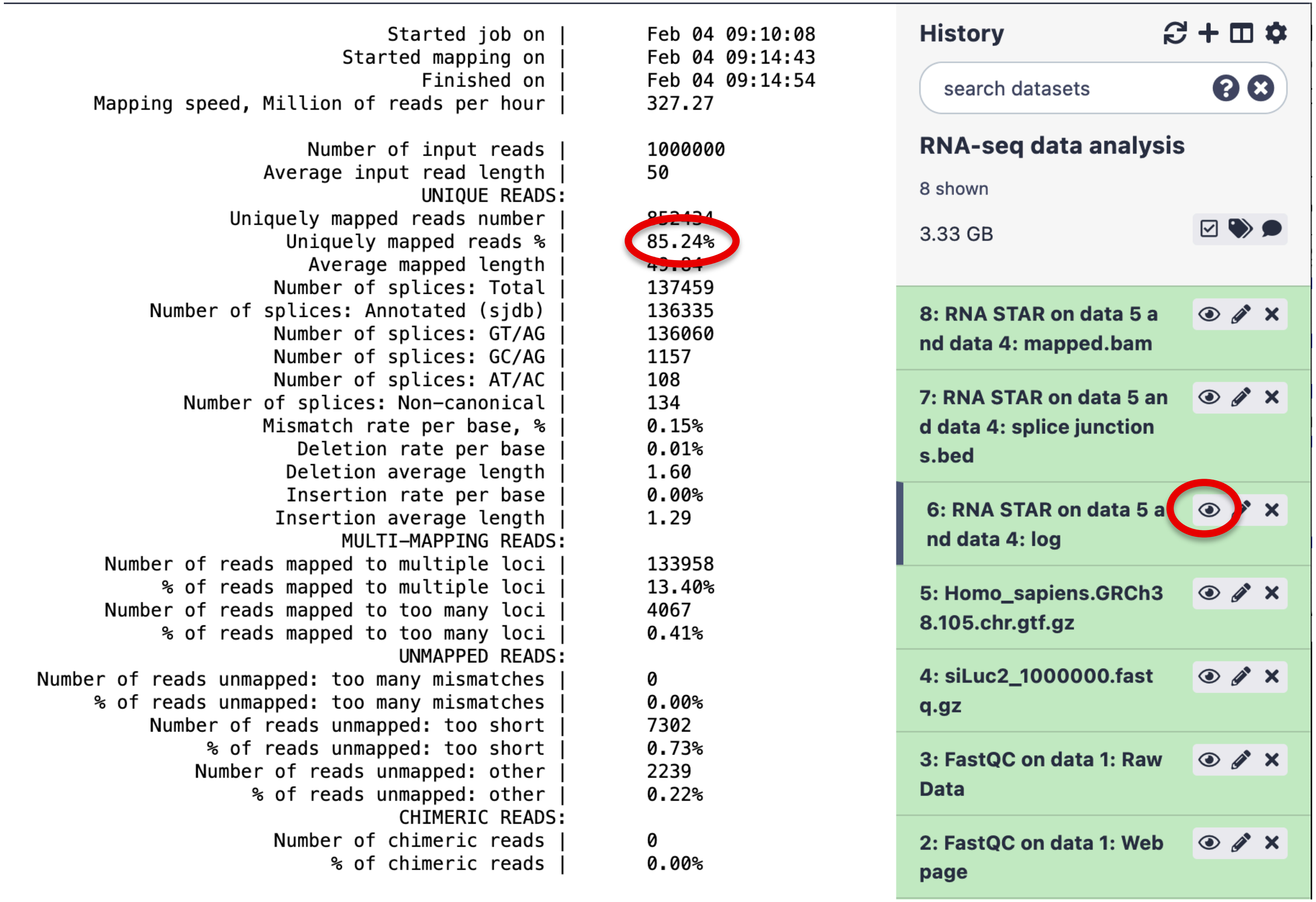
NGS read mapping : answers to questions

Céline Keime
keime@igbmc.fr

Exercise 1
1. Log file

Proportion of uniquely mapped reads :



```
                                 Started job on |      Feb 04 09:10:08
                             Started mapping on |      Feb 04 09:14:43
                                    Finished on |      Feb 04 09:14:54
       Mapping speed, Million of reads per hour |      327.27

                          Number of input reads |      1000000
                      Average input read length |      50
                                    UNIQUE READS:
                   Uniquely mapped reads number |      852434
                        Uniquely mapped reads % |      85.24%
                          Average mapped length |      49.84
                       Number of splices: Total |      137459
            Number of splices: Annotated (sjdb) |      136335
                       Number of splices: GT/AG |      136060
                       Number of splices: GC/AG |      1157
                       Number of splices: AT/AC |      108
               Number of splices: Non-canonical |      134
                      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 |      133958
             % of reads mapped to multiple loci |      13.40%
        Number of reads mapped to too many loci |      4067
             % of reads mapped to too many loci |      0.41%
                                  UNMAPPED READS:
  Number of reads unmapped: too many mismatches |      0
       % of reads unmapped: too many mismatches |      0.00%
            Number of reads unmapped: too short |      7302
                 % of reads unmapped: too short |      0.73%
                Number of reads unmapped: other |      2239
                     % of reads unmapped: other |      0.22%
                                  CHIMERIC READS:
                       Number of chimeric reads |      0
                            % of chimeric reads |      0.00%
```

History

search datasets

RNA-seq data analysis

8 shown

3.33 GB

8: RNA STAR on data 5 and data 4: mapped.bam

7: RNA STAR on data 5 and data 4: splice junctions.bed

6: RNA STAR on data 5 and data 4: log

5: Homo_sapiens.GRCh38.105.chr.gtf.gz

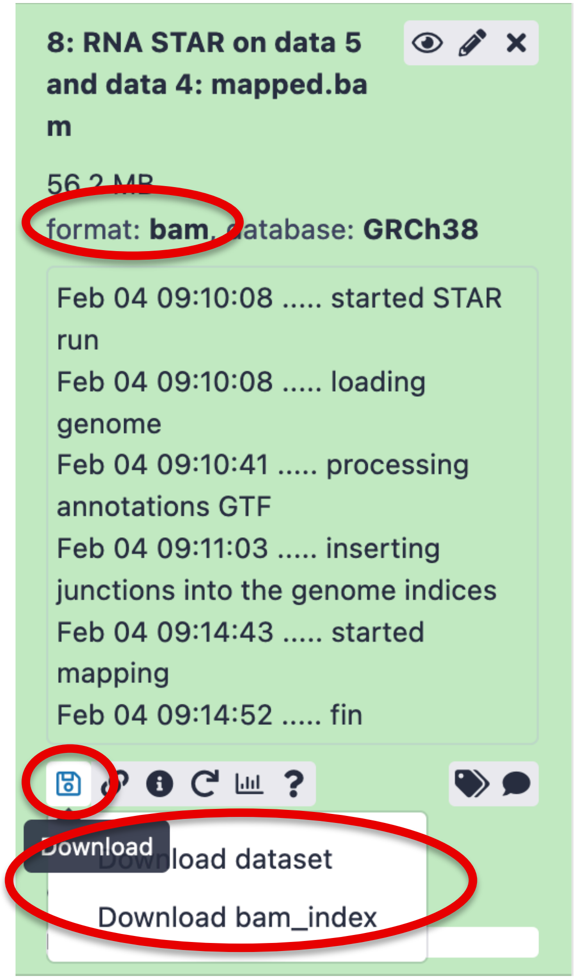
4: siLuc2_1000000.fastq.gz

3: FastQC on data 1: Raw Data

2: FastQC on data 1: Web page

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)



8: RNA STAR on data 5 and data 4: mapped.bam

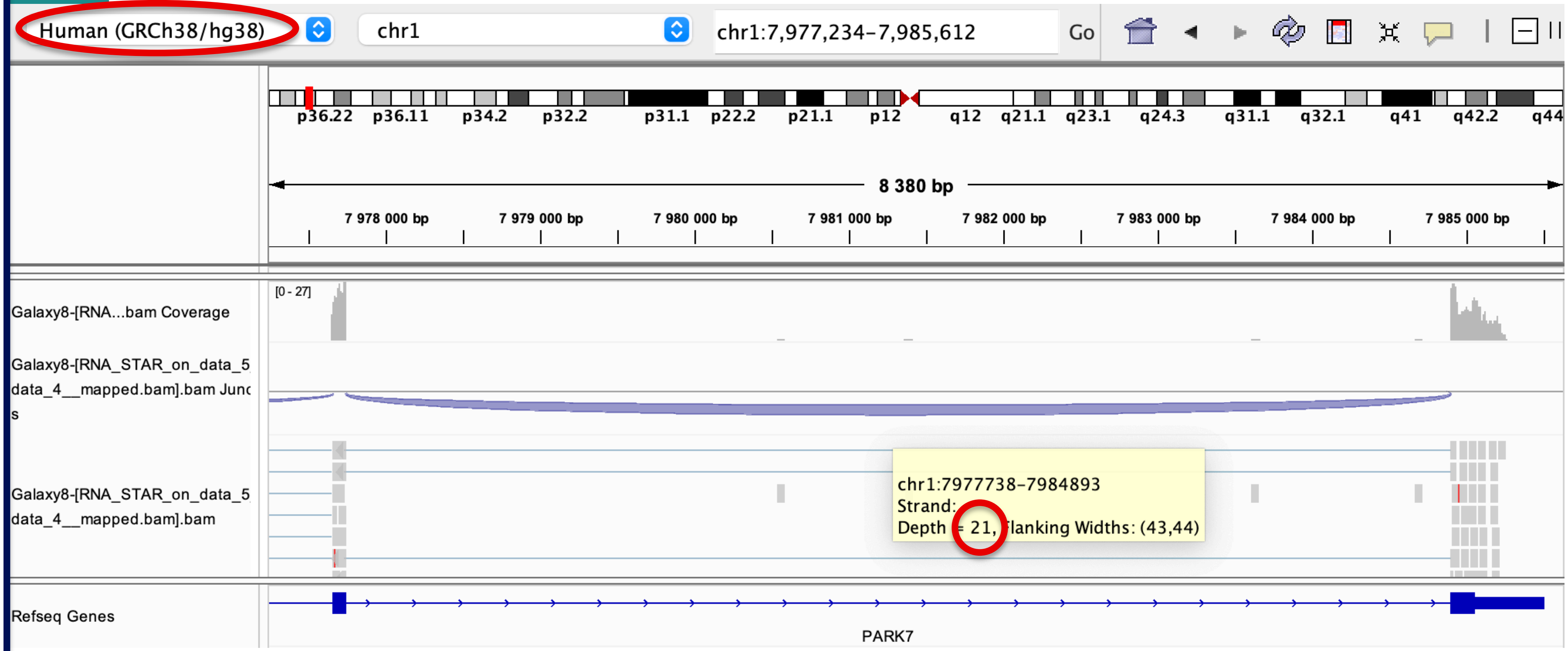
56.2 MB

format: **bam**, database: **GRCh38**

```
Feb 04 09:10:08 ..... started STAR run
Feb 04 09:10:08 ..... loading genome
Feb 04 09:10:41 ..... processing annotations GTF
Feb 04 09:11:03 ..... inserting junctions into the genome indices
Feb 04 09:14:43 ..... started mapping
Feb 04 09:14:52 ..... fin
```

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

Exercise 1
2. Splice junction



Human (GRCh38/hg38)

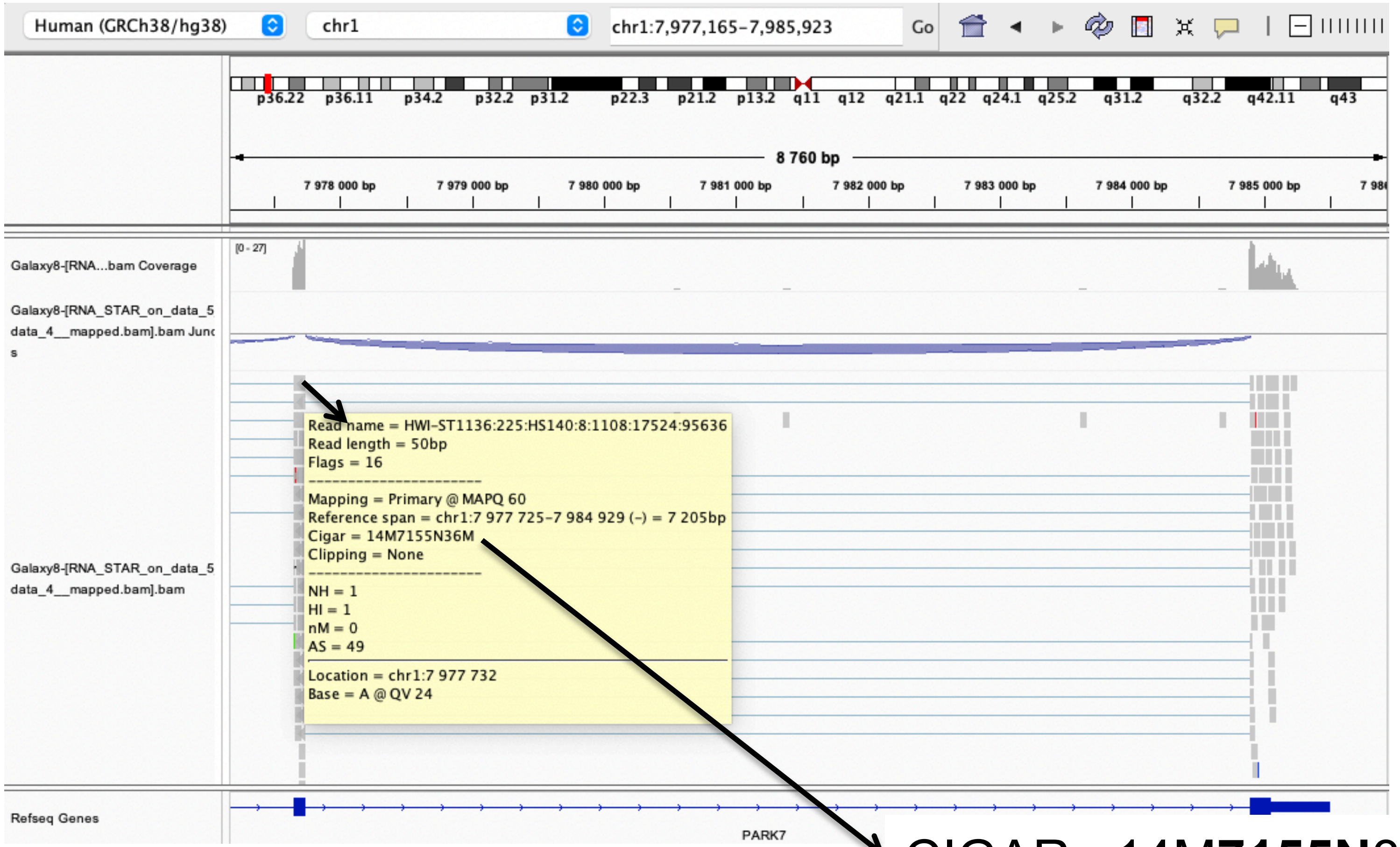
chr1

chr1:7,977,234-7,985,612

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

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

Exercise 1
2. Splice junction



Read name = HWI-ST1136:225:HS140:8:1108:17524:95636
Read length = 50bp
Flags = 16

Mapping = Primary @ MAPQ 60
Reference span = chr1:7 977 725-7 984 929 (-) = 7 205bp
Cigar = 14M7155N36M
Clipping = None

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

Location = chr1:7 977 732
Base = A @ QV 24

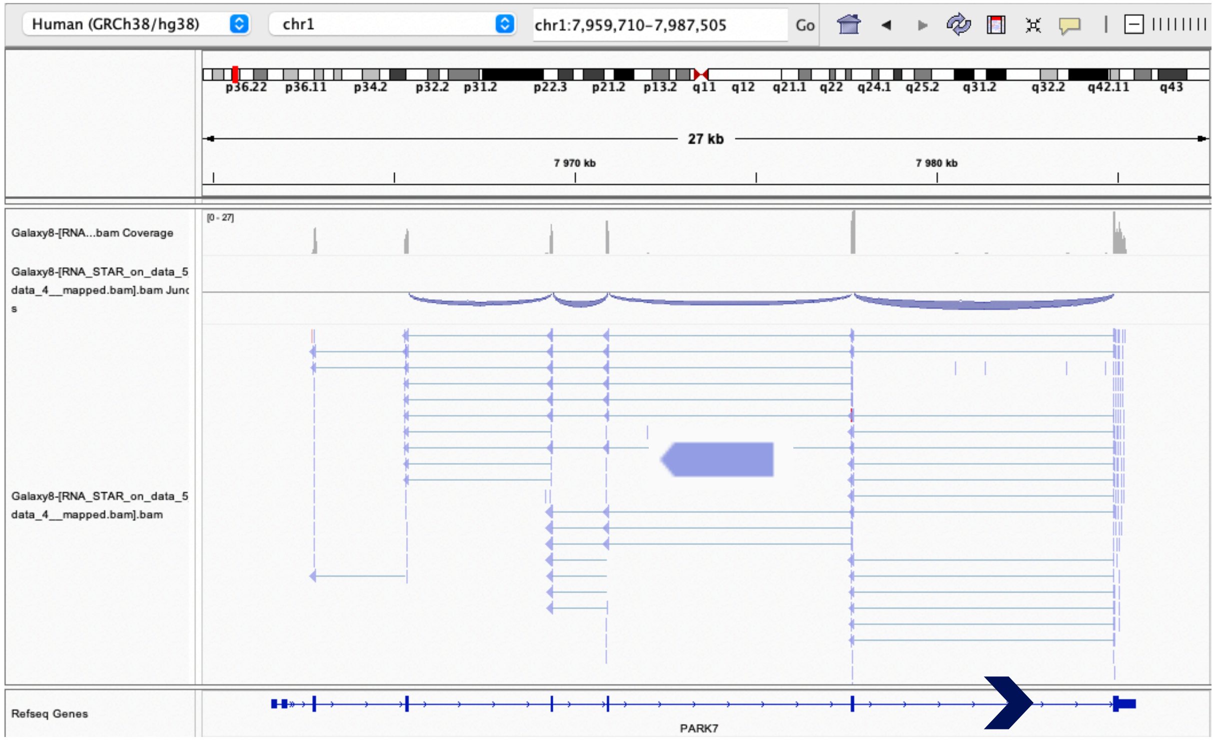
CIGAR : 14M**7155N**36M

Intron length :
7984893 - 7977738 = 7155

Exercise 1
2. Strand specificity

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

Park7 :



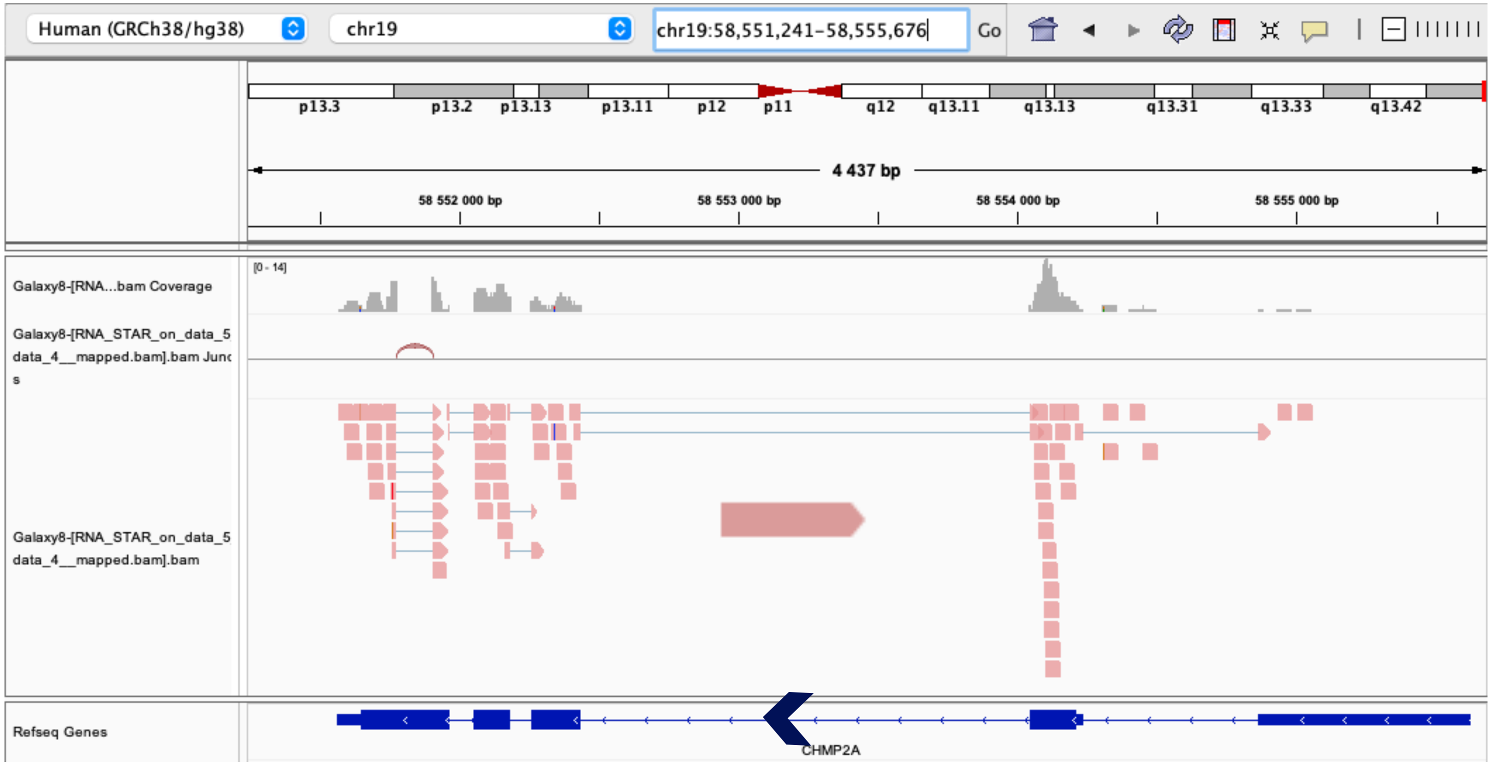
Human (GRCh38/hg38) chr1 chr1:7,959,710-7,987,505

Go
p36.22
p36.11
p34.2
p32.2
p31.2
p22.3
p21.2
p13.2
q11
q12
q21.1
q22
q24.1
q25.2
q31.2
q32.2
q42.11
q43
27 kb
7 970 kb
7 980 kb
[0 - 27]
Galaxy8-[RNA...bam Coverage
Galaxy8-[RNA_STAR_on_data_5 data_4__mapped.bam].bam Juncs
Galaxy8-[RNA_STAR_on_data_5 data_4__mapped.bam].bam
Refseq Genes
PARK7

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 :

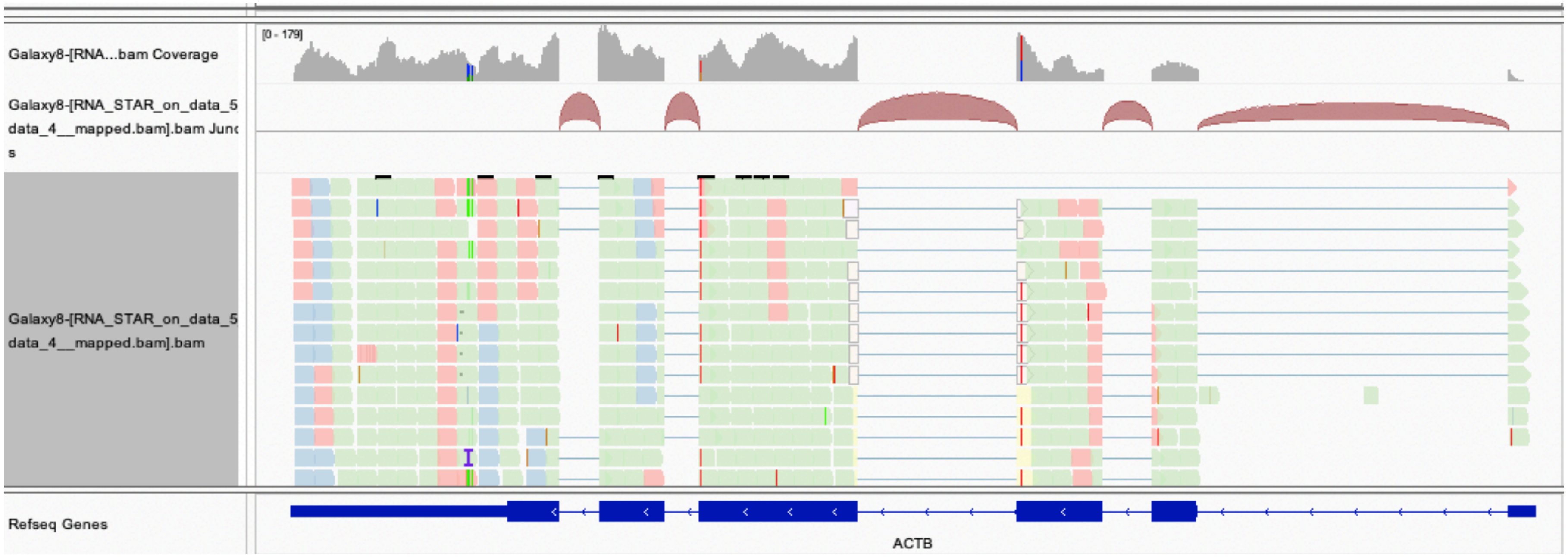


Human (GRCh38/hg38)   chr19   chr19:58,551,241-58,555,676   Go

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

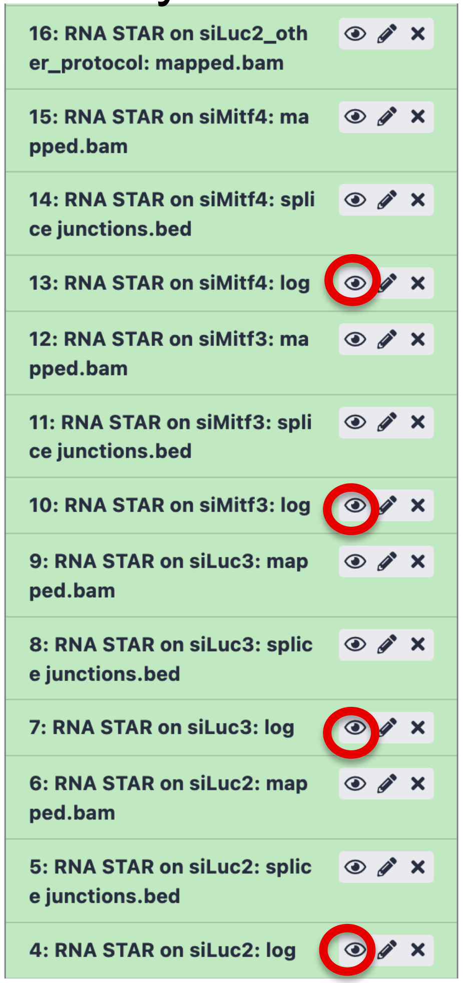

Galaxy8-[RNA...bam Coverage
[0 - 179]
Galaxy8-[RNA_STAR_on_data_5 data_4__mapped.bam].bam Junc s
Galaxy8-[RNA_STAR_on_data_5 data_4__mapped.bam].bam
Refseq Genes
ACTB

Number of reported alignments
→ see NH tag in pop-up windows to visualize
color-coding (that can be different from this one) : 🟩 1 🟥 2 🟦 3

There are multiple aligned reads on this gene

Exercise 2 - Question 1
Proportion of uniquely mapped reads

Galaxy : “NGS data analysis training Strasbourg” history



16: RNA STAR on siLuc2_other_protocol: mapped.bam

15: RNA STAR on siMitf4: mapped.bam

14: RNA STAR on siMitf4: splice junctions.bed

13: RNA STAR on siMitf4: log

12: RNA STAR on siMitf3: mapped.bam

11: RNA STAR on siMitf3: splice junctions.bed

10: RNA STAR on siMitf3: log

9: RNA STAR on siLuc3: mapped.bam

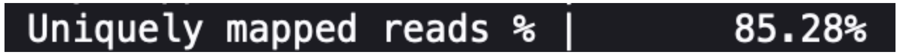
8: RNA STAR on siLuc3: splice junctions.bed

7: RNA STAR on siLuc3: log

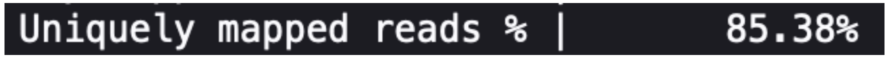
6: RNA STAR on siLuc2: mapped.bam

5: RNA STAR on siLuc2: splice junctions.bed

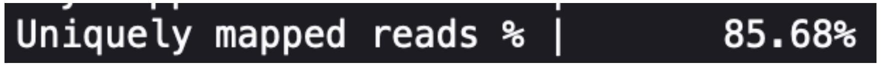
4: RNA STAR on siLuc2: log



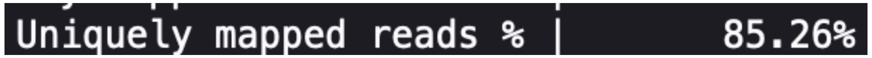
```
Uniquely mapped reads % |          85.28%
```



```
Uniquely mapped reads % |          85.38%
```



```
Uniquely mapped reads % |          85.68%
```

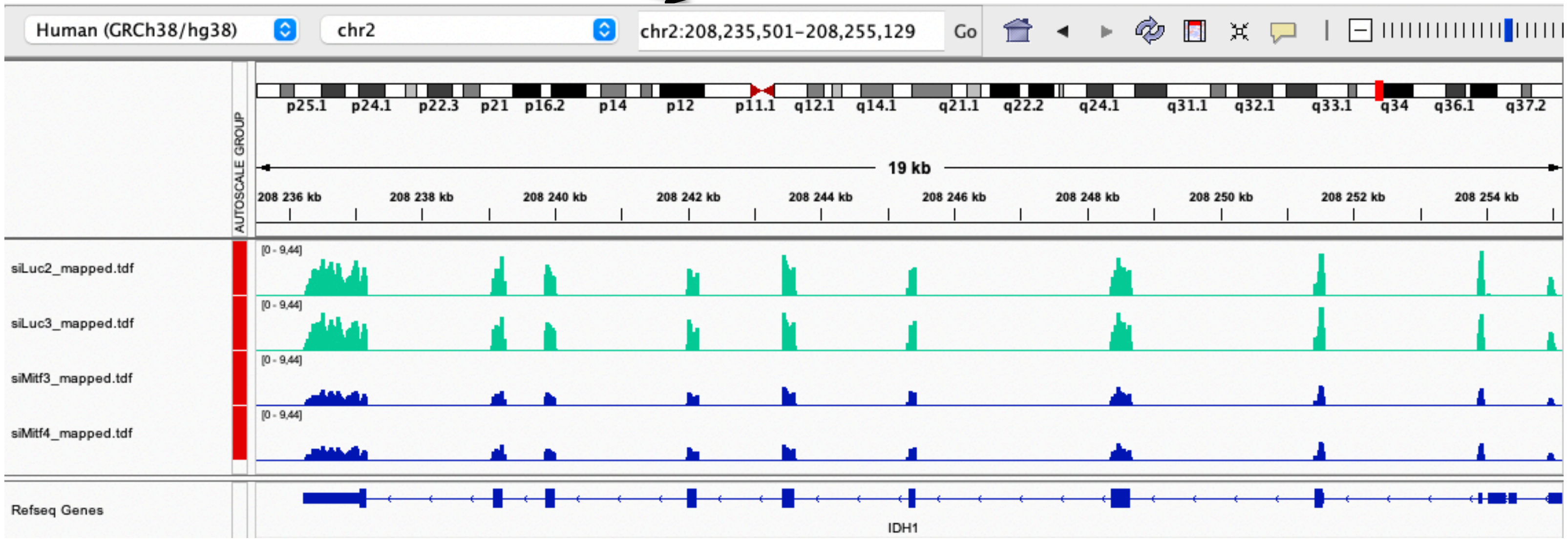


```
Uniquely mapped reads % |          85.26%
```

→ This proportion is consistent across samples

Exercise 2 – Question 2
Idh1 gene expression

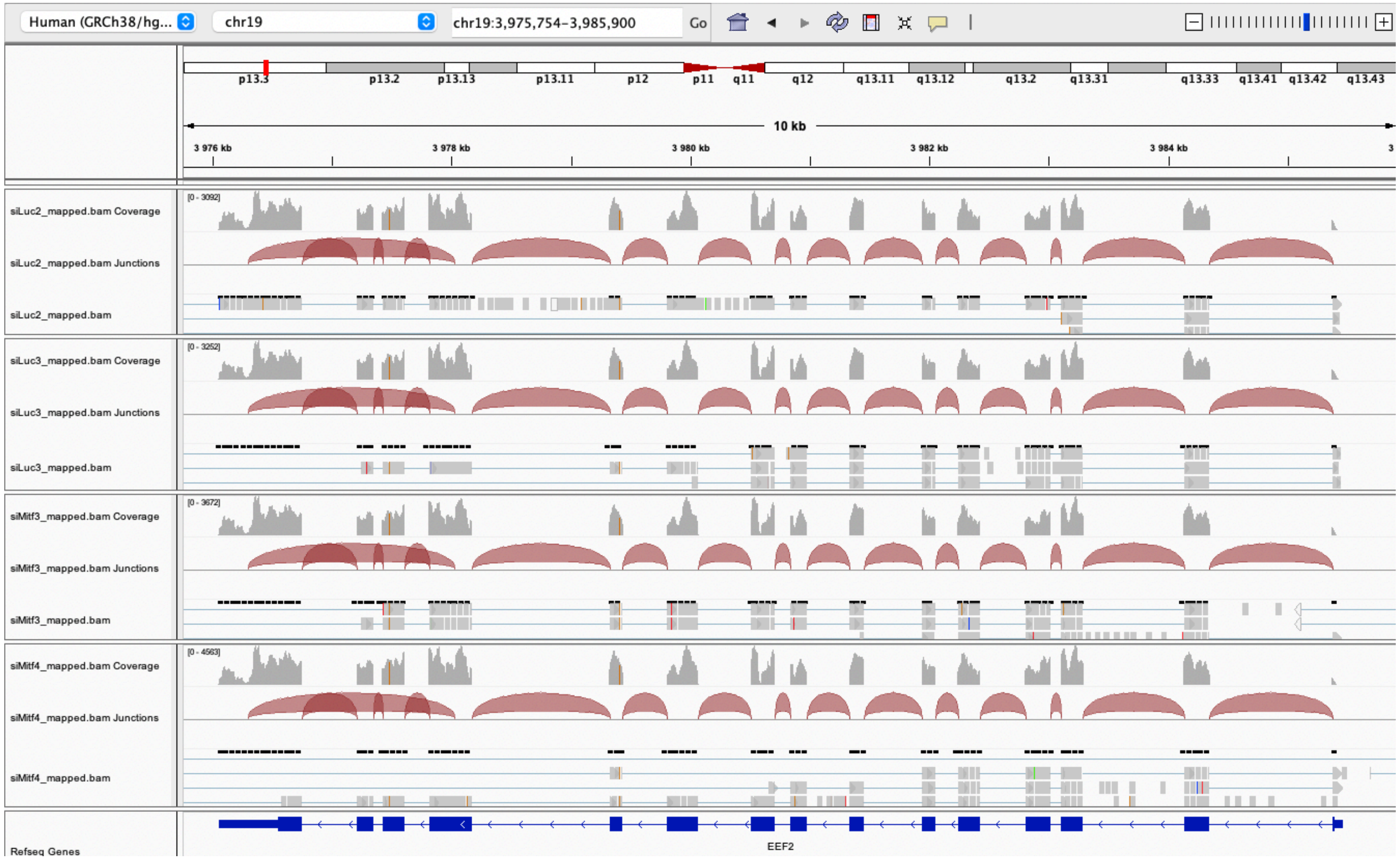
IGV : File → Load from file and select the 4 tdf files
Select all tdf tracks → Right-click → Group Autoscale :
→ IGV automatically adjusts the Y scale to the data range currently in view (this scaling continually adjusts as you move)
→ all tracks are on the same scale
Search for *Idh1*



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

Exercise 2 – Question 3

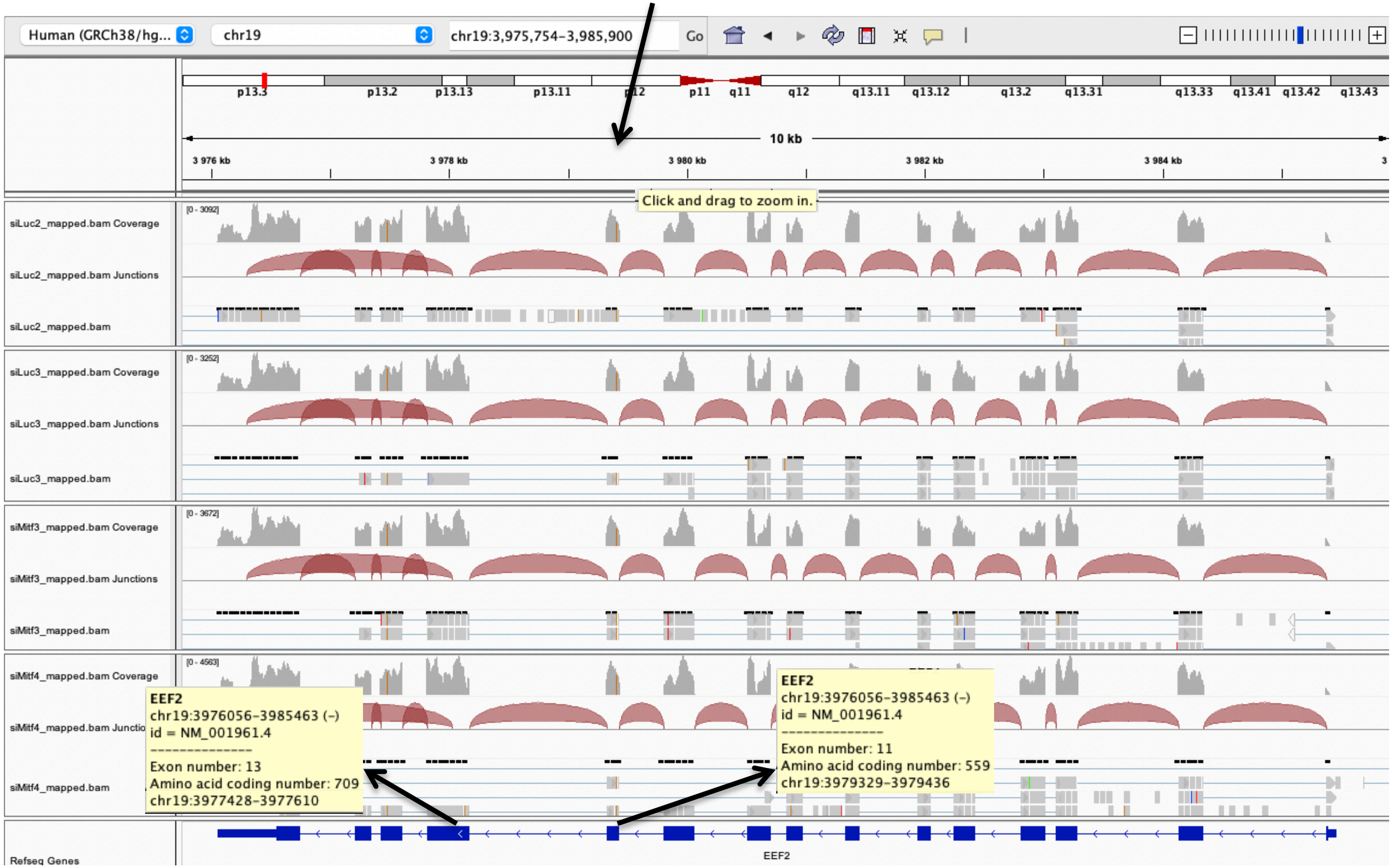
- File → new session
- File → load from files and load the 4 BAM files
- Search for *EEF2*



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_001961.4

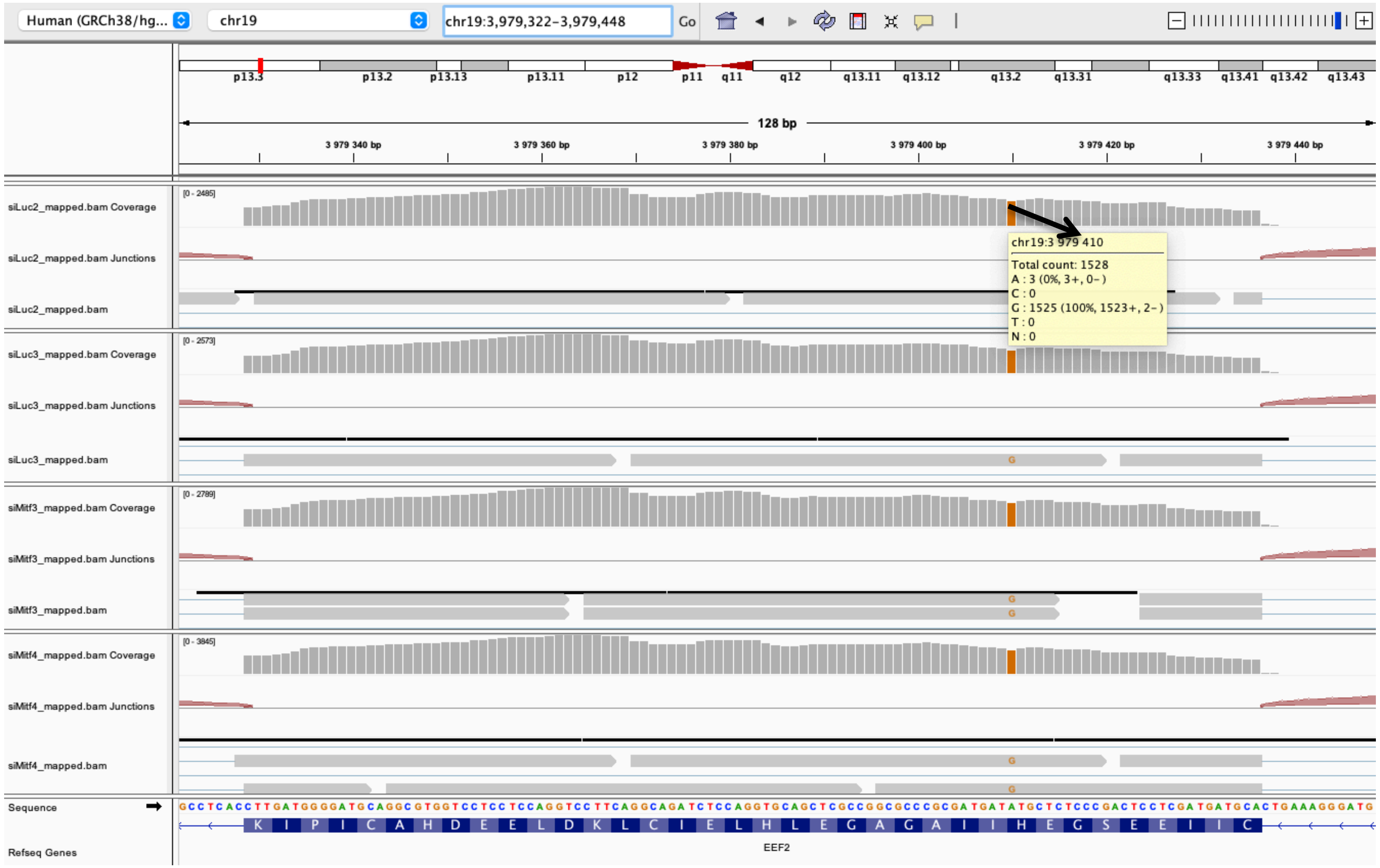
Exon number: 13
Amino acid coding number: 709
chr19:3977428-3977610

EEF2
chr19:3976056-3985463 (-)
id = NM_001961.4

Exon number: 11
Amino acid coding number: 559
chr19:3979329-3979436

Exercise 2 – Question 3

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



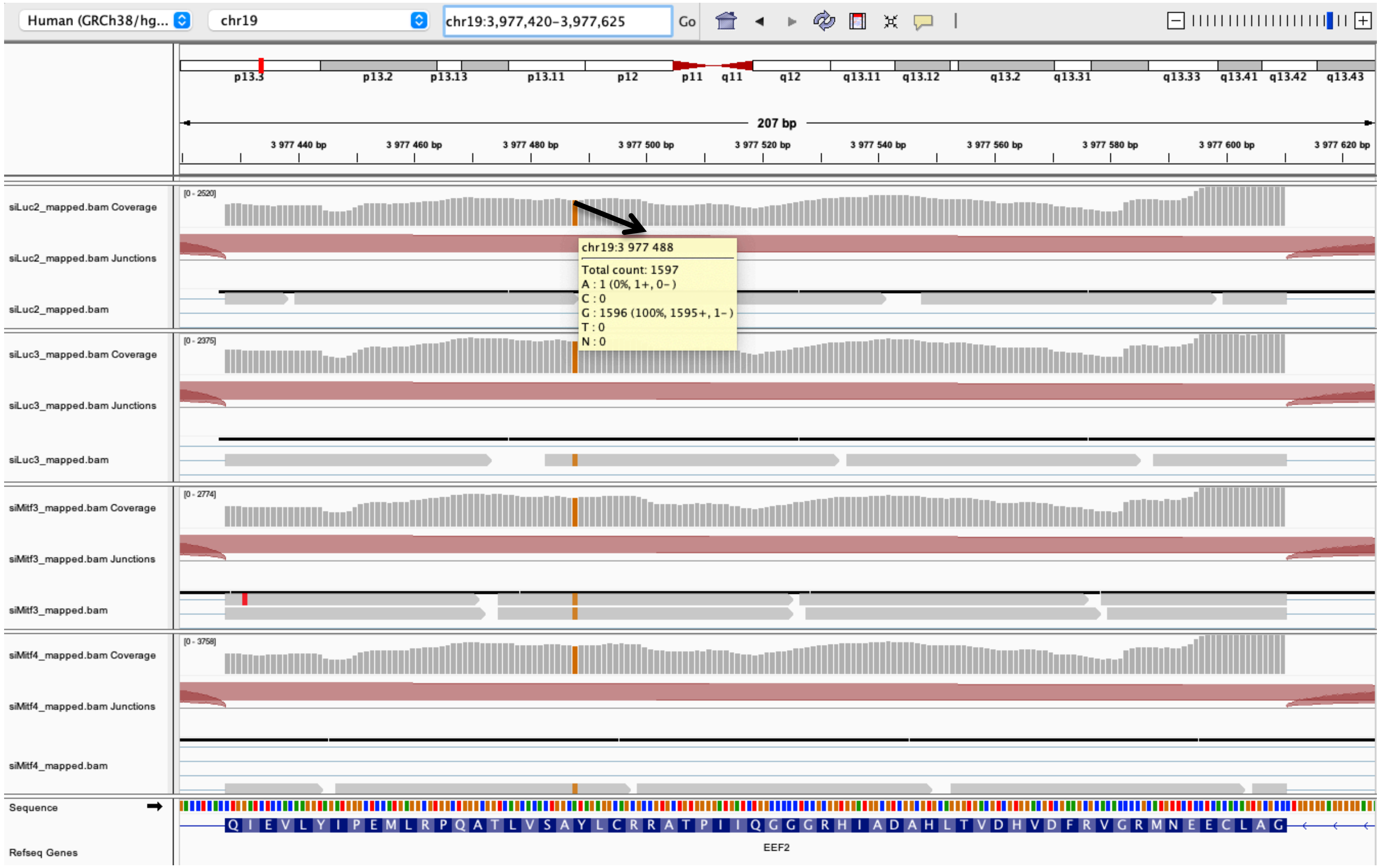
chr19:3 979 410

Total count: 1528
A : 3 (0%, 3+, 0–)
C : 0
G : 1525 (100%, 1523+, 2–)
T : 0
N : 0

Human (GRCh38/hg...
chr19
chr19:3,979,322-3,979,448
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
128 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_mapped.bam Coverage
[0 - 2485]
siLuc2_mapped.bam Junctions
siLuc2_mapped.bam
siLuc3_mapped.bam Coverage
[0 - 2573]
siLuc3_mapped.bam Junctions
siLuc3_mapped.bam
siMitf3_mapped.bam Coverage
[0 - 2789]
siMitf3_mapped.bam Junctions
siMitf3_mapped.bam
siMitf4_mapped.bam Coverage
[0 - 3845]
siMitf4_mapped.bam Junctions
siMitf4_mapped.bam
G
Sequence
GCCTCACCTTGATGGGGATGCAGGCGTGGTCCTCCTCCAGGTCCTTCAGGCAGATCTCCAGGTGCAGCTCGCCGGCGCCCGCGATGATATGCTCTCCCGACTCCTCGATGATGCACTGAAAGGGATG
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
Refseq Genes
EEF2

Exercise 2 – Question 3

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

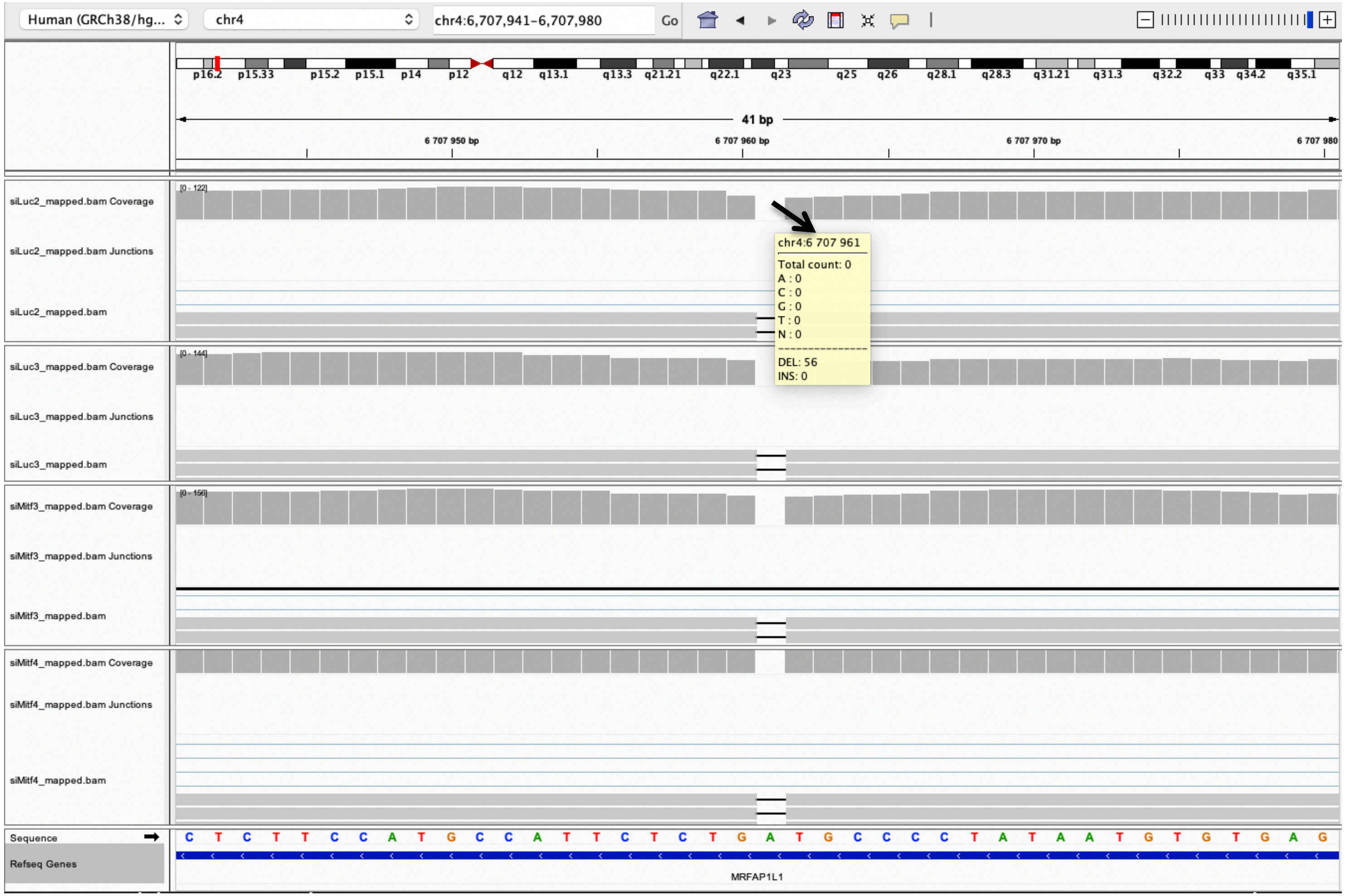


```
chr19:3 977 488

Total count: 1597
A : 1 (0%, 1+, 0- )
C : 0
G : 1596 (100%, 1595+, 1- )
T : 0
N : 0
```

Exercise 2 – Question 4

- Position chr4:6707961 :
  - Deletion vs reference genome



Human (GRCh38/hg...
chr4
chr4:6,707,941-6,707,980
Go
p16.2 p15.33 p15.2 p15.1 p14 p12 q12 q13.1 q13.3 q21.21 q22.1 q23 q25 q26 q28.1 q28.3 q31.21 q31.3 q32.2 q33 q34.2 q35.1
41 bp
6 707 950 bp
6 707 960 bp
6 707 970 bp
6 707 980
siLuc2_mapped.bam Coverage
[0 - 122]
siLuc2_mapped.bam Junctions
siLuc2_mapped.bam
siLuc3_mapped.bam Coverage
[0 - 144]
siLuc3_mapped.bam Junctions
siLuc3_mapped.bam
siMitf3_mapped.bam Coverage
[0 - 156]
siMitf3_mapped.bam Junctions
siMitf3_mapped.bam
siMitf4_mapped.bam Coverage
siMitf4_mapped.bam Junctions
siMitf4_mapped.bam
Sequence
Refseq Genes
C T C T T C C A T G C C A T T C T C T G A T G C C C C T A T A A T G T G T G A G
MRFAP1L1

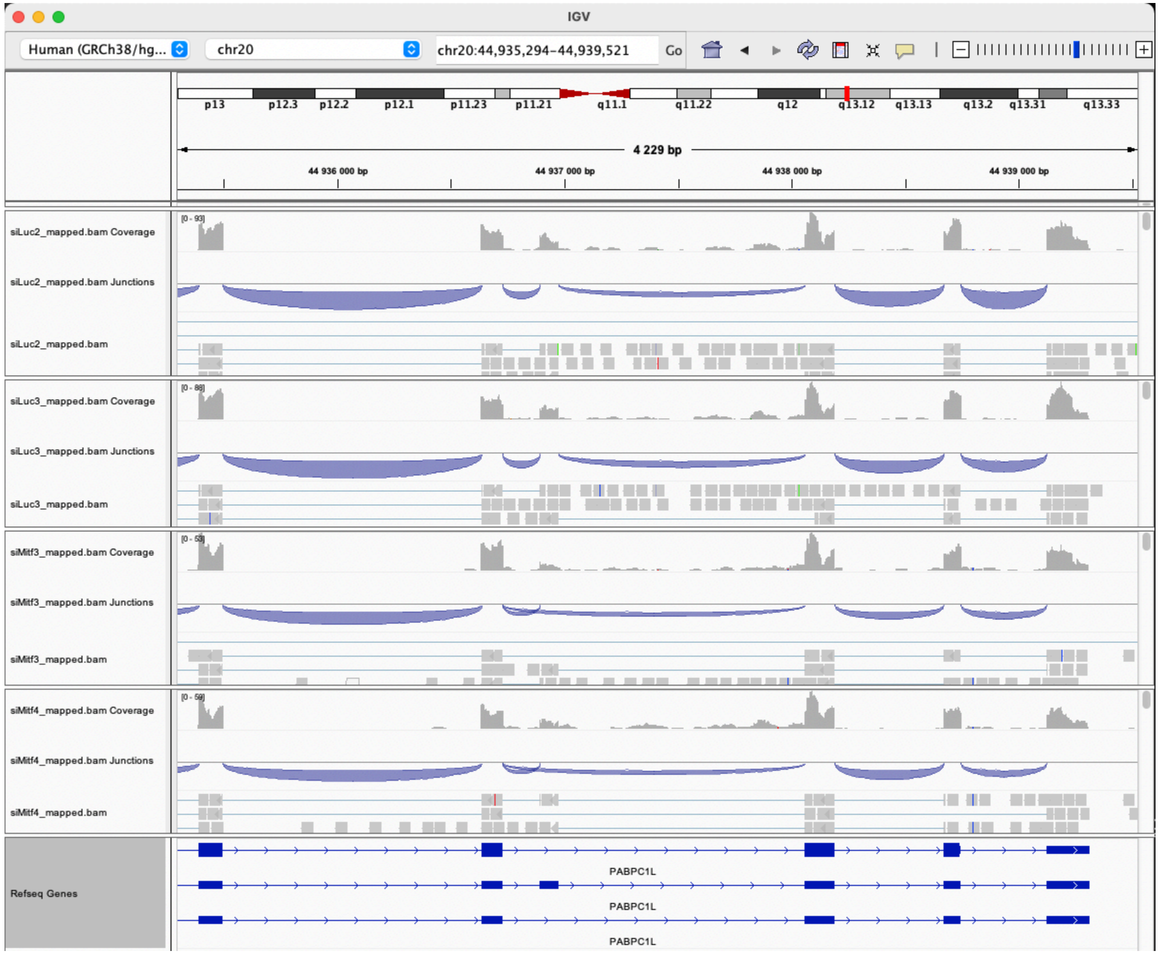
chr4:6 707 961

Total count: 0
A : 0
C : 0
G : 0
T : 0
N : 0

DEL: 56
INS: 0

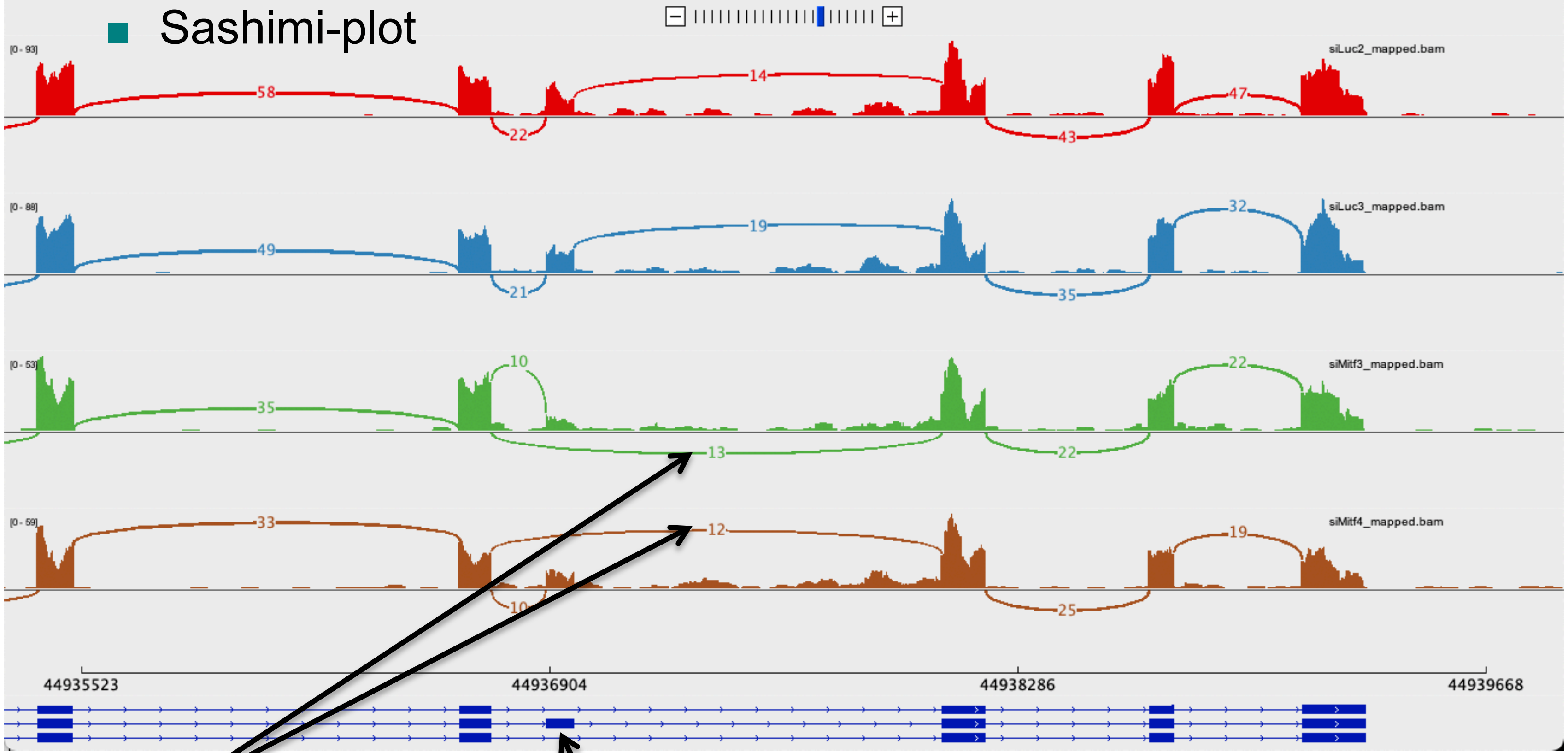
Exercise 2 – Question 5

- Region chr20:44,935,294-44,939,521 :
  - Right-click on Refseq Genes track → select Expanded to see all annotated isoforms



Exercise 2 – Question 5

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

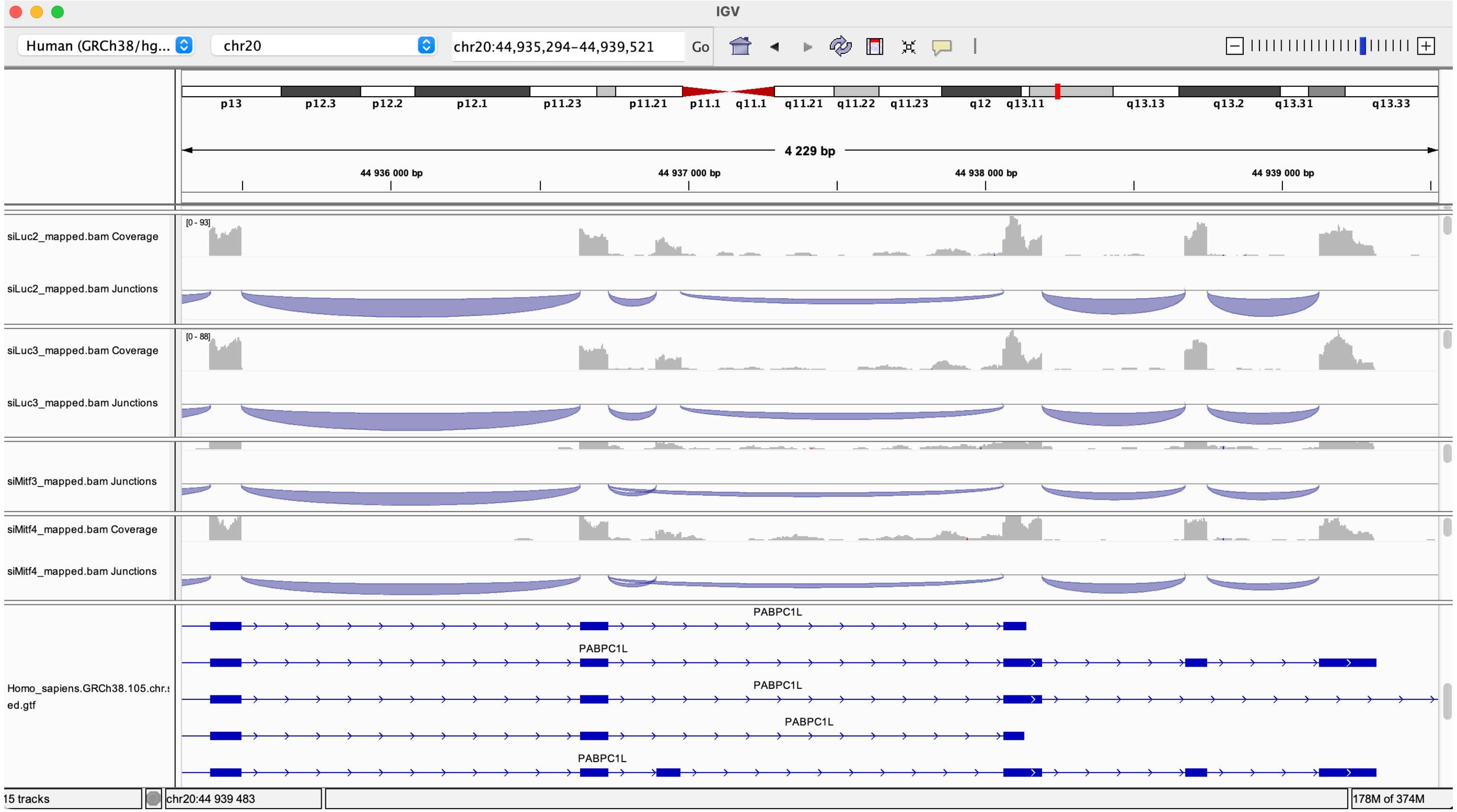


We detect an isoform without this exon in siMitf samples

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

Exercise 2 – Question 5

- If you would like to display Ensembl annotations, you can add this track
  - File → Load from file
  - Select Homo_sapiens.GRCh38.105.chr.sorted.gtf available in RNAseq/annotations folder

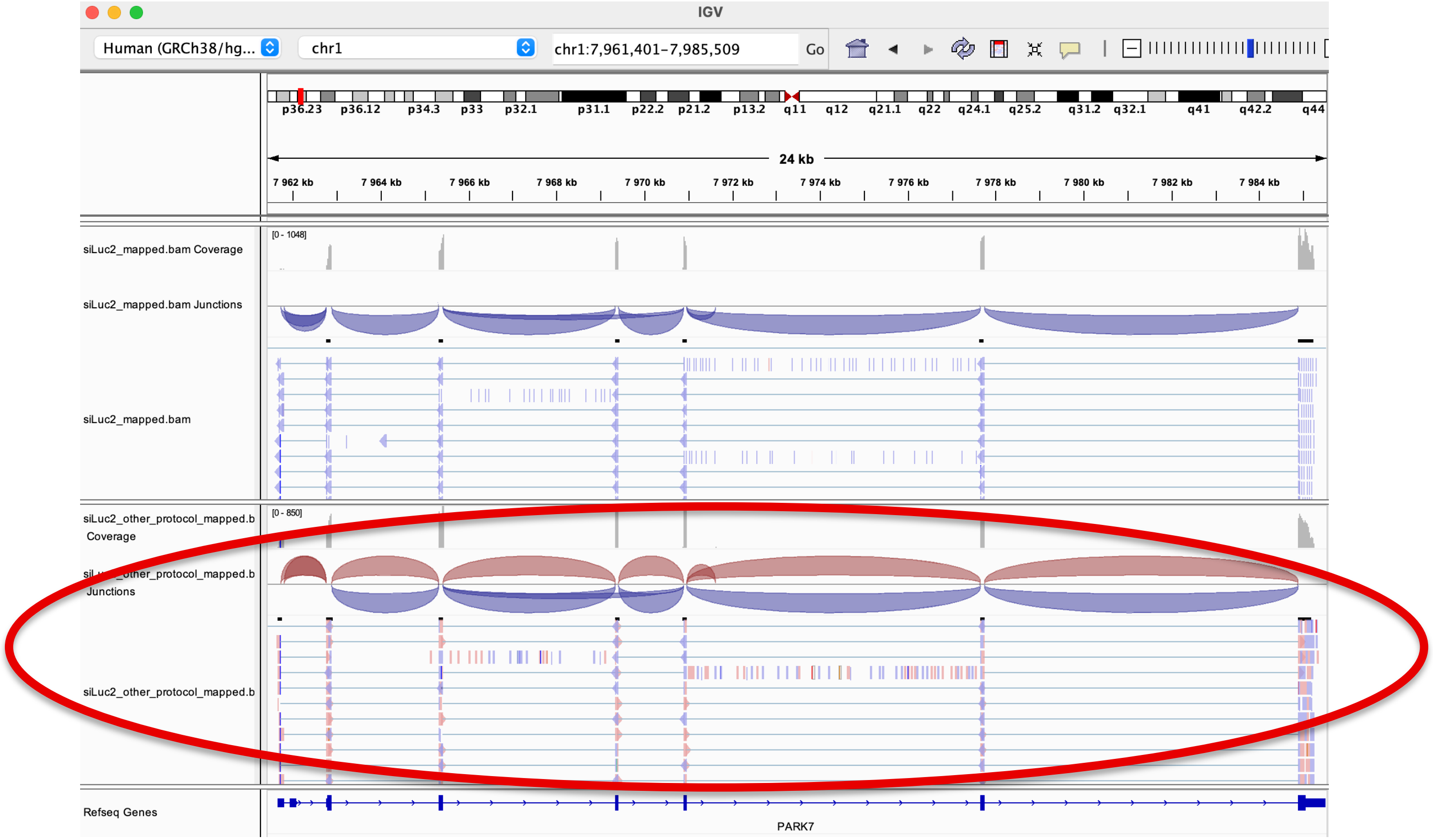


Exercise 2 – Question 5

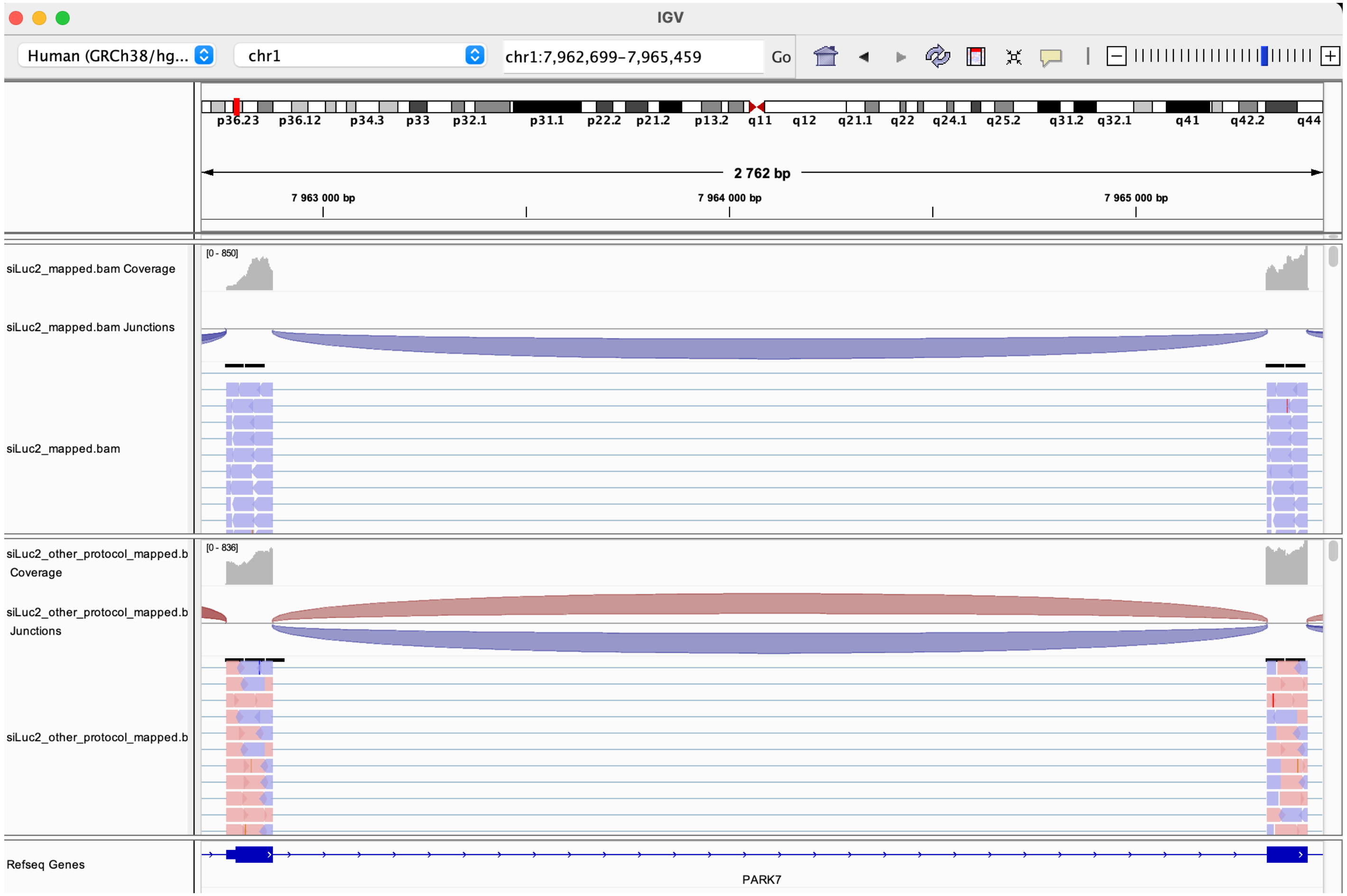
- You can save your IGV session
  - To save the current state of your IGV session to a named session file
  - File → Save Session
  - Data files must stay at the same location
- Use File → Open session to restore a saved session

Exercise 2 – Question 6

- 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


IGV
Human (GRCh38/hg...
chr1
chr1:7,961,401-7,985,509
Go
p36.23 p36.12 p34.3 p33 p32.1 p31.1 p22.2 p21.2 p13.2 q11 q12 q21.1 q22 q24.1 q25.2 q31.2 q32.1 q41 q42.2 q44
24 kb
7 962 kb 7 964 kb 7 966 kb 7 968 kb 7 970 kb 7 972 kb 7 974 kb 7 976 kb 7 978 kb 7 980 kb 7 982 kb 7 984 kb
siLuc2_mapped.bam Coverage
[0 - 1048]
siLuc2_mapped.bam Junctions
siLuc2_mapped.bam
siLuc2_other_protocol_mapped.b Coverage
[0 - 850]
siLuc2_other_protocol_mapped.b Junctions
siLuc2_other_protocol_mapped.b
Refseq Genes
PARK7

Exercise 2 – Question 6



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