NGS read mapping : answers to questions

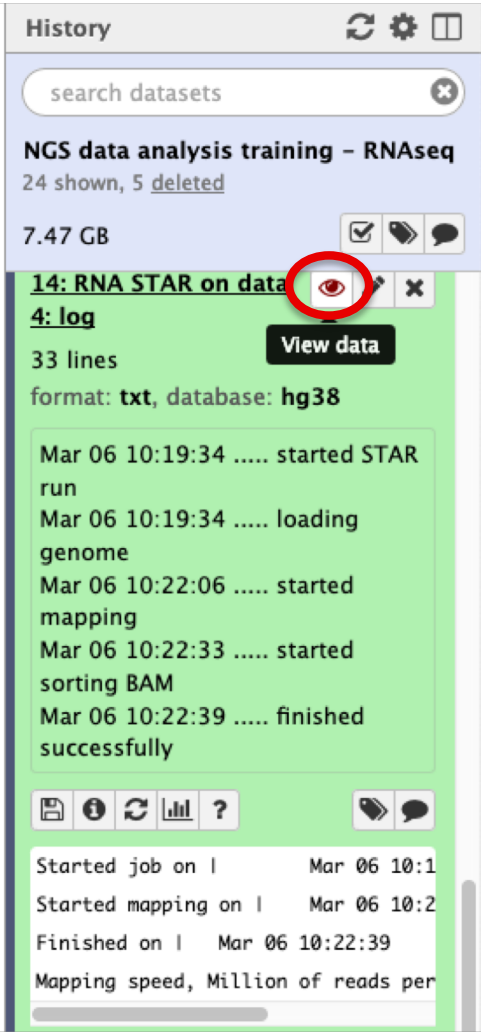
Céline Keime
keime@igbmc.fr

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 |       [illegible]
                 Uniquely mapped reads % |       85.28%
                   Average mapped length |       [illegible]
                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%
```

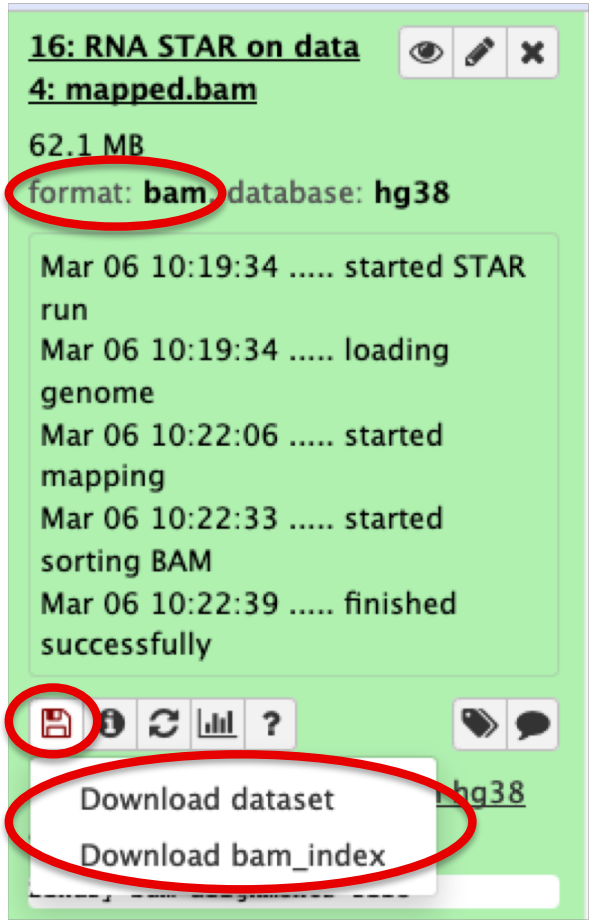


Exercise 1
2. Alignment file

- Galaxy

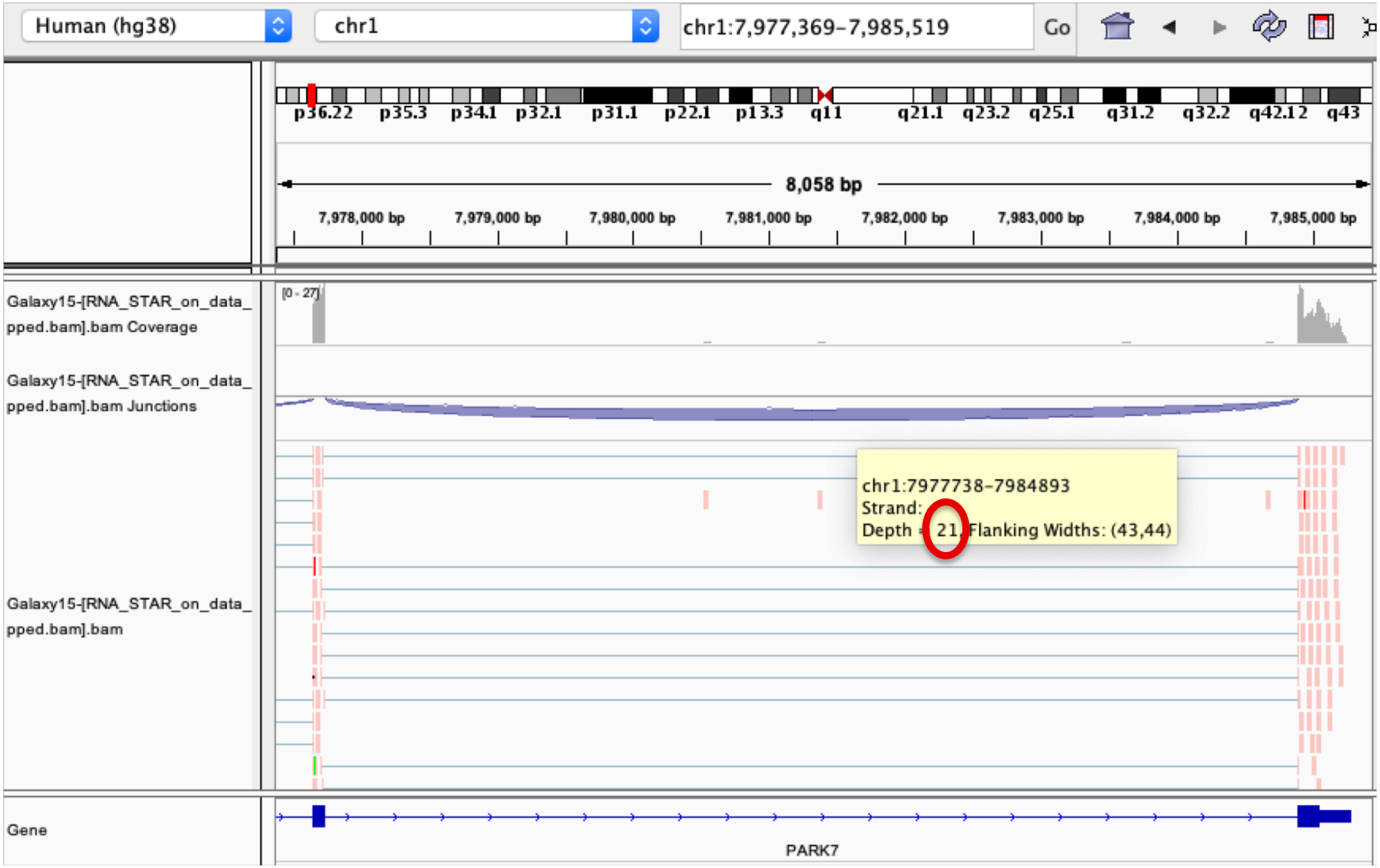
  3.1. STAR provides an alignment in BAM format

  3.2. Download this file together with the corresponding index (in the same directory)



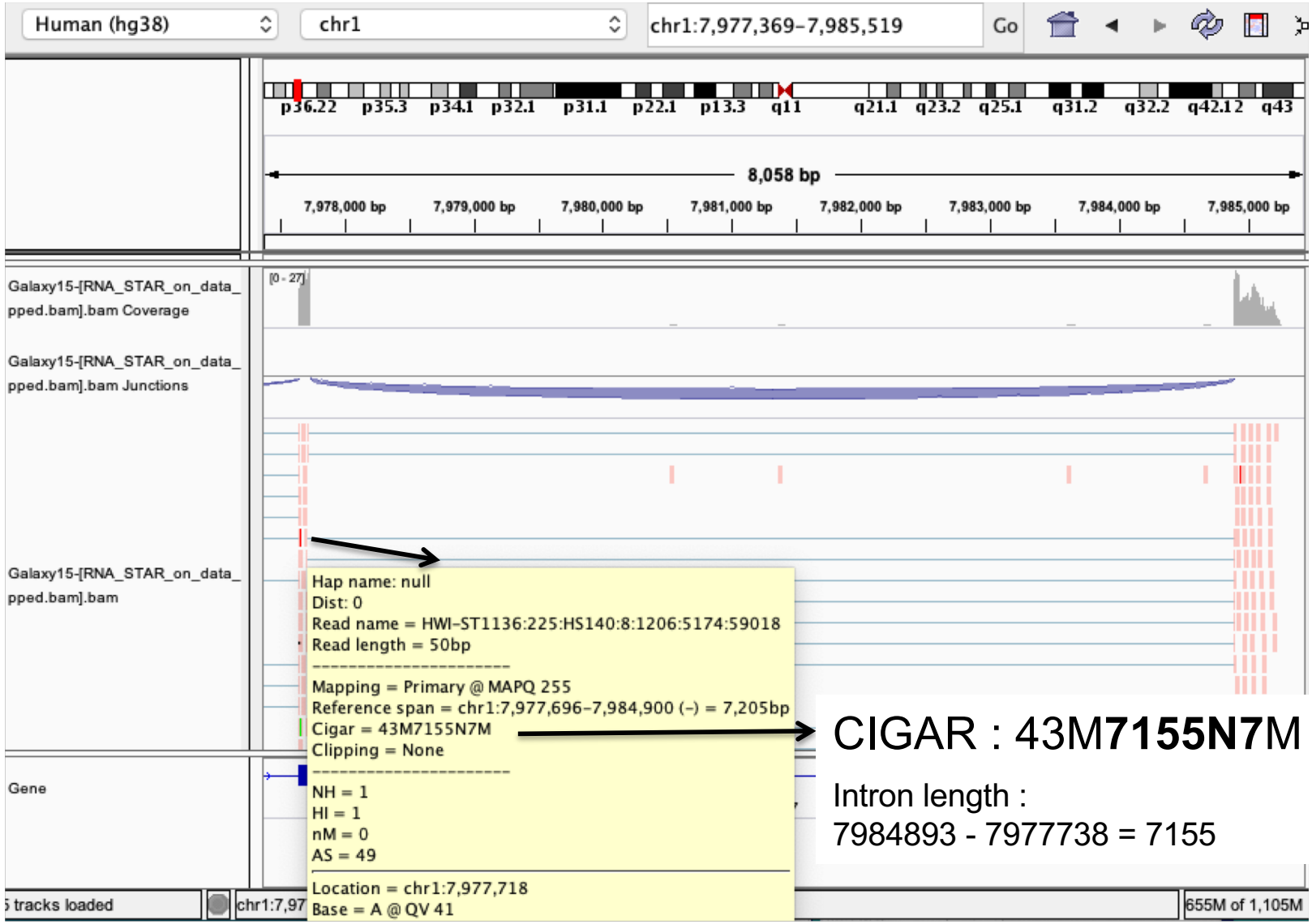
- IGV
  - File → Load from file and choose the downloaded BAM file

Exercise 1
2. Splice junction



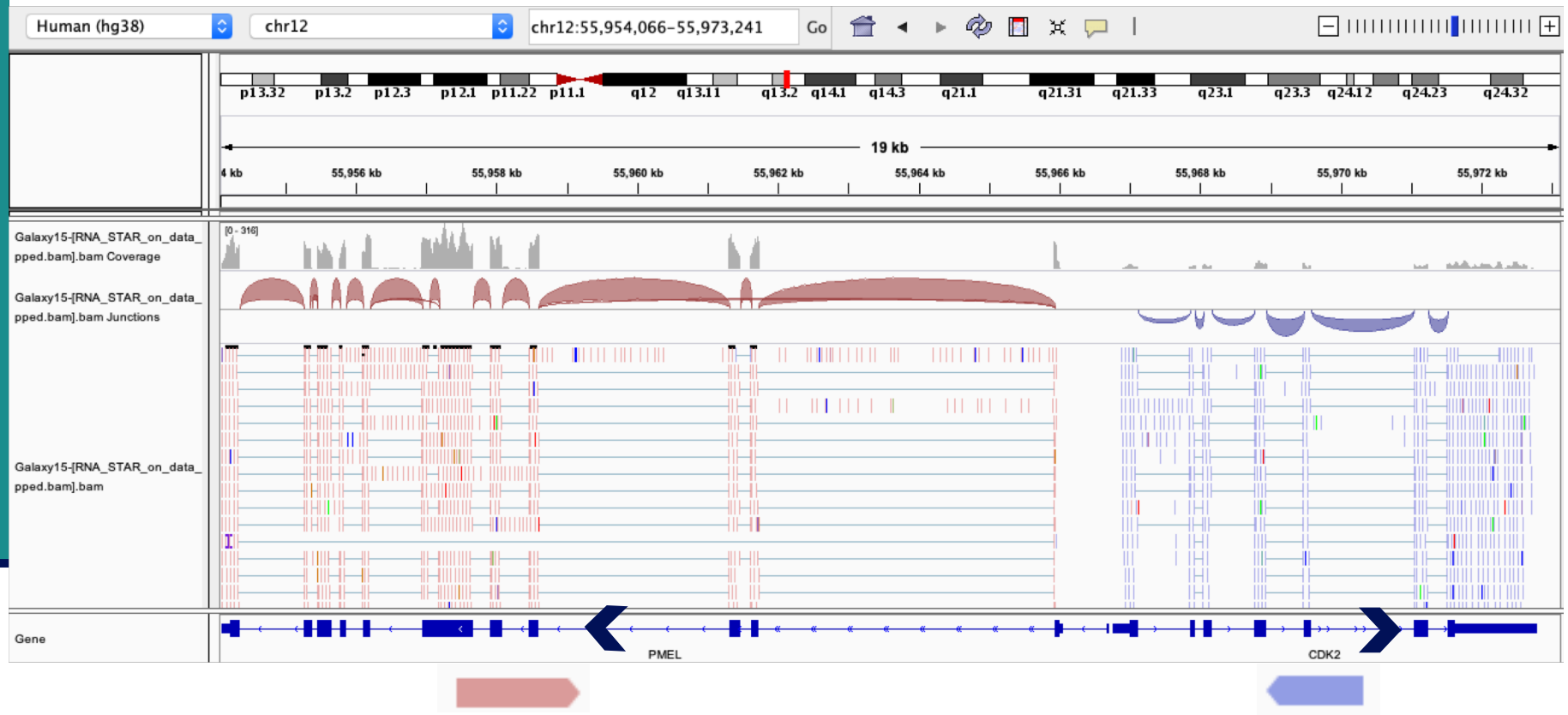
→ 21 alignments span the junction that joins the last 2 exons of *Park7* gene

Exercise 1
2. Splice junction



Exercise 1
2. Strand specificity

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

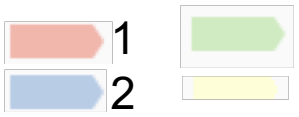


The library has been prepared with a directional mRNAseq protocol which retains strand information :
all reads are in the opposite direction as the transcribed strand

Exercise 1
2. Multiple mapped reads

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

Number of reported alignments :



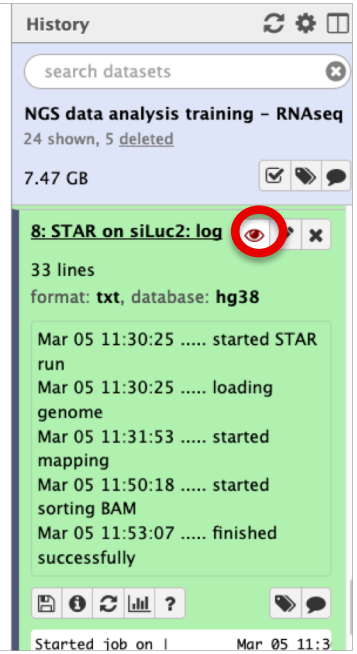
There are multiple aligned reads on this gene

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 |       [illegible]
                 Uniquely mapped reads % |       85.30%
                   Average mapped length |       [illegible]
                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%
```



STAR on siLuc2:	Uniquely mapped reads %	85.30%
STAR on siLuc3:	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

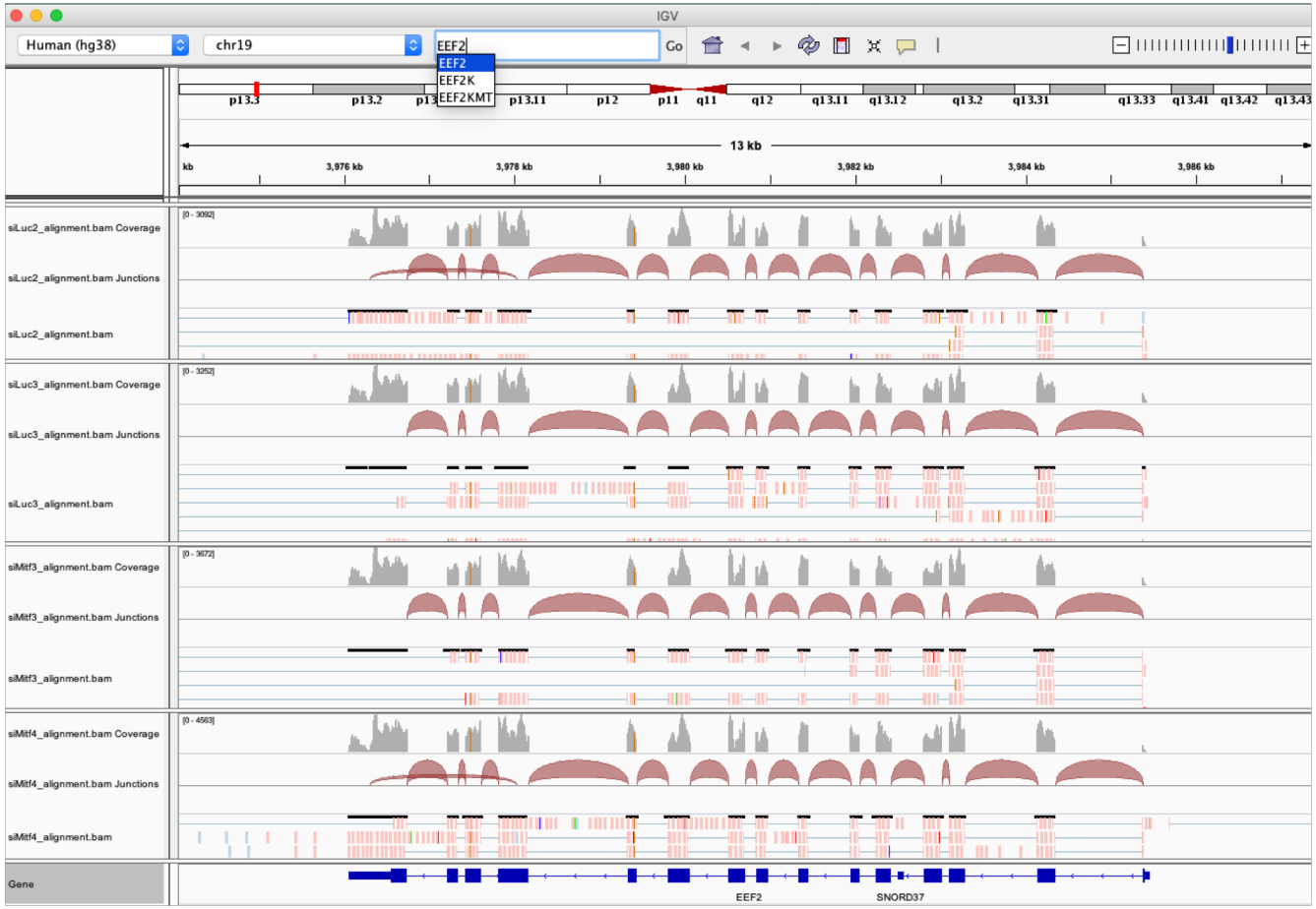
Exercise 2 – Question 2
Idh1 gene expression

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

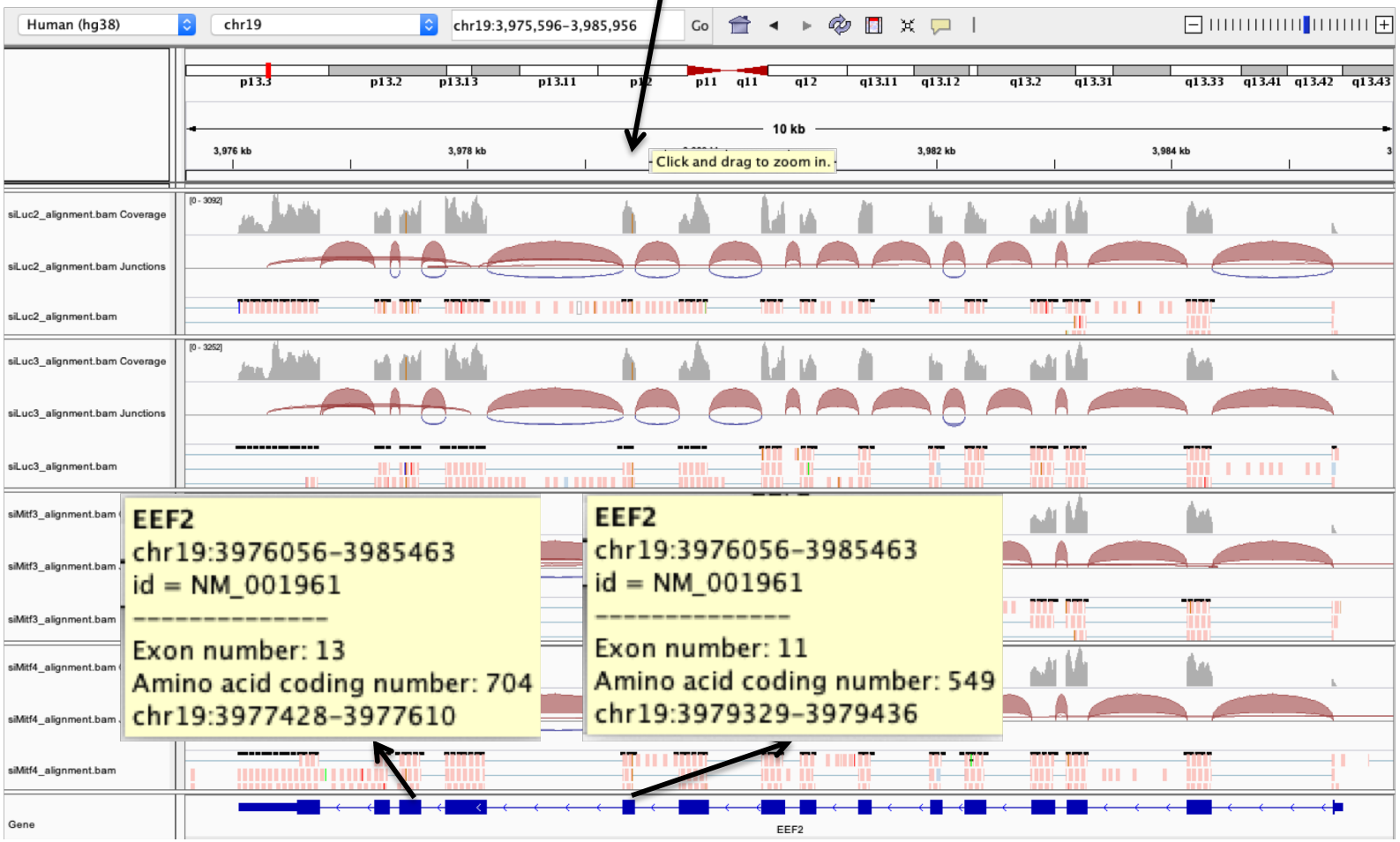
- File → new session
- File → load from files and load the 4 BAM files
- Search for EEF2



Exercise 2 – Question 3

Exon numbers are provided on annotation track

Click and drag on a region to zoom in



Exercise 2 – Question 3

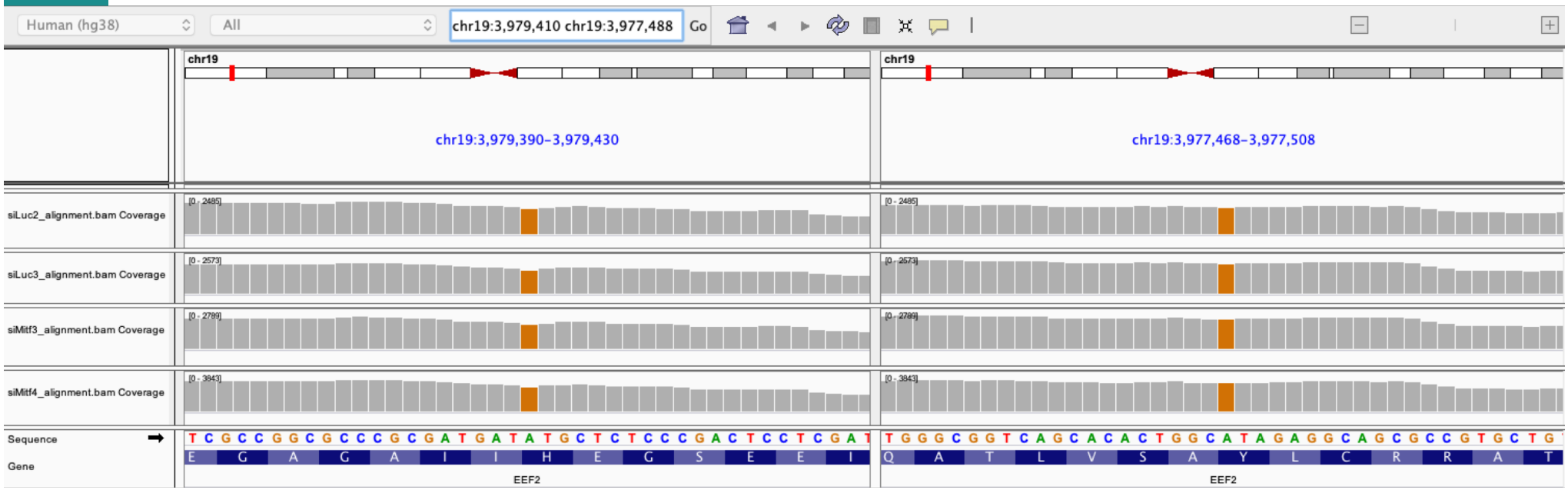
- *Eef2* exon 11
  - chr19:3,979,410 : G in ~100% of the reads, A in the genome

Exercise 2 – Question 3

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

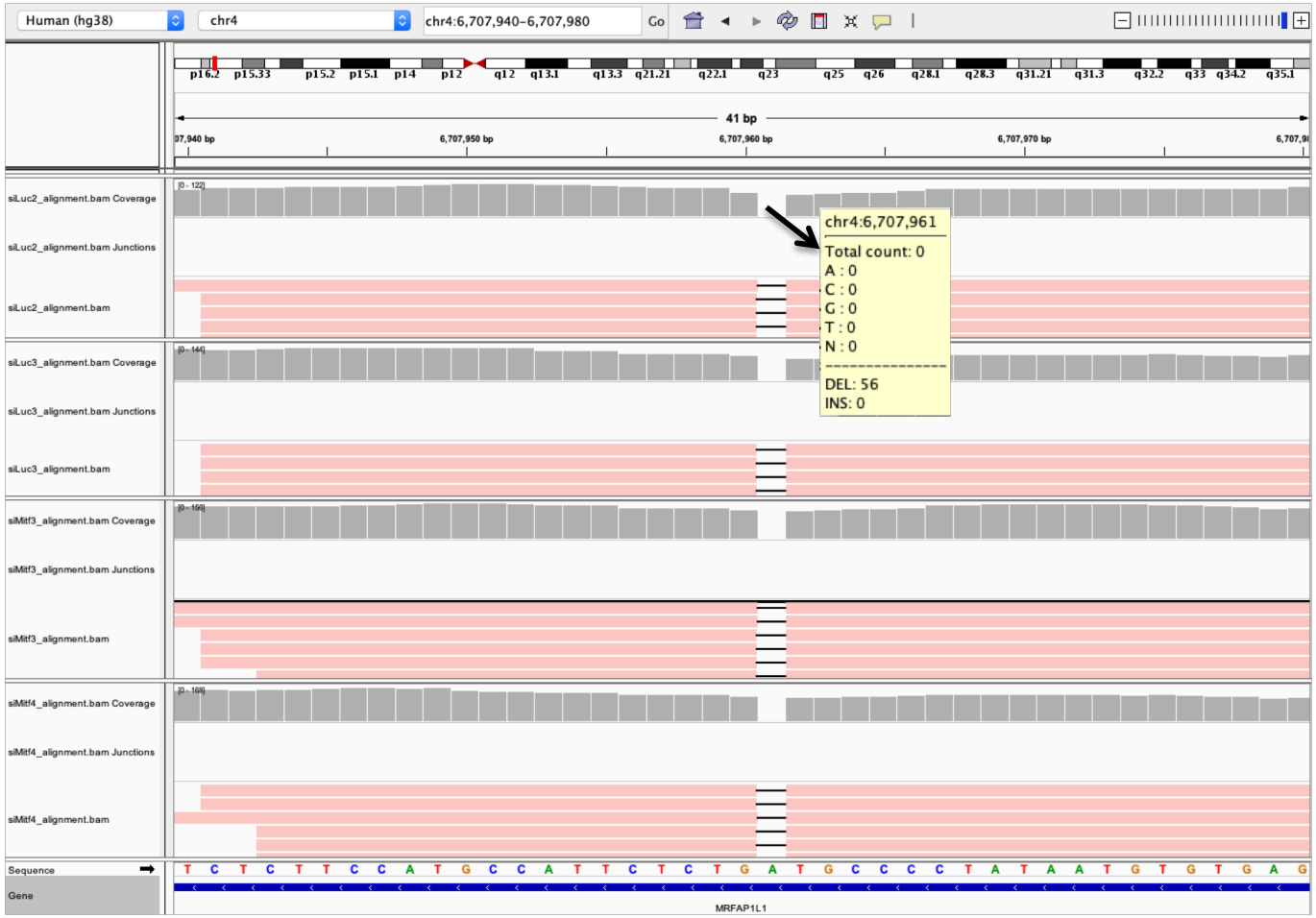
Exercise 2 – Question 3

- 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



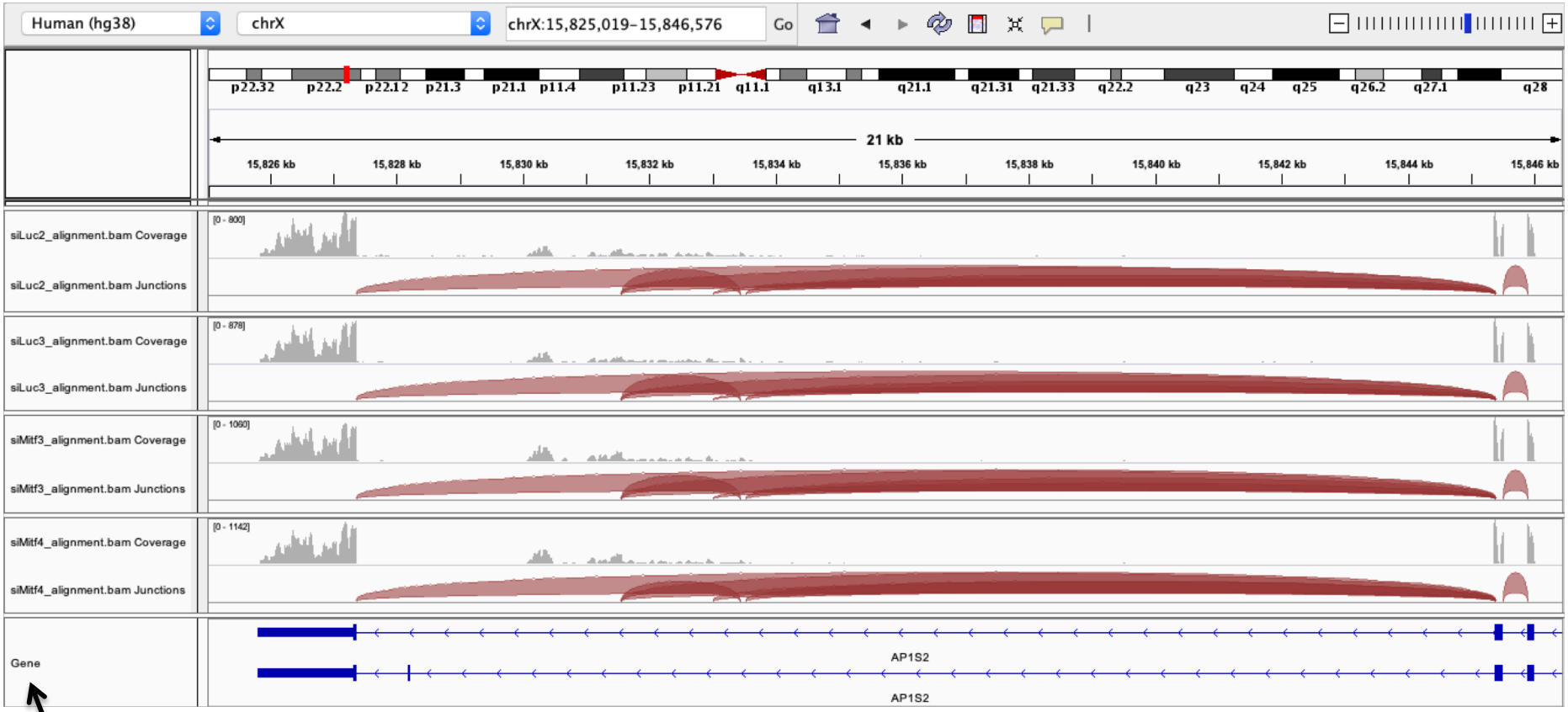
Exercise 2 – Question 4

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



Exercise 2 – Question 5

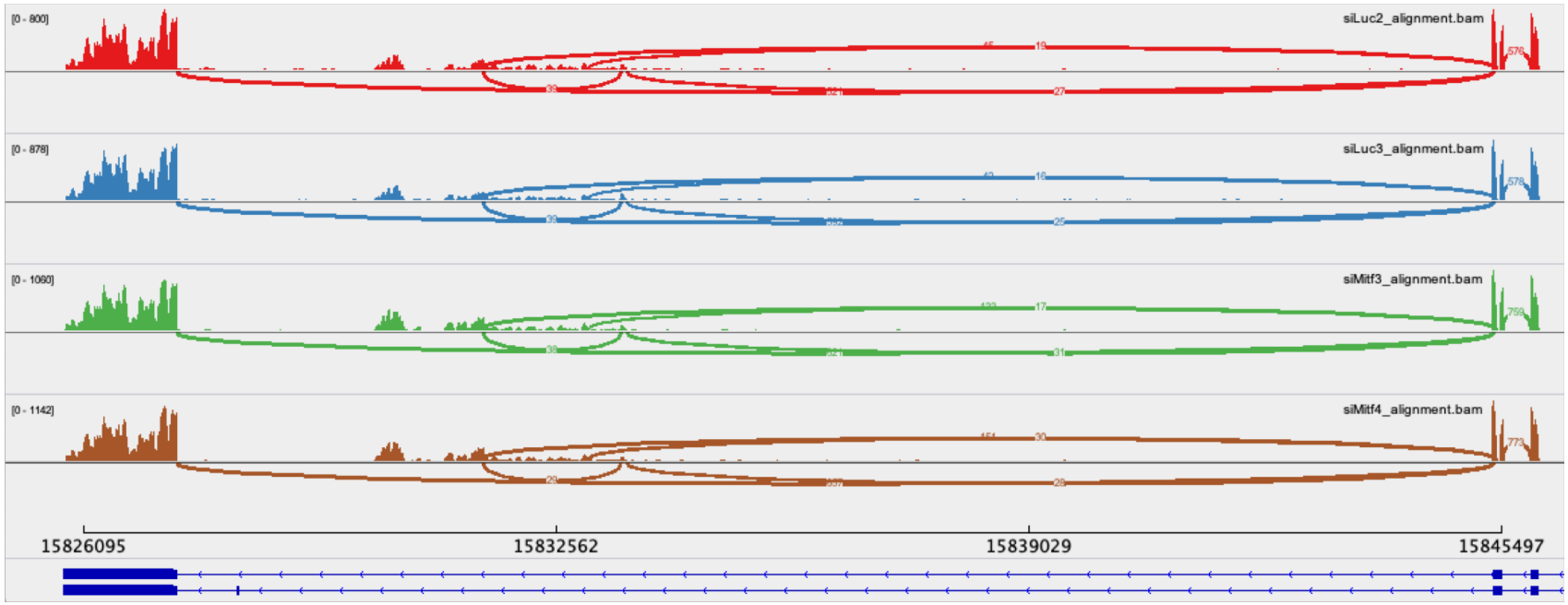
- Region chrX:15,825,019-15,846,576 :
  - Junctions corresponding to exons not annotated in Refseq



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

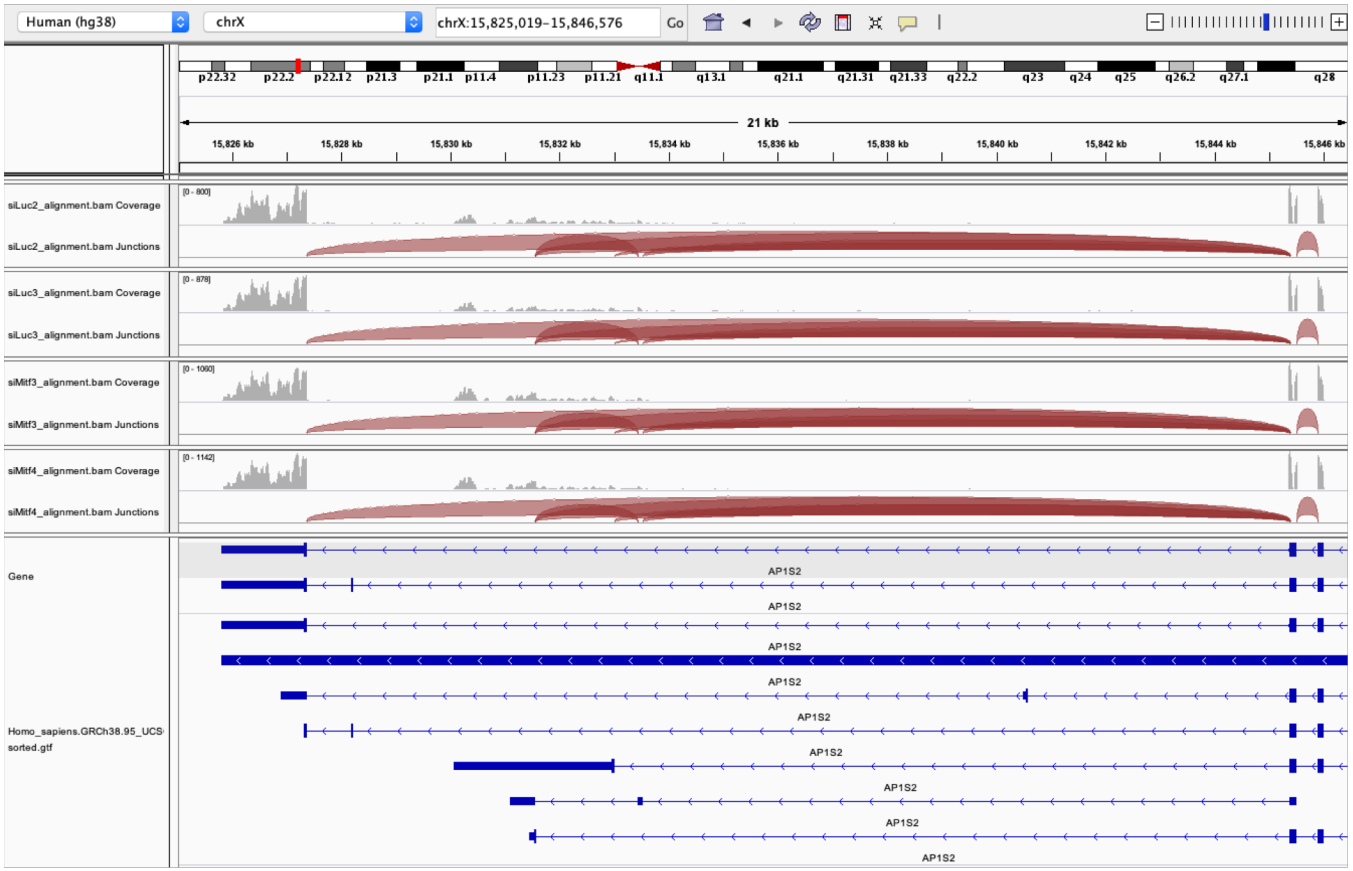
Exercise 2 – Question 5

- Region chrX:15,825,019-15,846,576 :
  - Junctions corresponding to exons not annotated in Refseq
  - Sashimi-plot :
    - Right-click on a BAM track → Sashimi plot → Select Alignment Tracks : all alignments



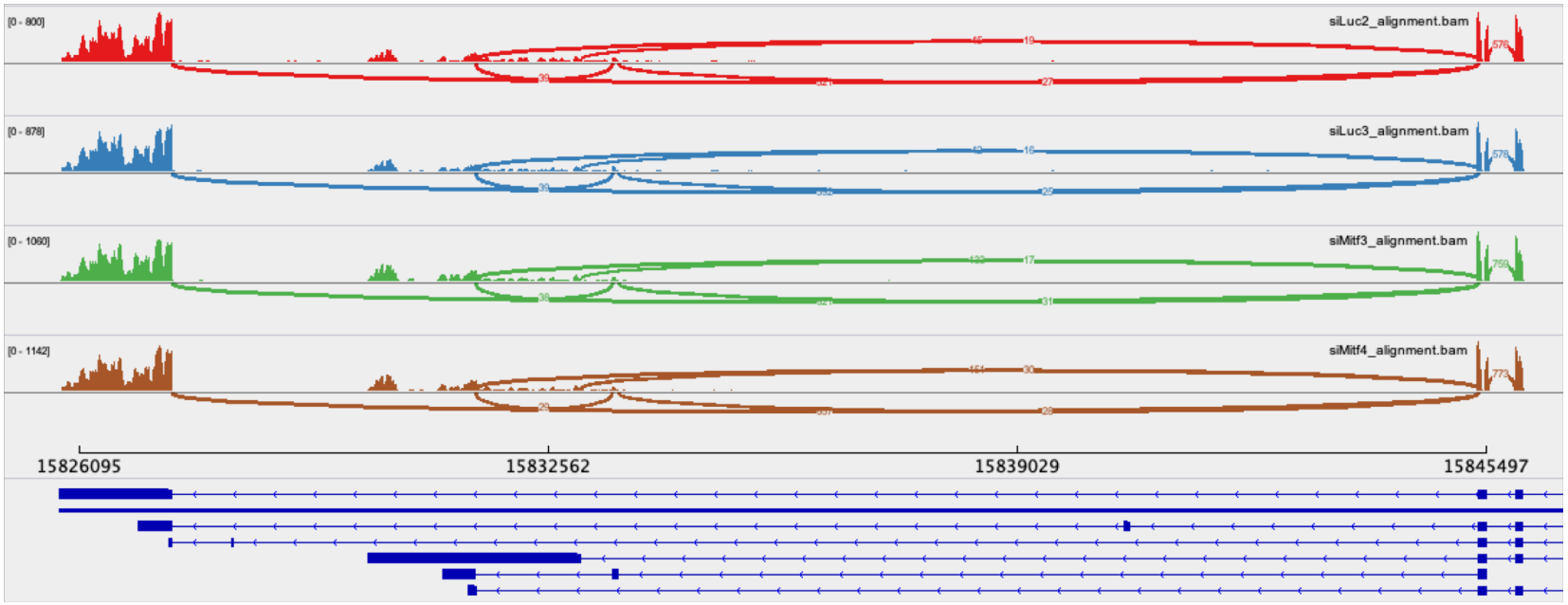
Exercise 2 – Question 5

- Region chrX:15,825,019-15,846,576 :
  - Ensembl annotations : more exons annotated in this region
    - File → load from file → Homo_sapiens.GRCh38.95_UCSC_chr.sorted.gtf
    - Right-click on the annotation track and select Expanded



Exercise 2 – Question 5

- Region chrX:15,825,019-15,846,576 :
  - Ensembl annotations :
    - Sashimi-plot : Right-click on a BAM track → Sashimi plot → Select Gene Track : Ensembl annotations → Select Alignment Tracks : all alignments

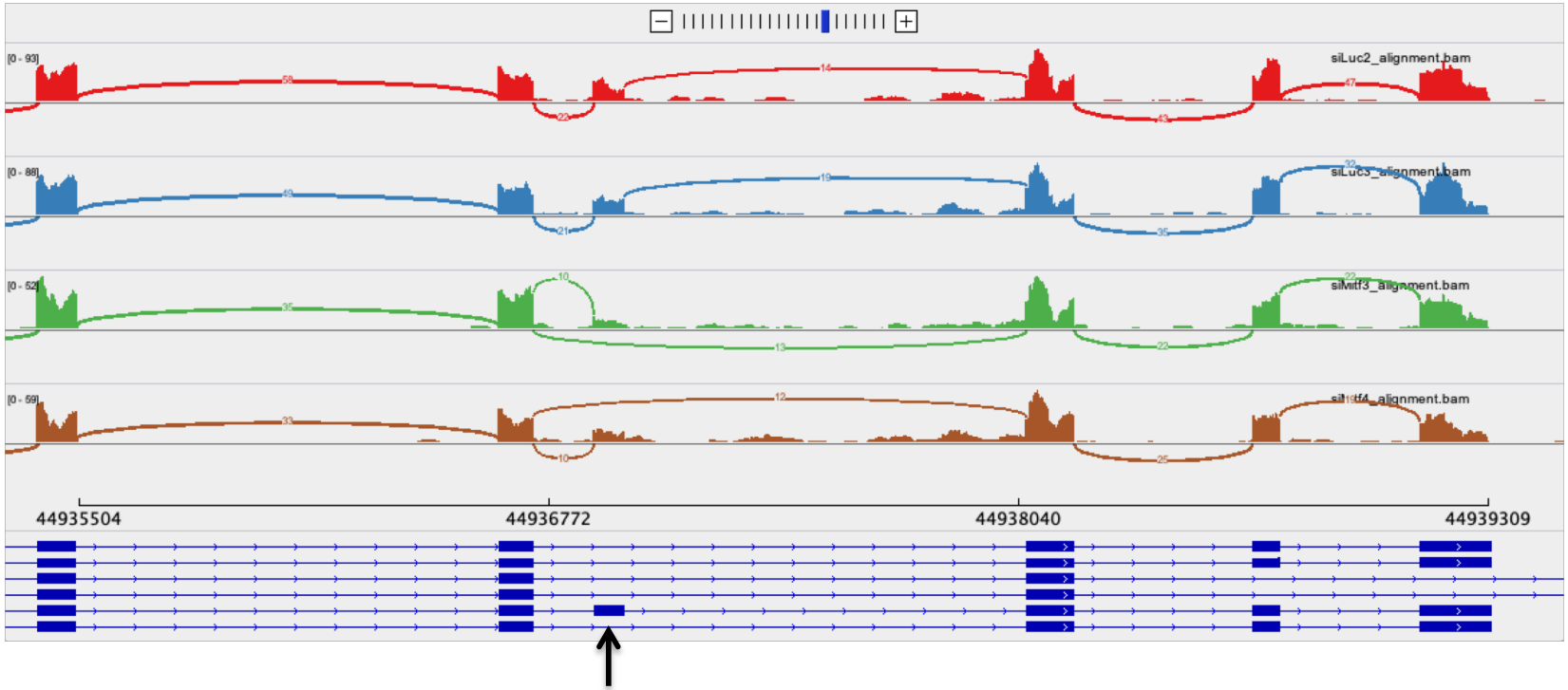


➔ Very useful to quickly visualize splicing events along genomic regions of interest

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

Exercise 2 – question 6

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

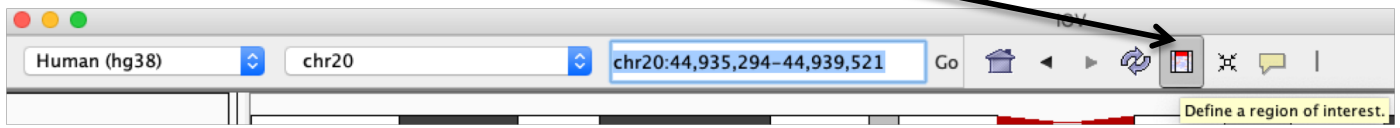


We detect an isoform without this exon in siMitf samples but not in siLuc ones

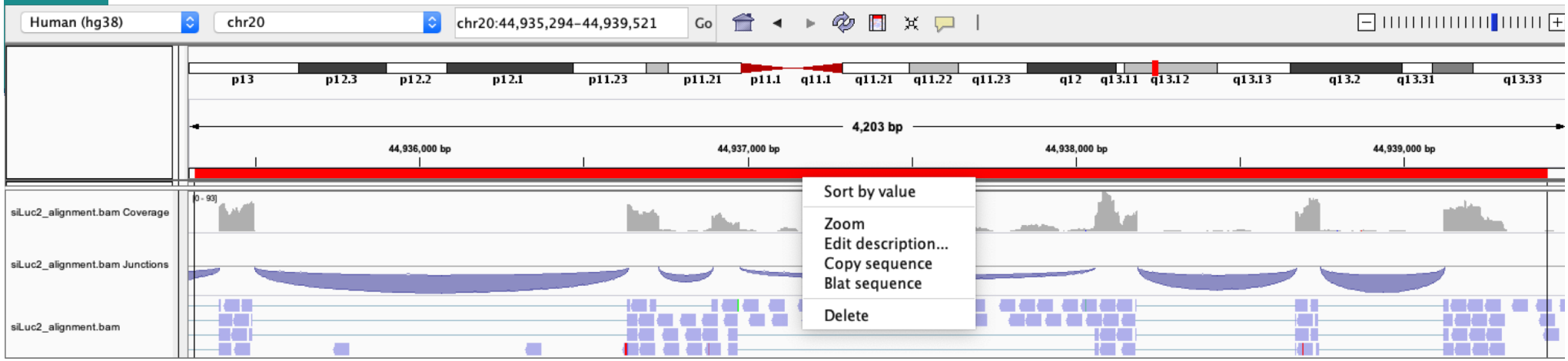
IGV is only a visualization tool
In-depth analysis using paired-end data with more coverage is needed

Exercise 2 – question 6

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



  - 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

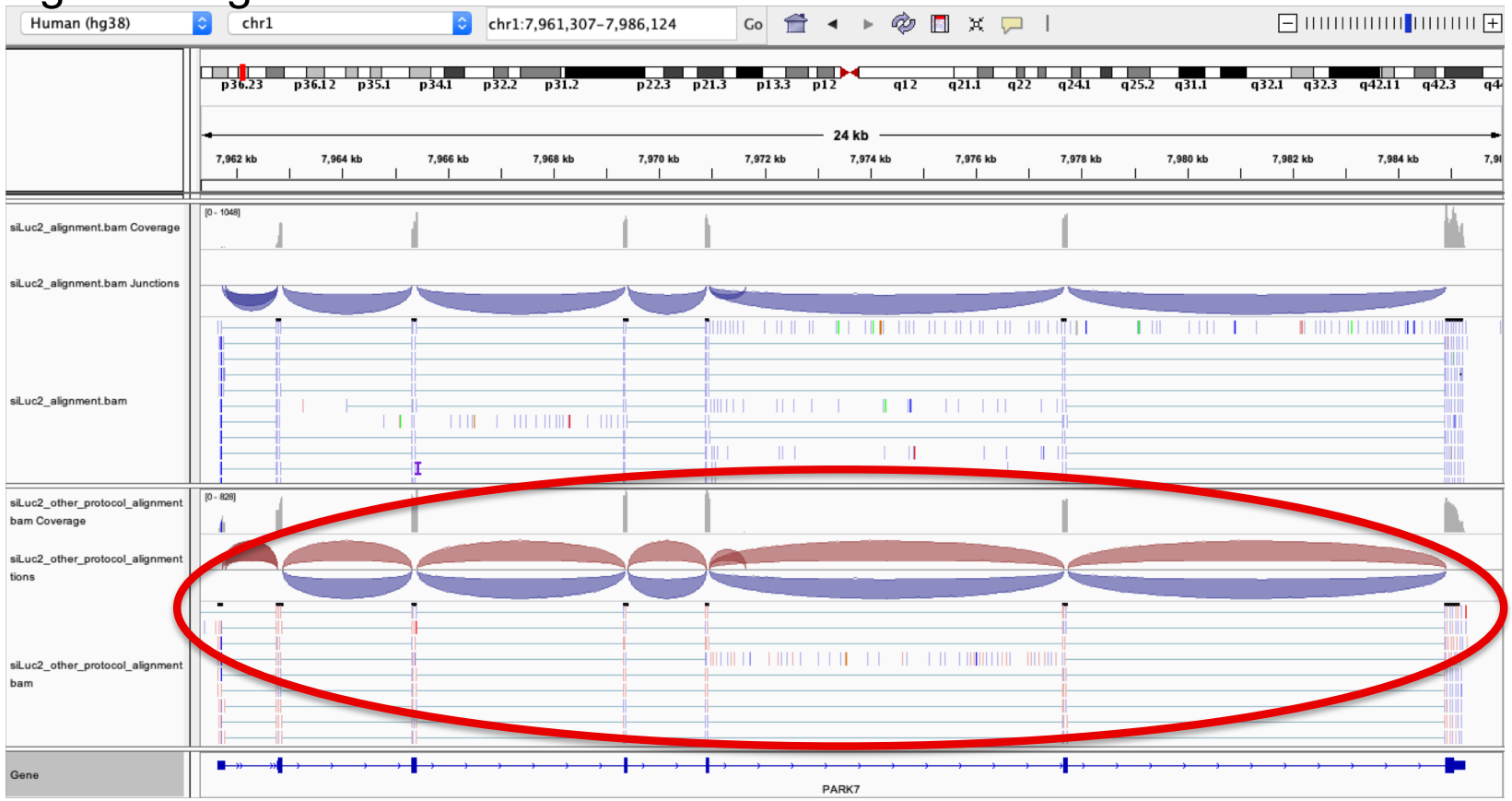


Exercise 2 – question 6

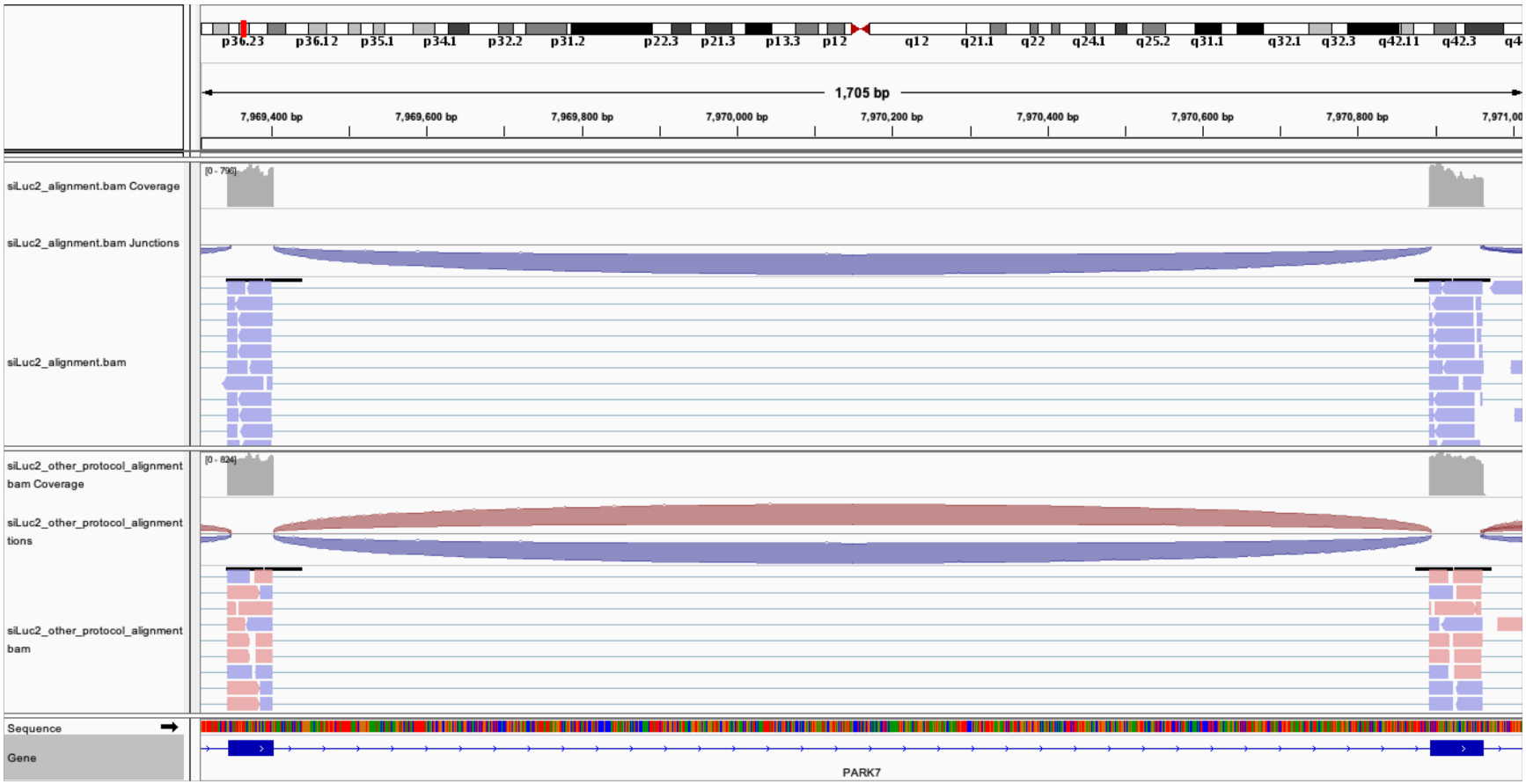
- You can save your IGV session
  - To save the current state of your IGV session to a named session file
  - File → Save Session
  - Data files must stay at the same location
- Use File → Open session to restore a saved session

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)