# 1 Module 7: RNA-Seq Expression Analysis (Human)

## 1.1 Practical Assignment - Answers

### 1.1.1 Task 1: Introducing the tutorial dataset

**1. Why is there more than one FASTQ per sample?**

There are 2 FASTQ files for each sample e.g. PT6_1.fastq and PT6_2.fastq. This is because this was paired-end sequence data.

### 1.1.2 Task 2: Mapping RNA-Seq reads to the genome using HISAT2

**2. How many index files were generated when you ran hisat2-build?**

There are **8** HISAT2 index files for our reference genome.

[ ]:
```
ls outputs/*.ht2 | wc -l
```

**3. What was the overall alignment rate for the PT2 sample to the reference genome?**

The overall alignment rate for PT2 was **95.30%**

```
801864 reads; of these:
  801864 (100.00%) were paired; of these:
    62170 (7.75%) aligned concordantly 0 times
    317242 (39.56%) aligned concordantly exactly 1 time
    422452 (52.68%) aligned concordantly >1 times
    ----
    62170 pairs aligned concordantly 0 times; of these:
      2085 (3.35%) aligned discordantly 1 time
    ----
    60085 pairs aligned 0 times concordantly or discordantly; of these:
      120170 mates make up the pairs; of these:
        75432 (62.77%) aligned 0 times
        16364 (13.62%) aligned exactly 1 time
        28374 (23.61%) aligned >1 times
95.30% overall alignment rate
```

**4. How does the alignment rate compare with that of the NP2 sample?**

The overall alignment rate for NP2 was **96.61%**. The alignment of the PT2 sample is 1.31% lower than the PT2 sample.

```
1180389 reads; of these:
  1180389 (100.00%) were paired; of these:
    60267 (5.11%) aligned concordantly 0 times
    442003 (37.45%) aligned concordantly exactly 1 time
    678119 (57.45%) aligned concordantly >1 times
    ----
    60267 pairs aligned concordantly 0 times; of these:
      1334 (2.21%) aligned discordantly 1 time
```

```
    ----
    58933 pairs aligned 0 times concordantly or discordantly; of these:
      117866 mates make up the pairs; of these:
        80006 (67.88%) aligned 0 times
        15578 (13.22%) aligned exactly 1 time
        22282 (18.90%) aligned >1 times
96.61% overall alignment rate
```

**5. How many NP2 reads were not aligned to the reference genome?**

Total number of unaligned reads in sample: **80,006**

**Here is a walkthrough of what the HISAT2 summary tells us for our NP2 sample and how we can tell which of the summary lines gives us this information:**

- We have 1,180,389 read pairs or 2,360,778 reads (2 x 1,180,389 pairs)

  `1180389 reads; of these:`

- All of our reads (100%) are paired - i.e. no reads without their mate

  `1180389 (100.00%) were paired; of these:`

- 1,120,122 pairs (94.9%) align concordantly one (37.45%) or more (57.45%) times

  ```
  442,003 (37.45%) aligned concordantly exactly 1 time
  678,119 (57.45%) aligned concordantly >1 times
  ```

- 60,267 pairs (5.11%) or 120,534 reads (2 x 60,267) did not align concordantly anywhere in the genome

  `60267 (5.11%) aligned concordantly 0 times`

- Of those 60,267 pairs, 1,334 pairs align discordantly (2.21% of the 60,267 pairs) `60267 pairs aligned concordantly 0 times; of these: 1334 (2.21%) aligned discordantly 1 time`

- This leave us with 58,933 pairs (60,267 - 1,334) where both reads in the pair do not align to the genome (concordantly or discordantly)

  `58933 pairs aligned 0 times concordantly or discordantly; of these:`

- Of those 117,866 reads (2 x 58,933) we have 37,860 reads (32.12%) which align to the genome without their mate `117866 mates make up the pairs; of these: .... 15578 (13.22%) aligned exactly 1 time 22282 (18.90%) aligned >1 times`

- Since we have 80,006 unaligned reads:

  `80006 (67.88%) aligned 0 times`

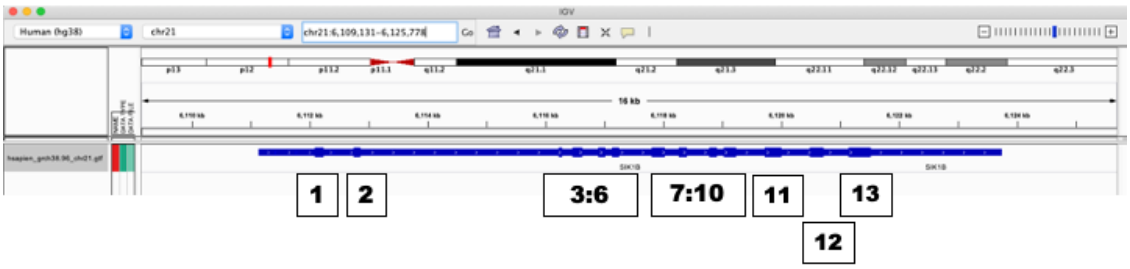
- The overall alignment rate is (2,360,778 - 80,006)/(2,360,778)*100= `96.61% overall alignment rate`

### 1.1.3 Task 3: Visualising transcriptomes with IGV

**6. How many CDS features are there in "SI1KB"?**

***SI1KB*** has **13** CDS features. You can get this in several ways:

- Count the number of exons/CDS features in the gene annotation:



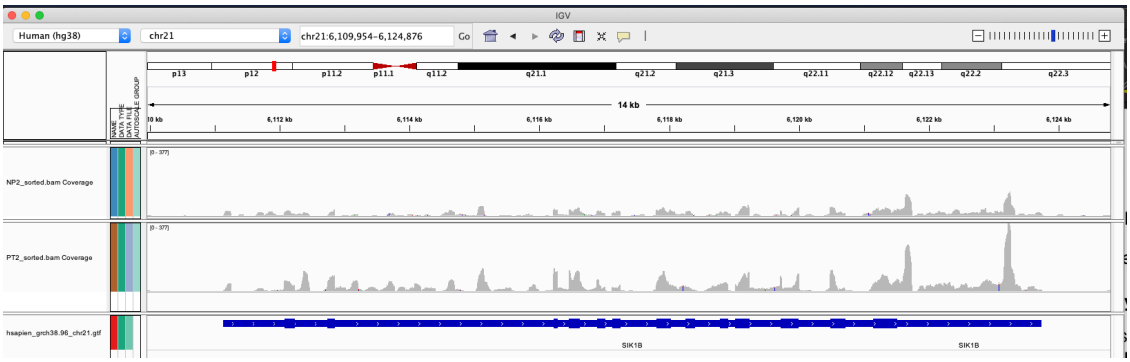

- Count the number of CDS features in the GFF file:

```
[ ]: grep -E "*SIK1B" data/hsapien_grch38.96_chr21.gtf | cut -f 3  | grep -c "CDS"
```

**7. Does the RNA-seq mapping agree with the gene model for SI1KB in blue?**

Yes. In general the peaks of the coverage tracks correspond to the annotated exon/CDS features.

**8. Do you think this gene (SIK1B) is differentially expressed between prostate cancer and normal adjacent tissue? Is looking at the coverage plots alone a reliable way to assess differential expression?**



Possibly. But, you can't tell differential expression by the counts alone as there may be differences in the sequencing depths of the samples. It is always better to do a statistical test.

### 1.1.4 Task 4:Transcript quantification with Kallisto

**9. What k-mer length was used to build the Kallisto index?**

A k-mer length of **31** was used.

Look at the output from kallisto index:

```
[build] k-mer length: 31
```

Or, look for the `-k` or `--kmer-size` option in the kallisto index usage:

```
kallisto index
```

**10. How many transcript sequences are there in hsapiens_chr21_transcripts.fa?**

There are **3,039** transcript sequences.

Look at the output from kallisto quant:

```
[index] number of targets: 3,039
```

Or, look for **n_targets** in one of the *run_info.json* files:

[ ]:
```
cat outputs/PT6/run_info.json
```

Or, you can run grep on the transcript FASTA file and count the number of header lines:

[ ]:
```
grep -c ">" data/hsapiens_chr21_transcripts.fa
```

**11. What is the transcripts per million (TPM) value for ENST00000399975.7 (USP16) in each of the samples?**

| Sample | TPM |
|---|---|
| NP10 | 0 |
| NP13 | 234.196 |
| NP2 | 35.5784 |
| NP4 | 333.785 |
| NP5 | 37.3378 |
| NP6 | 113.044 |
| PT10 | 88.171 |
| PT13 | 47.9757 |
| PT2 | 292.437 |
| PT4 | 68.9001 |
| PT5 | 88.3665 |
| PT6 | 126.898 |

You can look at each of the individual abundance files:

[ ]:
```
grep "^ENST00000399975.7" outputs/*/abundance.tsv |  awk -F"\t" '{print␣
↪$1"\t"$5}'
```

Or you can use a recursive grep:

[ ]:
```
grep -r "^ENST00000399975.7" outputs
```

**12. Do you think ENST00000399975.7 is differentially expressed?**

It is difficult to tell from simply looking at the TPM values. We would need to perform a statistical test to make that determination.

### 1.1.5 Task 5: Identifying differentially expressed genes with Sleuth"

**13: What is the most abundantly expressed transcript in the PT6 sample?**

ENST00000577708.1/ENST00000614492.1