philippine Genome Center Visayas*, 2025

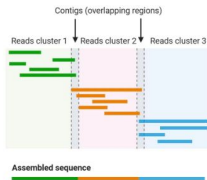
# Illumina Seq Workflow

### Step 3. Sequencing and Alignment



**Example of a short-read:**

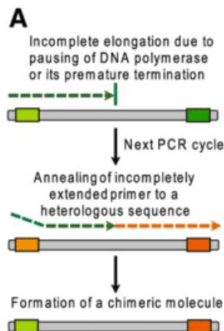
```
@HIS00123:45:HSTMLBXX:1:11101:12345:1000 1:N:0:ATCG  
GATTTCAGCTCCATGGTACCGTAATGCCTAGGTATCTATGCGTACCGTATG  
+  
AAAAFFFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
```



**Example of an assembled sequence:**

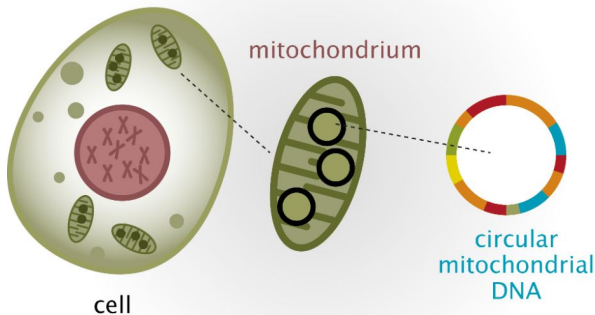
```
>assembled_mitochondrial_genome
GATCAGAGGCTATACCCCTATTAAACCATCAGGGAGCTCTCCATGCA
TTGGTATTTTTCGTGAGGGGTATGCACGGCATGACATTCGAGACAGCTG
AGCCCTCTTAAACATGAGGAGGAAACCTTGATCATGATGACATTGTGCTC
TAGGGTCAGGTAGAGGAGACGCTGTGATCTCAGTAGGACTTGCTCTTGT
GGGCTATTATTCGACGGACATACTCTCAACCTGAGCCCTAGTCTCATGGA
ACACTGAGGCTAGCAGCTAGGCTCTACCGTAGACCTTAGCTATAGATC
TCTCTGATGCTGCTTCTAGTCAGTTGATGTAGCTACCCCTATGCTACTCT
```

## PCR-Chimera Formation



*Source: Omelina et al., 2024*

## The Mitochondrial Genome

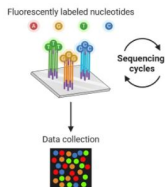


*Source: UZ Brussel., 2020*



## Disrupts Genome Assembly

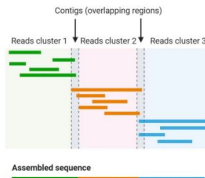
### ③ DNA library sequencing



**Example of a short-read:**

@H5500123.45.HS76L.BXGX.1:11101:12345:1000 1:N:0:ATCG  
GATTGGACTCCATGGTACCGTAATGCCGTAGGTATCTATGGGTACCGTATG  
+  
AAAAAEEEEE11111111111111111111111111111111111111

#### ④ Alignment and data analysis



**Example of an assembled sequence:**

```
>assembled_mitochondrial_genome
GATCAGAGGCTATACCCATTATACCACTACGGGAGCTCTCCATGCA
TTGGTATTTTTCCTCGGGGGTATGCACCGCATGACATTGCGAGACCTG
AGCCCTCTTAAACACAAACGAGGAAACCTGATCATGATGACCTTTTGCTC
TAGGGTCAGGTAGAGGAGACCTGGTAGCTCAGTAGGACTTGCTCTGTG
GGGCTATTCTCGCAGGCACTACTCTCAACCTGAGCCCTAGTCTCATGGA
ACACTGAAGCTAGCAGCTAGCAGTCTACCGTAGACCTTAGCCTATAGC
TTCTCGTAGCTCGCTTCTAGTCAGTTAGTAGTACCCCTACCTCTACCT
```

# Existing Approaches

Table 2.1: Comparison of Chimera Detection Approaches and Tools

| Method / Tool                    | Core Approach  | Key Limitations   |
|----------------------------------|--|---|
| <b>Reference-based Detection</b> | 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.  |
| <b>De novo Detection</b>         | 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.                      |
| <b>UCHIME</b>                    | 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.          |
| <b>UCHIME2</b>                   | 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. |
| <b>CATCh</b>                     | 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.                  |
| <b>ChimPipe</b>                  | 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. lemur* 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

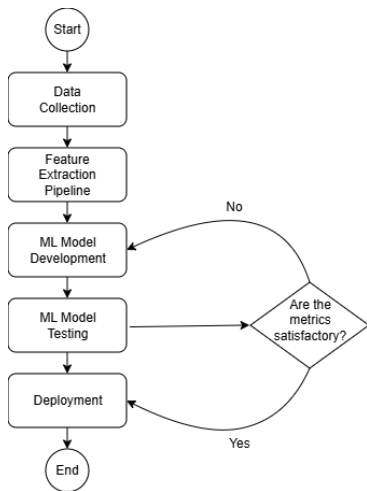


Figure: Process Diagram of the Special Project

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

- 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.

- 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
TGGTGTAGCTTAAACAAGCATAAAGCTGAAGATGTTACGATGGGCGGTGATAAGCCCAACGACCTGAAAGGTTTGGTCTGGCTTTATTATCAGCTTTACCGGAATTTACACACCGAGAGCTCCGCGGCGCGGTGAGGATGGCTCA
..... NM:i:8 ms:i:220
AS:i:220 nn:i:0 tp:A:P cm:i:8 s1:i:164 s2:i:0 de:f:0.0533 rl:i:0
NC_039553.1_4_430_13:0:0_11:0:0_243d 163 NC_039553.1 4 60 150M = 281 427
GGTGTAGCTTAAACAAGCATAAAGCTGAGATGATCCGCTGGGCGGTGATAAGCCGACGAGGAGTGAAGTTTGGTCCAGGCTTTATTATCAGCTTTACCCCAATTTACACATGCGAGCTCCGCGGCGCGGTGAGGATGGCTCAG
..... NM:i:13 ms:i:170
AS:i:170 nn:i:0 tp:A:P cm:i:9 s1:i:135 s2:i:0 de:f:0.0867 rl:i:0
NC_039553.1_5_495_6:0:0_11:0:0_1d49 163 NC_039553.1 5 60 150M = 346 491
GTGTAGCTTACACAAGCATAAAGCTGAAGATGTTAAGATGGGCGGTGATAGCCCAACGACCTGAAAGGTTAGGTCCTGGCTTTATTATCAGGTTTCCCCCAATTTACACATGCGAGCTCCGCGGCGCGGTGAGGATGGCTCAGC
..... NM:i:6 ms:i:240
AS:i:240 nn:i:0 tp:A:P cm:i:12 s1:i:148 s2:i:0 de:f:0.04 rl:i:0
NC_039553.1_6_523_6:0:0_9:0:0_82c 163 NC_039553.1 6 60 150M = 374 518
TGTAGCTTAAACAAGCATAAAGCTGAAGATGTTAAGATGGGCGGTGATAAGCCCAACGACCTGCAAGGTTTGGTCTGGCTATATTACAGCTTTACCCCAATTTACACATGCGGCTCCGCGGCGCGGTGAGGATGGCTCAGCC
..... NM:i:6 ms:i:240
AS:i:240 nn:i:0 tp:A:P cm:i:10 s1:i:157 s2:i:0 de:f:0.04 rl:i:0
NC_039553.1_9_574_7:0:0_7:0:0_181b 163 NC_039553.1 9 60 150M = 425 566
AGCTTAAACAAGCATAAAGCTGAGATGTTAAGCTGGGCGGTGATAAGCCCAACGACCTGAAAGGTTTGGTCTGGCTTTATTATCAGCTTTACCGCAATTTACACATGCGAGCTCCGCGGCGCGGTGAGGCTGCCCTCCGCTCC
..... NM:i:7 ms:i:230
AS:i:230 nn:i:0 tp:A:P cm:i:12 s1:i:176 s2:i:0 de:f:0.0467 rl:i:0
NC_039553.1_10_391_9:0:0_8:0:0_256b 99 NC_039553.1 10 60 150M = 242 382
GCTTAAACAAGCATAAAGCTGAGATGTTAAGATGGGCGGTGATAAGCCCAACGACCTGAAAGGTTTGGTCTGGCTTTATTATCAGCTTTACCCCAATTTAGACATGCGAGCTCCGCGGCGCGGTGATGCTGCCCTCAGCTCCC
..... NM:i:9 ms:i:210
AS:i:210 nn:i:0 tp:A:P cm:i:15 s1:i:156 s2:i:0 de:f:0.06 rl:i:0
NC_039553.1_11_509_6:0:0_11:0:0_a19 99 NC_039553.1 11 60 150M = 360 499
CTTCAACAAGCATAAAGCTGAAGATGTTAAGATGGGCGGTGATAAGCCCAACGACCTGAAAGGTTAGGTCCTGGCTTTATTATGAGCTTTACCCCAATTTACACATGCGATCTCCGCGGCGCGGTGAGGATGCCCTCAGCTCCCG
..... NM:i:6 ms:i:242
AS:i:242 nn:i:0 tp:A:P cm:i:10 s1:i:150 s2:i:0 de:f:0.04 rl:i:0
NC_039553.1_12_427_9:0:0_9:0:0_157 163 NC_039553.1 12 60 150M = 278 416
TTAAACAAGCATAAAGCTGAAGATTTAGATGGGCGGTGATAAGCCCAACGACCTGAAAGTTTGGTCTGGCTTTATTATCAGCTTTACCCCAATTTACACATGCGAGCTCCGCGGCGCGGTGAGGATGCCCTCCGCTCCCGT
..... NM:i:9 ms:i:210
AS:i:210 nn:i:0 tp:A:P cm:i:8 s1:i:150 s2:i:0 de:f:0.06 rl:i:0
```

Figure: SAM File of Clean Reads

# Data Preprocessing

|   |          |        |        |          |          |         |             |         |                              |        |  |  |
|---|----------|--------|--------|----------|----------|---------|-------------|---------|------------------------------|--------|--|--|
| chimer1   | A51:1240 | nm:i:0 | tp:A:P | cm:i:119 | sl:i:109 | s2:i:10 | de:f:0      | SA:Z:NC | 039553.1,2062,+3,4M165,1,0;  | rl:i:0 |  |  |
| <p> CTCAATTATATAGGAGGTCGCCGCTGCCCTGTGACCAAAAGTTTATATACAGCTTACCCCAATTTACACATGCGAGCTCCGCGGGCCCGTAGGATGCGCCTCAGCTCCGTCGGGAGATGAGGAGCGGGCATCAGGCACAGATGTCG<br/> NM:i:0 ms:i:218 </p>  |          |        |        |          |          |         |             |         |                              |        |  |  |
| chimer1   | A51:1218 | nm:i:0 | tp:A:P | cm:i:116 | sl:i:94  | s2:i:10 | de:f:0      | SA:Z:NC | 039553.1,2051,+4,5M105,15,0; | rl:i:0 |  |  |
| <p> TTTTATATACAGCTTACCCCAATTTACACATGCGAGCTCCGCGGGCCCGTAGGATGCGCCTCAGCTCCGTCGGGAGATGAGGAGCGGGCATCACCACCTTGACAGGCCCAACGCCCTGCATCAATTCGCGTTACAGCTTAGCACTCA<br/> NM:i:0 ms:i:192 </p> |          |        |        |          |          |         |             |         |                              |        |  |  |
| chimer1   | A51:1192 | nm:i:0 | tp:A:P | cm:i:115 | sl:i:87  | s2:i:10 | de:f:0      | SA:Z:NC | 039553.1,4313,+9,525M8,48,0; | rl:i:0 |  |  |
| <p> TATAGGAGGTCGCCGCTGCCCTGTGACCAAAAGTTTATATACAGCTTACCCCAATTTACACATGCGAGCTCCGCGGGCCCGTAGGATGCGCCTCAGCTCCGTCGGGAGATGAGGAGCGGGCATCAGGCACAGATGTCGCCGCCATGA<br/> NM:i:0 ms:i:234 </p> |          |        |        |          |          |         |             |         |                              |        |  |  |
| chimer1   | A51:1234 | nm:i:0 | tp:A:P | cm:i:119 | sl:i:109 | s2:i:10 | de:f:0      | SA:Z:NC | 039553.1,2059,-3,7M135,1,0;  | rl:i:0 |  |  |
| <p> TTATATACAGCTTACCCCAATTTACACATGCGAGCTCCGCGGGCCCGTAGGATGCGCCTCAGCTCCGTCGGGAGATGAGGAGCGGGCATCAGGCACAGATGTCGCCGCCATGACGCCCTGTTTAGCCACACCCCCAAGGGAATTCAG<br/> NM:i:0 ms:i:300 </p> |          |        |        |          |          |         |             |         |                              |        |  |  |
| chimer1   | A51:1300 | nm:i:0 | tp:A:P | cm:i:125 | sl:i:139 | s2:i:10 | de:f:0      | rl:i:0  |                              |        |  |  |
| <p> NM:i:0 ms:i:300 </p>  |          |        |        |          |          |         |             |         |                              |        |  |  |
| chimer1   | A51:1300 | nm:i:0 | tp:A:P | cm:i:125 | sl:i:139 | s2:i:10 | de:f:0      | rl:i:0  |                              |        |  |  |
| <p> NM:i:0 ms:i:300 </p>  |          |        |        |          |          |         |             |         |                              |        |  |  |
| chimer1   | A51:1300 | nm:i:0 | tp:A:P | cm:i:125 | sl:i:139 | s2:i:10 | de:f:0      | rl:i:0  |                              |        |  |  |
| <p> NM:i:0 ms:i:300 </p>  |          |        |        |          |          |         |             |         |                              |        |  |  |
| chimer1   | A51:1296 | nm:i:2 | tp:A:P | cm:i:125 | sl:i:139 | s2:i:10 | de:f:0.0133 | rl:i:0  |                              |        |  |  |
| <p> NM:i:2 ms:i:296 </p>  |          |        |        |          |          |         |             |         |                              |        |  |  |

### 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_base | ref_name  | ref_start | 1strand | mapq | cigar  | has_sa | sa_count | num_seg | sa_diff | co | sa_min | dk | sa_max | d | sa_mean | sa_same | sa_op | st | sa_max | r | sa_mean | sa_min | r | sa_mean | softclip_l | softclip_r | total_clip | breakpoint | kmer    | cool | kmer_jc | d | microhom | microhom |     |
|-----------|-------|-------------|-----------|-----------|-----------|---------|------|--------|--------|----------|---------|---------|----|--------|----|--------|---|---------|---------|-------|----|--------|---|---------|--------|---|---------|------------|------------|------------|------------|---------|------|---------|---|----------|----------|-----|
| NC_039502 | 0     | 150         | 13        | NC_039502 | 3         | 0       | 60   | 150M   | 0      | 0        | 0       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.9726     | 0.97143 | 1    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 4         | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98591    | 0.98571 | 1    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 5         | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.95887    | 0.95714 | 0    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 6         | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.97183    | 0.97143 | 1    | 1       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 9         | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98684    | 0.98571 | 0    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 10        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.97296    | 0.97143 | 0    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 11        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 1          | 1       | 1    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 12        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 1          | 1       | 1    | 1       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 12        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98640    | 0.98571 | 1    | 1       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 12        | 0       | 24   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.95889    | 0.95714 | 1    | 1       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 14        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 1          | 1       | 1    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 15        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98649    | 0.98571 | 1    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 17        | 0       | 60   | 148M4S | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 4       | 4          | 148        | 0          | 0.5        | 0       | 0    |         |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 18        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98649    | 0.98571 | 3    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 18        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.97221    | 0.97143 | 3    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 18        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98629    | 0.98571 | 3    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 19        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 1          | 1       | 3    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 20        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.97221    | 0.97143 | 0    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 21        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98611    | 0.98571 | 1    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 23        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98607    | 0.98571 | 0    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 25        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98629    | 0.98571 | 0    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 28        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98603    | 0.98571 | 1    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 32        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.97258    | 0.97143 | 2    | 1       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 34        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 1          | 1       | 1    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 34        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98611    | 0.98571 | 0    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 35        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98611    | 0.98571 | 1    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 36        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98648    | 0.98571 | 0    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 38        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98611    | 0.98571 | 1    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 39        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.98684    | 0.98571 | 0    | 0       |   |          |          |     |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 41        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 0 | 0       | 0       | 0     | 0  | 0      | 0 | 0       | 0      | 0 | 0       | 0          | 0          | 75         | 0.97296    | 0.97143 | 2    | 0       |   |          |          | 0.5 |
| NC_039502 | 0     | 150         | 13        | NC_039502 | 43        | 0       | 60   | 150M   | 0      | 0        | 1       | 0       | 0  | 0      | 0  | 0      | 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

|       | read     | label | read_hl | mean        | ref_nml | sa | strand | id   | cigar   | has_aa | sa_cdu | narrn_hl | sa_dft | sa_mrn | sa_mml | sa_smr | sa_posp | sa_rme | sa_mre | sa_rme | softscd | softscg | total_c | bwaaln  | c       | kmer    | microc  | kitlog | sp |   |
|-------|----------|-------|---------|-------------|---------|----|--------|------|---------|--------|--------|----------|--------|--------|--------|--------|---------|--------|--------|--------|---------|---------|---------|---------|---------|---------|---------|--------|----|---|
| 19985 | chimer_1 | 1     | 150     | 40 NC.03956 | 40      | 1  | +      | 60   | 150M    | 0      | 0      | 0        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 0.9848  | 0.96571 | 0       | 0       | 0      |    |   |
| 19986 | chimer_1 | 1     | 150     | 40 NC.03956 | 53      | 0  | +      | 60   | 150M    | 0      | 0      | 1        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 1       | 1       | 1       | 0       | 0      | 0  |   |
| 19987 | chimer_1 | 1     | 150     | 40 NC.03956 | 65      | 0  | +      | 150M | 0       | 0      | 1      | 0        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 0.95774 | 0.95714 | 0       | 0       | 0      | 0  |   |
| 19988 | chimer_1 | 1     | 150     | 40 NC.03956 | 68      | 0  | +      | 60   | 150M    | 0      | 0      | 0        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 0.95774 | 0.95714 | 0       | 0       | 0      | 0  |   |
| 19989 | chimer_1 | 1     | 150     | 40 NC.03956 | 67      | 0  | +      | 60   | 150M    | 0      | 0      | 1        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 0.95774 | 0.95714 | 0       | 0       | 0      | 0  |   |
| 19990 | chimer_1 | 1     | 150     | 40 NC.03956 | 67      | 1  | 0      | 60   | 118M32S | 1      | 1      | 2        | 0      | 4246   | 4246   | 4246   | 0       | 1      | 10     | 10     | 0       | 0       | 32      | 32      | 118     | 1       | 1       | 0      | 0  |   |
| 19991 | chimer_1 | 1     | 150     | 40 NC.03956 | 69      | 1  | 0      | 150M | 0       | 0      | 0      | 0        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 0.95774 | 0.95714 | 0       | 0       | 0      | 0  |   |
| 19992 | chimer_1 | 1     | 150     | 40 NC.03956 | 78      | 1  | 0      | 60   | 100M44S | 1      | 1      | 2        | 0      | 4237   | 4237   | 4237   | 0       | 16     | 16     | 0      | 0       | 0       | 41      | 41      | 100     | 1       | 1       | 0      | 0  |   |
| 19993 | chimer_1 | 1     | 150     | 40 NC.03956 | 77      | 1  | 0      | 150M | 0       | 0      | 0      | 0        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 0.94306 | 0.84296 | 0       | 0       | 0      | 0  |   |
| 19994 | chimer_1 | 1     | 150     | 40 NC.03956 | 79      | 0  | 0      | 60   | 100M44S | 1      | 1      | 2        | 0      | 4234   | 4234   | 4234   | 0       | 1      | 17     | 17     | 0       | 0       | 44      | 44      | 106     | 1       | 1       | 0      | 0  |   |
| 19995 | chimer_1 | 1     | 150     | 40 NC.03956 | 84      | 0  | +      | 60   | 112M38S | 1      | 1      | 2        | 0      | 5197   | 5197   | 5197   | 0       | 1      | 10     | 10     | 0       | 0       | 38      | 38      | 112     | 0.98377 | 0.98438 | 0      | 0  |   |
| 19996 | chimer_1 | 1     | 150     | 40 NC.03956 | 85      | 0  | 0      | 60   | 111M39S | 1      | 1      | 2        | 0      | 5196   | 5196   | 5196   | 0       | 1      | 20     | 20     | 0       | 0       | 99      | 99      | 111     | 0.98394 | 0.98447 | 0      | 0  |   |
| 19997 | chimer_1 | 1     | 150     | 40 NC.03956 | 88      | 0  | +      | 60   | 150M    | 0      | 0      | 0        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 0.95689 | 0.95654 | 1       | 1       | 0      | 0  |   |
| 19998 | chimer_1 | 1     | 150     | 40 NC.03956 | 89      | 0  | +      | 60   | 155I13S | 0      | 0      | 1        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 15      | 15      | 15      | 1       | 1       | 0      | 0  |   |
| 19999 | chimer_1 | 1     | 150     | 40 NC.03956 | 89      | 0  | 0      | 60   | 306I20M | 1      | 1      | 2        | 0      | 1973   | 1973   | 1973   | 0       | 1      | 1      | 1      | 0       | 0       | 30      | 30      | 30      | 0.98197 | 0.98352 | 0      | 0  |   |
| 20000 | chimer_1 | 1     | 150     | 40 NC.03956 | 89      | 0  | +      | 60   | 41S108M | 1      | 1      | 2        | 0      | 1962   | 1962   | 1962   | 0       | 1      | 15     | 15     | 0       | 0       | 41      | 41      | 41      | 0.98411 | 0.98482 | 0      | 0  |   |
| 20001 | chimer_1 | 1     | 150     | 40 NC.03956 | 89      | 0  | +      | 60   | 96N64S  | 1      | 1      | 2        | 0      | 4234   | 4234   | 4234   | 0       | 48     | 48     | 0      | 0       | 0       | 54      | 54      | 96      | 0       | 0       | 0      | 0  |   |
| 20002 | chimer_1 | 1     | 150     | 40 NC.03956 | 89      | 1  | +      | 60   | 35S117H | 1      | 1      | 2        | 0      | 1970   | 1970   | 1970   | 0       | 1      | 1      | 1      | 0       | 0       | 33      | 33      | 33      | 0.98275 | 0.98389 | 0      | 0  |   |
| 20003 | chimer_1 | 1     | 150     | 40 NC.03956 | 90      | 0  | +      | 60   | 150M    | 0      | 0      | 1        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 0.95832 | 0.95714 | 1       | 1       | 0      | 0  |   |
| 20004 | chimer_1 | 1     | 150     | 40 NC.03956 | 90      | 1  | 0      | 150M | 0       | 0      | 0      | 0        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 0.95832 | 0.95714 | 1       | 1       | 0      | 0  |   |
| 20005 | chimer_1 | 1     | 150     | 40 NC.03956 | 90      | 1  | 0      | 150M | 0       | 0      | 0      | 0        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 0.95832 | 0.95714 | 1       | 1       | 0      | 0  |   |
| 20006 | chimer_1 | 1     | 150     | 40 NC.03956 | 91      | 0  | +      | 60   | 150M    | 0      | 0      | 1        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 0.95774 | 0.95652 | 0       | 0       | 0      | 0  |   |
| 20007 | chimer_1 | 1     | 150     | 40 NC.03956 | 91      | 0  | +      | 60   | 150M    | 0      | 0      | 1        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 0.95832 | 0.95714 | 0       | 0       | 0      | 0  |   |
| 20008 | chimer_1 | 1     | 150     | 40 NC.03956 | 91      | 1  | 0      | 150M | 0       | 0      | 0      | 1        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 75      | 0.95832 | 0.95714 | 1       | 1       | 0      | 0  |   |
| 20009 | chimer_1 | 1     | 150     | 40 NC.03956 | 91      | 1  | 0      | 60   | 94N66S  | 1      | 1      | 2        | 0      | 4222   | 4222   | 4222   | 0       | 1      | 52     | 52     | 0       | 0       | 56      | 56      | 94      | 1       | 1       | 0      | 0  |   |
| 20010 | chimer_1 | 1     | 150     | 40 NC.03956 | 92      | 0  | +      | 60   | 28S121M | 1      | 1      | 2        | 0      | 3684   | 3684   | 3684   | 0       | 72     | 72     | 0      | 0       | 0       | 29      | 29      | 28      | 0.98197 | 0.98388 | 0      | 0  |   |
| 20011 | chimer_1 | 1     | 150     | 40 NC.03956 | 92      | 0  | +      | 60   | 72S78M  | 1      | 1      | 2        | 0      | 3064   | 3064   | 3064   | 0       | 1      | 60     | 60     | 0       | 0       | 72      | 72      | 72      | 0.98088 | 0.98071 | 1      | 1  | 0 |
| 20012 | chimer_1 | 1     | 150     | 40 NC.03956 | 92      | 0  | 0      | 60   | 66S84M  | 1      | 1      | 2        | 0      | 3070   | 3070   | 3070   | 0       | 1      | 59     | 59     | 0       | 0       | 66      | 66      | 66      | 0.98611 | 0.98665 | 1      | 0  |   |
| 20013 | chimer_1 | 1     | 150     | 40 NC.03956 | 92      | 0  | 0      | 60   | 115I39M | 0      | 0      | 1        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 11      | 11      | 11      | 1       | 1       | 0      | 0  |   |
| 20014 | chimer_1 | 1     | 150     | 40 NC.03956 | 92      | 0  | 0      | 60   | 165I31M | 0      | 0      | 1        | 0      | 0      | 0      | 0      | 0       | 0      | 0      | 0      | 0       | 0       | 16      | 16      | 16      | 0.97424 | 0.98041 | 0      | 0  |   |
| 20015 | chimer_1 | 1     | 150     | 40 NC.03956 | 92      | 0  | +      | 60   | 38S47M  | 1      | 1      | 2        | 0      | 3082   | 3082   | 3082   | 0       | 1      | 30     | 30     | 0       | 0       | 54      | 54      | 54      | 0.98505 | 0.98534 | 1      | 0  | 0 |
| 20016 | chimer_1 | 1     | 150     | 40 NC.03956 | 92      | 0  | +      | 60   | 149I39M | 1      | 1      | 2        | 0      | 3082   | 3082   | 3082   | 0       | 1      | 30     | 30     | 0       | 0       | 54      | 54      | 54      | 0.98426 | 0.98534 | 1      | 0  | 0 |

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

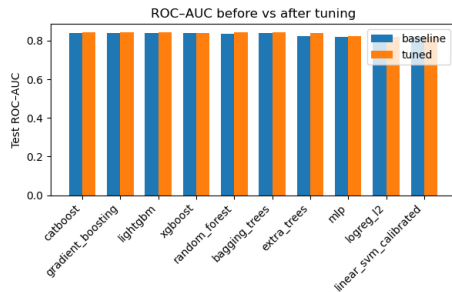
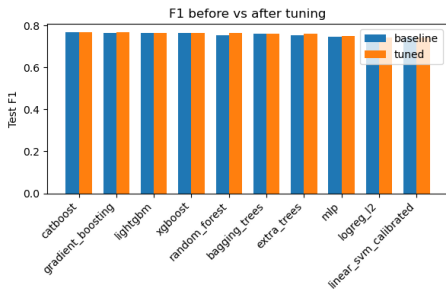
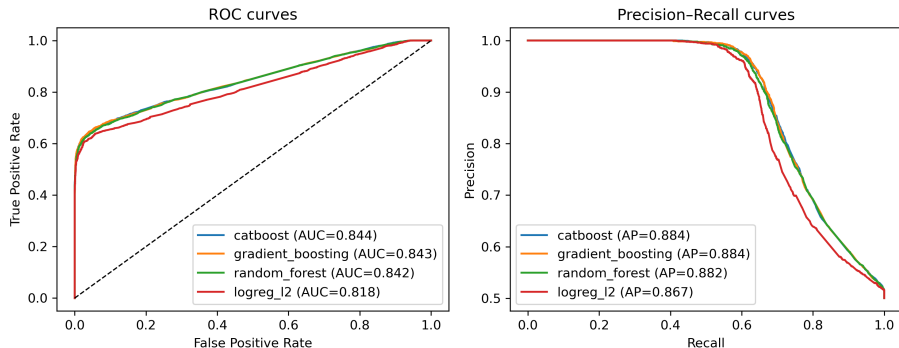


Figure: Test F1: baseline vs tuned.

Figure: Test ROC-AUC: baseline vs tuned.

- Tuning done with RandomizedSearchCV on training set
- Small but consistent gains ( $\Delta F1$ ,  $\Delta AUC \approx 0.001-0.01$ )
- 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  $\approx 0.5$ – $0.6$



# Confusion matrix: CatBoost (test set)

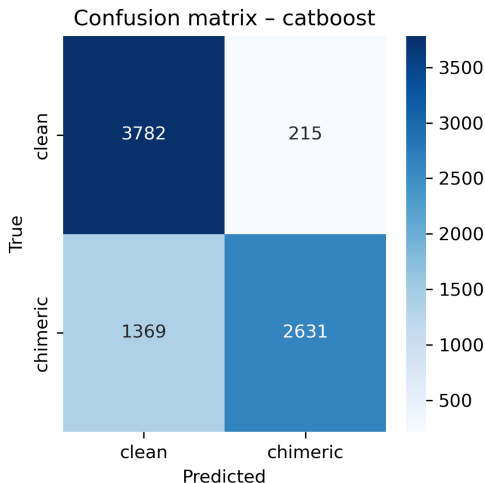
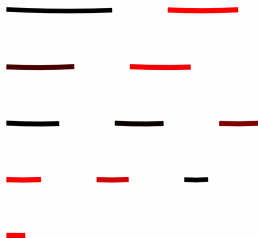
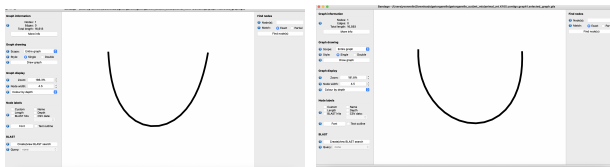


Figure: Confusion matrix heatmap for CatBoost.

- 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 ( $\sim 39.7\%$  mapped)
- **Implication:** missed chimeras can degrade assembly completeness/contiguity.

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

# Top features for CatBoost

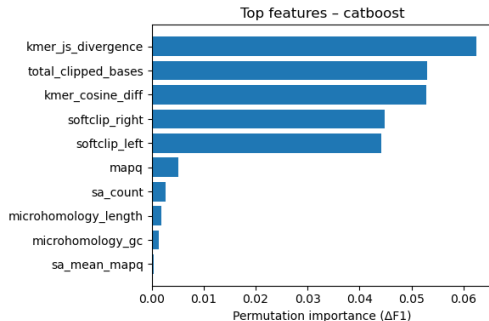
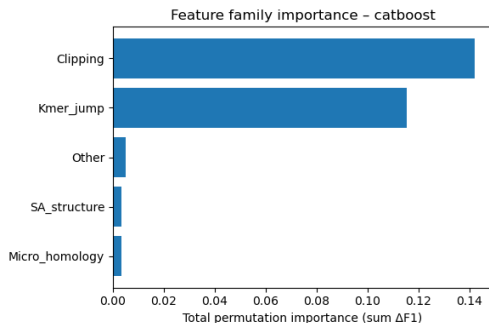


Figure: Permutation importance ( $\Delta F1$ ) for CatBoost.

- 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:
  - $F1 \approx 0.77$
  - $ROC-AUC \approx 0.84$
- 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.