The Impact of RNA Quality on Isoform and Fusion Detection by Long-Read Sequencing

Allison Brookhart[1†], Christophe Georgescu[1†], Brian Haas[1], Alexandre Melnikov[1], Houlin Yu[1], Asa Shin[1], Niall J. Lennon[1], Aziz M. Al'Khafaji[1]
[1]Broad Clinical Labs, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA | [†]Corresponding Authors

Background

- Alternative splicing plays a crucial role in natural development and in the progression of many diseases, such as neurodegenerative disorders and cancer. Long-read sequencing enables the capture of full-length isoforms, providing novel insights into the complexity of isoform expression.
- Despite the recent advances in long-read sequencing technologies, sample quality is often a limiting factor in obtaining high-quality transcriptomics data. Clinical samples are often low quality due to prolonged ischemic time prior to sample collection and storage.
- Many samples with low quality RNA still contain valuable information, but it is unclear which samples are sufficiently high quality to sequence. We aim to characterize how RNA Integrity Number (RIN) affects gene and isoform expression in long-read sequencing data.
- We also examine the impact of RNA quality on fusion transcript detection. Gene fusions are common, clinically relevant drivers of cancer that are difficult to detect with short-read sequencing.

Approach

High quality total RNA from 2 cell lines was fragmented in a buffer containing magnesium. RIN was measured using the TapeStation. Reverse transcription with an oligo dT primer was performed in triplicate. Samples were amplified through PCR, underwent library prep (Native Barcoding V14, Oxford Nanopore Technologies), and sequenced to a depth of approximately 4 million reads per sample.

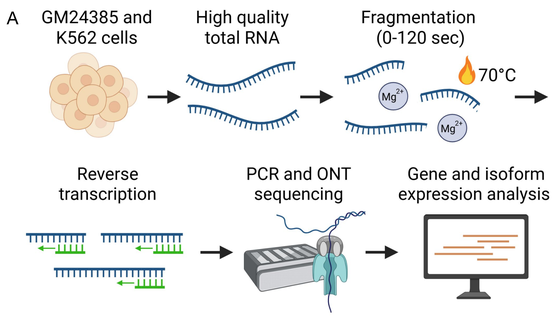

Figure 1. A) Diagram of sample preparation. B) Diagram of sequencing data analysis.

Acknowledgements

Oxford Nanopore Technologies provided reagent support for this work.

Sample Fragmentation

RIN scores ranging from 5-10 were produced by 0-120 sec of fragmentation (2A-B). Fragmentation has a greater impact on long transcripts, and decreases in RIN score are associated with shorter reads (2C).

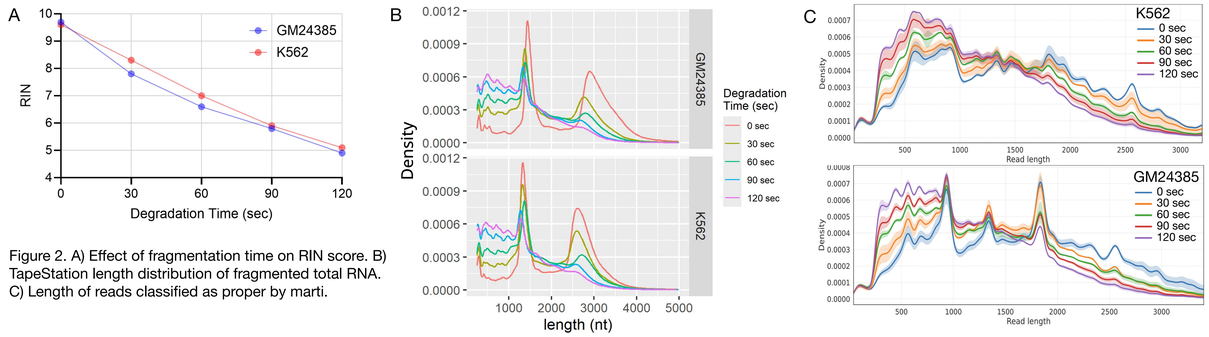

Figure 2. A) Effect of fragmentation time on RIN score. B) TapeStation length distribution of fragmented total RNA. C) Length of reads classified as proper by marti.

Gene and Isoform Detection

As RIN score decreases, the proportion of full splice match (FSM) reads decreases, and incomplete splice matches and novel not in catalog reads increase (3A). Lower quality samples have fewer unique FSM isoforms, and isoform diversity is lost as RIN decreases (3B). Coverage of the 5' end of genes decreases as a function of read length and sample quality (3C).

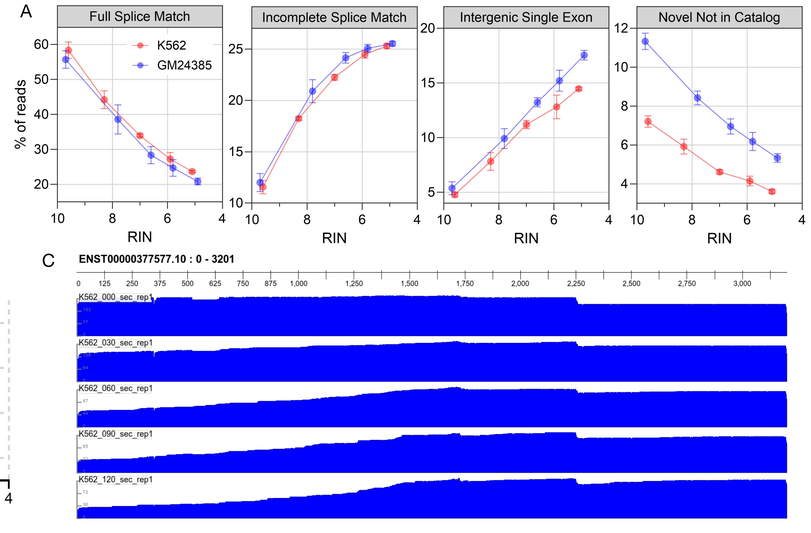

Figure 3. A) LRAA read classifications. B) Number of unique FSM coding isoforms detected by RIN. C) Example of gene coverage in Integrated Transcriptome Viewer (ITV).

Differential Gene and Isoform Expression

Differences in sample quality result in the artificial detection of differentially expressed genes and isoforms (4A). Fragmentation has a stronger effect on observed differential isoform expression (DIE) than differential gene expression (DGE, 4B). At RIN ~8, approximately 500 and 1,500 isoforms are differentially expressed in K562 and GM24385 samples respectively, and at RIN ~5, more than 6,000 isoforms are differentially expressed in each cell line.

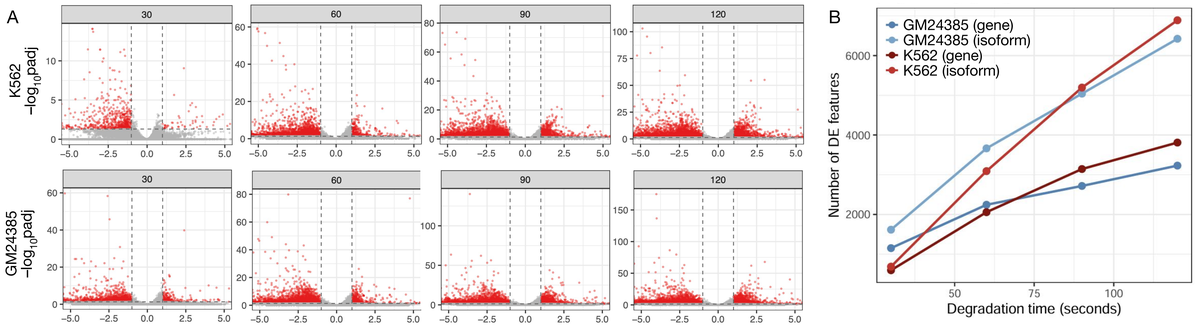

Figure. 4 A) Differential expression of K562 and GM24385 total RNA at different levels of degradation, compared to undegraded RNA. Red represents differentially expressed genes. B) Number of differential expressed (padj <0.05) genes and isoforms.

Fusion Detection

Fragmentation was found to have a significant impact on detection of 2 gene fusions (5A). In lower RIN samples, fewer reads span the fusion breakpoint (5B-C).

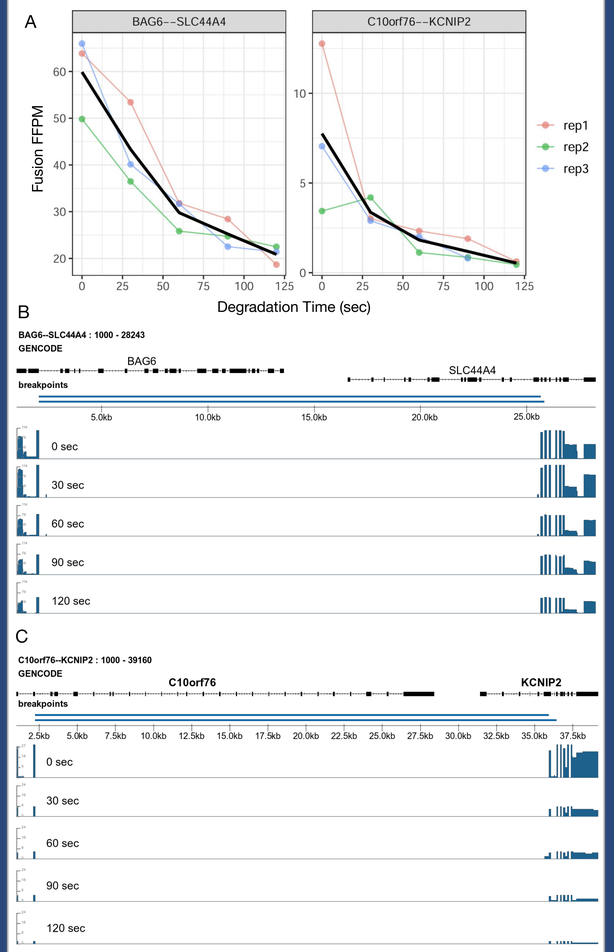

Figure 5. A) Number of fusion fragments per million reads detected using CTAT-LR Fusion. DESeq2 was used to identify two fusions where degradation time had a statistically significant effect on fusion FFPM. Reads aligned to B) BAG6-SLC44A4 and C) C10orf76-KCNIP2 in ITV.

Conclusions

- Detection of thousands of isoforms is lost as RIN deceases.
- Differences in sample quality can cause artificial differences in gene and isoform expression. RIN 7 is considered good quality for short-read sequencing, but in long-read sequencing data, more than 2,000 genes and 3,000 isoforms are differentially expressed at RIN 7 compared to RIN 10.
- Some fusions become more difficult to detect in lower RIN samples, as RNA molecules spanning the fusion breakpoint are fragmented.
- New molecular and computational tools should be developed to control for sample degradation and rescue low quality samples.