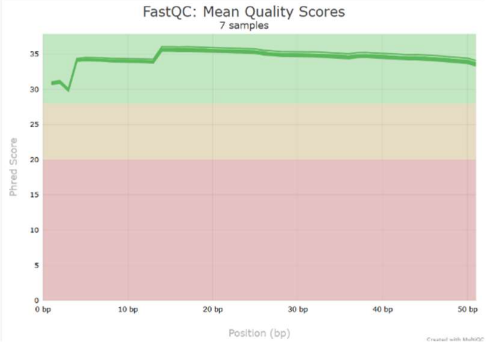
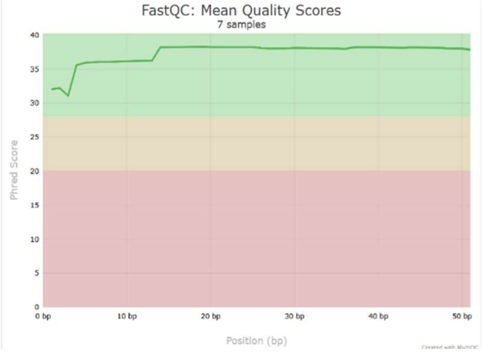
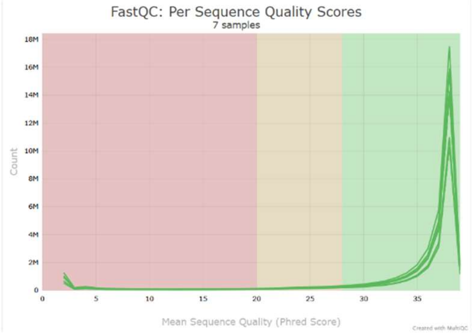
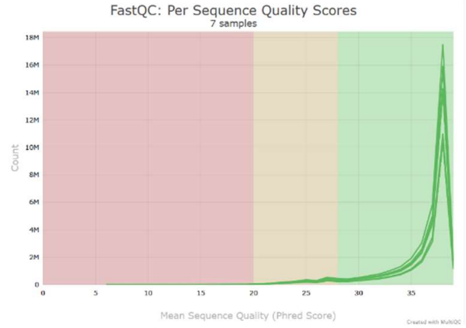
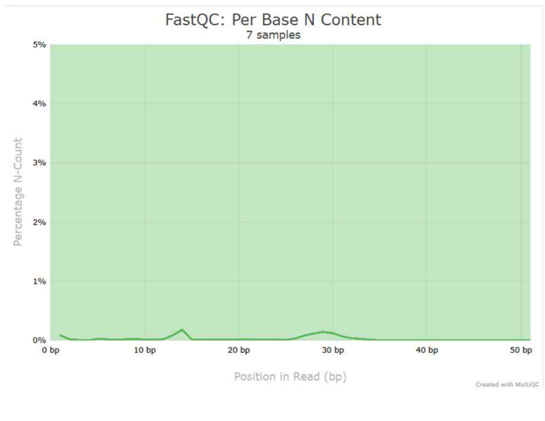
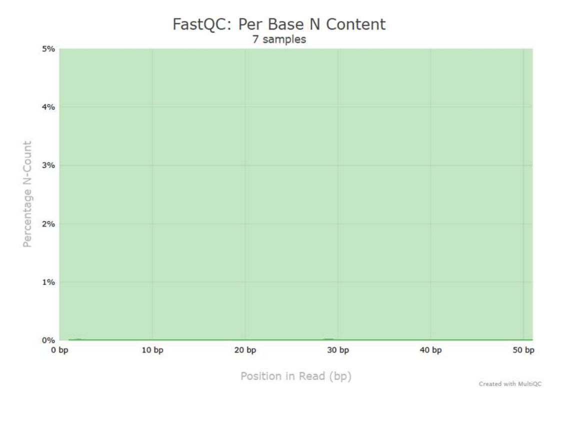
# Week 4 – Quality control and HISAT2 alignment

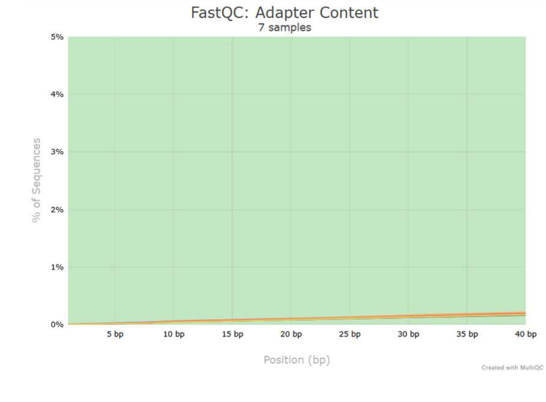
The single end transcriptome datasets (BioProject: PRJEB32551) were first analysed using FASTQc and a combined report was generated using MultiQC. The overall quality of the data was found to be very good with majority of the datapoints of the mean quality score lying in the green region and Per Sequence Quality Scores being >35 phred score. GC content was roughly 50% consistently amongst samples. Substantial amount of poly A adapter were found and small amount of N content (0.1-0.2%) was also found.

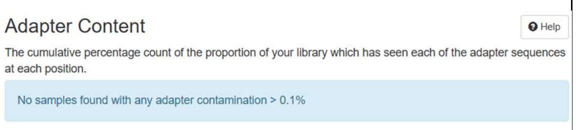
In order to remove these imperfections, Trimmomatic was used. The TreuSeq3 mode was used for single end sequences, with a SLIDINGWINDOW of 4, average quality required to keep read of 20, LEADING and TRIALING base pair values of 3. Although Trimming removed the N contents, adapters were still seen on examinations and were removed using the Cutadpt, with both 3' and 5' removal options enabled.

| **MultiQc before Trimming and adapter removal** | **MultiQc after Trimming and adapter removal** |
|---|---|
|  |  |
|  |  |







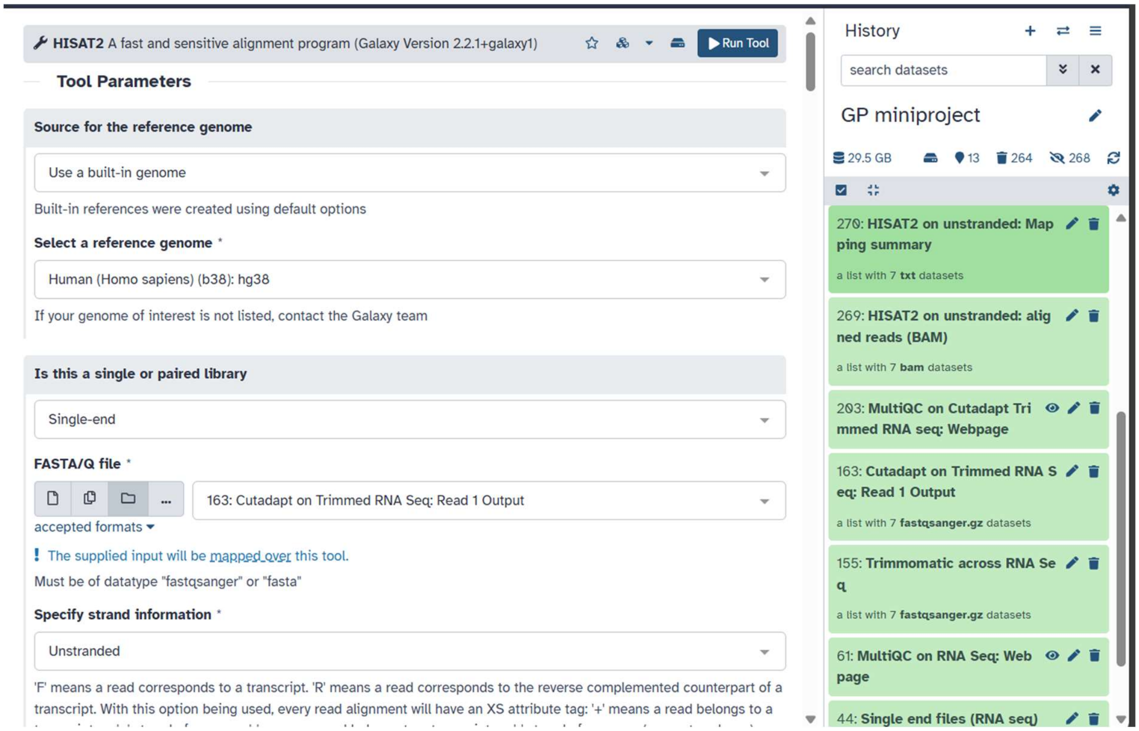


Adapter Content

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

No samples found with any adapter contamination > 0.1%

HISAT2 was used to map the cleaned transcriptome sequences against the human reference sequence hg38. The overall alignment was found to be >90% for forward, reverse and unstranded strandedness, therefore the sequences are assumed to be unstranded



**HISAT2** A fast and sensitive alignment program (Galaxy Version 2.2.1+galaxy1)

**Tool Parameters**

**Source for the reference genome**

Use a built-in genome

Built-in references were created using default options

**Select a reference genome** *

Human (Homo sapiens) (b38): hg38

If your genome of interest is not listed, contact the Galaxy team

**Is this a single or paired library**

Single-end

**FASTA/Q file** *

163: Cutadapt on Trimmed RNA Seq: Read 1 Output

accepted formats ▾

! The supplied input will be mapped over this tool.

Must be of datatype "fastqsanger" or "fasta"

**Specify strand information** *

Unstranded

'F' means a read corresponds to a transcript. 'R' means a read corresponds to the reverse complemented counterpart of a transcript. With this option being used, every read alignment will have an XS attribute tag: '+' means a read belongs to a

History

search datasets

GP miniproject

29.5 GB 13 264 268

270: **HISAT2 on unstranded: Mapping summary**
a list with 7 **txt** datasets

269: **HISAT2 on unstranded: aligned reads (BAM)**
a list with 7 **bam** datasets

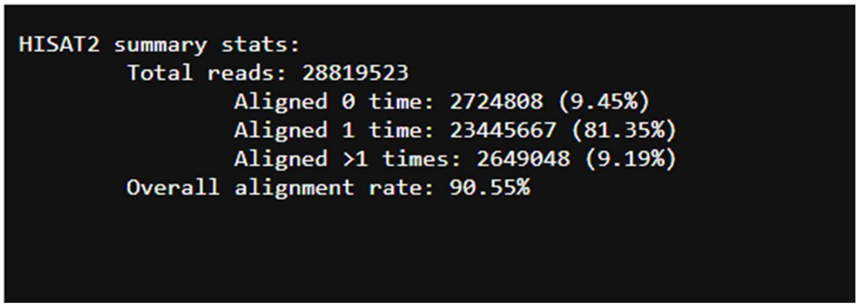
203: **MultiQC on Cutadapt Trimmed RNA seq: Webpage**

163: **Cutadapt on Trimmed RNA Seq: Read 1 Output**
a list with 7 **fastqsanger.gz** datasets

155: **Trimmomatic across RNA Seq**
a list with 7 **fastqsanger.gz** datasets

61: **MultiQC on RNA Seq: Webpage**

44: **Single end files (RNA seq)**



```
HISAT2 summary stats:
        Total reads: 28819523
                Aligned 0 time: 2724808 (9.45%)
                Aligned 1 time: 23445667 (81.35%)
                Aligned >1 times: 2649048 (9.19%)
        Overall alignment rate: 90.55%
```