Quality control pipeline: Trimming and filtering features using Flexbar

Introduction

PL-Grid instance of Galaxy provides all the tools necessary for quality control of your raw reads, using FASTQC software, and for preparation to alignment to the reference throughout trimming and filtering, using Flexbar software. This tutorial will give you an introduction to how to use these tools and it will guide you through the process.

Input Dataset

In this tutorial we will use the file '`test_dataset.fastq`' which is available for download from the Bismark homepage (it contains 10,000 reads in FastQ format, Phred33 qualities, 50 bp long reads, from a human directional BS-Seq library).

Link to input dataset:

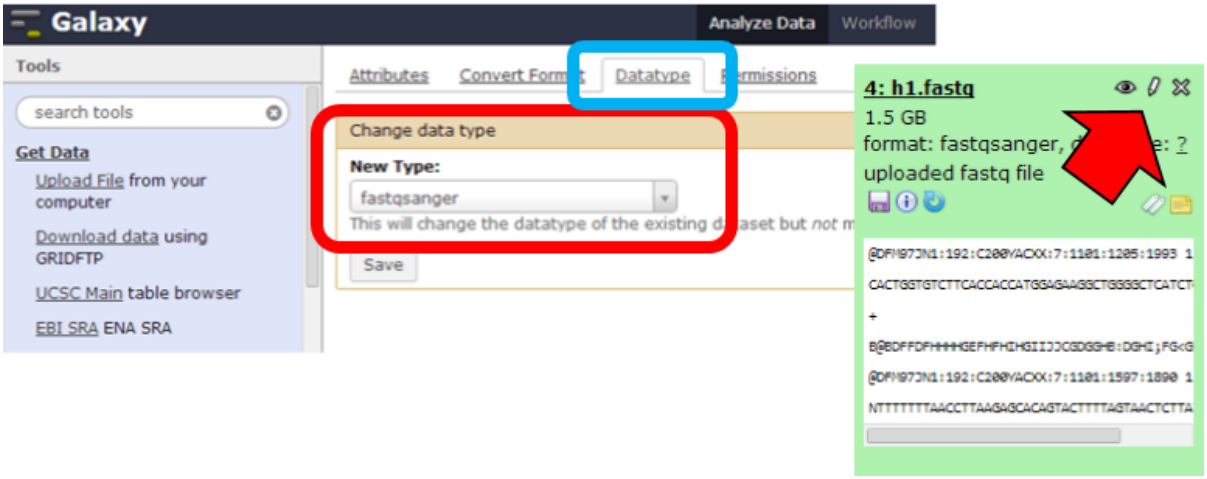
http://www.bioinformatics.babraham.ac.uk/projects/bismark/test_data.fastq

Also, you may try using your own input files. In such a case, please use the Upload File tool.

> **Tools**
>
> Get data -> Upload File

File format: fastqsanger [HINT: this is important, because galaxy will recognize all kinds of fastq files as generic fastq format. However, most tools require more specific fastqsanger format]

You could do it also later using "Edit Attributes" in your history Data/History window.



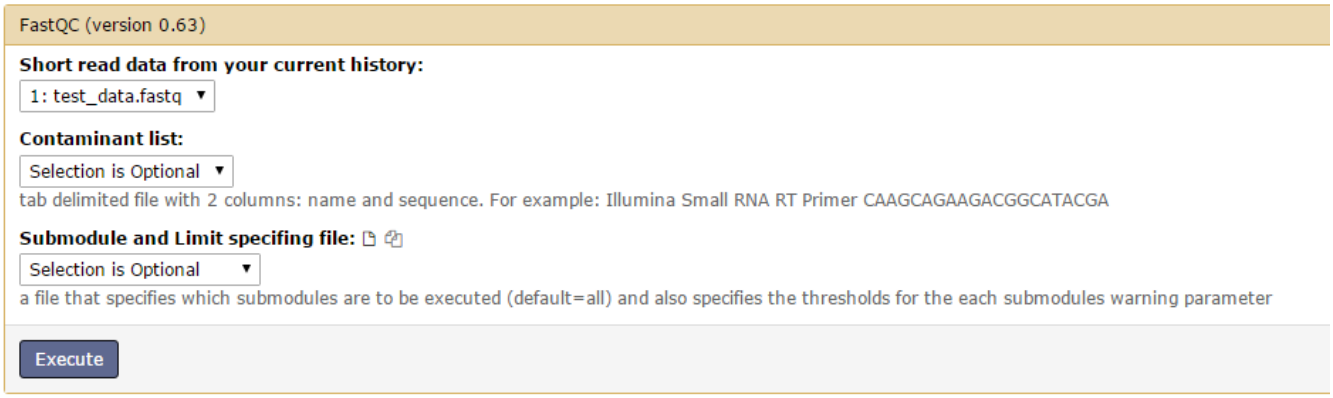
Raw data quality control

For visualization of the quality of your sequenced reads we use FASTQC software

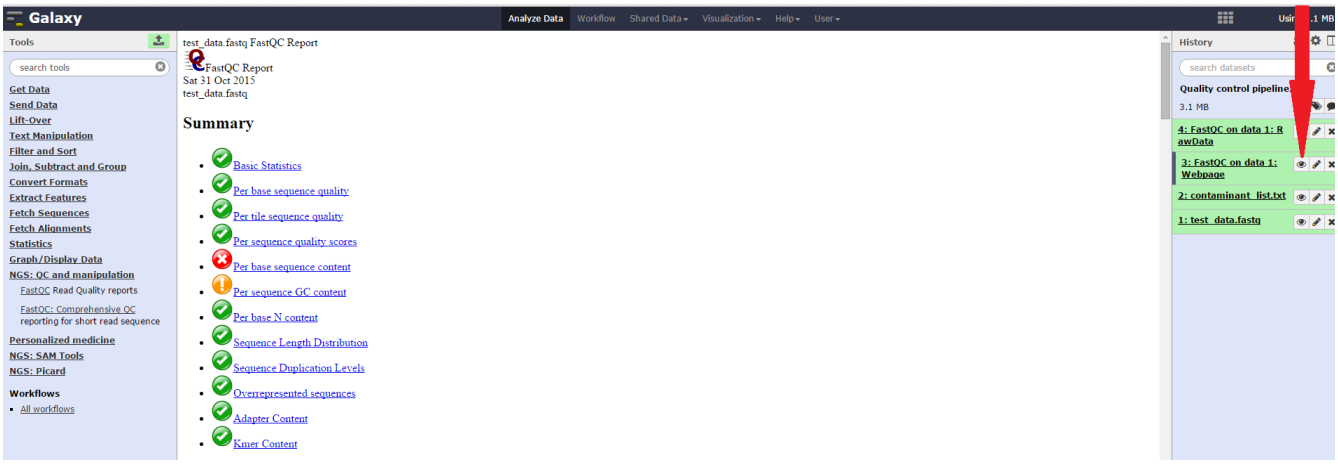
> **Tools**
>
> NGS: QC and manipulation -> FastQC: Read Quality reports

- *Short read data from your current history*: select uploaded fastq file.

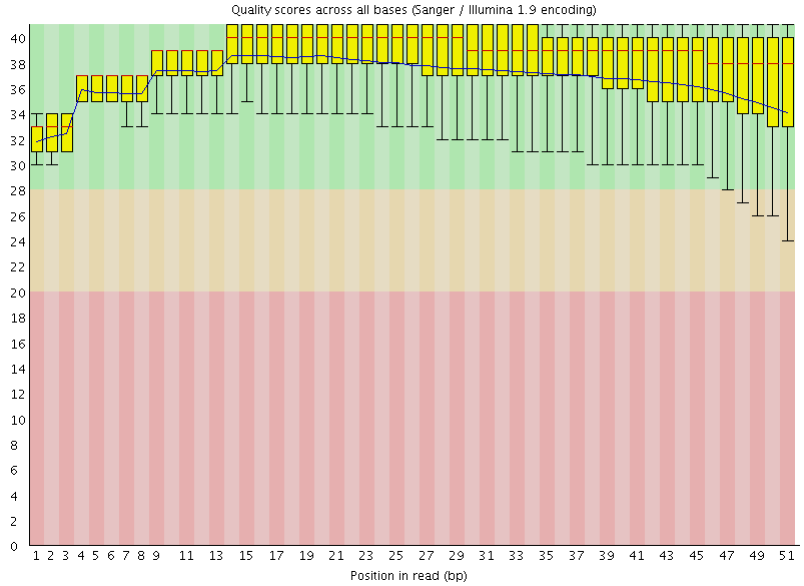
You could also add the "Contaminant list" if you know basic assumptions of library preparation step, or add the list provided us by FASTQC authors (https://github.com/csf-ngs/fastqc/blob/master/Contaminants/contaminant_list.txt).

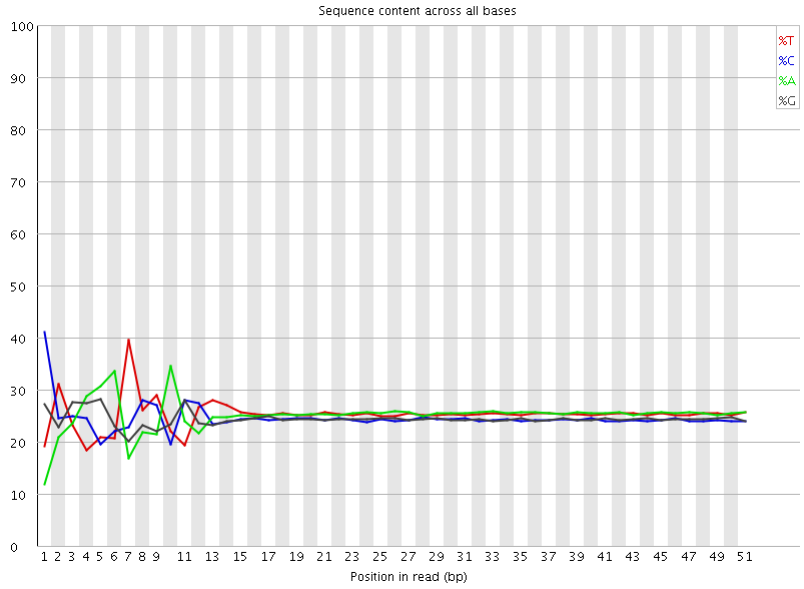


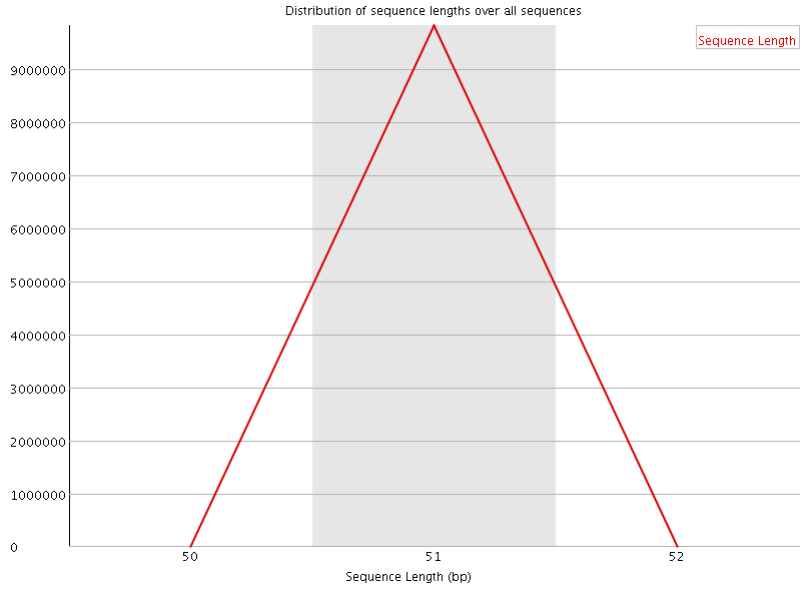
The FASTQC results should be available in your history window (click the eye icon near the name of the history step related to the executed FastQC run):



Some of these results are presented below:







In order to learn how to interpret these charts, please refer to the official help on FastQC and interpretation of QC metrics.

Quality trimming and filtering sequences

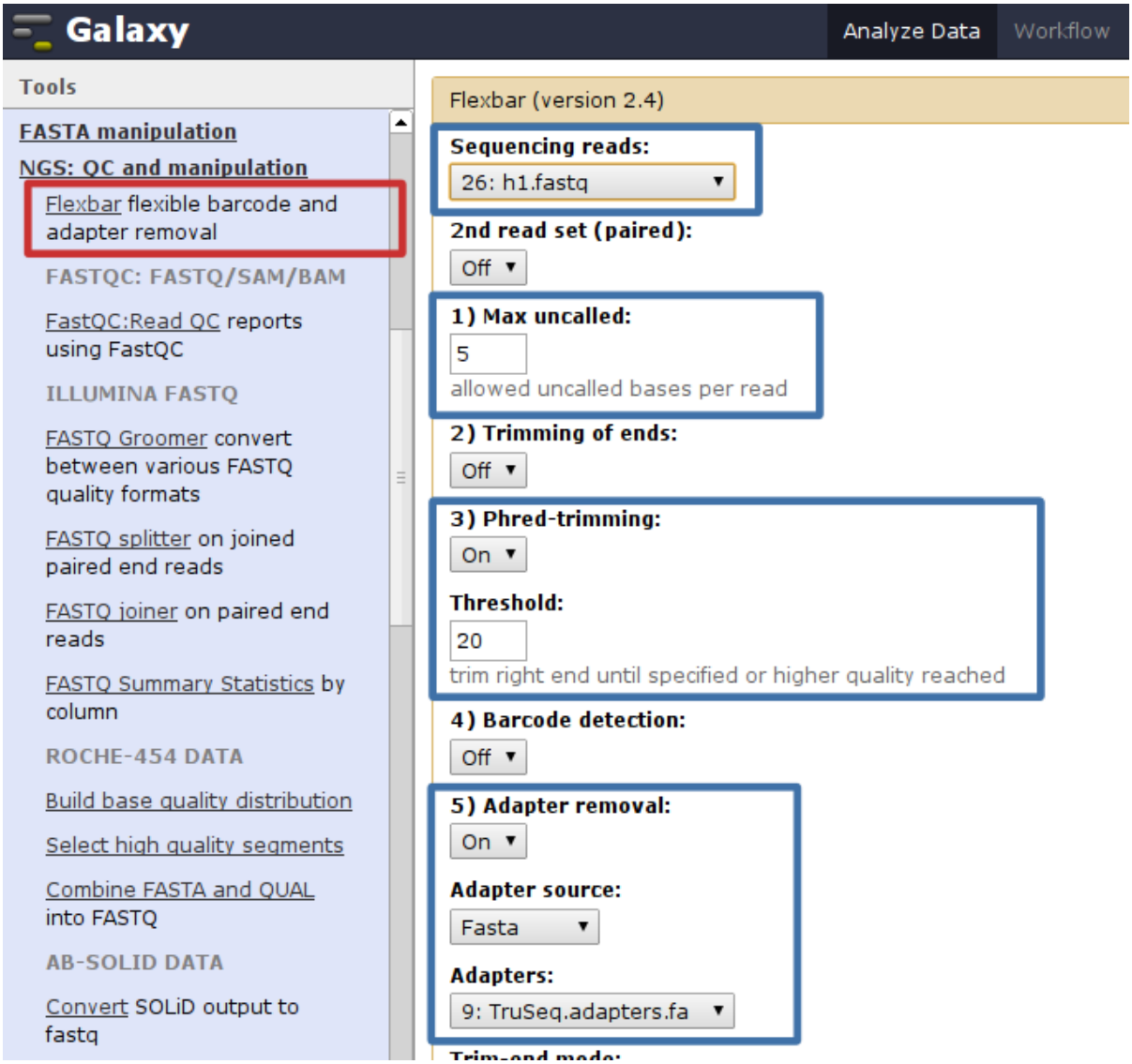
Flexbar software demultiplexes barcoded runs and removes adapter sequences. Moreover, trimming and filtering features are provided. Flexbar increases read mapping rates and improves genome and transcriptome assemblies. It supports next-generation sequencing data in fasta/q and csfasta/q format from Illumina, Roche 454, and the SOLiD platform.

Tools

Personalized medicine -> Flexbar flexible barcode and adapter removal

- *Sequencing reads*: select uploaded fastq file (for SE, for paired end (PE) you must select "2nd read set (paired)").
- *1) Max uncalled*: 5 allowed uncalled bases per read.
- *3) Phred-trimming*: ON and Threshold: 20.
- *5) Adapter removal*: ON, Adapter source: Fasta and Adapters: choose uploaded adapters file in FASTA format -> Nextera adapters and TruSeq adapters.
- *6) Trimming to length*: ON and Length: 50 trim reads to certain length from right.

Then please click the Execute button in order to start a Flexbar run. You will need to wait some time for completion, as the tool has a relatively high consumption of resources.



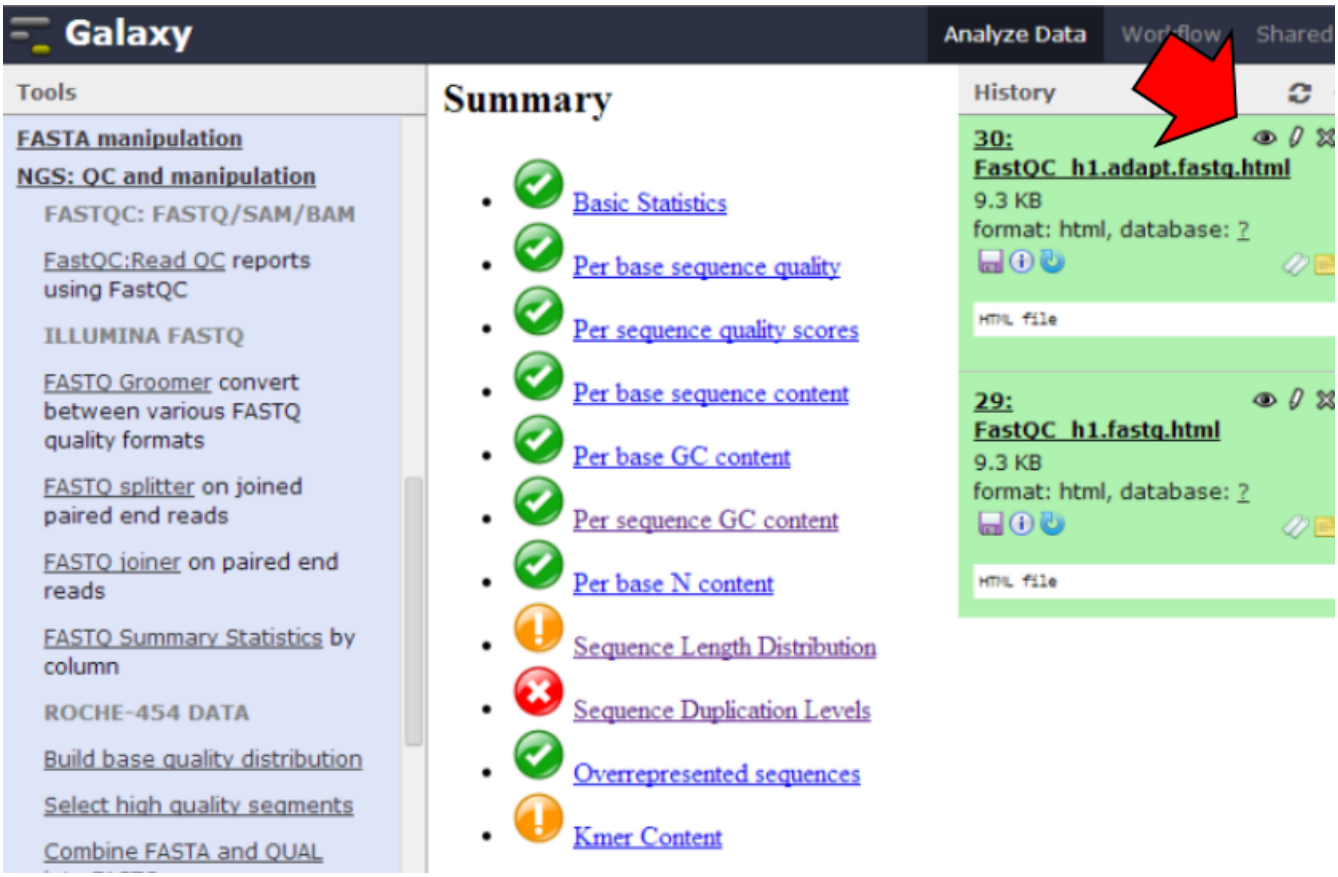
The summary of adapter removal and trimming is presented in the Flexbar output:

```
Adapter removal statistics
==========================
Adapter:              Overlap removal:    Full length:
TruSeq Universal Adapter  2711062            0
TruSeq Adapter, Index 1  2754956            0
TruSeq Adapter, Index 2  3342               0
TruSeq Adapter, Index 3  567                0
TruSeq Adapter, Index 4  151                0
TruSeq Adapter, Index 5  34                 0
TruSeq Adapter, Index 6  45                 0
TruSeq Adapter, Index 7  39                 0
TruSeq Adapter, Index 8  7                  0
TruSeq Adapter, Index 9  14                 0
TruSeq Adapter, Index 10  4496              0 -> specified for this sample
TruSeq Adapter, Index 11  7                 0
TruSeq Adapter, Index 12  12                0
TruSeq Adapter, Index 13  14                0
TruSeq Adapter, Index 14  8                 0
TruSeq Adapter, Index 15  3                 0
TruSeq Adapter, Index 16  2                 0
TruSeq Adapter, Index 18  2                 0
TruSeq Adapter, Index 19  8                 0
TruSeq Adapter, Index 20  0                 0
TruSeq Adapter, Index 21  0                 0
TruSeq Adapter, Index 22  1                 0
TruSeq Adapter, Index 23  0                 0
TruSeq Adapter, Index 25  12                0
TruSeq Adapter, Index 27  4                 0

Min, max, mean and median adapter overlap: 1 / 58 / 2 / 2

Output file statistics
======================
Read file:              FlexbarTargetFile.fastq
  written reads         10124610
  skipped short reads   7265

Filtering statistics
====================
Processed reads                  10148633
  skipped due to uncalled bases     16758
  trimmed due to low quality       748711
  short prior adapter removal        2046
  finally skipped short reads        7265
Discarded reads overall            24023
Remaining reads                  10124610   (99% of input reads)
```

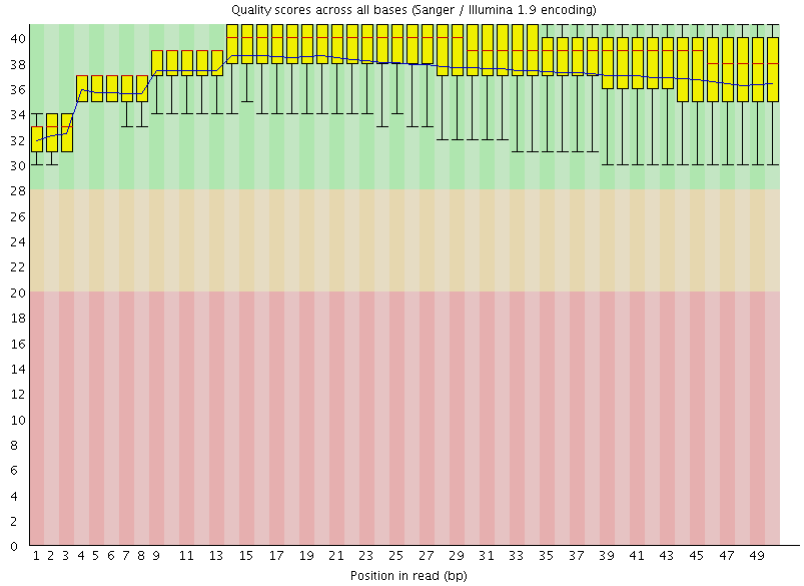
Trimmed data quality control (optional)

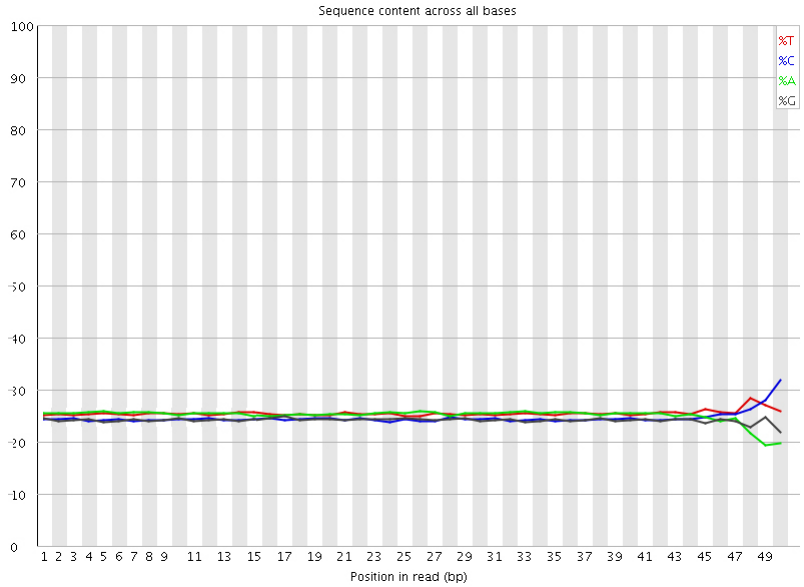
To check quality trimming and filtering results, and to compare our raw data and data after the trimming, we must repeat the "Raw data quality control" step for trimmed data.

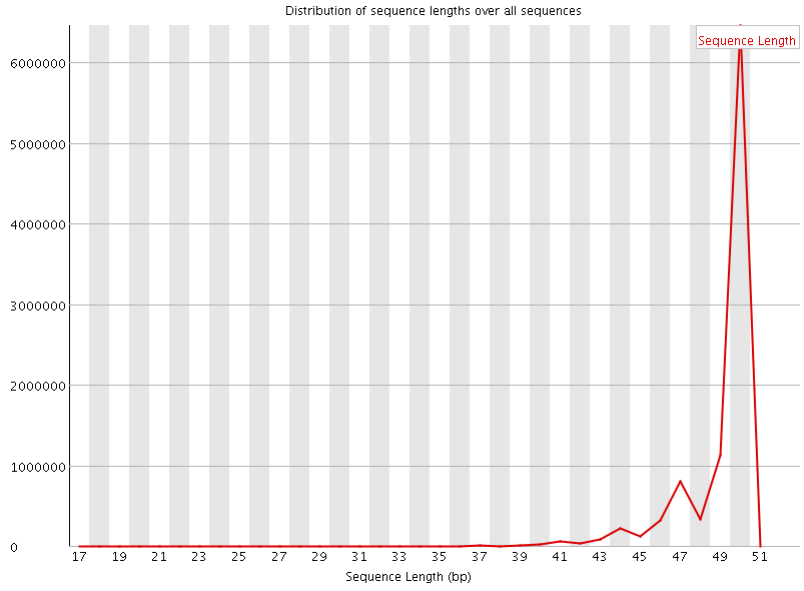
The results for trimmed data:



Some of these results are presented below:


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)


Sequence content across all bases
%T
%C
%A
%G
Position in read (bp)


Distribution of sequence lengths over all sequences
Sequence Length
Sequence Length (bp)

We could observe that "Per base sequence quality" figure (to the left) is better in comparison to the figure before trimming. Some of the reads with quality lower than 20 were excluded from the analysis, especially that at the end of reads (46-50 base pair). Similar changes can also be seen at "Sequence length (bp)" figure before and after trimming (to the right). The main part of reads remained without trimming (50 bp) but from some of them adapters (probably those which lower than 45 bp) and reads which have poor quality (46-49 bp) were cut off. The middle figure, "Per base sequence content" shows the mean percentage of nucleotides (A, C, G, T). We could observe that after trimming, mainly of adapters which have repetitive sequences, there was equalization of the mean nucleotide percentage in comparison to reads before trimming.

Closing remarks

This tutorial covers trimming adapters and low quality reads using Flexbar. Additionally, we used FastQC at every step of analysis to control quality of our raw reads. It should be the first step before you proceed with the alignment of your raw sequences to the reference Short reads alignment to human reference genome using BWA or Short reads alignment to mouse transcriptome using Tophat2.