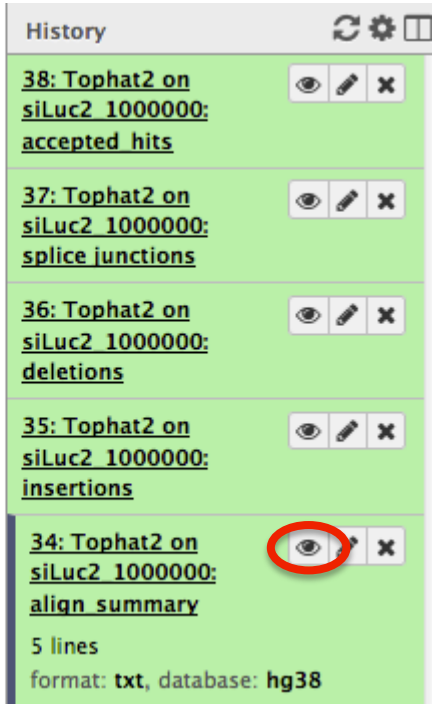
NGS read mapping : answers to questions

Céline Keime
keime@igbmc.fr

Exercise 1
1. Alignment summary statistics

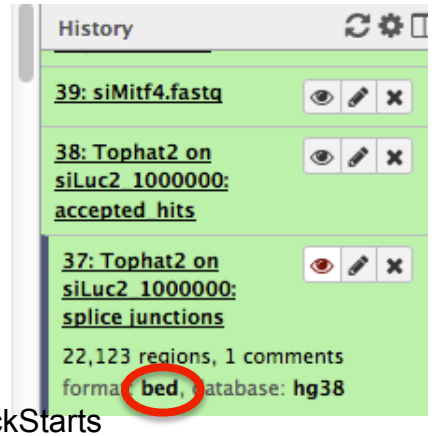
```
Reads:
          Input      :   1000000
           Mapped    :    947951 (94.8% of input)
            of these:    126745 (13.4%) have multiple alignments (1 have >20)
94.8% overall read mapping rate.
```



1.1. 947,951 reads mapped onto hg38
1.2. 13.4% of these reads have multiple alignments

Exercise 1
2. Splice junctions

1	2	3	4	5	6	7	8	9	10	11	12
track name=junctions description="TopHat junctions"											
chr1	15015	15822	JUNC00000001	1	-	15015	15822	255,0,0	2	23,27	0,780
chr1	18337	18521	JUNC00000002	1	-	18337	18521	255,0,0	2	25,25	0,159
chr1	18337	24758	JUNC00000003	2	-	18337	24758	255,0,0	2	29,21	0,6400
chr1	18337	195283	JUNC00000004	2	-	18337	195283	255,0,0	2	29,21	0,176925
chr1	18354	18947	JUNC00000005	1	-	18354	18947	255,0,0	2	15,35	0,558
chr1	164755	165897	JUNC00000006	1	-	164755	165897	255,0,0	2	36,14	0,1128
chr1	188860	189044	JUNC00000007	1	-	188860	189044	255,0,0	2	25,25	0,159
chr1	188860	195283	JUNC00000008	2	-	188860	195283	255,0,0	2	29,21	0,6402
chr1	523351	524493	JUNC00000009	1	-	523351	524493	255,0,0	2	36,14	0,1128



Chr Start End Name Score Strand thickStart thickEnd itemRgb blockCount blockSizes blockStarts (relative to *chromStart)*

Number of alignments spanning the junction

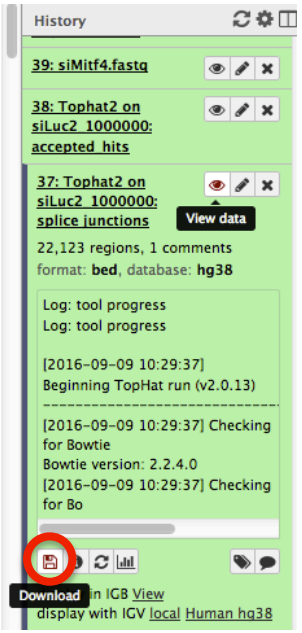
Each junction consists of 2 connected BED blocks → Each block is as long as the maximal overhang of any read spanning the junction

2.1. Splice junctions provided in a BED file

Exercise 1
2.2. Splice junctions visualization

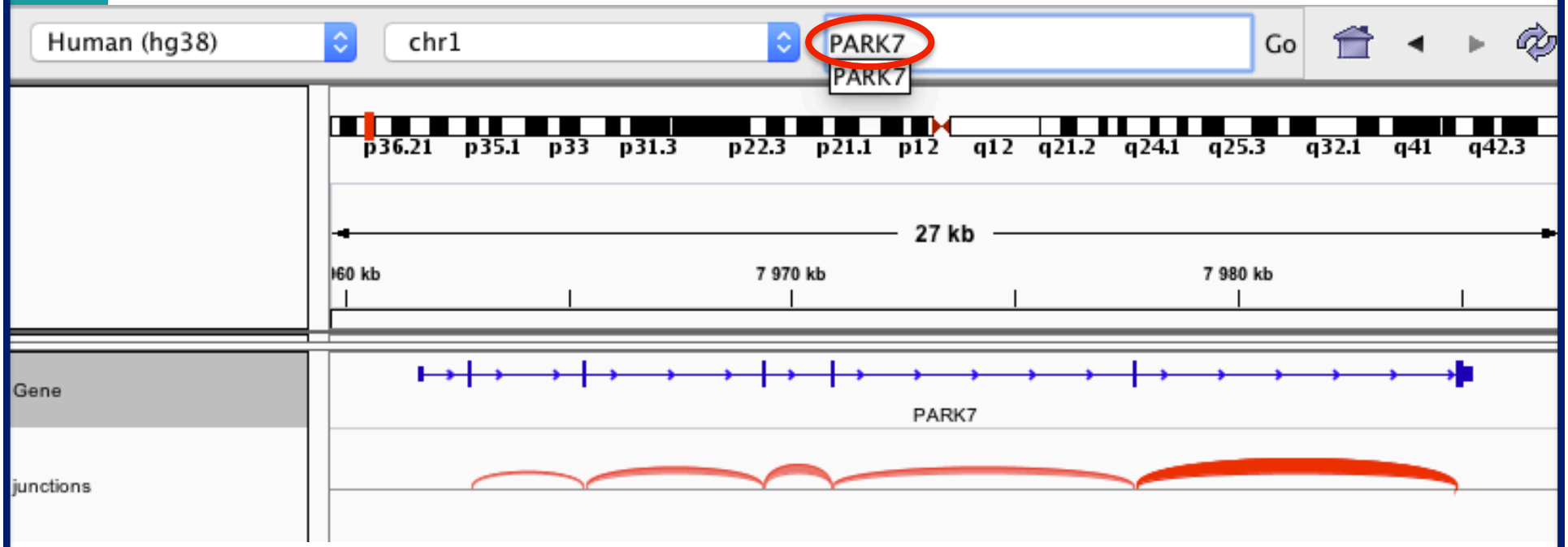
- Galaxy
  - Download splice junctions BED file

1	2	3	4	5	6	7	8	9	10	11	12
track name=junctions description="TopHat junctions"											
chr1	15015	15822	JUNC00000001	1	-	15015	15822	255,0,0	2	23,27	0,780
chr1	18337	18521	JUNC00000002	1	-	18337	18521	255,0,0	2	25,25	0,159
chr1	18337	24758	JUNC00000003	2	-	18337	24758	255,0,0	2	29,21	0,6400
chr1	18337	195283	JUNC00000004	2	-	18337	195283	255,0,0	2	29,21	0,176925
chr1	18354	18947	JUNC00000005	1	-	18354	18947	255,0,0	2	15,35	0,558
chr1	164755	165897	JUNC00000006	1	-	164755	165897	255,0,0	2	36,14	0,1128
chr1	188860	189044	JUNC00000007	1	-	188860	189044	255,0,0	2	25,25	0,159
chr1	188860	195283	JUNC00000008	2	-	188860	195283	255,0,0	2	29,21	0,6402
chr1	523351	524493	JUNC00000009	1	-	523351	524493	255,0,0	2	36,14	0,1128
chr1	765211	766342	JUNC00000010	1	-	765211	766342	255,0,0	2	36,14	0,1117
chr1	805866	808598	JUNC00000011	1	-	805866	808598	255,0,0	2	25,25	0,2707
chr1	945098	945566	JUNC00000012	8	-	945098	945566	255,0,0	2	48,49	0,419
chr1	946261	946426	JUNC00000013	1	-	946261	946426	255,0,0	2	25,25	0,140
chr1	946527	948172	JUNC00000014	2	-	946527	948172	255,0,0	2	18,42	0,1603
chr1	952095	952449	JUNC00000015	2	-	952095	952449	255,0,0	2	44,38	0,316
chr1	953857	954034	JUNC00000016	3	-	953857	954034	255,0,0	2	35,31	0,146
chr1	963227	963360	JUNC00000017	1	+	963227	963360	255,0,0	2	26,24	0,109
chr1	1047655	1047810	JUNC00000018	2	+	1047655	1047810	255,0,0	2	32,35	0,120
chr1	1049017	1049254	JUNC00000019	3	+	1049017	1049254	255,0,0	2	42,19	0,218
chr1	1050011	1050256	JUNC00000020	1	+	1050011	1050256	255,0,0	2	26,24	0,221
chr1	1051616	1051748	JUNC00000021	1	+	1051616	1051748	255,0,0	2	29,21	0,111
chr1	1051790	1053777	JUNC00000022	1	+	1051790	1053777	255,0,0	2	25,25	0,1962
chr1	1054512	1054850	JUNC00000023	3	+	1054512	1054850	255,0,0	2	39,27	0,311



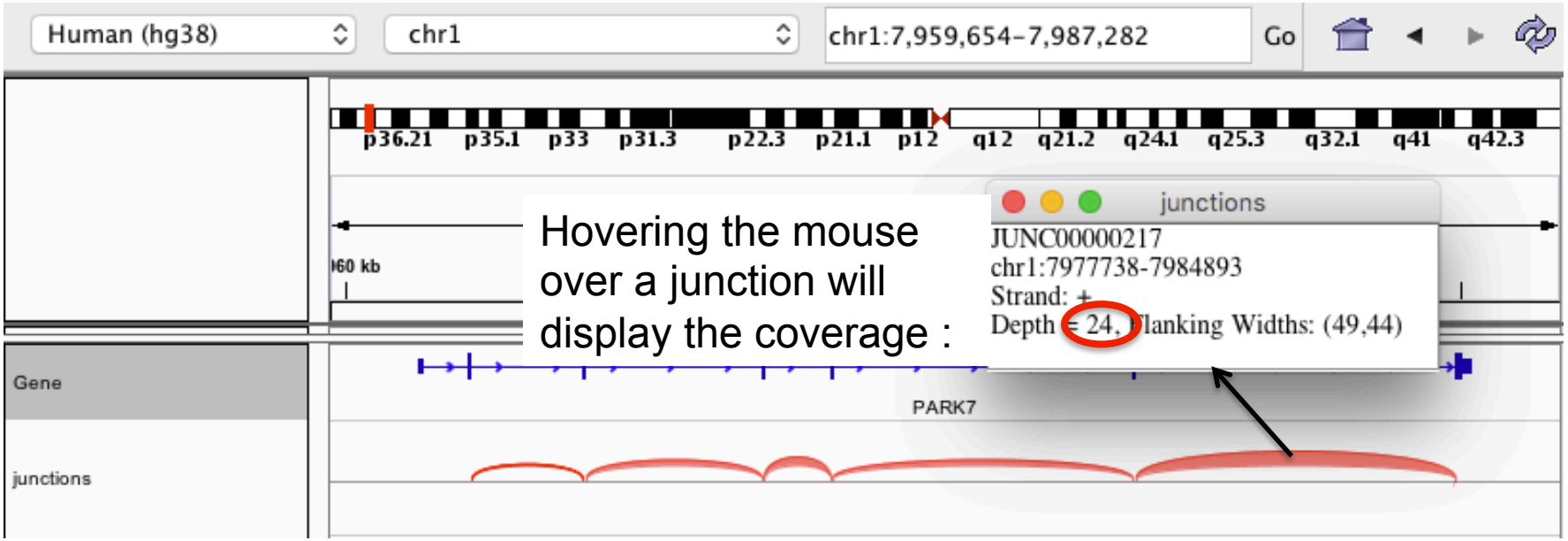
- IGV
  - Select the appropriate genome assembly (hg38)
  - File → Load from file and choose the downloaded BED file

Exercise 1
2.3. Splice junctions visualization



Exercise 1
2.3. Splice junctions visualization

- IGV



- BED file on Galaxy

chr1	7977689	7984937	JUNC00000217	24	+	7977689	7984937	255,0,0	2	49,44	0,7204

7977689+49 = 7977738 : junction start position
7984937-44 = 7984893 : junction end position

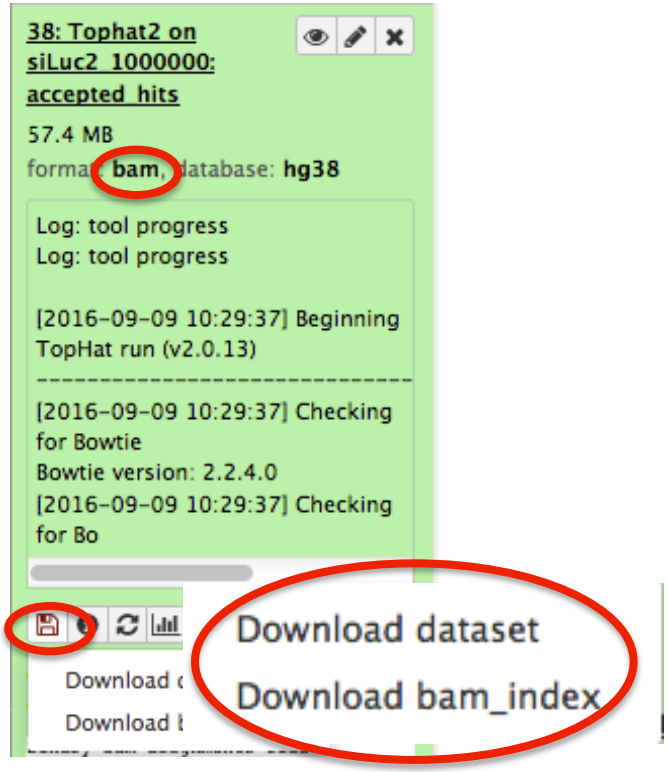
→ 24 alignments span the junction that joins the last 2 exons of *Park7* gene

Exercise 1
3. Alignment visualization

- Galaxy

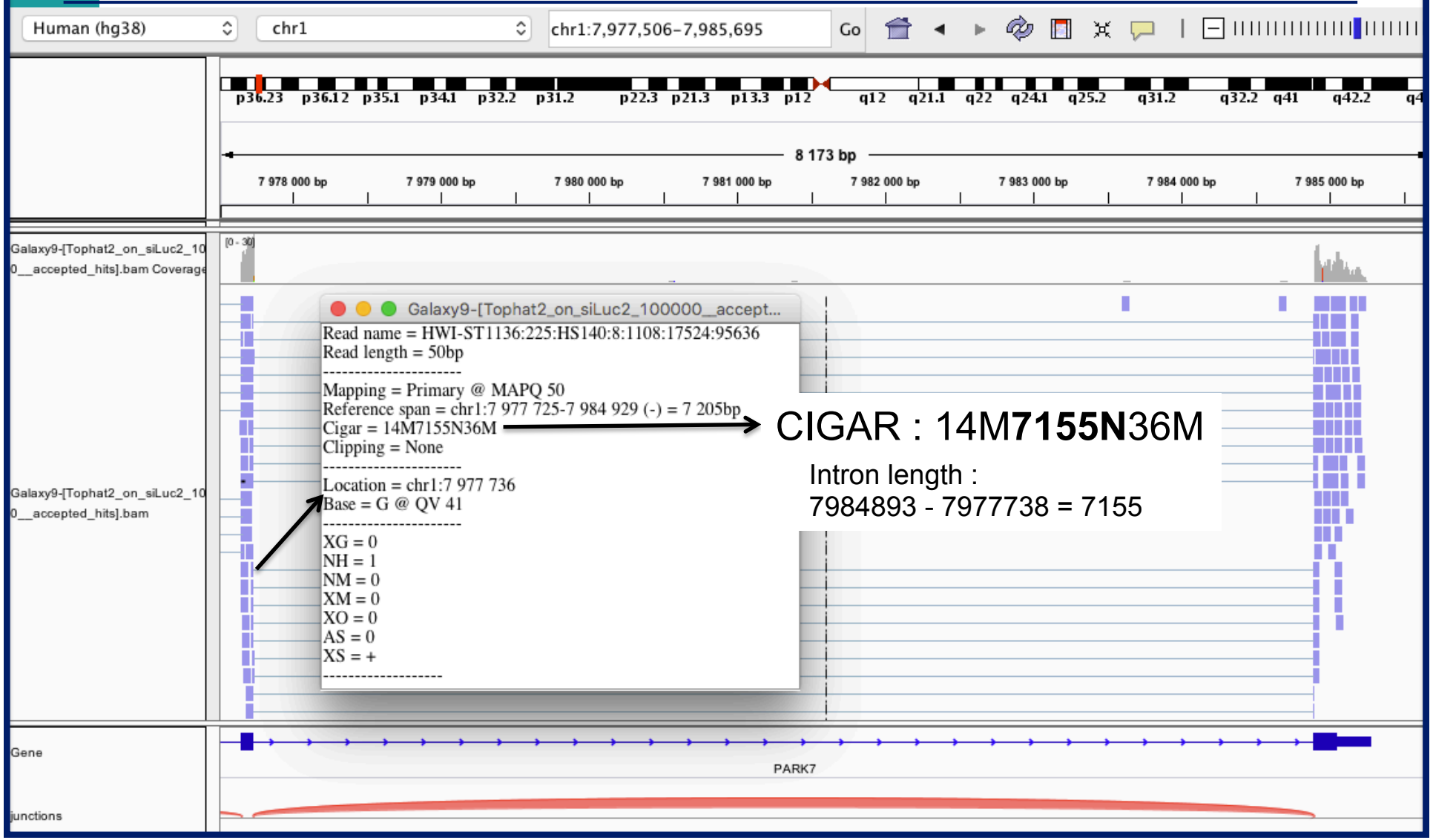
  3.1. Tophat2 provides an alignment in BAM format

  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
3.3. Reads aligned on a splice junction

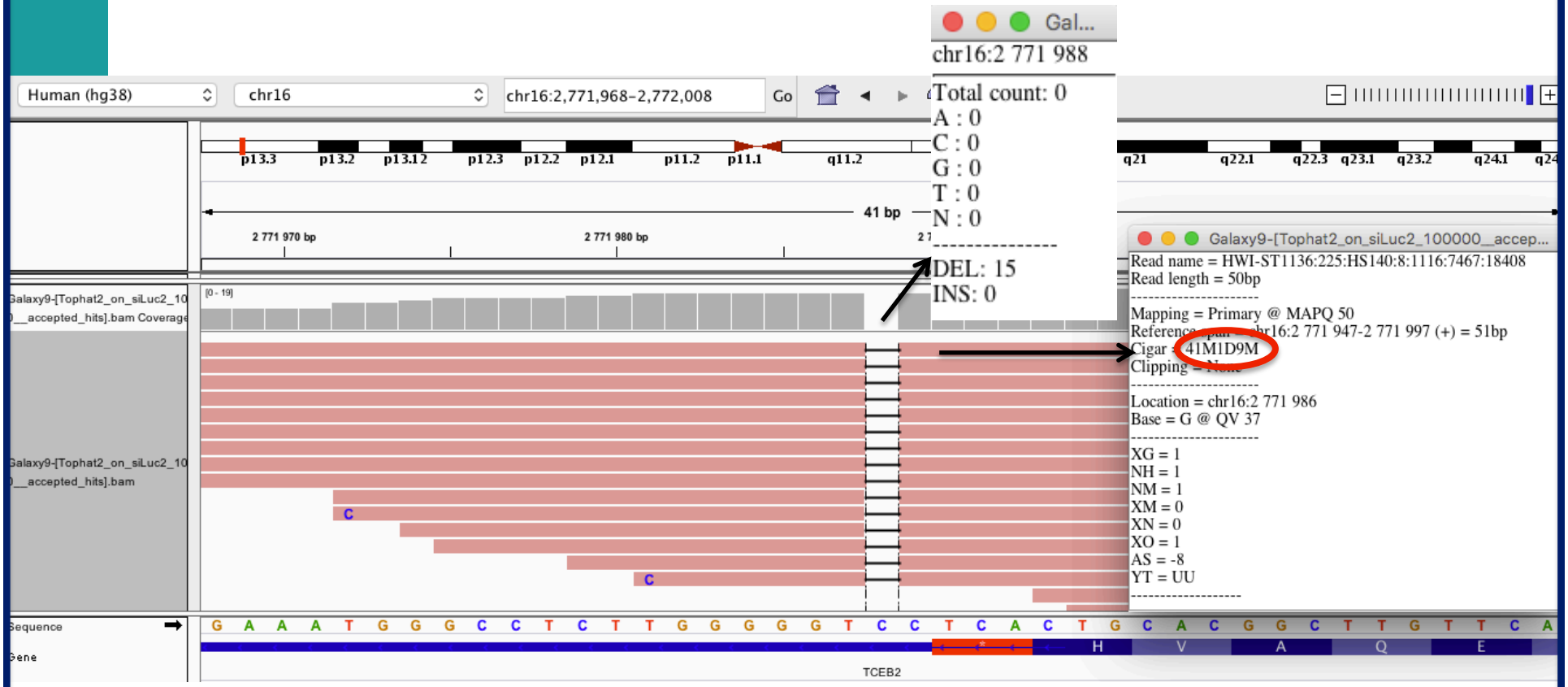


Exercise 1
3.4. Visualization of strand specificity

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
3.5. Visualization of a deletion



15 reads aligned with a deletion at position chr16:2,771,988

Exercise 1
3.6. Visualization of multiple mapped reads

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

Number of reported alignments : 1 2 3 4

There are multiple aligned reads on this gene

Exercise 2 - Question 1
Proportion of mapped reads in all samples

Galaxy : Shared Data → Data Libraries → CNRS training
RNAseq → alignment → align_summary :

Name Tophat2 on siLuc2: align_summary

```
Reads:
Input     :  43672265
Mapped    :  42797297 (98.0% of input)
of these:   5829092 (13.6%) have multiple alignments (1132 have >20)
98.0% overall read mapping rate.
```

Name Tophat2 on siLuc3: align_summary

```
Reads:
Input     :  46565834
Mapped    :  45633110 (98.0% of input)
of these:   6030755 (13.2%) have multiple alignments (861 have >20)
98.0% overall read mapping rate.
```

Name Tophat2 on siMitf3: align_summary

```
Reads:
Input     :  43985979
Mapped    :  43048694 (97.9% of input)
of these:   5763991 (13.4%) have multiple alignments (765 have >20)
97.9% overall read mapping rate.
```

Name Tophat2 on siMitf4: align_summary

```
Reads:
Input     :  51348313
Mapped    :  50317655 (98.0% of input)
of these:   6826164 (13.6%) have multiple alignments (643 have >20)
98.0% overall read mapping rate.
```

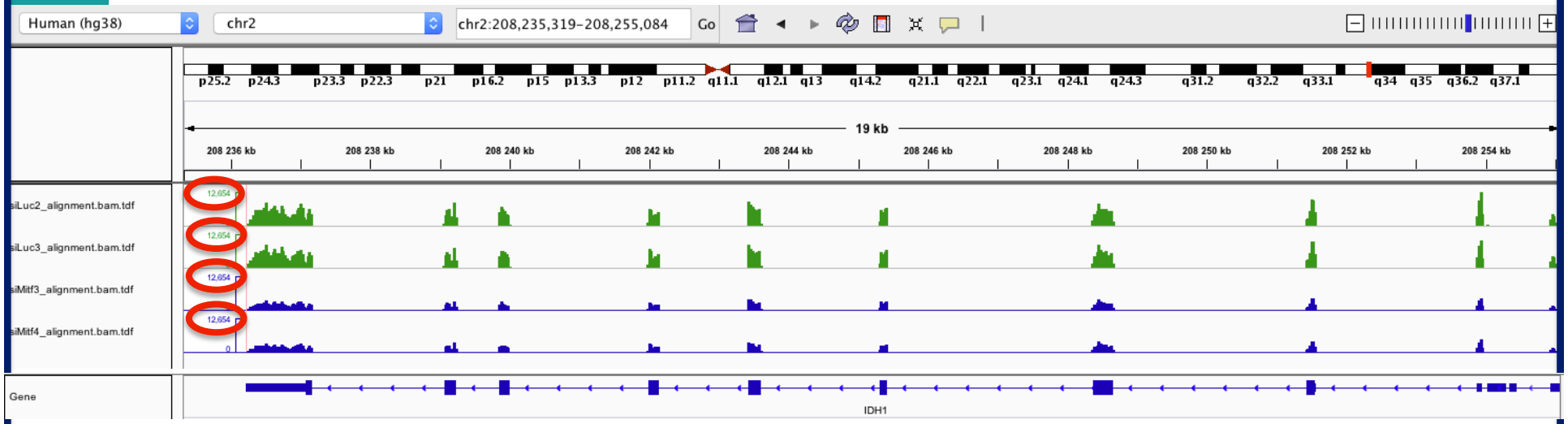
→ This proportion is high and consistent across samples

Exercise 2 – Question 2

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



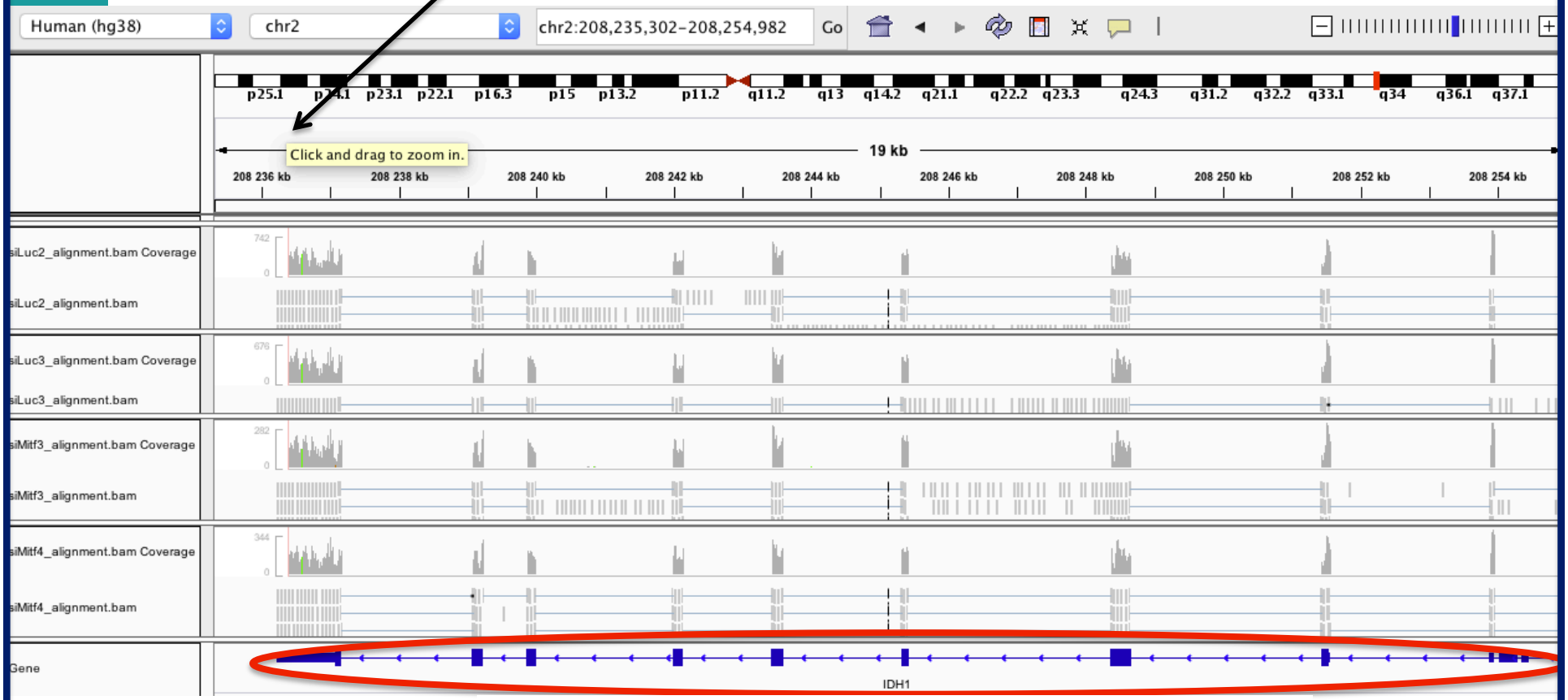
Idh1 is under-expressed in siMitf samples compared to siLuc ones

Exercise 2 – Question 3
Alignments visualization

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

Exercise 2 - Question 3

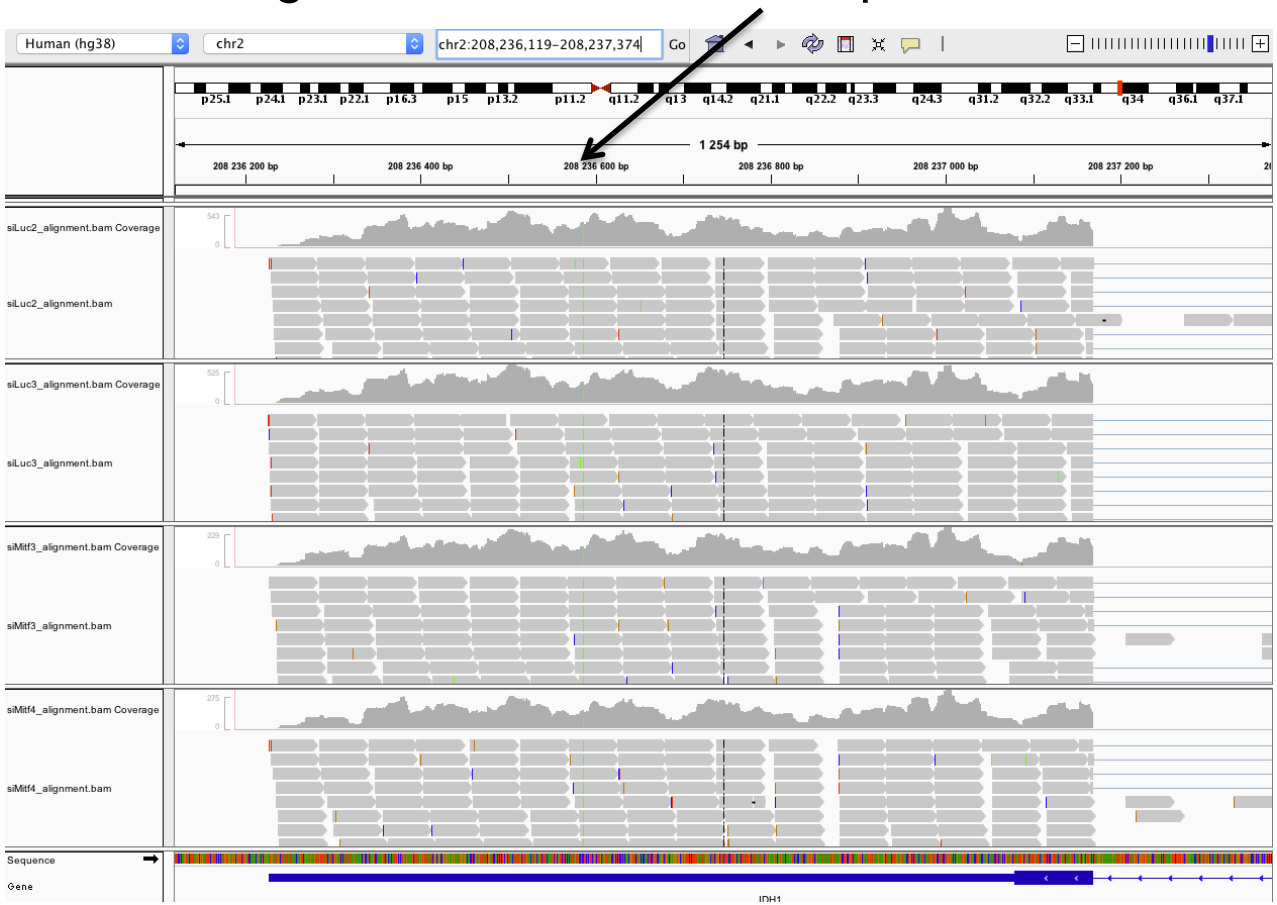
Click and drag to define a window around the last exon to zoom in



Arrows indicate annotated transcribed strand

Exercise 2 - Question 3

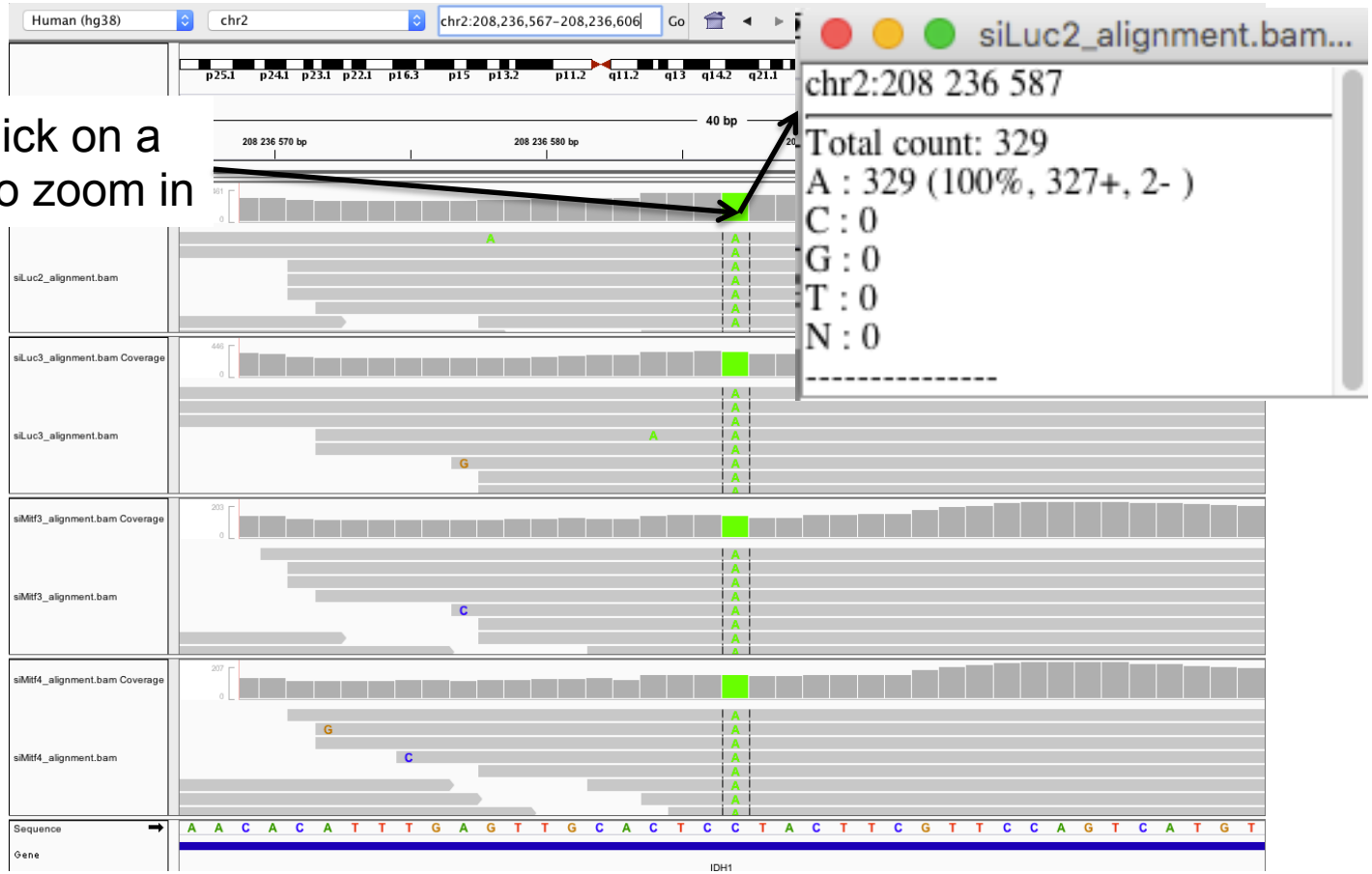
- You can see a nucleotide difference in green
- Click and drag to zoom in around this position



Exercise 2 - Question 3

- In the location chr2:208,236,587 :
  - A in 100% of the RNA-seq reads, C in the reference genome

Double click on a position to zoom in



Exercise 2 – Question 4

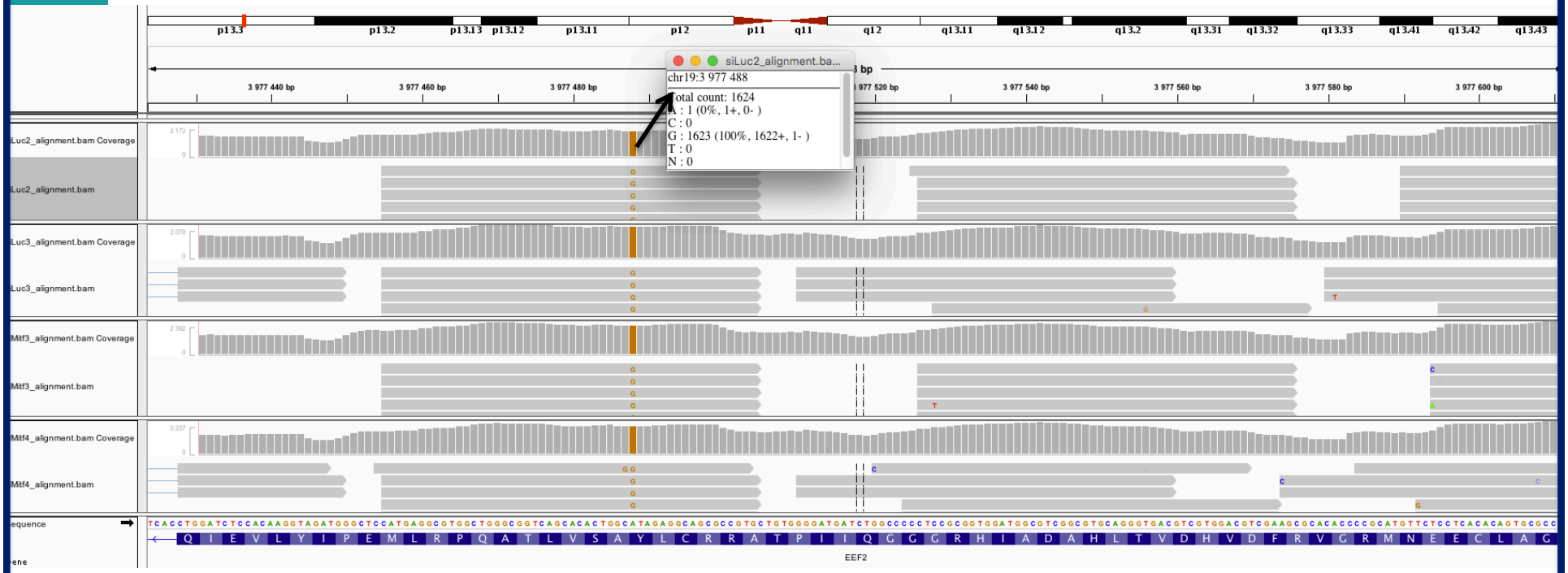
Exon numbers are provided on annotation track

Exercise 2 – Question 4

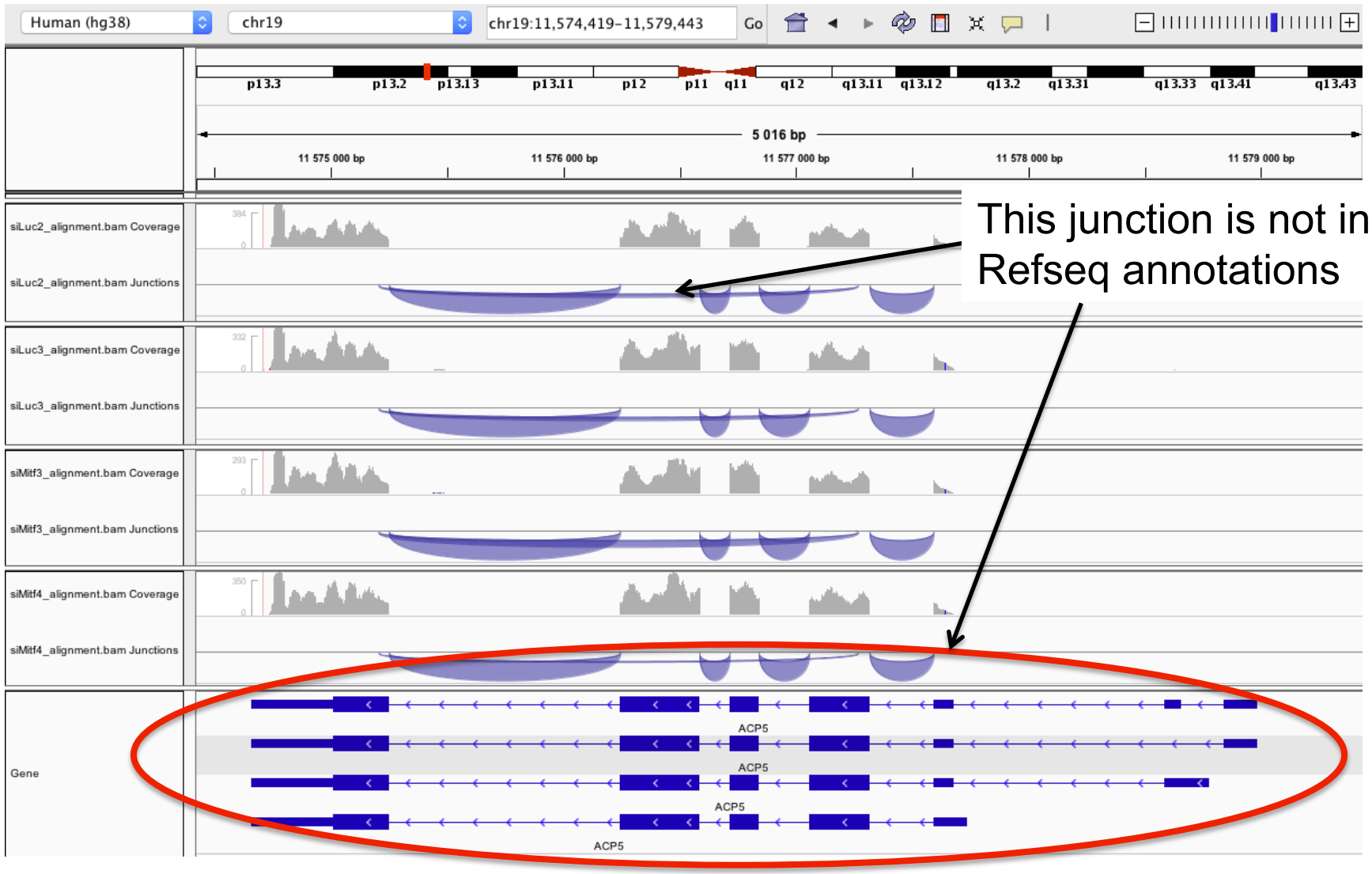
- *Eef2* exon 11

Exercise 2 – Question 4

- *Eef2* exon 13

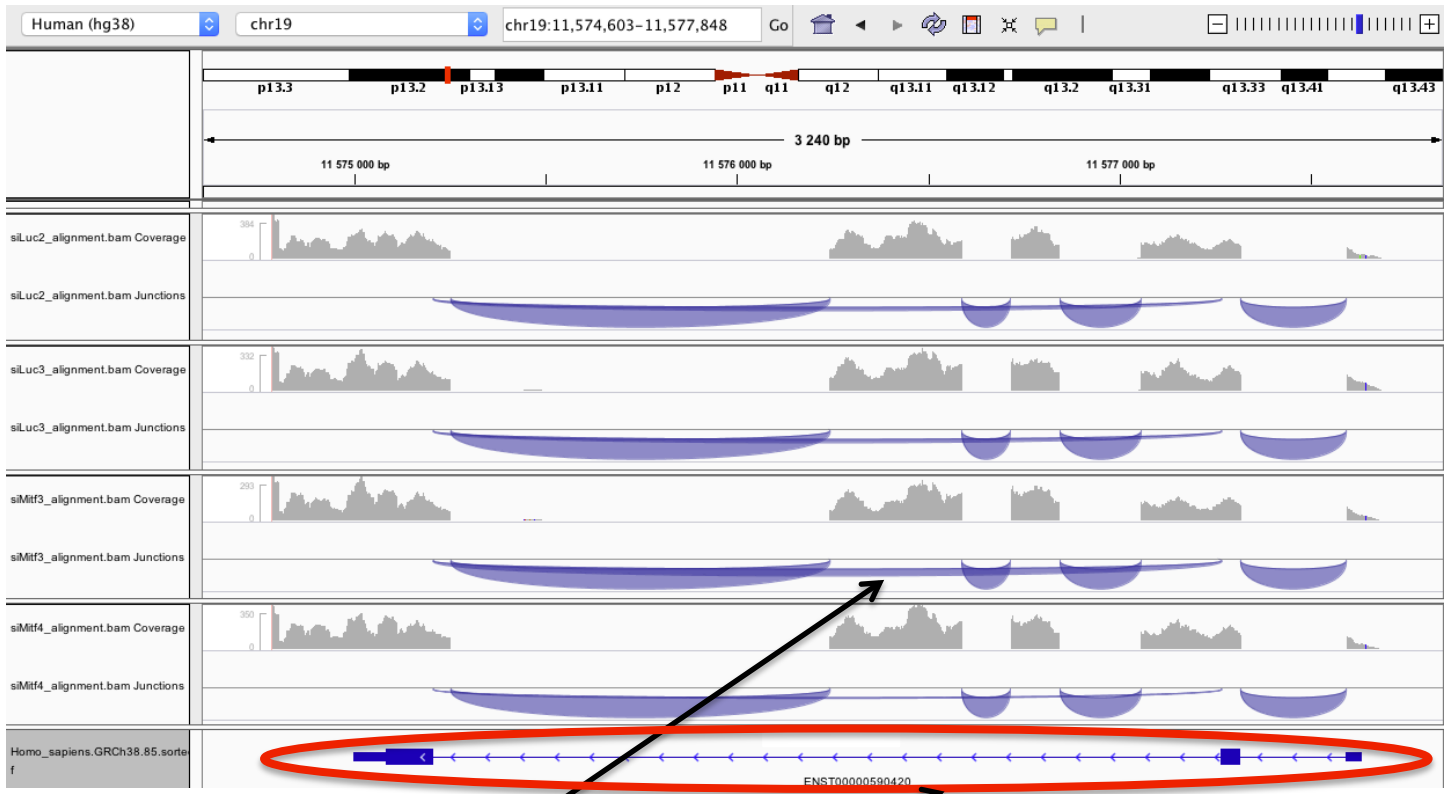


Exercise 2 – Question 5



Exercise 2 – Question 5

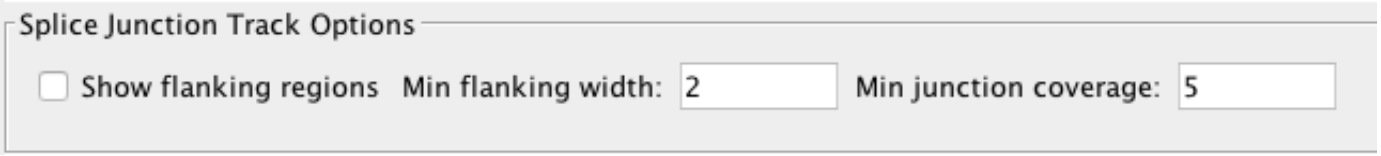
- File → Load from file and select Ensembl annotations (Homo_sapiens.GRCh38.85.sorted.gtf)
- Right click on Ensembl annotations track and select Expanded



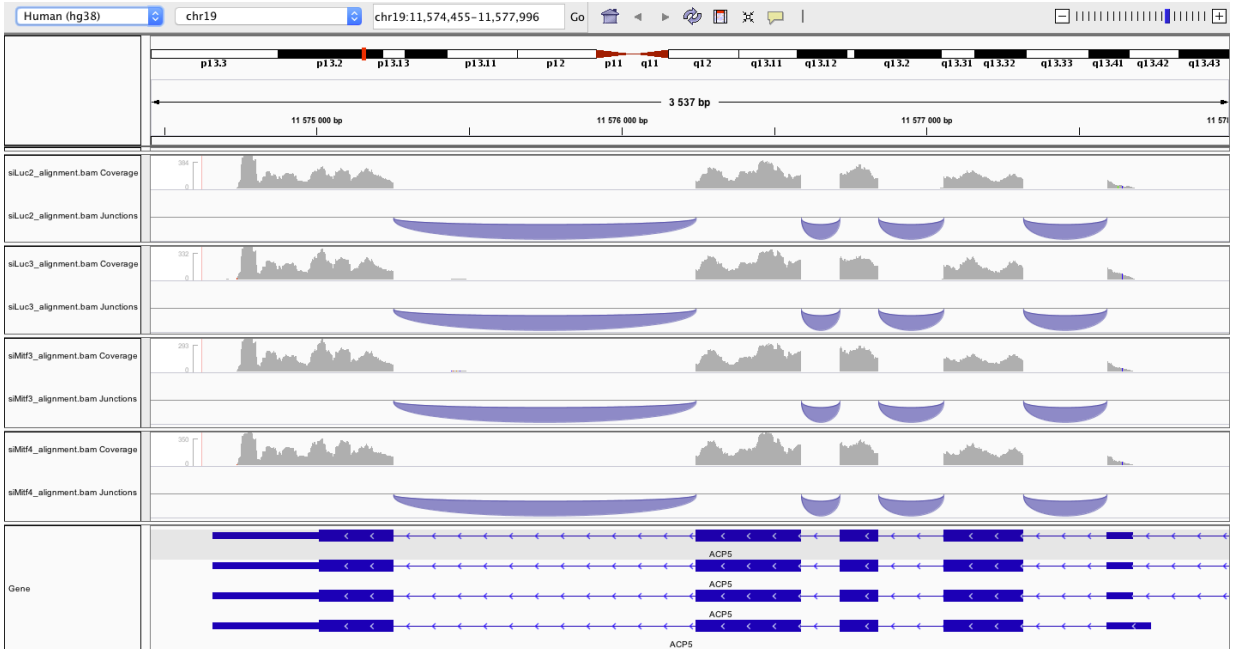
This junction is in Ensembl annotations

Exercise 2 – Question 5

- To modify the display of splice junctions :
  - View → Preferences → Alignments

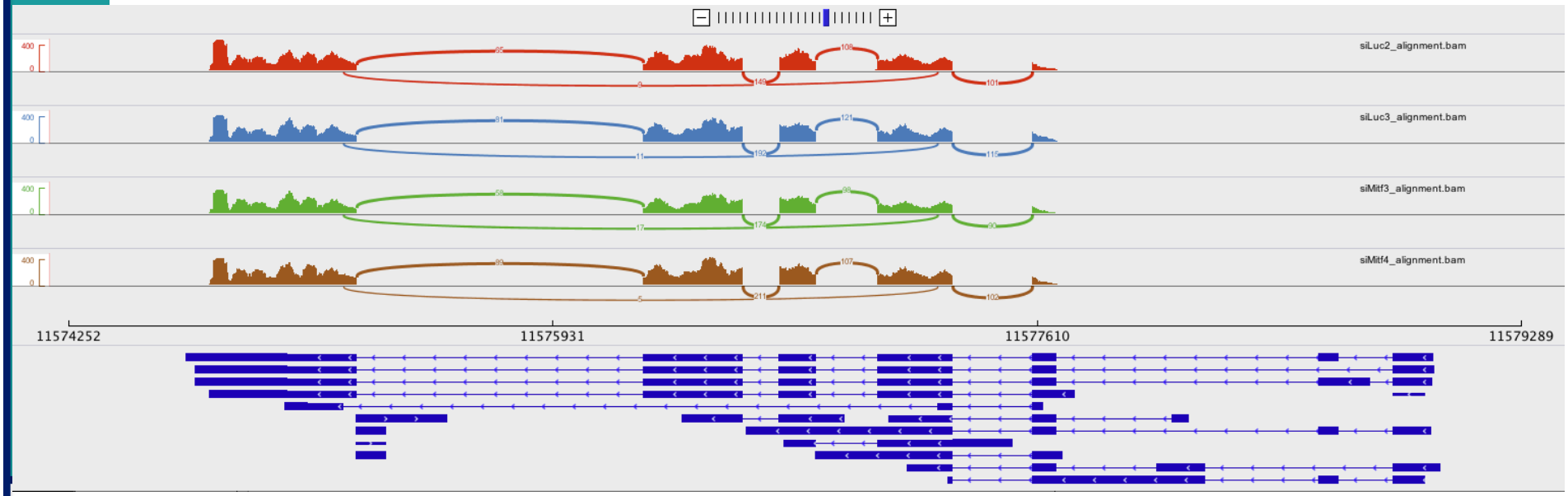


  - Example with a minimum junction coverage of 20



Exercise 2 – Question 5

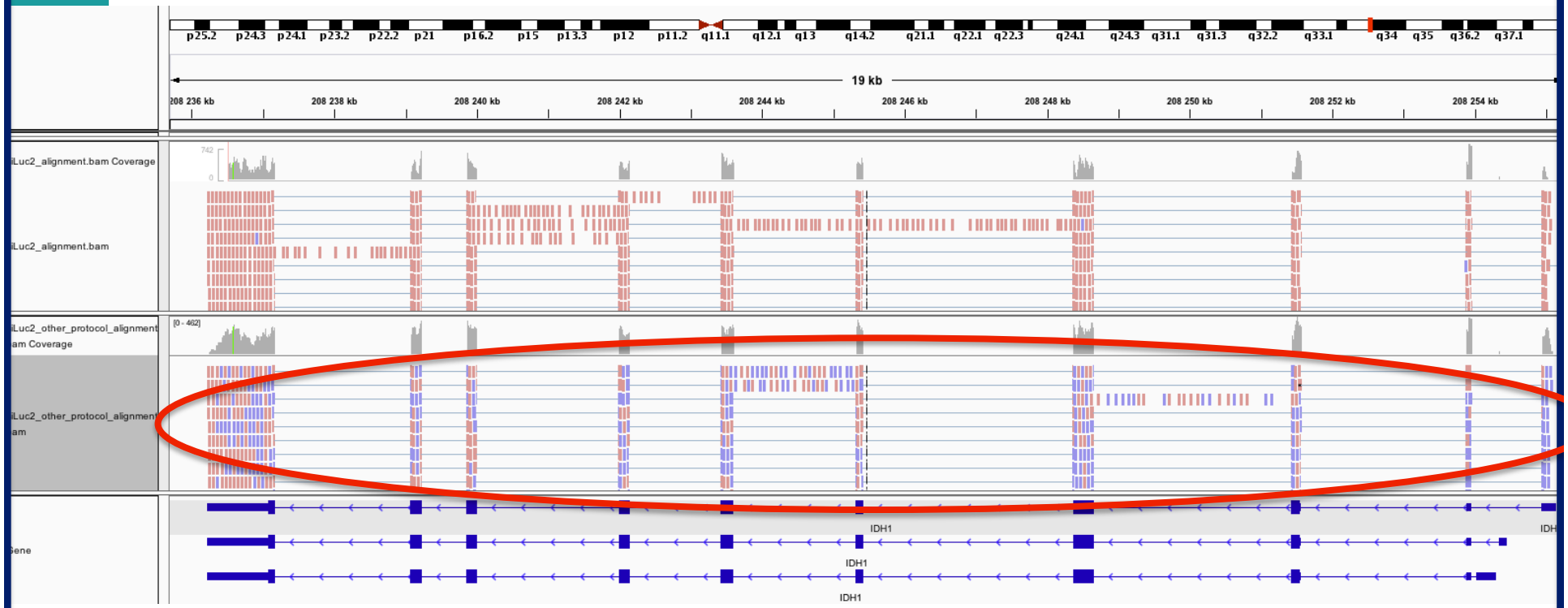
- Sashimi plot



➔ Very useful to quickly screen differentially spliced exons along genomic regions of interest (more accurate with paired-end data)

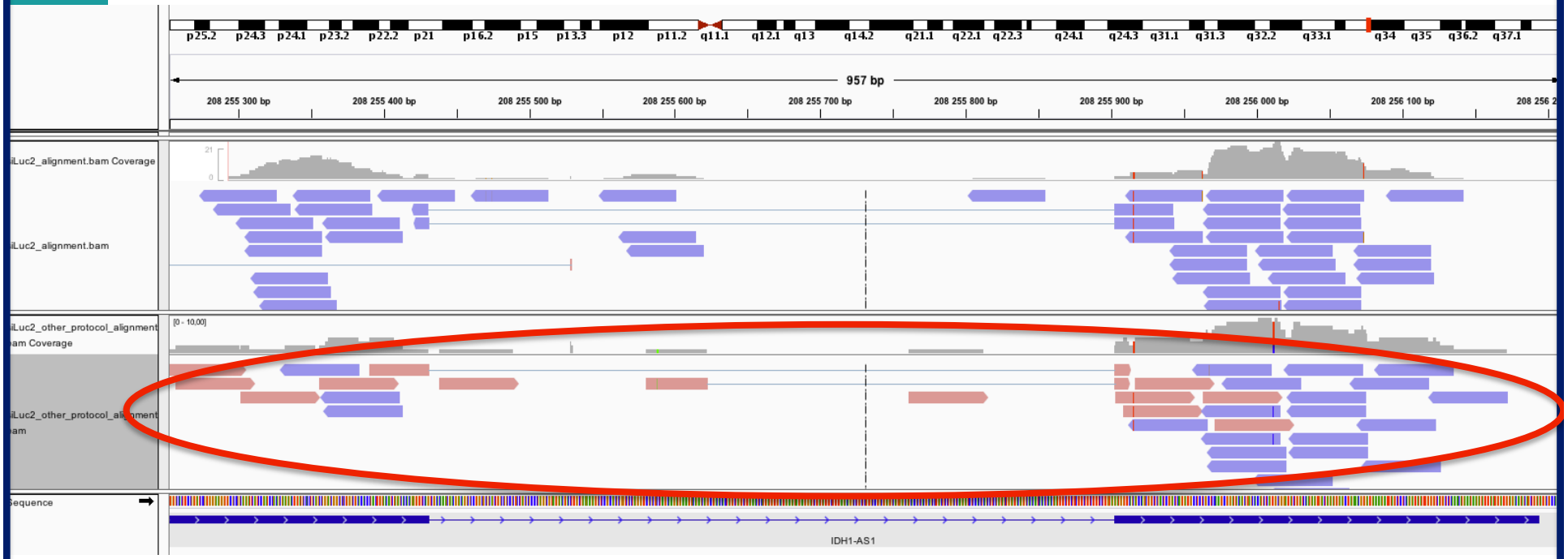
Exercise 2 – Question 6

- File → load from file and select siLuc2_other_protocol_alignment.bam
- Right-click on BAM file → Color alignments by → read strand
- e.g. *Idh1* gene



Exercise 2 – Question 6

- e.g. *Idh1-as1* gene



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