# Project scope and objects

## Aim

To determine the lowest possible input coverage of Oxoford Nanopore Technologies (ONT) sequence data that will maintain fidelity across multiple downstream applications.

### Objectives

1. Perform quality control on sequence read datasets using FastQC and MultiQC
2. Adjust quality of reads using NanoFilt (repeat obj. 1 & 2 until quality is satisfactory)
3. Align sequence reads of each dataset to the appropriate reference genome using Minimap2.
4. Compare called variants across datasets to ensure accuracy at lowest coverage. In-house scripts and/or programs will be written for this objective.

# Materials and Methods

The input data for this analysis are the files generated through ONT sequencing. The DNA in this analysis is a mixture of multiple individuals to yield the highest possible quality DNA for sequencing.

The first step in the analysis is to ensure the quality of the data is as good as possible. When bases are called, the base itself (A, C, T or G) is recorded as well as the surety of the sequencer that the base called is true, i.e. the quality of the base call. This the essence of a FASTQ file, which combines FASTA (molecular sequences) with quality scores. Straight off the sequencer, FastQC1 can interpret and summarize these quality scores for each resulting FASTQ file. Ease of interpretation is achieved with MultiQC2 which can collate the resulting quality descriptions into a single view. Low quality bases or reads need to be removed, as erroneous base calls will influence downstream analysis. For this, NanoFilt3 will be used, as it is designed to work with long reads. FASTQ files are large and contain many reads, so the quality files are averages across all reads contained within. The reports will therefore change as the individual reads are manipulated and quality needs to be tested again with the same two tools. If necessary, more adjustments can be made. Depending on the number of iterations required to achieve the desired quality, this step will last approximately 14 days with about 8 hours of hands-on analysis.

Once the quality is satisfactory, the data needs to be aligned to the appropriate reference genome. The more data there is, the longer this will take. However, besides the amount of data, alignment speed is also determined by the length of the read. Shorter reads have more possible mapping sites in repetitive genomes such as those of mammals. Longer reads will therefore take less time to map, as they tend to be more unique. The alignment will be done using Minimap2. This process is estimated to take a week once initiated and will require approximately 8 hours of analytical time to start.

The alignment will generate sequence alignment map (SAM) files. They contain the reads themselves as well as relevant information about where the read mapped to the reference genome. As can be expected, these files are quite large. To ease the storage burden, they are converted to BAM (binary SAM) before moving on. Other ease-of-use steps include sorting and indexing the BAM files. All of these steps will be done using samtools4 and will take two ours of analytical time with a day to finish after that.

Generally, the next step will be to call and annotate the variants. However, given the fact that these BAM files are an amalgamation of individuals instead of a single one, this step won’t be possible. The files will therefore need to be analysed using custom scripts/programs that will be written during the course of this analysis. This will all be hands-on analysis time and is estimated to take approximately two weeks.

Finally, a document will be generated summarising the findings. The three datasets will be compared with another in terms of number of variants detected as well as number of variants shared between all three.

Table 1: Analysis time estimation with number separate columns for hands-on analysis (working time) and total running tme.

|  |  |  |
| --- | --- | --- |
| Step | Working time required (h) | Running time required (d) |
| Quality control | 8 | 14 |
| Sequence alignment | 8 | 7 |
| S/BAM processing | 2 | 1 |
| Variant analysis | 112 | 14 |
| Admin | 13 | 1 |
| *Total* | *143 hours* | *37 days* |