

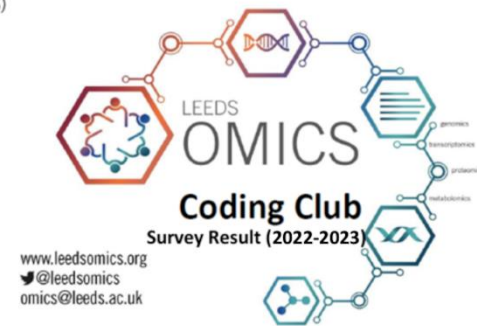
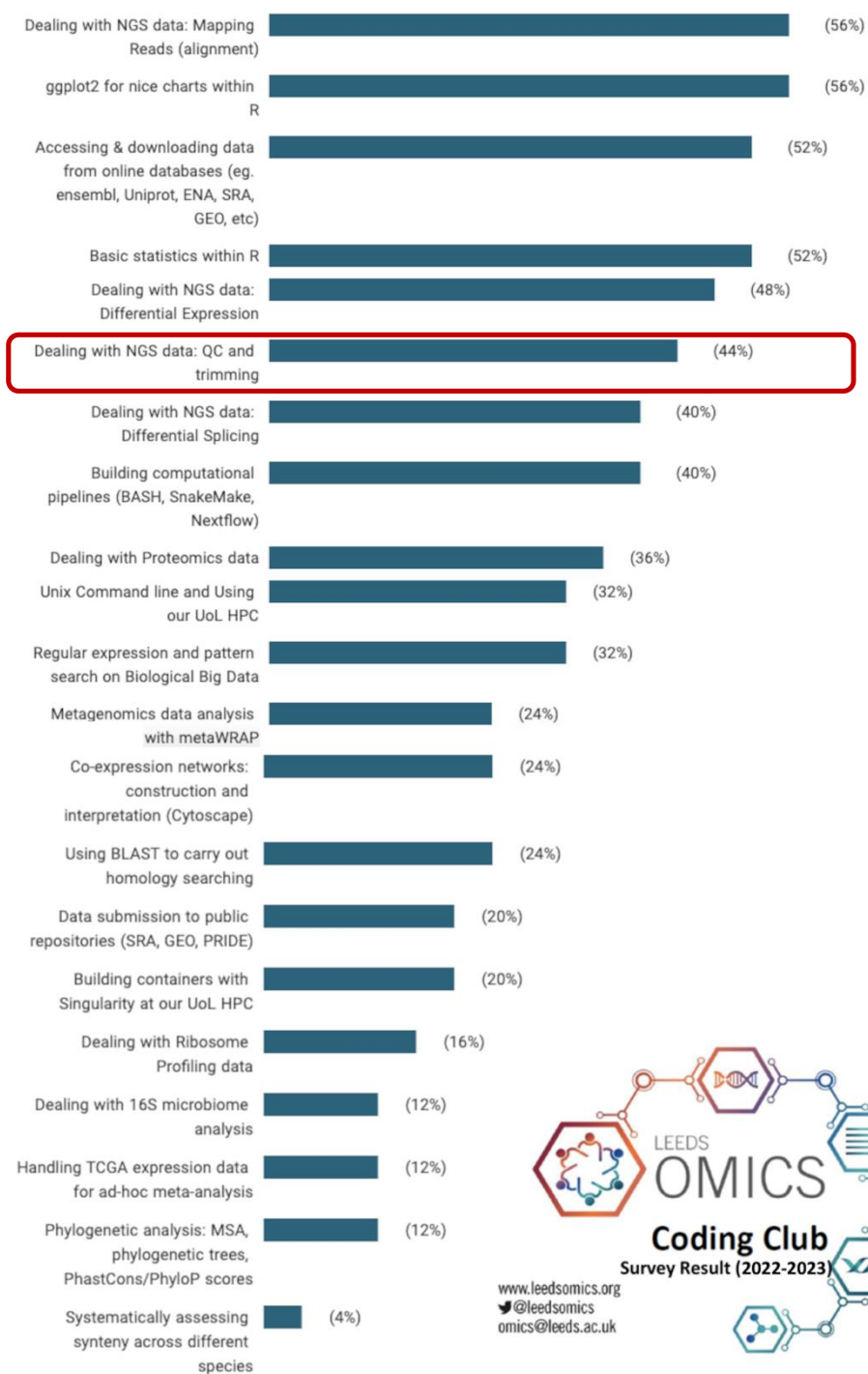
[www.leedsomics.org](http://www.leedsomics.org)  
@leedsomics  
[omics@leeds.ac.uk](mailto:omics@leeds.ac.uk)

## Dealing with NGS data: QC and trimming

Club Moderator(s): Elton Vasconcelos

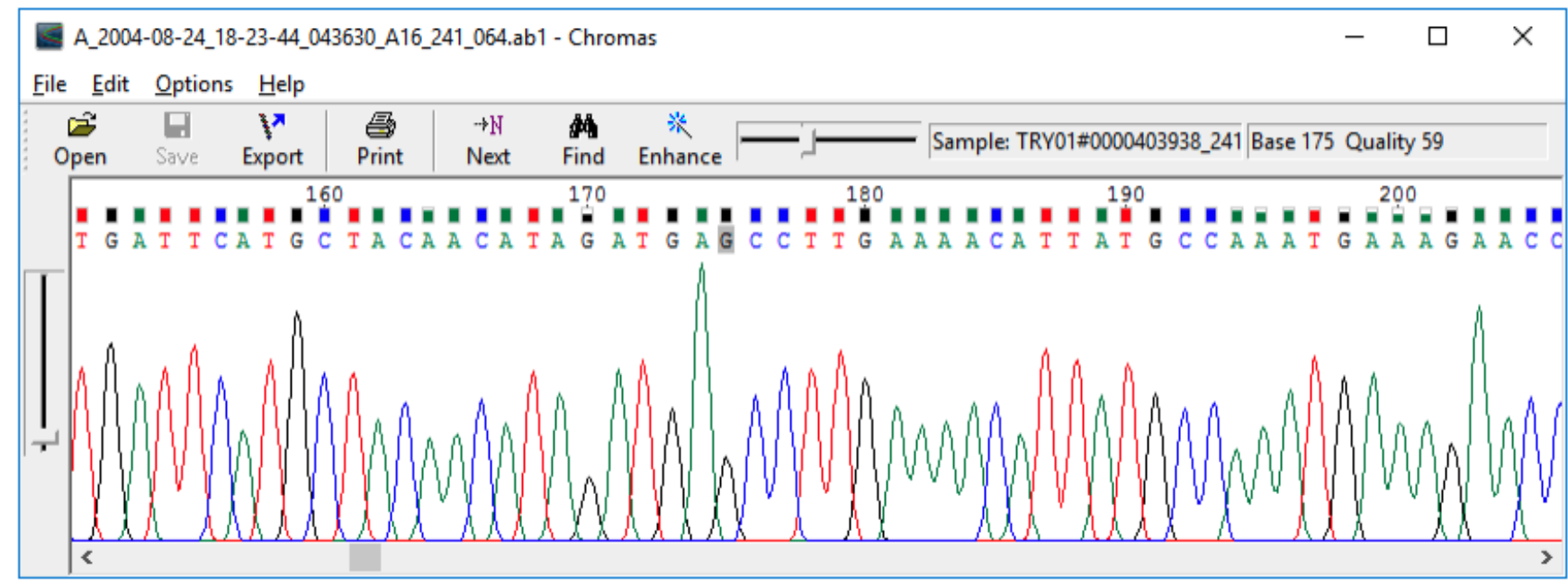
# Topics to be addressed on the 2022-23 season - Survey Result

6th session

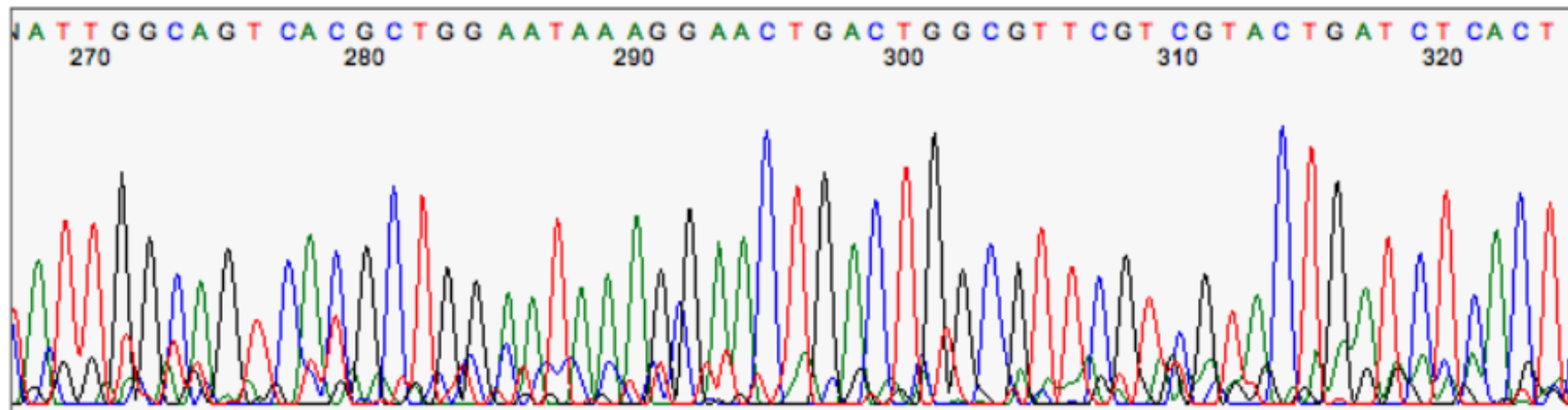


# Going back in time: Sanger Sequencing Strategy

Good quality sequencing



Not as good



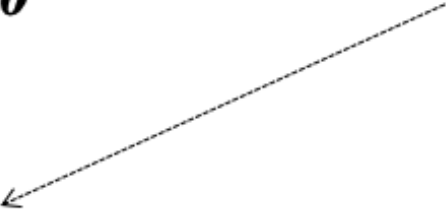
## Phred - Quality Base Calling

**Phred** is a base-calling program for DNA sequence traces. Phred reads DNA sequence chromatogram files and analyzes the peaks to call bases, assigning quality scores ("Phred scores") to each base call. Phred was developed by Drs. Phil Green and Brent Ewing, and is distributed by CodonCode Corporation under license from the University of Washington. Phred is widely used by the largest academic and commercial DNA sequencing laboratories. This page gives a brief description of Phred. For information about Phrap, Cross\_match, and Consed, please visit [www.phrap.com](http://www.phrap.com).

[Ewing et al.1998a, Genome Research 8:175-85](#)

$$QV = -10 \log_{10} EP$$

EP = error probability



| Phred quality score | Probability that the base is called wrong | Accuracy of the base call |
|---------------------|---|---------------------------|
| 10                  | 1 in 10                                   | 90%                       |
| 20                  | 1 in 100                                  | 99%                       |
| 30                  | 1 in 1,000                                | 99.9%                     |
| 40                  | 1 in 10,000                               | 99.99%                    |
| 50                  | 1 in 100,000                              | 99.999%                   |

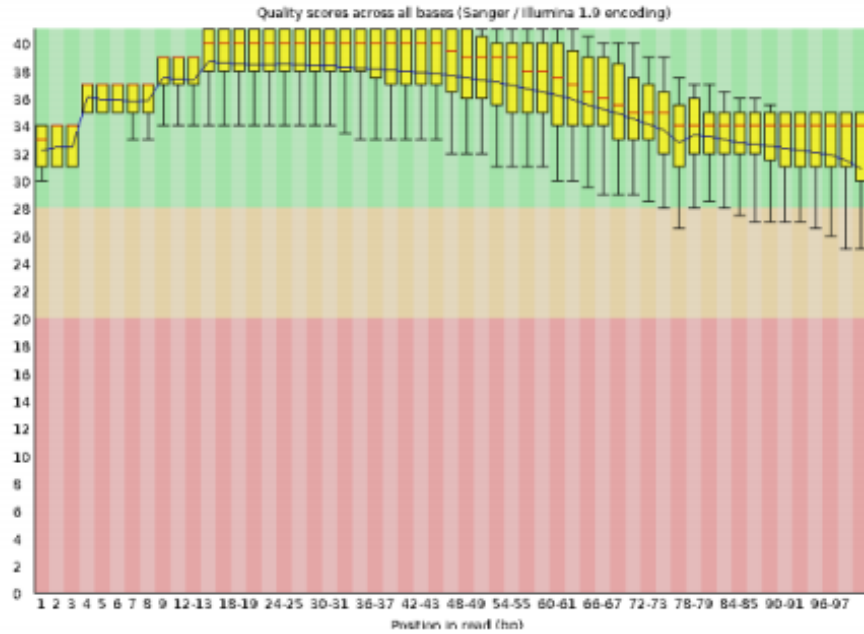






# FastQC for Quality Control (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)

## ✔ Per base sequence quality

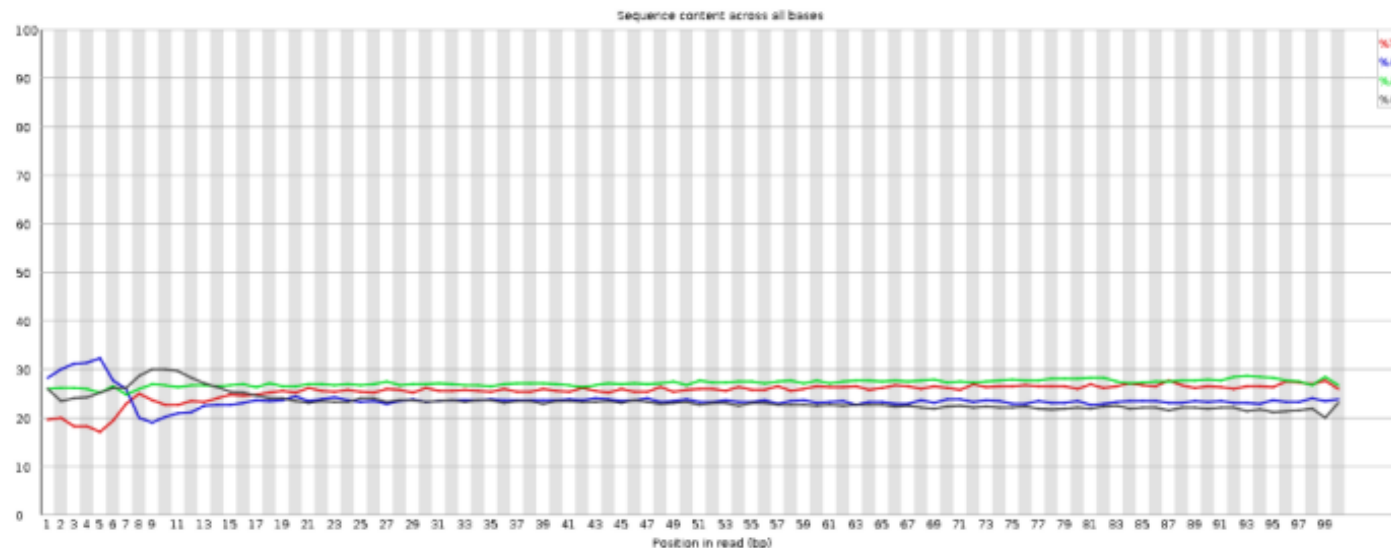


```
$ fastqc -t 4 --nogroup *.fastq.gz
```

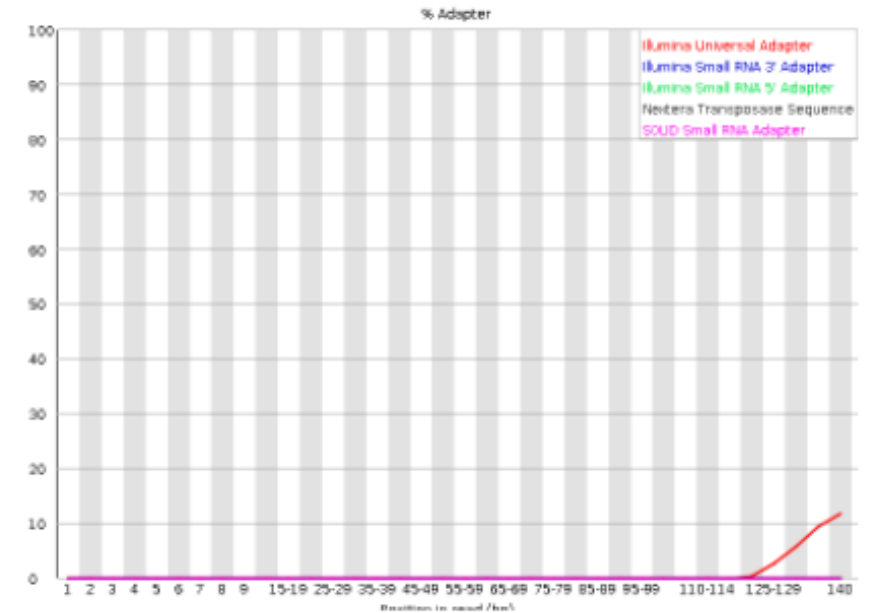
- **MultiQC** (<https://multiqc.info/>)

```
$ multiqc -o multiqc_out/ .
```

## ✔ Per base sequence content



## ✘ Adapter Content



# Trimming adapters and low quality bases at reads' extremities

- Trimmomatic (<http://www.usadellab.org/cms/?page=trimmomatic>)
- Cutadapt (<https://cutadapt.readthedocs.io/en/stable/>)
- Trim Galore ([https://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/))

Example of trimmomatic command line:

```
$ java -jar /your/trimmomatic/folder/path/trimmomatic-0.36.jar PE -threads 24 -phred33  
input_R1.fastq.gz input_R2.fastq.gz output_R1-trimmed_paired.fq.gz output_R1-  
trimmed_unpaired.fq.gz output_R2-trimmed_paired.fq.gz output_R2-trimmed_unpaired.fq.gz  
HEADCROP:15 ILLUMINACLIP:/your/trimmomatic/folder/path/adapters/user-  
defined_adapters_file.fasta:2:30:7 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:30
```

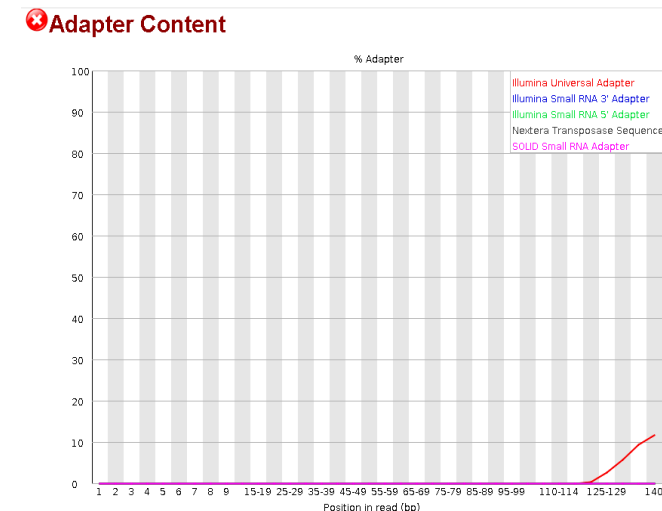
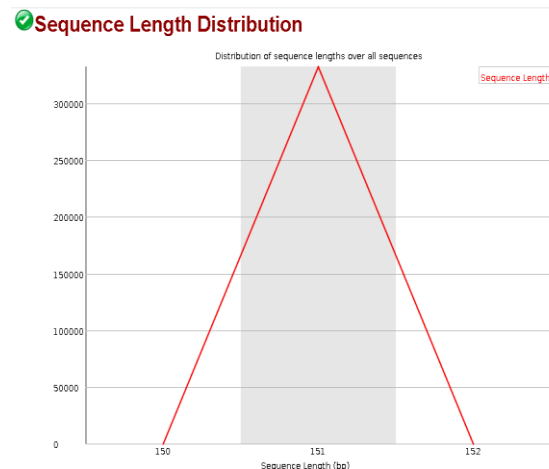
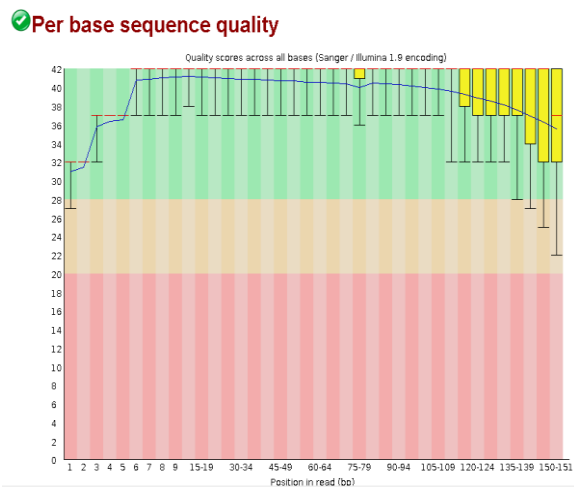


# Re-analyse the trimmed data using FastQC

## Before trimming

### Basic Statistics

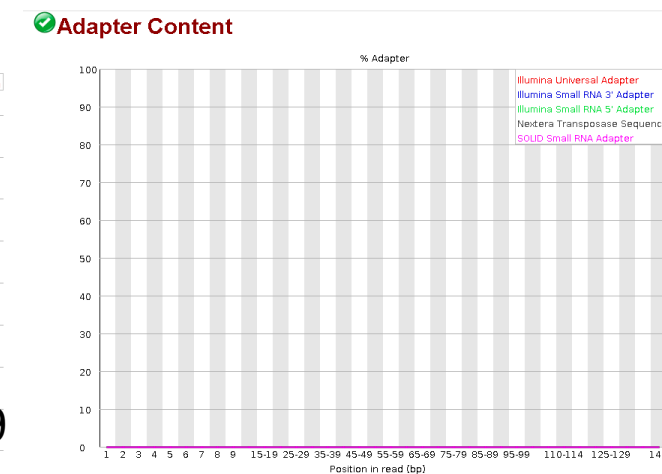
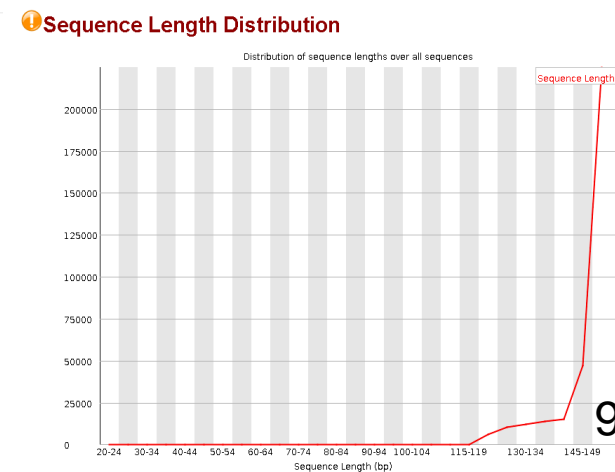
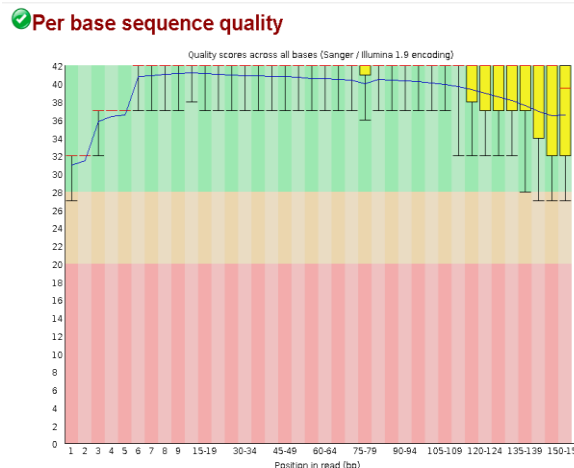
| Measure                           | Value                   |
|-----------------------------------|-------------------------|
| Filename                          | normal_rep1_r1.fastq    |
| File type                         | Conventional base calls |
| Encoding                          | Sanger / Illumina 1.9   |
| Total Sequences                   | 331958                  |
| Sequences flagged as poor quality | 0                       |
| Sequence length                   | 151                     |
| %GC                               | 54                      |



## After trimming

### Basic Statistics

| Measure                           | Value                   |
|-----------------------------------|-------------------------|
| Filename                          | normal_rep1_r1_val_1.fq |
| File type                         | Conventional base calls |
| Encoding                          | Sanger / Illumina 1.9   |
| Total Sequences                   | 331389                  |
| Sequences flagged as poor quality | 0                       |
| Sequence length                   | 21-151                  |
| %GC                               | 54                      |



Bring your issues on!