



NGS read mapping : answers to questions

Céline Keime
keime@igbmc.fr

Exercise 1

1. Log file

Proportion of uniquely mapped reads :

```
Started job on | Apr 05 10:43:32
Started mapping on | Apr 05 10:49:10
Finished on | Apr 05 10:49:25
Mapping speed, Million of reads per hour | 240.00

Number of input reads | 1000000
Average input read length | 50
UNIQUE READS:
Uniquely mapped reads number | 852134
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 + ↺ ▾

Rechercher des données ▾ ×

RNA-seq data analysis ✎

3.57 GB 📍 8 ↻

☑️ ⚙️

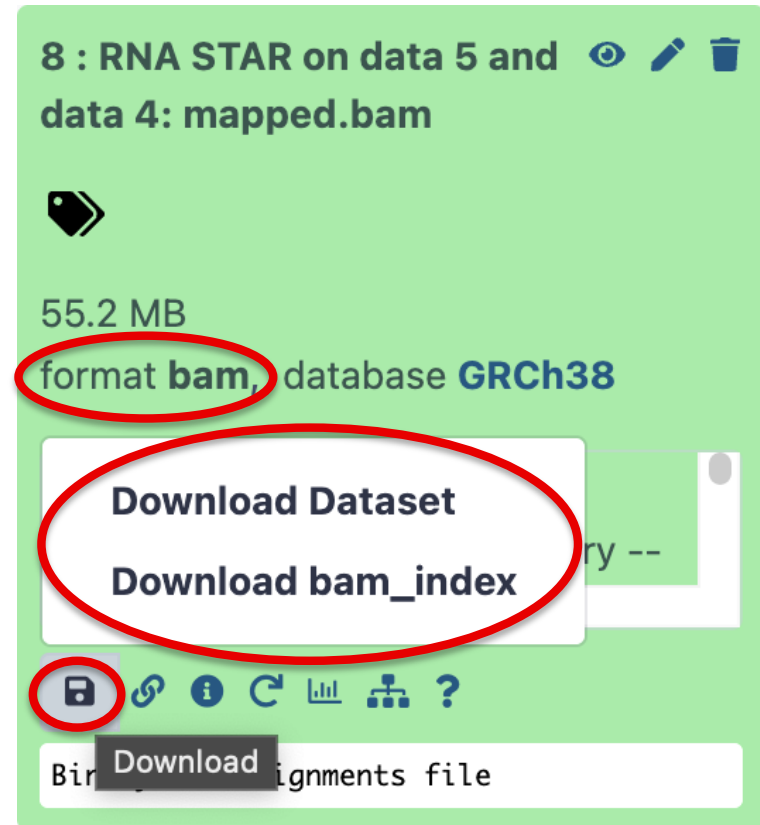
- 8 : RNA STAR on data 5 and data 4: mapped.bam 📄 ✎ 🗑️
- 7 : RNA STAR on data 5 and data 4: splice junctions.bed 📄 ✎ 🗑️
- 6 : RNA STAR on data 5 and data 4: log 📄 ✎ 🗑️
- 5 : Homo_sapiens.GRCh38.105.chr.gtf.gz 📄 ✎ 🗑️
- 4 : siLuc2_1000000.fastq.gz 📄 ✎ 🗑️
- 3 : FastQC on data 1: RawData 📄 ✎ 🗑️
- 2 : FastQC on data 1: Webp 📄 ✎ 🗑️

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

55.2 MB

format **bam**, database GRCh38

Download Dataset

Download bam_index

Download

Alignments file

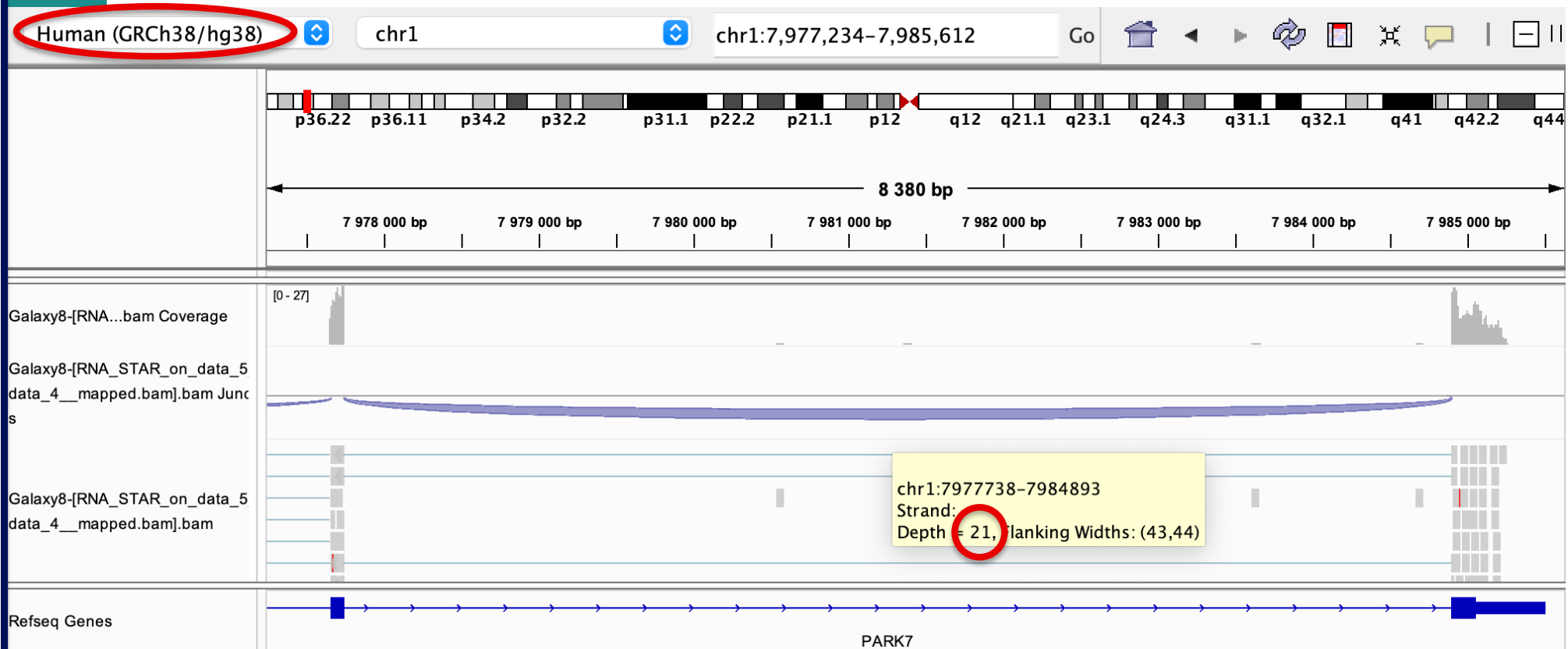
The screenshot shows a dataset card in Galaxy. The title is "8 : RNA STAR on data 5 and data 4: mapped.bam". Below the title, there is a tag icon, the size "55.2 MB", and the format "format bam, database GRCh38". A red circle highlights the word "bam". Below this, a white box contains two buttons: "Download Dataset" and "Download bam_index", both of which are circled in red. At the bottom, there is a toolbar with icons for download, share, info, refresh, and help. A red circle highlights the download icon. Below the toolbar, a text box contains the word "Download" and a dropdown menu showing "Alignments file".

■ IGV

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

Exercise 1

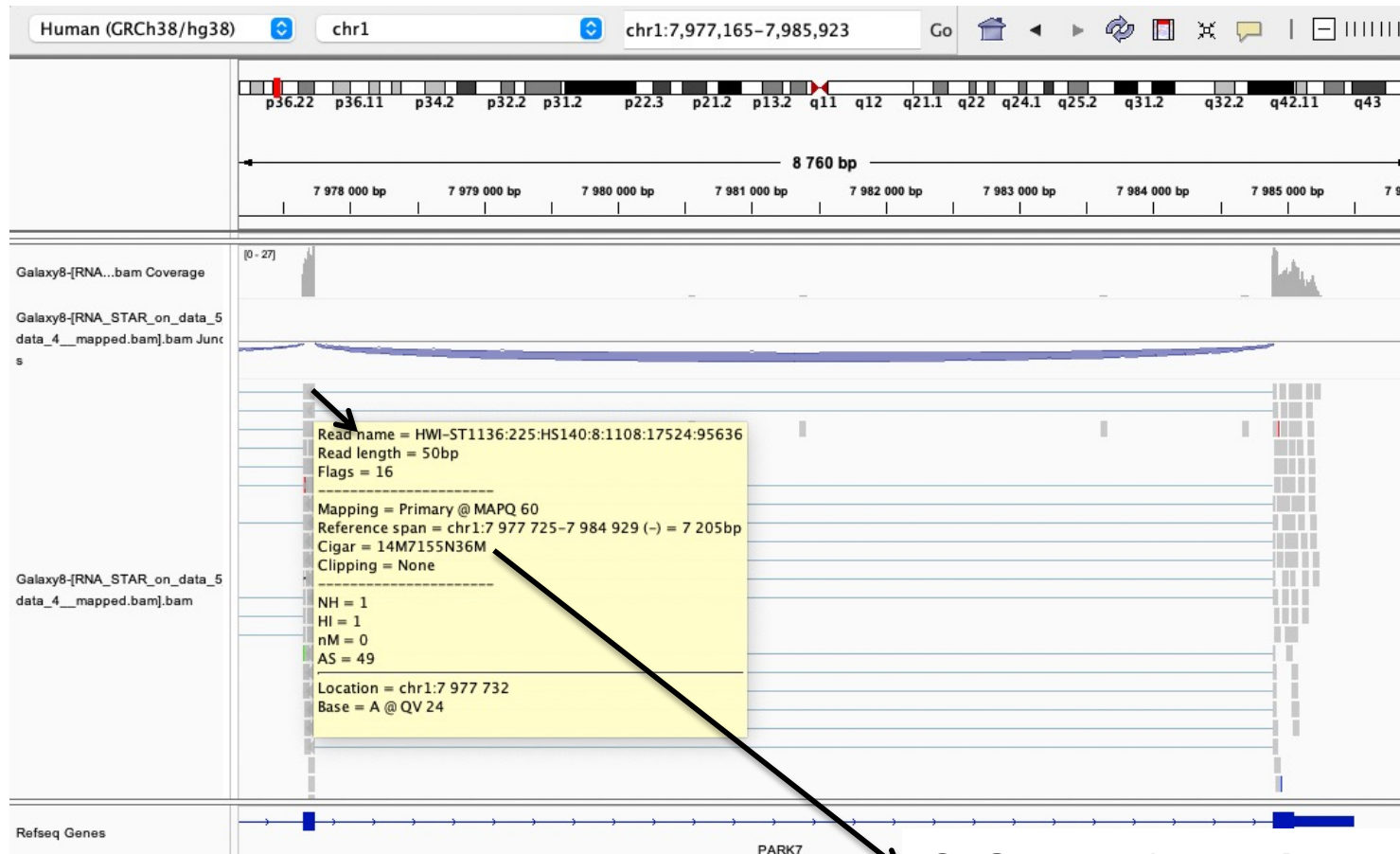
2. Splice junction



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

Exercise 1

2. Splice junction



CIGAR : 14M7155N36M

Intron length :

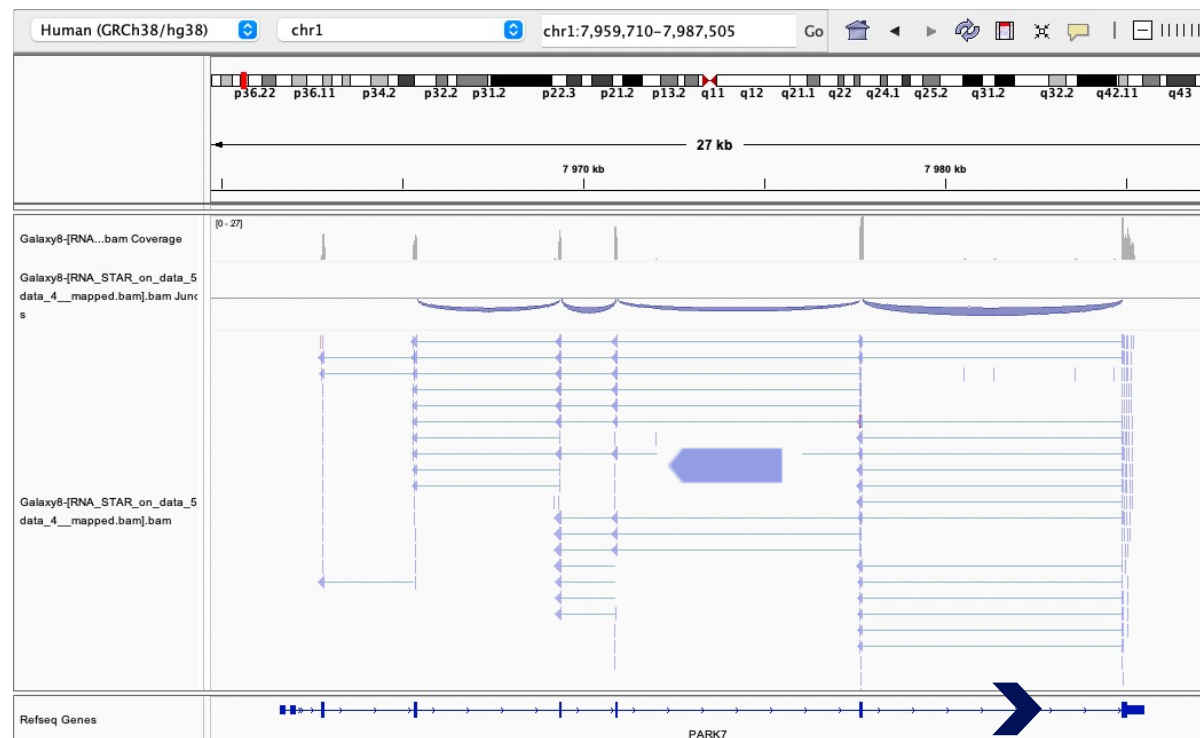
7984893 - 7977738 = 7155

Exercise 1

2. Strand specificity

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

Park7 :

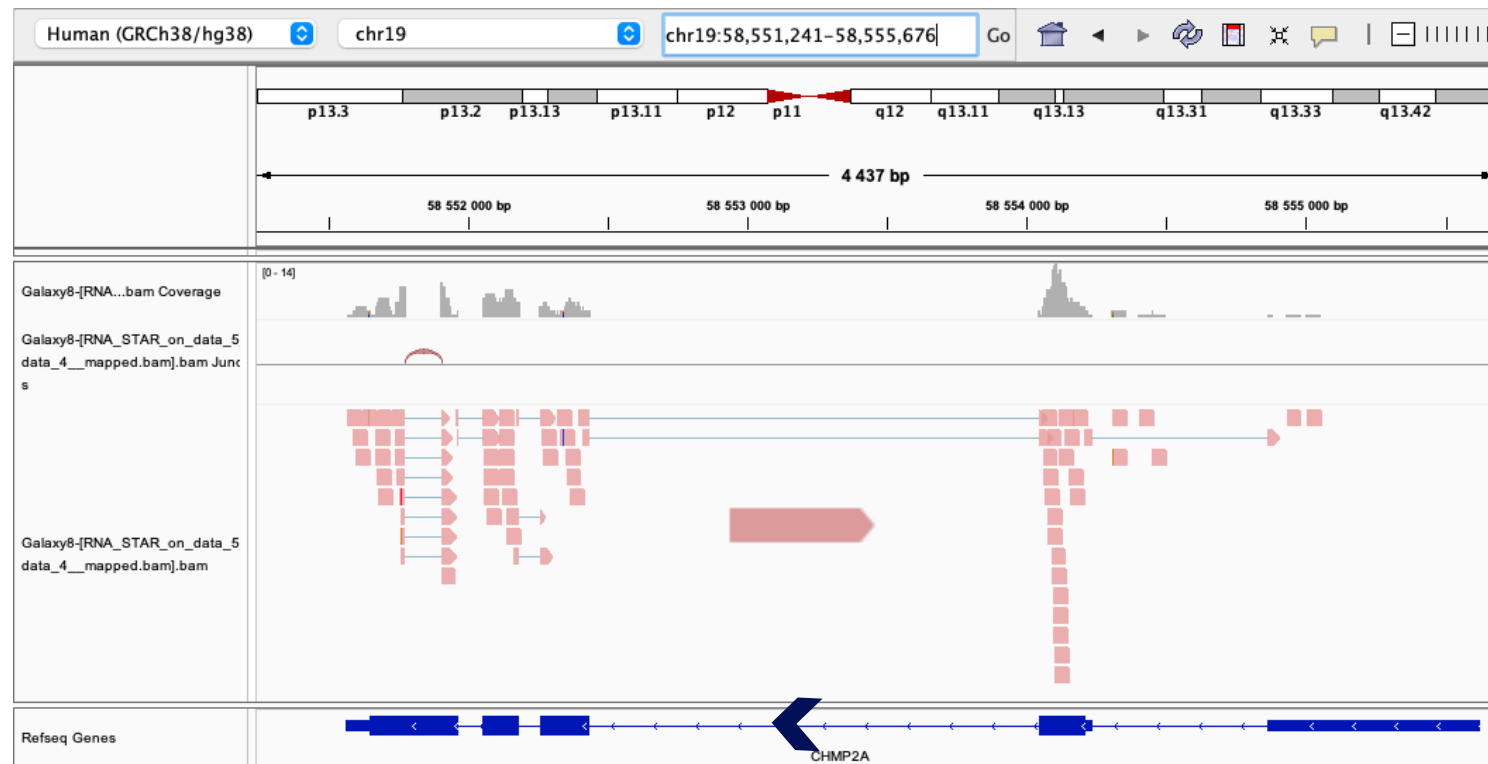


The library has been prepared with a directional mRNAseq protocol which retains strand information :
reads are in the opposite direction compared to the transcribed strand

Exercise 1

2. Strand specificity

Chmp2a :

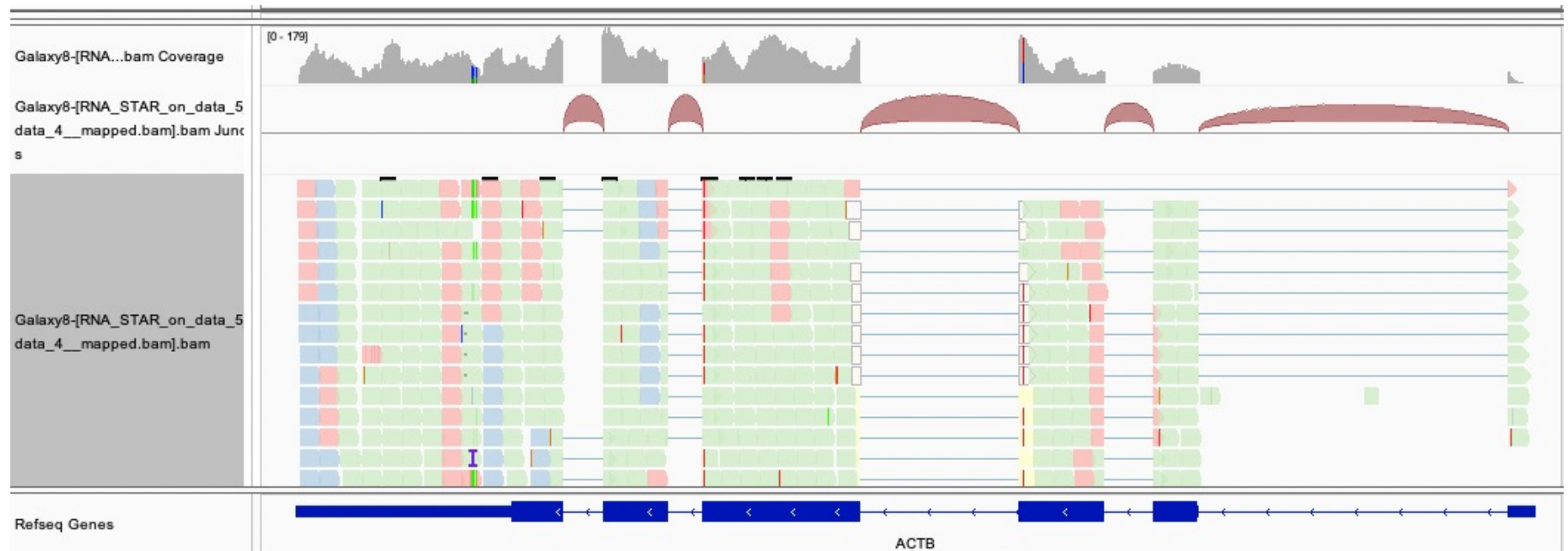


The library has been prepared with a directional mRNAseq protocol which retains strand information :
reads are in the opposite direction compared to the transcribed strand

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) : 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_oth er_protocol: mapped.bam	  
15: RNA STAR on siMitf4: ma pped.bam	  
14: RNA STAR on siMitf4: spli ce junctions.bed	  
13: RNA STAR on siMitf4: log	  
12: RNA STAR on siMitf3: ma pped.bam	  
11: RNA STAR on siMitf3: spli ce junctions.bed	  
10: RNA STAR on siMitf3: log	  
9: RNA STAR on siLuc3: map ped.bam	  
8: RNA STAR on siLuc3: spli ce junctions.bed	  
7: RNA STAR on siLuc3: log	  
6: RNA STAR on siLuc2: map ped.bam	  
5: RNA STAR on siLuc2: spli ce 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

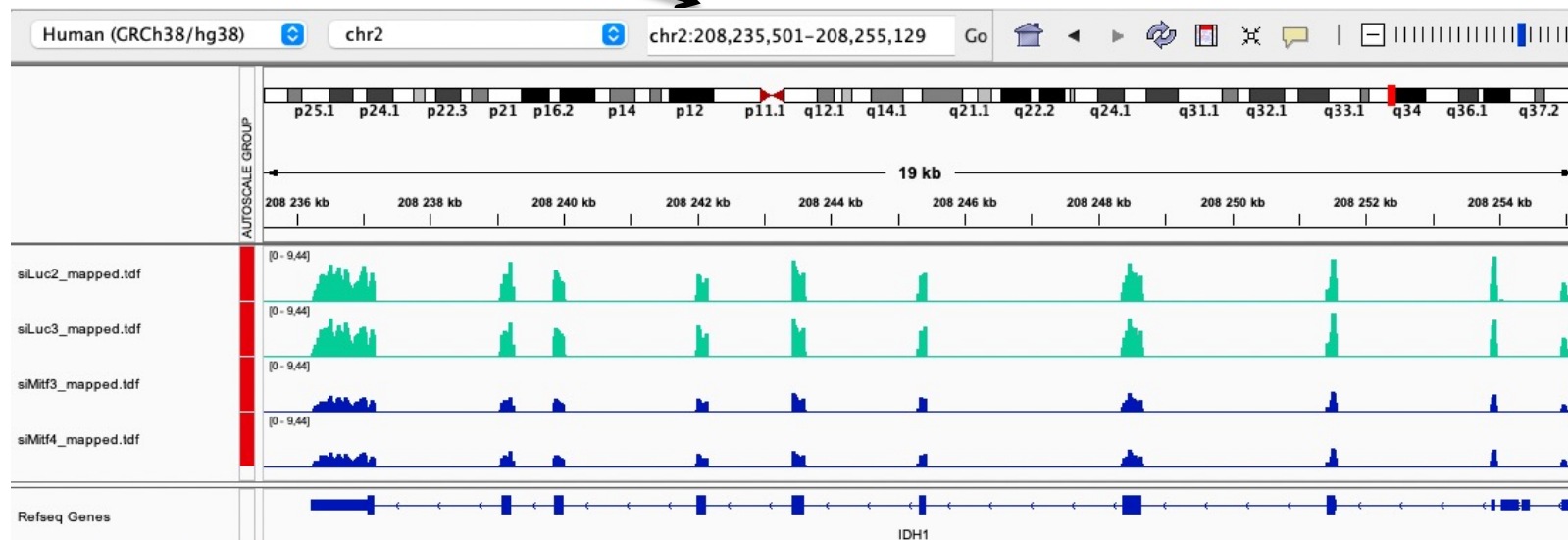
IGV : File → Load from file and select the 4 tdf files

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

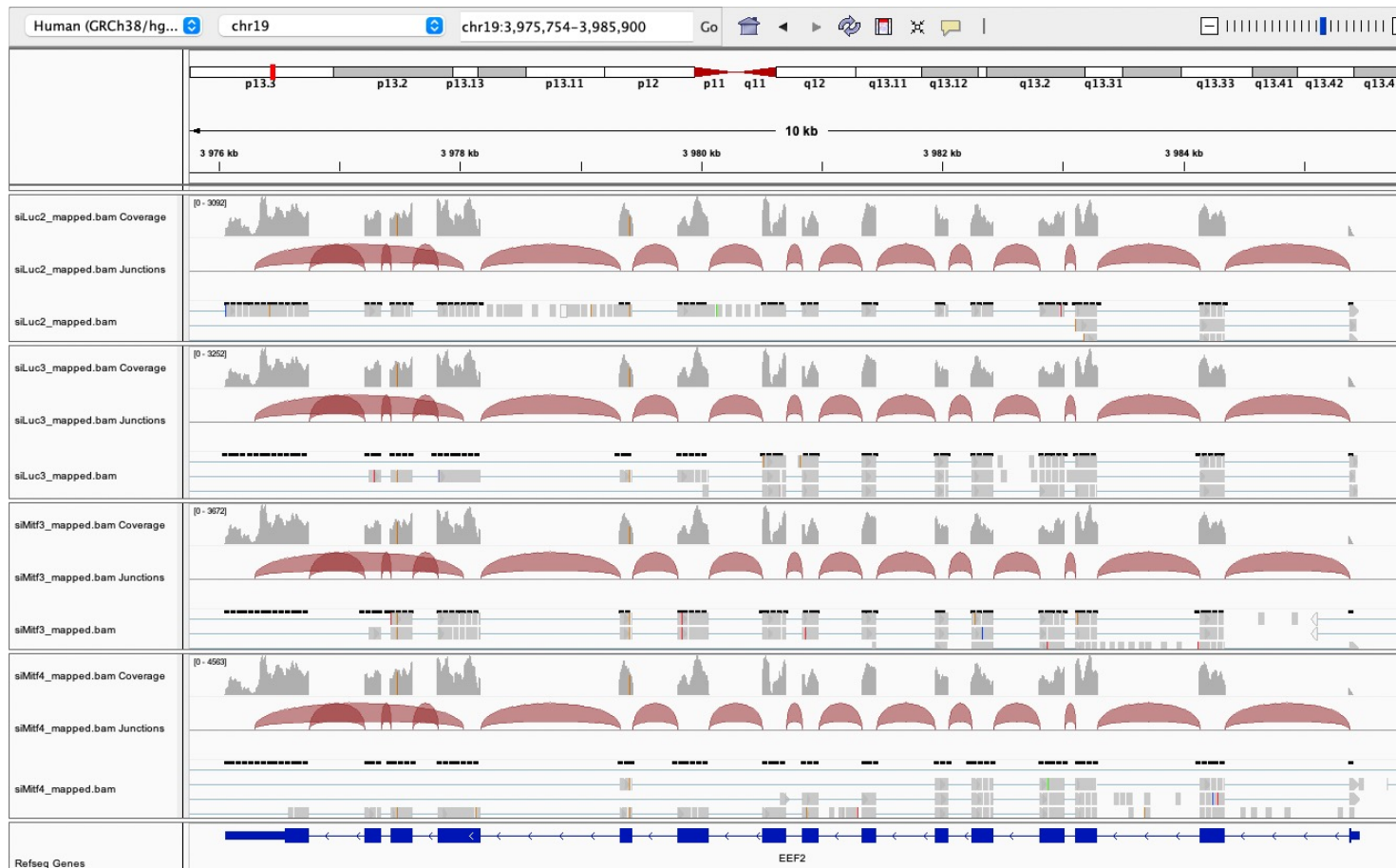
Search for *Idh1*



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

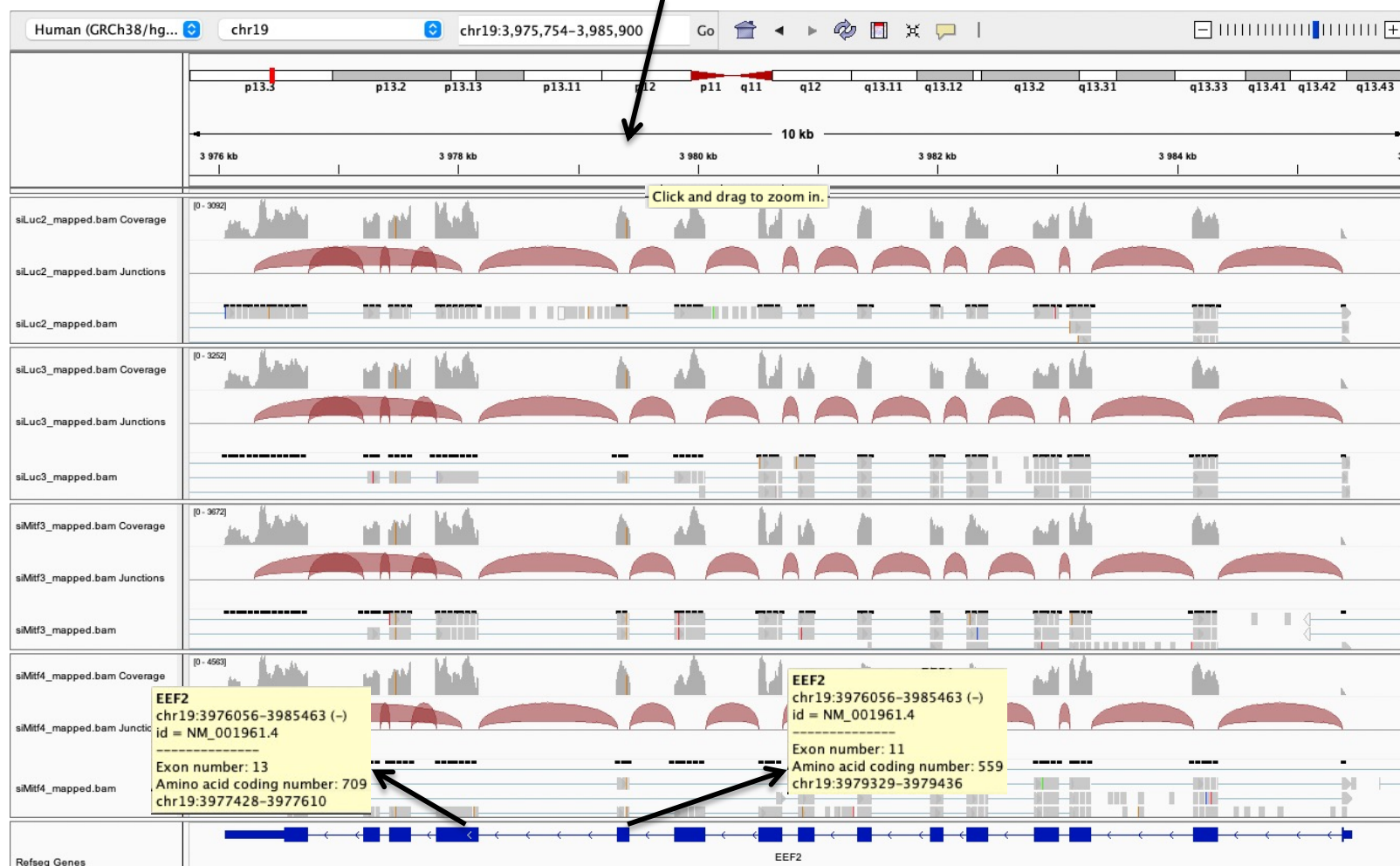
Exercise 2 – Question 3

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



Exercise 2 – Question 3

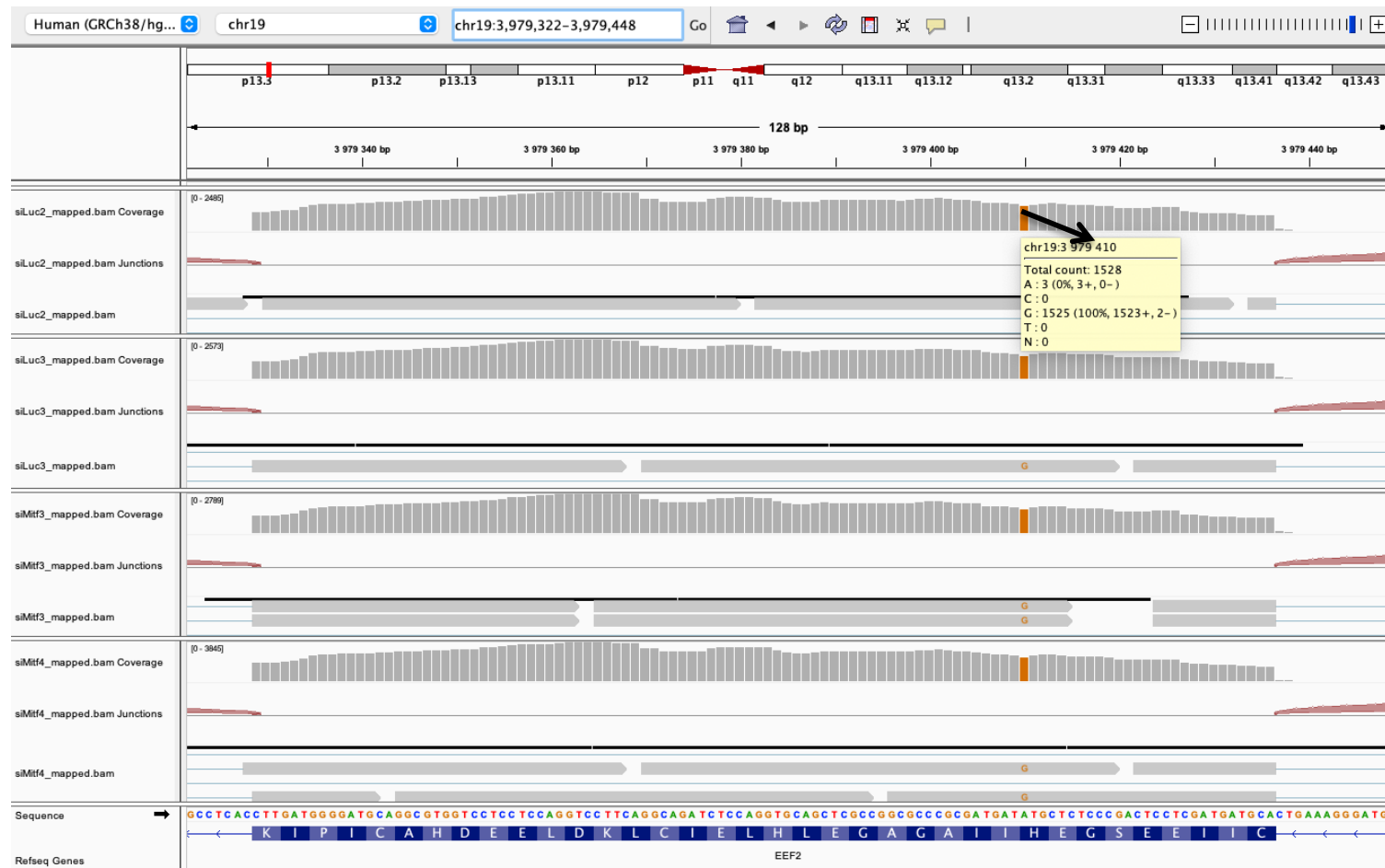
Exon numbers are provided on annotation track
Click and drag on a region to zoom in



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

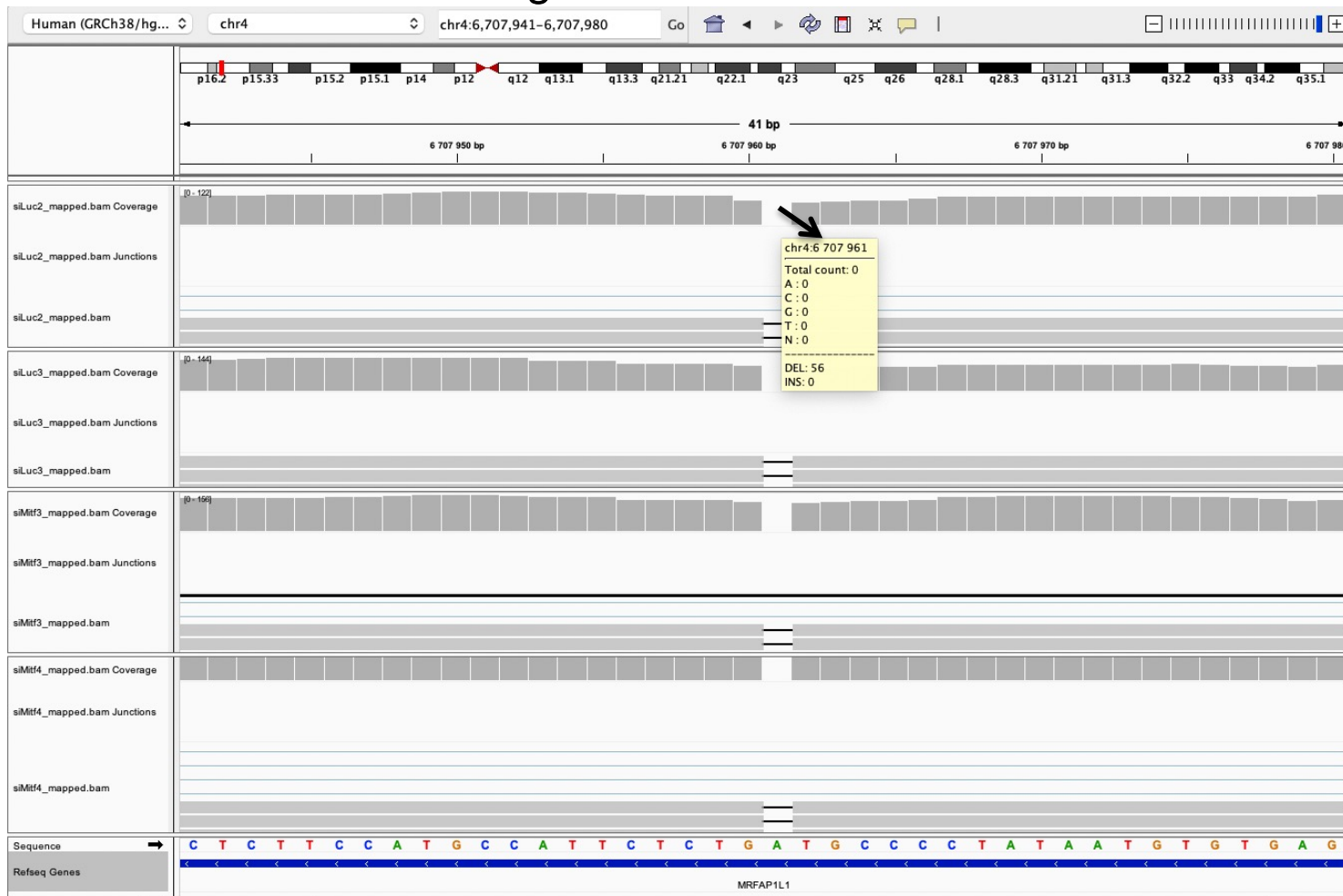
■ *Eef2* exon 13

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



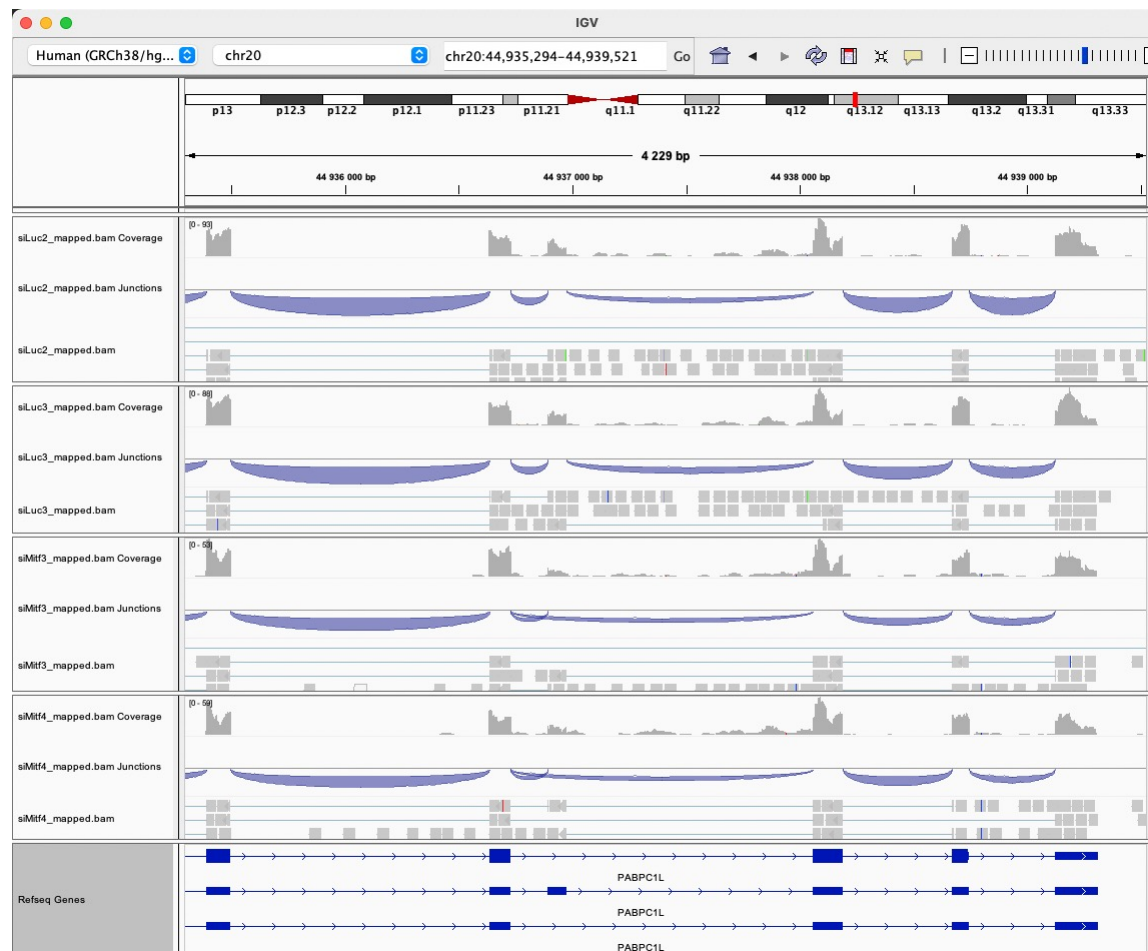
Exercise 2 – Question 4

- Position chr4:6707961 :
 - Deletion vs reference genome



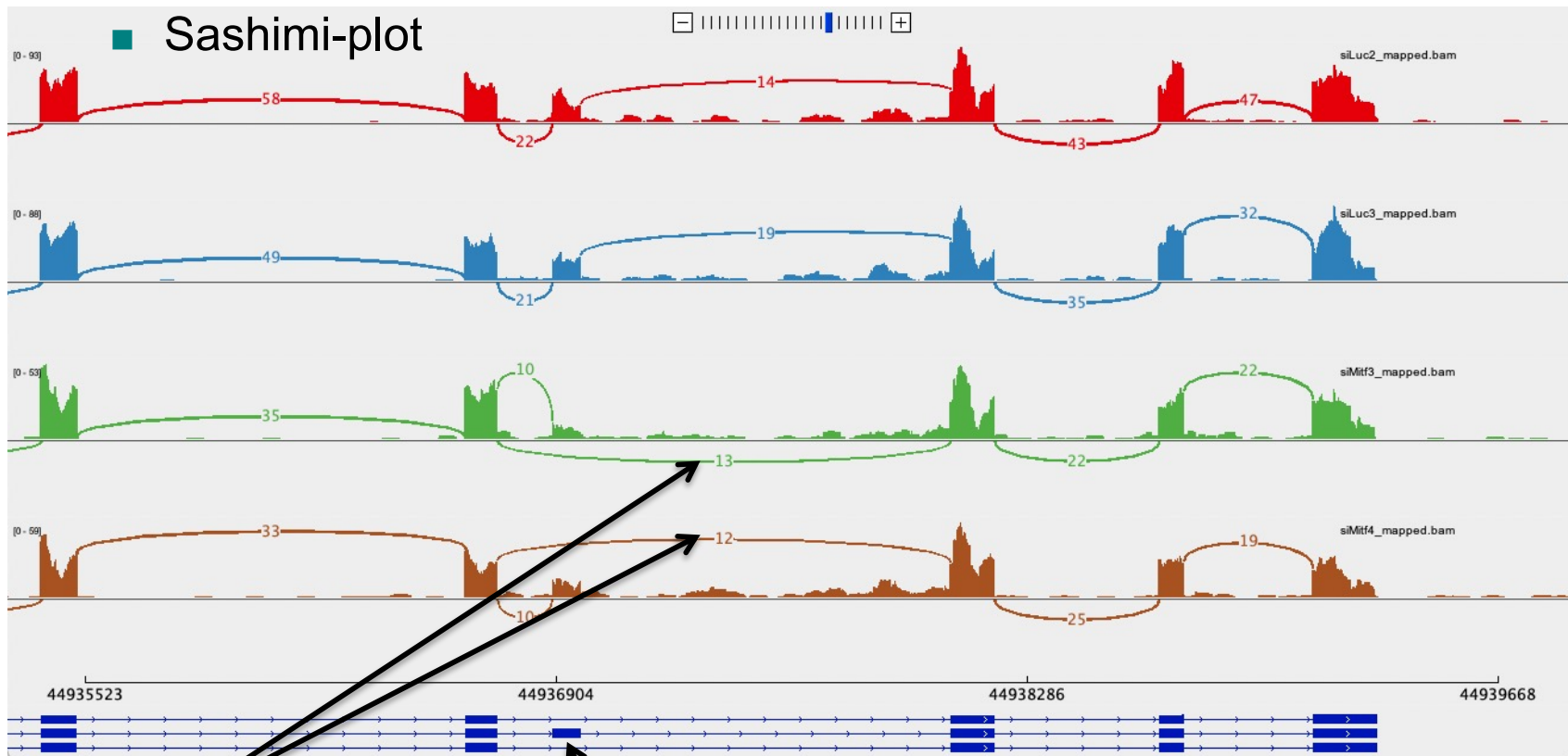
Exercise 2 – Question 5

- 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 :



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

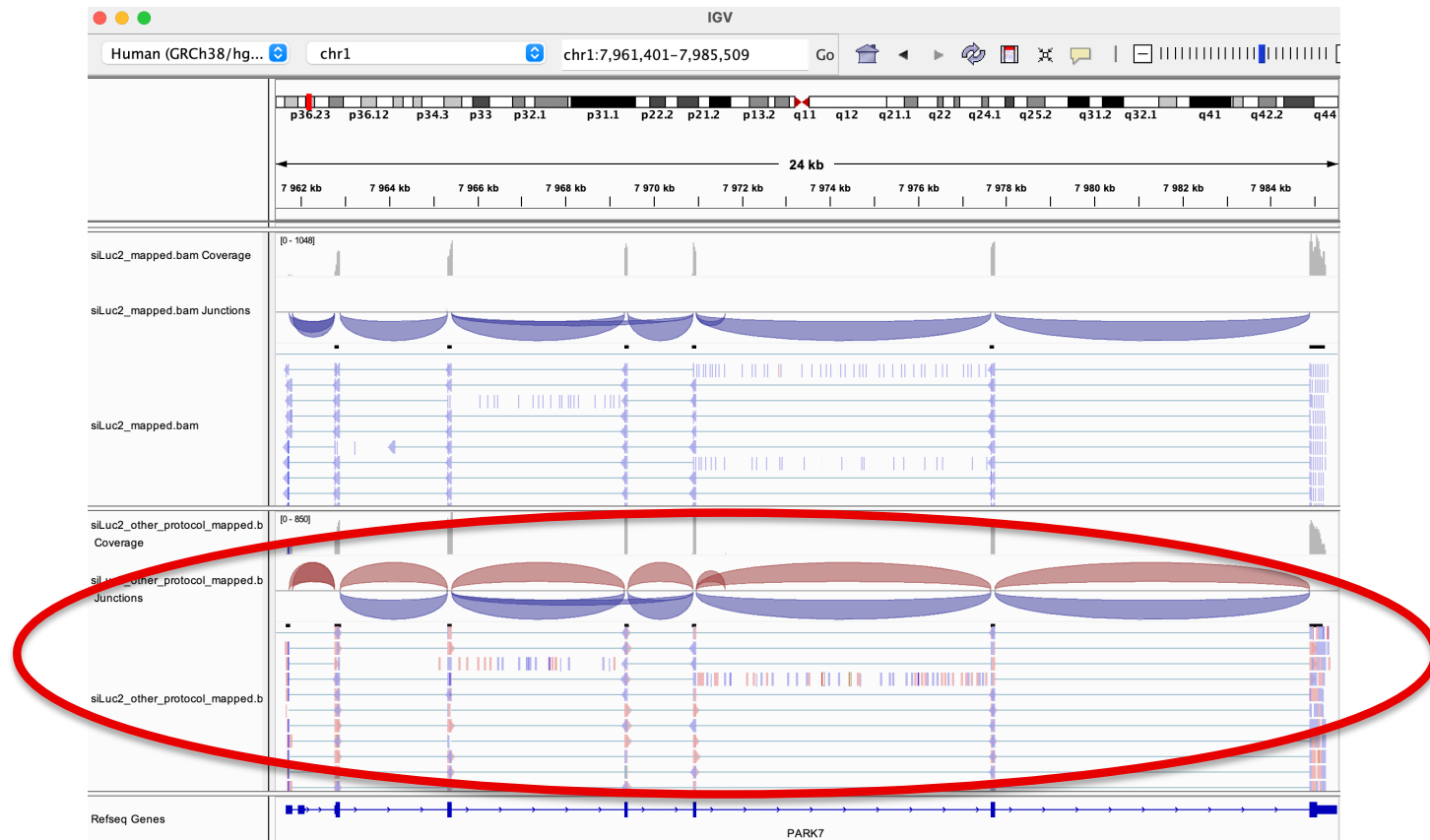


Exercise 2 – Question 5

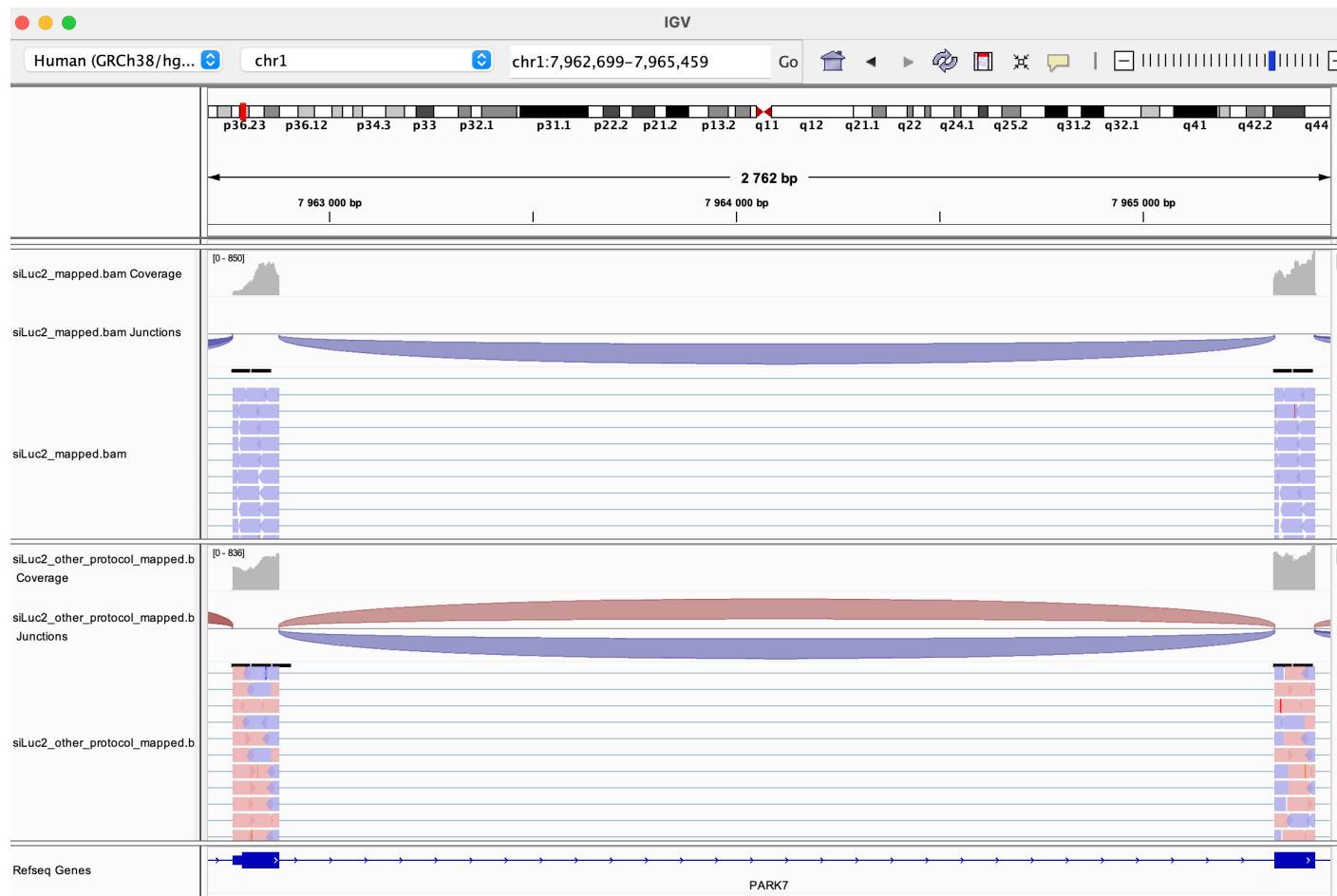
- You can save your IGV session
 - To save the current state of your IGV session to a named session file
 - File → Save Session
 - Data files must stay at the same location
- Use File → Open session to restore a saved session

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



Exercise 2 – Question 6



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