Project 2 Group 2 Result

Original paper and data sets

Research | Open access | Published: 22 September 2023

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

Juan-Carlos A. Padilla, Seda Barutcu, Ludovic Malet, Gabrielle Deschamps-Francoeur, Virginie Calderon, Eunjeong Kwon & Eric Lécuyer

BMC Genomics **24**, Article number: 564 (2023) | Cite this article

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.

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

- Cell_2: 1190 bp (Median is 749 bp)
- EV_2: 602 bp (Median is 499 bp)

→ These read lengths are relatively short for ONT sequencing, but still longer than Illumina sequencing

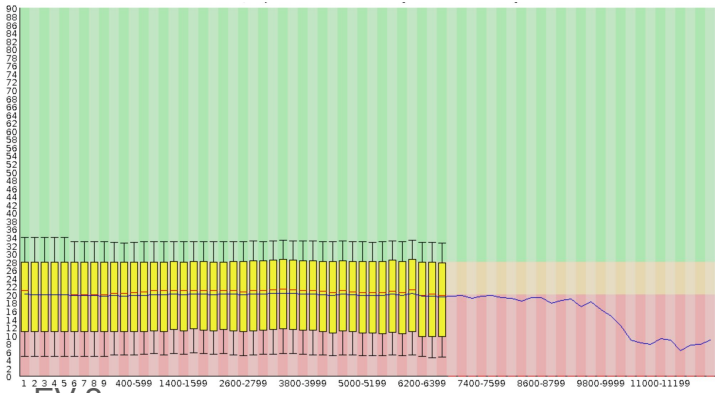
What is the average read quality? What kind of accuracy would you expect?

- Cell_2: 16.9851
- EV_2: 16.977

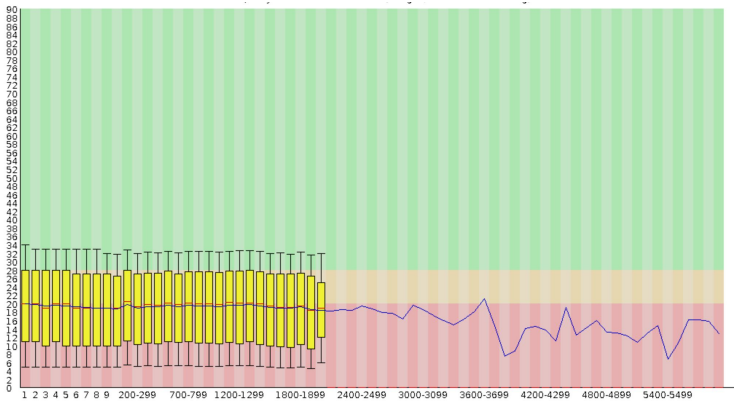
→ An average quality score of ~17 is within the expected range for raw ONT reads (base call accuracy of 98%)

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

Cell 2



EV 2



Mean read quality:
fastqc_cell **16.98** > nano_cell: **11.2**
fastqc_ev: **16.97** > nano_ev: **11.1**

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

minimap2 -x

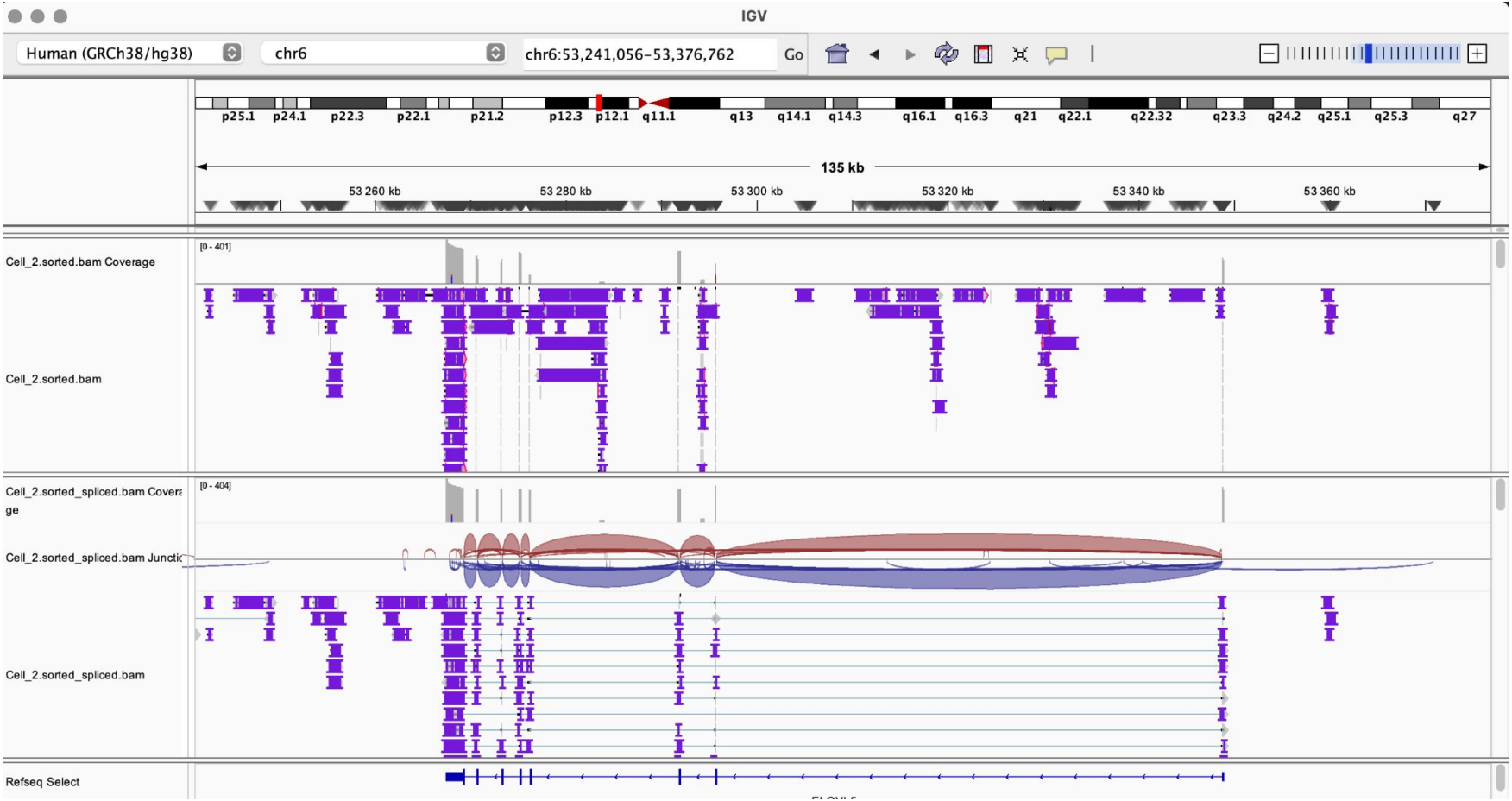
Nanopore DNA reads → `-x map-ont`
Long-read RNA-seq (Nanopore or PacBio)--> `-x splice`

Preset:

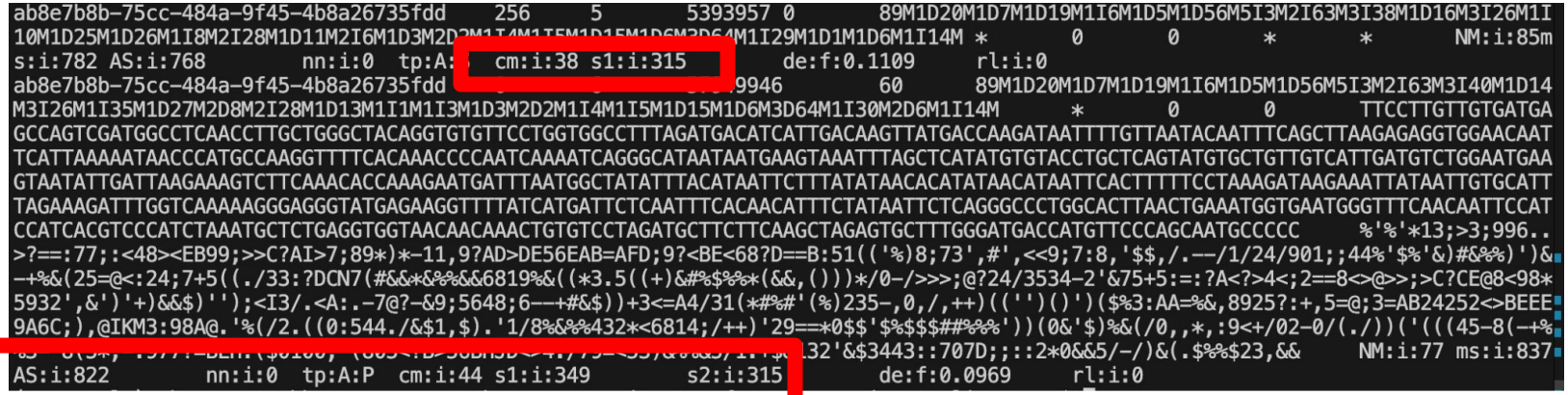
-x STR preset (always applied before other options; see minimap2.1 for details)

- 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

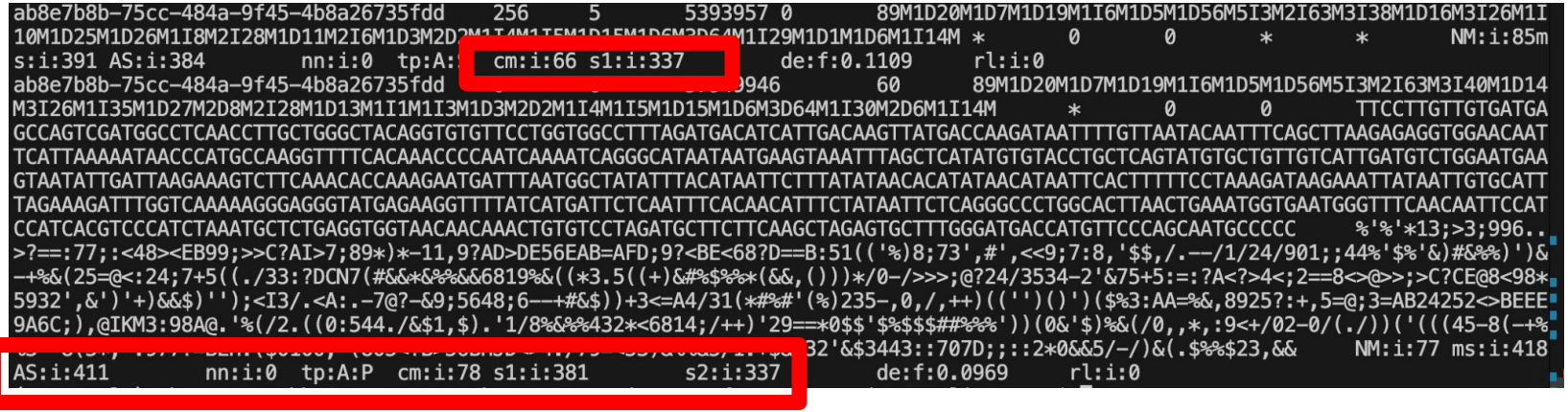
You might consider using `-x map-ont` or `-x splice`. Do you see differences in the alignment in e.g. IGV?



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



-x map-ont



-x splice

CIGAR Strings differences

```
samtools view Cell_2.sorted.bam | awk ' {print $6}' | grep N

samtools view Cell_2.sorted_spliced.bam | awk ' {print $6}' | grep N
```

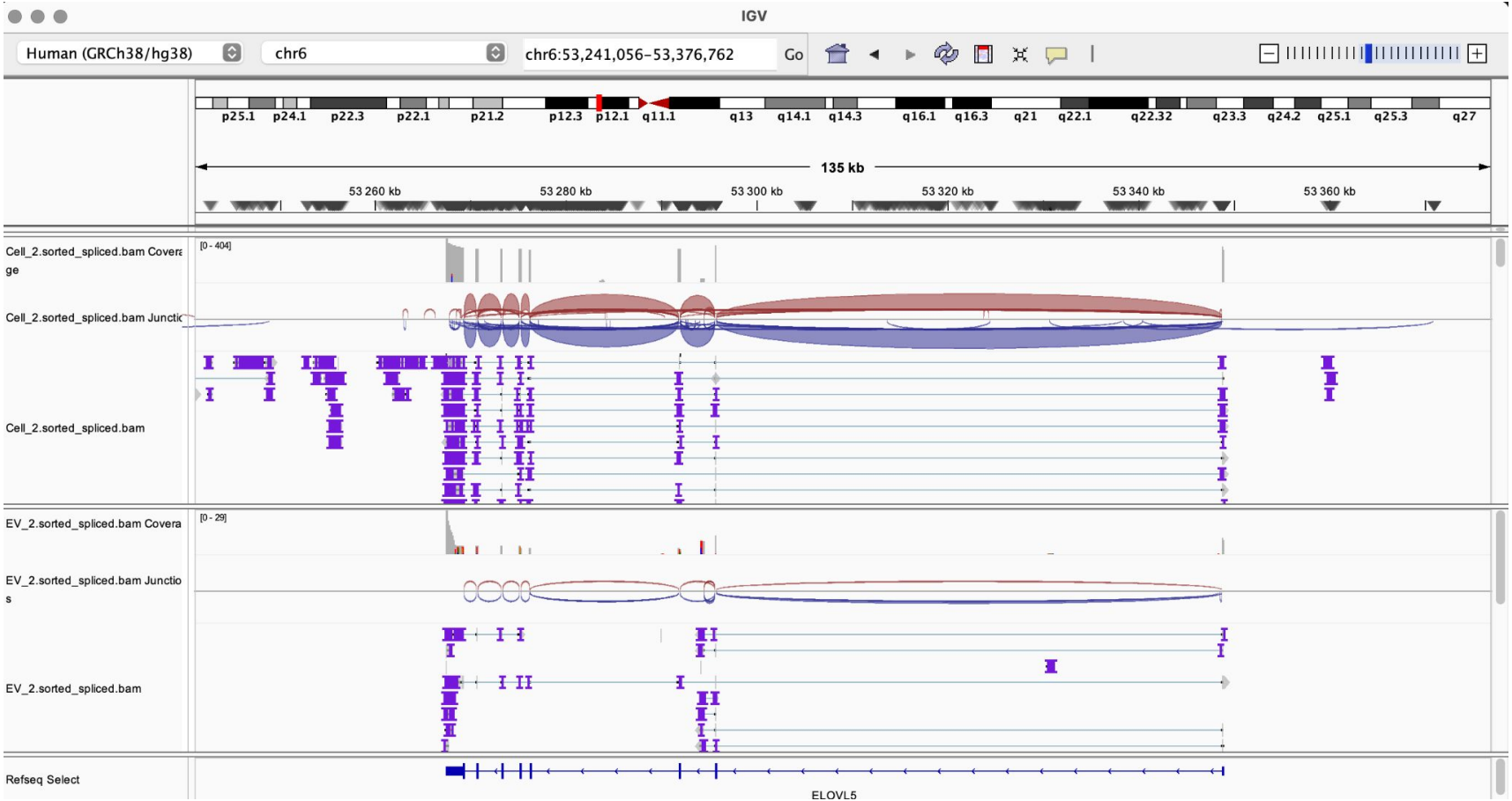
CIGAR with letter N only present in Cell_2.sorted_spliced.bam file

CIGAR strings

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

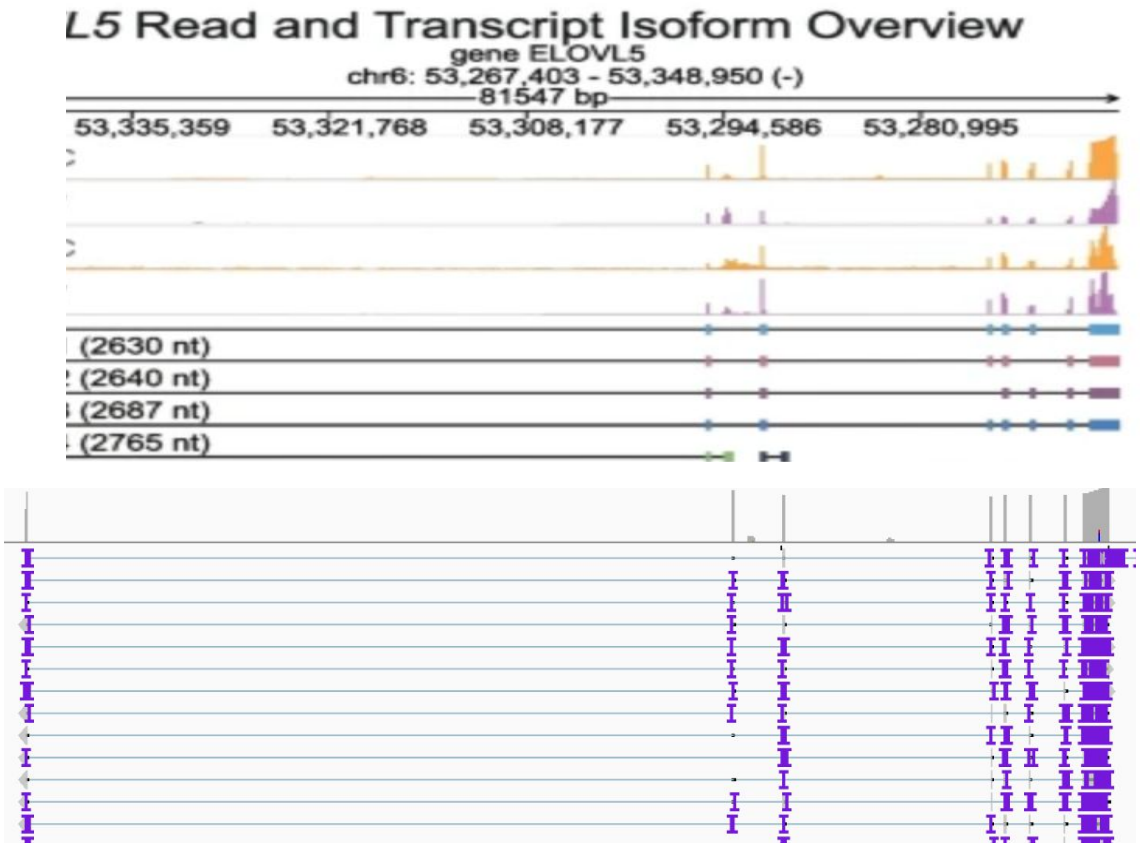
Almost never used

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



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

Supplementary materials


L5 Read and Transcript Isoform Overview
gene ELOVL5
chr6: 53,267,403 - 53,348,950 (-)
81547 bp
53,335,359
53,321,768
53,308,177
53,294,586
53,280,995
(2630 nt)
(2640 nt)
(2687 nt)
(2765 nt)

Interpretation of bam files generated by minimap2

High mapping rate:

- **Cell_2**: 99.67% mapped
- **EV_2**: 99.14% mapped
  That’s excellent and suggests reads are aligning well to the reference genome.

Primary alignment rate is also high:

- Over 98% of the primary reads are mapped in both samples. This is usually what you'd look at for downstream analyses.

No duplicates detected – which is typical for long-read data.

Interpretation of bam files generated by minimap2

1. High proportion of supplementary alignments:

- **Cell_2**: 66,342 supplementary reads
- **EV_2**: 2,793 supplementary reads
  Supplementary alignments are often split reads (e.g., structural variants, chimeras, or alignment across exon junctions). In long-read data, this is **normal**, but if it's unusually high (like in Cell_2), it might warrant a look.

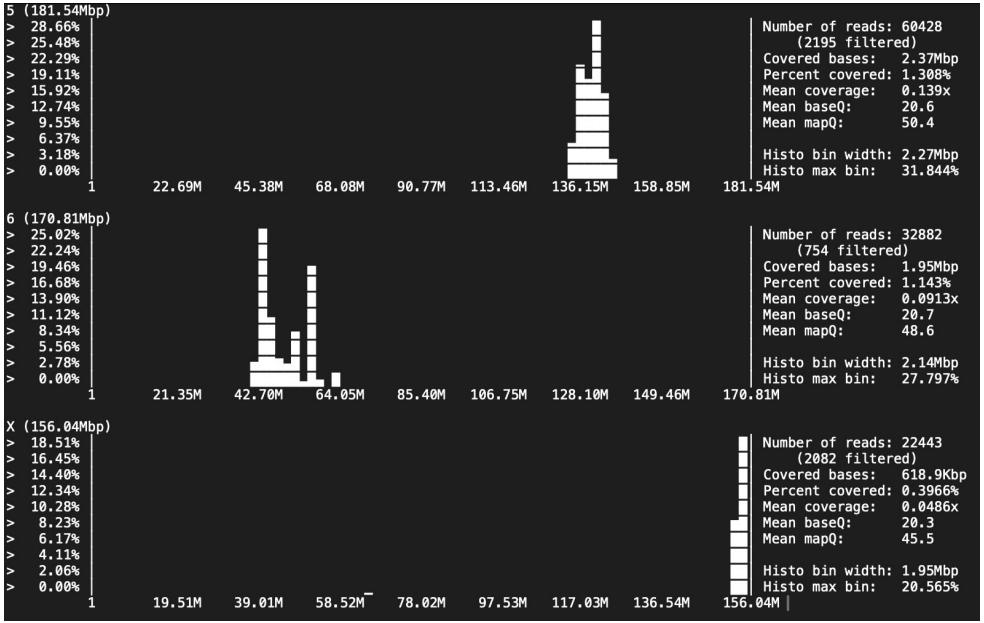
2. Total number of reads differs greatly between samples:

- **Cell_2**: ~121k reads
- **EV_2**: ~9.7k reads

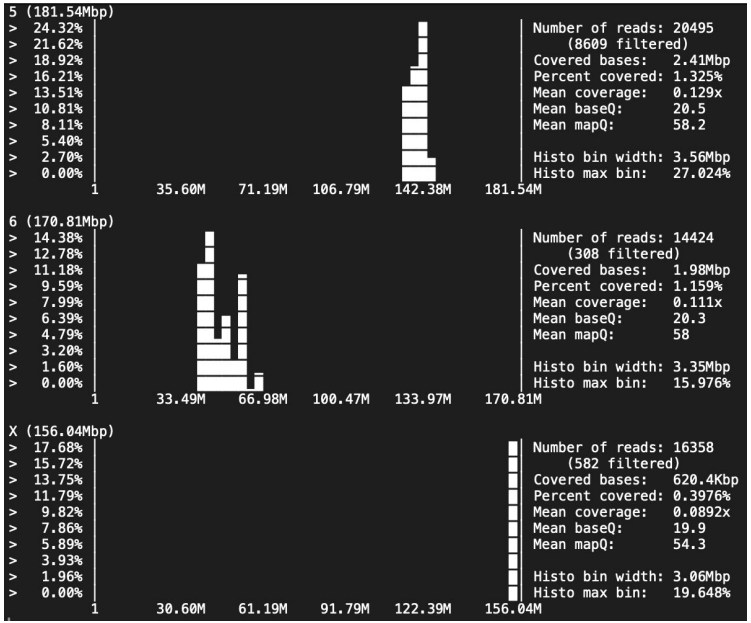
Check out the option `-x` of `minimap2`. Are the defaults appropriate?
You might consider using `-x map-ont` or `-x splice`. Do you see differences in the alignment in e.g. IGV?
How are spliced alignments stored in the SAM file with the different settings of `-x`?

Whole cell_2 coverage

map-ont

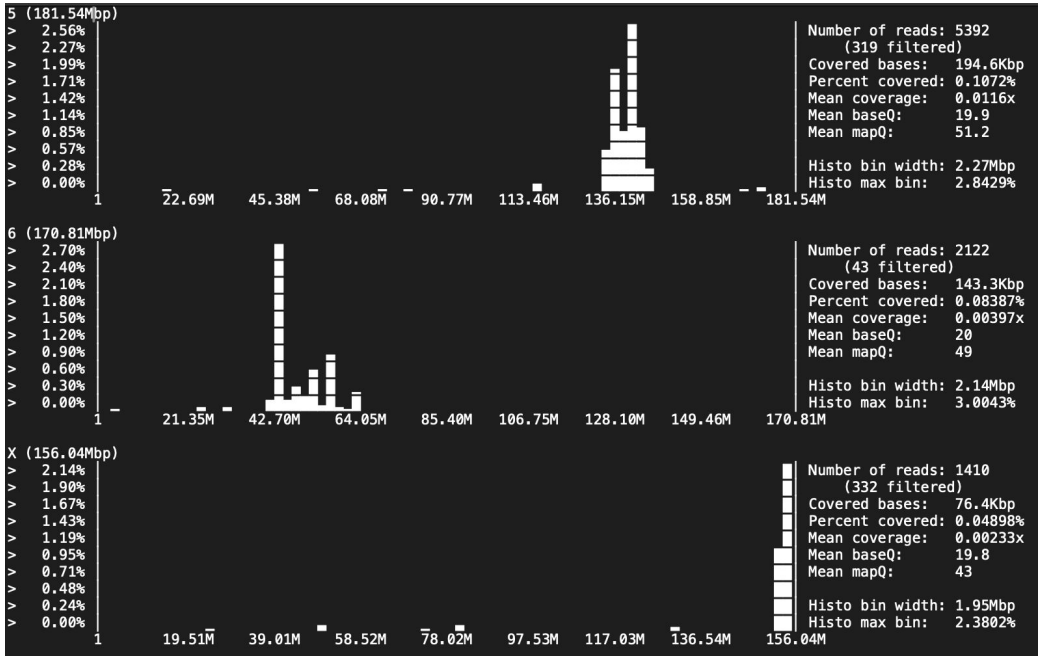


splice

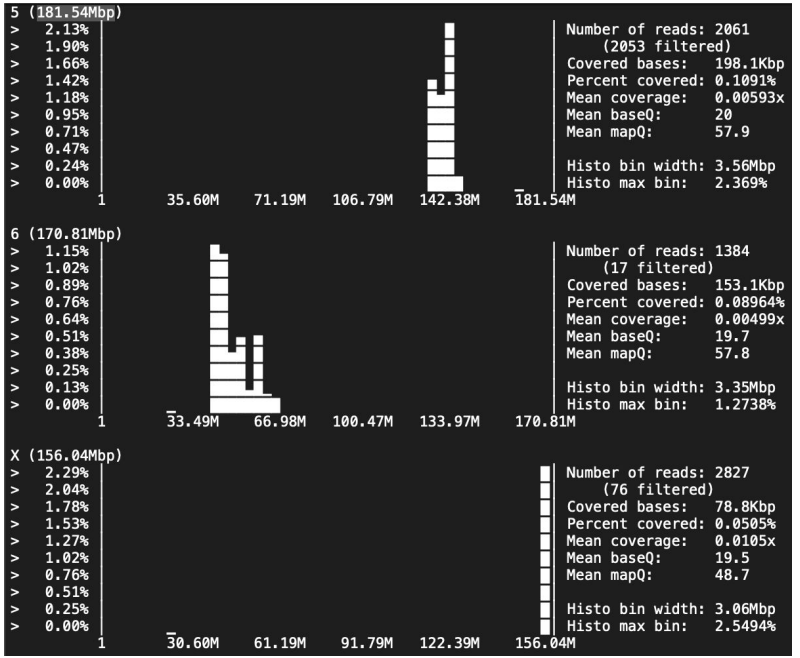


ev_2 coverage

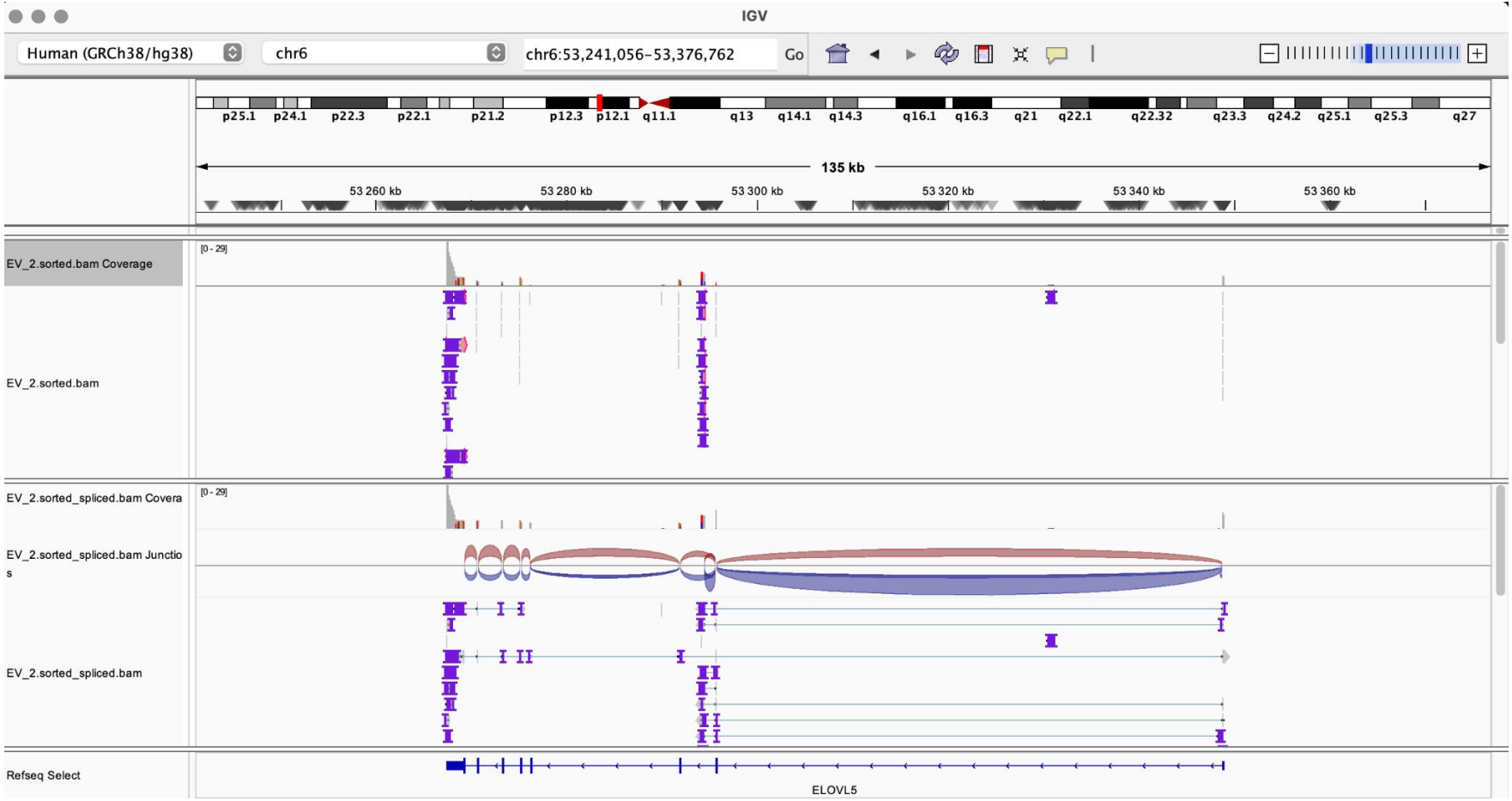
map-ont

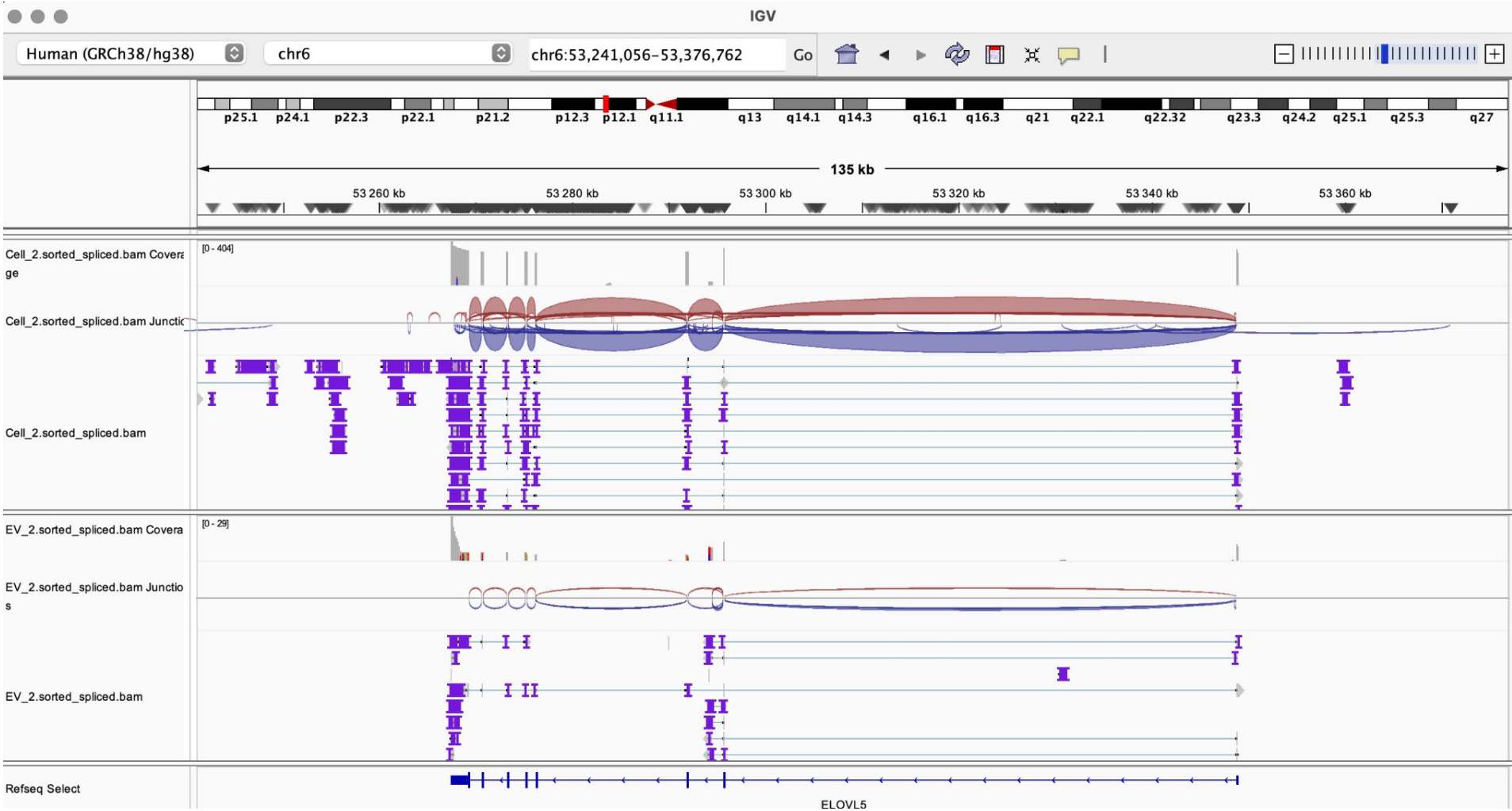


splice



IGV analysis (ELOVL5) differences in alignment


IGV
Human (GRCh38/hg38)
chr6
chr6:53,241,056-53,376,762
Go
p25.1 p24.1 p22.3 p22.1 p21.2 p12.3 p12.1 q11.1 q13 q14.1 q14.3 q16.1 q16.3 q21 q22.1 q22.32 q23.3 q24.2 q25.1 q25.3 q27
135 kb
53 260 kb
53 280 kb
53 300 kb
53 320 kb
53 340 kb
53 360 kb
EV_2.sorted.bam Coverage
[0 - 29]
EV_2.sorted.bam
EV_2.sorted_spliced.bam Covera
[0 - 29]
EV_2.sorted_spliced.bam Junctio
s
EV_2.sorted_spliced.bam
Refseq Select
ELOVL5


IGV
Human (GRCh38/hg38)
chr6
chr6:53,241,056-53,376,762
Go
p25.1 p24.1 p22.3 p22.1 p21.2 p12.3 p12.1 q11.1 q13 q14.1 q14.3 q16.1 q16.3 q21 q22.1 q22.32 q23.3 q24.2 q25.1 q25.3 q27
135 kb
53 260 kb
53 280 kb
53 300 kb
53 320 kb
53 340 kb
53 360 kb
Cell_2.sorted_spliced.bam Coverage
[0 - 404]
Cell_2.sorted_spliced.bam Junctio
Cell_2.sorted_spliced.bam
EV_2.sorted_spliced.bam Coverage
[0 - 29]
EV_2.sorted_spliced.bam Junctions
EV_2.sorted_spliced.bam
Refseq Select
ELOVL5

Cell_2.sorted.bam

```
(ngs-tools) abc@c108395bb627:/group_work/group2/project2/references/project2/alignments$ samtools view -h Cell_2.sorted.bam | he
ad -n 10
@HD     VN:1.6  SO:coordinate
@SQ     SN:5    LN:181538259
@SQ     SN:6    LN:170805979
@SQ     SN:X    LN:156040895
@PG     ID:minimap2     PN:minimap2     VN:2.28-r1209   CL:minimap2 -a -x map-ont /group_work/group2/project2/references/project
2/references/Homo_sapiens.GRCh38.dna.primary_assembly.chr5.chr6.chrX.fa /group_work/group2/project2/references/project2/reads/Ce
ll_2.fastq.gz
@PG     ID:samtools     PN:samtools     PP:minimap2     VN:1.21 CL:samtools view -bh
@PG     ID:samtools.1   PN:samtools     PP:samtools     VN:1.21 CL:samtools sort -o /group_work/group2/project2/references/proje
ct2/alignments/Cell_2.sorted.bam
@PG     ID:samtools.2   PN:samtools     PP:samtools.1   VN:1.21 CL:samtools view -h Cell_2.sorted.bam
fb0d4d61-ef19-48d4-9e49-40bc6132cb78    2048    5       1571455 60      1031H90M1I31M1D4M2D4M4D5M1I95M1I57M1D36M     *0      0
      GAGGAAATGTTCGTTTTCTCCATTTGCTCTCAGGCACCAACATTAATTTGAACTTAGAAATAAACTATAACAAAGCCAGGCATGGTGGTTCCACACCTATAATTCCAGCACTTTGGGAGGCC
AGGAGAGGCTTGAGGCCCAGGAGTCTGAGACCAGCTTAGGCAACACAGTGAGACCCTGTCCCTACAATTACAAAATAAACTAGCTGGGCGTGGTGGTGCACACCTGTAGCCCCCAACTACTCAGAAGG
CTGAGTTGGGAAGATCACCTGAGCTGCCCAGGAGTCTGAGCTGCAGTGAGCTGAGATTGTACCACTGCACCAATC       5/029;025AC>6D?LDAE.:569K5=C=*DCA91+2989GCGGDC
CIC=BDDK?866G@=:<11068+,78;@DGBD@A>9=+.723?<,/368BAHDDEEIKEMMGAC779LGJEHJ87.'.3/-'*+%%&'&%<?C?>@>94+5/42,<<?98221/5CB8:>C@>@ADCC
D79-66<=AD@E>;<86-*69BG:CF5;79ED858384>25&&&,<?@A?A=B=A;773221&)//-')895158@=>D6>=>=B@..67:A<EC=98EACAKMDBE8+C:=8906:6A>5@BF:931
8.7,.4<<<DI:<A?@>>C<;8;     NM:i:12 ms:i:592        AS:i:588        nn:i:0  tp:A:P  cm:i:30 s1:i:222        s2:i:131        de:f:0
.0243     SA:Z:X,154083452,+,13S481M10D862S,60,29;X,154083066,+,501S291M1D564S,60,8;    rl:i:134
63ef7e04-ddb7-498f-a3ec-583cc30ac395    272     5       2585154 0       48S45M4D15M1D12M1I2M1I3M3D3M5D7M9D28M2I12M2I30M1893S  *
      0       0       *       *       NM:i:47 ms:i:131        AS:i:108        nn:i:0  tp:A:S  cm:i:5  s1:i:42       de:f:0.1687
    rl:i:20
```

Cell_2_sorted_spliced.bam

```
(ngs-tools) abc@c108395bb627:/group_work/group2/project2/references/project2/alignments$ samtools view -h Cell_2.sorted_spliced.
bam | head -n 10
@HD     VN:1.6  SO:coordinate
@SQ     SN:5    LN:181538259
@SQ     SN:6    LN:170805979
@SQ     SN:X    LN:156040895
@PG     ID:minimap2     PN:minimap2     VN:2.28-r1209   CL:minimap2 -a -x splice /group_work/group2/project2/references/project2
/references/Homo_sapiens.GRCh38.dna.primary_assembly.chr5.chr6.chrX.fa /group_work/group2/project2/references/project2/reads/Cel
l_2.fastq.gz
@PG     ID:samtools     PN:samtools     PP:minimap2     VN:1.21 CL:samtools view -bh
@PG     ID:samtools.1   PN:samtools     PP:samtools     VN:1.21 CL:samtools sort -o /group_work/group2/project2/references/proje
ct2/alignments/Cell_2.sorted_spliced.bam
@PG     ID:samtools.2   PN:samtools     PP:samtools.1   VN:1.21 CL:samtools view -h Cell_2.sorted_spliced.bam
ab8e7b8b-75cc-484a-9f45-4b8a26735fdd    256     5       5393957 0       89M1D20M1D7M1D19M1I6M1D5M1D56M5I3M2I63M3I38M1D16M3I26M1I
10M1D25M1D26M1I8M2I28M1D11M2I6M1D3M2D2M1I4M1I5M1D15M1D6M3D64M1I29M1D1M1D6M1I14M *       0       0       *       *       NM:i:85m
s:i:391 AS:i:384        nn:i:0  tp:A:S  cm:i:66 s1:i:337        de:f:0.1109     rl:i:0
d47da1d5-c11c-4cd5-8ef3-e1140f389a91    256     5       6848307 0       19M1I14M3D22M1D41M2I4M1D30M     *       0       0       *
*       NM:i:14 ms:i:95 AS:i:94 nn:i:0  tp:A:S  cm:i:8  s1:i:57 de:f:0.0815     rl:i:0
```

featureCounts

Cell_2_sorted:

Geneid	Chr	Start	End	Strand	Length	alignments/Cell_2.sorted.bam
ENSG00000012660	6	53277646	53277743	-	98	3
ENSG00000012660	6	53294088	53294491	-	404	47

Cell_2_sorted_spliced:

Geneid	Chr	Start	End	Strand	Length	alignments/Cell_2.sorted_spliced.bam
ENSG00000012660	6	53277646	53277743	-	98	3
ENSG00000012660	6	53294088	53294491	-	404	5

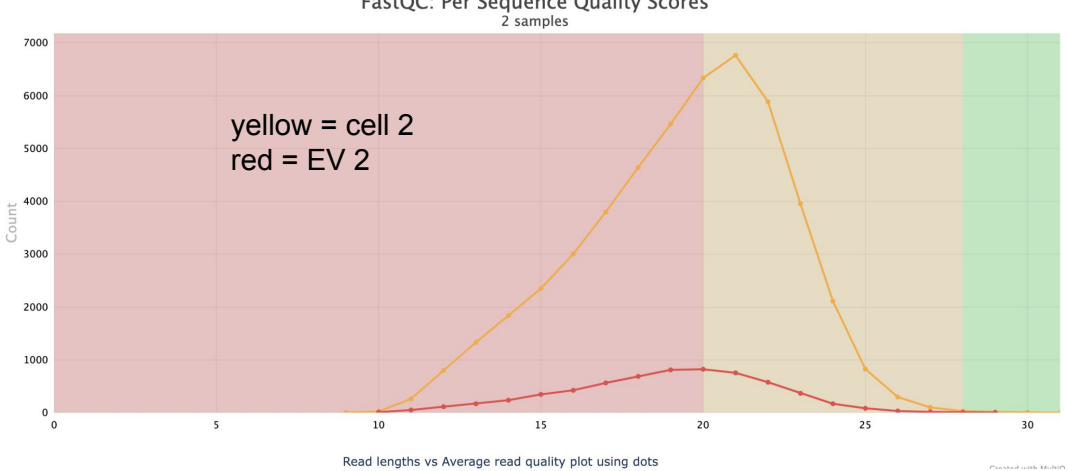
featureCounts

EV_2_sorted:

Geneid	Chr	Start	End	Strand	Length	alignments/EV_2.sorted.bam
ENSG00000012660	6	53294088	53294491	-	404	10

EV_2_sorted_spliced:

ENSG00000012660	6	53294088	53294491	-	404	1



cell 2

ev 2

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

Quality is lower in nanoplot

Mean read quality:
fastqc_cell 16.98 >nano_cell: 11.2
fastqc_ev: 16.97 >nano_ev: 11.1