Machine Learning Pipeline for Detecting PCR-Induced Chimeric Reads

MitoChime: Organellar Chimera Detection from Per-Read Features

Duran, Lin, Pailden

University of the Philippines Visayas
Philippine Genome Center Visayas

December 8, 2025

Outline

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, junction complexity, and split-alignment counts from both clean and chimeric reads
3. Train, validate, and compare supervised machine learning models for classifying reads as clean or chimeric.
4. Determine feature importance and identify indicators of PCR-induced chimerism.
5. Integrate the optimized classifier into a modular and interpretable pipeline deployable on standard computing environments at PGC Visayas.

Scope of the Study

- Focuses on PCR-induced chimeric reads in *Sardinella lemuru* mitochondrial sequencing data to:
  - to limit interspecific variation in mitochondrial genome size, GC content, and repetitive regions so that differences in read patterns can be attributed more directly to PCR-induced chimerism
  - to align the analysis with relevant *S. lemuru* sequencing projects at PGC Visayas
  - to take advantage of the availability of *S. lemuru* mitochondrial assemblies and raw datasets in public repositories such as the National Center for Biotechnology Information (NCBI), which facilitates reference selection and benchmarking
  - to develop a tool that directly supports local studies on *S. lemuru* population structure and fisheries management produce tools applicable to local population and fisheries studies

Scope of the Study

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

Key Exclusions

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

Other Limitations

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

Methodology

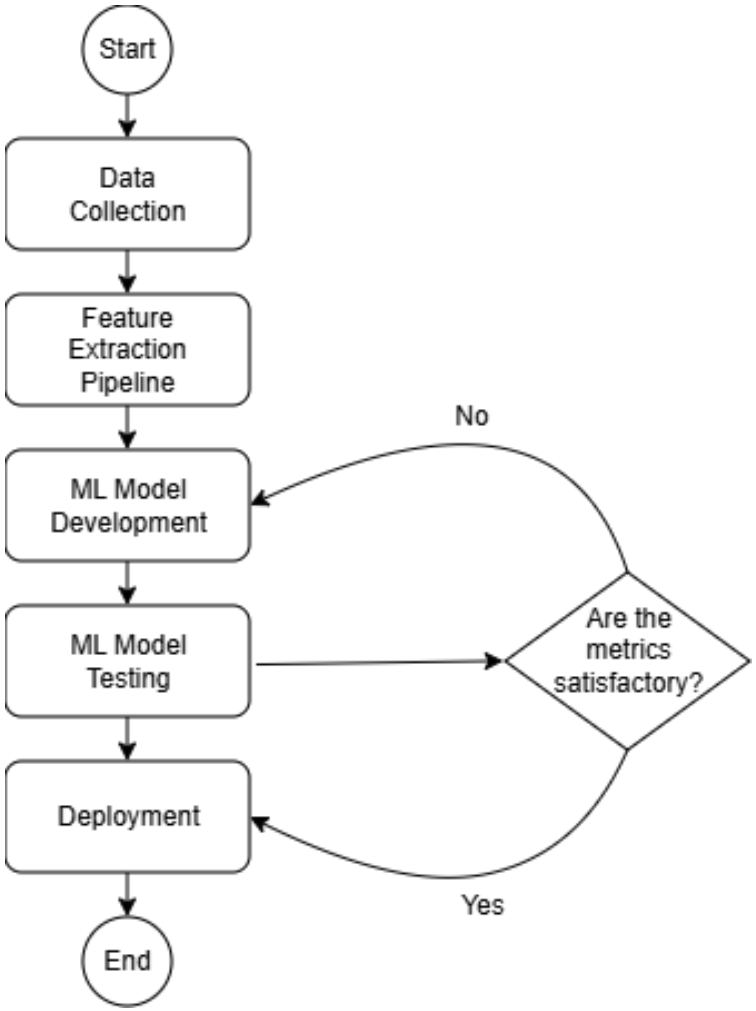

Figure: Process Diagram of the Special Project

Data Collection

The *S. lemuru* mitochondrial reference genome (NCBI: NC_039553.1) was downloaded in FASTA format and used as the basis for generating simulated reads.

Data Preprocessing

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

Data Preprocessing

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

Data Preprocessing

```
NC_039553.1_3_540_8:0:0_6:0:0_ef2       163     NC_039553.1     3       60      150M    =       391     538
        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
        GGTGTAGCTTAAACAAAGCATAACACTGCAGATGATCCGCTGGGCCGTGATAAGCCGCAGCAGGAGTGAAAGTTTTGGTCCAGGCTTTATTATCAGCTTTACCCCAAATTTACACATGCGAGCCTCCGCGGCCCCGTGAGGATGCCCTCAG
        ........................................................................................................................................................ 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

```
        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
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  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
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  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
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  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
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  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
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  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
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  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
        IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  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 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

Stratified Train–Test Split

- First step: **create a held-out test set** for final evaluation.
- Use `build_datasets.py`:
  1. Combine clean and chimeric feature tables.
  2. Attach labels (0 = clean, 1 = chimeric) if missing.
  3. Shuffle and perform **stratified** split:

$$\text{Train} : \text{Test} = 80\% : 20\%$$

     with the same class proportions in each split.
- Output:
  - `train.tsv` (used for model selection and cross-validation).
  - `test.tsv` (kept untouched until the very end).

5-Fold Stratified Cross-Validation

- On the **training set only**, we perform:

$$\text{5-fold stratified cross-validation}$$

- Procedure:
  1. Split training data into 5 folds with balanced 0/1 labels.
  2. For each fold:
     - Train the model on 4 folds.
     - Evaluate on the remaining fold.
  3. Average metrics across the 5 folds:

$$\text{mean F1} \pm \text{std}, \quad \text{mean accuracy} \pm \text{std}$$

- This tells us:
  - **Typical performance** on unseen data.
  - **Stability** of each model (via standard deviation).
  - Helps guide which algorithms are promising before going to the test set.

Model Zoo: Algorithms Compared

- We implemented a panel of 13 classifiers using scikit-learn and gradient boosting libraries:
  - **Baseline**: Dummy (always predicts most frequent class).
  - **Linear models**: Logistic regression (`logreg_l2`), linear SVM with calibration.
  - **Tree ensembles**:
    - Random Forest, Extra Trees.
    - Gradient Boosting (sklearn).
    - XGBoost, LightGBM, CatBoost.
    - Bagging with decision trees.
  - **Others**: k-NN, Gaussian Naive Bayes, shallow MLP.
- All models use the same preprocessing pipeline:

$$\text{Imputer (median)} \rightarrow \text{StandardScaler} \rightarrow \text{Classifier}$$

Hyperparameter Tuning for Top Models

- For the 10 strongest families, we perform **RandomizedSearchCV** with 5-fold CV:
  - Logistic regression, linear SVM (calibrated).
  - Random Forest, Extra Trees, Gradient Boosting.
  - XGBoost, LightGBM, CatBoost.
  - Bagging (trees), MLP.
- Each search explores combinations of:
  - Tree depth, number of estimators, learning rate, subsample ratios, etc.
  - For MLP: hidden layer sizes, regularization (α), learning rate.
- Selection criterion:
  - Choose the hyperparameters with the best **cross-validated F1-score**.
  - Re-fit the best model on the **full training set**, then evaluate on the held-out test set.

Classification Metrics (Per-Read)

- For each model, on the test set we compute:
  - **Accuracy**:
$$\frac{\#\text{ correct predictions}}{\#\text{ all predictions}}$$
  - **Precision** (for chimeras):
$$\frac{TP}{TP + FP}$$
    Of the reads we call "chimeric", how many are truly chimeric?
  - **Recall** (for chimeras):
$$\frac{TP}{TP + FN}$$
    Of all true chimeric reads, how many did we detect?
  - **F1-score** (for chimeras):
$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$
    Harmonic mean: high only if both precision and recall are high.

Threshold-Free Metrics: ROC–AUC and PR Curves

- Our models output a **score** per read (probability of being chimeric).
- By sweeping a threshold on this score, we can draw:
  - **ROC curve**:
    - x-axis: False Positive Rate (FPR).
    - y-axis: True Positive Rate (TPR = recall).
    - **ROC–AUC** = area under the curve.
  - **Precision–Recall (PR) curve**:
    - x-axis: Recall.
    - y-axis: Precision.
    - **Average Precision (AP)** = area under PR curve.
- Intuition for ROC–AUC:

  AUC $\approx 0.84 \Rightarrow$ 84% chance a random chimera is scored higher than a ra

Overall Performance Across Models (Test Set)

Model	CV Acc	CV F1	Test Acc	Test F1	ROC–AUC
Dummy baseline	0.50	0.67	0.50	0.67	0.50
Logistic regression	0.79	0.75	0.79	0.74	0.82
Linear SVM (cal.)	0.79	0.75	0.79	0.74	0.82
Random Forest	0.80	0.77	0.79	0.75	0.83
Extra Trees	0.80	0.77	0.79	0.75	0.82
Gradient Boosting	0.81	0.78	0.80	0.77	0.84
XGBoost	0.81	0.77	0.80	0.76	0.84
LightGBM	0.81	0.77	0.80	0.76	0.84
CatBoost	0.81	0.78	0.80	0.77	0.84
k-NN	0.78	0.75	0.78	0.75	0.81
Gaussian NB	0.75	0.66	0.74	0.65	0.82
Bagging (trees)	0.80	0.77	0.79	0.76	0.84
MLP	0.79	0.75	0.79	0.75	0.82

Table: Summary of cross-validation and test performance (chimeric class F1).

ROC and PR Curves (Placeholder)

ROC curves (CatBoost, GBM, RF, logreg)

Precision–Recall curves

- ROC–AUC for top models: ≈ 0.84.
- Curves pushed towards the top-left / top-right illustrate strong separation between clean and chimeric reads across thresholds.

Confusion Matrix and Class-Wise Behaviour (CatBoost)

Placeholder: confusion matrix for CatBoost on test set

CatBoost (test set, illustrative)

clean: precision ≈ 0.73, recall ≈ 0.95
chimeric: precision ≈ 0.92, recall ≈ 0.66
overall accuracy ≈ 0.80

- **Clean reads**:
  - Very high recall: most true clean reads are correctly kept.
- **Chimeric reads**:
  - High precision: when we call a read chimeric, it is usually correct.
  - Moderate recall: we detect about two-thirds of all chimeras.
- Practical behaviour: **conservative chimera filter** that prioritizes not discarding clean reads.

Effect of Hyperparameter Tuning (F1 and ROC–AUC)

Model	F1 (base)	AUC (base)	F1 (tuned)	AUC (tuned)
CatBoost	0.767	0.839	0.769	0.844
Gradient Boosting	0.766	0.840	0.767	0.843
LightGBM	0.764	0.838	0.766	0.842
XGBoost	0.765	0.839	0.765	0.839
Random Forest	0.755	0.834	0.763	0.842
Bagging (trees)	0.760	0.837	0.763	0.842
Extra Trees	0.753	0.824	0.760	0.837
MLP	0.748	0.819	0.749	0.821
Logistic reg.	0.744	0.821	0.743	0.818
Linear SVM (cal.)	0.744	0.820	0.743	0.818

Table: Test F1 and ROC–AUC before vs after hyperparameter tuning.

- Tuning yields **modest but consistent gains** in F1 and ROC–AUC.
- Confirms that the initial defaults were already reasonable, but performance can be further refined.

Permutation Feature Importance (Placeholder)

Placeholder: permutation importance for CatBoost

- Top features across CatBoost, GBM, RF:
  - `total_clipped_bases`
  - `kmer_js_divergence`, `kmer_cosine_diff`
  - `softclip_left`, `softclip_right`
  - `mapq`
- Interpretation:
  - Chimeras are characterized by **large clipped segments** and **abrupt k-mer composition shifts**.
  - Aligners are already "seeing" the breakpoint signal; the ML model learns to combine these signals into a chimera score.

Summary of Findings

- We built a per-read feature table capturing:
  - Alignment and clipping patterns.
  - Supplementary alignments and breakpoint distances.
  - Sequence-level k-mer divergence and microhomology.
- A broad panel of ML models was evaluated:
  - Tree-based ensembles (CatBoost, Gradient Boosting, Random Forest, LightGBM, XGBoost) achieved the **best performance**.
  - Test F1 for chimeras ≈ 0.76–0.77, ROC–AUC ≈ 0.84.
- Model behaviour:
  - Conservative on clean reads (high recall).
  - High precision on chimeric reads, moderate recall.

Implications for Mitochondrial Assembly

- The ML classifier can be used as a **pre-filter** before assembling mitochondrial genomes:
  - Remove high-confidence chimeric reads to reduce false junctions.
  - Retain the majority of clean reads to preserve coverage.
- Especially useful for:
  - Small, circular, and repetitive organellar genomes where chimeras are particularly harmful.
  - Scenarios without high-quality reference genomes or abundance information.
- The feature importance analysis provides biological insight:
  - Confirms the role of soft-clipping, supplementary alignments, and k-mer jumps as core signals of chimeric structure.

Limitations and Future Work

- Current study uses **simulated** chimeras and a single species:
  - Need to validate on real experimental datasets.
  - Extend to other organellar genomes and library preparations.
- Classifier currently treats each read independently:
  - Future work: incorporate read-pair information, local read depth, or graph features.
- Integration into practical pipelines:
  - Wrap as a command-line tool interfacing with standard BAM/FASTQ workflows.
  - Benchmark impact on final assembly quality (contiguity, misassemblies).

Conclusion

- We developed a **machine learning pipeline** that:
  - Learns from alignment- and sequence-based features.
  - Achieves strong separation between clean and chimeric reads.
- Tree-based gradient boosting models (CatBoost, GBM, RF) provide:
  - High test F1 and ROC–AUC.
  - Interpretable feature importance aligned with known chimera mechanisms.
- This framework is a step towards **reference-free chimera detection** tailored for organellar genomes and low-resource settings.

Thank You

Questions?