SO YOU SEQUENCED SOMETHING .....

Now what?

01/02/2024

ONT DATA

ONT raw files are output as .fast5 files
Illumina raw data files are output as .fastq files

FAST5 based on the hierarchical data format HDF5 format which enables storage of large and comples data

FASTQ A FASTQ file is a text file that contains the sequence data from the clusters that pass filter on a flow cel

FAST5 IN MORE DETAIL

HDF5 files uses a chunking strategy for storing and accessing multidimensional data.

- Advantage is that instead of unzipping a large file to access one component of the file, you can specifically access the component you are interested in
- FASST5 refers to the structure imposed on the HDF5
- There are 3 main branches of data stored in the fast5, Analysis, Raw, and UniqueGlobalKey.
    - Raw stores the raw signal levels,
        - Analysis stores analysis results such as base-calls, signal correction and segmentation information.

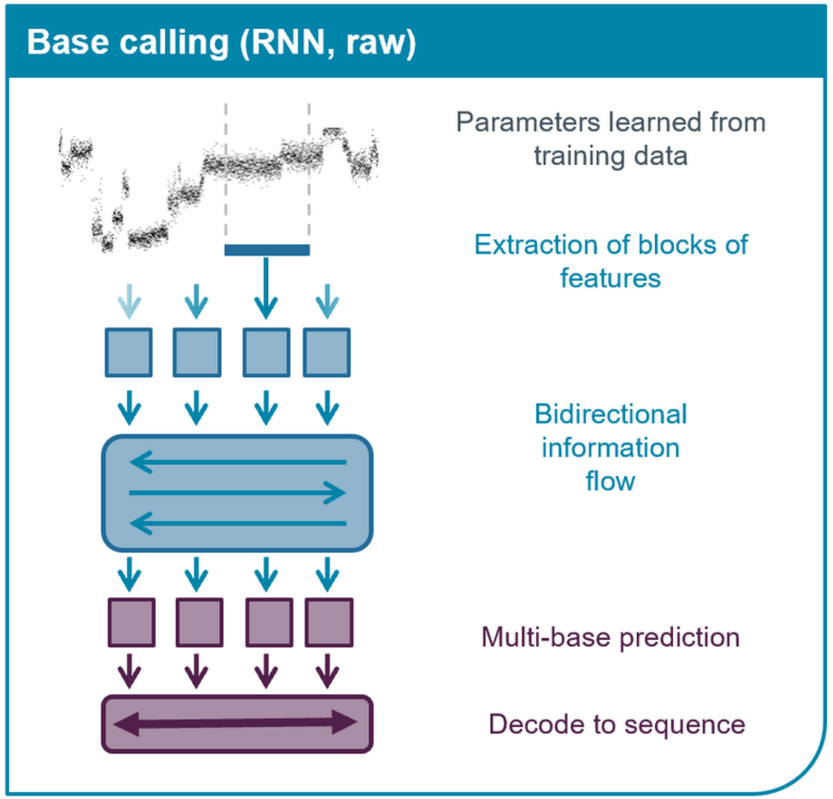
FASTQ FILES

For each cluster that passes filter, a single sequence is written to the corresponding sample's R1 FASTQ file, and, for a paired-end run, a single sequence is also written to the sample's R2 FASTQ file. <u>Each entry in a FASTQ files consists of 4 lines:</u>

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

FAST5 - WHERE TO NEXT?

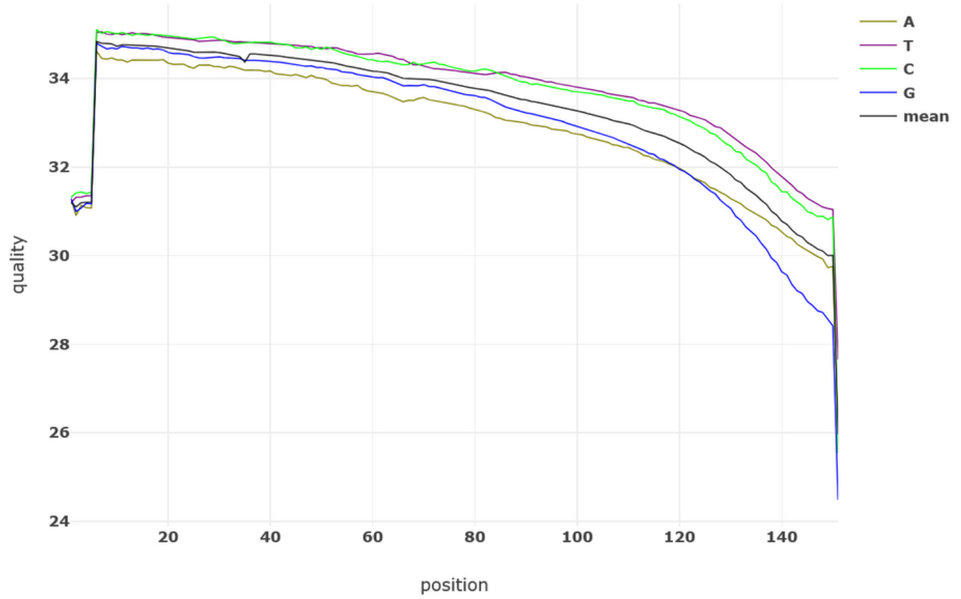
- FAST5 files are large and do not contain nucleotide sequences
- We can convert FAST5 files to SLOW5, POD5 to reduce the size for long term storage
- We then use a **basecaller** tool such as **Guppy, Bonito or Dorado** to translate the electrical signals to nucleotide sequences



QUALITY CONTROL

Illumina - Fastp package

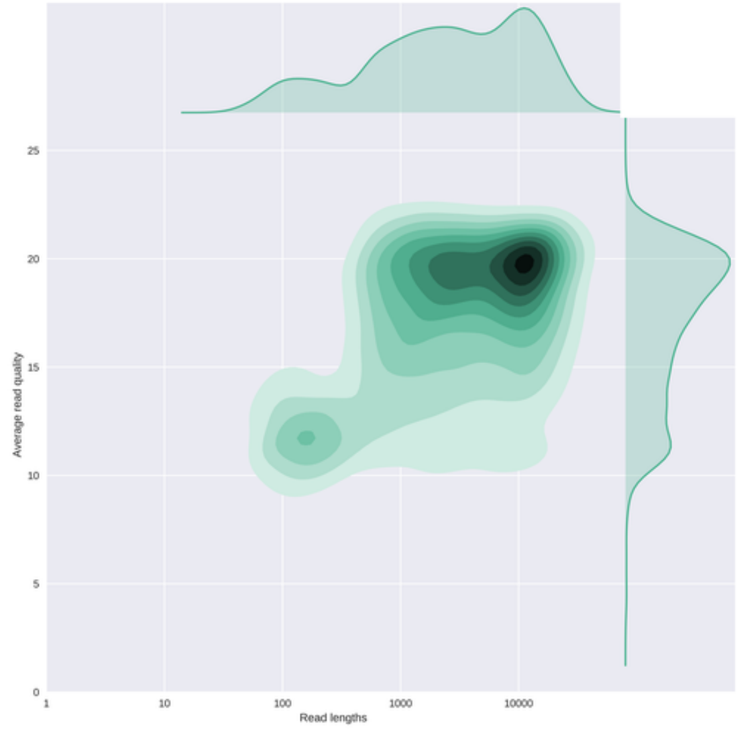
- Insert size
- Duplication rate
- Quality per base/per position
- Overrepresented sequences
- Presence of adapters



QUALITY CONTROL

ONT - Nanopack

- Read length
- Quality per base/per position
- Read length distribution



QUALITY CONTROL

You can increase the accuracy and reliability of your data and decrease **error** by ensuring you have only analyse **good quality** sequencing data.

During quality control, we should remove reads that are too short, too long, have too low Phred scores or still have adapters attached to them.

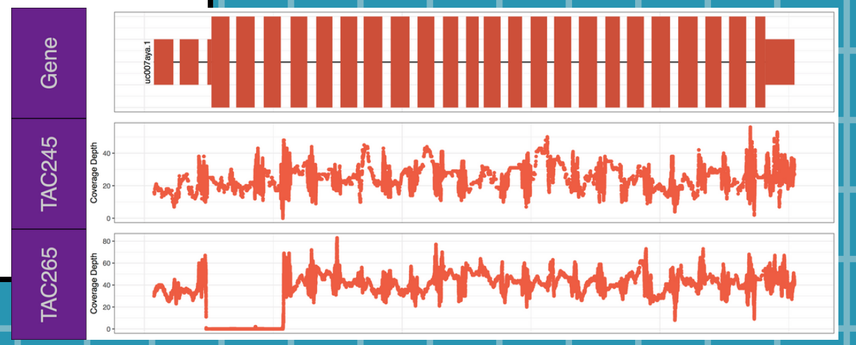
QUALITY SCORE (Q-SCORE)

Represents the confidence or accuracy of a base call in DNA sequencing data. It quantifies the probability of an error in the base call. Higher quality scores indicate higher confidence in the base call.

GENOME COVERAGE

Genome coverage refers to the **proportion or percentage** of a reference genome that is covered by sequencing reads.

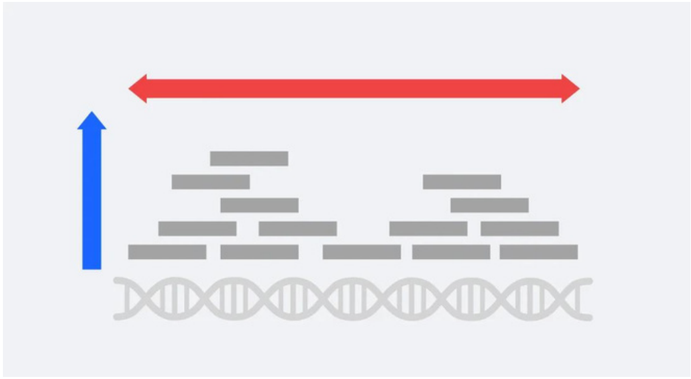
Higher genome coverage indicates that more regions of the genome have been sequenced, which can be important for various downstream analyses such as variant calling and genome assembly



DEPTH OF COVERAGE (READ DEPTH)

refers to the **average numbe**r of sequencing reads that align to a particular position in the genome.

Higher depth of coverage provides greater confidence in the accuracy of base calls and increases the ability to **detect variants** or mutations.



TASK

- Fork the github repository for the course
- Create a SLURM script and fetch a dataset you want to work with using SRA toolkit
- Check the quality of your dataset, using both statistical tools and visual tools in NanoPack
- Fetch a HEV sequence (SLURM) and compare it to your HEV database using DIAMOND.