

# 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](#)

3147 Accesses | 1 Altmetric | [Metrics](#)

#### 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

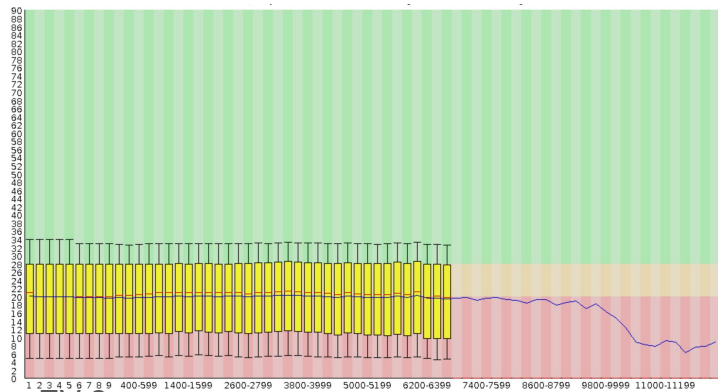
**What is the average read quality? What kind of accuracy would you expect?**

- Cell\_2: 16.9851
- EV\_2: 16.977

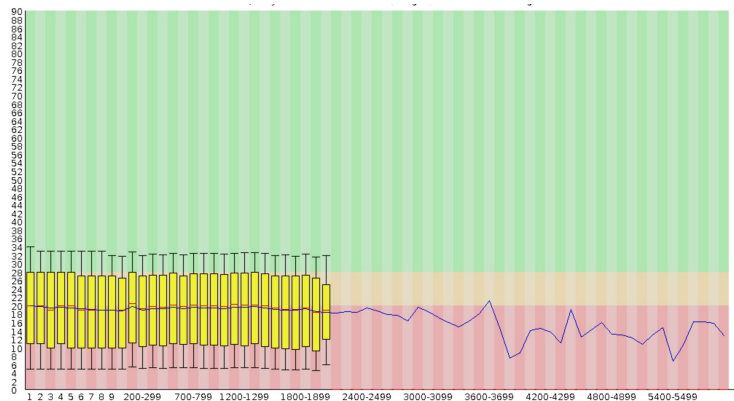
→ An average quality score of ~17 is within the expected range for raw ONT reads (base call accuracy of 98%)

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

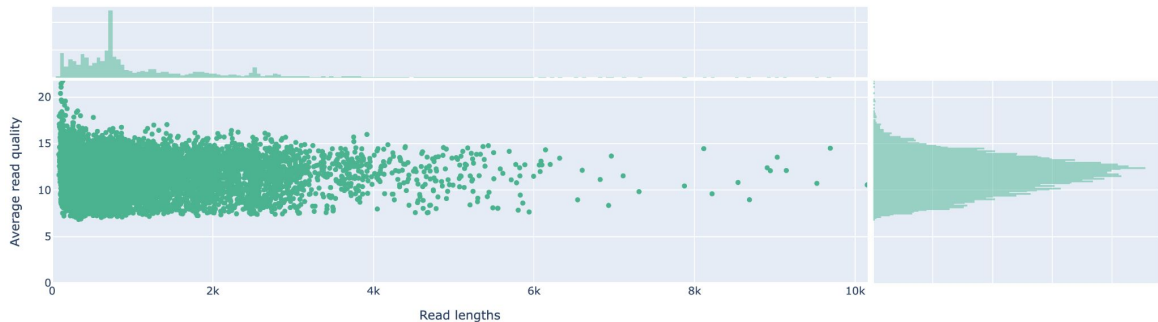
Cell 2



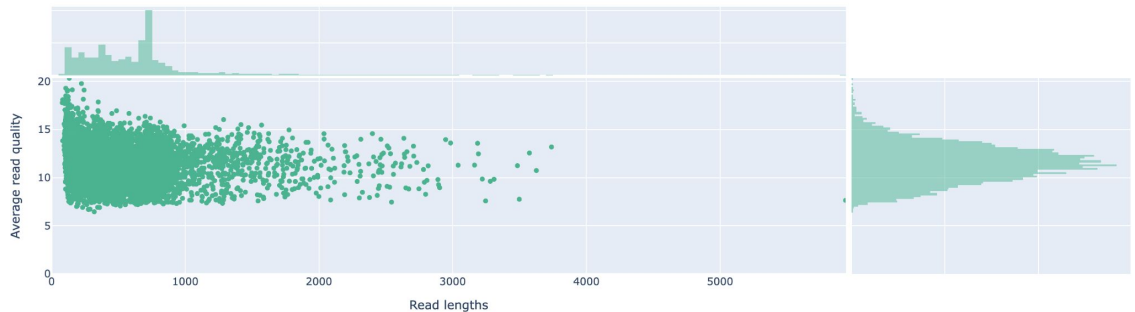
EV 2



Read lengths vs Average read quality plot using dots



Read lengths vs Average read quality plot using dots



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

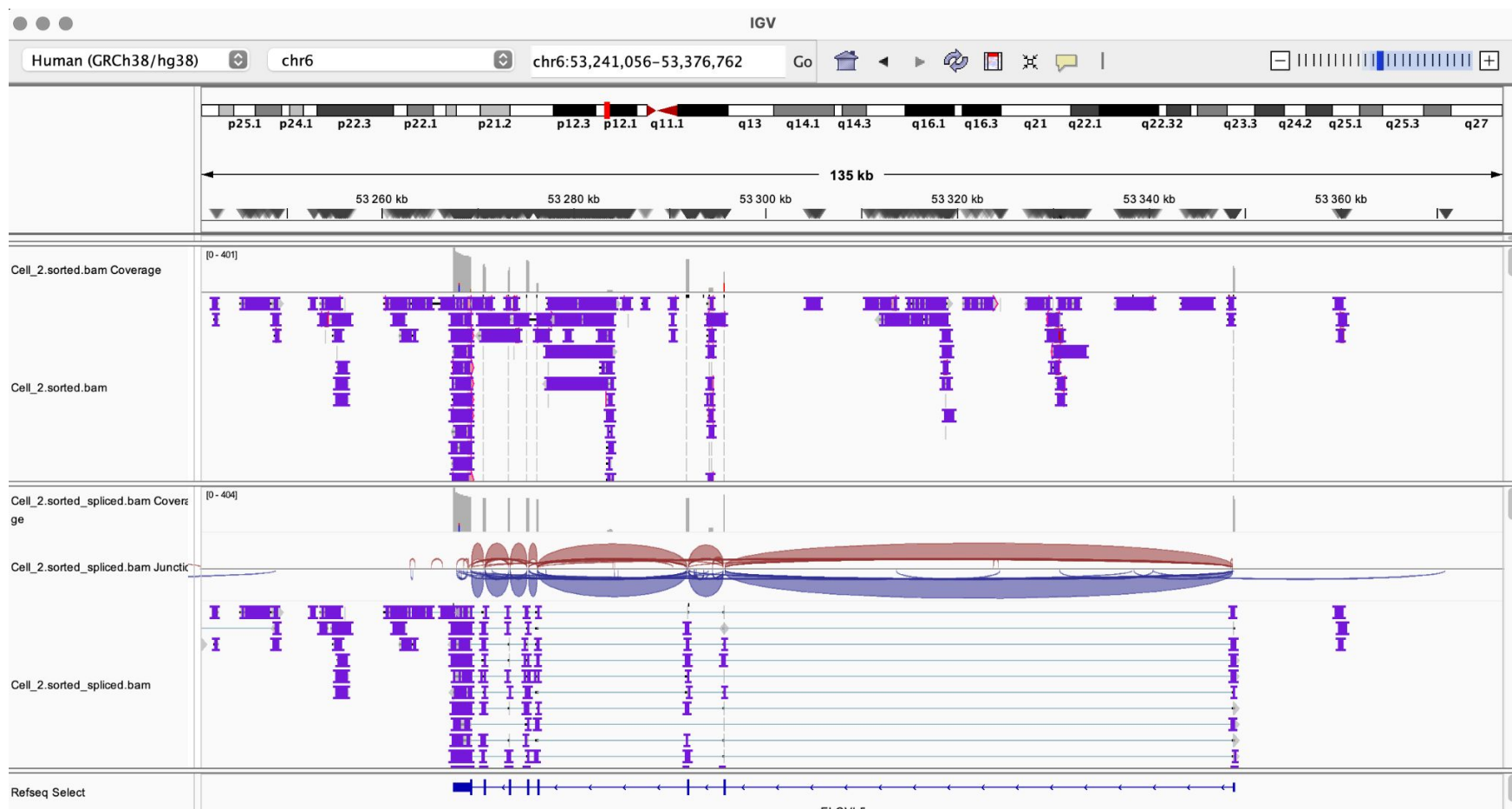
- asm5/asm10/asm20 - asm-to-ref mapping, for ~0.1/1/5% sequence divergence

- 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
```

```
1430H44M1195M2157M1111M1145M3032M508M2031M110M1021M1118M109M206M1564N8M2023M1159M4065N12M1023M3027M1138M1132M1134301M1925N10M50
24M1035M1032M2018N27M101M108M1084M1012M1019M3040M2030M102M16843N4M1057M109M1151M108M109M2022M214M108M
1H10M1164M14744N64M102M1012M202M1114M206M1010M2822G6M213M6I7M204M2010M8044M198492N35M1152M1047H
48S12M6110M2115M1308N217M2010M10169M114M113M2042M116M5126M10103M1058M1012M307M11105M2018M11133M117M1174M6S
18M2166M93N19M2038M517M2039N11M2032M1126M202M116M102M49595N18M2111M107M1151M1042M1112M115M1123M301M1710N13M1015M1020M1022M2140M
10137M1059M3111M1120M1118M3111M1023M3160M117M1117M106M217M106M21116M116M103M10118M1013M2142M11113M107M
91M1013M2119M93N19M1039K317M2039N111M2011M1056M2024217M101M8643N5M112M1076M1162M101M2113M1012M141M1087M1110M2166M1018M1036M1020
4202M1011M12018M1080M103M103M10103M208N33M1144M2010M202M110M1113M212M2115M1011M1139M2010M101710N50M1113M215M1044M10131M2036M107
7M3032M2113M2161M207M1185M1012M1094M3140M2112M1065M1085M2010M2076M109M20120M
39M1113M93N25M109M112M1112M1016M2039N1M211M19M1026M114M103M1037M8643N55M1135M1171M1110M1031M3112M1039M1140M101M1039M2136M104M20
55M207M1034M1907M1062M2115M109M1016M114M3035M209M342N5M1111M219M104M1129M119M1212M115M102M1023M1110M1710N1M3118M114M2122M10124
M1077M1020M1145M114M211M1026M101M3123M1034M112M11162M1036M1020M106M104M2110M1038M1020M1131M2045M103M3029M1117M106M2150M1057M113M
2111M105M2012M1140M113M2012M5020M105M1014M
2257M112M1026M1110M114M1151M2122M108M211M106M201M102M1036M1131M1161M102M1034M114M316M2138M3127M103M107M112M3147M2046M1018M1042M
1113M1114M113M1038M1110M4811212N13M202M112M3013M106M1115M3114M302M405M309M303M206M102M1026M514M1025M704M702M104M1084M3018M2022M1
13M112M1123M109M2123M1184M4021M114M1057M7067M114M1170M
12578M114M218M1113M1030M209M205M1417N130M114M2021M5015M1023M
165S113M706M1013M3009M607M114M104M102M1020M105M114M115M107M104M1076M111M2136M1033M102M112M1144M113M1156M1013M6018M1014M1152M10
34M114M1143M115M1090M1022M1018M1020M698N1164M1022M1129M1118M2125M2119M2018M1134M1019M1122M11346M2161M95
1043S25M113M114M305M1012M116032M109M1016M1018M11419N3M1019M215M202M1013M2013M112M1009M106M1010M117M108M1023M1061M1139M308M2123M113
5M214M1133M1024M114M213M20115M1014M202M107M2110M1147M113M8057M55
```

## CIGAR strings

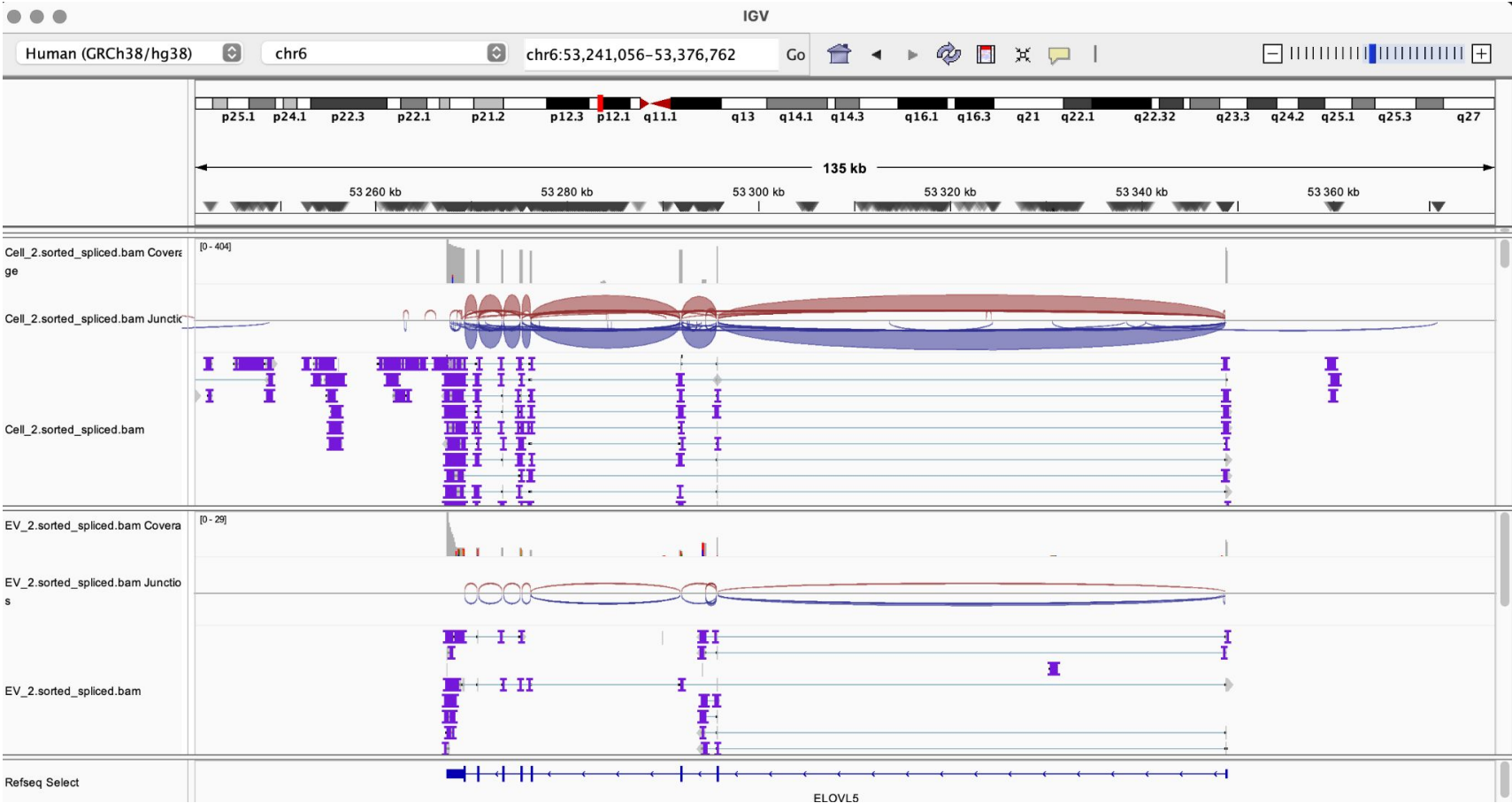
| Op | BAM | Description   |
|----|-----|---|
| M  | 0   | alignment match (can be a sequence match or mismatch) |
| I  | 1   | insertion to the reference                            |
| D  | 2   | deletion from the reference                           |
| N  | 3   | skipped region from the reference                     |
| 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

## CIGAR with letter N only present in Cell\_2.sorted\_spliced.bam file



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

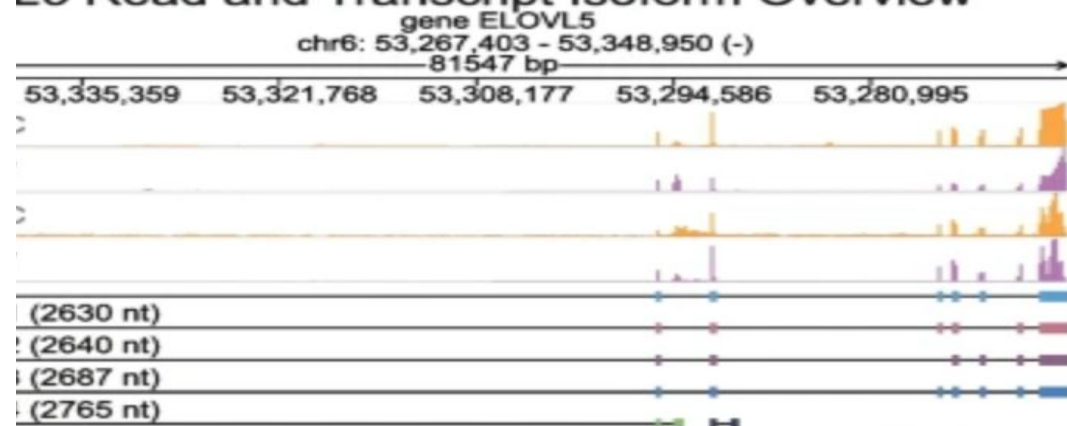


Do you already see evidence for splice variants in the alignments?



# Supplementary materials

## L5 Read and Transcript Isoform Overview



# 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

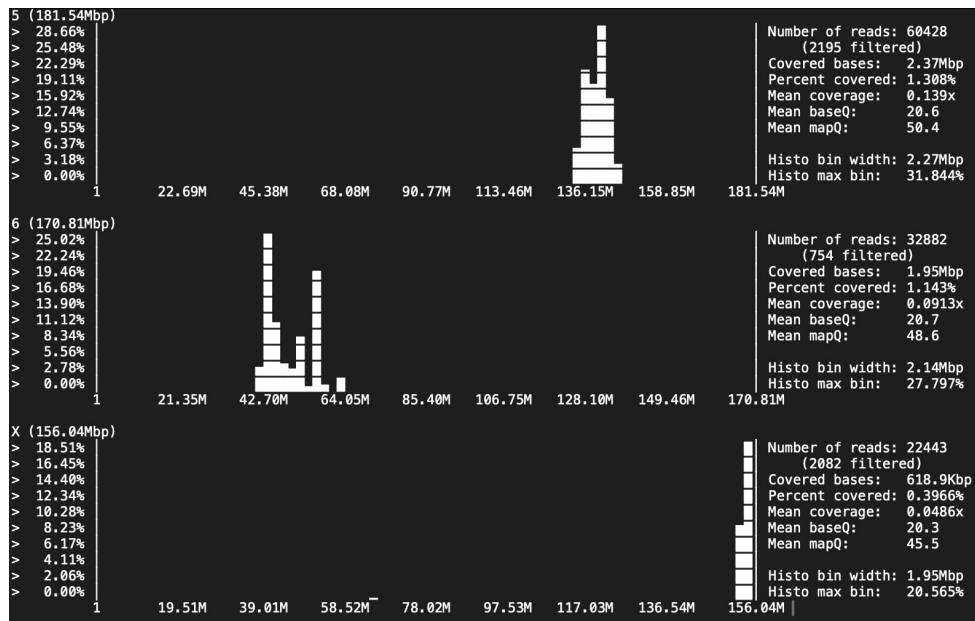
Check out the option `-x` of `minimap2`. Are the defaults appropriate?

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

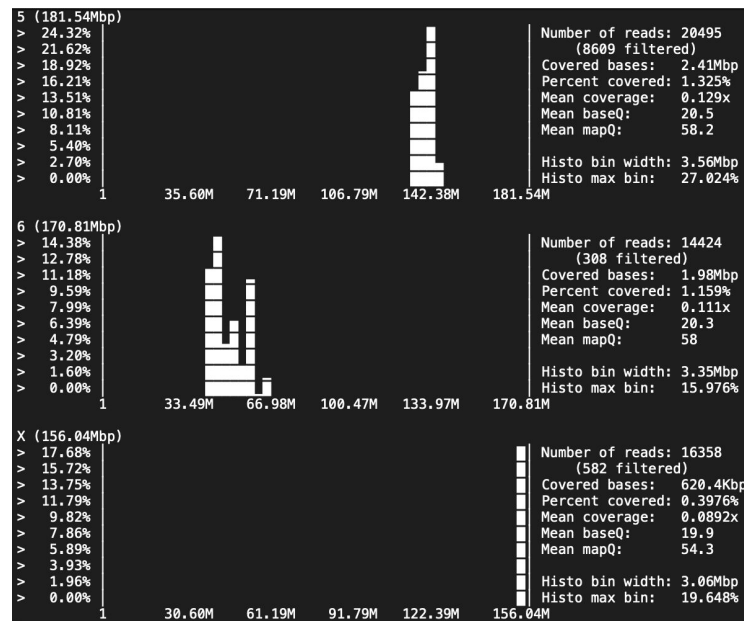
How are spliced alignments stored in the SAM file with the different settings of `-x`?

# Whole cell\_2 coverage

map-ont



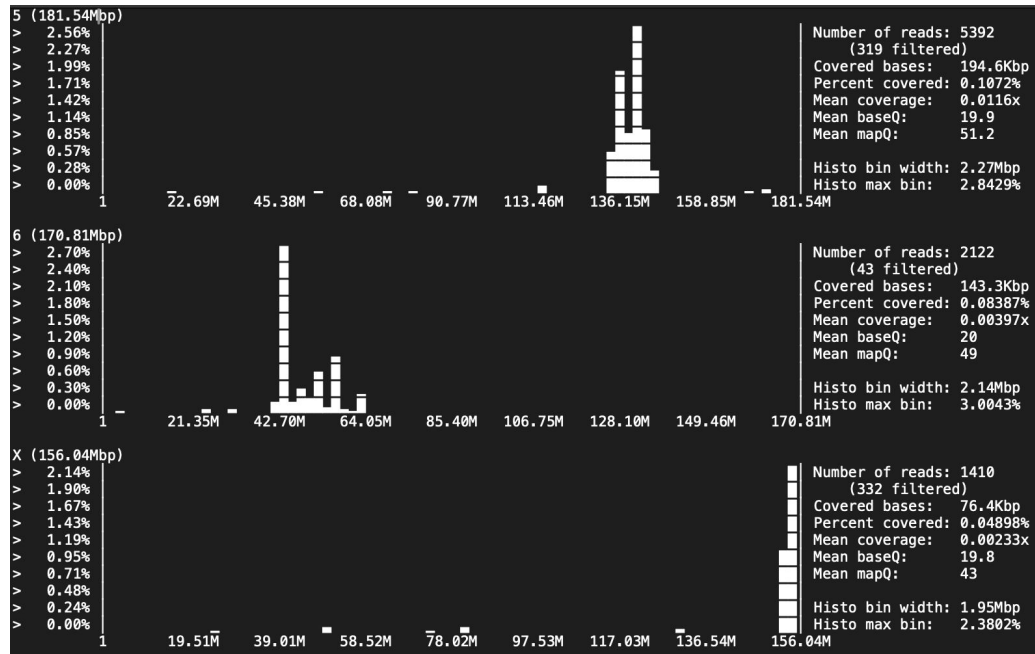
splice



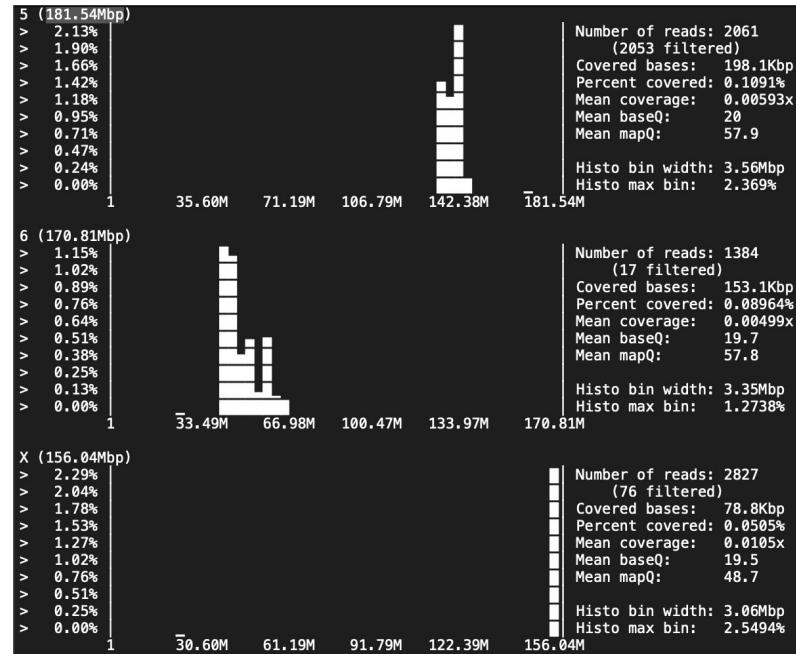


# ev\_2 coverage

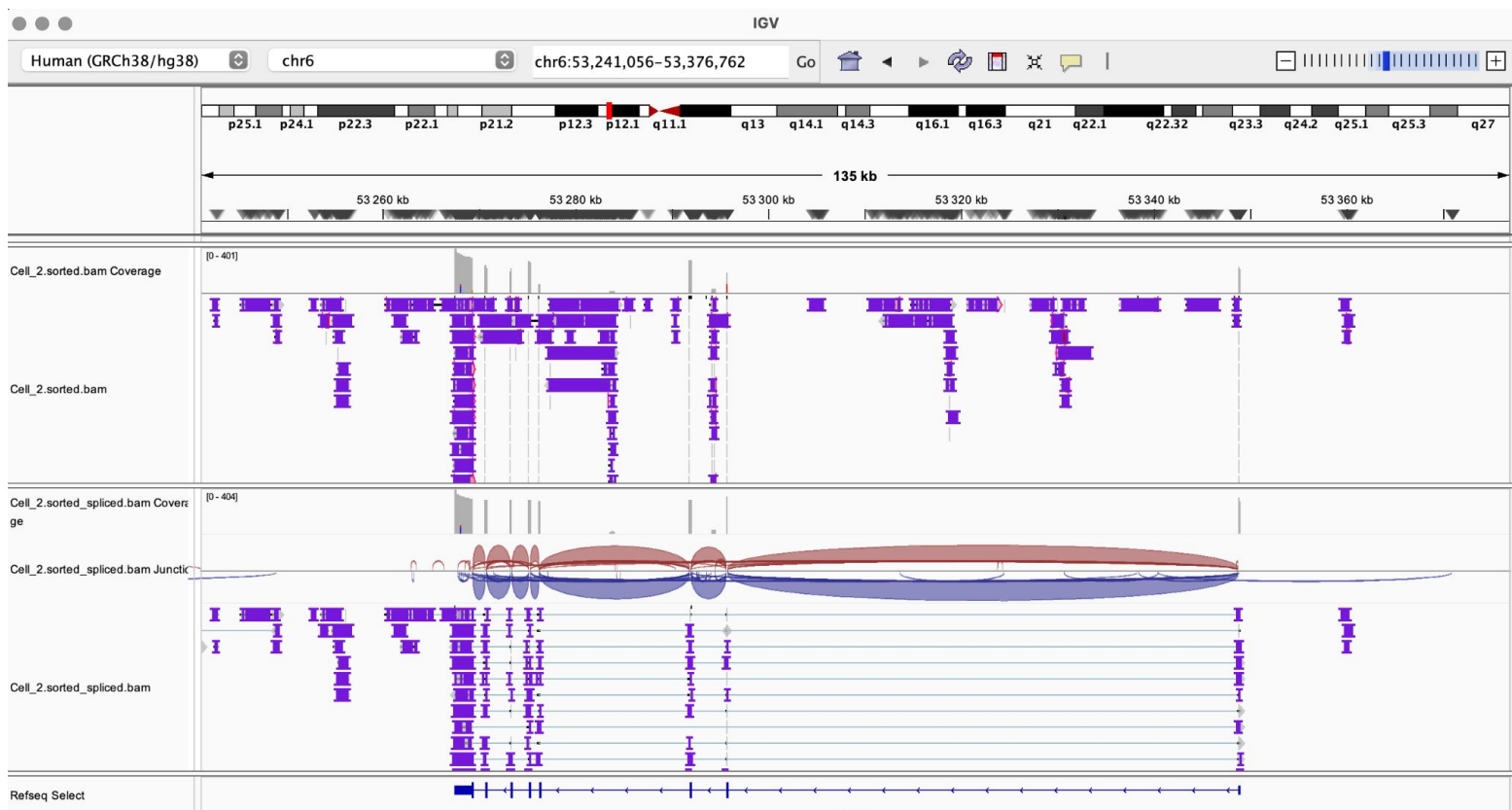
## map-ont

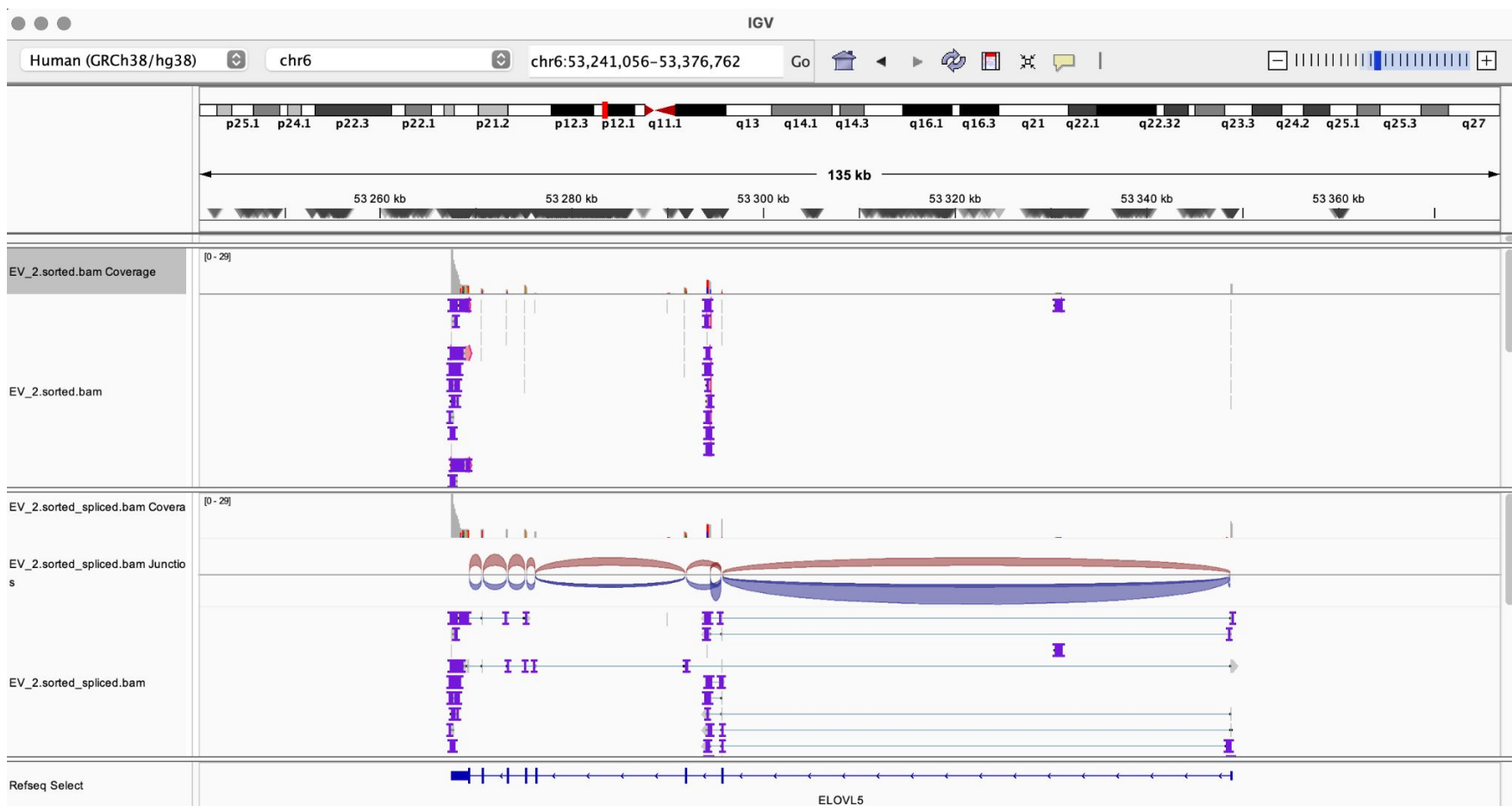


## splice



# IGV analysis (ELOVL5) differences in alignment



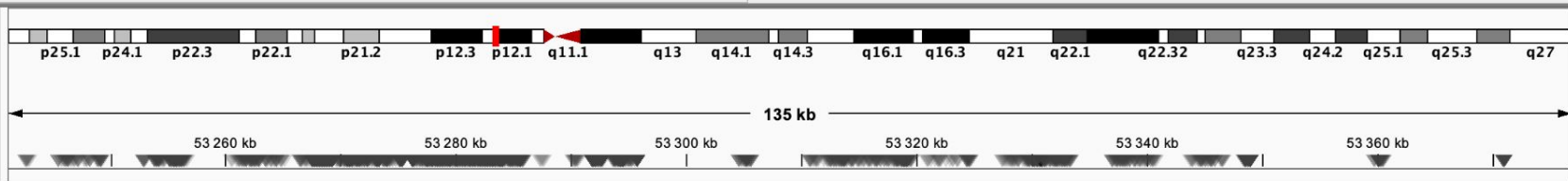


Human (GRCh38/hg38)

chr6

chr6:53,241,056-53,376,762

Go



Cell\_2.sorted\_spliced.bam Coverage

[0 - 404]

Cell\_2.sorted\_spliced.bam Junctions

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 | head -n 10
@HD      VN:1.6   S0: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/project2/references/Homo_sapiens.GRCh38.dna.primary_assembly.chr5.chr6.chrX.fa /group_work/group2/project2/references/project2/reads/Cell_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/project2/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
      GAGGAAATGTTCTGTTTTCTCCATTTGCTCTCAGGCACCAACATTAATTTGAACTTAGAAATAAACTATAACAAAGCCAGGCATGGTGGTTCACACCTATAATTCCAGCACTTTGGGAGGCC
AGGAGAGGCTTGAGGCCAGGAGTCTGAGACCAGCTTAGGCAACACAGTGAGACCCTGTCCCTACAATTACAAAATAAACTAGCTGGGCGTGGTGGTGCACACCTGTAGCCCCCACTACTCAGAAGG
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 |



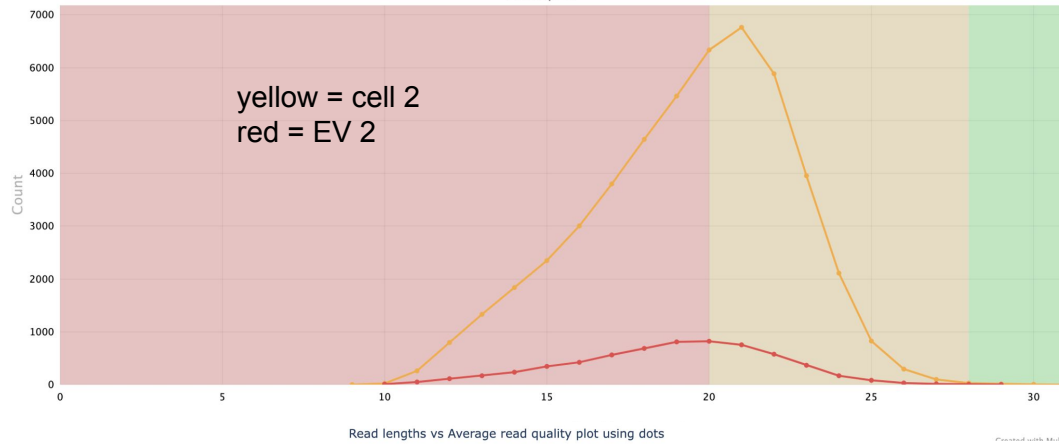
# featureCounts

EV\_2\_sorted:

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

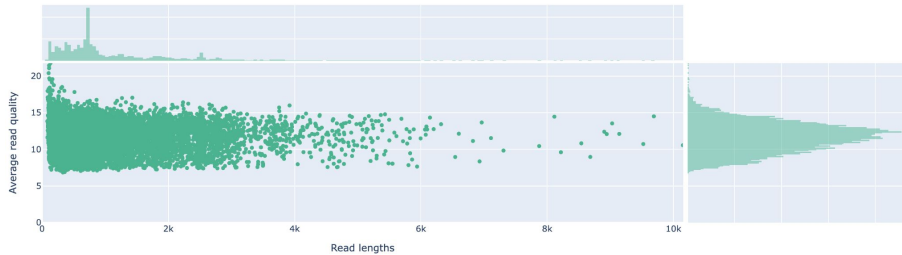
EV\_2\_sorted\_spliced:

|                 |   |  |          |  |          |   |     |   |
|-----------------|---|--|----------|--|----------|---|-----|---|
| ENSG00000012660 | 6 |  | 53294088 |  | 53294491 | - | 404 | 1 |
|-----------------|---|--|----------|--|----------|---|-----|---|



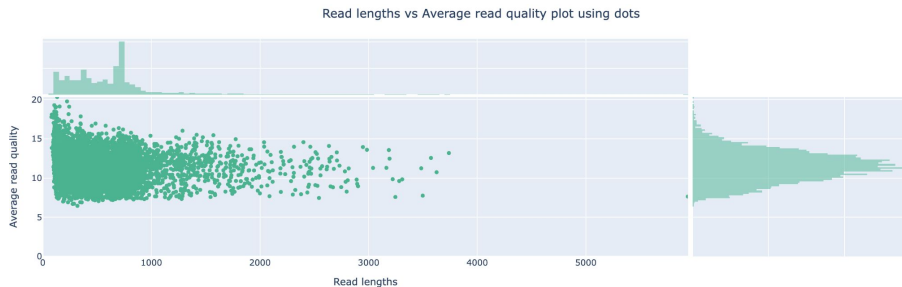
Note any differences between `fastqc` and `NanoPlot`? How is that compared to the publication?

Quality is lower in nanoplot



cell 2

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



ev 2