Machine Learning Pipeline for Detecting PCR-Induced Chimeric Reads

MitoChime: Organellar Chimera Detection from Per-Read Features

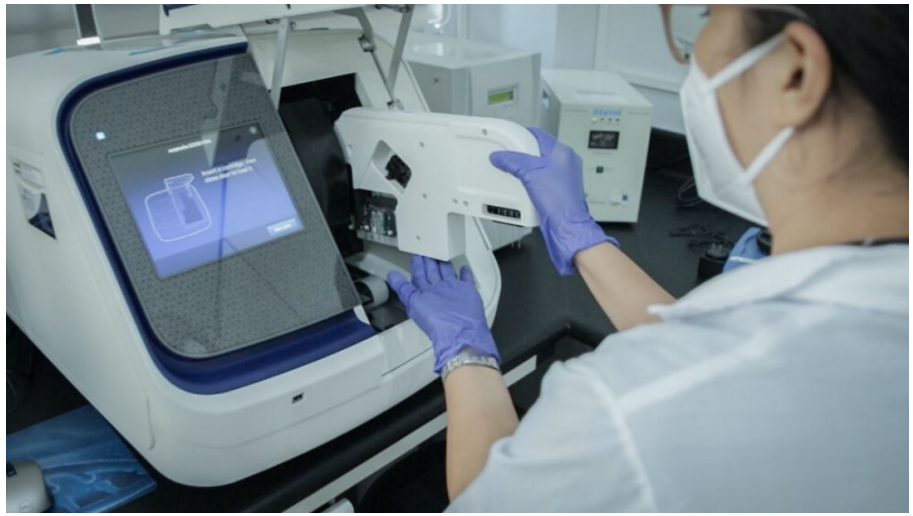
Duran, Lin, Pailden

University of the Philippines Visayas
Philippine Genome Center Visayas

December 9, 2025

Outline

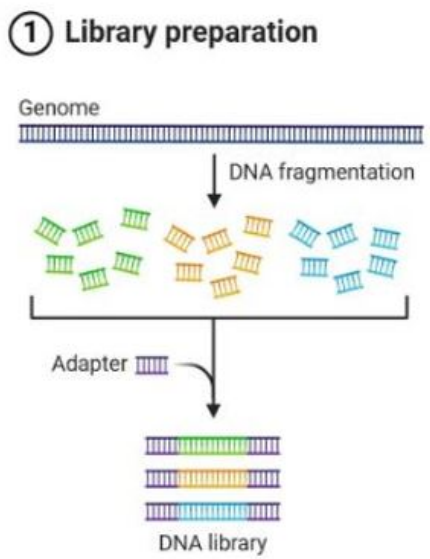
Next Generation Sequencing (NGS)



Source: University of the Philippines Visayas, 2022

Illumina Seq Workflow

Step 1. Library Preparation

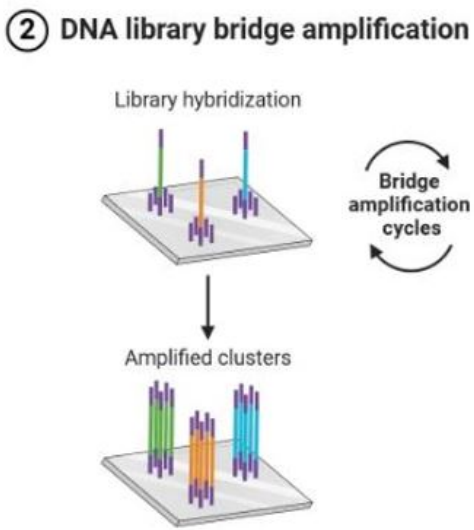


Source: Microbe Notes, 2024

Source: Philippine Genome Center Visayas, 2025

Illumina Seq Workflow

Step 2. Library Bridge Amplification (PCR)

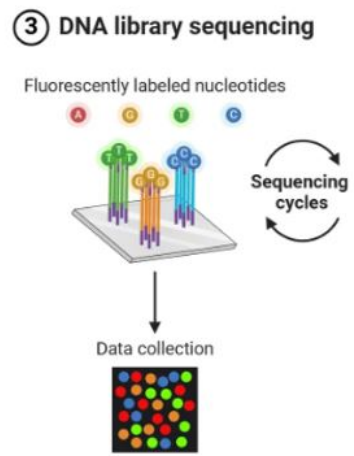


Source: Microbe Notes, 2024

Source: Philippine Genome Center Visayas, 2025

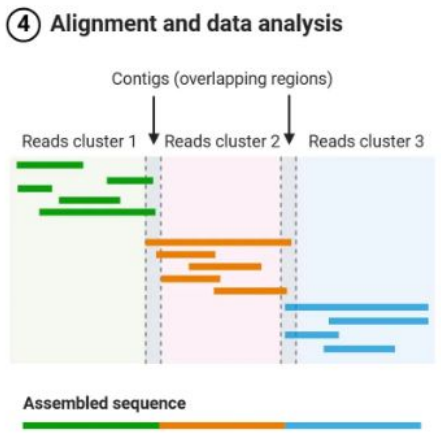
Illumina Seq Workflow

Step 3. Sequencing and Alignment



Example of a short-read:

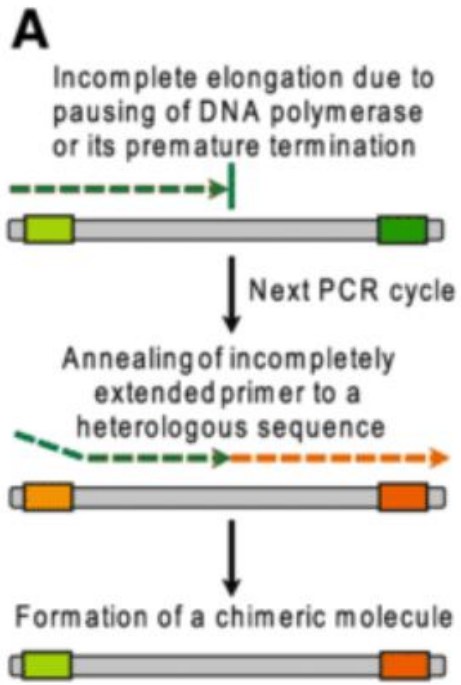
```
@NS500123:45:H5TWLBGXX:1:11101:12345:1000 1:N:0:ATCG
GATTTGACTCCATGGTACCGTAATGCGTAGGTATCTATGCGTACCGTATG
+
AAAAAFFFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
```



Example of an assembled sequence:

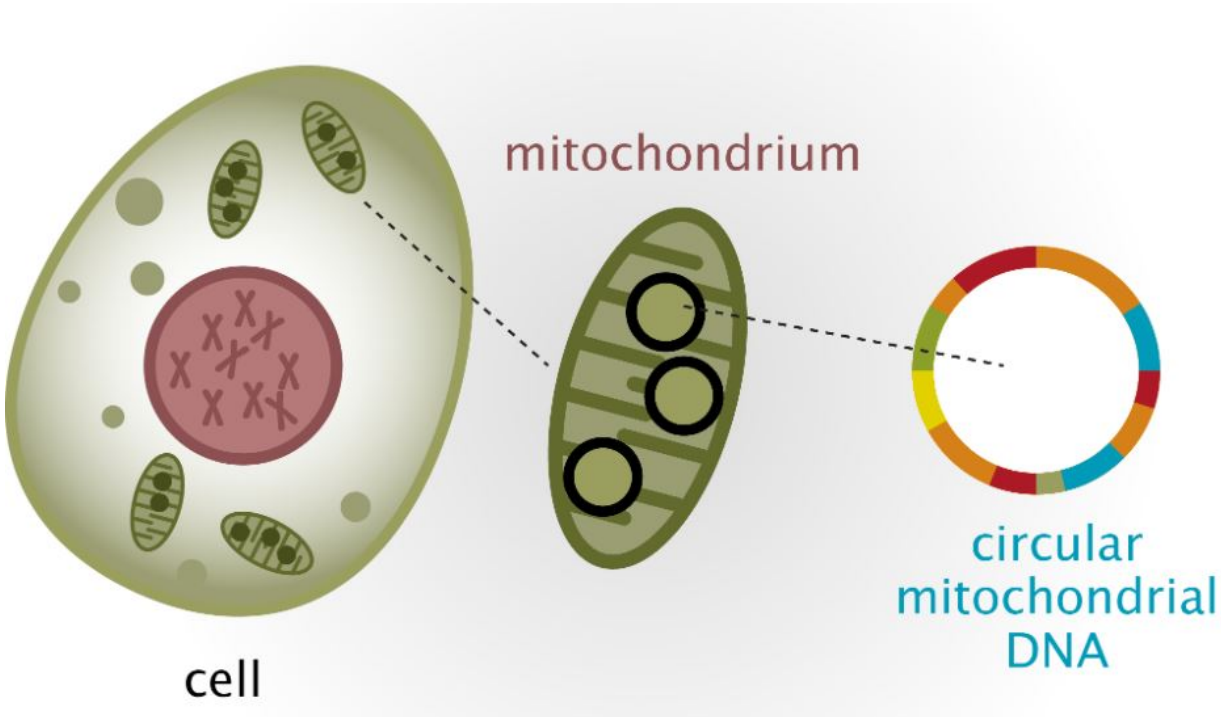
```
>assembled_mitochondrial_genome
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCAT
TTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTG
AGCCCTTCTAAACACAAACGAGGAAACCTGATCATGATGAACTTTGGCTC
TAGGGTCAAGGTAGAGGAGACGCTGGTAGCTCAGTAGGACTTGCTCTTGT
GGGCTATTCTCGCACGGACATACTCTCAACCTGAGCCTAGTGCTCATGGA
ACACTGAAGCCTAGCAGCTAGCAGTTCTACCGTACAAGTTAGCGTATAGC
TTCTCGTAGCTCGCTTCCTAGTCAAGTTAAGCTACCCCTACCCTCTACCT
```

PCR-Chimera Formation



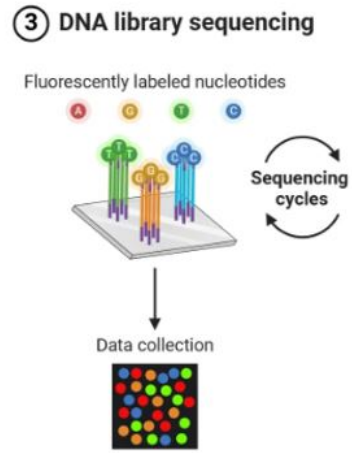
Source: Omelina et al., 2024

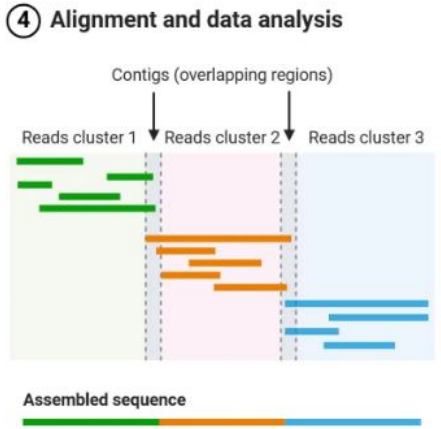
The Mitochondrial Genome



Source: UZ Bruzzel., 2020

Disrupts Genome Assembly





Example of a short-read:

```
@NS500123:45:H5TWLBGXX:1:11101:12345:1000 1:N:0:ATCG
GATTTGACTCCATGGTACCGTAATGCGTAGGTATCTATGCGTACCGTATG
+
AAAAAFFFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
```

Example of an assembled sequence:

```
>assembled_mitochondrial_genome
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCAT
TTGGTATTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTG
AGCCCTTCTAAACACAAACGAGGAAACCTGATCATGATGAACTTTGGCTC
TAGGGTCAAGGTAGAGGAGACGCTGGTAGCTCAGTAGGACTTGCTCTTGT
GGGCTATTCTCGCACGGACATACTCTCAACCTGAGCCTAGTGCTCATGGA
ACACTGAAGCCTAGCAGCTAGCAGTTCTACCGTACAAGTTAGCGTATAGC
TTCTCGTAGCTCGCTTCCTAGTCAAGTTAAGCTACCCCTACCCTCTACCT
```

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 contain ing 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

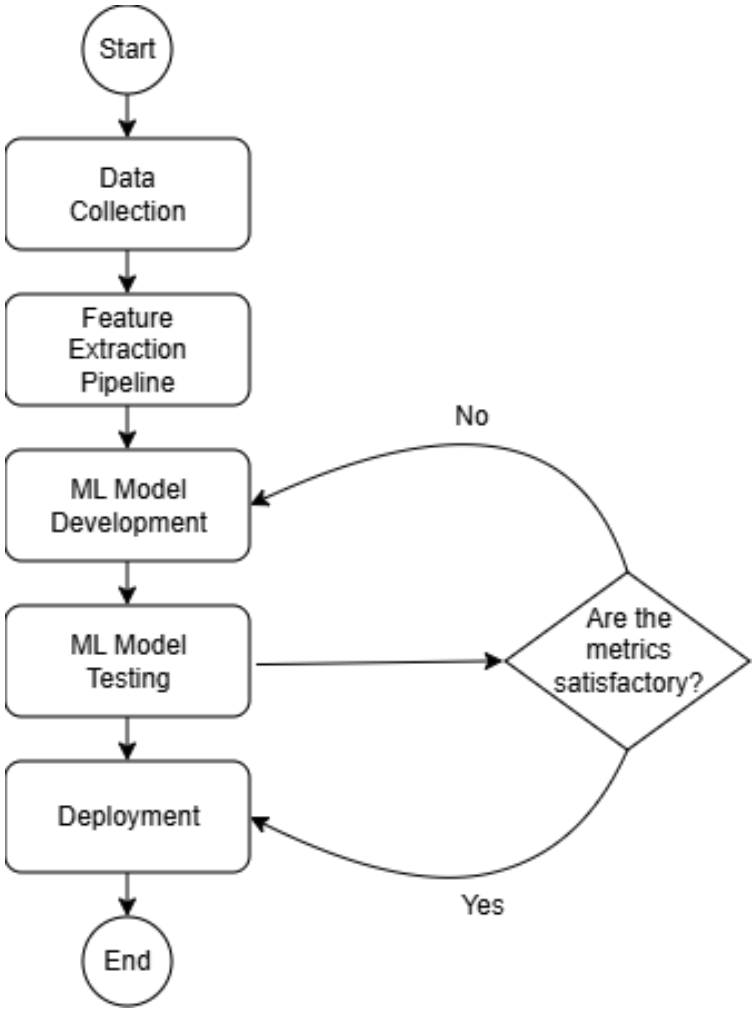


Figure: Process Diagram of the Special Project

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
        TGGTGTAGCTTAAACAAAGCATAACACTGAAGATGTTACGATGGGCCGTGATAAGCCCCACACGCACTGAAAGGTTTGGTCCTGGCTTTATTATCAGCTTTACCGGAATTTACACCAGCGAGCCTCCGCGGCGCCGTGAGGATGGCCTCA
        ........................................................................................................................................................ 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
        GGTGTAGCTTAAACAAAGCATAACACTGCAGATGATCCGCTGGGCCGTGATAAGCCGCAGCAGGAGTGAAAGTTTTGGTCCAGGCTTTATTATCAGCTTTACCCCAAATTACACATGCGAGCCTCCGCGGCCCCGTGAGGATGCCCTCAG
        ........................................................................................................................................................ 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
        GTGTAGCTTACACAAAGCATAACACTGAAGATGTTAAGATGGGCCGTGATCAGCCCCACAAGCACTGAAAGGTTAGGTCCTGGCTTTATTATCAGGTTTCCCCCAATTTACACATGCGAGCCTCCGCGGCCCCGTGAGGATGCGCTCAGC
        ........................................................................................................................................................ 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
        TGTAGCTTAAACAAAGCATAACACTGAAGATGTTAAGATGGGCCGTGATAAGCCCCCCAAGCACTGCAAGGTTTGGTCGTGGCTATATTAACAGCTTTACCCCAATTTACACATGCGCGCCTCCGCGGCCCCGTGAGGATGCCCTCAGCC
        ........................................................................................................................................................ 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
        AGCTTAAACAAAGCATAACACTGCAGATGTTAAGCTGGGCCGTGATAAGCCCCACAAGCACTGAAAGGTTTGGTCGAGGCTTTATTATCAGCTTTACCGCAATTTACACATGCGAGCCTCCGCGGCCCCGTGAGGCTGCCCTCCGCCTCC
        ........................................................................................................................................................ 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
        GCTTAAACAAAGCCTAACACTGAAGATGTAAAGATGGGCCGTGATAAGCCCCACAAGCACAGAAAGGTTTGTTCCTGGCTTTATTAACAGCTTTACCCCAATTTAGACATGCGAGCCTCCGCGGCCCCGTGATGCTGGCCTCAGCCTCCC
        ........................................................................................................................................................ NM:i:9  ms:i:210
        AS:i:210        nn:i:0  tp:A:P  cm:i:5  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
        CTTCAACAAAGCATAACACTGAAGATGTTAAGATGGGCCGTTATAAGCCCGACAAGCACTGAAAGGTTAGGTCCTGGCTTTATTATGAGCTTTACCCCAATTTACACATGCGATCCTCCGCGGCCCCGTGAGGATGCCCTCAGCCTCCCG
        ........................................................................................................................................................ 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
        TTAAACAAAGCATAACACTGAAGAATTTCAGATGGGCCGTGCTAAGGCCCACAAGCACTGAAAGTTTTGGTCCTGGCTTTATTATCAGCTTTACCCCAATTTACACATGCGAGCCTCGGGGGCCCCGTGAGGATGCCCTCCGCCTCCCGT
        ........................................................................................................................................................ NM:i:9  ms:i:210
        AS:i:210        nn:i:0  tp:A:P  cm:i:8  s1:i:158        s2:i:0  de:f:0.06       rl:i:0
```

Figure: SAM File of Clean Reads

Data Preprocessing

```
        AS:i:240        nn:i:0  tp:A:P  cm:i:19 s1:i:109        s2:i:0  de: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_0:0:0_1047     161     NC_039553.1     89      60      41S109M =       7383    7444
        CCTAATTATATAGGAGGTCCCGCCTGCCCTGTGACCAAAAGTTTATTATCAGCTTTACCCCAATTTACACATGCGAGCCTCCGCGGCCCCGTGAGGATGCCCTCAGCCTCCCGTCCGGAGATGAGGAGCGGGCATCAGGCACGATGTTCG
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  NM:i:0  ms:i:218
        AS:i:218        nn:i:0  tp:A:P  cm:i:16 s1:i:94 s2:i:0  de: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_0:0:0_e2e    81      NC_039553.1     89      60      96M54S  =       13068   12885
        TTTATTATCAGCTTTACCCCAATTTACACATGCGAGCCTCCGCGGCCCCGTGAGGATGCCCTCAGCCTCCCGTCCGGAGATGAGGAGCGGGCATCACCCACTTGACAAGCCCCACCGCCTGTACAATTGCCGTTACAGCTCTAGCACTCA
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  NM:i:0  ms:i:192
        AS:i:192        nn:i:0  tp:A:P  cm:i:15 s1:i:87 s2:i:0  de:f:0  SA:Z:NC_039553.1,4313,-,92S58M,48,0;    rl:i:0
chimera_1_A9831-10051_B14983-15061_MH0_40665_41142_0:0:0_0:0:0_2371     81      NC_039553.1     89      60      33S117M =       3362    3158
        TATAGGAGGTCCCGCCTGCCCTGTGACCAAAAGTTTATTATCAGCTTTACCCCAATTTACACATGCGAGCCTCCGCGGCCCCGTGAGGATGCCCTCAGCCTCCCGTCCGGAGATGAGGAGCGGGCATCAGGCACGATGTTCGGCCCATGA
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  NM:i:0  ms:i:234
        AS:i:234        nn:i:0  tp:A:P  cm:i:19 s1:i:109        s2:i:0  de:f:0  SA:Z:NC_039553.1,2059,-,37M113S,1,0;    rl:i:0
chimera_1_A9831-10051_B14983-15061_MH0_41027_41581_0:0:0_0:0:0_ae9      97      NC_039553.1     90      60      150M    =       7450    7467
        TTATTATCAGCTTTACCCCAATTTACACATGCGAGCCTCCGCGGCCCCGTGAGGATGCCCTCAGCCTCCCGTCCGGAGATGAGGAGCGGGCATCAGGCACGATGTTCGGCCCATGACGCCTTGTTTAGCCACACCCCCAAGGGAATTCAG
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  NM:i:0  ms:i:300
        AS:i:300        nn:i:0  tp:A:P  cm:i:25 s1:i:139        s2:i:0  de:f:0  rl:i:0
chimera_1_A9831-10051_B14983-15061_MH0_5784_6251_0:0:0_0:0:0_133d       145     NC_039553.1     90      60      150M    =       6133    5895
        TTATTATCAGCTTTACCCCAATTTACACATGCGAGCCTCCGCGGCCCCGTGAGGATGCCCTCAGCCTCCCGTCCGGAGATGAGGAGCGGGCATCAGGCACGATGTTCGGCCCATGACGCCTTGTTTAGCCACACCCCCAAGGGAATTCAG
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  NM:i:0  ms:i:300
        AS:i:300        nn:i:0  tp:A:P  cm:i:25 s1:i:139        s2:i:0  de:f:0  rl:i:0
chimera_1_A9831-10051_B14983-15061_MH0_5788_6251_0:0:0_0:0:0_1913       81      NC_039553.1     90      60      150M    =       6133    5895
        TTATTATCAGCTTTACCCCAATTTACACATGCGAGCCTCCGCGGCCCCGTGAGGATGCCCTCAGCCTCCCGTCCGGAGATGAGGAGCGGGCATCAGGCACGATGTTCGGCCCATGACGCCTTGTTTAGCCACACCCCCAAGGGAATTCAG
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  NM:i:0  ms:i:300
        AS:i:300        nn:i:0  tp:A:P  cm:i:25 s1:i:139        s2:i:0  de:f:0  rl:i:0
chimera_1_A9831-10051_B14983-15061_MH0_32227_32777_0:0:0_0:0:0_be3      161     NC_039553.1     91      60      150M    =       6793    6812
        NNTTATCAGCTTTACCCCAATTTACACATGCGAGCCTCCGCGGCCCCGTGAGGATGCCCTCAGCCTCCCGTCCGGAGATGAGGAGCGGGCATCAGGCACGATGTTCGGCCCATGACGCCTTGTTTAGCCACACCCCCAAGGGAATTCAGC
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  NM:i:2  ms:i:296
        AS:i:296        nn:i:2  tp:A:P  cm:i:25 s1:i:139        s2:i:0  de:f: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_lengt	mean_bas	ref_name	ref_start_1	strand	mapq	cigar	has_sa	sa_count	num_segn	sa_diff_co	sa_min_de	sa_max_d	sa_mean_	sa_same_	sa_opp_st	sa_max_n	sa_mean_	sa_min_nr	sa_mean_	softclip_le	softclip_ri	total_clipp	breakpoint	kmer_cosi	kmer_js_d	microhom	microhomolc
NC_0395E	0	150	13	NC_0395E	3	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.9726	0.97143	1	0
NC_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	9	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.98664	0.98571	0	0
NC_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	11	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	12	0	24	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.9589	0.95714	1	1
NC_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	17	0	60	146M4S	0	0	1	0	0	0	0	0	0	0	0	0	0	0	4	4	146	0	0.5	0	0
NC_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	19	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	1	1	3	0
NC_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	28	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_0395E	0	150	13	NC_0395E	32	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.97258	0.97143	2	1
NC_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	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_0395E	0	150	13	NC_0395E	41	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.97296	0.97143	2	0.5
NC_0395E	0	150	13	NC_0395E	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

	read_id	label	read_le	mean_	ref_nar	ref_sta	strand	mapq	cigar	has_sa	sa_cou	num_s	sa_diff	sa_min	sa_ma	sa_me	sa_san	sa_opp	sa_ma	sa_me	sa_min	sa_me	softclip	softclip	total_c	breakp	kmer_c	kmer_j	microh	microh	ology_gc
19985	chimera_1	1	150	40	NC_0395	40	1	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	
19986	chimera_1	1	150	40	NC_0395	53	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	1	1	0	0	
19987	chimera_1	1	150	40	NC_0395	65	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.95774	0.95714	0	0	
19988	chimera_1	1	150	40	NC_0395	65	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.95774	0.95714	0	0	
19989	chimera_1	1	150	40	NC_0395	67	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.95774	0.95714	0	0	
19990	chimera_1	1	150	40	NC_0395	67	1	60	118M32S	1	1	2	0	4246	4246	4246	0	1	10	10	0	0	0	32	32	118	1	1	0	0	
19991	chimera_1	1	150	40	NC_0395	69	1	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.95774	0.95714	0	0	
19992	chimera_1	1	150	40	NC_0395	76	0	60	109M41S	1	1	2	0	4237	4237	4237	0	1	16	16	0	0	0	41	41	109	1	1	0	0	
19993	chimera_1	1	150	40	NC_0395	77	1	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.94366	0.94286	0	0	
19994	chimera_1	1	150	40	NC_0395	79	0	60	106M44S	1	1	2	0	4234	4234	4234	0	1	17	17	0	0	0	44	44	106	1	1	0	0	
19995	chimera_1	1	150	40	NC_0395	84	0	60	112M38S	1	1	2	0	5197	5197	5197	0	1	10	10	0	0	0	38	38	112	0.98377	0.98438	0	0	
19996	chimera_1	1	150	40	NC_0395	85	0	60	111M39S	1	1	2	0	5196	5196	5196	0	1	20	20	0	0	0	39	39	111	0.98394	0.98447	0	0	
19997	chimera_1	1	150	40	NC_0395	88	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.95683	0.95554	1	1	
19998	chimera_1	1	150	40	NC_0395	89	0	60	15S135M	0	0	1	0	0	0	0	0	0	0	0	0	0	15	0	15	15	1	1	0	0	
19999	chimera_1	1	150	40	NC_0395	89	0	60	30S120M	1	1	2	0	1973	1973	1973	0	1	1	1	0	0	30	0	30	30	0.98197	0.98352	0	0	
20000	chimera_1	1	150	40	NC_0395	89	0	60	41S109M	1	1	2	0	1962	1962	1962	0	1	15	15	0	0	41	0	41	41	0.98411	0.98462	0	0	
20001	chimera_1	1	150	40	NC_0395	89	1	60	96M54S	1	1	2	0	4224	4224	4224	0	1	48	48	0	0	0	54	54	96	1	1	0	0	
20002	chimera_1	1	150	40	NC_0395	89	1	60	33S117M	1	1	2	0	1970	1970	1970	0	1	1	1	0	0	33	0	33	33	0.98275	0.98389	0	0	
20003	chimera_1	1	150	40	NC_0395	90	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.95832	0.95714	1	1	
20004	chimera_1	1	150	40	NC_0395	90	1	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.95832	0.95714	1	1	
20005	chimera_1	1	150	40	NC_0395	90	1	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.95832	0.95714	1	1	
20006	chimera_1	1	150	40	NC_0395	91	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.95774	0.95652	0	0	
20007	chimera_1	1	150	40	NC_0395	91	0	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.95832	0.95714	0	0	
20008	chimera_1	1	150	40	NC_0395	91	1	60	150M	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0.95832	0.95714	0	0	
20009	chimera_1	1	150	40	NC_0395	91	1	60	94M56S	1	1	2	0	4222	4222	4222	0	1	52	52	0	0	0	56	56	94	1	1	0	0	
20010	chimera_1	1	150	40	NC_0395	92	0	60	29S121M	0	0	1	0	0	0	0	0	0	0	0	0	0	29	0	29	29	0.98167	0.98338	1	0	
20011	chimera_1	1	150	40	NC_0395	92	0	60	72S78M	1	1	2	0	3064	3064	3064	0	1	60	60	0	0	72	0	72	72	0.98608	0.98571	1	0	
20012	chimera_1	1	150	40	NC_0395	92	0	60	66S84M	1	1	2	0	3070	3070	3070	0	1	59	59	0	0	66	0	66	66	0.98611	0.98565	1	0	
20013	chimera_1	1	150	40	NC_0395	92	0	60	11S139M	0	0	1	0	0	0	0	0	0	0	0	0	0	11	0	11	11	1	1	1	0	
20014	chimera_1	1	150	40	NC_0395	92	0	60	16S134M	0	0	1	0	0	0	0	0	0	0	0	0	0	16	0	16	16	0.97424	0.98041	1	0	
20015	chimera_1	1	150	40	NC_0395	92	0	60	3S147M	0	0	1	0	0	0	0	0	0	0	0	0	0	3	0	3	3	0	0.5	1	0	
20016	chimera_1	1	150	40	NC_0395	92	0	60	54S96M	1	1	2	0	3082	3082	3082	0	1	30	30	0	0	54	0	54	54	0.9855	0.98534	1	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

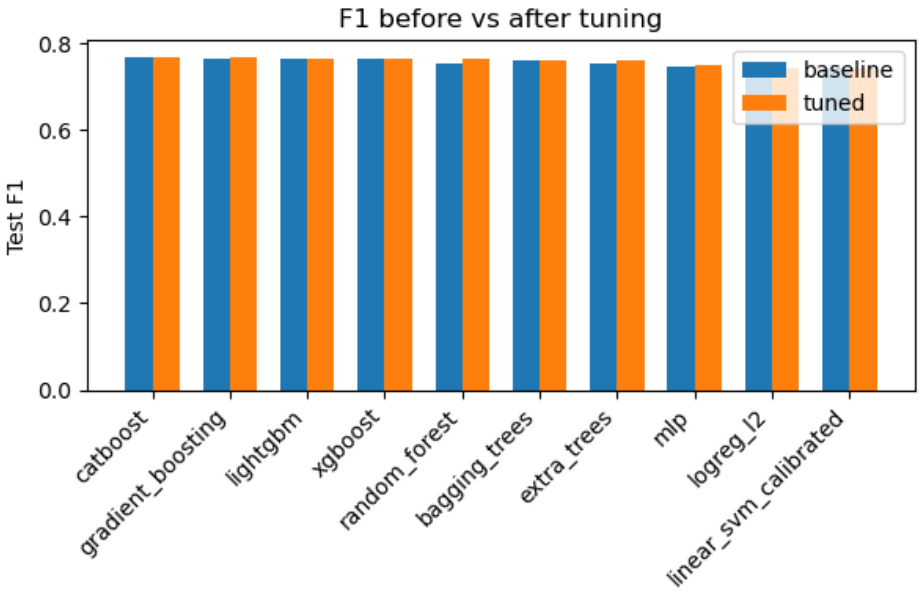


Figure: Test F1: baseline vs tuned.

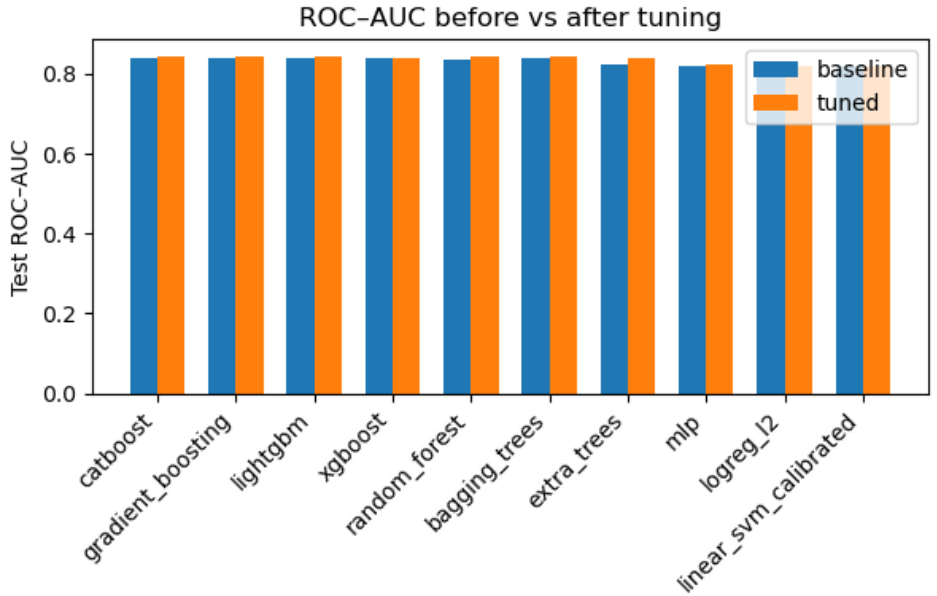


Figure: Test ROC–AUC: baseline vs tuned.

- Tuning done with `RandomizedSearchCV` on training set
- Small but consistent gains (ΔF1, ΔAUC $\approx$ 0.001–0.01)
- Top-ranked models remain the same (CatBoost, Gradient Boosting, LightGBM)

ROC and precision–recall curves

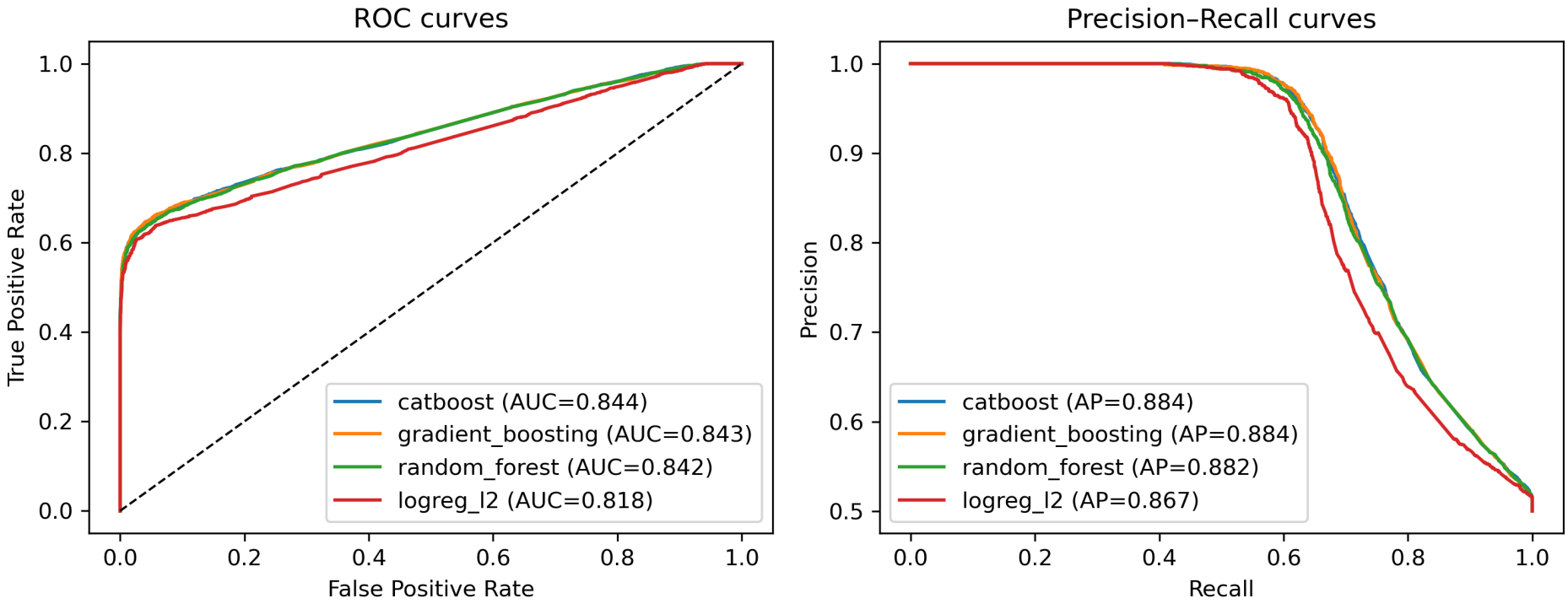
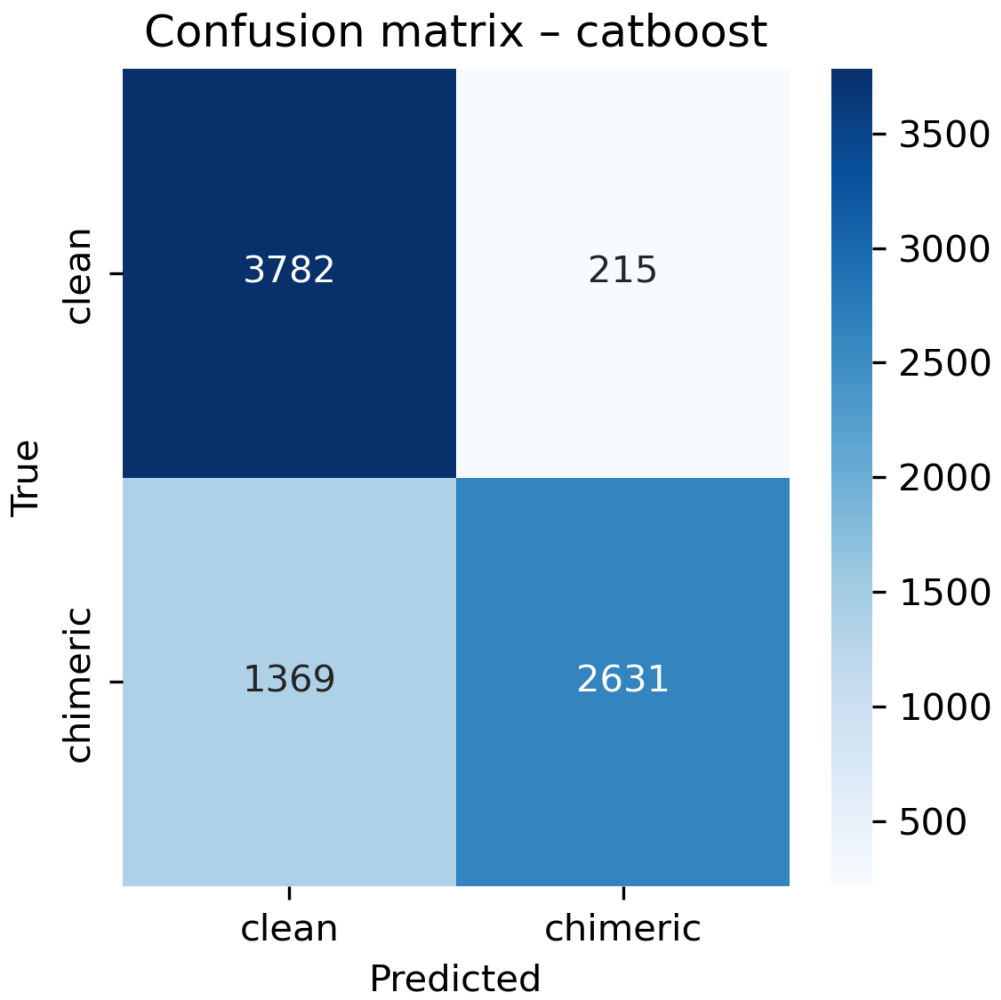


Figure: ROC (left) and PR (right) curves for CatBoost, Gradient Boosting, Random Forest, and logistic regression.

- Ensembles: ROC–AUC ≈ 0.84; logreg: ≈ 0.82
- Average precision ≈ 0.88 for ensembles
- Precision > 0.9 up to recall ≈ 0.5–0.6

Confusion matrix: CatBoost (test set)



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

Figure: Confusion matrix heatmap for CatBoost.

Top features for CatBoost

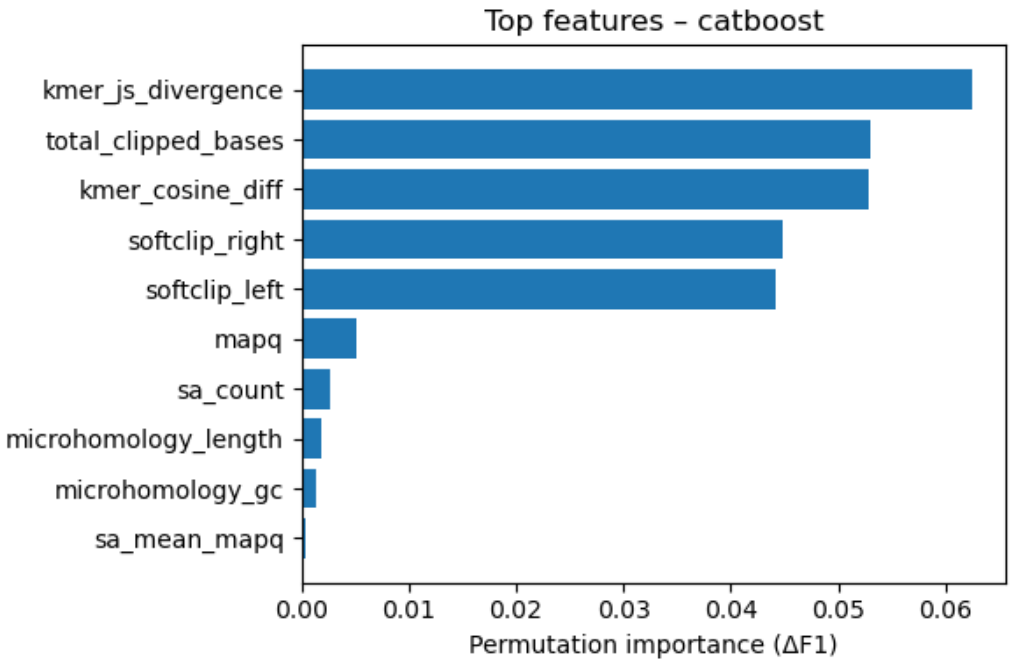


Figure: Permutation importance (ΔF1) for CatBoost.

- 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

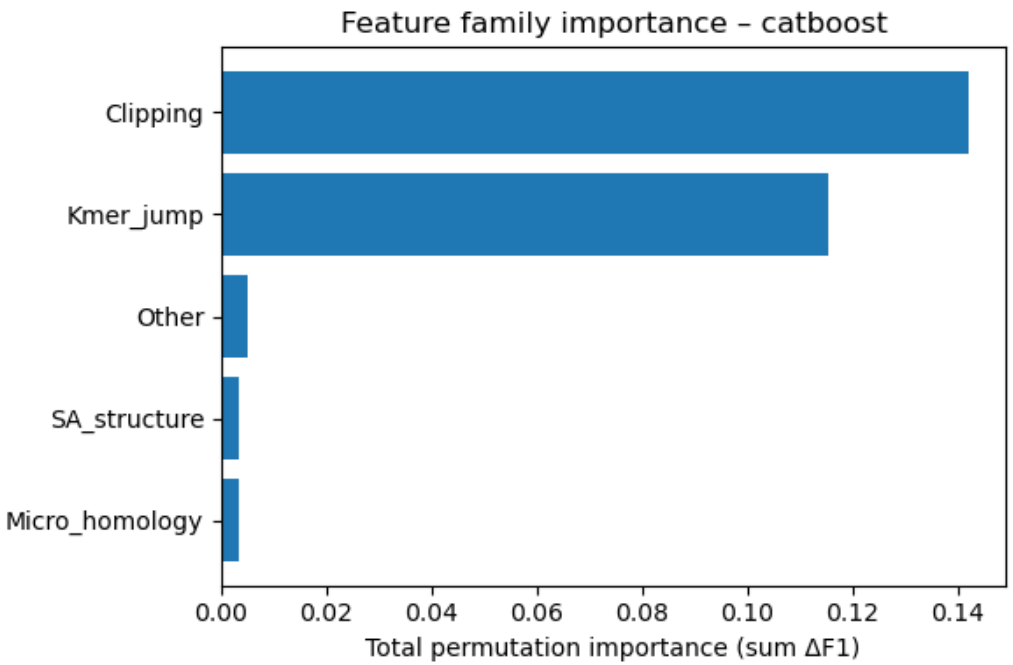


Figure: Aggregated feature families for CatBoost.

- 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

Take-Home Messages

- Tree-based ensembles (CatBoost, Gradient Boosting, LightGBM) consistently outperform linear baselines.
- Best models reach F1 ≈ 0.77 and ROC–AUC ≈ 0.84 on held-out reads.
- Key predictive signals match chimera biology:
  - k-mer composition jumps along the read
  - extensive soft-clipping and total clipped bases
- At the default threshold, the filter is conservative:
  - preserves most clean reads
  - removes a substantial fraction of chimeras
- Overall: our feature set and model ensemble provide a practical pre-filter before mitochondrial assembly.