Introduction to Oxford Nanopore Technologies (ONT) sequencing bioinformatics pipeline

Dr John Tembo

We will record these sessions and put them online so you can refer back to them later on

We will also put the slides up online so you can access the notes (links and image credits)

Let’s Download the data (5 fast5 files)

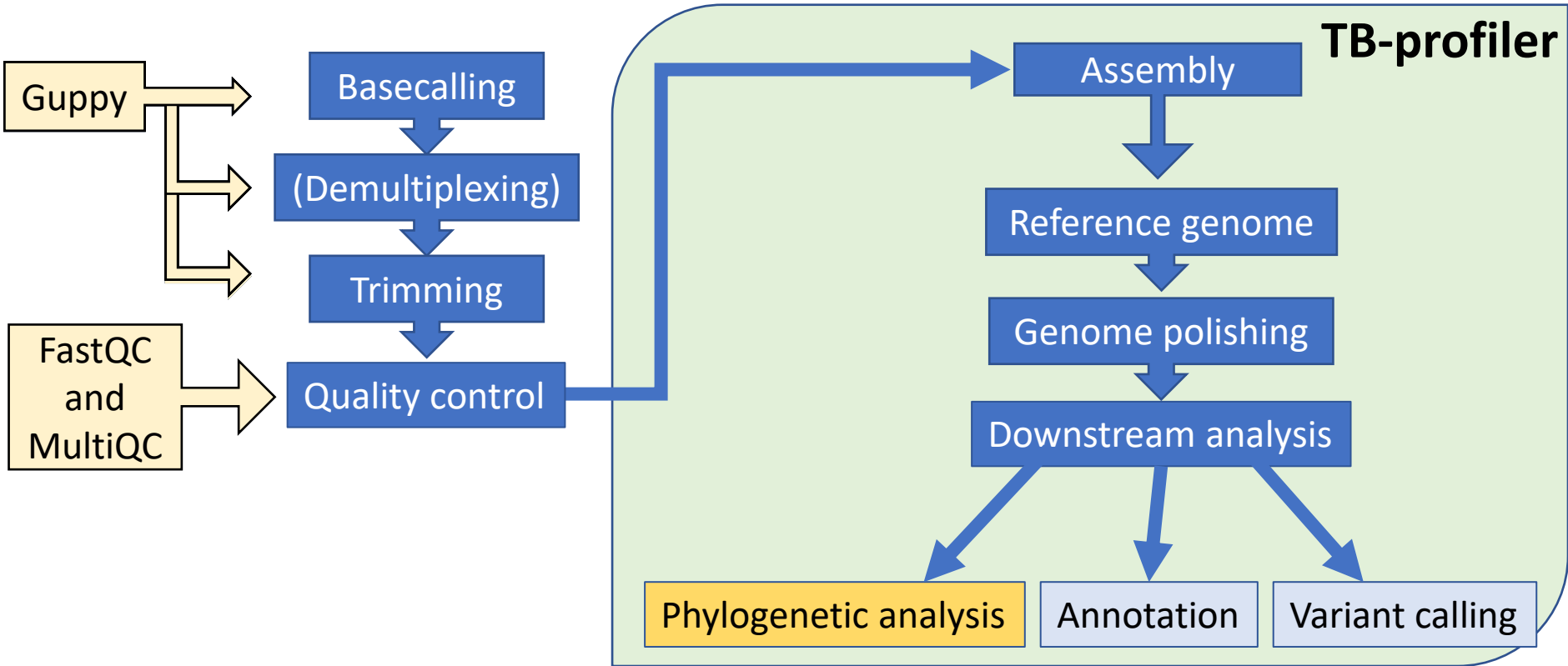
Aims of this session

- Learn how to install and run commands using the Windows version of Guppy
- Learn how to basecall ONT raw files (.fast5) into .fastq files
- Learn how to combine .fastq files into one file (concatenation)
- Check the quality of your sequencing files using FastQC

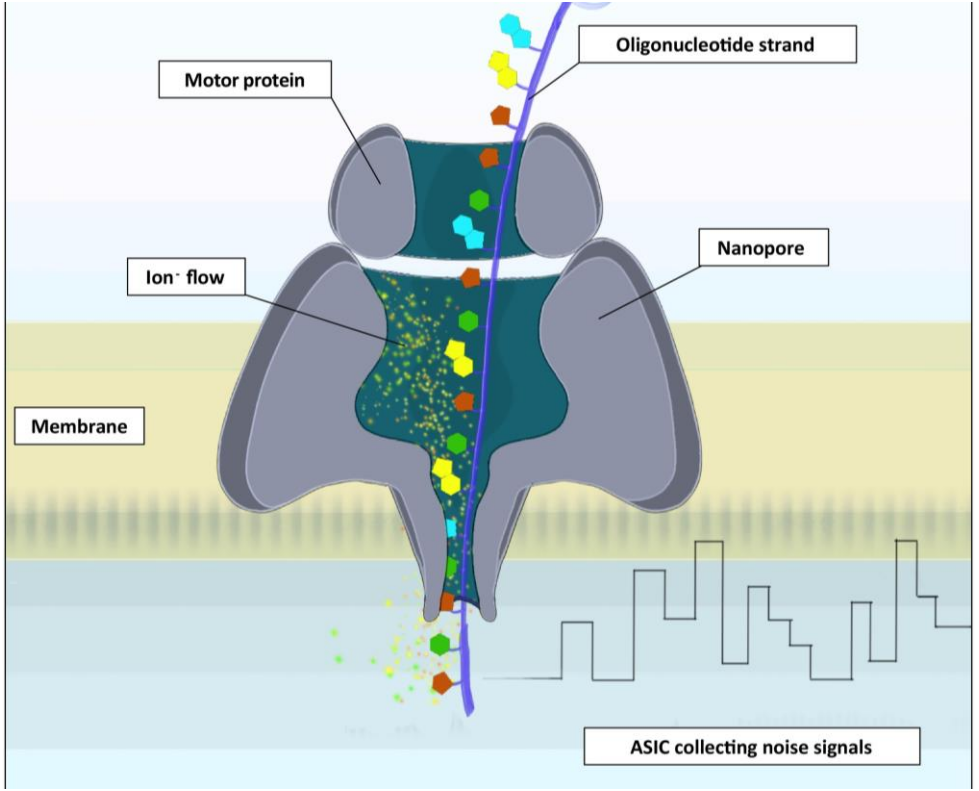
Before you start

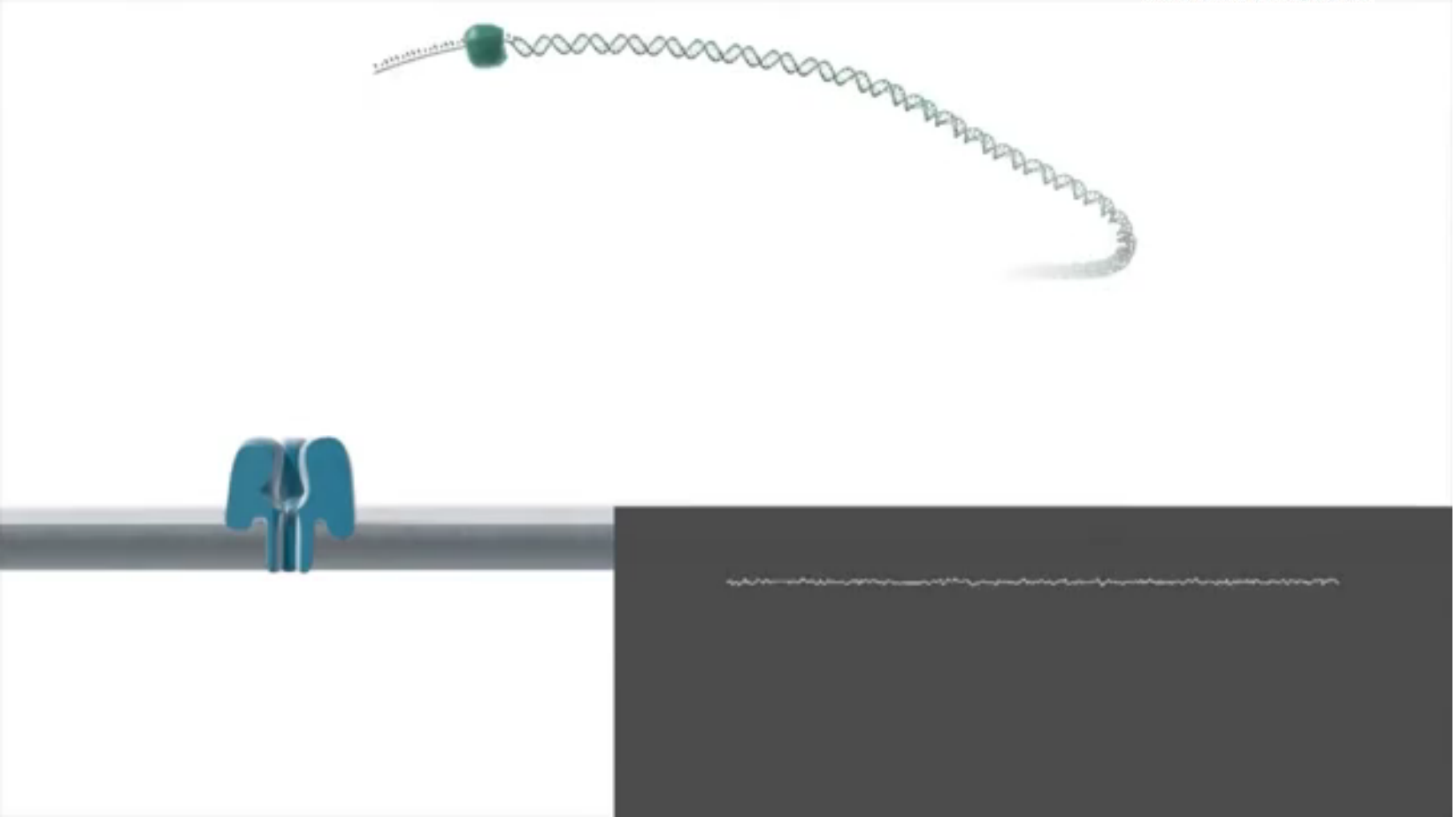
- What operating system are you using? Be aware that most bioinformatics tools only run on Linux and MacOS
- For Windows 10 and above users they can use Windows Sub Linux System (WSL)
- How powerful is your computer? Some programmes require a minimum amount of memory (RAM ≥16 GB), Processor (how many CPUs), Disk Space
- Power – processing huge files can take a long time (days!) so you will need to make sure you have a continuous power supply throughout, otherwise you will have to start again
- Internet– CLI's don't generally need internet to run, but if you are uploading to cloud based server, downloading and installing bioinformatics tools you need a good connection

ONT MTB Bioinformatics Pipeline



- Fast5 files are used by the MinKNOW instrument software to store the primary raw sequencing data
- Also known as squiggle data, chromatogram peaks or ‘raw data’
- Guppy is a basecalling software that converts the primary raw data to Sequences bases data (A, G, T, C)





Gfycat

Fast5 files

- File extension .fast5
- Direct output from ONT primary sequencing data
- Large files, which store complex data
- Binary, cannot be opened with a text editor (e.g. notepad)
- Contains the ‘squiggles’ (disruption in electrical signal when a base moves across the pore)

https://www.frontiersin.org/articles/10.3389/fgene.2019.01332/full

Some tips before we start

- When you are writing the code script Never use Microsoft word (it has hidden characters)
- At least use simple text editor (e.g. Notepad or TextEdit)
- Recommended install Notepad++ (Windows-Users) or (Xcode for MacOS)
- It’s always useful to write down your code (and annotate with #) so you know what to do when you come back to it later
- What is your working directory?
- Can you set it to your home?
- And How to check what is your work directory?

Some Tips before we Start

- In any command you will have to provide **< Path>** of the file / folder : This is the long address of the file or folder location with all the upper directories in the hierarchy separated by a
  - forward slash “**/** ” in MacOS and Linux

  E.g. /Users/sylviarofael/miniconda3/share/tbprofiler/tbdb.fasta
  - Backward slash “ **** “ in Windows

  E.g.

  “C:\Users\WSL_shared\ONT_workshop\MTB_fastq\barcode01*.fastq”
  - * is a wildcard that can replace text in the name
- When to use **" "** if the path has spaces in the file name this can break the Syntax so in this case we can put it between vertical quotation marks
- “**.**” means current directory

Let’s Install Guppy!

How code is structured (Syntax)?

- **Programme command** – the name of the programme you want to execute the action (e.g. guppy_basecaller)
- **Argument(s) or parameter(s)** – instructions to execute the command in a way you need separated by spaces **(Required) –i (--input) –s (--save_path) –c (--config)**
- **Flag(s)** – enable you to specify options **(optional) you can see by running the tool with "--help" to see the available flags and options.**
- **Option(s)** – used to explain to the programme what argument/parameter you want it to execute (e.g. filter out all files with less than 4000 reads)

How to deal with errors

- Errors are common, especially when you are learning, so don’t panic!
- Read through the error – they are designed to help you identify the problem
- Someone will have come across it before: check forums on e.g. Stack Overflow and Biostar and GitHub
- Google! You can copy and paste the error

Guppy

Guppy is a data processing toolkit that contains ONT basecalling algorithms, and several bioinformatic post-processing features

It is integrated with MinKNOW, the ONT device control software

Requirements:

- At least 8 GB RAM
- Windows, MacOS or Linux command line

Guppy has three functions:

- Basecaller (converts .fast5 files into .fastq files)
- Barcoder (demultiplexes/separates into barcode files, and trims Adapters and barcodes)
- Aligner (can align basecalled reads to a reference genome)

Today's 'test' data

- Used SQK-RBK004
- Used a R9.4.1 flow cell

Guppy basic code – basecalling (Windows)

```
<path_guppy_basecaller>
-i <fast5_folder_address>
-s <fastq_output_folder_address>
-c dna_r9.4.1_450bps_fast.cfg
```

Address to where guppy basecaller is stored

Input folder and address

Output folder and address

Configuration file and information (flow cell type, kit and fast or high accuracy basecalling)

Guppy example code (Windows) - basecalling

```
“C:\Program Files\OxfordNanopore\ont-guppy-cpu\bin\guppy_basecaller.exe”
-i C:\Users\rekglel\Documents\Sequencing\tutorial_tests\fast5
-s C:\Users\rekglel\Documents\Sequencing\tutorial_tests\fastq
-c dna_r9.4.1_450bps_fast.cfg
```

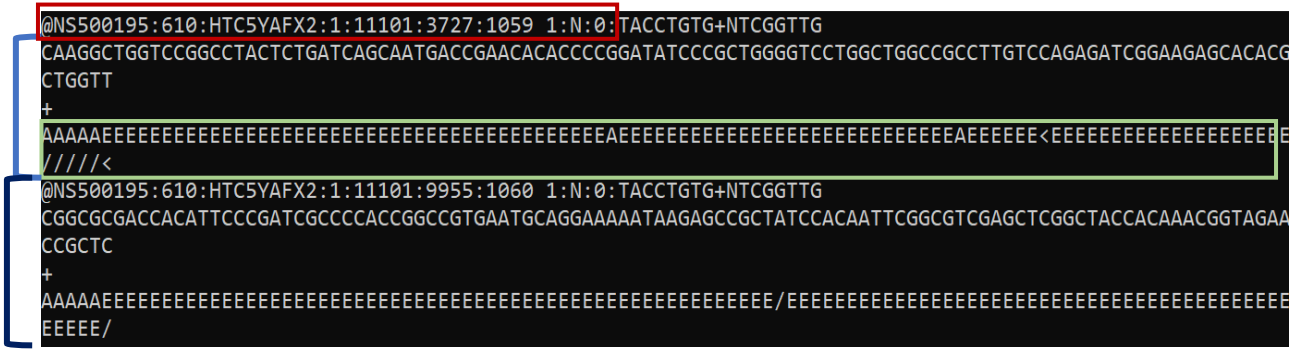
Guppy example code (MacOS) - basecalling

```
guppy_basecaller
-i c/Users/rekglel/Documents/Sequencing/tutorial_tests/fast5
-s c/Users/rekglel/Documents/Sequencing/tutorial_tests/fastq
-c dna_r9.4.1_450bps_fast.cfg
```

Let’s basecall some .fast5 files

Fastq files

- File extension .fastq
- fastq = fasta + quality scores, result of conversion of squiggles to bases
- Can have multiple sequences in a list
- Each sequence read is defined in 4 lines: Header + sequence code+ Name field (optional, usually empty line starts with "+" sign + ‘sequence quality’ information each nucleotide base is corresponding to a code that describes its quality)



Example of Fastq File content containing 2 sequence reads; each read is defined in 4 line red square: Header/Sequence ID, green box is the quality scores corresponding to each base

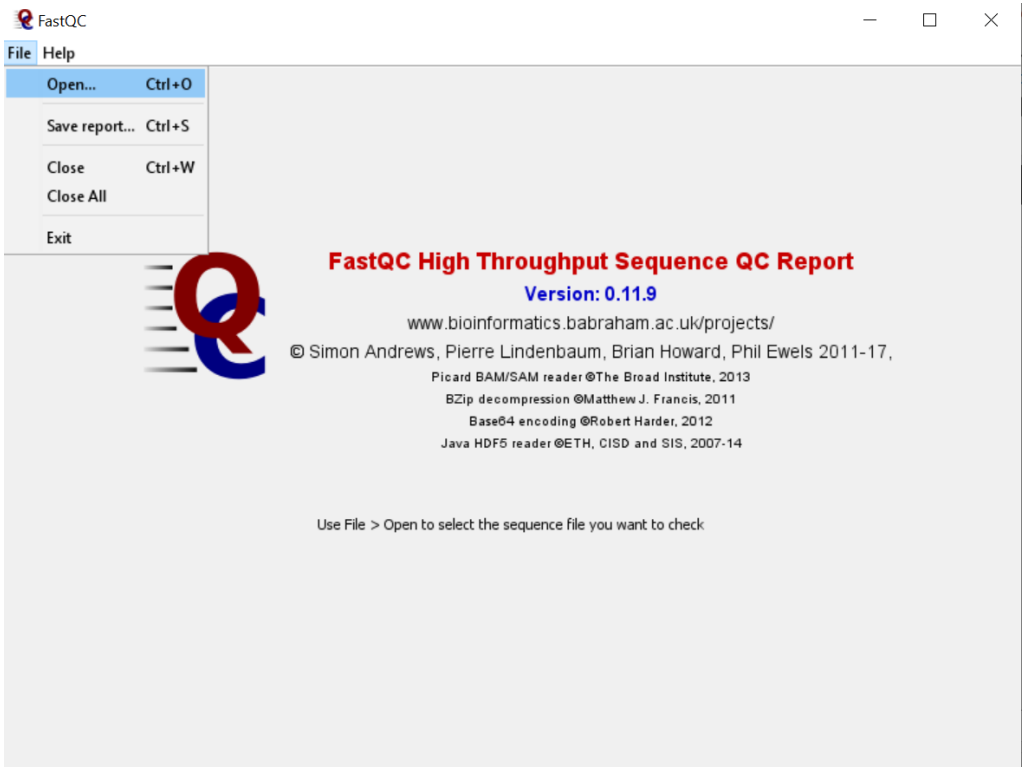
FastQC

- FastQC aims to assess the quality control checks on raw sequence data coming from high throughput sequencing pipelines
- It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.
- Requirements:
  - Java installed (check you have the correct version)
  - At least 8 GB RAM
  - Operating systems: Linux, MacOS or Windows
- A simple way to do some quality control checks on raw sequence data

https://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc

FastQC (GUI)

- Open the ‘run_fastqc.bat’ file
- Click on ‘open’ and add your file



Example FastQC (Windows) command

(navigate to folder with executable file first)

```
C:\Users\rekglel\Documents\Sequencing\Programmes\fastqc_v0.11.9\FastQC\run_fastqc.bat
C:\Users\rekglel\Documents\Sequencing\tb\name.fastq
```

Example FastQC (MacOS) command

(navigate to folder with executable file first)

```
fastqc
c/Users/rekglel/Documents/Sequencing/tuberculosis/name.fastq
```


Thank You