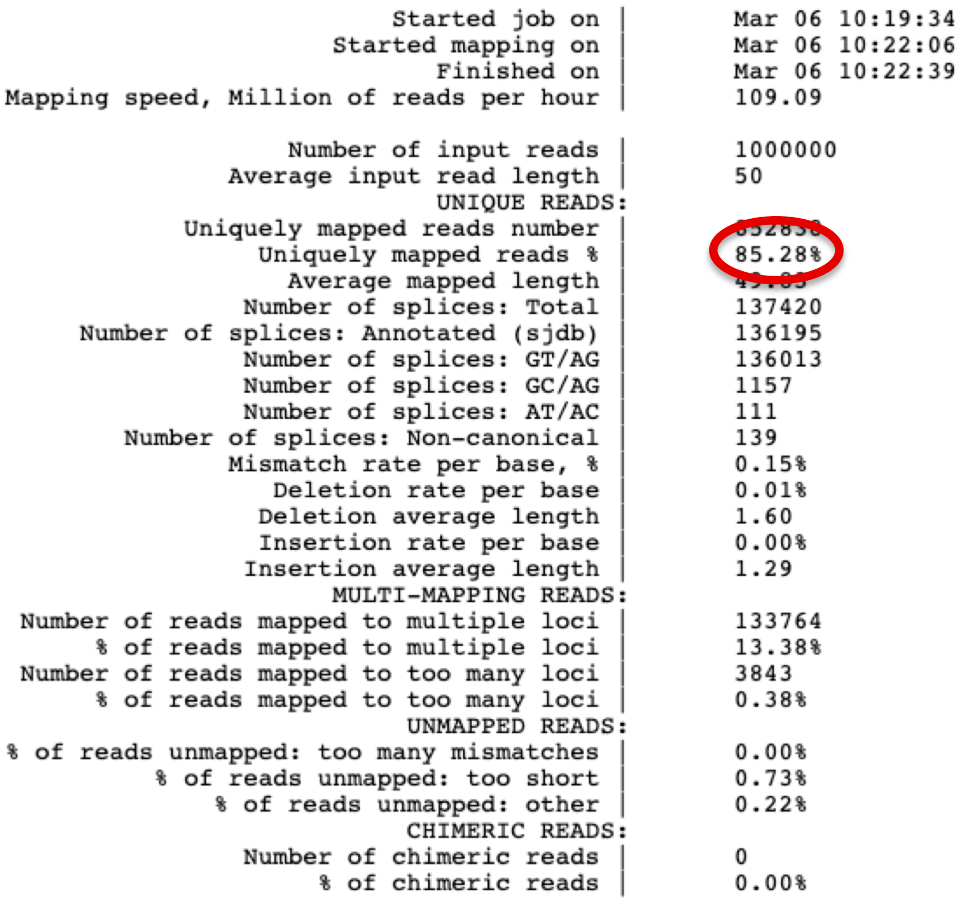
NGS read mapping : answers to questions

Céline Keime
keime@igbmc.fr

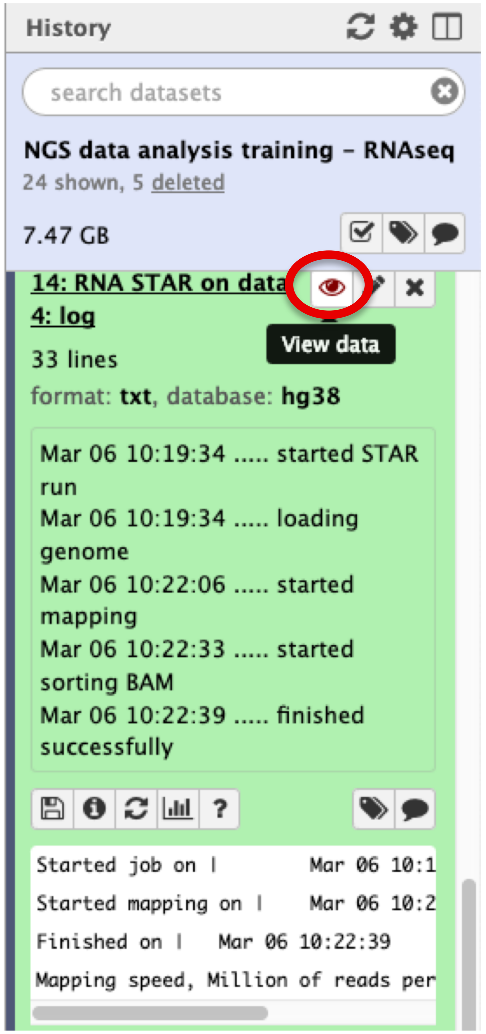
Exercise 1
1. Log file

Proportion of uniquely mapped reads :



```
                                 Started job on |       Mar 06 10:19:34
                             Started mapping on |       Mar 06 10:22:06
                                    Finished on |       Mar 06 10:22:39
       Mapping speed, Million of reads per hour |       109.09

                          Number of input reads |       1000000
                      Average input read length |       50
                                    UNIQUE READS:
                   Uniquely mapped reads number |       852838
                        Uniquely mapped reads % |       85.28%
                          Average mapped length |       49.83
                       Number of splices: Total |       137420
            Number of splices: Annotated (sjdb) |       136195
                       Number of splices: GT/AG |       136013
                       Number of splices: GC/AG |       1157
                       Number of splices: AT/AC |       111
               Number of splices: Non-canonical |       139
                      Mismatch rate per base, % |       0.15%
                         Deletion rate per base |       0.01%
                        Deletion average length |       1.60
                        Insertion rate per base |       0.00%
                       Insertion average length |       1.29
                             MULTI-MAPPING READS:
        Number of reads mapped to multiple loci |       133764
             % of reads mapped to multiple loci |       13.38%
        Number of reads mapped to too many loci |       3843
             % of reads mapped to too many loci |       0.38%
                                  UNMAPPED READS:
       % of reads unmapped: too many mismatches |       0.00%
                 % of reads unmapped: too short |       0.73%
                     % of reads unmapped: other |       0.22%
                                  CHIMERIC READS:
                       Number of chimeric reads |       0
                            % of chimeric reads |       0.00%
```



NGS data analysis training – RNAseq

24 shown, 5 deleted

7.47 GB

14: RNA STAR on data 4: log

33 lines

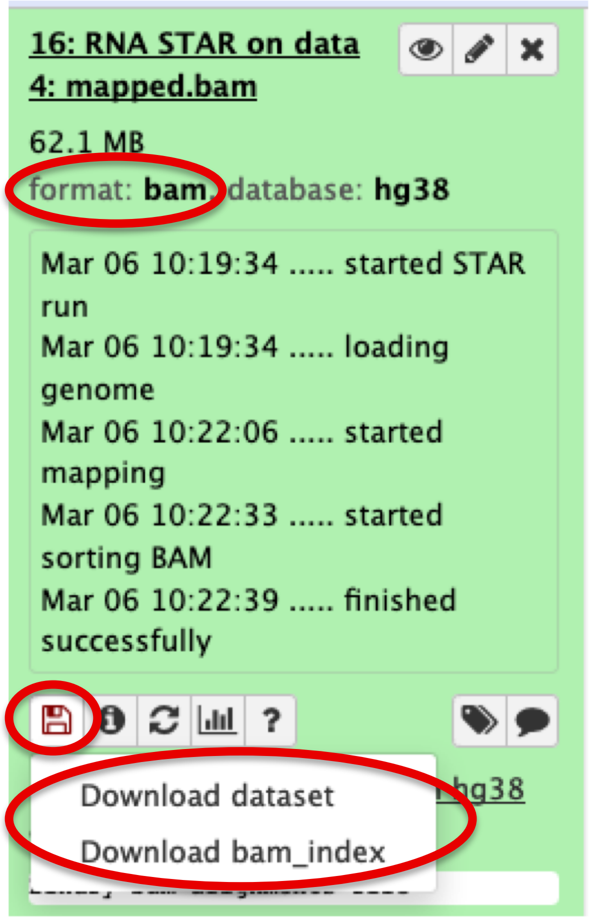
format: txt, database: hg38

```
Mar 06 10:19:34 ..... started STAR run
Mar 06 10:19:34 ..... loading genome
Mar 06 10:22:06 ..... started mapping
Mar 06 10:22:33 ..... started sorting BAM
Mar 06 10:22:39 ..... finished successfully
```

```
Started job on |        Mar 06 10:1
Started mapping on |    Mar 06 10:2
Finished on |    Mar 06 10:22:39
Mapping speed, Million of reads per
```

Exercise 1
2. Alignment file

- Galaxy
  - STAR provides an alignment in BAM format
  - Download this file together with the corresponding index (in the same directory)



16: RNA STAR on data 4: mapped.bam

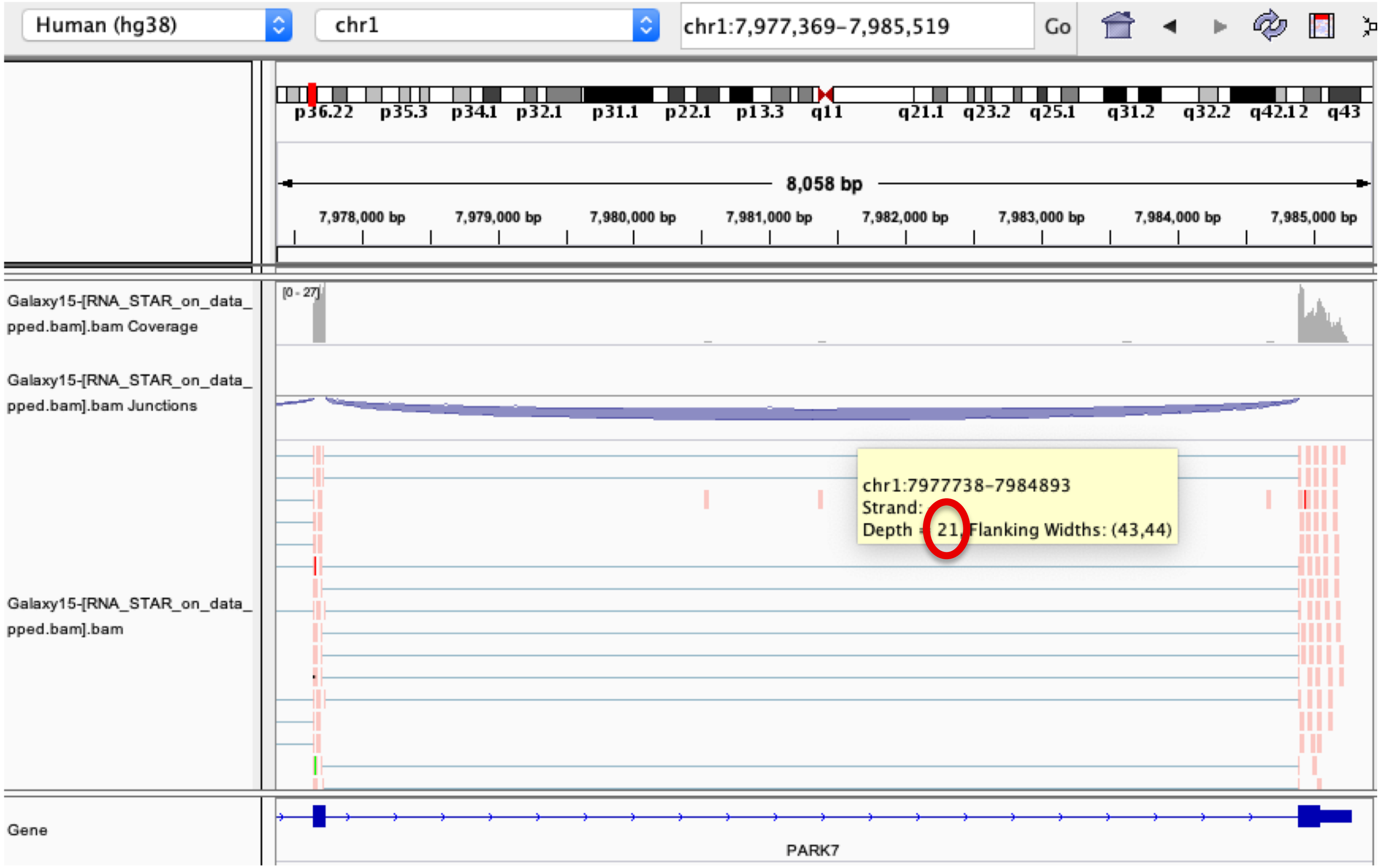
62.1 MB

format: **bam**, database: **hg38**

```
Mar 06 10:19:34 ..... started STAR run
Mar 06 10:19:34 ..... loading genome
Mar 06 10:22:06 ..... started mapping
Mar 06 10:22:33 ..... started sorting BAM
Mar 06 10:22:39 ..... finished successfully
```

- IGV
  - File → Load from file and choose the downloaded BAM file

Exercise 1
2. Splice junction



Human (hg38)

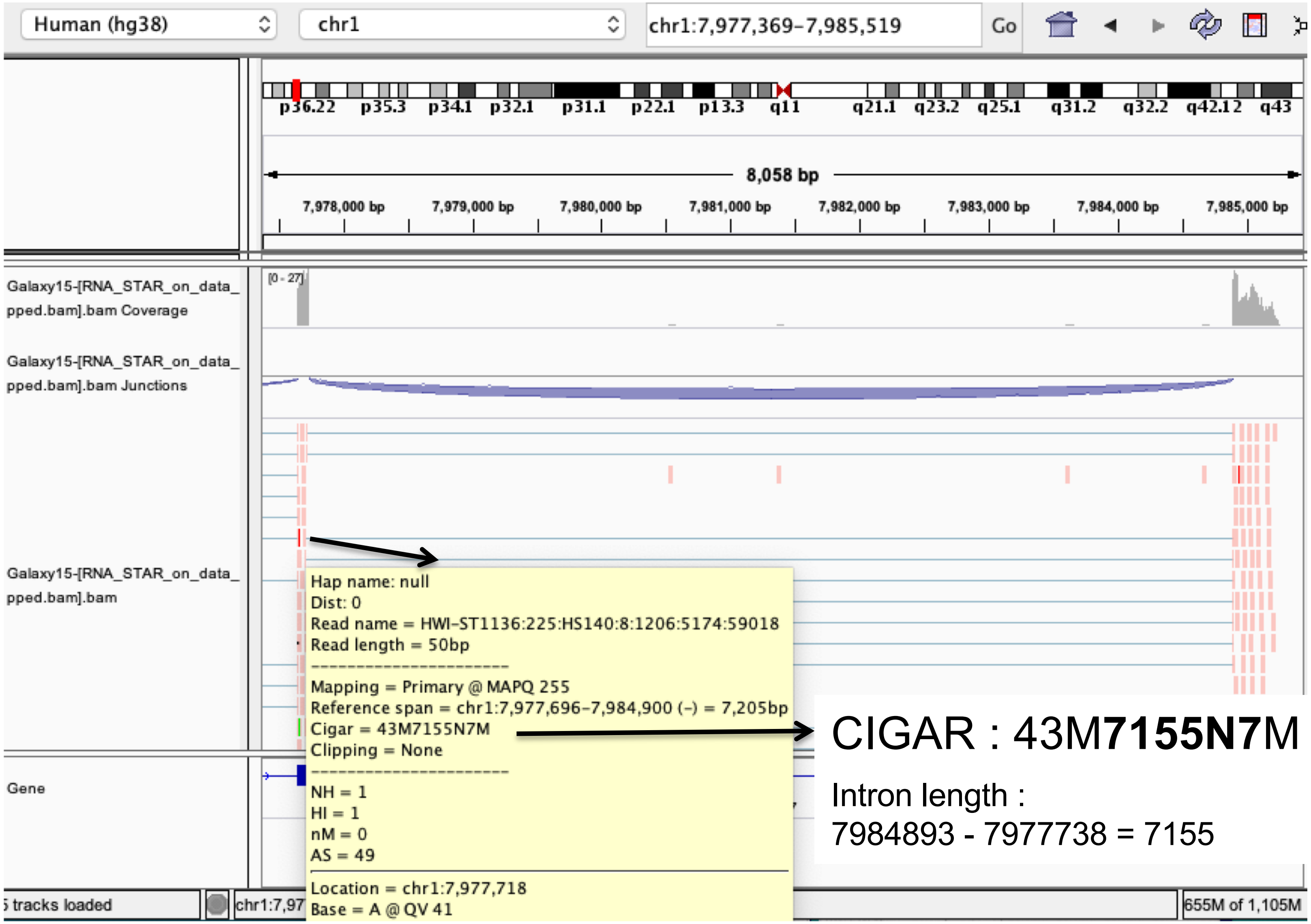
chr1

chr1:7,977,369-7,985,519

chr1:7977738-7984893
Strand:
Depth = 21, Flanking Widths: (43,44)

→ 21 alignments span the junction that joins the last 2 exons of *Park7* gene

Exercise 1
2. Splice junction


Human (hg38)
chr1
chr1:7,977,369-7,985,519
Go
p36.22 p35.3 p34.1 p32.1 p31.1 p22.1 p13.3 q11 q21.1 q23.2 q25.1 q31.2 q32.2 q42.12 q43
8,058 bp
7,978,000 bp
7,979,000 bp
7,980,000 bp
7,981,000 bp
7,982,000 bp
7,983,000 bp
7,984,000 bp
7,985,000 bp
[0 - 27]
Galaxy15-[RNA_STAR_on_data_pped.bam].bam Coverage
Galaxy15-[RNA_STAR_on_data_pped.bam].bam Junctions
Galaxy15-[RNA_STAR_on_data_pped.bam].bam
Gene
5 tracks loaded
chr1:7,97
655M of 1,105M

Hap name: null
Dist: 0
Read name = HWI-ST1136:225:HS140:8:1206:5174:59018
Read length = 50bp

Mapping = Primary @ MAPQ 255
Reference span = chr1:7,977,696-7,984,900 (-) = 7,205bp
Cigar = 43M7155N7M
Clipping = None

NH = 1
HI = 1
nM = 0
AS = 49

Location = chr1:7,977,718
Base = A @ QV 41

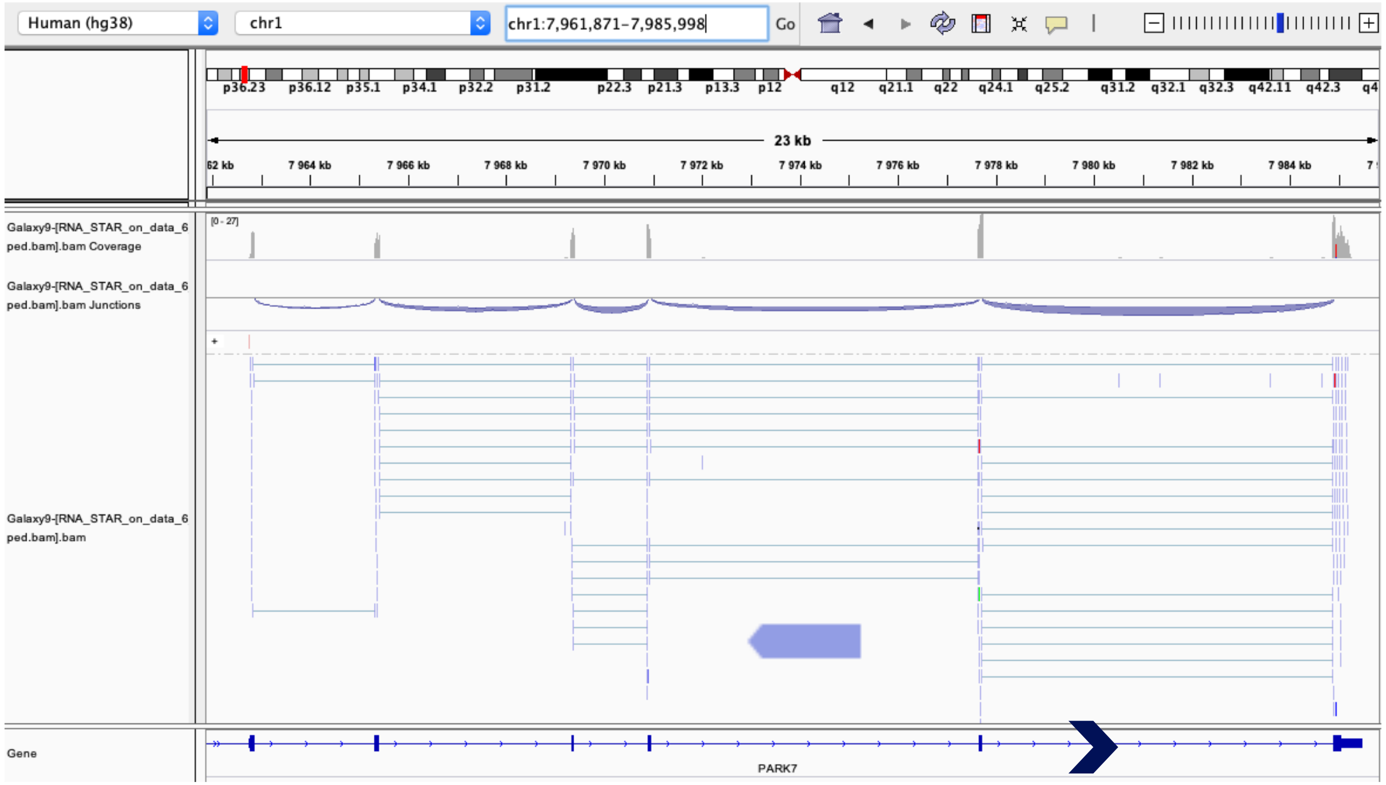
CIGAR : 43M**7155N**7M

Intron length :
7984893 - 7977738 = 7155

Exercise 1
2. Strand specificity

Right click on BAM file → Color alignments by → read strand

Park7 :



Human (hg38)

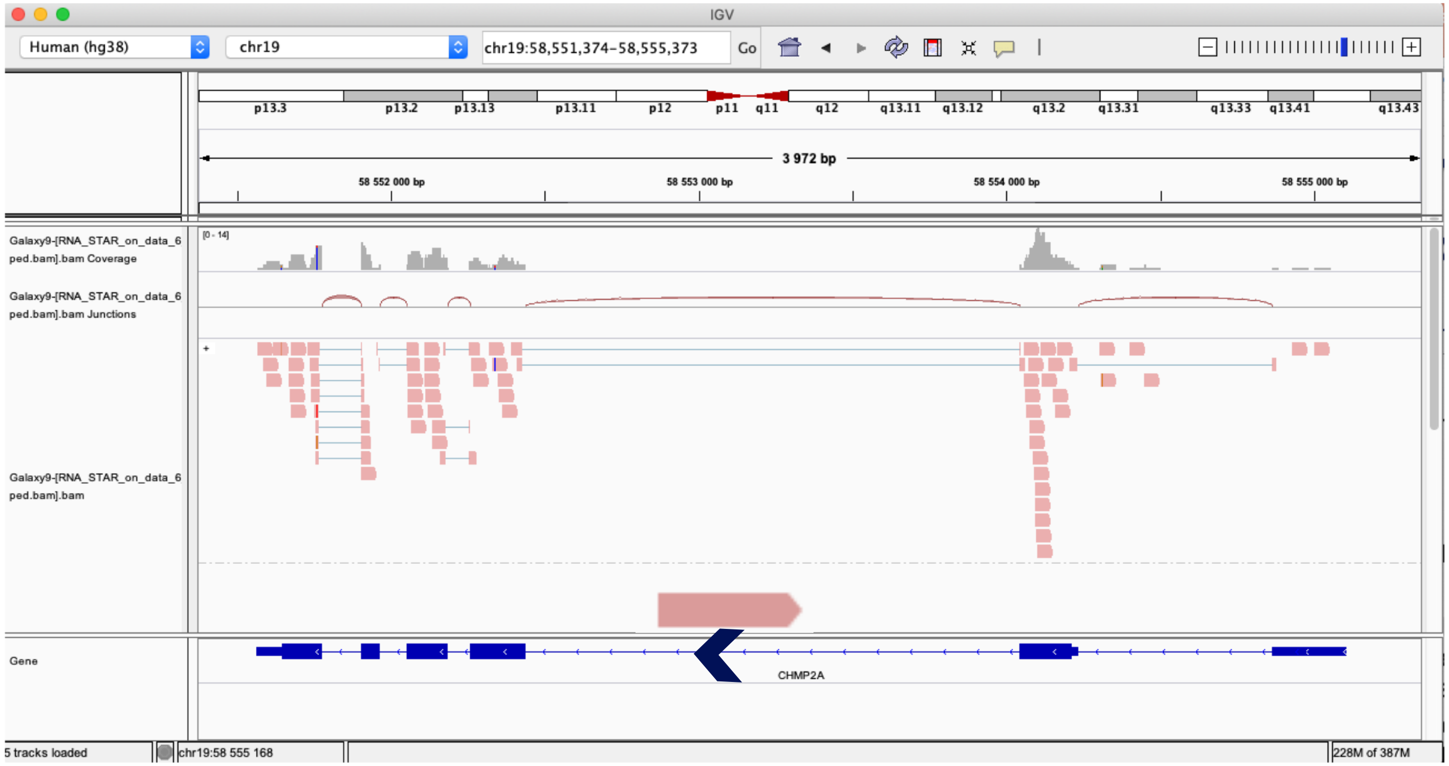
chr1

chr1:7,961,871-7,985,998

The library has been prepared with a directional mRNAseq protocol which retains strand information :
reads are in the opposite direction as the transcribed strand

Exercise 1
2. Strand specificity

Chmp2a :



Human (hg38)

chr19

chr19:58,551,374-58,555,373

5 tracks loaded

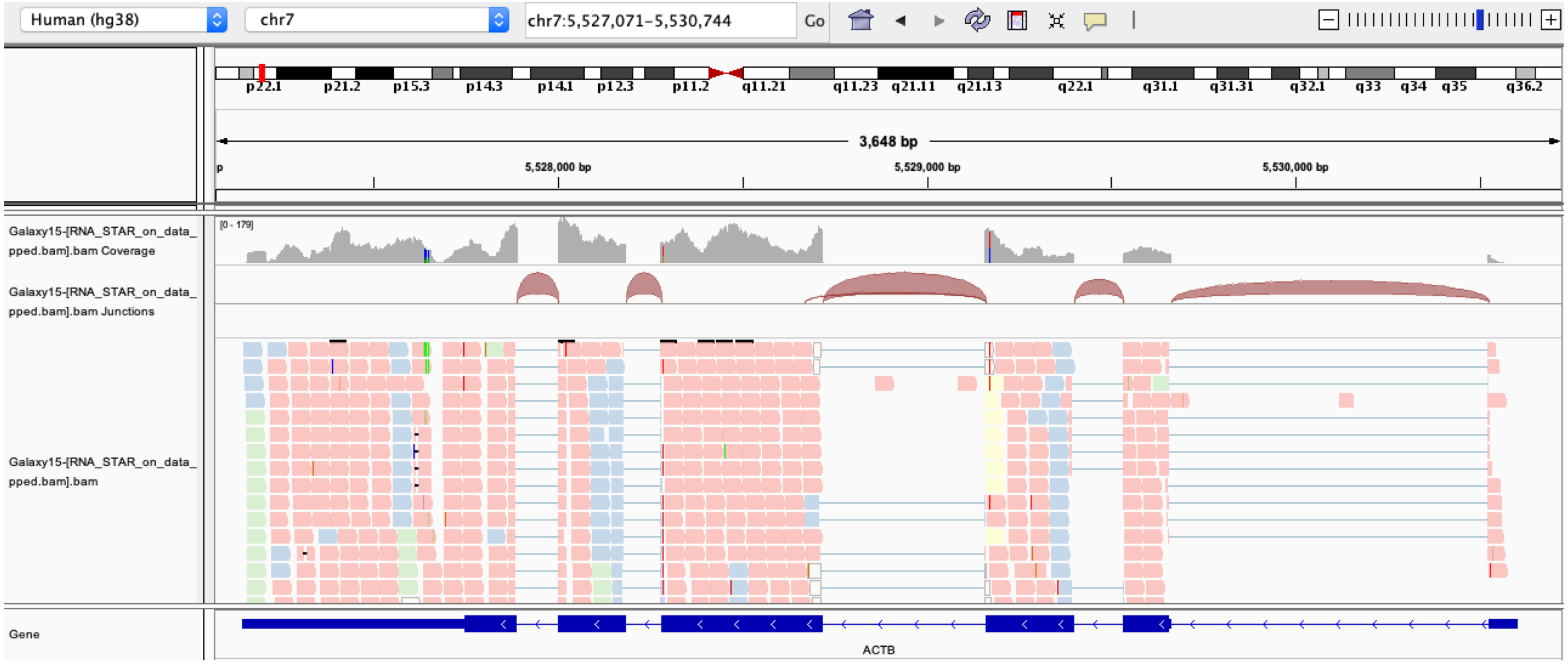
chr19:58 555 168

228M of 387M

The library has been prepared with a directional mRNAseq protocol which retains strand information :
reads are in the opposite direction as the transcribed strand

Exercise 1
2. Multiple mapped reads

Right click on BAM file → Color alignments by → tag → NH

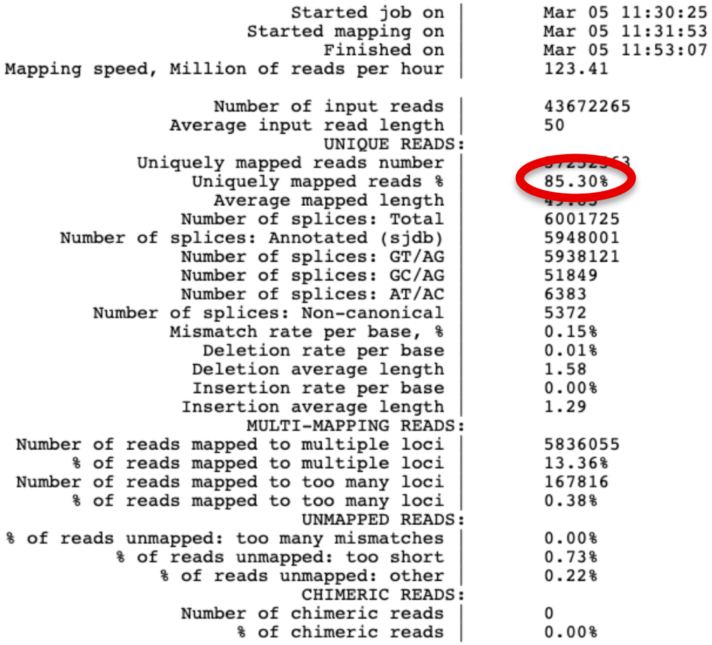


Number of reported alignments : 🟥 1 🟩 3
🟦 2 🟨 4

There are multiple aligned reads on this gene

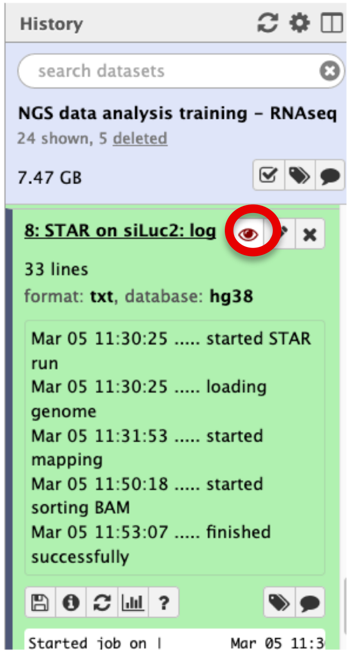
Exercise 2 - Question 1
Proportion of uniquely mapped reads

Galaxy : Shared Data → Data Libraries → NGS data analysis training RNAseq → alignment → log files :



```
                                 Started job on |       Mar 05 11:30:25
                             Started mapping on |       Mar 05 11:31:53
                                    Finished on |       Mar 05 11:53:07
       Mapping speed, Million of reads per hour |       123.41

                          Number of input reads |       43672265
                      Average input read length |       50
                                    UNIQUE READS:
                   Uniquely mapped reads number |       37252963
                        Uniquely mapped reads % |       85.30%
                          Average mapped length |       49.63
                       Number of splices: Total |       6001725
            Number of splices: Annotated (sjdb) |       5948001
                       Number of splices: GT/AG |       5938121
                       Number of splices: GC/AG |       51849
                       Number of splices: AT/AC |       6383
               Number of splices: Non-canonical |       5372
                      Mismatch rate per base, % |       0.15%
                         Deletion rate per base |       0.01%
                        Deletion average length |       1.58
                        Insertion rate per base |       0.00%
                       Insertion average length |       1.29
                             MULTI-MAPPING READS:
        Number of reads mapped to multiple loci |       5836055
             % of reads mapped to multiple loci |       13.36%
        Number of reads mapped to too many loci |       167816
             % of reads mapped to too many loci |       0.38%
                                  UNMAPPED READS:
       % of reads unmapped: too many mismatches |       0.00%
                 % of reads unmapped: too short |       0.73%
                     % of reads unmapped: other |       0.22%
                                  CHIMERIC READS:
                       Number of chimeric reads |       0
                            % of chimeric reads |       0.00%
```

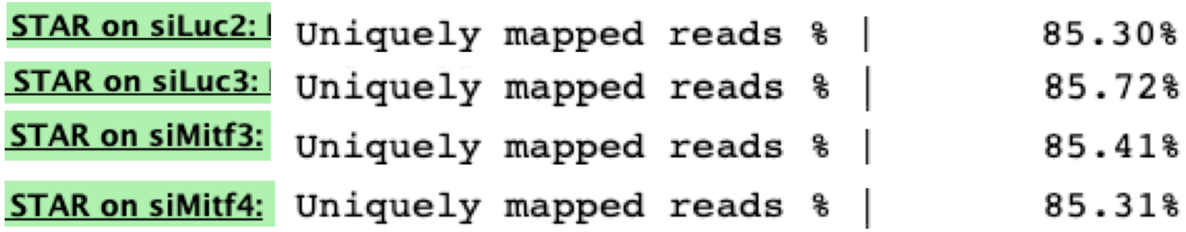


NGS data analysis training – RNAseq
24 shown, 5 deleted
7.47 GB

8: STAR on siLuc2: log
33 lines
format: txt, database: hg38

```
Mar 05 11:30:25 ..... started STAR run
Mar 05 11:30:25 ..... loading genome
Mar 05 11:31:53 ..... started mapping
Mar 05 11:50:18 ..... started sorting BAM
Mar 05 11:53:07 ..... finished successfully
```

Started job on | Mar 05 11:3



```
STAR on siLuc2: l   Uniquely mapped reads % |       85.30%
STAR on siLuc3: l   Uniquely mapped reads % |       85.72%
STAR on siMitf3:    Uniquely mapped reads % |       85.41%
STAR on siMitf4:    Uniquely mapped reads % |       85.31%
```

→ This proportion is consistent across samples

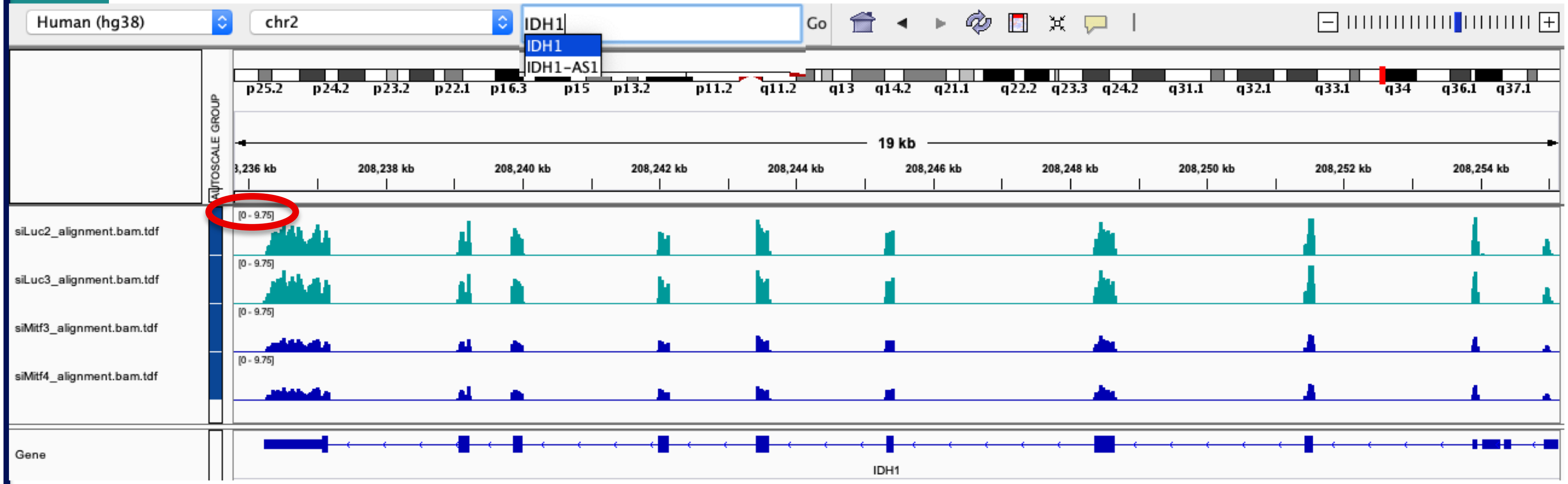
Exercise 2 – Question 2
Idh1 gene expression

IGV : File → Load from file and select the 4 tdf files
Select all tdf tracks → Right-click → Group Autoscale :

→ IGV automatically adjusts the Y scale to the data range currently in view (this scaling continually adjusts as you move)

→ all tracks are on the same scale

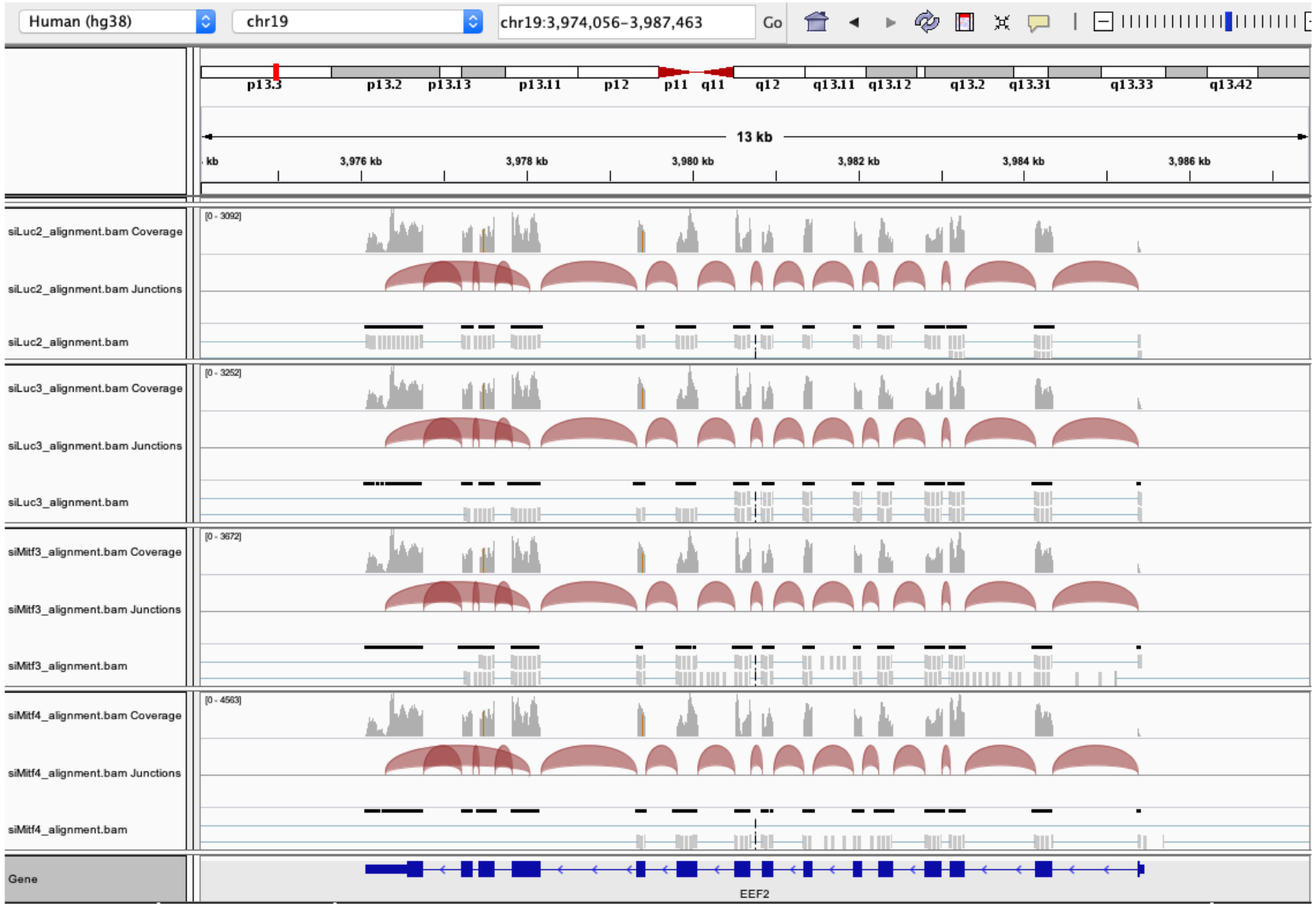
Search for Idh1



Idh1 is under-expressed in siMitf samples compared to siLuc ones

Exercise 2 – Question 3

- File → new session
- File → load from files and load the 4 BAM files
- Search for EEF2



Human (hg38)

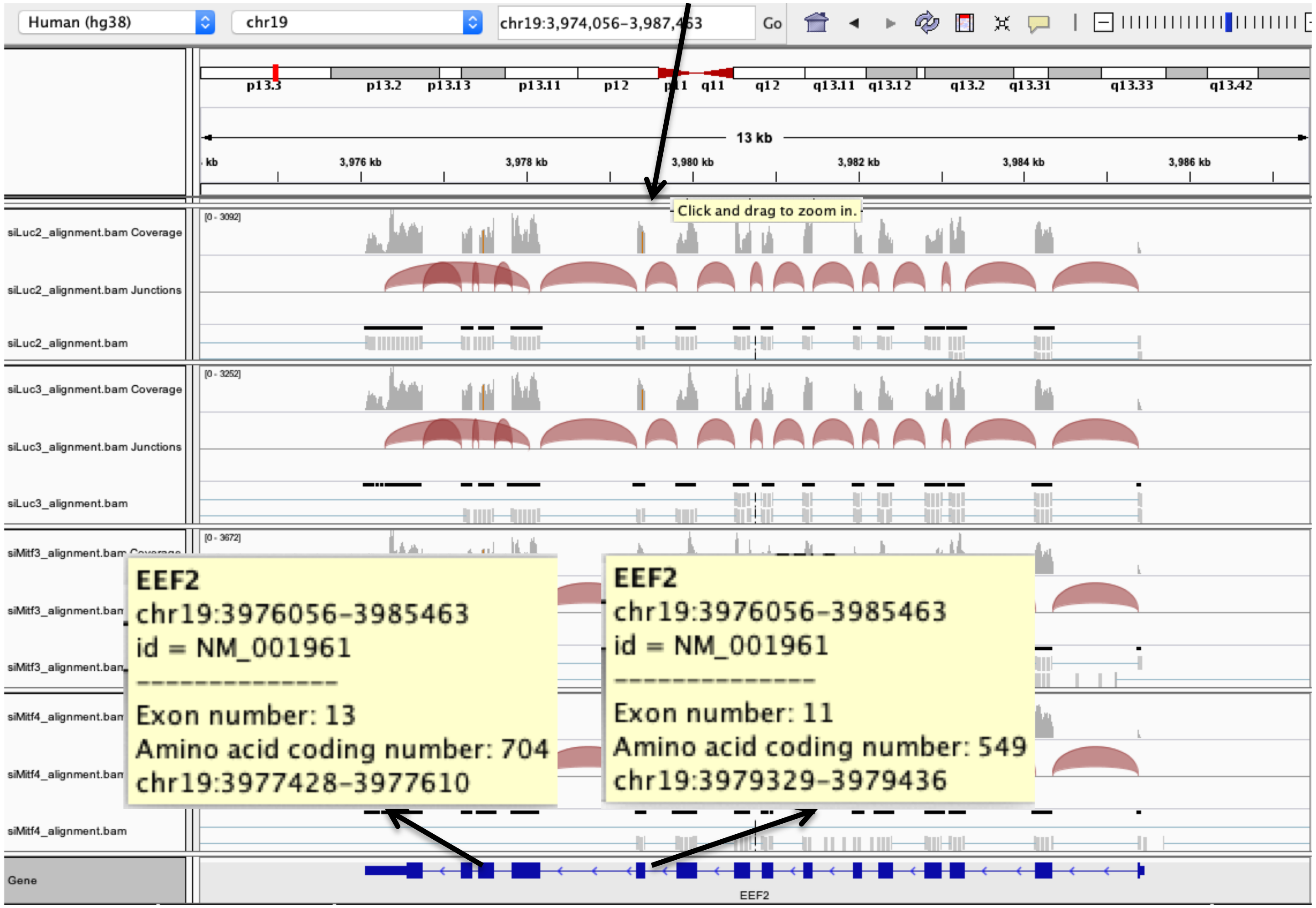
chr19

chr19:3,974,056-3,987,463

Exercise 2 – Question 3

Exon numbers are provided on annotation track

Click and drag on a region to zoom in



EEF2
chr19:3976056-3985463
id = NM_001961

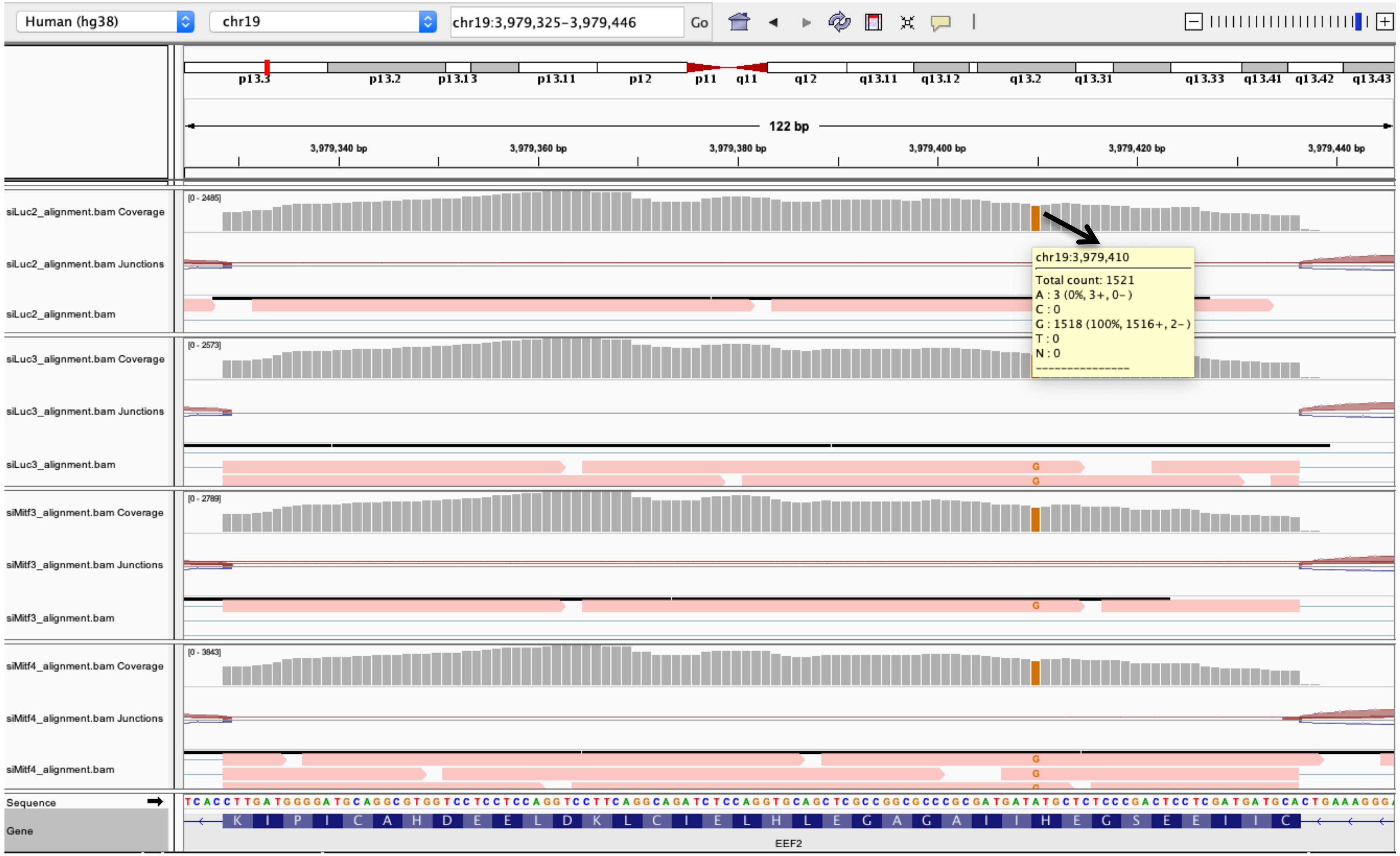
Exon number: 13
Amino acid coding number: 704
chr19:3977428-3977610

EEF2
chr19:3976056-3985463
id = NM_001961

Exon number: 11
Amino acid coding number: 549
chr19:3979329-3979436

Exercise 2 – Question 3

- *Eef2* exon 11
  - chr19:3,979,410 : G in ~100% of the reads, A in the genome

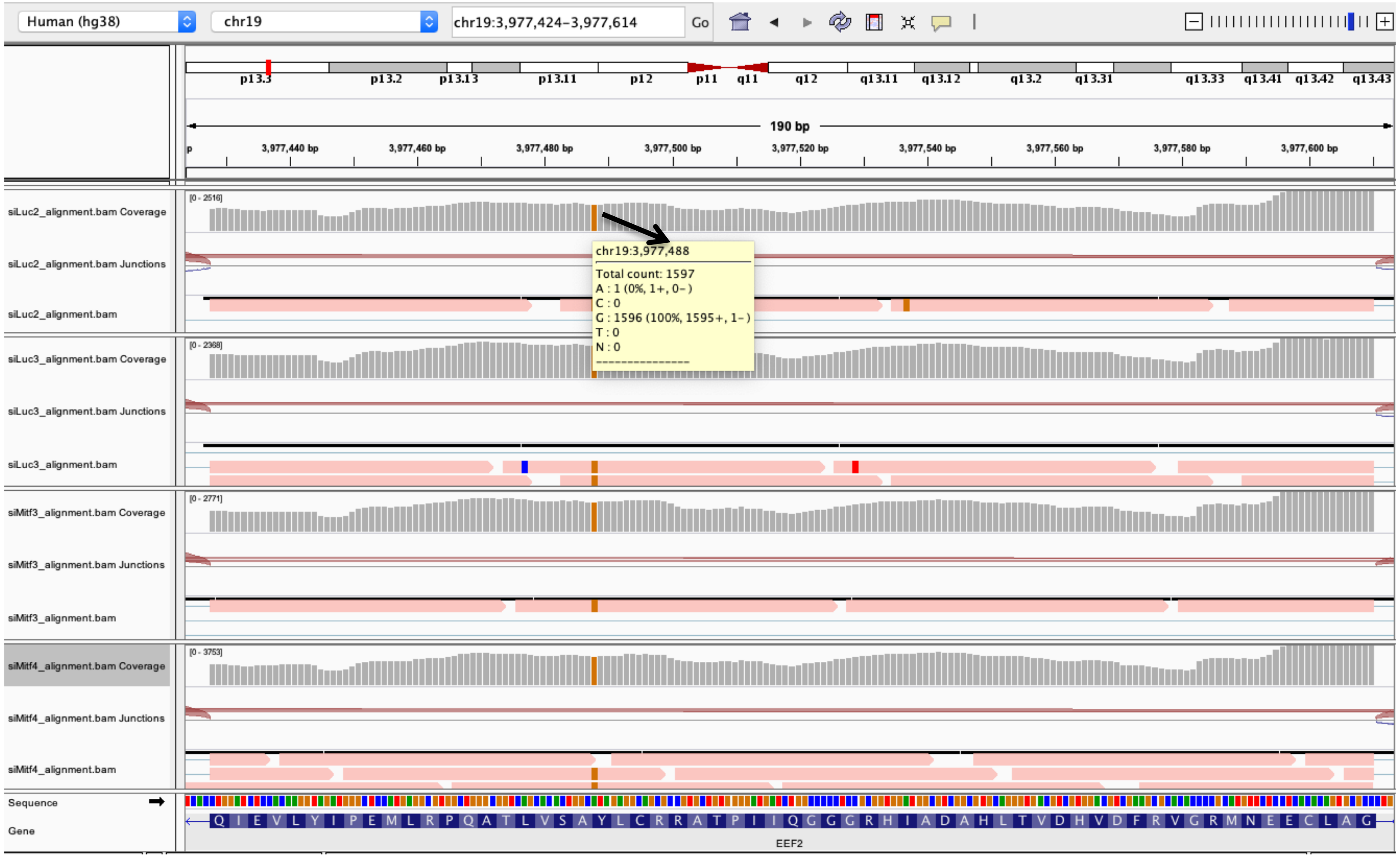


```
chr19:3,979,410
Total count: 1521
A : 3 (0%, 3+, 0– )
C : 0
G : 1518 (100%, 1516+, 2– )
T : 0
N : 0
---------------
```

Human (hg38)
chr19
chr19:3,979,325-3,979,446
Go
p13.3
p13.2
p13.13
p13.11
p12
p11
q11
q12
q13.11
q13.12
q13.2
q13.31
q13.33
q13.41
q13.42
q13.43
122 bp
3,979,340 bp
3,979,360 bp
3,979,380 bp
3,979,400 bp
3,979,420 bp
3,979,440 bp
siLuc2_alignment.bam Coverage
[0 - 2485]
siLuc2_alignment.bam Junctions
siLuc2_alignment.bam
siLuc3_alignment.bam Coverage
[0 - 2573]
siLuc3_alignment.bam Junctions
siLuc3_alignment.bam
siMitf3_alignment.bam Coverage
[0 - 2789]
siMitf3_alignment.bam Junctions
siMitf3_alignment.bam
siMitf4_alignment.bam Coverage
[0 - 3843]
siMitf4_alignment.bam Junctions
siMitf4_alignment.bam
G
Sequence
TCACCTTGATGGGGATGCAGGCGTGGTCCTCCTCCAGGTCCTTCAGGCAGATCTCCAGGTGCAGCTCGCCGGCGCCCGCGATGATATGCTCTCCCGACTCCTCGATGATGCACTGAAAGGG
Gene
K I P I C A H D E E L D K L C I E L H L E G A G A I I H E G S E E I I C
EEF2

Exercise 2 – Question 3

- *Eef2* exon 13
  - chr19:3,977,488 : G in ~100% of the reads, A in the genome



Human (hg38)

chr19

chr19:3,977,424-3,977,614

chr19:3,977,488

Total count: 1597

A : 1 (0%, 1+, 0-)

C : 0

G : 1596 (100%, 1595+, 1-)

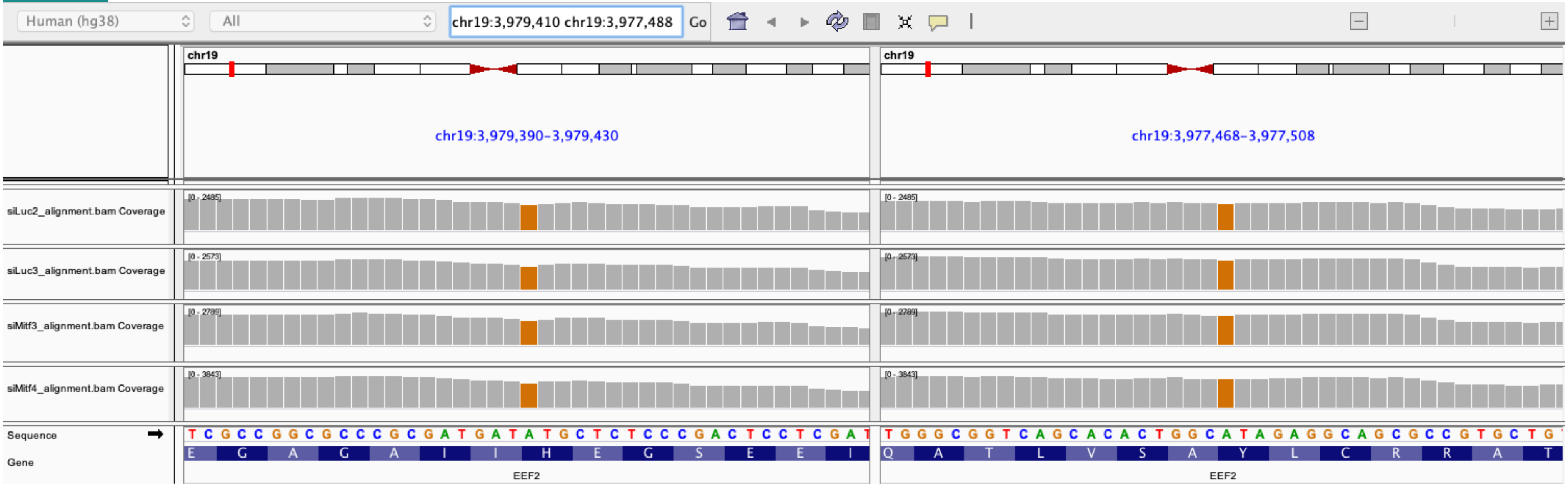
T : 0

N : 0

Go
p13.3
p13.2
p13.13
p13.11
p12
p11
q11
q12
q13.11
q13.12
q13.2
q13.31
q13.33
q13.41
q13.42
q13.43
190 bp
3,977,440 bp
3,977,460 bp
3,977,480 bp
3,977,500 bp
3,977,520 bp
3,977,540 bp
3,977,560 bp
3,977,580 bp
3,977,600 bp
siLuc2_alignment.bam Coverage
[0 - 2516]
siLuc2_alignment.bam Junctions
siLuc2_alignment.bam
siLuc3_alignment.bam Coverage
[0 - 2368]
siLuc3_alignment.bam Junctions
siLuc3_alignment.bam
siMitf3_alignment.bam Coverage
[0 - 2771]
siMitf3_alignment.bam Junctions
siMitf3_alignment.bam
siMitf4_alignment.bam Coverage
[0 - 3753]
siMitf4_alignment.bam Junctions
siMitf4_alignment.bam
Sequence
Gene
Q I E V L Y I P E M L R P Q A T L V S A Y L C R R A T P I I Q G G G R H I A D A H L T V D H V D F R V G R M N E E C L A G
EEF2

Exercise 2 – Question 3

- It is also possible to visualize several regions on IGV
  - Enter several locations or genes in the search box, separated by space
  - Click on 🏠 to go back to genome view

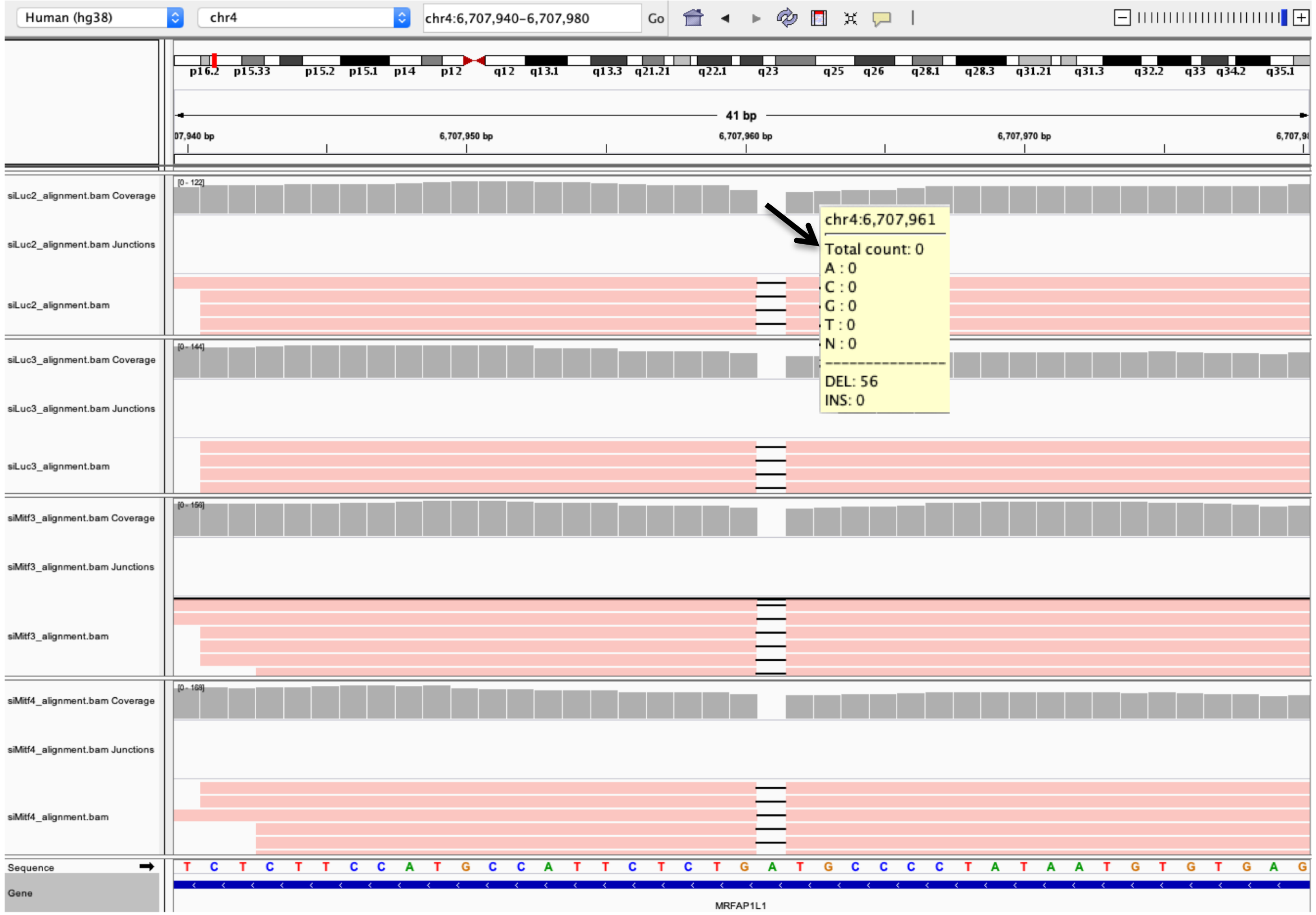


chr19:3,979,410 chr19:3,977,488

Human (hg38)
All
Go
chr19
chr19:3,979,390-3,979,430
chr19
chr19:3,977,468-3,977,508
siLuc2_alignment.bam Coverage
[0 - 2485]
siLuc3_alignment.bam Coverage
[0 - 2573]
siMitf3_alignment.bam Coverage
[0 - 2789]
siMitf4_alignment.bam Coverage
[0 - 3843]
Sequence
Gene
T C G C C G G C G C C C G C G A T G A T A T G C T C T C C C G A C T C C T C G A T
E G A G A I I H E G S E E I
EEF2
T G G G C G G T C A G C A C A C T G G C A T A G A G G C A G C G C C G T G C T G
Q A T L V S A Y L C R R A T
EEF2

Exercise 2 – Question 4

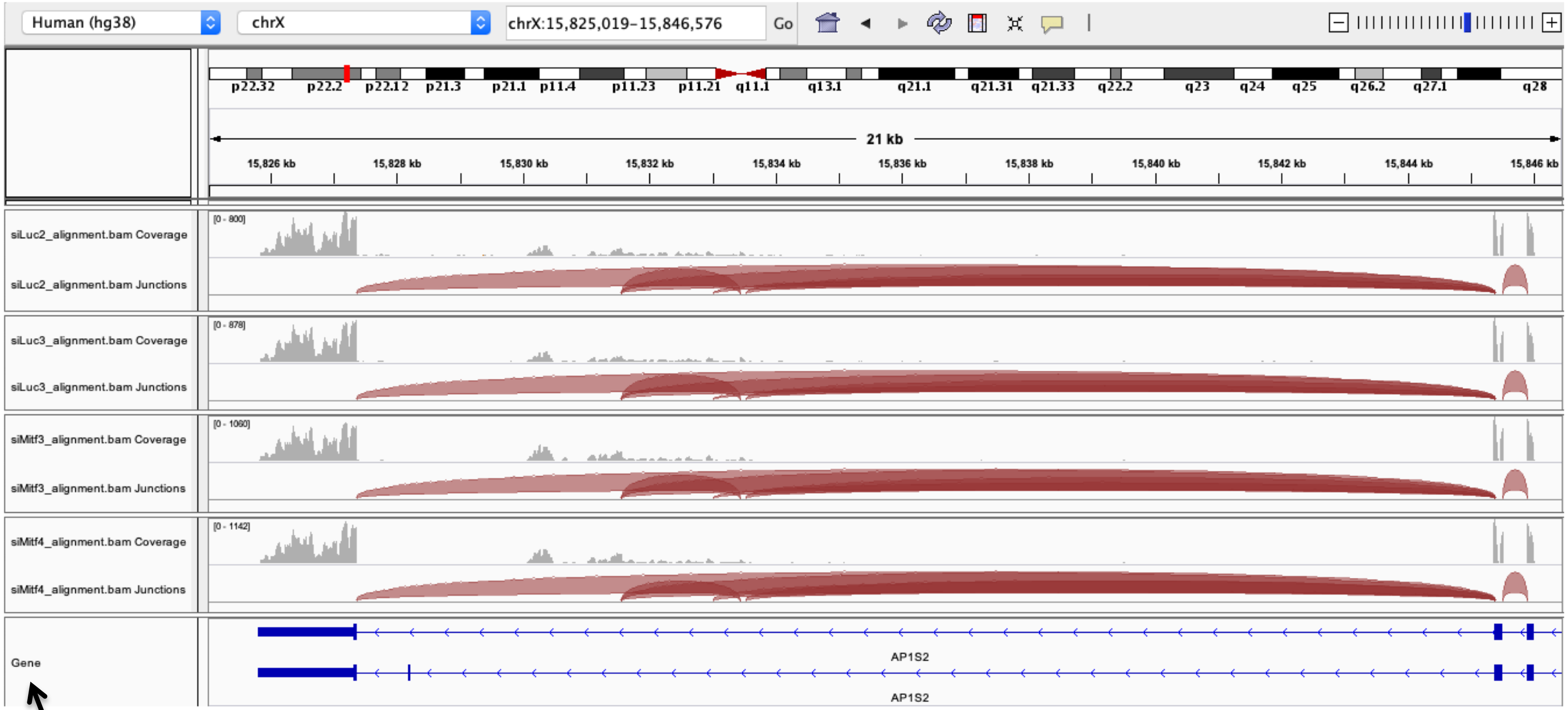
- Position chr4:6707960-6707961 :
  - Deletion vs reference genome



```
chr4:6,707,961
Total count: 0
A : 0
C : 0
G : 0
T : 0
N : 0
---------------
DEL: 56
INS: 0
```

Exercise 2 – Question 5

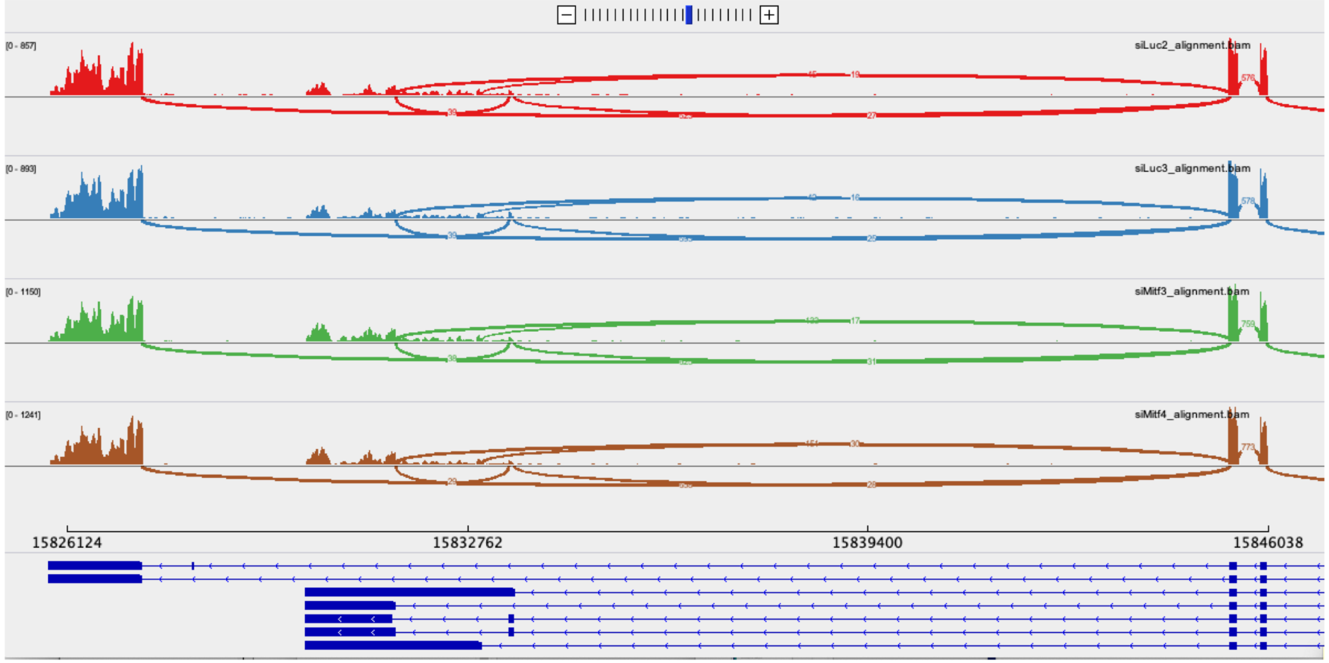
- Region chrX:15,825,019-15,846,576 :
  - We observe junctions corresponding to several isoforms of AP1S2



Right click on the annotation track and select Expanded to visualize all isoforms

Exercise 2 – Question 5

- Region chrX:15,825,019-15,846,576 :
  - We observe junctions corresponding to several isoforms of AP1S2
  - Sashimi-plot :
    - Right-click on a BAM track → Sashimi plot → Select Alignment Tracks : all alignments

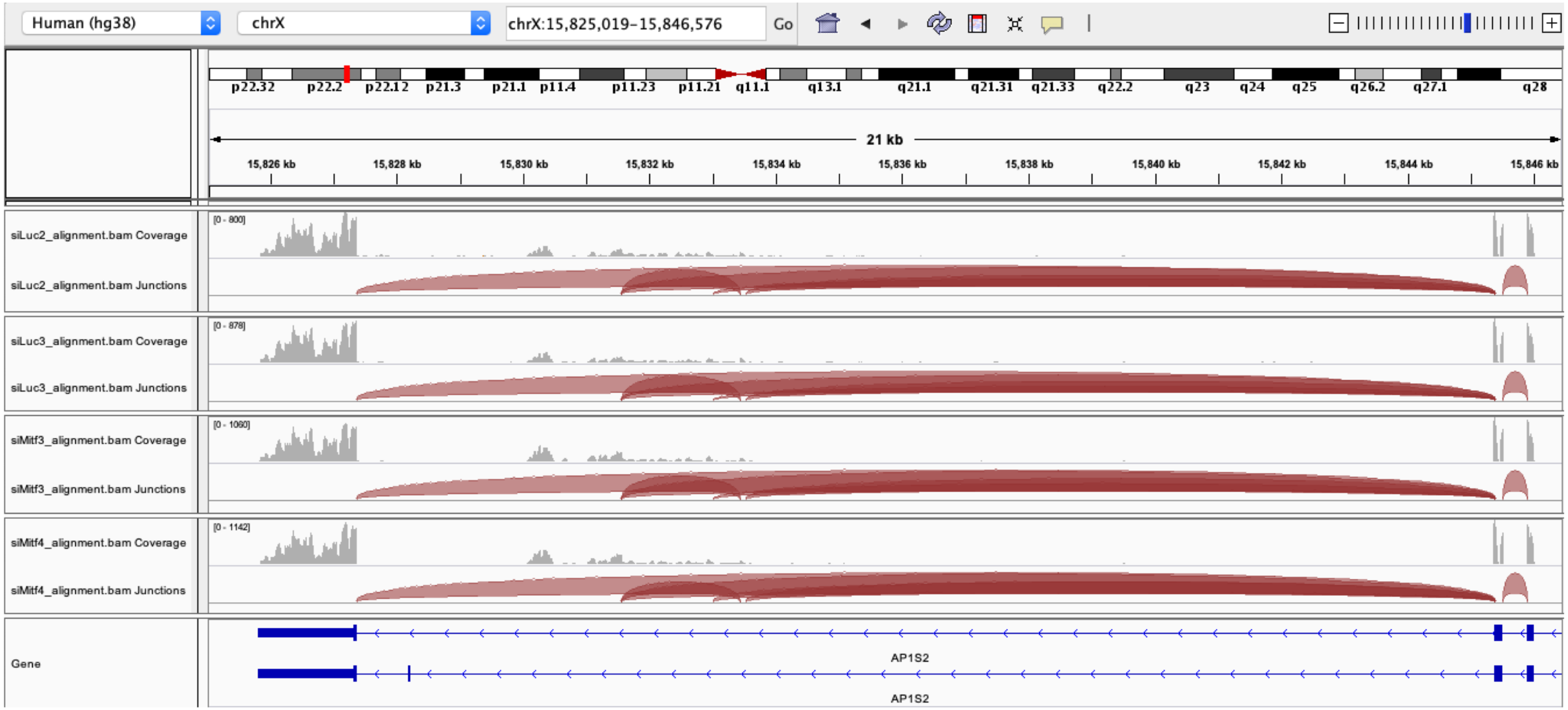


→ Very useful to quickly visualize splicing events along genomic regions of interest

→ **More accurate with paired-end data**

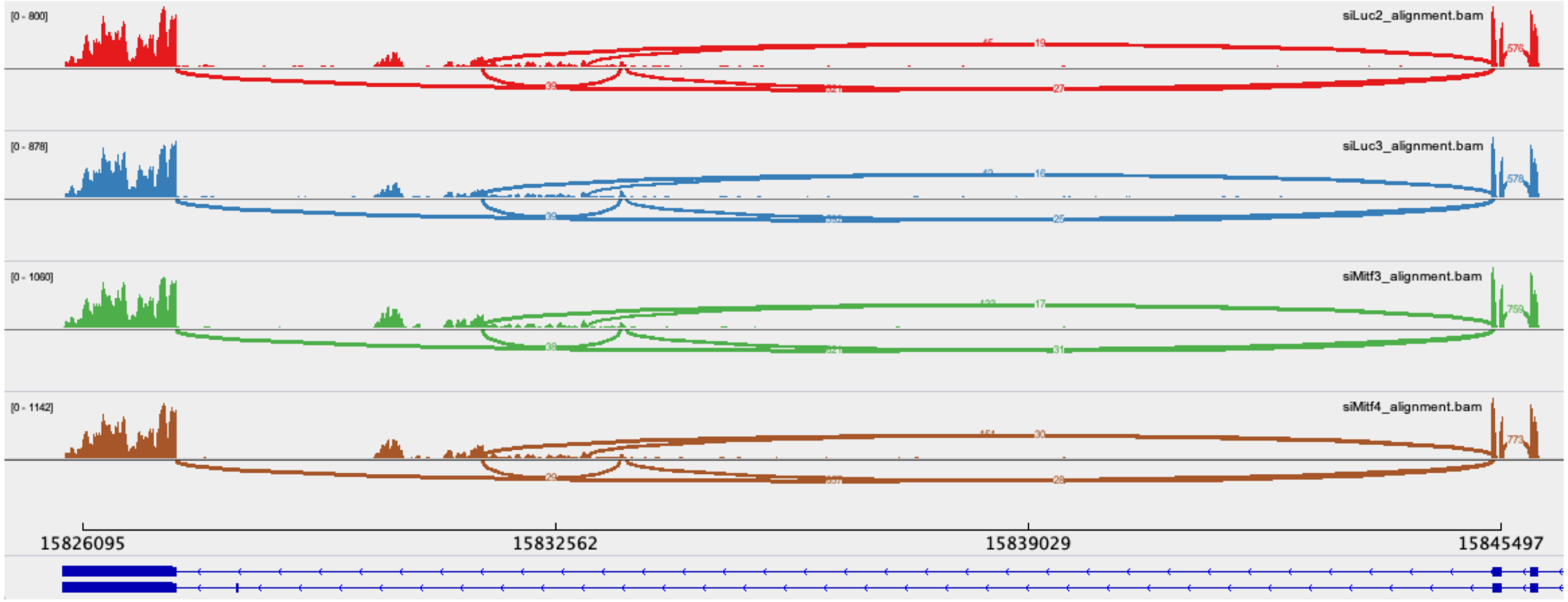
Exercise 2 – Question 5

- March 2019 : these isoforms were not annotated in Refseq



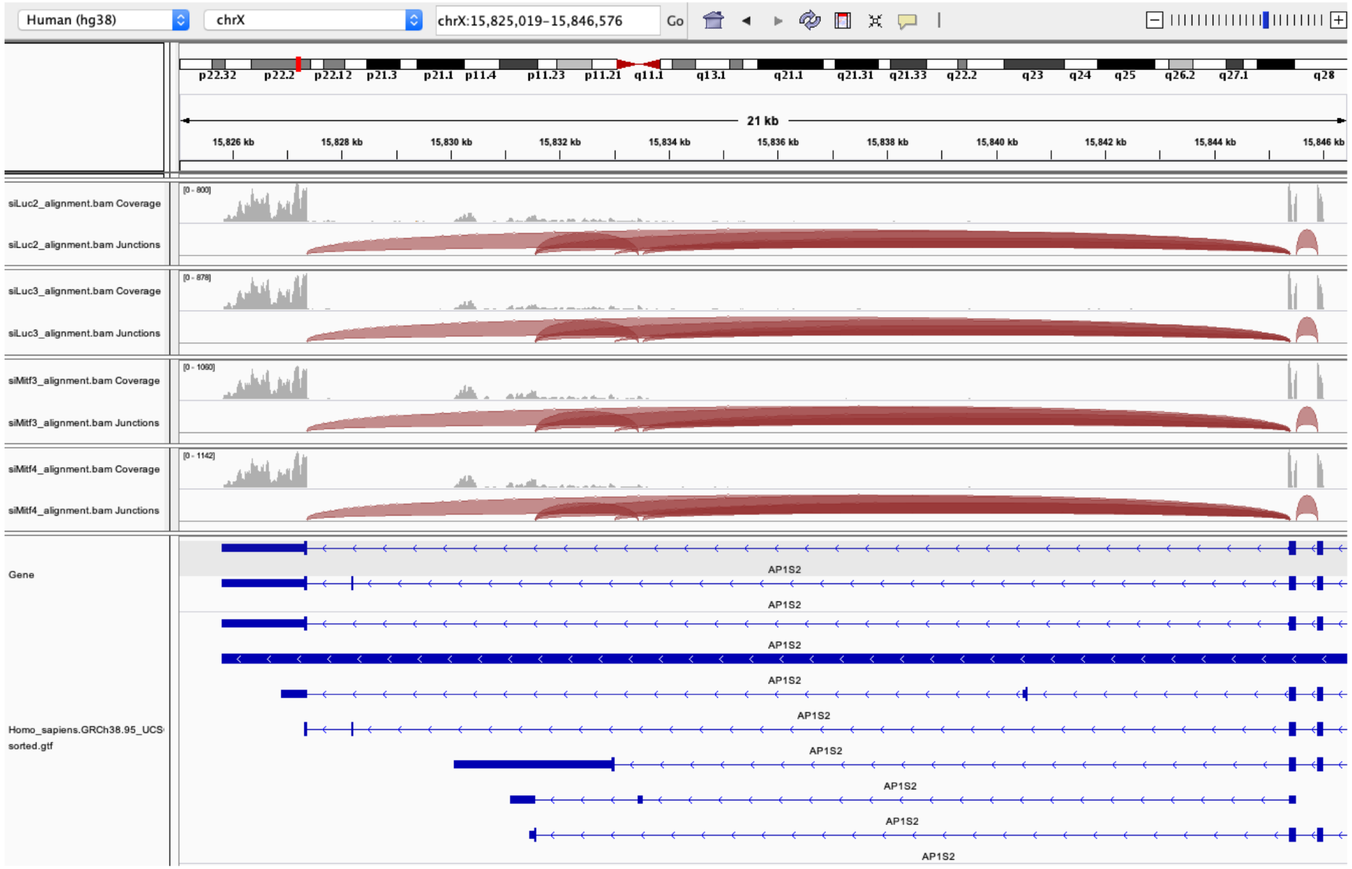
Exercise 2 – Question 5

- March 2019 : these isoforms were not annotated in Refseq
  - Sashimi plot :



Exercise 2 – Question 5

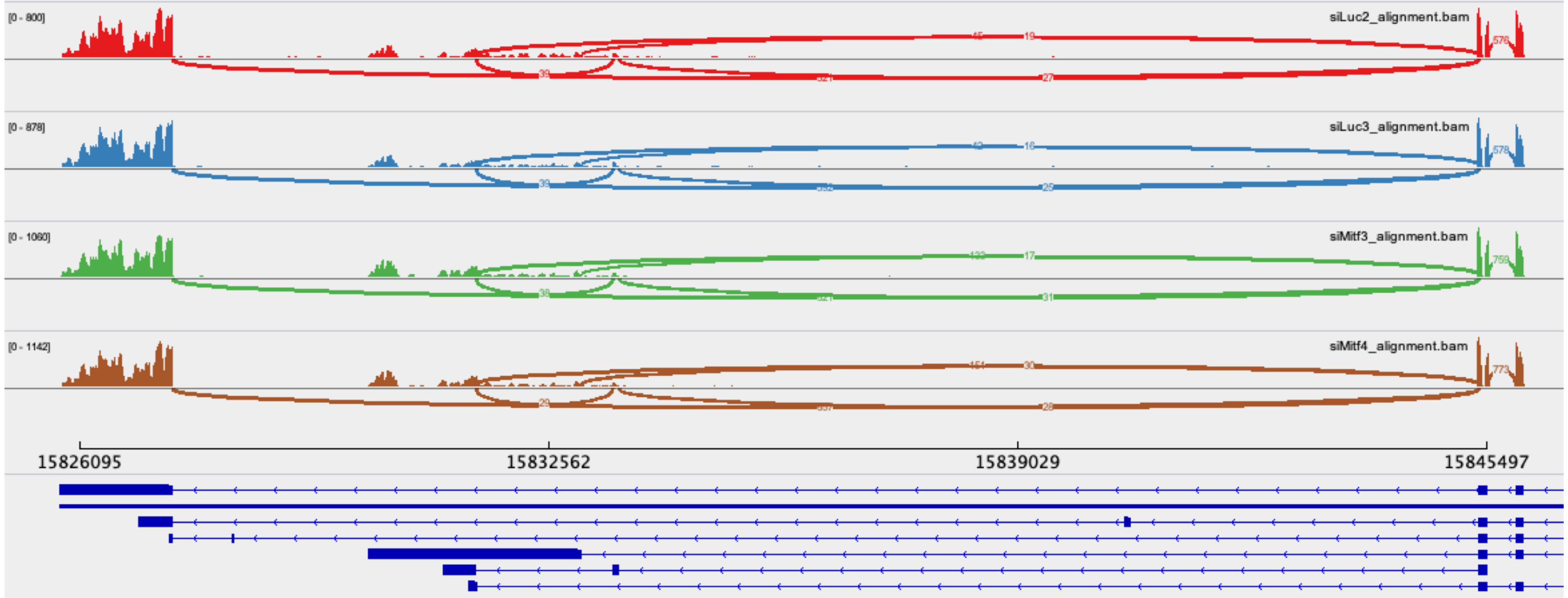
- March 2019 : these isoforms were not annotated in Refseq
  - But more exons annotated in this region in Ensembl
    - File → load from file → Homo_sapiens.GRCh38.95_UCSC_chr.sorted.gtf
    - Right-click on the annotation track and select Expanded



Human (hg38) | chrX | chrX:15,825,019-15,846,576 | Go

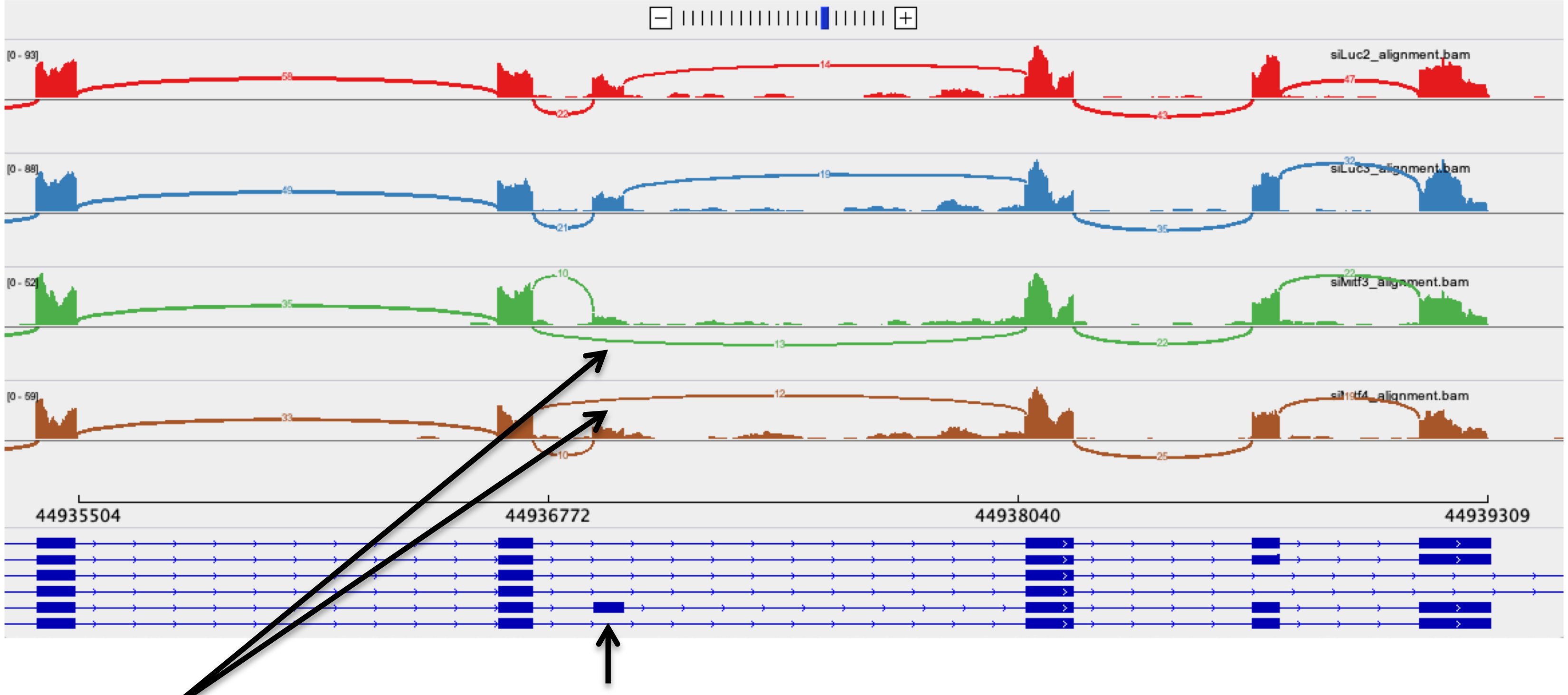
Exercise 2 – Question 5

- March 2019 : these isoforms were not annotated in Refseq
  - But more exons annotated in this region in Ensembl
    - Sashimi plot with Ensembl annotations :



Exercise 2 – question 6

- Region chr20:44,935,294-44,939,521 :
  - Sashimi-plot

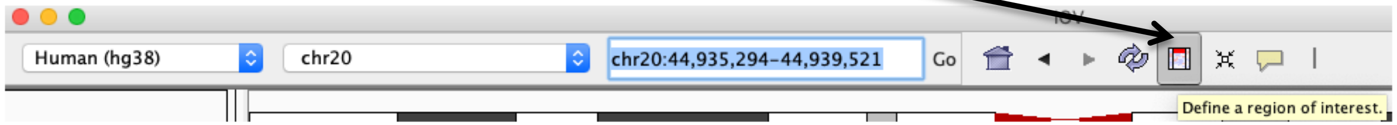


We detect an isoform without this exon in siMitf samples

IGV is only a visualization tool
In-depth analysis using paired-end data with more coverage is needed

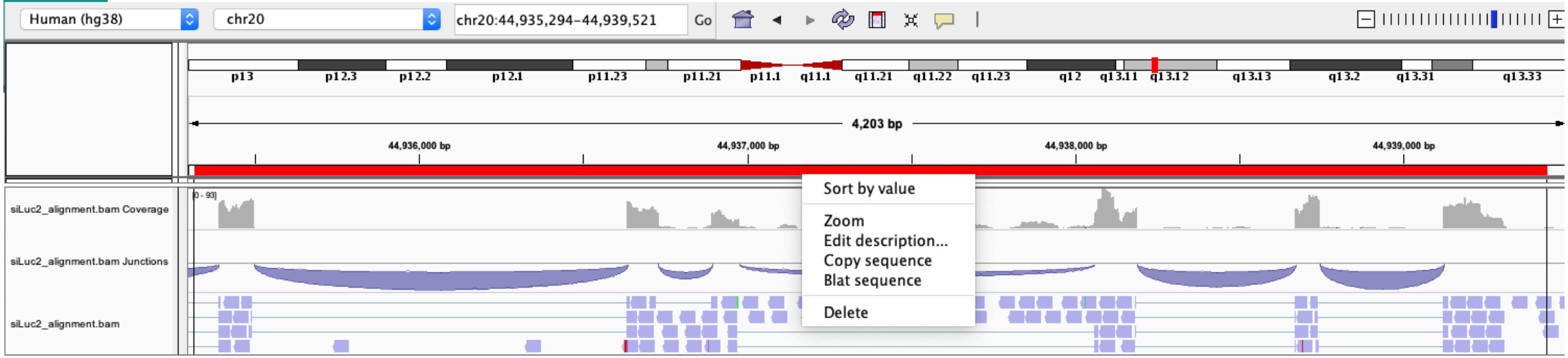
Exercise 2 – question 6

- If you want to save this region :
  - Click on define a region of interest



Human (hg38) | chr20 | chr20:44,935,294-44,939,521

  - Click on a track to define the start and end position of your region of interest → a red bar appears
  - Give a name to this region (Right-click on the bar → edit description)
  - Go to Regions → Region Navigator to display again this region



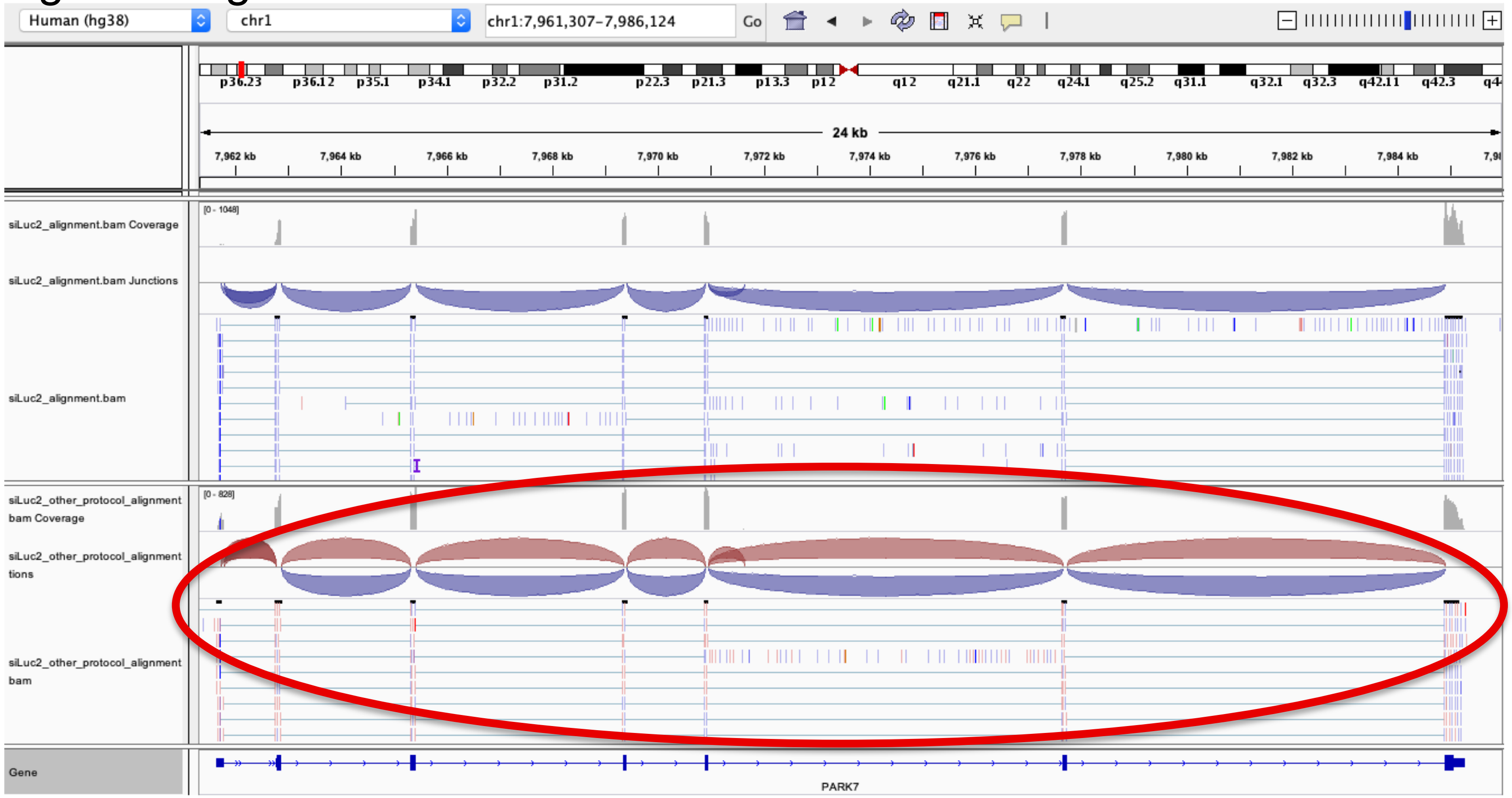
Human (hg38) | chr20 | chr20:44,935,294-44,939,521

Exercise 2 – question 6

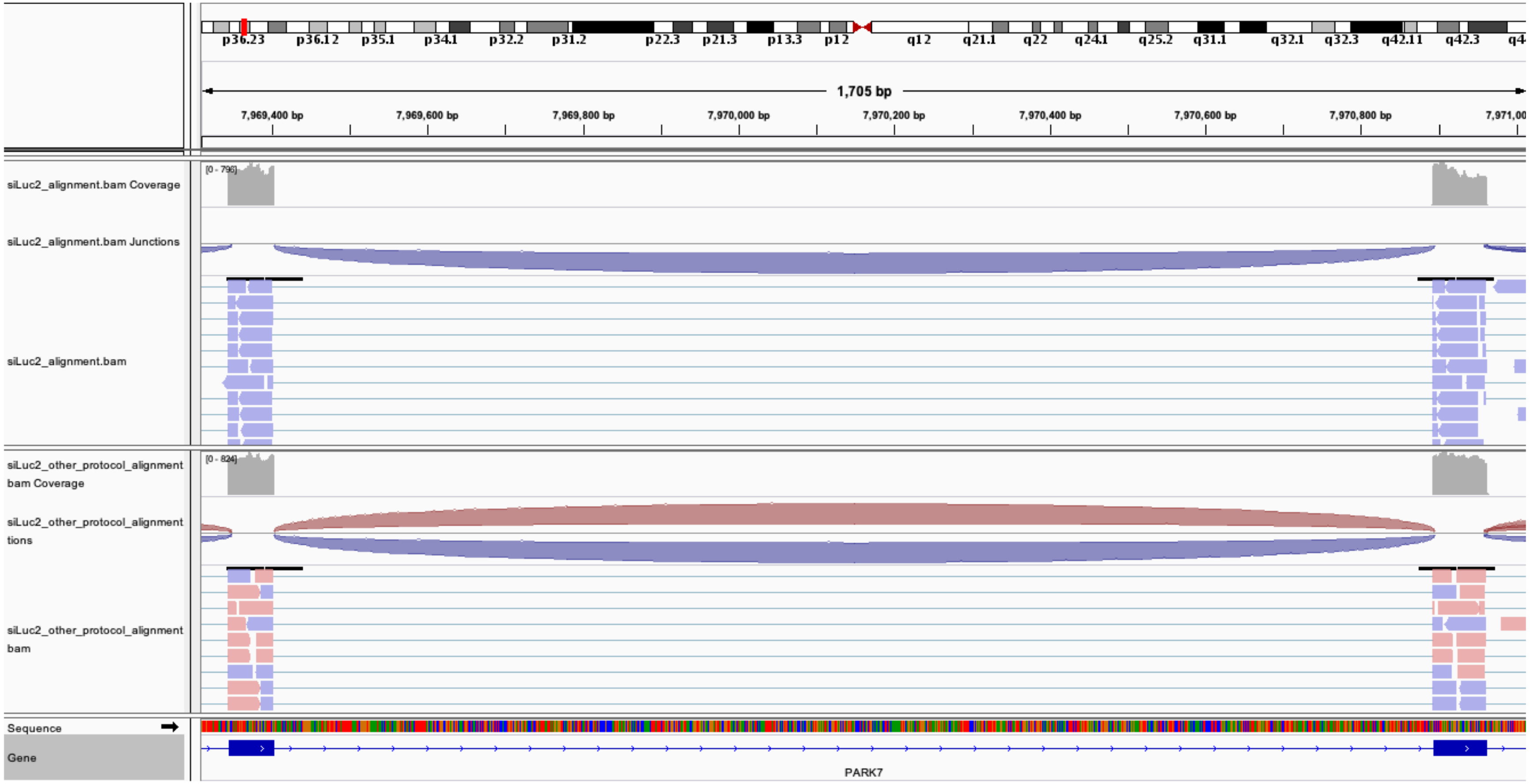
- You can save your IGV session
  - To save the current state of your IGV session to a named session file
  - File → Save Session
  - Data files must stay at the same location
- Use File → Open session to restore a saved session

Exercise 2 – Question 7

- Remove siLuc3 and siMitf3/4 tracks (Right click on tracks → Remove track)
- File → load from file and select siLuc2_other_protocol_alignment.bam
- Right-click on BAM file → Color alignments by → read strand
- e.g. *Park7* gene



Exercise 2 – Question 7



→ This protocol is not directional (it does not preserve strand information)

You can display alignments grouped by read strand
(right-click on BAM track → Group alignments by → read strand)