End Reason Filtering in Single Molecule Sequencing of Native DNA by Oxford Nanopore Technologies

Diya Kannappan[1], Monica Holmes[1], Pranjal Srivastava[1], Rachel Case[1], Chloe Ziemelis[1], Nideesh Srivatsavan[1], Jackson Wasylewski[1], Diego Carrillo[1], Yisang Moon[1], Anika M. Patel[1], Gregory A. Farnum[1], Brian Athey[1,2]

[1]Gilbert S. Omenn Department of Computational Medicine and Bioinformatics, [2]Department of Psychiatry, University of Michigan, Ann Arbor

Abstract

Oxford Nanopore sequencing is a single molecule sequencing technology that measures single-molecules of DNA. The signal obtained for each molecule is used to predict the DNA sequence, and each prediction ("read") is stored and annotated with an end-reason that describes how the end of a read was identified from the signal during sequencing. While this metadata is retained in POD5/FAST5 files, it is not carried on in FASTQ/BAM formats. We show that only signal_positive reads reliably represent complete molecules, whereas other classes (e.g., mux_change, unblock_mux_change, signal_negative) indicate that the read was derived from only part of a molecule, producing truncated sequence predictions. Across multiple datasets, these non-signal-positive reads comprised ~3–6% of total reads present in the final BAM output file and were enriched for shorter lengths and atypical quality distributions. These findings support incorporating end-reason filtering into standard nanopore workflows to enhance data accuracy and reliability.

Introduction

Oxford Nanopore sequencing measures changes in ionic current as DNA strands pass through a protein pore. These raw electrical signals are segmented into single-molecule events and then basecalled into nucleotide sequences. Segmentation is critical because it defines the start and end of each read. Metadata about why the end of a read was determined from the signal (end reason) is recorded for each read in pod5 or fast5 file types but are not retained in standard FASTQ/BAM outputs. Our study investigates how retaining and filtering on end-reason metadata can improve data quality by distinguishing reads where signal was generated across the full molecule from truncated reads that did not pass end to end through the pore.

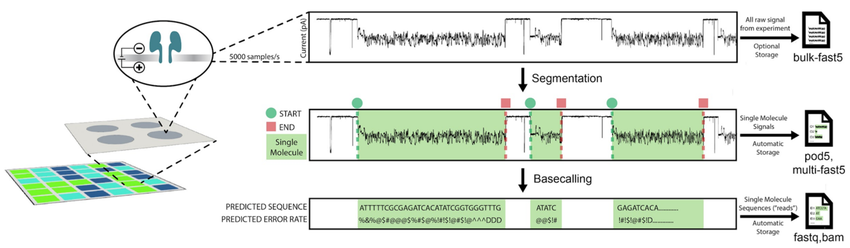

Figure 1. Nanopore sequencing workflow. Continuous ionic current traces are recorded and stored in bulk-fast5 files. Signals are segmented into single-molecule reads (POD5/multi-fast5), each annotated with an end reason. Basecalling converts these reads into nucleotide sequences (FASTQ/BAM), but the end-reason information is discarded, potentially obscuring the distinction between complete and truncated reads. End reasons can be recovered from pod5 or fast5 files and then used to tag the corresponding read in bam files.

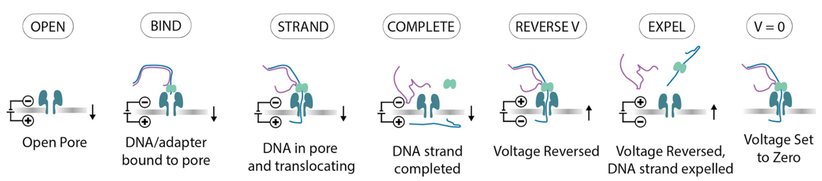

Figure 2. Schematic of nanopore sequencing states. DNA enters the pore (BIND) and translocates through the pore causing a partial blockage of current (STRAND) until completion (COMPLETE) or interruption. Interruptions include software-induced voltage reversal and ejection (REVERSE V, EXPEL) or routine pore scans (V=0). Only COMPLETE states result in signal generation across the full molecule.

End Reasons

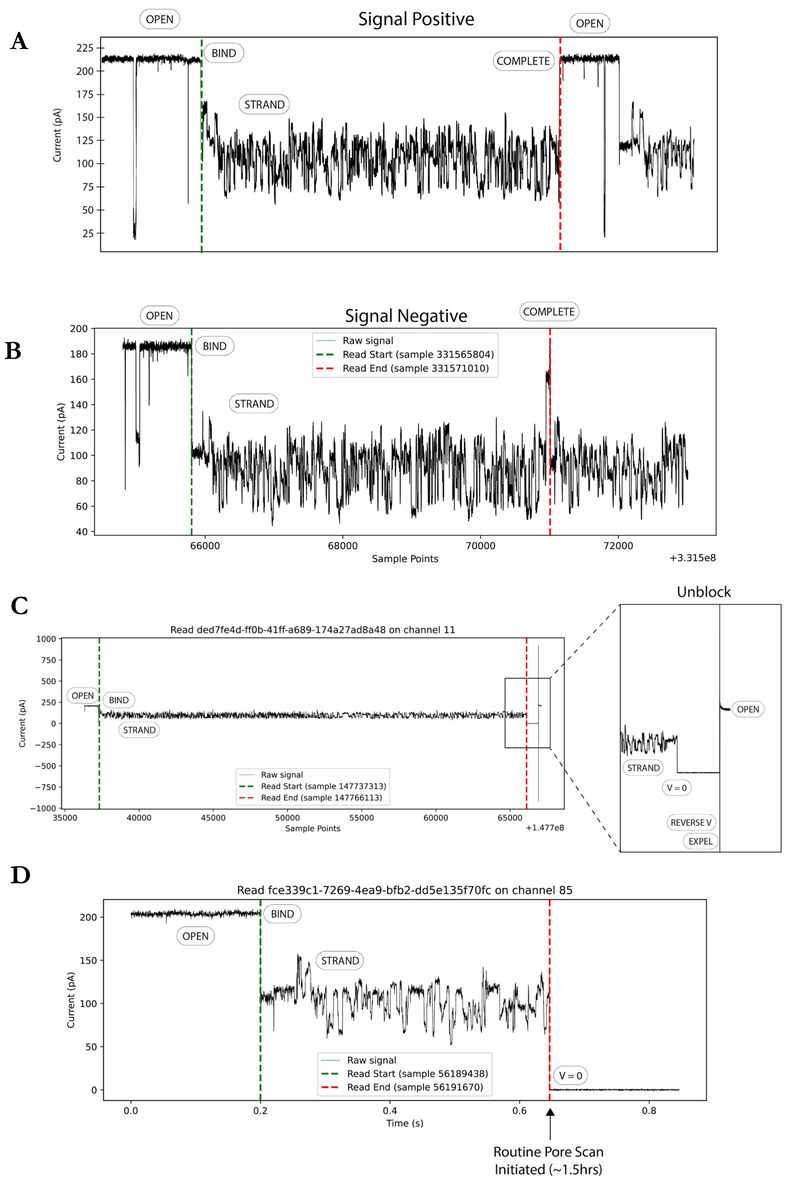

Figure 3. Representative ionic-current traces for nanopore end-reason classes. (A) Signal Positive: Full translocation of DNA strand through pore: current shifts from OPEN → BIND → STRAND → COMPLETE, returning to open current baseline upon end; represents full single-molecule read. (B) Signal Negative: Anomalous ending; uncertain of cause or trend. (C) Unblock MUX Change: Software-initiated voltage reversal (REVERSE V, EXPEL) ejects the DNA strand, producing reads with truncated sequences. (D) MUX Change: Routine pore scan interrupts sequencing, abruptly ending the signal at V=0 without full molecule passage.

Results

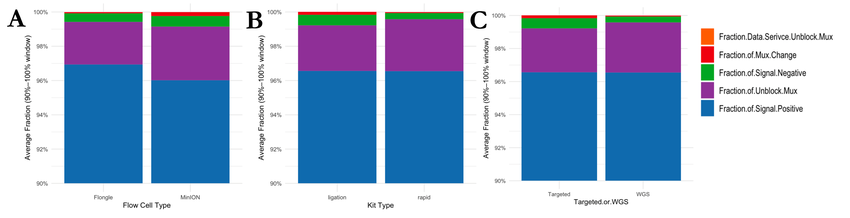

Figure 4. Prevalence of End Reasons Across Diverse Experimental Conditions. Stacked bar charts show the mean fraction of reads for each end reason, grouped by (A) flow-cell type, (B) library preparation kit, and (C) assay type. Data was generated from ~30 sequencing experiments over approximately 1 year. 8 external sequencing runs were obtained from exemplar publicly available data provided by nanopore and other labs. Both internal and external sequencing experiments had similar proportions of each end reason type, with non signal_positive reads constituting a 3-6% minority on average.

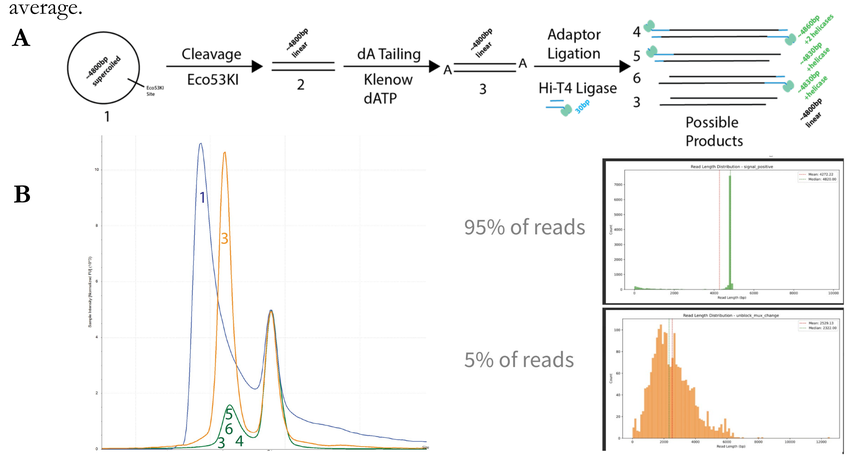

Figure 5. Non-signal positive Reads Exhibit Atypical Length Profiles (A) Plasmid was linearized with Eco53KI (blunt cut at unique site) followed by dA tailing and ligation to ONT-specific sequencing adapters. Possible molecular species are shown, with the sequenceable species shown in green (B) TapeStation Peaks confirm expected size shifts throughout library preparation steps. (C) The distribution of basecalled read lengths for signal positive reads (top panel) shows a sharp peak at the expected read length and represent 95% of all reads present in the final .bam file. The distribution of basecalled read lengths for unblock_mux_change reads (bottom panel) shows a broad range of truncated reads that do not correspond to the single-molecules produced by the library prep.

Discussion and Conclusions

Only signal_positive reads represent full single molecules; others are truncated. Most bioinformatic workflows implicitly assume that the sequence of a read corresponds with the sequence predicted for the full molecule. Filter out reads that have non-signal positive end reasons to avoid erroneous consequences during inference. If you do not filter, then most workflows will need to be modified with additional statistical terms to appropriately model the additionally uncertainty involved in interpreting sequence predictions of truncated reads.

Acknowledgements: Supported by Phenomics Health Inc. (Plymouth Township, MI)