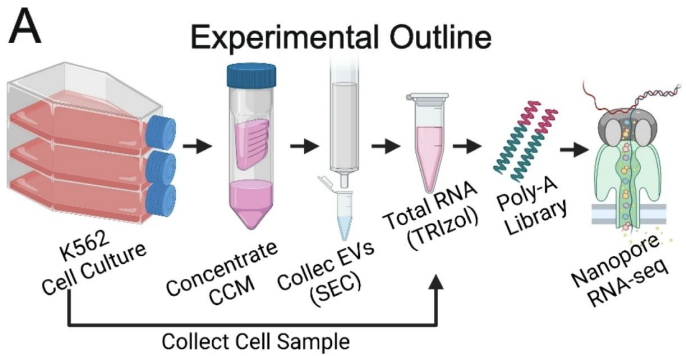
Project 2: Long read genome sequencing

Orlando Yanez, Justine Epiney, Claudia Nguyen

Aim: Align long reads from RNA-seq data to a reference genome

The authors used RNA sequencing with Oxford Nanopore Technology of both extracellular vesicles and whole cells from cell culture.



Padilla *et al. BMC Genomics* (2023) 24:564
https://doi.org/10.1186/s12864-023-09552-6

BMC Genomics

RESEARCH Open Access

Profiling the polyadenylated transcriptome of extracellular vesicles with long-read nanopore sequencing

Juan-Carlos A. Padilla[1,2], Seda Barutcu[1], Ludovic Malet[1], Gabrielle Deschamps-Francoeur[1], Virginie Calderon[1], Eunjeong Kwon[1] and Eric Lécuyer[1,2,3*]

Abstract

Background While numerous studies have described the transcriptomes of extracellular vesicles (EVs) in different cellular contexts, these efforts have typically relied on sequencing methods requiring RNA fragmentation, which limits interpretations on the integrity and isoform diversity of EV-targeted RNA populations. It has been assumed that mRNA signatures in EVs are likely to be fragmentation products of the cellular mRNA material, and the extent to which full-length mRNAs are present within EVs remains to be clarified.

Results Using long-read nanopore RNA sequencing, we sought to characterize the full-length polyadenylated (poly-A) transcriptome of EVs released by human chronic myelogenous leukemia K562 cells. We detected 443 and 280 RNAs that were respectively enriched or depleted in EVs. EV-enriched poly-A transcripts consist of a variety of biotypes, including mRNAs, long non-coding RNAs, and pseudogenes. Our analysis revealed that 10.58% of all EV reads, and 18.67% of all cellular (WC) reads, corresponded to known full-length transcripts, with mRNAs representing the largest biotype for each group (EV = 58.13%, WC = 43.93%). We also observed that for many well-represented coding and non-coding genes, diverse full-length transcript isoforms were present in EV specimens, and these isoforms were reflective-of but often in different ratio compared to cellular samples.

Conclusion This work provides novel insights into the compositional diversity of poly-A transcript isoforms enriched within EVs, while also underscoring the potential usefulness of nanopore sequencing to interrogate secreted RNA transcriptomes.

Keywords Extracellular vesicles, Long-Read RNA Sequencing, Nanopore sequencing, Polyadenylated transcriptome, Poly-A, mRNA, lncRNA, Transcriptomics, Transcript Isoforms, RNA-seq

Have a look at the quality report. What are the average read lengths? Is that expected?

Average read lengths

- Cell_2: 1186.7 bp
- EV_2: 607.9 bp

Both runs are form cDNA. Transcripts are usually around 1-2kb. The average read length is therefore quite short in sample EV_2.

What is the average read quality? What kind of accuracy would you expect?

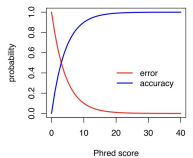
The median base quality is for both around **12** (Cell_2: 11.8 and EV_2: 11.5).

This means that the error probability is about 10^(-**12**/10) = 0.06, so an **accuracy of 94%**.

fasta + basequality (fasta + q = fastq)

$$BASEQ = -10log_{10}\Pr\{base\ is\ wrong\}$$

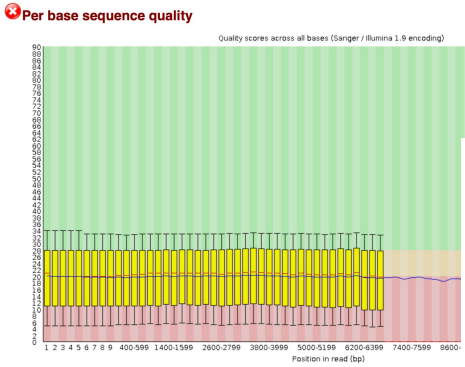
$$-10log_{10}(0.01) = 20$$
$$-10log_{10}(0.1) = 10$$
$$-10log_{10}(0.5) = 3$$

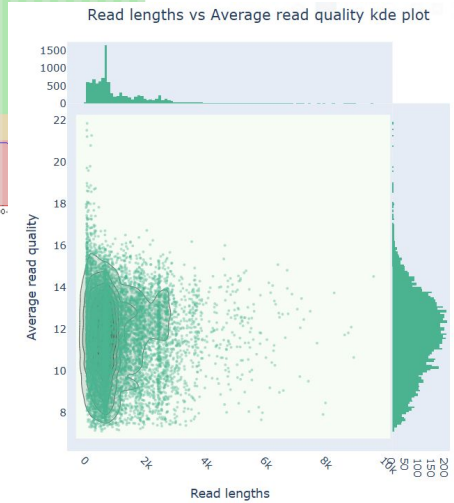


Note any differences between `fastqc` and `NanoPlot`? How is that compared to the publication?

y-axis scales automatically with length of x-axis (long-read)

→ Indicating bad phred score, although 12 is good





fastqc: Focussing more on general output of seqdata

NanoPlot: seq_parameters interesting for ONT sequencing (num_reads, yield, distributions, etc..)

Check out the option `-x` of `minimap2`. Are the defaults appropriate?

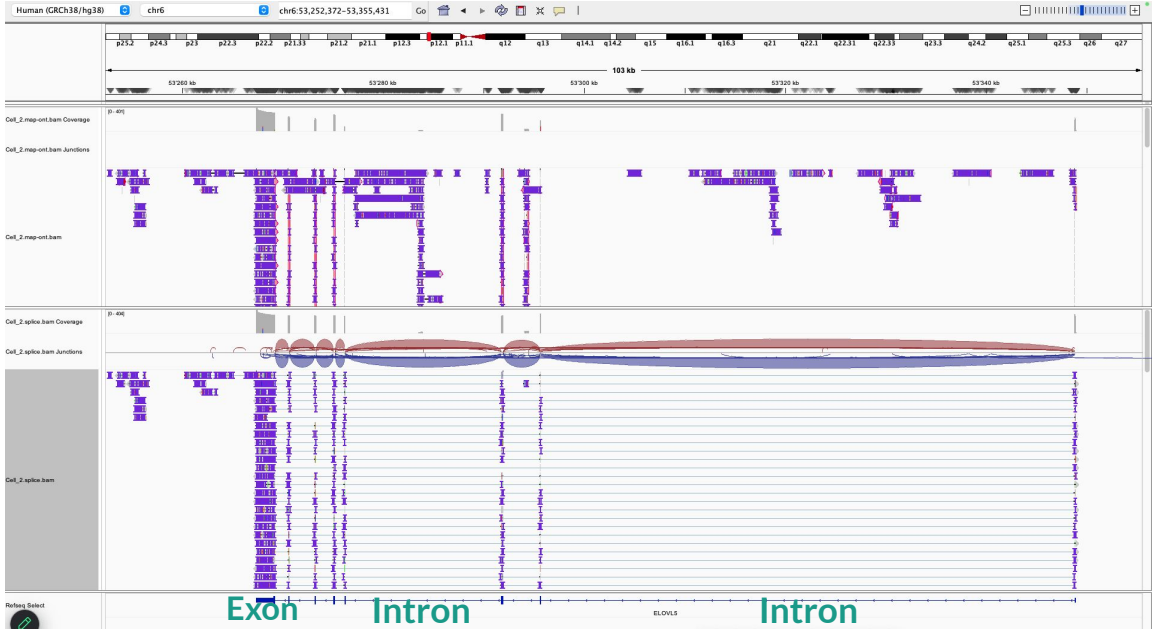
```
Preset:
  -x STR        preset (always applied before other options; see minimap2.1 for details) []
                - lr:hq - accurate long reads (error rate <1%) against a reference genome
                - splice/splice:hq - spliced alignment for long reads/accurate long reads
                - asm5/asm10/asm20 - asm-to-ref mapping, for ~0.1/1/5% sequence divergence
                - sr - short reads against a reference
                - map-pb/map-hifi/map-ont/map-iclr - CLR/HiFi/Nanopore/ICLR vs reference mapping
                - ava-pb/ava-ont - PacBio CLR/Nanopore read overlap
```

splice: spliced alignment for long reads

You might consider using `-x map-ont` or `-x splice`. Do you see differences in the alignment in e.g. IGV?

-x map-ont

-x splice



-x map-ont: Errors in the alignment are observed. Read sections aligned in the Intron regions

How are spliced alignments stored in the SAM file with the different settings of `-x`?

In x- splice, the splice alignments are stored in the CIGAR string as a N

CIGAR symbol	Meaning
`M`	Match (aligned bases)
`N`	**Skipped region** (typically an intron!)
`S`	Soft clip
`H`	Hard clip
`D`	Deletion
`I`	Insertion

with x- map-ont, which is not expecting introns, these regions are considered a structural variant (eg soft clip, deletion, poorly mapped …)

We can find this in our SAM file by checking for the “N” in the cigar string

In the case of the cells we find

- 37704 N = skipped region in x-splice
- 0 N for x- mapont

How deep is the gene ELOVL5 sequenced in both samples?

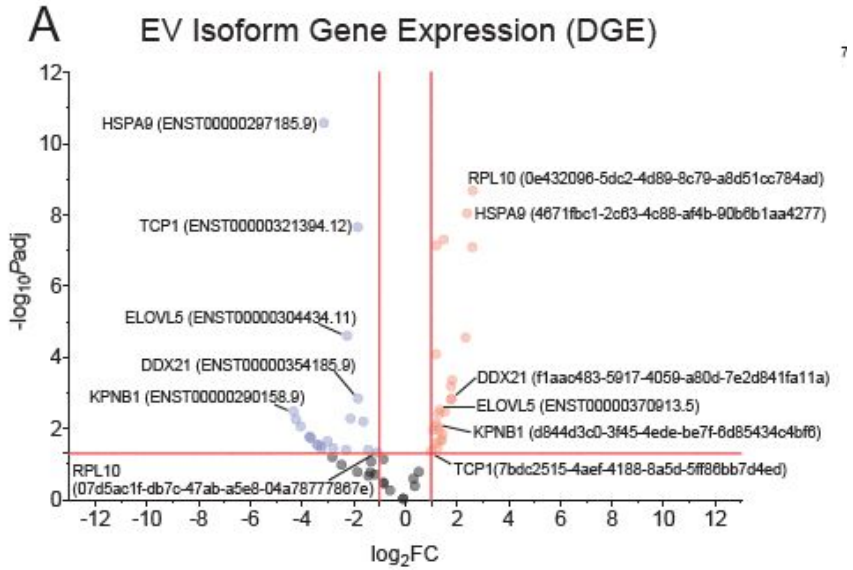
Mean depth:

- **Cell_2:** 11.624
- **EV_2:** 0.365

group5 > scripts > $ 08_sequencing_depth.sh

```
#!/usr/bin/env bash

cd /group_work/group5/project2/

mkdir -p results

for sample in EV_2 Cell_2; do
    # Compute mean depth from samtools depth output
    samtools depth -r 6:53265404-53350950 alignments/"$sample".splice.bam \
    | awk '{sum+=$3} END {if (NR>0) print sum/NR; else print 0}' \
    > results/"$sample".mean_depth.txt
done

```

ELOVL5 shows a shift in isoform representation in EV versus cellular specimens



For some genes, EV specimens display differential recruitment of transcript isoforms relative to their cells of origin

Do you already see evidence for splice variants in the alignments?

Cells

EVs