



# 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%

History

search datasets

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

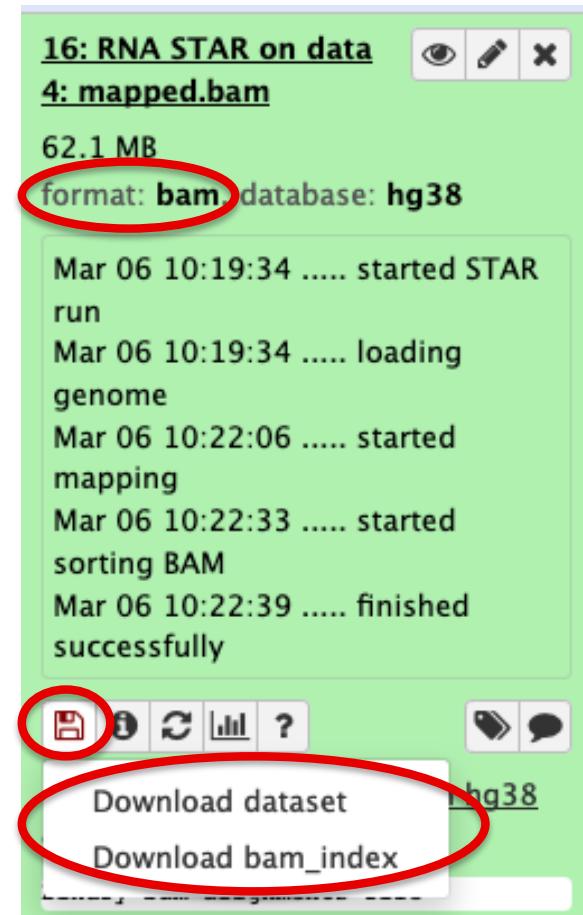
Mapping speed, Million of reads per

# 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)



The screenshot shows a Galaxy workflow step titled "16: RNA STAR on data" with a sub-step "4: mapped.bam". The output is a 62.1 MB file in BAM format, using the hg38 database. The execution log shows the STAR process starting at 10:19:34, loading the genome, starting mapping at 10:22:06, starting sorting BAM at 10:22:33, and finishing successfully at 10:22:39. At the bottom, a red circle highlights the save icon, and another red circle highlights the "Download dataset" and "Download bam\_index" options in the dropdown menu.

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

Download dataset hg38

Download bam\_index

### ■ IGV

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

# Exercise 1

## 2. Splice junction



→ 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

8,058 bp

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

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 : 43M7155N7M**

Intron length :  
7984893 - 7977738 = 7155

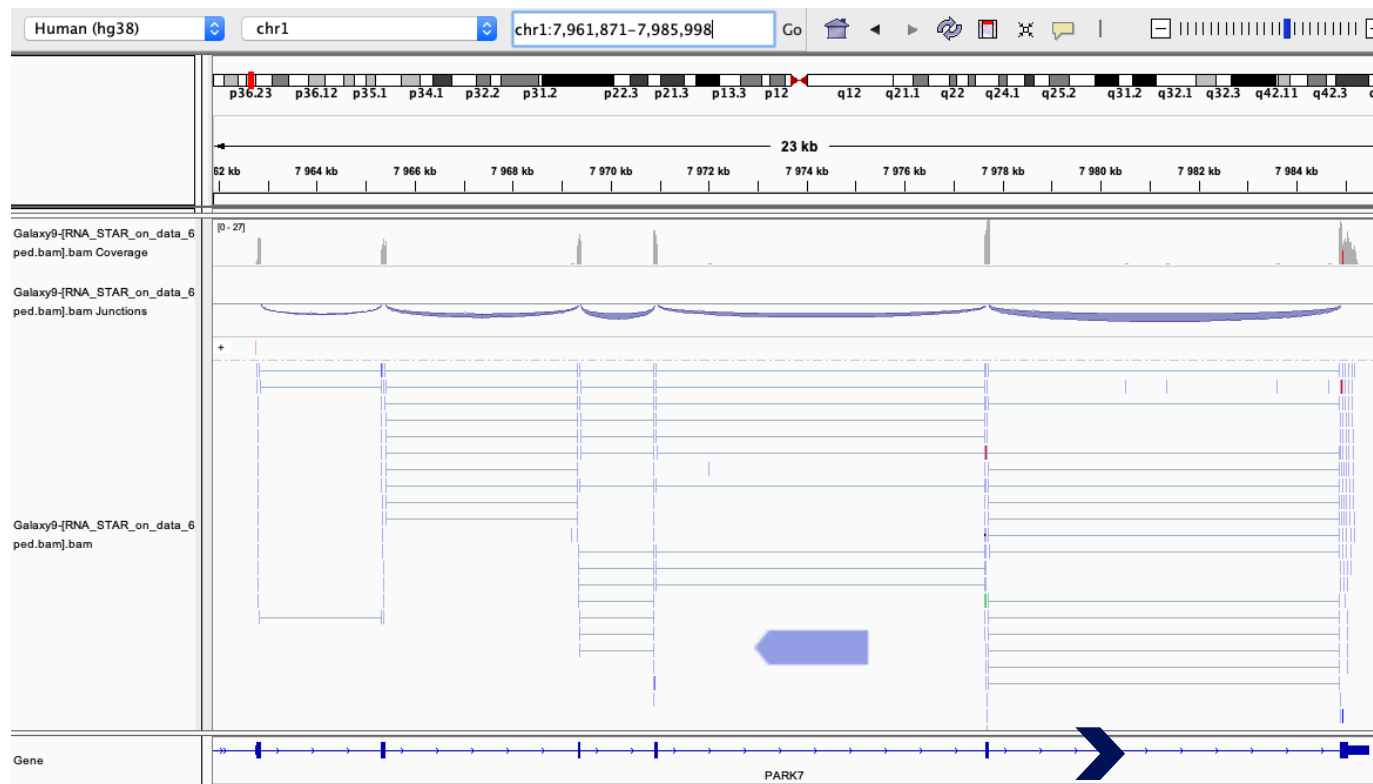
5 tracks loaded chr1:7,977,718 655M of 1,105M

# Exercise 1

## 2. Strand specificity

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

*Park7 :*

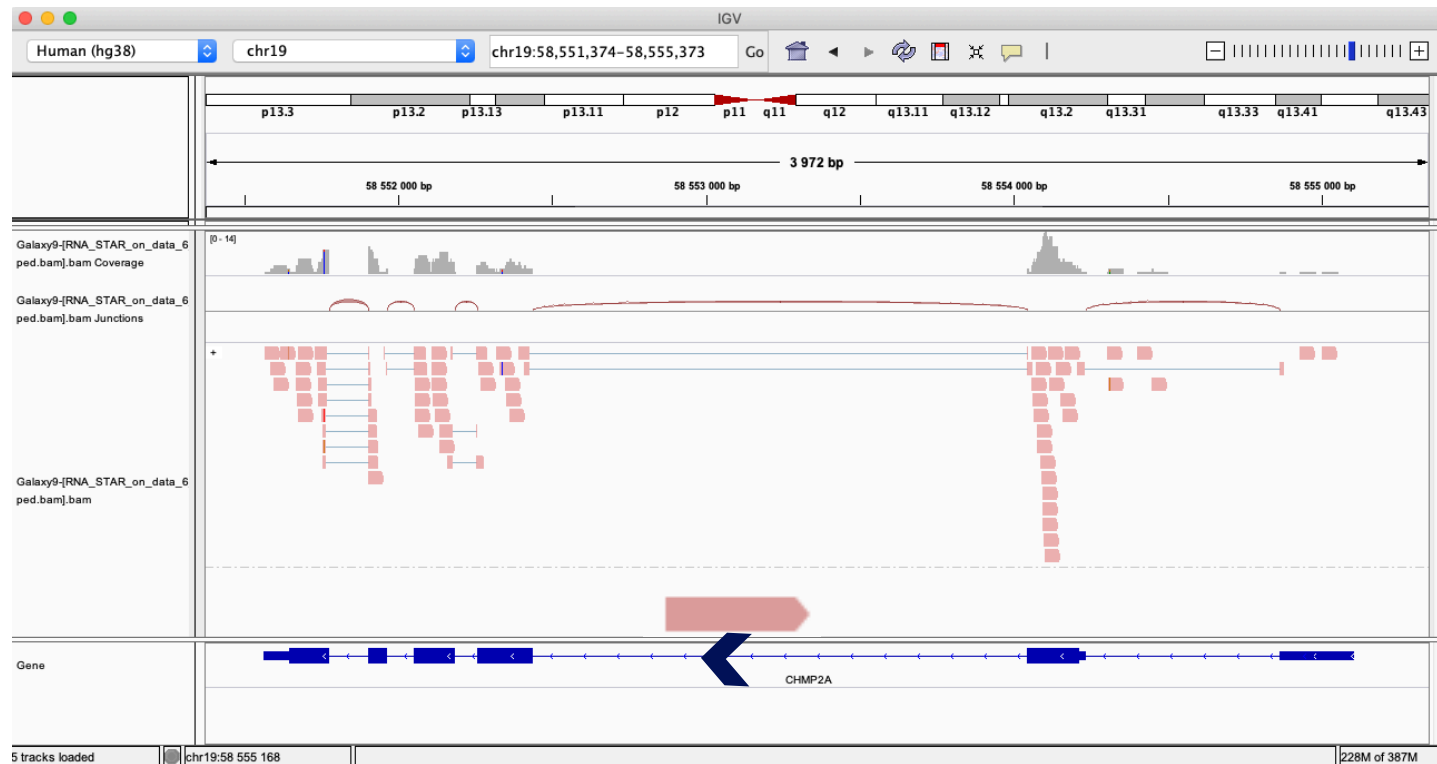


The library has been prepared with a directional mRNAseq protocol which retains strand information :  
reads are in the opposite direction as the transcribed strand

# Exercise 1

## 2. Strand specificity

*Chmp2a* :

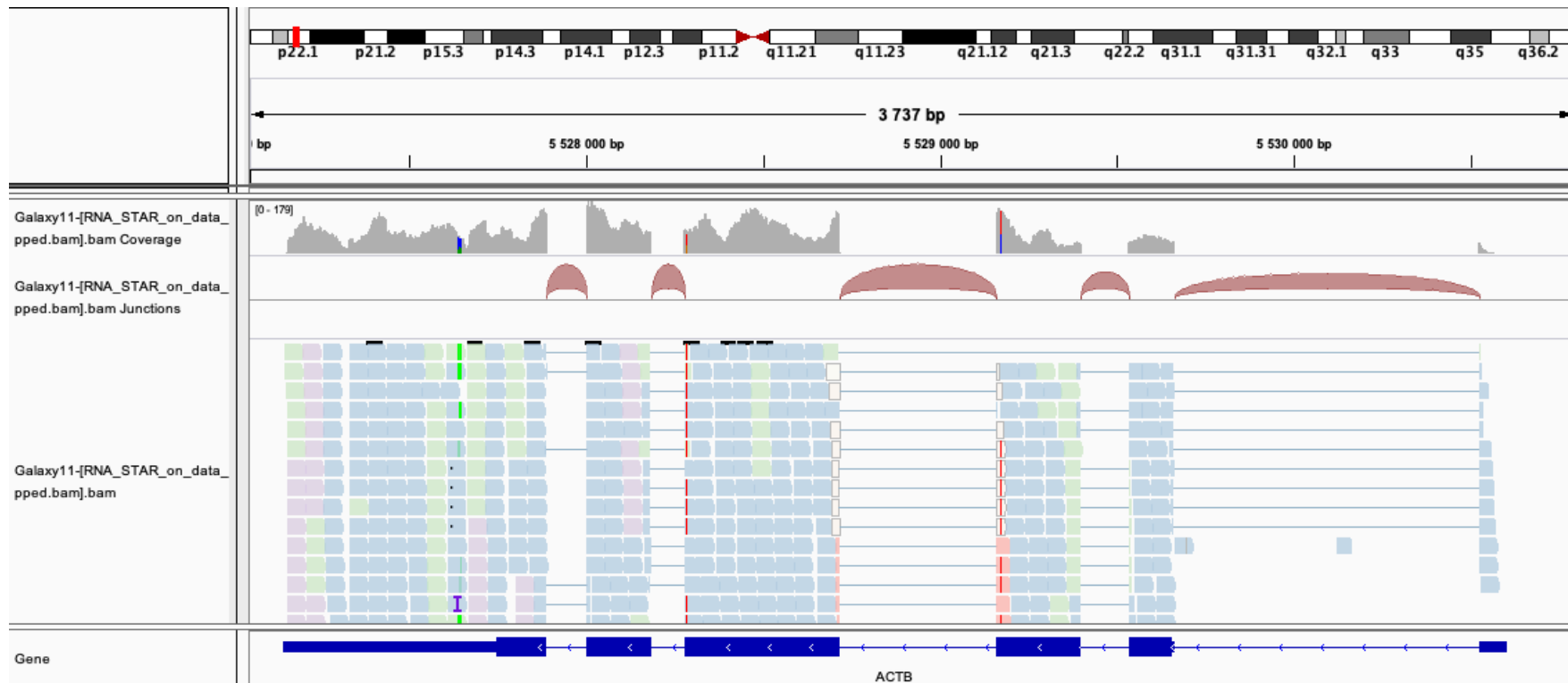


The library has been prepared with a directional mRNAseq protocol which retains strand information :  
reads are in the opposite direction as the transcribed strand

# Exercise 1

## 2. Multiple mapped reads

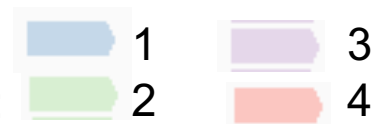
Right click on BAM file → Color alignments by → tag → NH



Number of reported alignments

→ see NH tag in pop-up windows to visualize

color-coding (that can be different from this one) :



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 8722563
Uniquely mapped reads %    85.30%
Average mapped length      47105
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%
  
```

<b>STAR on siLuc2:</b>	Uniquely mapped reads %	85.30%
<b>STAR on siLuc3:</b>	Uniquely mapped reads %	85.72%
<b>STAR on siMitf3:</b>	Uniquely mapped reads %	85.41%
<b>STAR on siMitf4:</b>	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

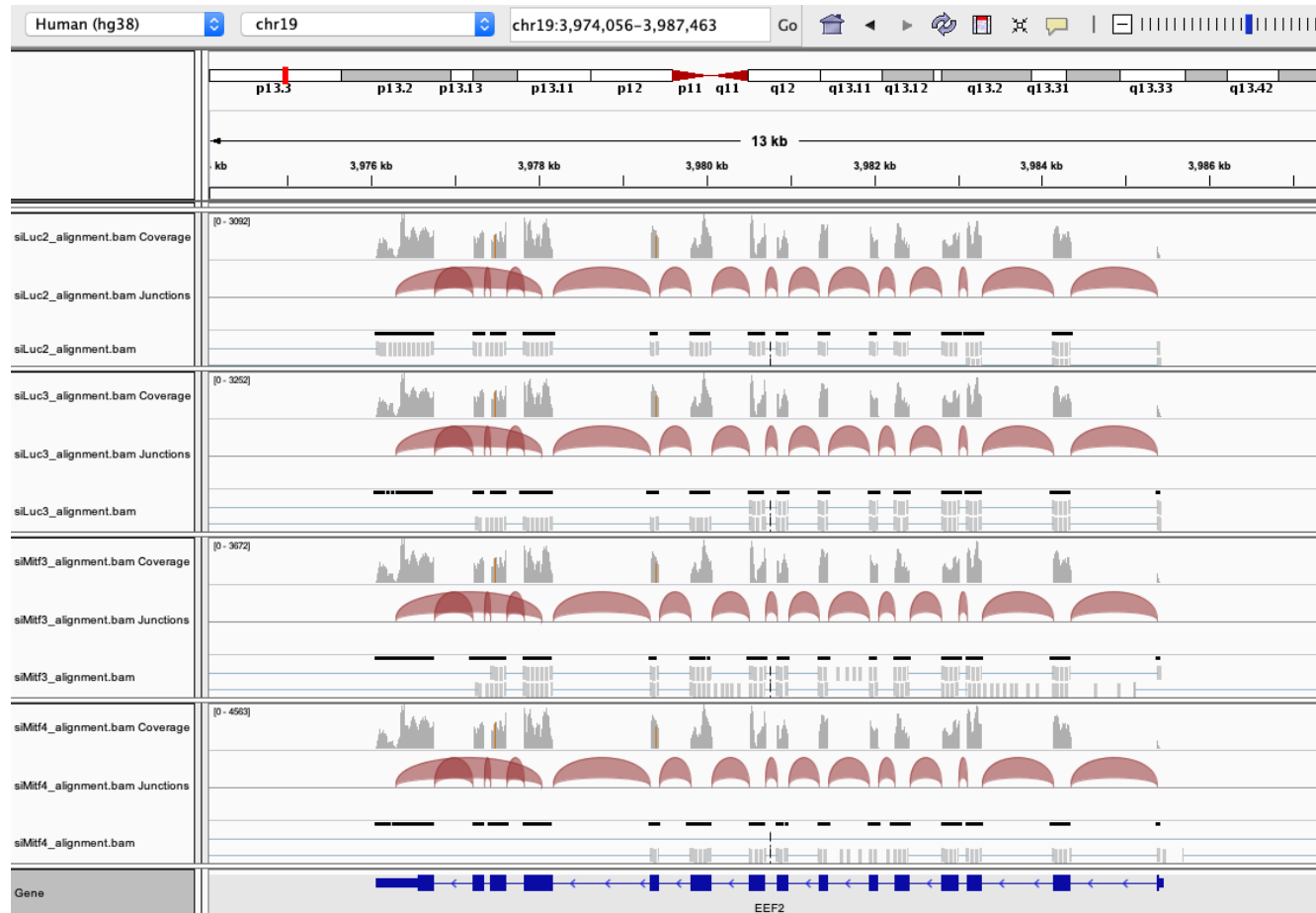
Search for *Idh1*



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

# Exercise 2 – Question 3

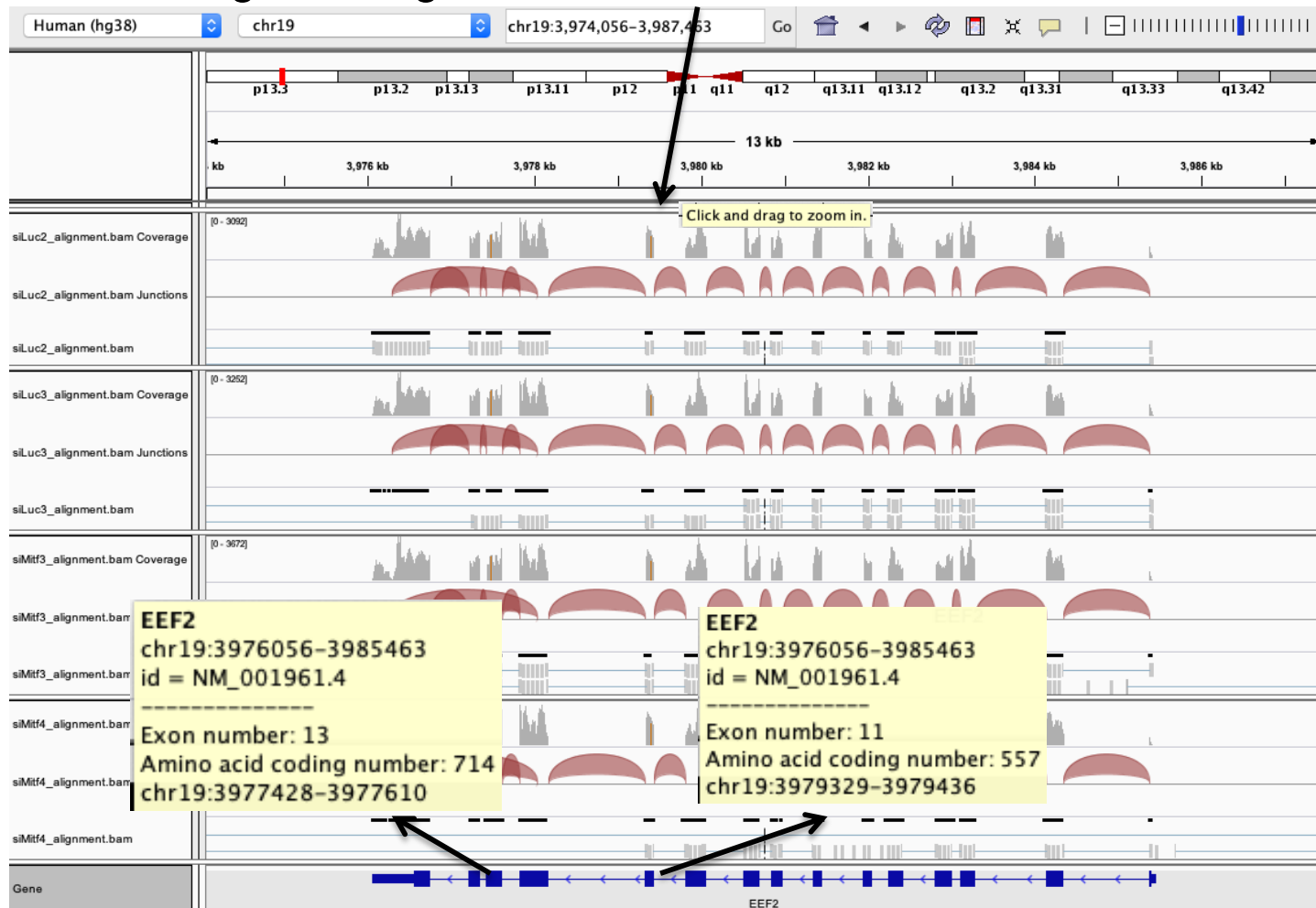
- File → new session
- File → load from files and load the 4 BAM files
- Search for *EEF2*



# Exercise 2 – Question 3

Exon numbers are provided on annotation track

Click and drag on a region to zoom in



# Exercise 2 – Question 3

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




# Exercise 2 – Question 3

## ■ *Eef2* exon 13

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



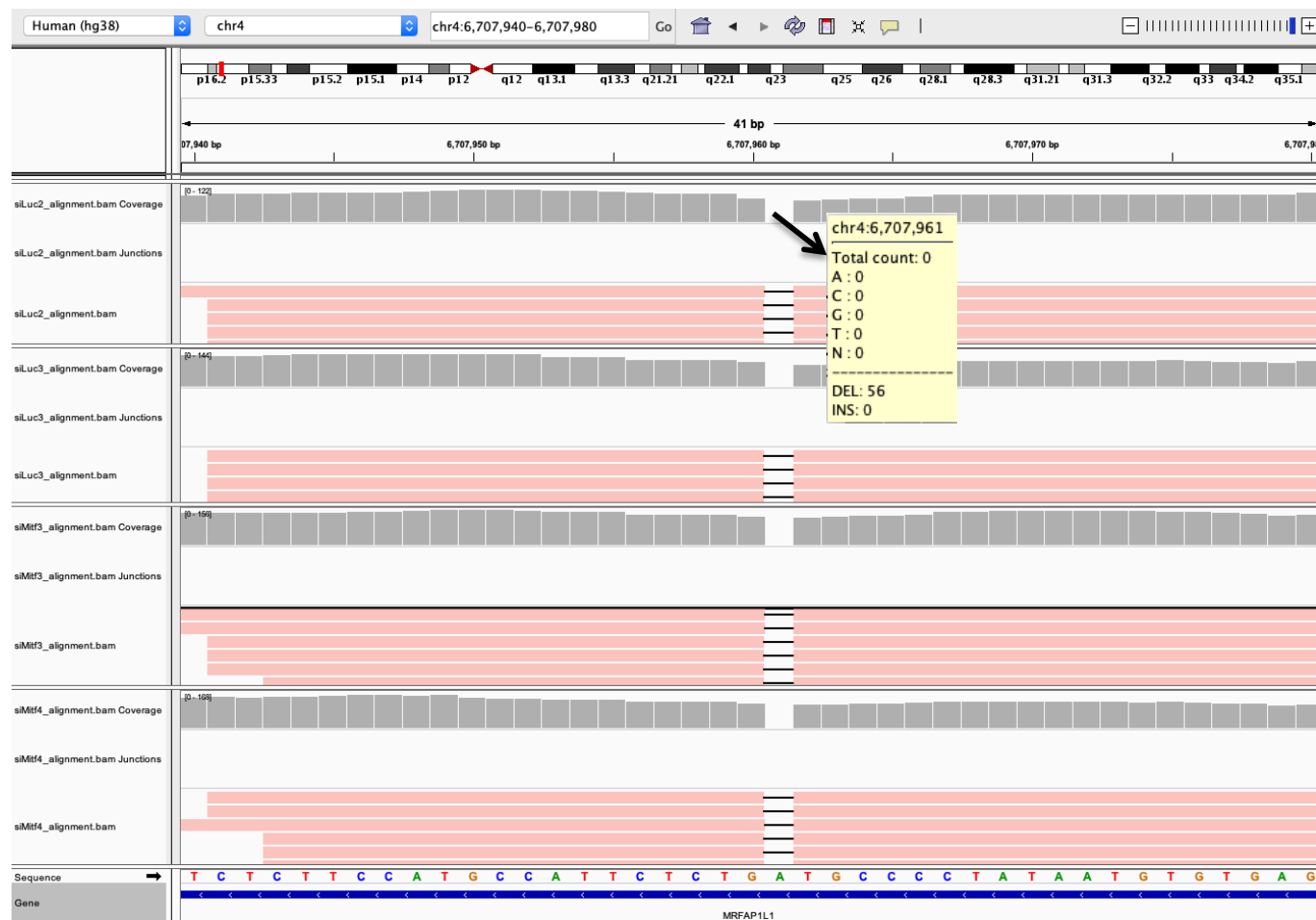
# Exercise 2 – Question 3

- 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



# Exercise 2 – Question 4

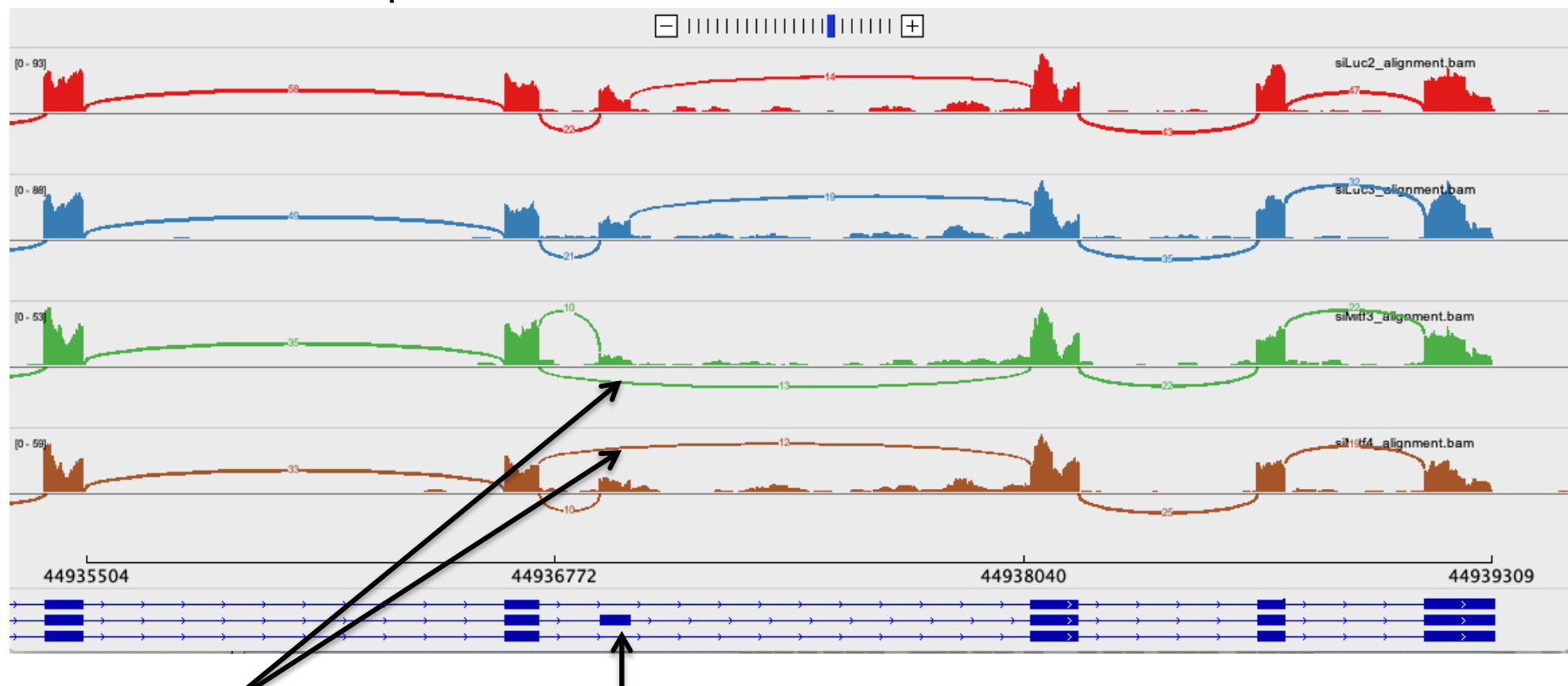
- Position chr4:6707960-6707961 :
  - Deletion vs reference genome





# Exercise 2 – Question 5

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



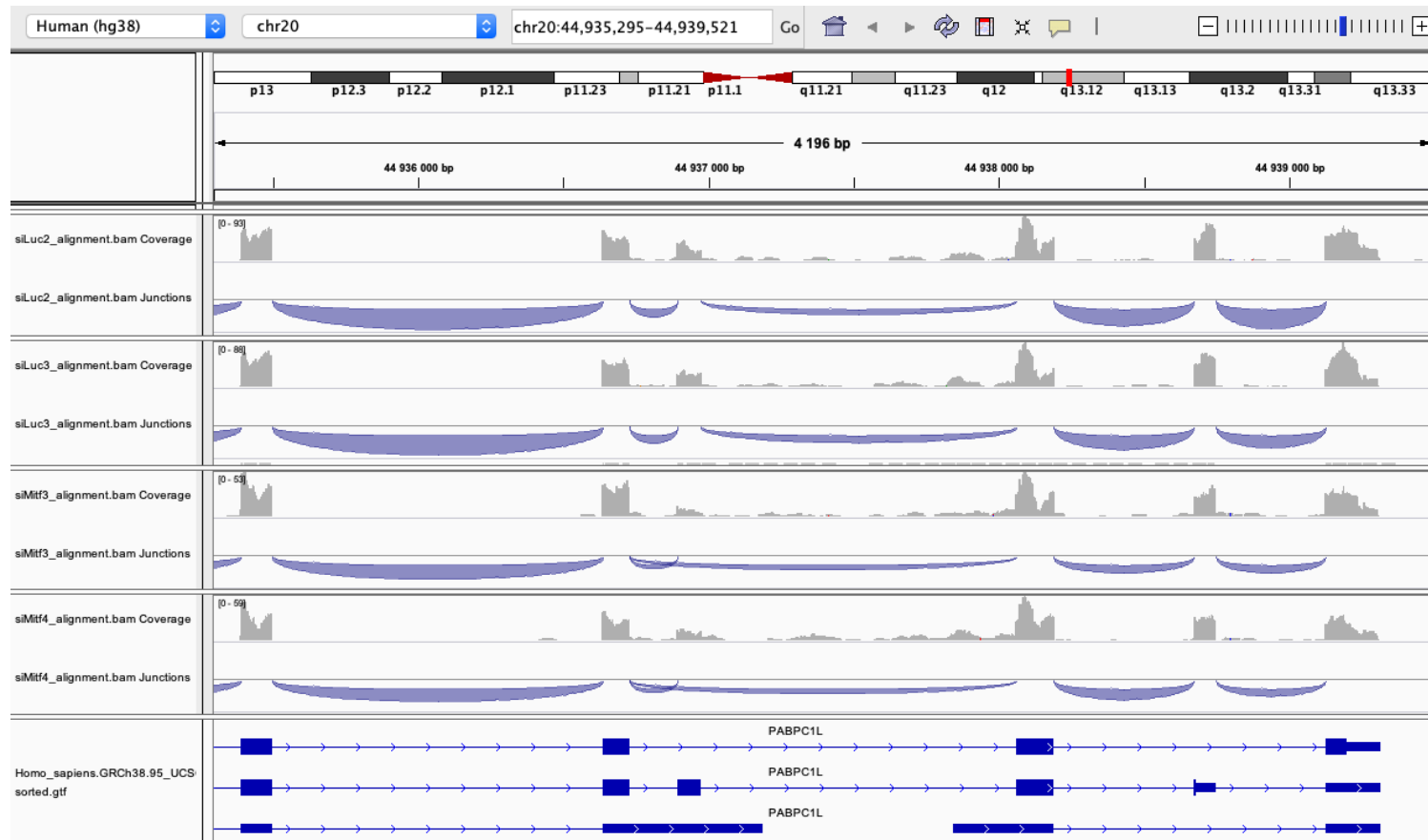
We detect an isoform without this exon in siMitf samples

**IGV is only a visualization tool**

**In-depth analysis using paired-end data with more coverage is needed**

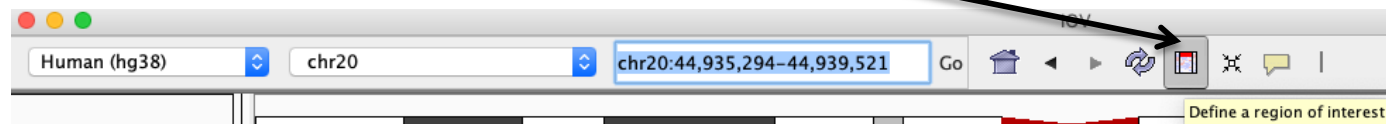
# Exercise 2 – Question 5

- If you would like to display Ensembl annotations, you can add this track
  - File → Load from file
  - Select Homo\_sapiens.GRCh38.95\_UCSC\_chr.sorted.gtf available in RNAseq/annotations folder

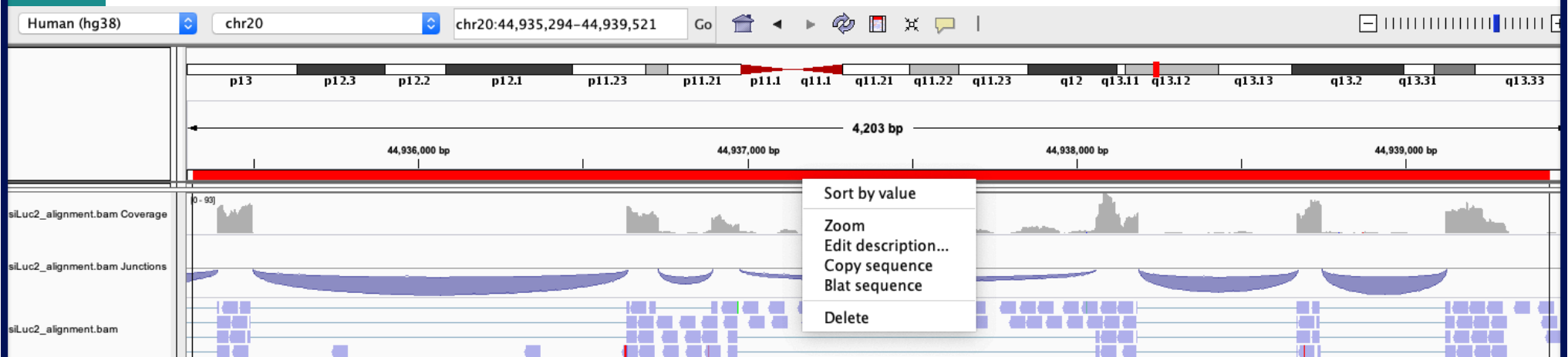


# Exercise 2 – Question 5

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



- 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



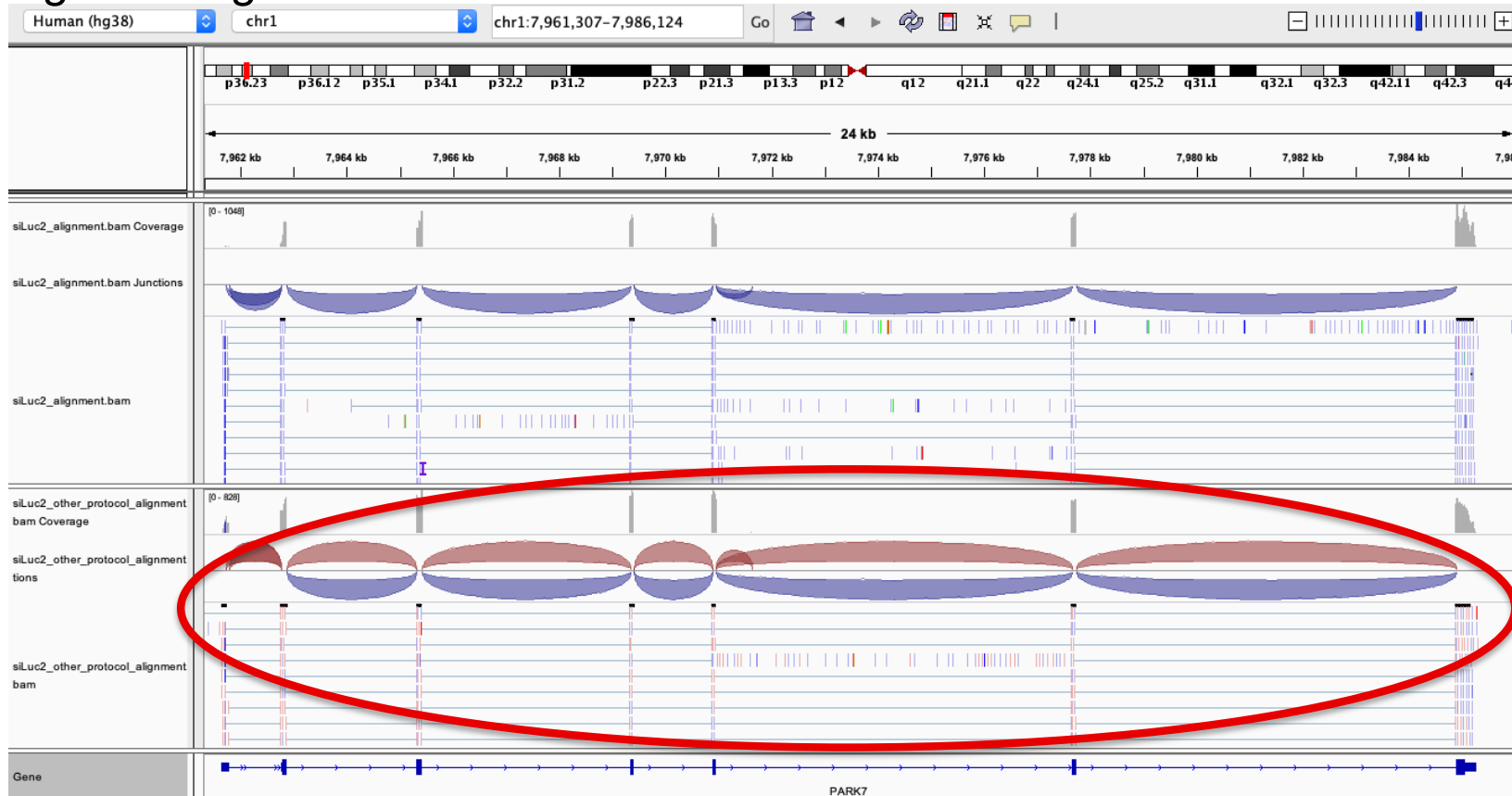
# Exercise 2 – Question 5

---

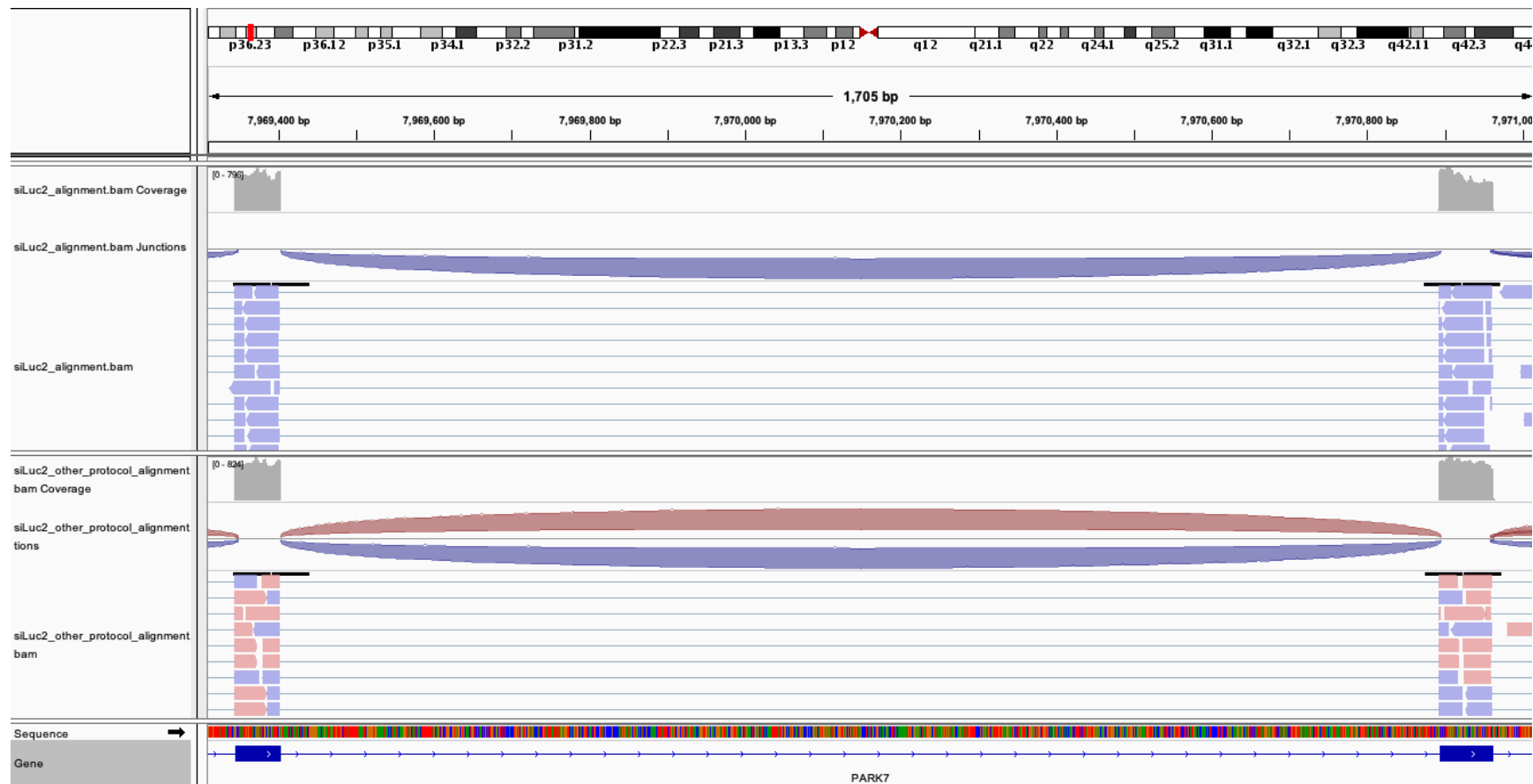
- You can save your IGV session
  - To save the current state of your IGV session to a named session file
  - File → Save Session
  - Data files must stay at the same location
- Use File → Open session to restore a saved session

# 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



# Exercise 2 – Question 6



→ 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)

# Exercise 3 – Question 1

```
Total Reads          43088618
Total Tags*           49986051
Total Assigned Tags°  47196194
```

```
=====
Group                Total_bases      Tag_count      Tags/Kb
CDS_Exons            96264295       35939434      373.34
5'UTR_Exons          6950804        419760        60.39
3'UTR_Exons          31717723       7309834       230.47
Introns              1512979984     3120459       2.06
TSS_up_1kb           29147844       43305         1.49
TSS_up_5kb           130226667      86698         0.67
TSS_up_10kb          232911303     120174        0.52
TES_down_1kb         31104985       146931        4.72
TES_down_5kb         134802199      239250        1.77
TES_down_10kb        236957571     286533        1.21
=====
```

\* reads spliced once are counted as 2 tags, reads spliced twice are counted as 3 tags, ...

° number of tags that can be assigned to the 10 above groups

Tags assigned to “TSS\_up\_1kb” are also assigned to “TSS\_up\_5kb” and “TSS\_up\_10kb”

Tags assigned to “TSS\_up\_5kb” are also assigned to “TSS\_up\_10kb”

# Exercise 3 – Question 2

---

- RSeQC infer experiment

- on siLuc2 library prepared with a directional protocol :

```
This is SingleEnd Data
```

```
Fraction of reads explained by "++,--": 0.0086
```

```
Fraction of reads explained by "+-,-+": 0.9914
```

```
Fraction of reads explained by other combinations: 0.0000
```

- On siLuc2 library prepared with a non directional protocol :

```
This is SingleEnd Data
```

```
Fraction of reads explained by "++,--": 0.4988
```

```
Fraction of reads explained by "+-,-+": 0.5012
```

```
Fraction of reads explained by other combinations: 0.0000
```