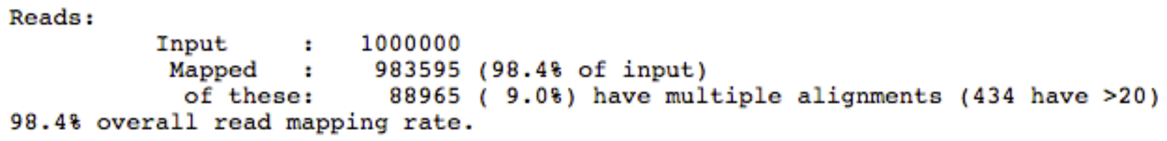
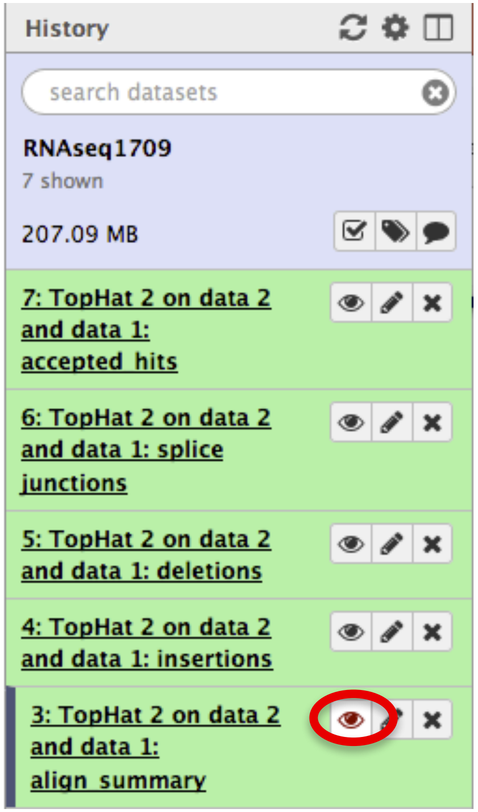
NGS read mapping : answers to questions

Céline Keime
keime@igbmc.fr

Exercise 1
1. Alignment summary statistics



```
Reads:
          Input     :   1000000
           Mapped   :    983595 (98.4% of input)
            of these:     88965 ( 9.0%) have multiple alignments (434 have >20)
98.4% overall read mapping rate.
```



History

search datasets

RNAseq1709

7 shown

207.09 MB

7: TopHat 2 on data 2 and data 1: accepted_hits

6: TopHat 2 on data 2 and data 1: splice junctions

5: TopHat 2 on data 2 and data 1: deletions

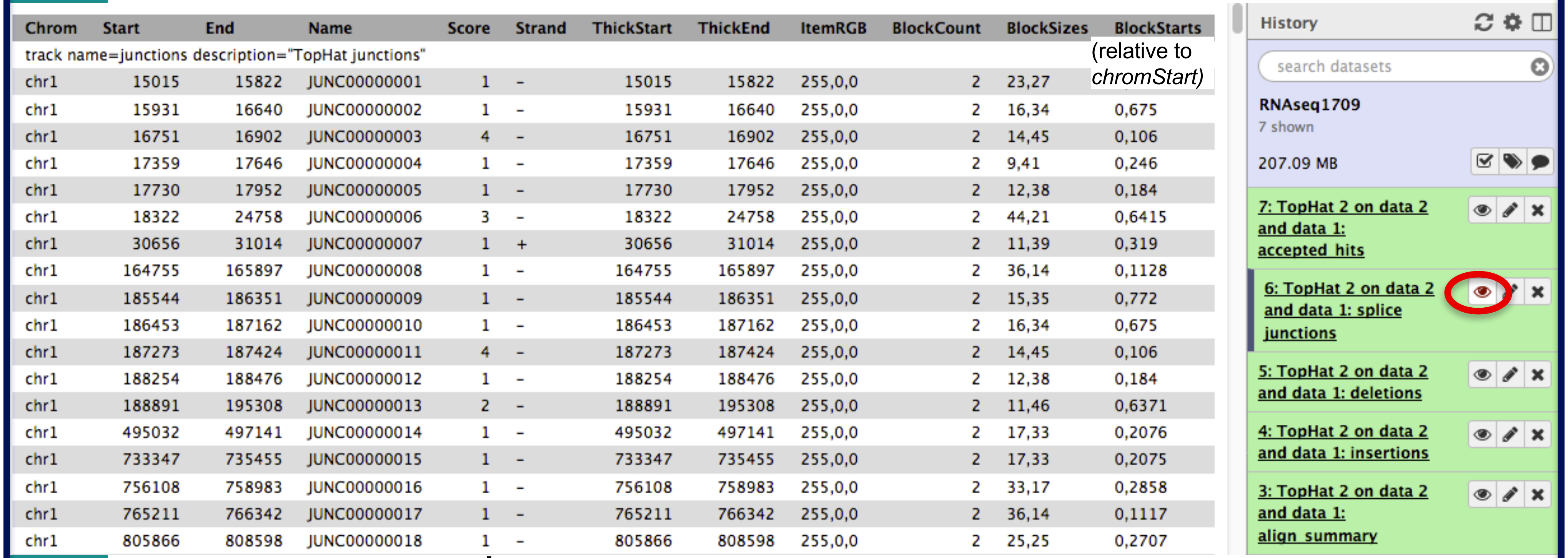
4: TopHat 2 on data 2 and data 1: insertions

3: TopHat 2 on data 2 and data 1: align_summary

1.1. 983,595 reads mapped onto hg38

1.2. 9% of these reads have multiple alignments

Exercise 1
2. Splice junctions



Chrom	Start	End	Name	Score	Strand	ThickStart	ThickEnd	ItemRGB	BlockCount	BlockSizes	BlockStarts
track name=junctions description="TopHat junctions"											
chr1	15015	15822	JUNC00000001	1	-	15015	15822	255,0,0	2	23,27	
chr1	15931	16640	JUNC00000002	1	-	15931	16640	255,0,0	2	16,34	0,675
chr1	16751	16902	JUNC00000003	4	-	16751	16902	255,0,0	2	14,45	0,106
chr1	17359	17646	JUNC00000004	1	-	17359	17646	255,0,0	2	9,41	0,246
chr1	17730	17952	JUNC00000005	1	-	17730	17952	255,0,0	2	12,38	0,184
chr1	18322	24758	JUNC00000006	3	-	18322	24758	255,0,0	2	44,21	0,6415
chr1	30656	31014	JUNC00000007	1	+	30656	31014	255,0,0	2	11,39	0,319
chr1	164755	165897	JUNC00000008	1	-	164755	165897	255,0,0	2	36,14	0,1128
chr1	185544	186351	JUNC00000009	1	-	185544	186351	255,0,0	2	15,35	0,772
chr1	186453	187162	JUNC00000010	1	-	186453	187162	255,0,0	2	16,34	0,675
chr1	187273	187424	JUNC00000011	4	-	187273	187424	255,0,0	2	14,45	0,106
chr1	188254	188476	JUNC00000012	1	-	188254	188476	255,0,0	2	12,38	0,184
chr1	188891	195308	JUNC00000013	2	-	188891	195308	255,0,0	2	11,46	0,6371
chr1	495032	497141	JUNC00000014	1	-	495032	497141	255,0,0	2	17,33	0,2076
chr1	733347	735455	JUNC00000015	1	-	733347	735455	255,0,0	2	17,33	0,2075
chr1	756108	758983	JUNC00000016	1	-	756108	758983	255,0,0	2	33,17	0,2858
chr1	765211	766342	JUNC00000017	1	-	765211	766342	255,0,0	2	36,14	0,1117
chr1	805866	808598	JUNC00000018	1	-	805866	808598	255,0,0	2	25,25	0,2707

(relative to *chromStart)*

History

search datasets

RNAseq1709

7 shown

207.09 MB

7: TopHat 2 on data 2 and data 1: accepted_hits

6: TopHat 2 on data 2 and data 1: splice junctions

5: TopHat 2 on data 2 and data 1: deletions

4: TopHat 2 on data 2 and data 1: insertions

3: TopHat 2 on data 2 and data 1: align_summary

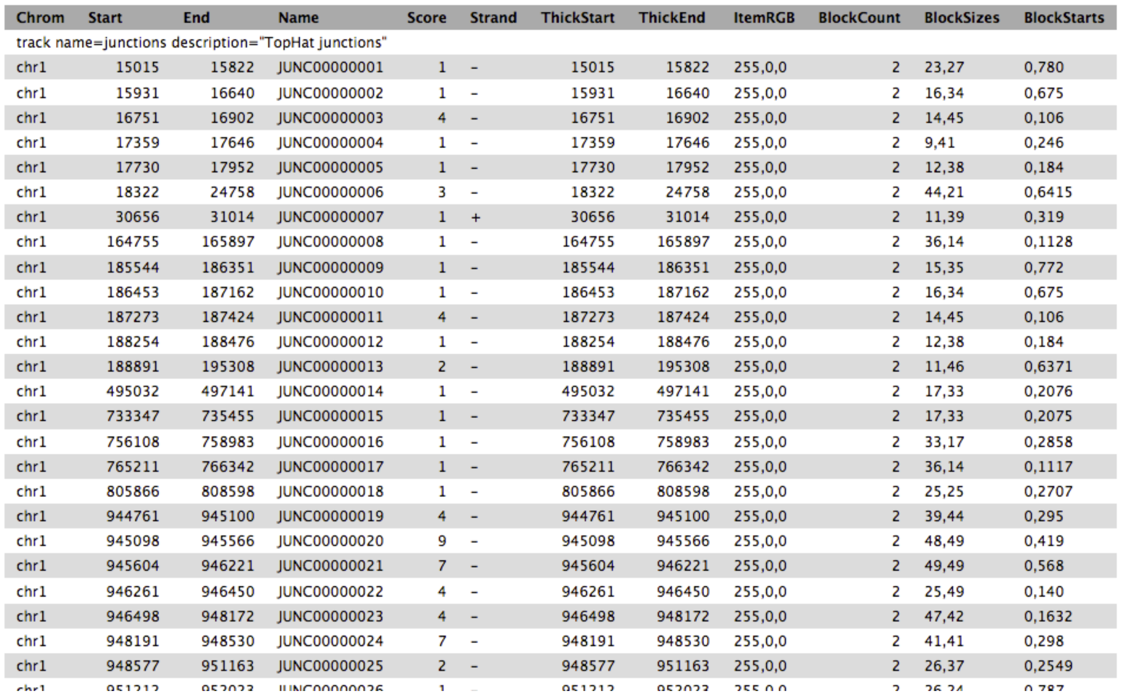
Number of alignments spanning the junction

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

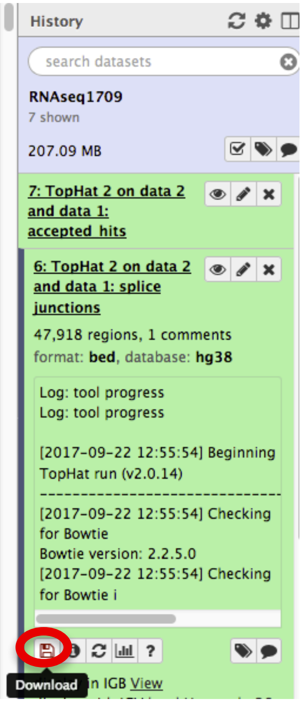
2.1. Splice junctions provided in a BED file

Exercise 1
2.2. Splice junctions visualization

- Galaxy
  - Download splice junctions BED file



Chrom	Start	End	Name	Score	Strand	ThickStart	ThickEnd	ItemRGB	BlockCount	BlockSizes	BlockStarts
track name=junctions description="TopHat junctions"											
chr1	15015	15822	JUNC00000001	1	-	15015	15822	255,0,0	2	23,27	0,780
chr1	15931	16640	JUNC00000002	1	-	15931	16640	255,0,0	2	16,34	0,675
chr1	16751	16902	JUNC00000003	4	-	16751	16902	255,0,0	2	14,45	0,106
chr1	17359	17646	JUNC00000004	1	-	17359	17646	255,0,0	2	9,41	0,246
chr1	17730	17952	JUNC00000005	1	-	17730	17952	255,0,0	2	12,38	0,184
chr1	18322	24758	JUNC00000006	3	-	18322	24758	255,0,0	2	44,21	0,6415
chr1	30656	31014	JUNC00000007	1	+	30656	31014	255,0,0	2	11,39	0,319
chr1	164755	165897	JUNC00000008	1	-	164755	165897	255,0,0	2	36,14	0,1128
chr1	185544	186351	JUNC00000009	1	-	185544	186351	255,0,0	2	15,35	0,772
chr1	186453	187162	JUNC00000010	1	-	186453	187162	255,0,0	2	16,34	0,675
chr1	187273	187424	JUNC00000011	4	-	187273	187424	255,0,0	2	14,45	0,106
chr1	188254	188476	JUNC00000012	1	-	188254	188476	255,0,0	2	12,38	0,184
chr1	188891	195308	JUNC00000013	2	-	188891	195308	255,0,0	2	11,46	0,6371
chr1	495032	497141	JUNC00000014	1	-	495032	497141	255,0,0	2	17,33	0,2076
chr1	733347	735455	JUNC00000015	1	-	733347	735455	255,0,0	2	17,33	0,2075
chr1	756108	758983	JUNC00000016	1	-	756108	758983	255,0,0	2	33,17	0,2858
chr1	765211	766342	JUNC00000017	1	-	765211	766342	255,0,0	2	36,14	0,1117
chr1	805866	808598	JUNC00000018	1	-	805866	808598	255,0,0	2	25,25	0,2707
chr1	944761	945100	JUNC00000019	4	-	944761	945100	255,0,0	2	39,44	0,295
chr1	945098	945566	JUNC00000020	9	-	945098	945566	255,0,0	2	48,49	0,419
chr1	945604	946221	JUNC00000021	7	-	945604	946221	255,0,0	2	49,49	0,568
chr1	946261	946450	JUNC00000022	4	-	946261	946450	255,0,0	2	25,49	0,140
chr1	946498	948172	JUNC00000023	4	-	946498	948172	255,0,0	2	47,42	0,1632
chr1	948191	948530	JUNC00000024	7	-	948191	948530	255,0,0	2	41,41	0,298
chr1	948577	951163	JUNC00000025	2	-	948577	951163	255,0,0	2	26,37	0,2549
chr1	951212	952023	JUNC00000026	1	-	951212	952023	255,0,0	2	26,24	0,787



RNAseq1709
7 shown
207.09 MB

7: TopHat 2 on data 2 and data 1: accepted_hits

6: TopHat 2 on data 2 and data 1: splice junctions

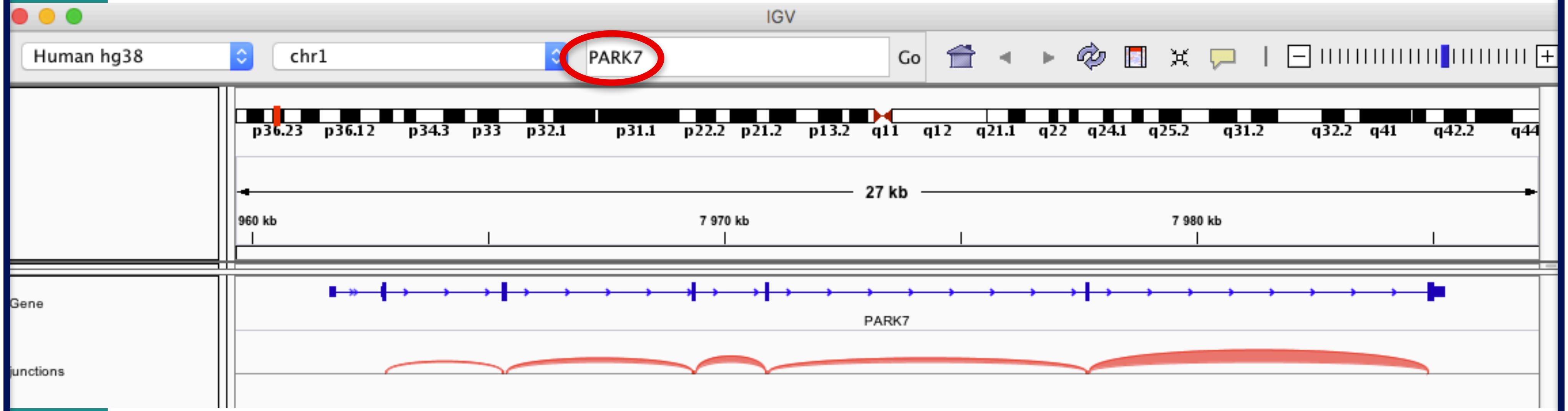
47,918 regions, 1 comments
format: **bed**, database: **hg38**

```
Log: tool progress
Log: tool progress

[2017-09-22 12:55:54] Beginning
TopHat run (v2.0.14)
-----------------------------------
[2017-09-22 12:55:54] Checking
for Bowtie
Bowtie version: 2.2.5.0
[2017-09-22 12:55:54] Checking
for Bowtie i
```

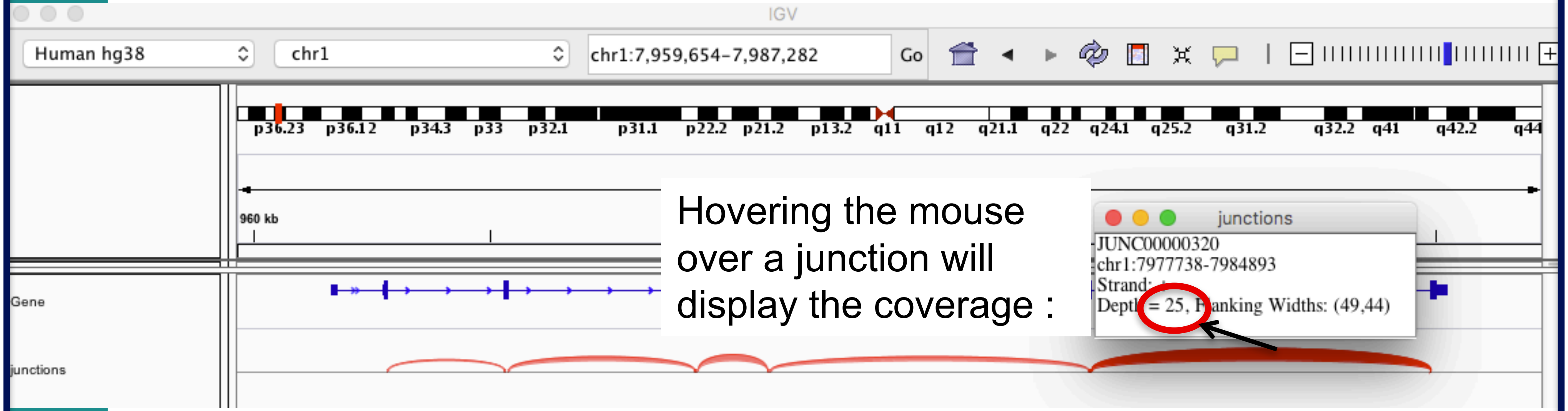
- IGV
  - Select the appropriate genome assembly (hg38)
  - File → Load from file and choose the downloaded BED file

Exercise 1
2.3. Splice junctions visualization



Exercise 1
2.3. Splice junctions visualization

- IGV



chr1:7,959,654-7,987,282

Hovering the mouse over a junction will display the coverage :

junctions

JUNC00000320
chr1:7977738-7984893
Strand: +
Depth = 25, Flanking Widths: (49,44)

- BED file on Galaxy

```
chr1     7977689    7984937    JUNC00000320    25    +        7977689    7984937    255,0,0          2    49,44         0,7204
```

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

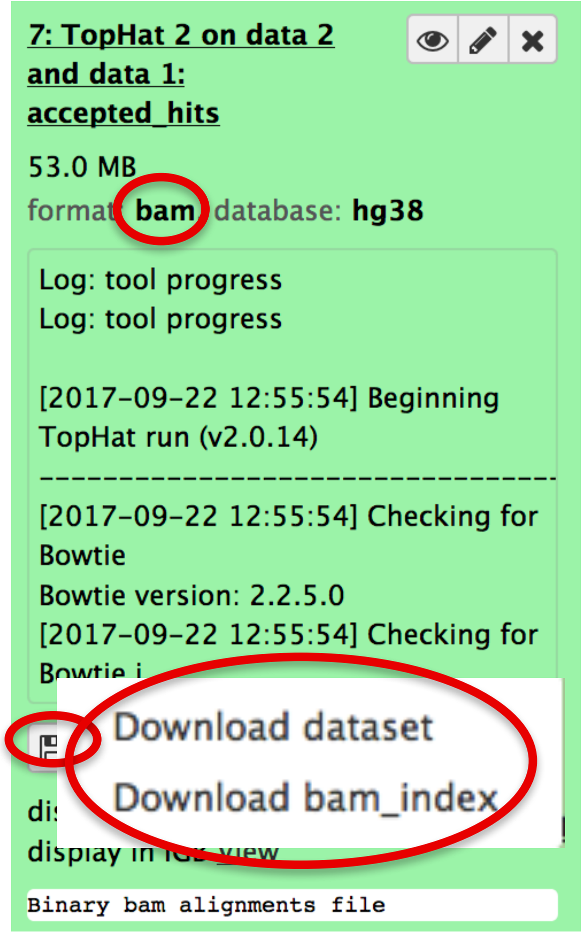
→ 25 alignments span the junction that joins the last 2 exons of *Park7* gene

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)



7: TopHat 2 on data 2 and data 1: accepted_hits

53.0 MB

format: bam, database: hg38

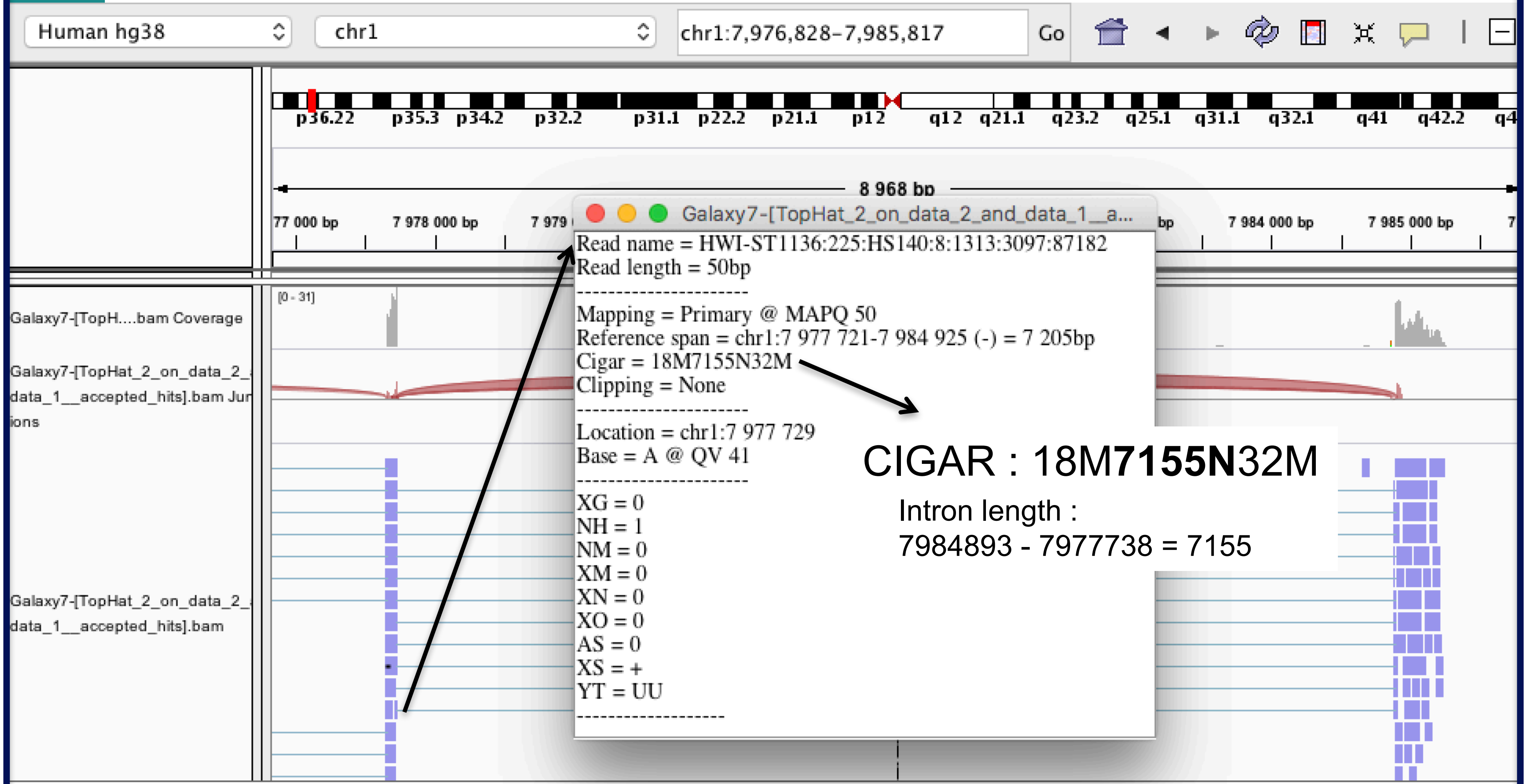
```
Log: tool progress
Log: tool progress

[2017-09-22 12:55:54] Beginning
TopHat run (v2.0.14)
-----------------------------------
[2017-09-22 12:55:54] Checking for
Bowtie
Bowtie version: 2.2.5.0
[2017-09-22 12:55:54] Checking for
Bowtie i
```

Binary bam alignments file

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

Exercise 1
3.3. Reads aligned on a splice junction



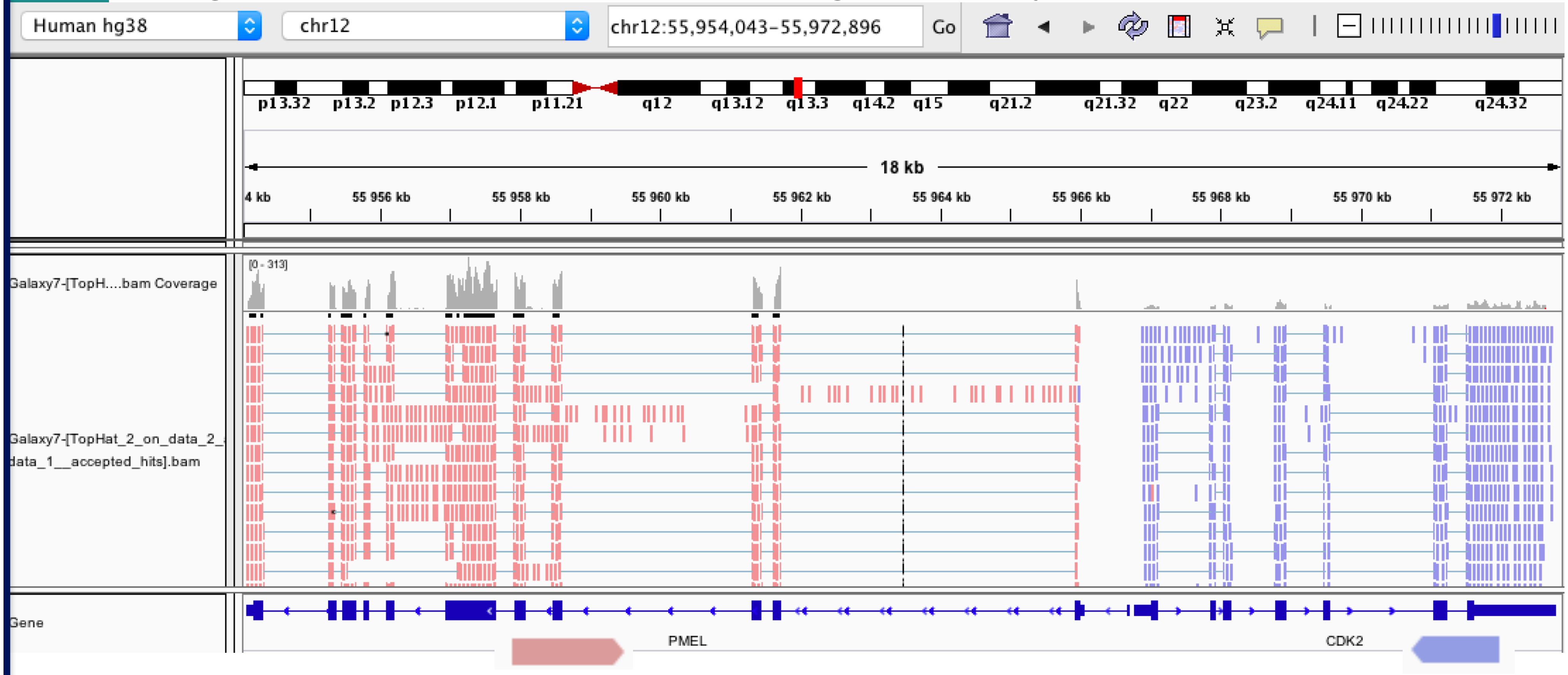
```
Galaxy7-[TopHat_2_on_data_2_and_data_1__a...
Read name = HWI-ST1136:225:HS140:8:1313:3097:87182
Read length = 50bp
----------------------
Mapping = Primary @ MAPQ 50
Reference span = chr1:7 977 721-7 984 925 (-) = 7 205bp
Cigar = 18M7155N32M
Clipping = None
----------------------
Location = chr1:7 977 729
Base = A @ QV 41
----------------------
XG = 0
NH = 1
NM = 0
XM = 0
XN = 0
XO = 0
AS = 0
XS = +
YT = UU
-------------------
```

CIGAR : 18M**7155N**32M

Intron length :
7984893 - 7977738 = 7155

Exercise 1
3.4. Visualization of strand specificity

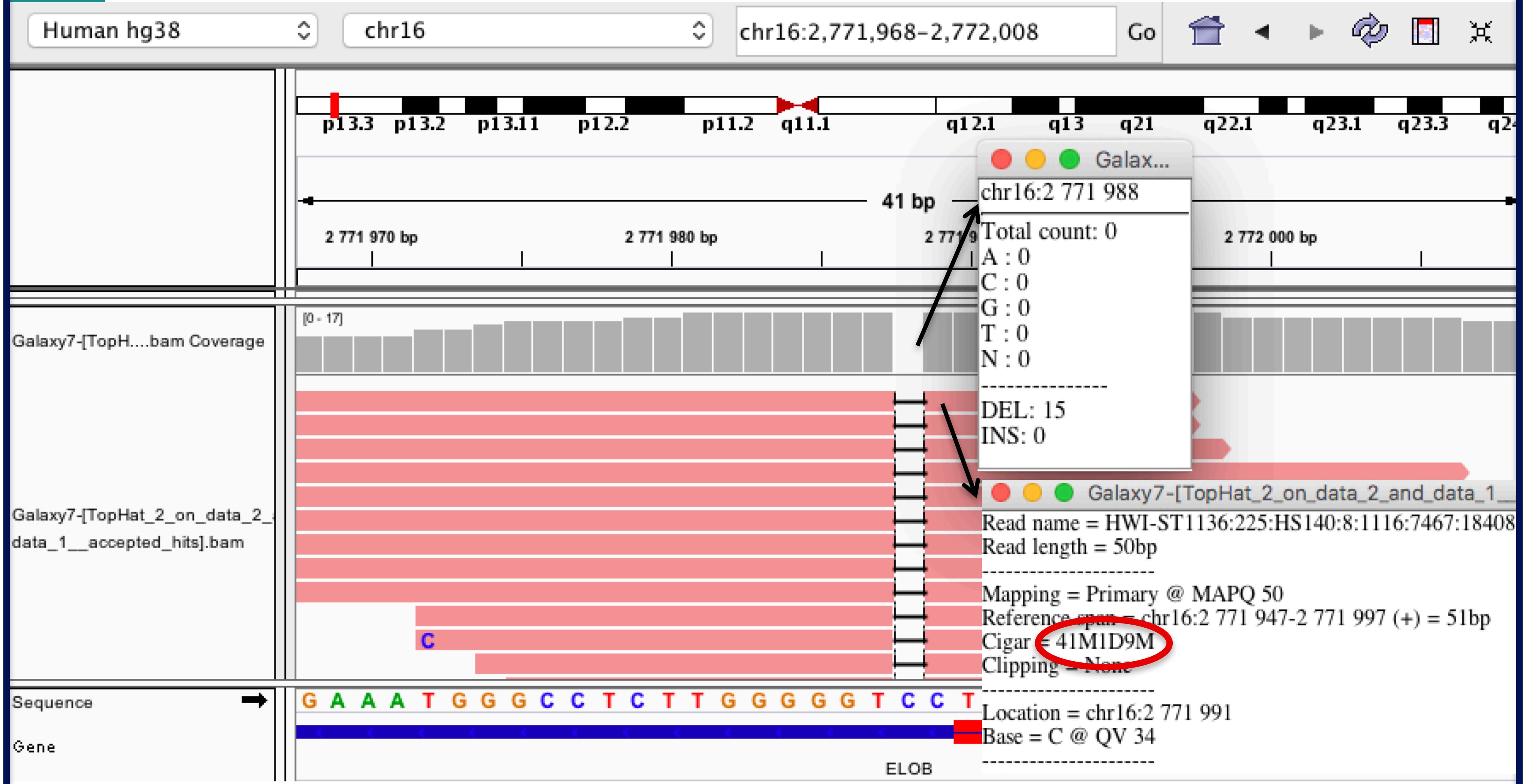
Right click on BAM file → Color alignments by → read strand



The library has been prepared with a directional mRNAseq protocol which retains strand information :

all reads are in the opposite direction compared to the transcribed strand

Exercise 1
3.5. Visualization of a deletion



Human hg38
chr16
chr16:2,771,968-2,772,008
Go
p13.3
p13.2
p13.11
p12.2
p11.2
q11.1
q12.1
q13
q21
q22.1
q23.1
q23.3
41 bp
2 771 970 bp
2 771 980 bp
2 772 000 bp
Galaxy7-[TopH....bam Coverage
[0 - 17]
Galaxy7-[TopHat_2_on_data_2_data_1__accepted_hits].bam
C
Sequence
G A A A T G G G C C T C T T G G G G T C C T
Gene
ELOB
Galax...
Galaxy7-[TopHat_2_on_data_2_and_data_1_

chr16:2 771 988

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

DEL: 15
INS: 0

Read name = HWI-ST1136:225:HS140:8:1116:7467:18408
Read length = 50bp

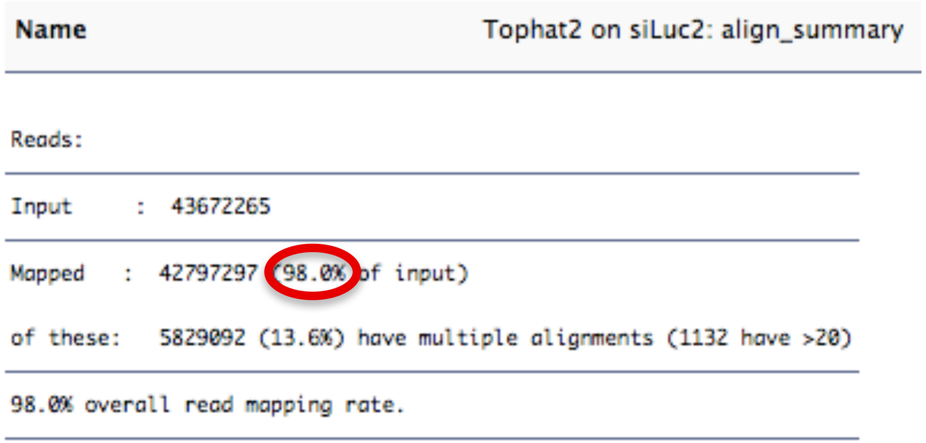
Mapping = Primary @ MAPQ 50
Reference span = chr16:2 771 947-2 771 997 (+) = 51bp
Cigar = 41M1D9M
Clipping = None

Location = chr16:2 771 991
Base = C @ QV 34
----

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

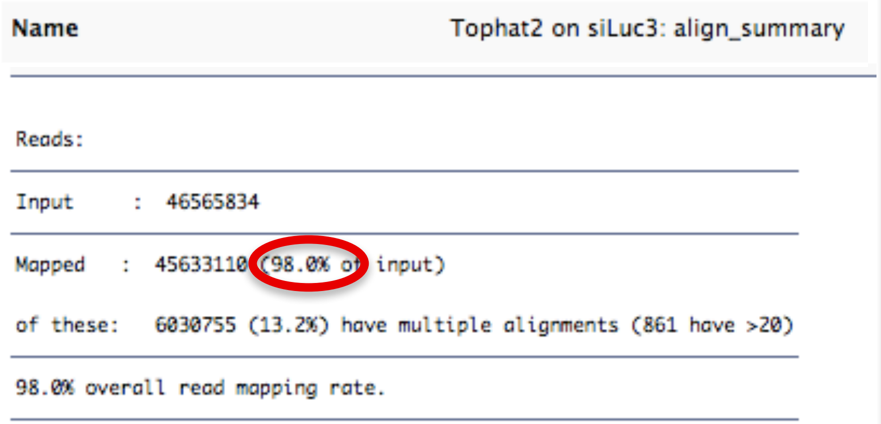
Exercise 2 - Question 1
Proportion of mapped reads in all samples

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



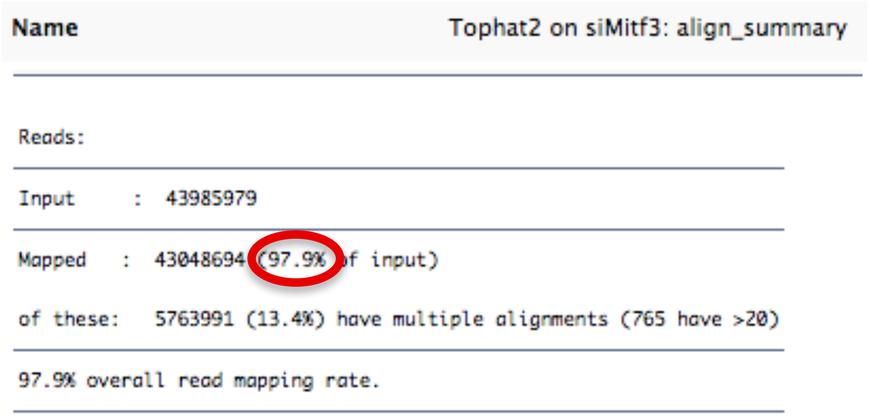
```
Name                        Tophat2 on siLuc2: align_summary

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.
```



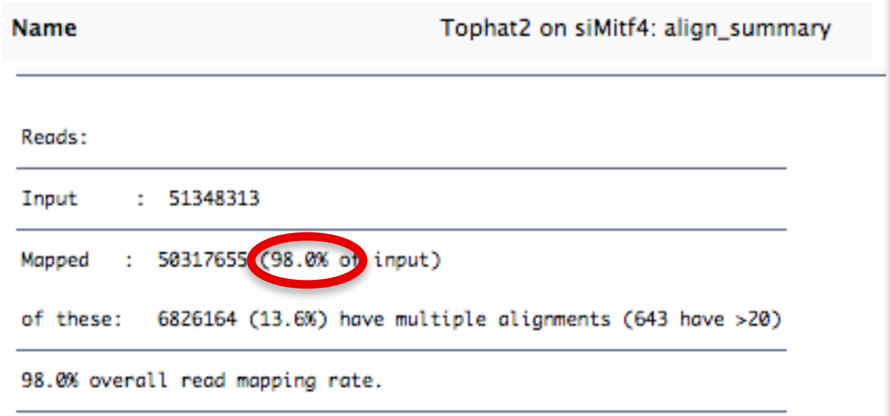
```
Name                        Tophat2 on siLuc3: align_summary

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

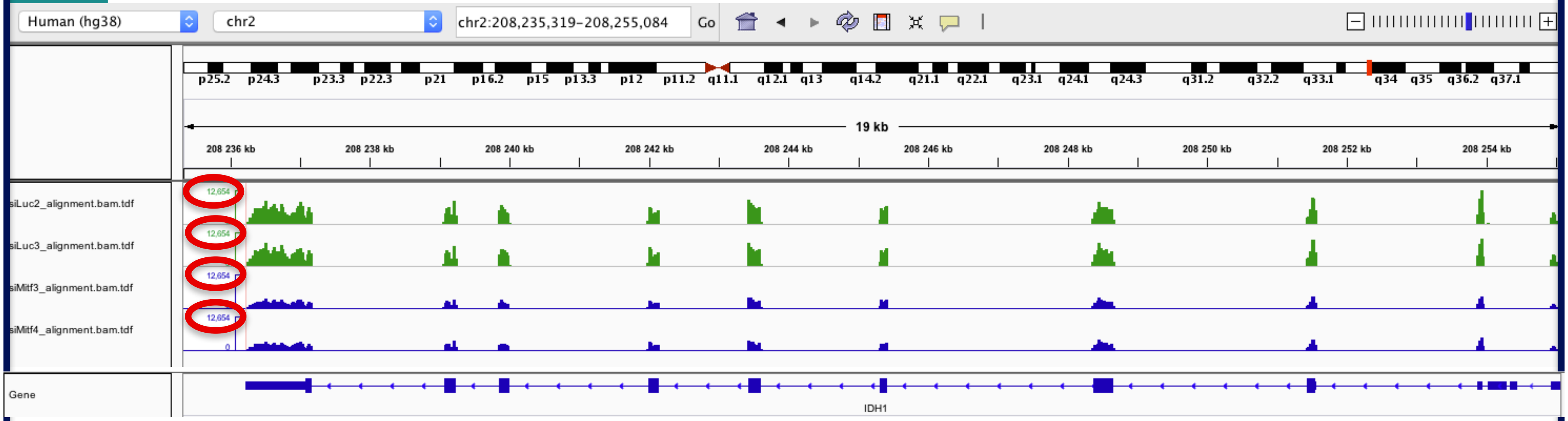
Exercise 2 – Question 2

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



Human (hg38)

chr2

chr2:208,235,319-208,255,084

Go

siLuc2_alignment.bam.tdf

siLuc3_alignment.bam.tdf

siMitf3_alignment.bam.tdf

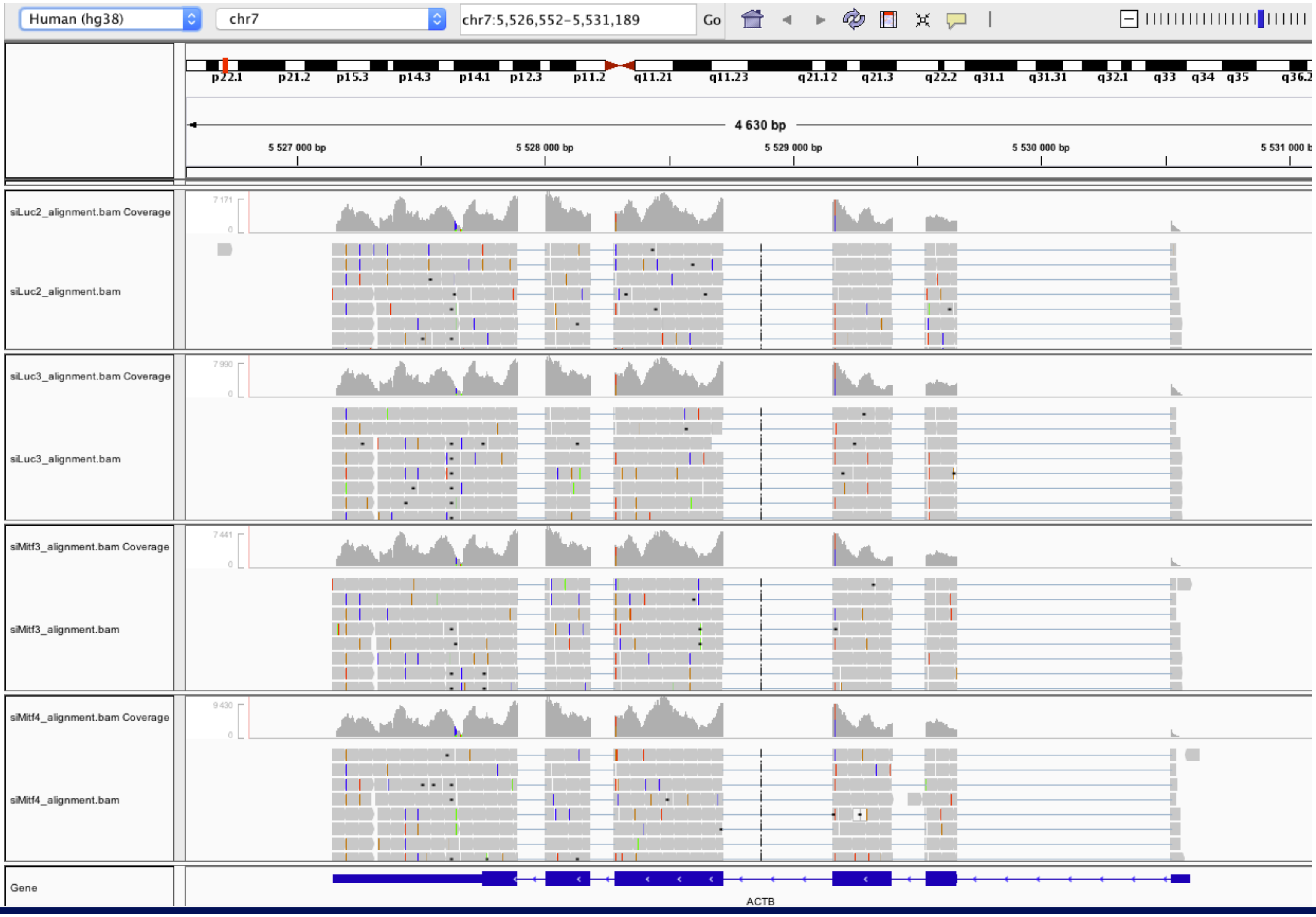
siMitf4_alignment.bam.tdf

Gene

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

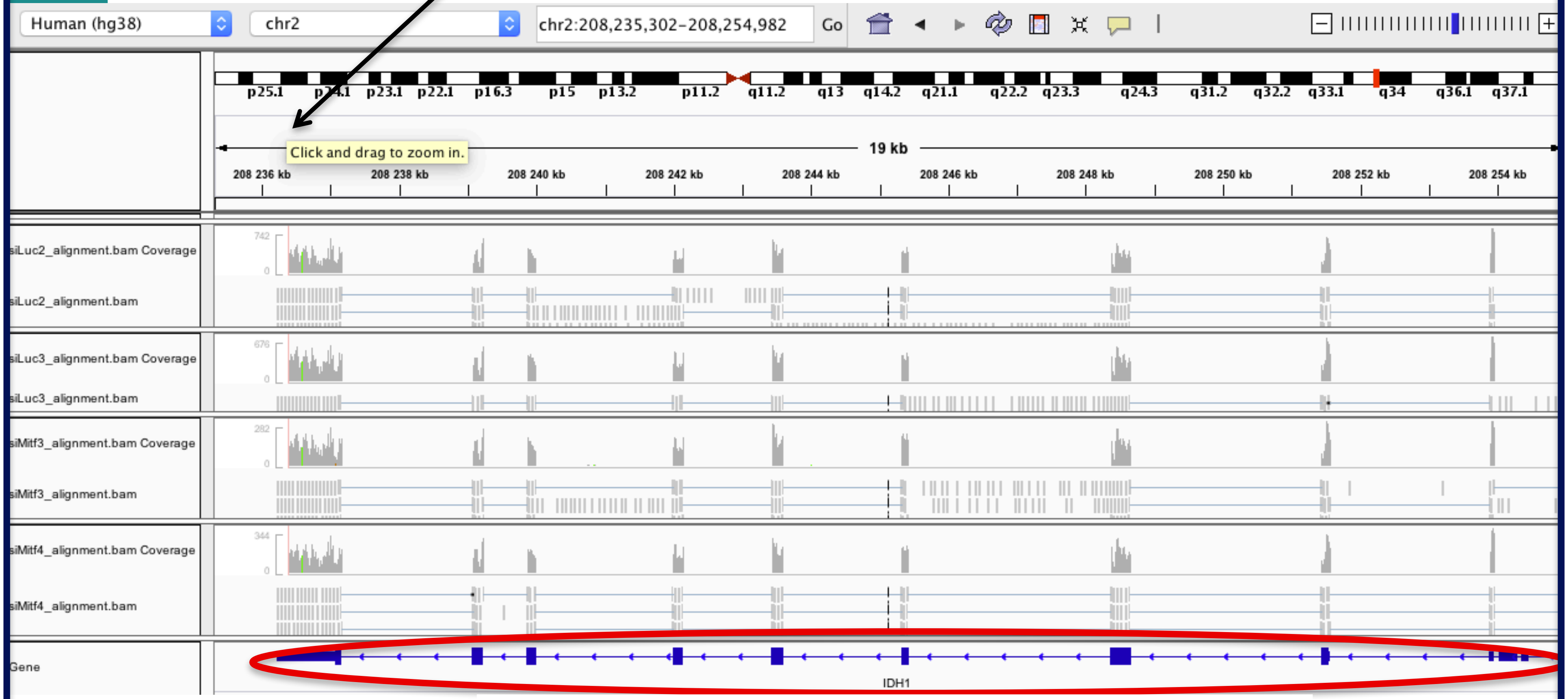
Exercise 2 – Question 3
Alignments visualization

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



Exercise 2 - Question 3

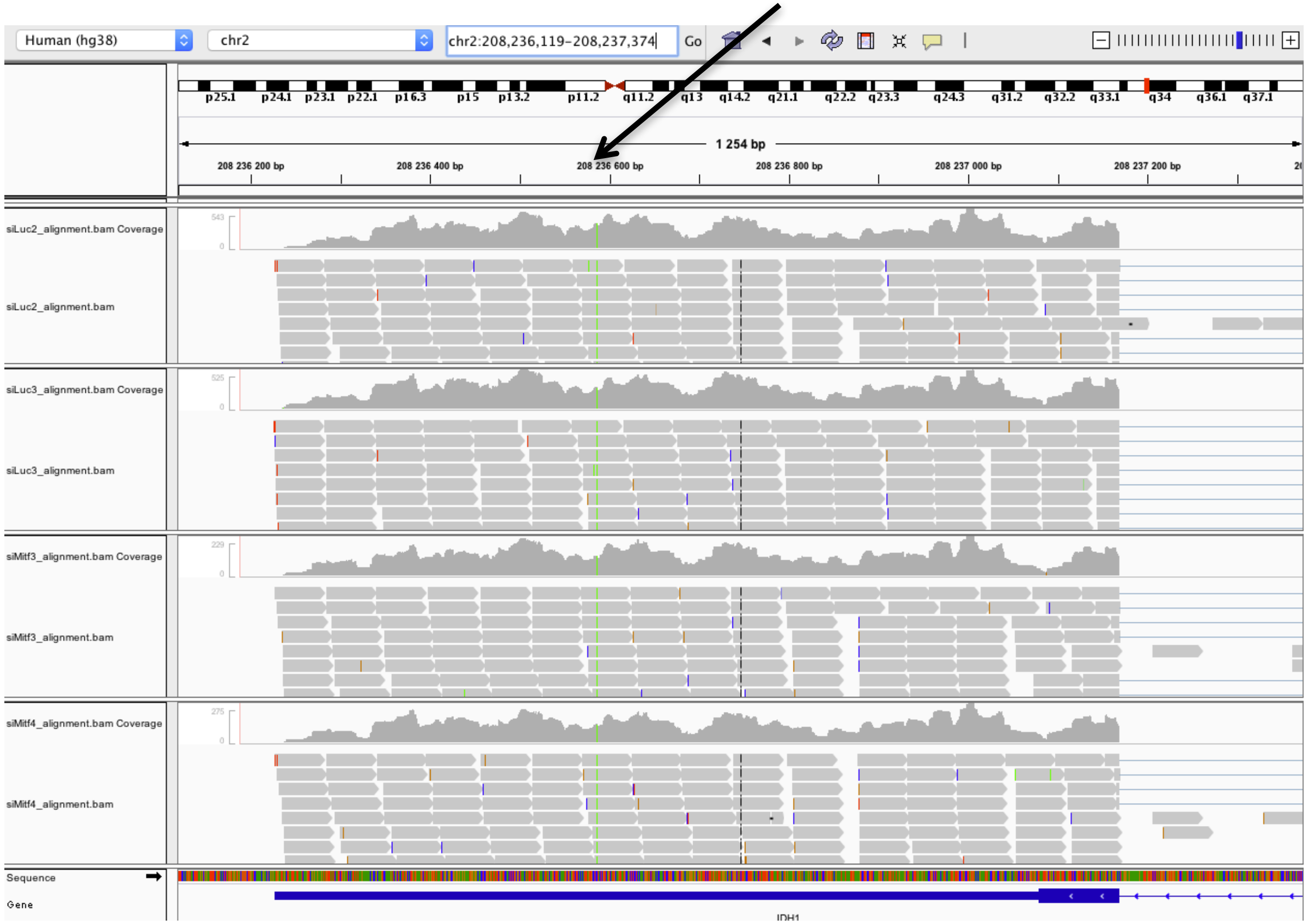
Click and drag to define a window around the last exon to zoom in



Arrows indicate annotated transcribed strand

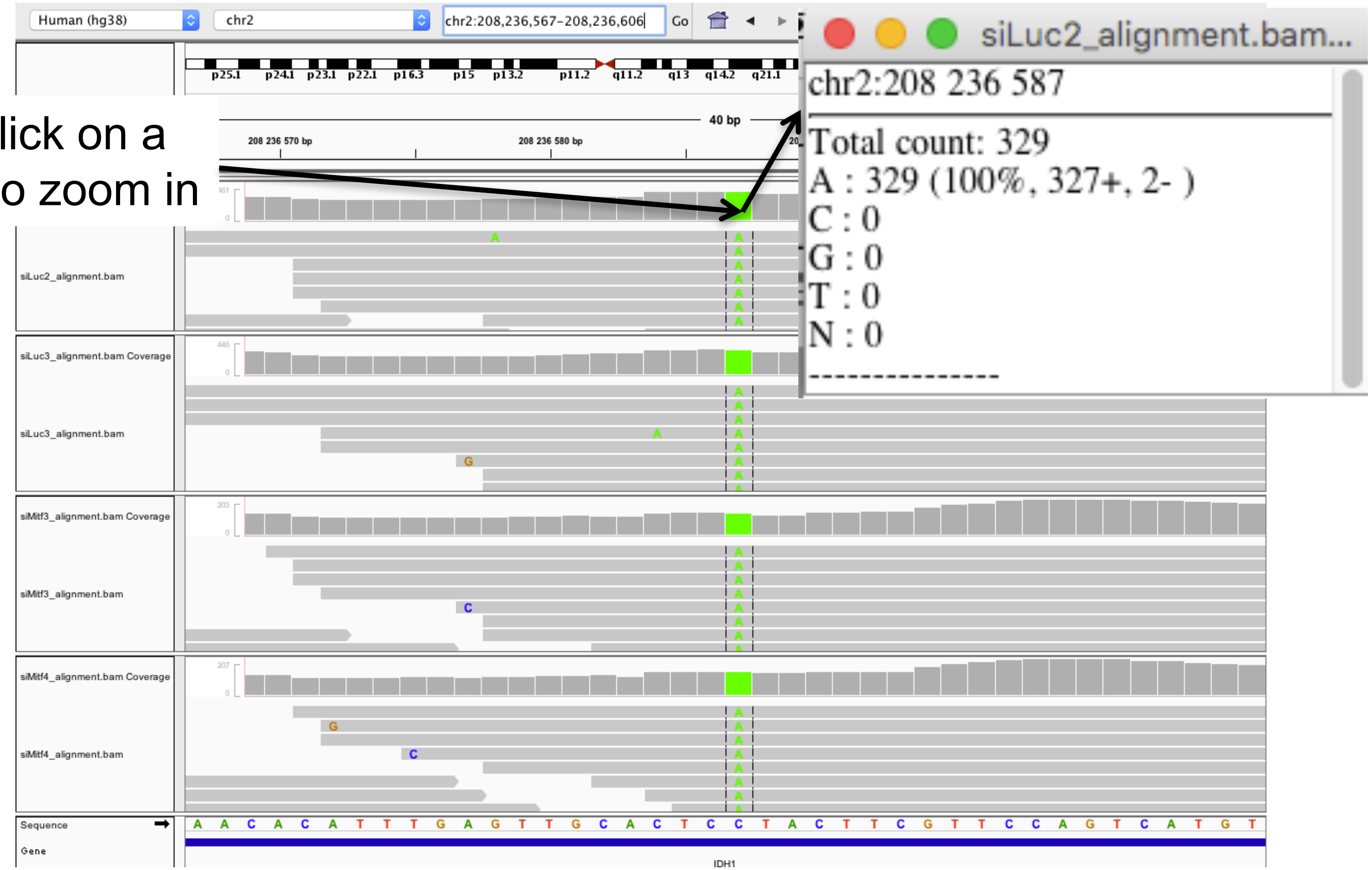
Exercise 2 - Question 3

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


Human (hg38)
chr2
chr2:208,236,119-208,237,374
Go
p25.1 p24.1 p23.1 p22.1 p16.3 p15 p13.2 p11.2 q11.2 q13 q14.2 q21.1 q22.2 q23.3 q24.3 q31.2 q32.2 q33.1 q34 q36.1 q37.1
1 254 bp
208 236 200 bp
208 236 400 bp
208 236 600 bp
208 236 800 bp
208 237 000 bp
208 237 200 bp
siLuc2_alignment.bam Coverage
siLuc2_alignment.bam
siLuc3_alignment.bam Coverage
siLuc3_alignment.bam
siMitf3_alignment.bam Coverage
siMitf3_alignment.bam
siMitf4_alignment.bam Coverage
siMitf4_alignment.bam
Sequence
Gene
IDH1

Exercise 2 - Question 3

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

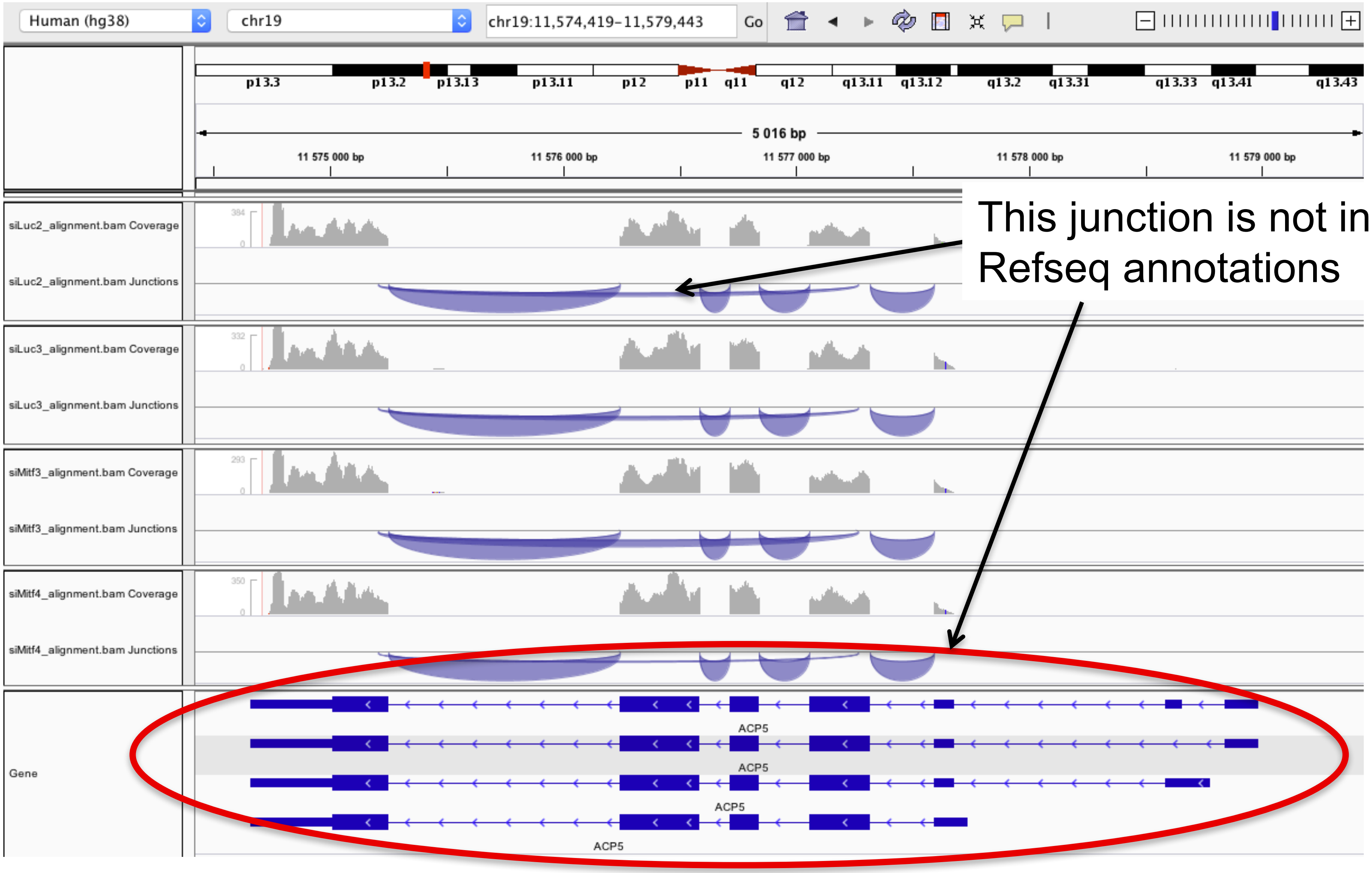


chr2:208 236 587

Total count: 329
A : 329 (100%, 327+, 2-)
C : 0
G : 0
T : 0
N : 0

Double click on a position to zoom in

Exercise 2 – Question 4



Human (hg38)

chr19

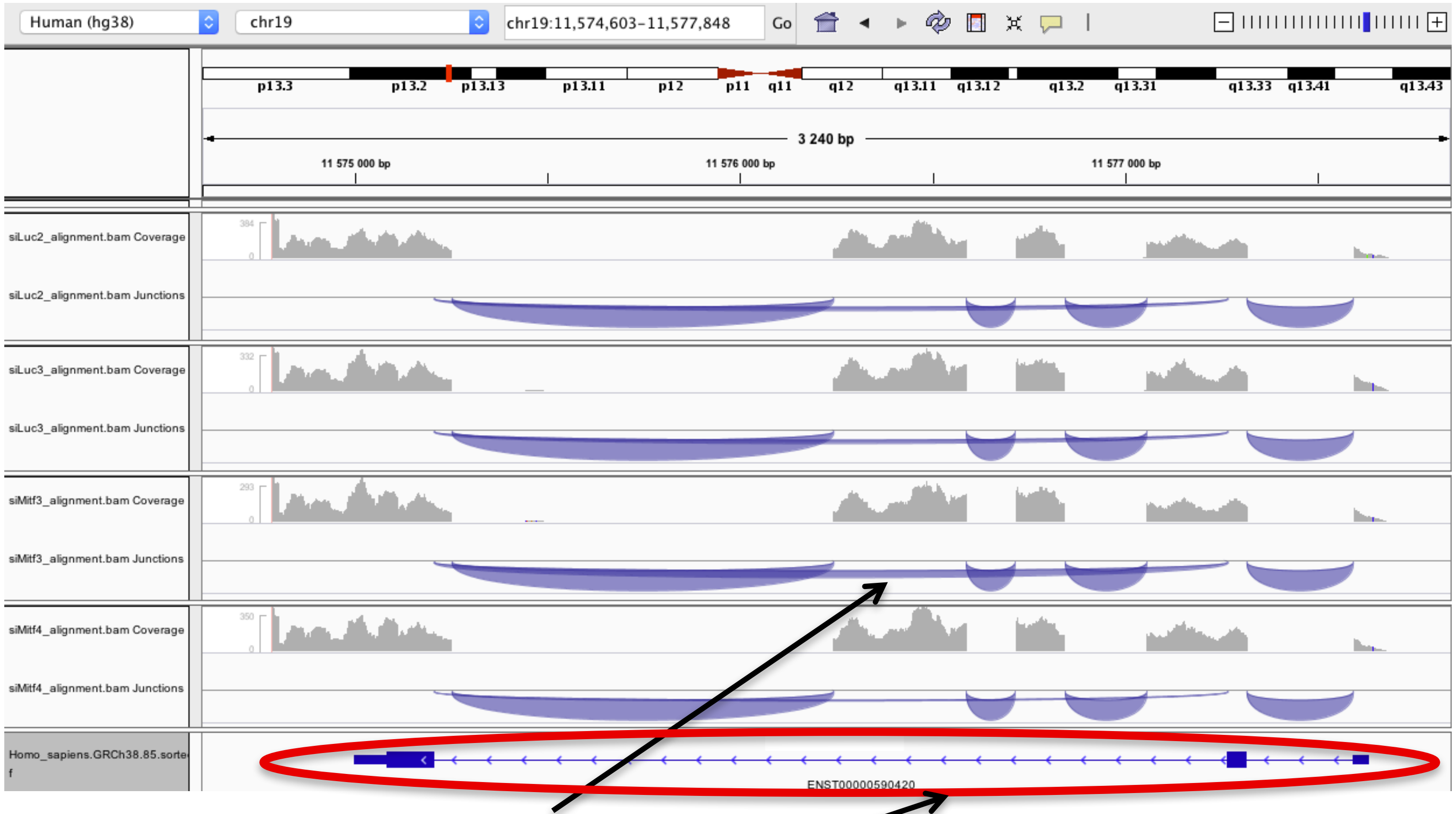
chr19:11,574,419-11,579,443

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
q13.43
5 016 bp
11 575 000 bp
11 576 000 bp
11 577 000 bp
11 578 000 bp
11 579 000 bp
siLuc2_alignment.bam Coverage
siLuc2_alignment.bam Junctions
siLuc3_alignment.bam Coverage
siLuc3_alignment.bam Junctions
siMitf3_alignment.bam Coverage
siMitf3_alignment.bam Junctions
siMitf4_alignment.bam Coverage
siMitf4_alignment.bam Junctions
Gene
ACP5
ACP5
ACP5
ACP5

This junction is not in Refseq annotations

Exercise 2 – Question 4

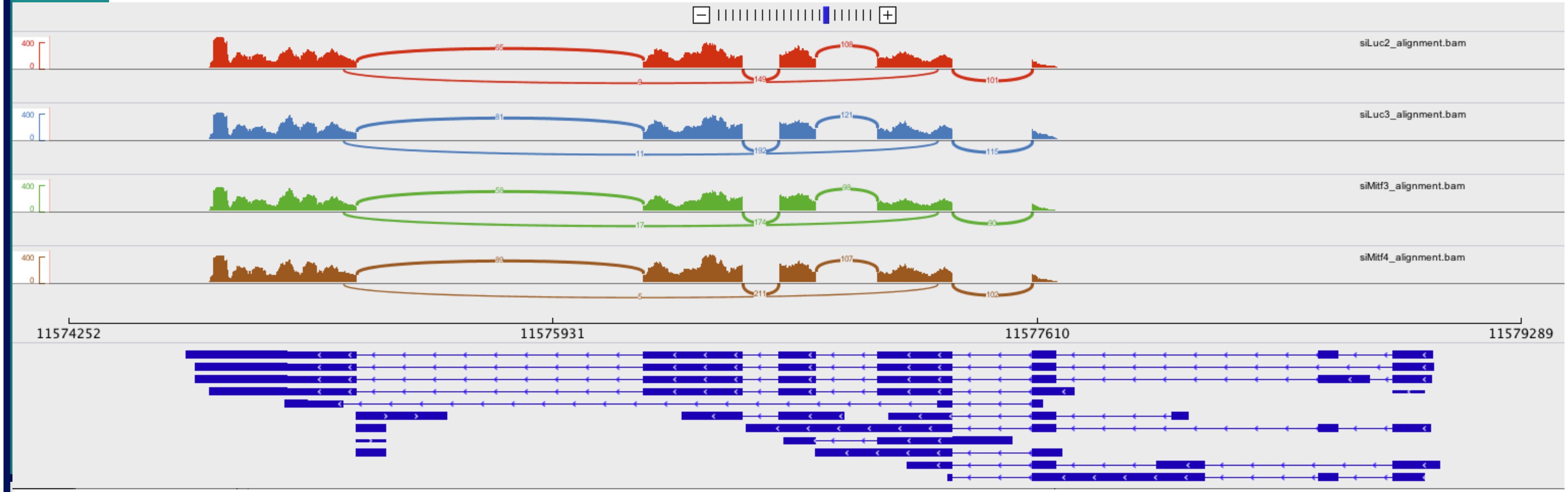
- File → Load from file and select Ensembl annotations (Homo_sapiens.GRCh38.85.sorted.gtf)
- Right click on Ensembl annotations track and select Expanded


Human (hg38)
chr19
chr19:11,574,603-11,577,848
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
q13.43
3 240 bp
11 575 000 bp
11 576 000 bp
11 577 000 bp
siLuc2_alignment.bam Coverage
siLuc2_alignment.bam Junctions
siLuc3_alignment.bam Coverage
siLuc3_alignment.bam Junctions
siMitf3_alignment.bam Coverage
siMitf3_alignment.bam Junctions
siMitf4_alignment.bam Coverage
siMitf4_alignment.bam Junctions
Homo_sapiens.GRCh38.85.sorted.gtf
ENST00000590420

This junction is in Ensembl annotations

Exercise 2 – Question 4

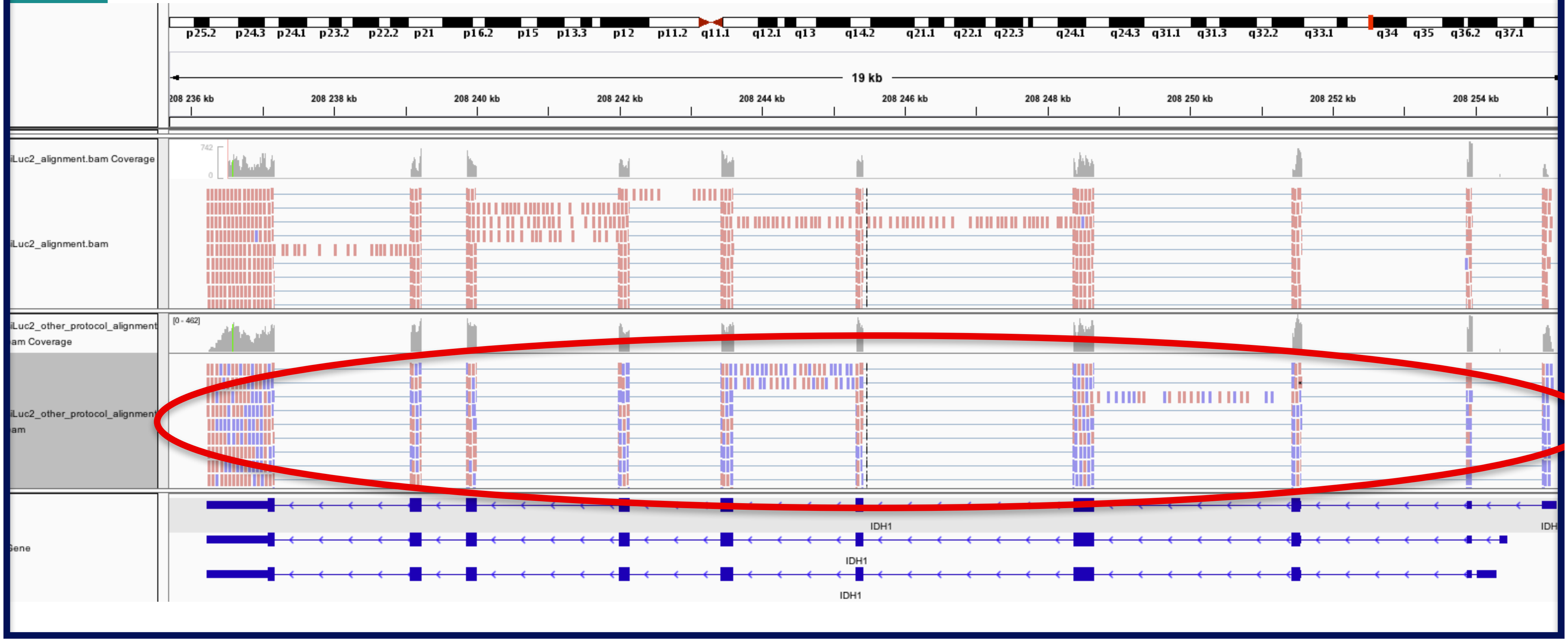
- Sashimi plot



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

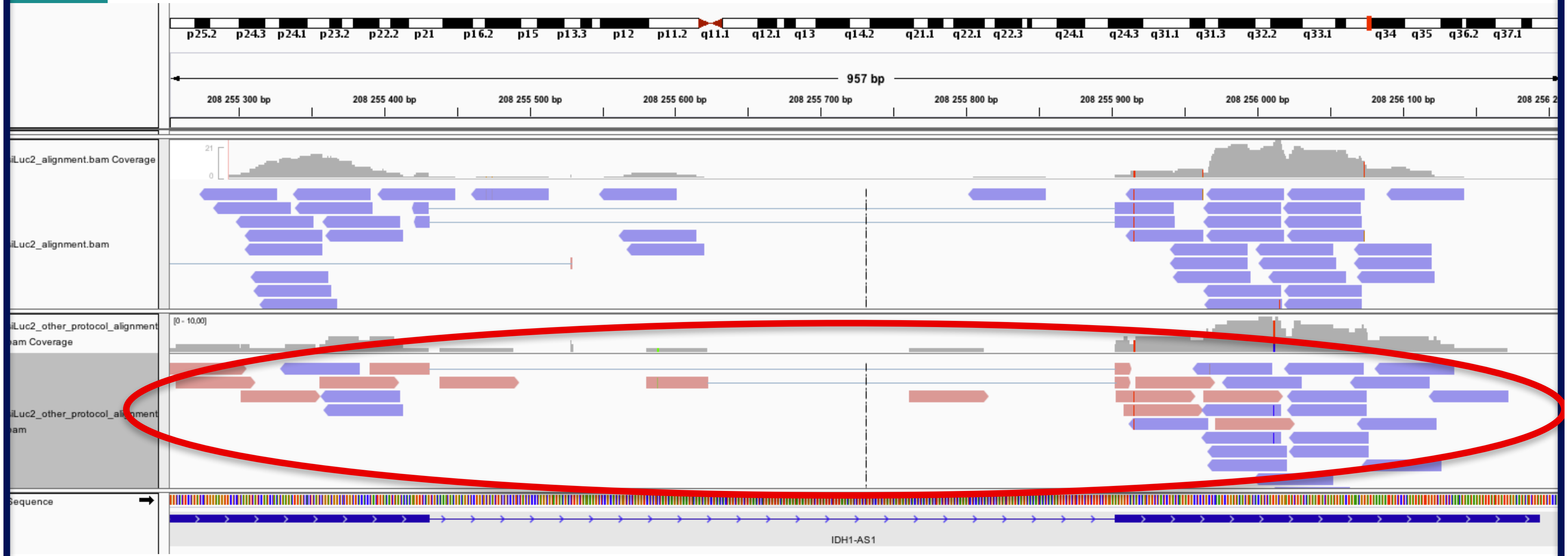
Exercise 2 – Question 5

- 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


p25.2 p24.3 p24.1 p23.2 p22.2 p21 p16.2 p15 p13.3 p12 p11.2 q11.1 q12.1 q13 q14.2 q21.1 q22.1 q22.3 q24.1 q24.3 q31.1 q31.3 q32.2 q33.1 q34 q35 q36.2 q37.1
19 kb
208 236 kb
208 238 kb
208 240 kb
208 242 kb
208 244 kb
208 246 kb
208 248 kb
208 250 kb
208 252 kb
208 254 kb
iLuc2_alignment.bam Coverage
742
0
iLuc2_alignment.bam
iLuc2_other_protocol_alignment
am Coverage
[0 - 462]
iLuc2_other_protocol_alignment
am
ene
IDH1
IDH1
IDH1
IDH

Exercise 2 – Question 5

- e.g. *Idh1-as1* gene


p25.2 p24.3 p24.1 p23.2 p22.2 p21 p16.2 p15 p13.3 p12 p11.2 q11.1 q12.1 q13 q14.2 q21.1 q22.1 q22.3 q24.1 q24.3 q31.1 q31.3 q32.2 q33.1 q34 q35 q36.2 q37.1
957 bp
208 255 300 bp
208 255 400 bp
208 255 500 bp
208 255 600 bp
208 255 700 bp
208 255 800 bp
208 255 900 bp
208 256 000 bp
208 256 100 bp
208 256 2
iLuc2_alignment.bam Coverage
21
0
iLuc2_alignment.bam
iLuc2_other_protocol_alignment
am Coverage
[0 - 10.00]
iLuc2_other_protocol_alignment
am
Sequence
IDH1-AS1

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