# Practical: Quality control of long-read data

## Background

Quality control of sequencing data is pivotal to ensure high quality downstream analyses such as genome assembly, typing, and annotation. If the quality of the sequencing data is low or if it contains contaminants the conclusions drawn based on downstream analyses can be wrong.

In this practical you are going to work with sequencing depth and base calling errors.

## Overview

Step 1: NanoPlot

Step 2: Quality filtering with Chopper

Step 3: NanoPlot

## Step 1: Nanoplot

[NanoPlot](https://github.com/wdecoster/NanoPlot) is a tool to summarize and visualize Nanopore data. It is installed in the conda environment **fastp,** activate this and type ‘NanoPlot -h’ to see the help message.

1. Which input types does NanoPlot take?
2. Consider which options might be useful to include and what to call your output\_dir
3. Choose a Nanopore fastq.gz file
4. Run NanoPlot on the file you chose. This should take ~5 min or just enough time for a short coffee break.

EXAMPLE:
NanoPlot –fastq\_rich [input\_reads.fastq.gz] -o [output\_dir]

1. Look in your new NanoPlot folder
	1. Open the NanoPlot-report.html in a browser
	2. What is the sequencing depth of the genome if this is an E. coli?
	3. Which Q-score has this data been filtered on?
	4. Do you think this data could be filtered further?

## Step 2: Quality filtering with Chopper

Lets try to filter the data on quality score. For this you will use [Chopper](https://github.com/wdecoster/chopper), a filtering tool for long read sequencing data. Chopper is also installed in the conda environment **fastp,** activate this and type ‘chopper -h’ to see the help message.

EXAMPLE:
gunzip -c reads.fastq.gz | chopper -q 10 -l 500 | gzip > filtered\_reads.fastq.gz

1. Use Chopper to filter the reads on a quality score you choose. This should take ~2.5 min.
	1. Remember to consider what to call your output
2. What does Chopper give you of information and can it be used to infer a new sequencing depth after filtering?

## Step 3: Nanoplot again..

1. Run NanoPlot again and compare your result from before and after filtering.
	1. What is the mean read quality before filtering
	2. What is the mean read quality after filtering
2. At what depth do the filtered reads cover your E. coli genome?
3. What is most important, high quality or high sequencing depth?

## Extra: Sequencing depth

It is important to consider sequencing depth when working with long-read data. A high sequencing depth can help reduce errors and lead to more accurate consensus sequences but if the depth is too high it can lead to unnecessary long processing times.
In this exercise you should find out how many bases are in a given fastq.gz file without using NanoPlot.

1. Choose a fastq.qz file from the Nanopore E. coli data.
2. Look at the size of your gzipped FASTQ file and compare it to the number of bases in it.

du -sh [input\_reads.fastq.gz]