A Computational Pipeline for Detecting RNA Modifications from Oxford Nanopore Direct RNA Sequencing Data

Salvador A. Rodarte [1], Eda Koculi [2], Jonathon E. Mohl [1,3]

[1] Computational Science Program, The University of Texas at El Paso, El Paso, Texas.
[2] Department of Chemistry & Biochemistry, The University of Texas at El Paso, El Paso, Texas
[3] Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, Texas.

RNA molecules carry chemical modifications that regulate protein synthesis, yet detecting these modifications computationally remains an open problem. Traditional sequencing platforms digitize RNA into short complementary DNA reads, discarding the chemical signals that distinguish modified from unmodified bases. Oxford Nanopore Technologies (ONT) Direct RNA Sequencing offers an alternative: native RNA strands pass through a protein nanopore, producing a real-time ionic current signal that encodes both sequence and chemical modification state. The challenge shifts from wet-lab detection to a signal processing and classification problem—extracting modification signatures from noisy, high-dimensional current traces.

We applied a multi-stage computational pipeline to ONT direct RNA sequencing data from *Escherichia coli* ribosomal RNA (rRNA). Raw signal data were basecalled using Guppy, aligned to a reference genome with Minimap2, and processed through SAMtools for quality filtering. Per-position error profiles were then computed using EpiNano, which quantifies mismatch, insertion, and deletion rates at each nucleotide. To identify modification sites, we calculated a sum-of-errors metric (ΔX_i) comparing biologically modified RNA against an unmodified in vitro transcribed control, applied Z-score normalization across all positions, and flagged sites exceeding a Z-score ≥ 3 and $\log_2$ fold-change ≥ 2 threshold. This approach detected 16 of 25 known modification sites in 23S rRNA and 9 of 11 in 16S rRNA, including two sites not previously reported by orthogonal methods. Under ribosome assembly stress conditions, three modification sites showed statistically significant changes in signal intensity, demonstrating the pipeline's ability to capture condition-dependent modification dynamics.

The current threshold-based detection has two key limitations: 1) it produces binary calls that miss lower-confidence modification sites, and 2) signal spillover across adjacent modified nucleotides (±3 nt) generates false positives. We are developing a supervised machine learning classifier that uses the per-position feature vector—ΔX_i, Z-score, fold-change, and local 5-mer sequence context—to output per-position modification probabilities rather than binary calls. By training on experimentally validated modification sites as labeled examples, this model aims to improve sensitivity for currently undetected sites while filtering spillover artifacts, producing a generalizable tool applicable beyond ribosomal RNA.