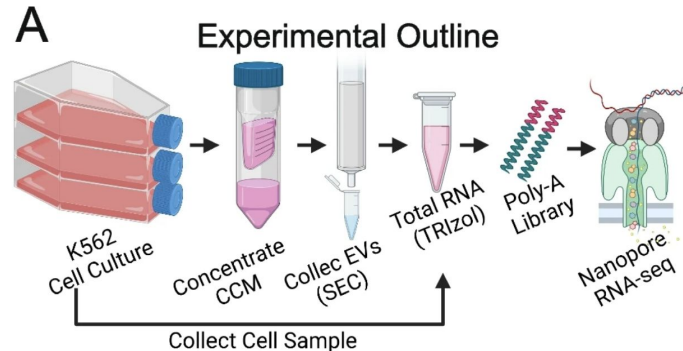


Project 2: Long read genome sequencing

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 Barutcu¹, Ludovic Malet¹, Gabrielle Deschamps-Francoeur¹, Virginie Calderon¹, Eunjeong Kwon¹ 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 from 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?

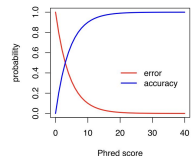
The median base quality is for both around **12** (Cell_2: 11.8 and EV_2: 11.5).

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

fasta + basequality (fasta + q = fastq)

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

$$\begin{aligned} -10\log_{10}(0.01) &= 20 \\ -10\log_{10}(0.1) &= 10 \\ -10\log_{10}(0.5) &= 3 \end{aligned}$$

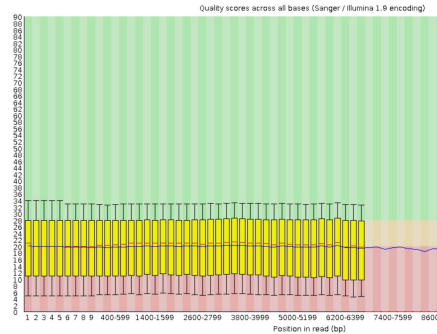


Note any differences between **fastqc** and **NanoPlot**? How is that compared to the publication?

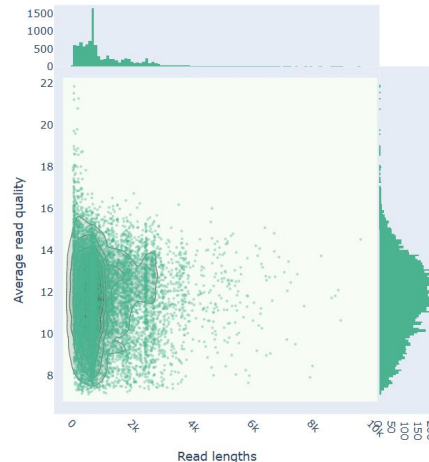
y-axis scales
automatically
with length of
x-axis (long-read)

→ Indicating bad
phred score,
although 12 is
good

Per base sequence quality



Read lengths vs Average read quality kde plot



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

```
Example spliced reads in -x splice:
bbf69cb6-1046-4502-be9a-84e1a0e19bc0 10S20M2I24M1I43M1I10M1I7M2D10M5D34M1D13M1I72M2D14M2I31M3I13M2I4M1D15M1D20M1D8M1D76M1I52M1D22M2D16M3D48M1D37M1I21M1D16M2I50M1I39M2D103M3I76M
1D48428N26M1I48M4D89M2D8M1D4M1I1M1I32M1I79M1D50M1I1M1I1M2I64M2D3M1D41M1D12M1I26M2D9M1D5M1D3M1I1M1I28M2I13M1D32M2I24M1I4M1D21M2I6M1D60M1I88M1D7M1D22M1D11M1I59M2I17M6D16M3I32M1D1
7M1D30M1D23M1D3M1D21M1D58M1I3M1D2M2D15M1I24M1I10M1I6M1I13M3I12M1D66M1D15M1I3M3I13M2I3M1I6M2D18M1D45M3I13M1I3M1D49M1I19M1I29M1D34M5I22M1I24M2I29M1D2M1I8M1D2M1I4M1D6M3I19M1I24M2D
2M1D21M1D32M1D14M1D16M1I38M4D7M1I6M1D7M2D29M1I21M2I38M1I3M4I5M1D10M2I19M1D8M1D2M2D13M4D9M1I30M1D30M1I10M1I11M1D42M1D19M2I17M1D26M3D4M4D6M2D3M1D25M1I7M2I46M1D21M2D1M1I2M1D5M1D5M
2D8M3I42M2D7M3D3M1I50M1D27M1D15M2I150M7D5M2D22M1D1M3I3M1D16M1D23M3D27M1D5M1D4M1D59M2D36M1D15M1I7M1D13M1I4M1I1M2D8M1D5M1I10M2D10M1D73M1I40M1D13M1D4M1D21M1I4M1D11M2D38M1D1M1
D7M1D8M1I10M2I1M1D7M1D87M1I9M1D6M1D4M1I7M1I3M1D3M2I62M5D58M1D9M1D34M2D30M2D10M1I34M3I14M1D14M2D10M2D8M1D39M1D14M6D6M1I2M1D40M2I7M2D24M1D41M1D44M2I29M1I23M1I6M1I25M1I37M2I2M2I34
M1I83M3I41M1D60M2I60M3I12M2I8M3D1M1D42M1D86M1D22M2D9M1I5M1I7M1I32M3D7M2I18M1I30M1I32M3D2M2I46M2D20M1I31M1I22M1D11M1I40M3I51M1D6M2D70M2D33M2I3M4I1M1I26M3D1M1D36M1D20M1D12M1I79M4
D5M2D6M1D2M2I80M1D30M1D12M1D39M1D14M3D26M
c19909f9-6f1b-4a62-9f6d-933ebc4a7784 24S12M1D11M2D12M244I101N72M6D31M2I62M1D25M1D21M3D2M2I6M1D37M2D40M7D17M1D5M1D7M1I3M1D11M4877N1I26M2D4M3D21M1I16M2I1M1D17M1S
75b31424-6efd-457a-baaf-9a72ecb2e57f 17M1I3M1I18M1I17M1I12M1I6M1D3M1D13M2D8M3I17M1D4M1D4M1D45M2I34M3I49M1D7M1I12M2I1M8172N8M1D23M4I7M1I15M1D22M1D7M
7b843d09-29ca-40fe-a9c8-e850d8dca0bf 7M1D7M3I96M1D33M8I22M1D21M1I7M2D32M1I27M2I1M3D8M5278N2M1D9M1I31M1D8M1D16M1I2M1D10M2I11M2I63M1I17M1I2M1D10M1D5M1D4M2D22M1D6M1D41M1I51M1D41M1
I25M1I19M2D8M1I41M1D13M1I56M1D16M1I10M1I6M1D12M1I30M1I33M5I36M1D3M1I72M3D59M1I24M1I4M1D11M1D35M1D12M1D5M2I7M1D7M1D2M1I6M1D13M1D4M1D4M1D2M2D21M1D25M1I37M1I20M6D4M1I12M1I21M2I4M2
D19M1D27M1I14M1D14M2I41M1I14M2I41M1D3M1D16M1I46M1I56M1D10M2I6M1I6M1D5M1D5M2I45M3D6M1D34M1I79M1I30M2D13M1I5M1D4M3D2M1D54M2D45M2D61M1D4M2I18M1D2M1D30M1I15M2I17M2I31M1D2M1D27M2I19
M2D1M1D7M1D8M1D40M1I20M4D26M1D5M3D26M1D8M2I7M1D6M2D4M1D8M1D22M3D31M1D8M2I3M1D10M2D25M1D14M1D7M1D7M1I46M2D15M3D8M2D19M1D19M1D23M1I1M2D37M1D13M2D26M
efalbc43-f493-4706-93f8-e796772e0ed2 17M1D47M1D43M1D78M1I16M4D63M2I3M5D2M1D1M3D4M1D8155N7M1D33M1I2M1I19M1D90M3I52M1D6M1D13M2D6M1I26M
```

```
Example spliced reads in -x map-ont:
```



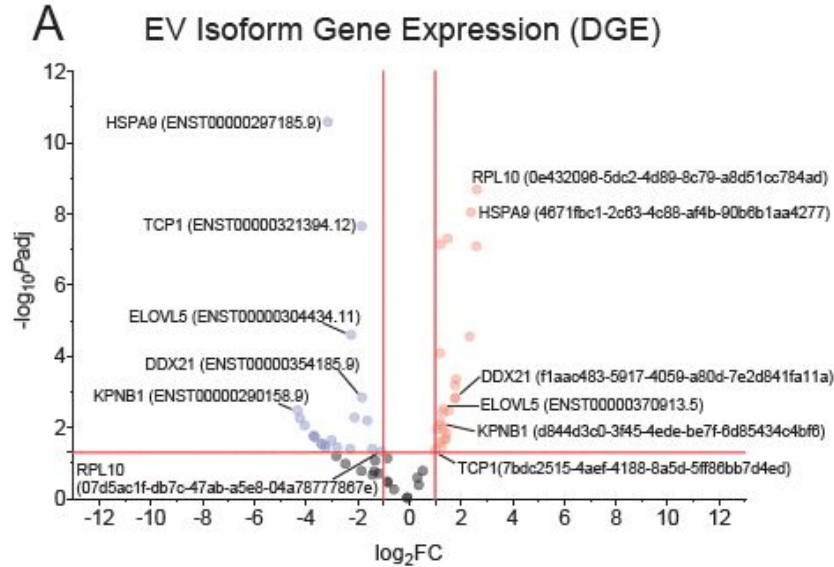
How deep is the gene **ELOVL5** sequenced in both samples?

Mean depth:

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

```
group5 > scripts > $ 08_sequencing_depth.sh
1  #!/usr/bin/env bash
2
3  cd /group_work/group5/project2/
4
5  mkdir -p results
6
7  for sample in EV_2 Cell_2; do
8      # Compute mean depth from samtools depth output
9      samtools depth -r 6:53265404-53350950 alignments/"$sample".splice.bam \
10      | awk '{sum+=$3} END {if (NR>0) print sum/NR; else print 0}' \
11      > results/"$sample".mean_depth.txt
12  done
13
```

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

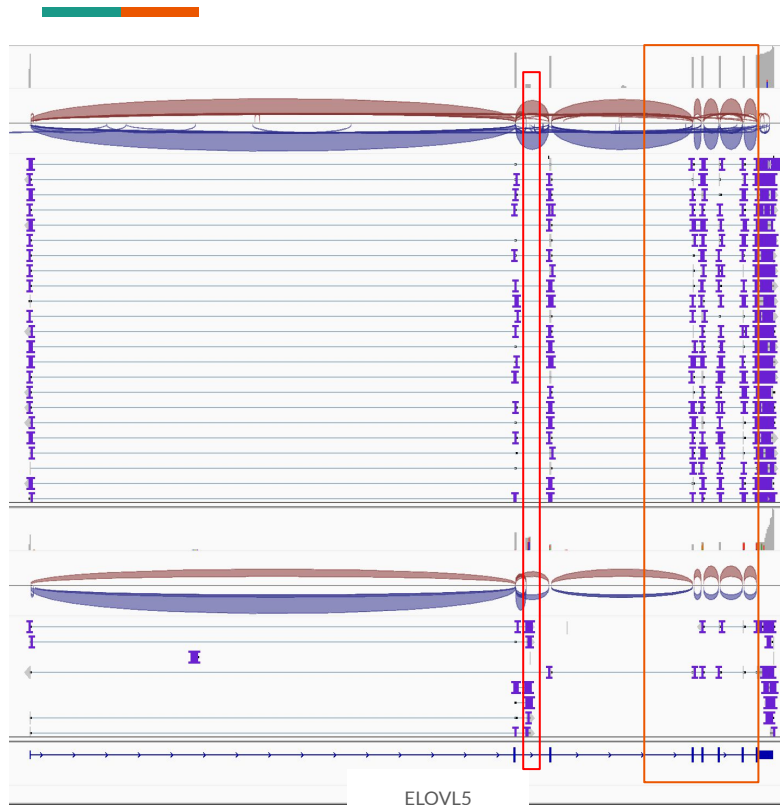


For some genes, EV specimens display differential recruitment of transcript isoforms relative to their cells of origin

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

Cells

EVs



B ELOVL5 Read and Transcript Isoform Overview

