Long-range genomic language models identify extensive chimera artifacts in Nanopore direct RNA sequencing

**Figure 1**A, schematic of LLM

Figure 1B, simulation (internal and terminal), CNN, CNN+transformer, HyenaDNA

Figure 1C, real data VCaP002/VCaP004, detected chimeric reads when applying Dorado and Deepchopper, validation of chimeric read by direct cDNA, (Bar plot, Dorado trim, Dorado w/o trim, Deepchopper), Explain the coverage difference between true events and artifact, we expect true events have high coverage than false positive events, at least in RNA004 and PromethION data.

Figure 1D, E, F: BLAT, quality and PolyA motif for false positive chimeric artifacts. (possible in supplementary)

Figure 1G: other platforms, such as ONT cDNA(direct or PCR), R2C2, CapTrap etc. Testing this using H1 and WTC11 from ENCODE. Test if other platform has less chimeric artifacts predicted by DeepChopper and reduced artifact reads almost cannot be validated by other platform.

(https://www.encodeproject.org/matrix/?type=Experiment&control\_type!=\*&assay\_term\_name=long+read+RNA-seq&status=released)

Figure 2: A,B,C, distribution of false positive across chr, gene expression and size distribution for those with artifacts. Gene GO enrichement suggest Ribosome and ChrM genes are frequently involved.

~~Figure 1J, gene rescue, including the chrM genes.~~

Figure 2D, gene fusion due to chimeric artifacts, highlight by IGV and current signals, for example open pore signal, adapter feature signals!

Figure 2E: Count the percentage of Deepchopper splitted reads account for original chimeric reads and unmapped reads. Count how many of the unmapped reads can be re-mapped after deepchopper split. Could the relative (and/or absolute) coverage increase of genes ranked by these ratio or count.

**Figure 1 Supplementary**

Terminal adapter quality, soft-clipping mapping comparison