Centre for Clinical Microbiology

Introduction to Oxford Nanopore Technologies (ONT) sequencing bioinformatics pipeline

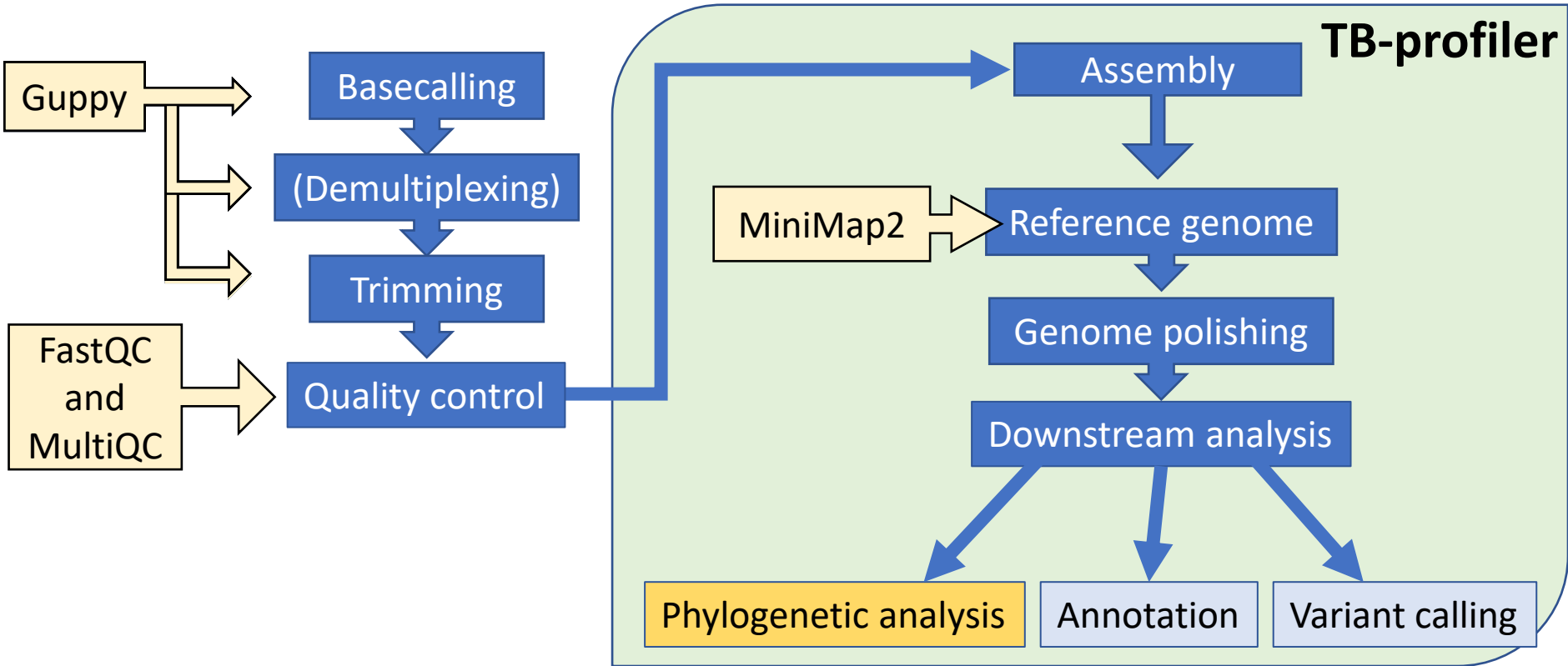
Dr Linzy Elton, Prof Neil Stoker, Dr Sylvia Rofael

Let’s Download the data (5 fast5 files)

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



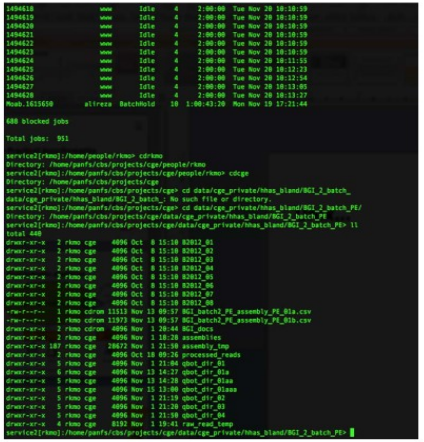
Bioinformatic platforms

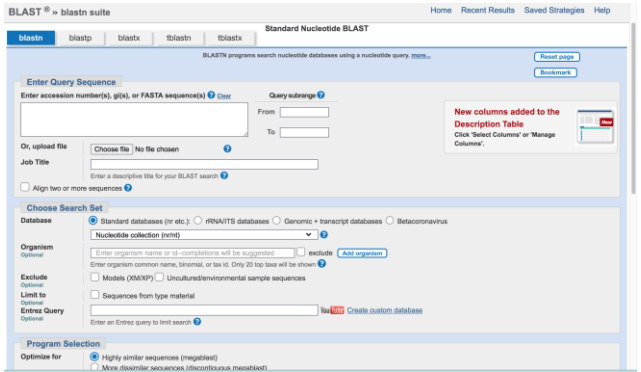
Unix...

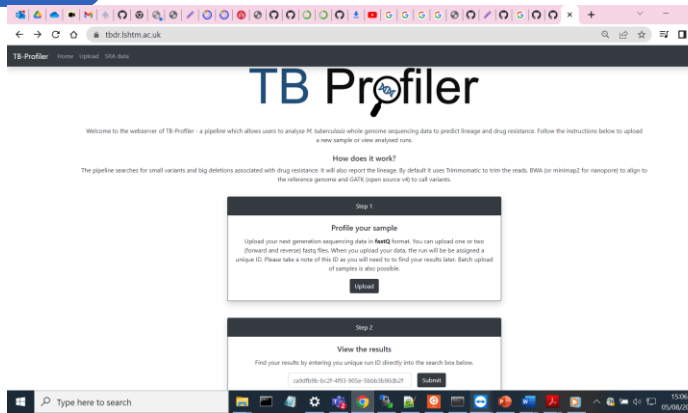
Data Analysis

Bioinformatics tools

- command line
  - Unix
  - Python
- Web-based interface
  - user friendly
  - e.g. NCBI BLAST tools

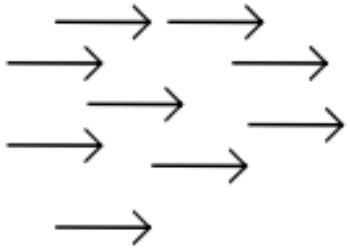


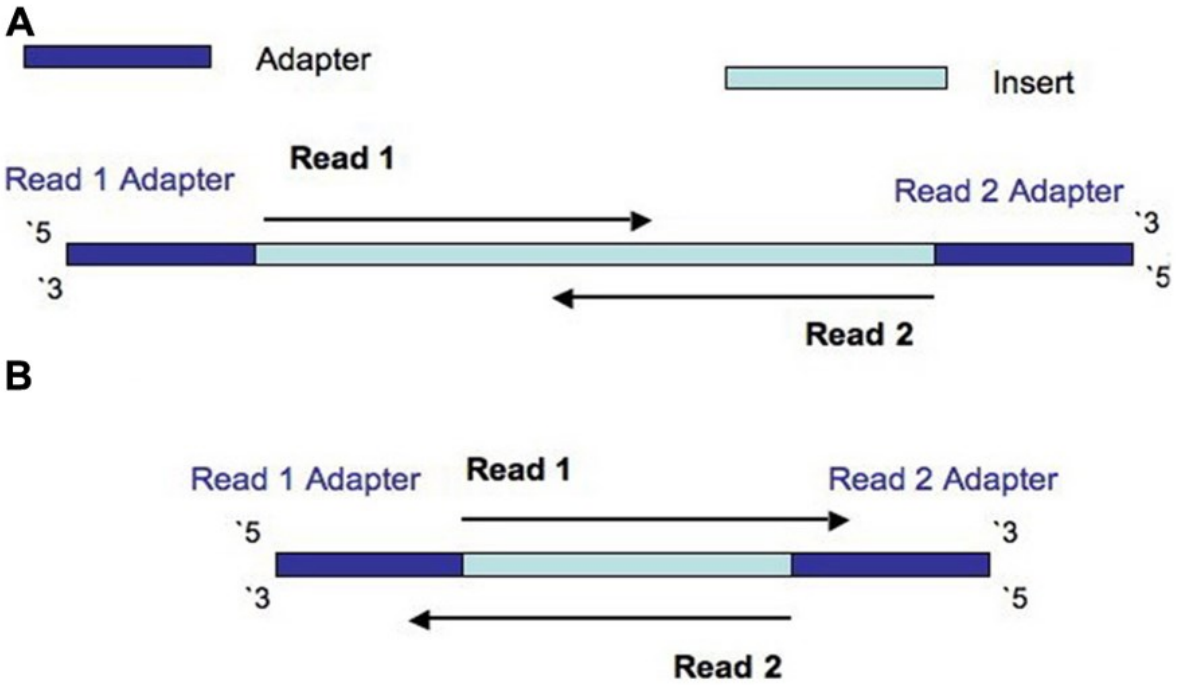




Paired read

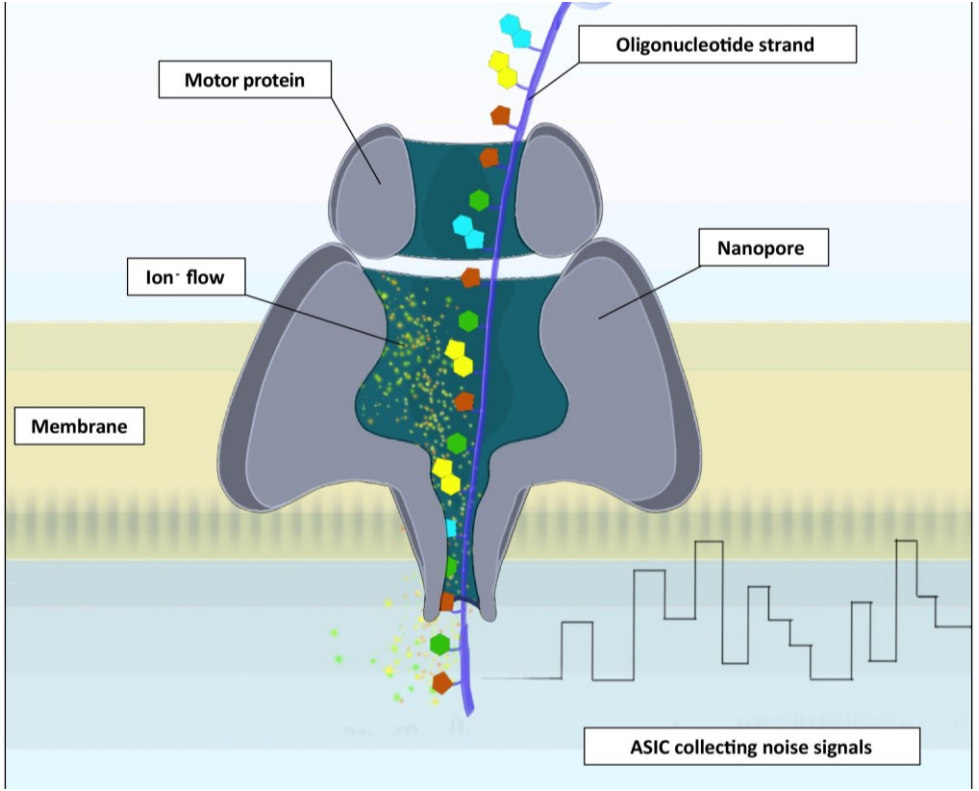
Single end reads



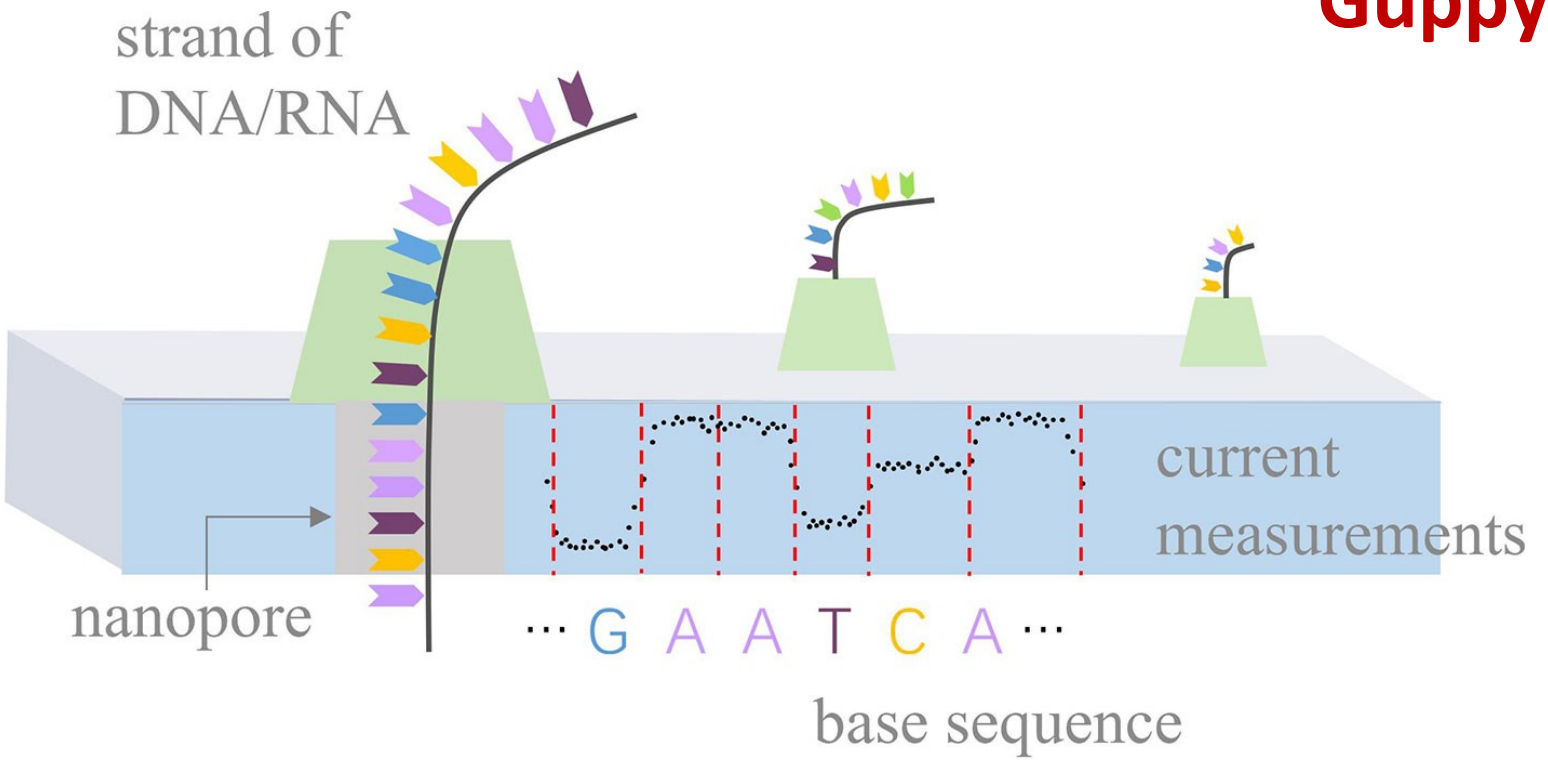


Turner, F. S. 2014. Assessment of insert sizes and adapter content in fastq data from NexteraXT libraries. Front Genet, 5, pp. 5.

- Fast5 files are used by the MinKNOW instrument software to store the primary raw sequencing data (squiggles data, chromatogram peaks ‘raw data’) from ONT sequencing devices
- Guppy is a basecalling software that converts the primary raw data to Sequences bases data (A, G, T, C)



Guppy



- Basecalling
- Quality control
- Alignment/ assembly
- Polishing
- Downstream analysis
- Annotation
- Variant calling

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

Some Tips before we Start

- When your are writing the code script Never use Microsoft word (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?

Use the cd and pwd command

```
cd ~
pwd
cd ..        #-> Go up one directory
mkdir  ONT_workshop
```

You can copy and Paste commands from "**Script_ONT.txt**" shared with you

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 Please refer to Script_ONT.txt (line 47)

Then Download the data (5 fast5 files)

What are Checksums? Please Compare it with the values in the script (line 21)

Open Excel spreadsheet, Copy and paste the checksums provided in the script in column A, the checksums you get in column B, in column C use “Exact” function

= Exact (A1, B1)

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)

Other points

- You can ‘pipe’ code “|” together so that each command is linked together and they follow on
- The path (address) of the tools/programmes can be saved in the $PATH variable of your working environment this allows you to call the tool/programme by its name only without providing its path each time you need to use it
- **https://www.baeldung.com/linux/path-variable**

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

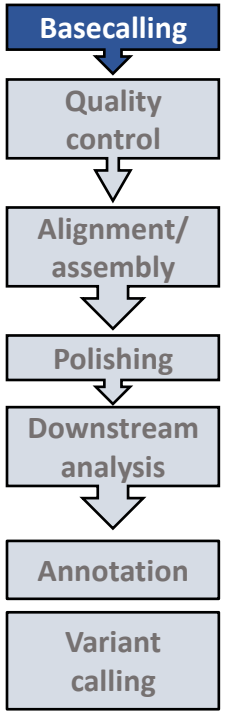
It is integrated with MinKNOW, the ONT device control software

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)

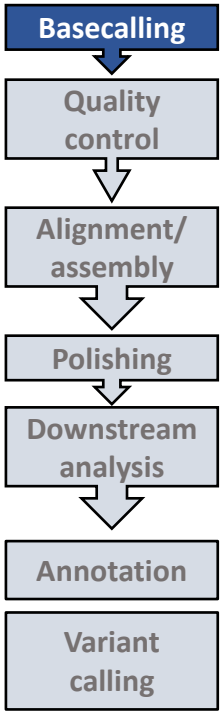


Today's 'test' data

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

Guppy – how to install

- Windows – download straight from community website
- Can also install directly from the command window (directions can be found in the Guppy document on the community website)



Guppy basic code – basecalling (Windows)

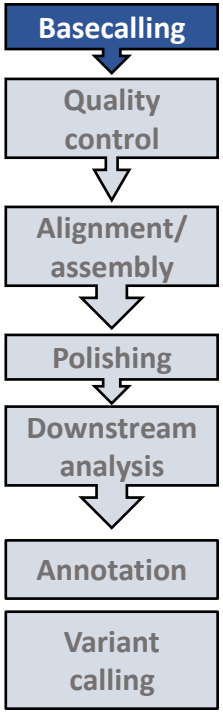
```
<path_guppy_basecaller>
-i <fast5_skip_folder_address>
-s <fastq_output_folder_address>
-c dna_r9.4.1_450bps_fast.cfg
```

Address to where guppy basecaller is stored

Input folder and address

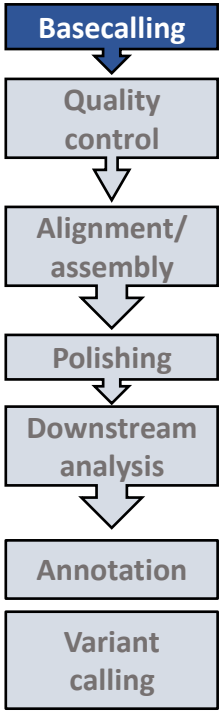
Output folder and address

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 basic code – trimming (Windows)

```
<location_of_guppy_barcoder>
      -i <fast5_skip_folder_address>
      -s <fastq_output_folder_address>
      -r
      -c configuration.cfg
      --trim_barcodes
      --barcode_kits SQK-RBK004
```

Address to where guppy barcoder is stored

Address of input folder

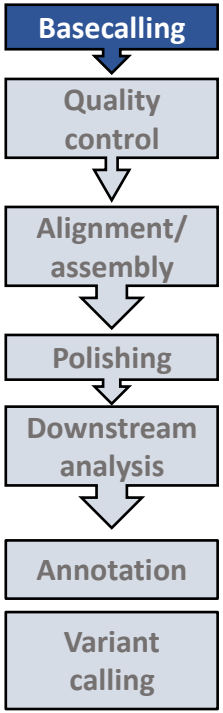
Address of output folder

Configuration file

Command to 'trim'

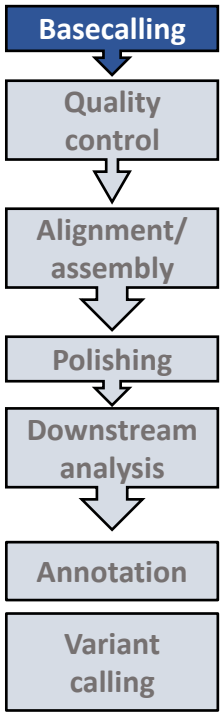
Tell it what type of barcodes used

-r recursive ('do this to all folders within this one') useful for barcode folders



Guppy example code (Windows) - trimming

```
"C:\Program Files\OxfordNanopore\ont-guppy-cpu\bin\guppy_barcoder.exe"  -i C:\Users\rekglel\Documents\Sequencing\tutorial_tests\fastq -s C:\Users\rekglel\Documents\Sequencing\tutorial_tests\trimmed -r -c configuration.cfg –trim_barcodes –barcode_kits SQK-RBK004
```



Let’s run Guppy Please refer to Script_ONT.txt (line 57)

Please refer to Script_ONT.txt (line 126)

Then Download ready 12 fastq files

What are Checksums? Please Compare it with the values in the script

Open Excel spreadsheet, Copy and paste the checksums provided in the script in column A, the checksums you get in column B, in column C use “Exact” function

= Exact (A1, B1)

How many reads in each fastq file?

Let’s look into one fastq file barcode01.fastq use the less command

FASTA files

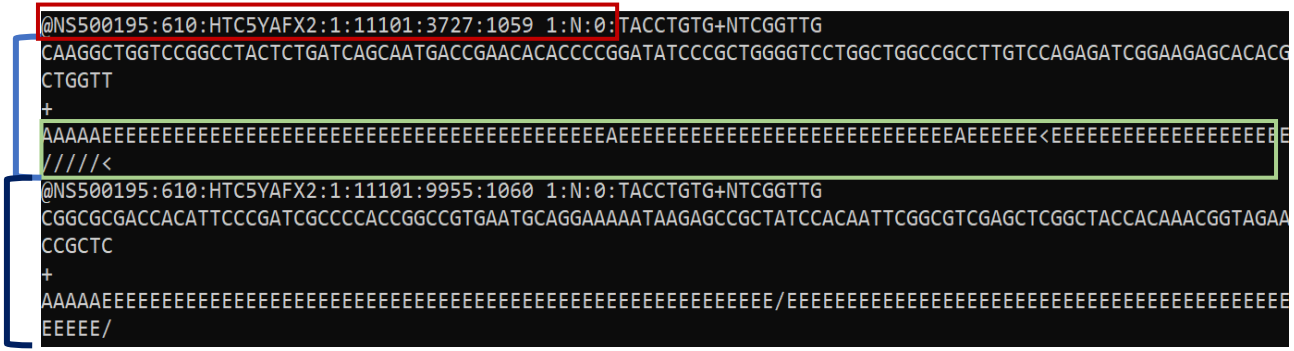
- File extension .fna
- This file format is used for reference genomes or reference database sequences
- In this file format, each sequence read is defined in 2 lines: header + sequence code

Sequence header/ID

```
>NC_015758.1 Mycobacterium tuberculosis variant africanum GM041182, complete genome
TTGACCGATGACCCCGGTTCAGGCTTCACCACAGTGTGGAACGCGGTCGTCTCCGAACTTAACGGCGACC
CTAAGGTTGACGACGGACCCAGCAGTGATGCTAATCTCAGCGCTCCGCTGACCCCTCAGCAAAGGGCTTG
```

Fastq files

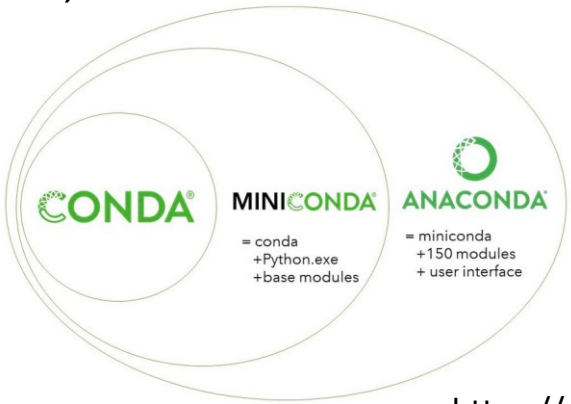
- File extension .fastq
- fastq = fasta + quality scores, result of conversion of squiggles to bases
- Can have multiple sequences in a list
- Each sequence read is defined in 4 lines: Header + sequence code+ Name field (optional, usually empty line starts with "+" sign + ‘sequence quality’ information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

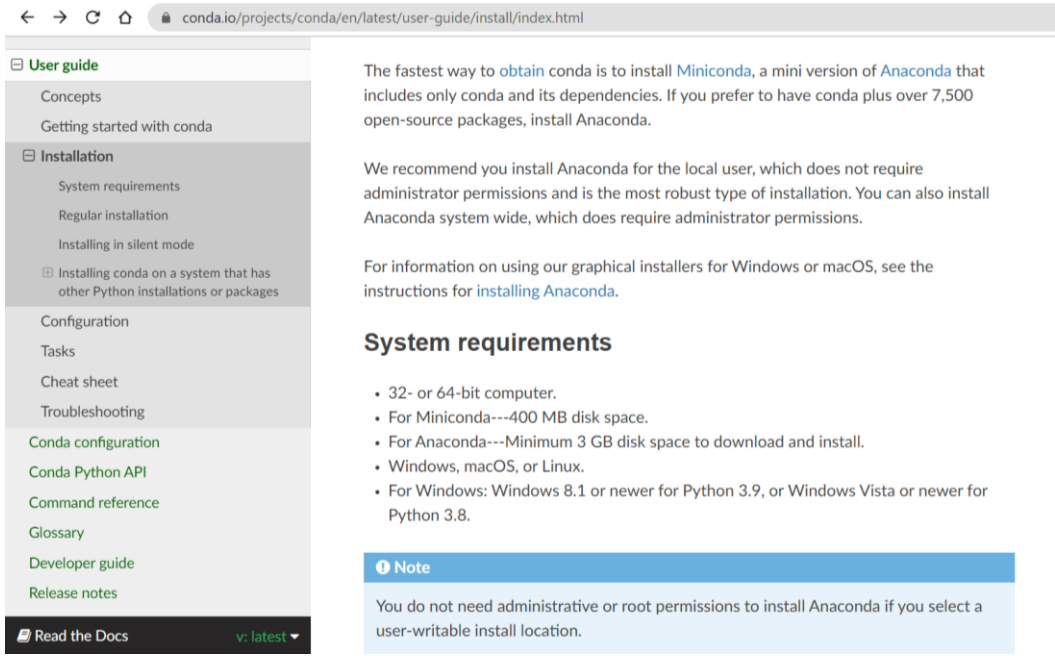


Example of Fastq File content containing 2 sequence reads; each read is defined in 4 line red square: Header/Sequence ID, green box is the quality scores corresponding to each base

Anaconda (3 GB) /Miniconda (400 MB)

- A repository for >7,500 open source packages
- Package installing tool (Conda)
- Creating virtual environments (isolated environments for different projects to solve the problem of different tools' versions required for each)





https://conda.io/projects/conda/en/latest/user-guide/install/index.html

https://repo.anaconda.com/miniconda/

https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html

FastQC

FastQC aims to assess the quality control checks on raw sequence data coming from high throughput sequencing pipelines

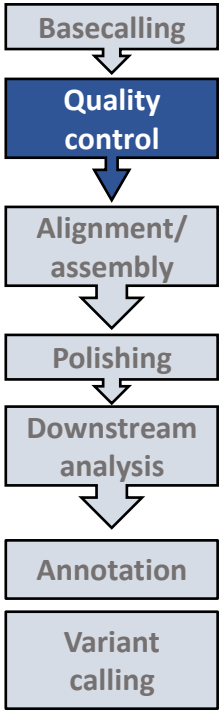
It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

Requirements:

- Java installed
- At least 8 GB RAM
- Operating systems: Linux, MacOS and 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 – how to install (Linux/MacOs/WSL) line 177

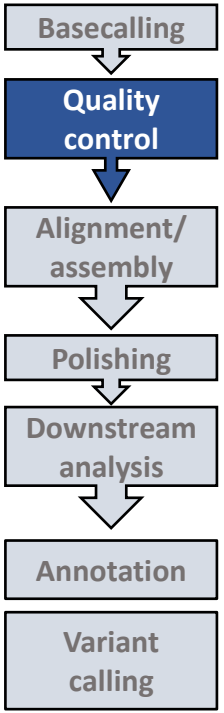
Download the .zip file from the website

You will need to navigate to the folder you downloaded it to and then run:

```
sudo apt install fastqc
```

if you get an error installing it, you may need to update your cache:

```
sudo apt update
```



FastQC basic code (Linux) (line 191)

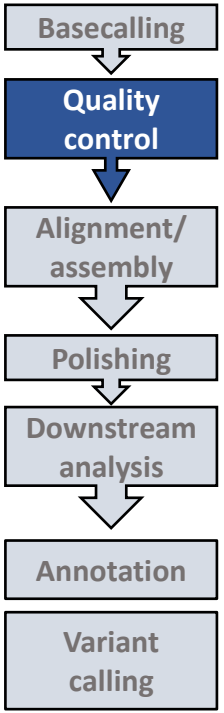
```
fastqc <filepath>/*.fastq -t 2
```

*.fastq indicates you want to process every file with the extension .fastq

Number of threads you are using

Name of the programme

File address for the .fastq files you want to QC



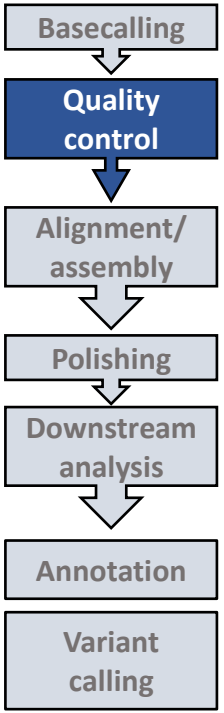
MultiQC

MultiQC searches a given directory(folder) for analysis logs (e.g. from FastQC) and compiles a HTML report

It's a general use tool for summarising the output from numerous bioinformatics tools, analysing across many samples into a single report

Requirements:

- Java installed
- At least 8 GB RAM
- Operating Systems: Linux, MacOS and Windows

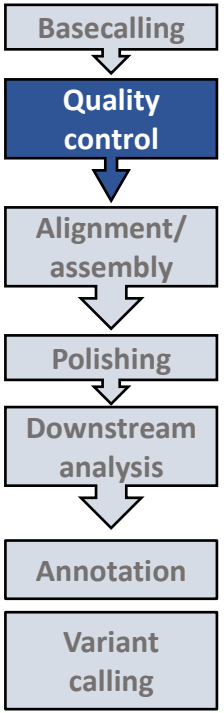


MultiQC – how to install (Linux/MacOs)

Install multiQC:

```
(line 198)
conda install -c bioconda -c multiqc

conda install -c bioconda/label/cf201901 multiqc
```

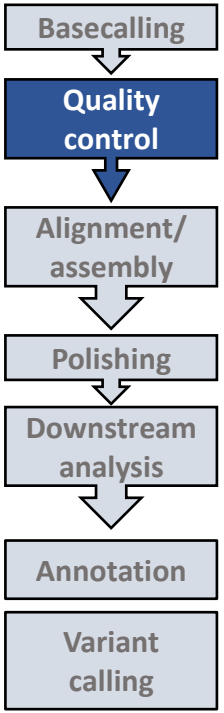


MultiQC basic code

```
multiqc fastqc
```

‘do this to all the files in this foler’

Name of the programme



Please refer to Script_ONT.txt (line 174)

Let’s Install the fastqc

Run the fastq files of Barcode 08 and Barcode 09: which is a better Sequence

UCL

ASANTE SANA!