Agenda

- Why is quality important
- Read Quality
- Assembly Quality
- QC Summary: The 4Cs

Next Generation Sequencing

Phases of Whole Genome Sequencing (WGS)

- DNA/RNA Extraction
- Library Preparation
- Sequencing
- Bioinformatics Analysis
- Genomic Epidemiology

Bioinformatic Analysis

Phases of Bioinformatics Analysis

- **Data Preprocessing**: Remove low-quality reads, adapters, and trim sequences (read cleaning)
- **Assembly**: Assemble reads into longer contigs or consensus
- **Annotation**: Annotate genetic variants and predict their functional impact
- **Genomic Characterization**: Identify genomic features that confer phenotypic qualities such as virulence
- **Phylogenetics**: Infer evolutionary relationships

Data Expectation

Data Reality

Garbage In, Garbage Out!

- The quality of data limits what you can confidently say about the data and how you can subsequently use it.
- Cleaning and preprocessing data are critical steps in data analysis.
- High-quality data ensures that analyses and results can be reproduced by others, which is crucial for public health, scientific research and credibility.

What QC thresholds should we use?

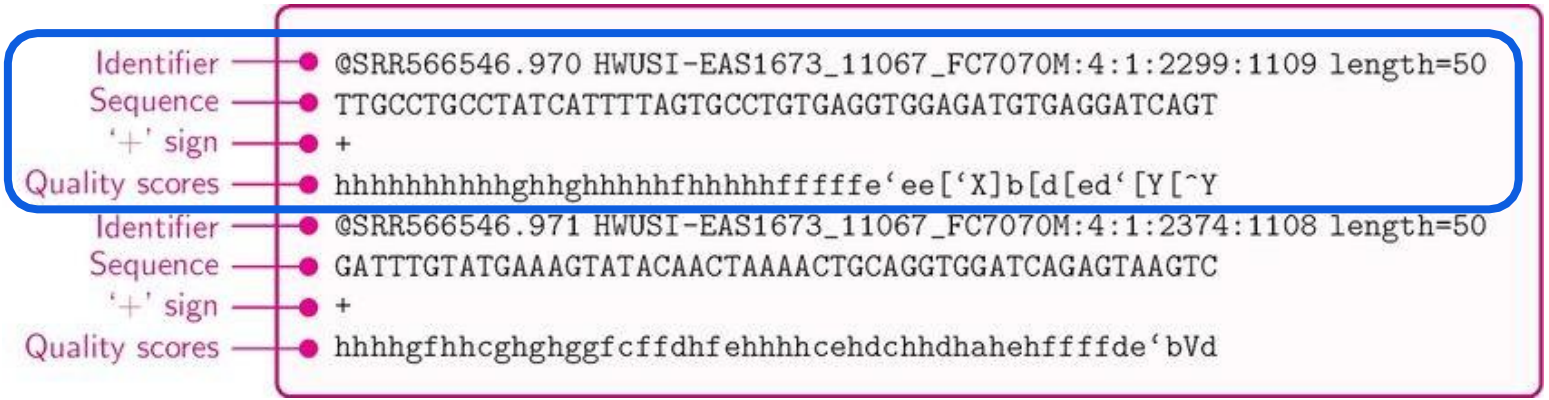
- Comprehensive quality thresholds have NOT been defined for most pathogens
- Guidance is available for some pathogens from some programmes, e.g. PulseNet

- Thresholds may need to be self-defined and agreed in your lab
- Need to evaluate QC metrics against reasonable **biological** and **technical** expectations for the **organism** and **sequencing approach**

Challenges

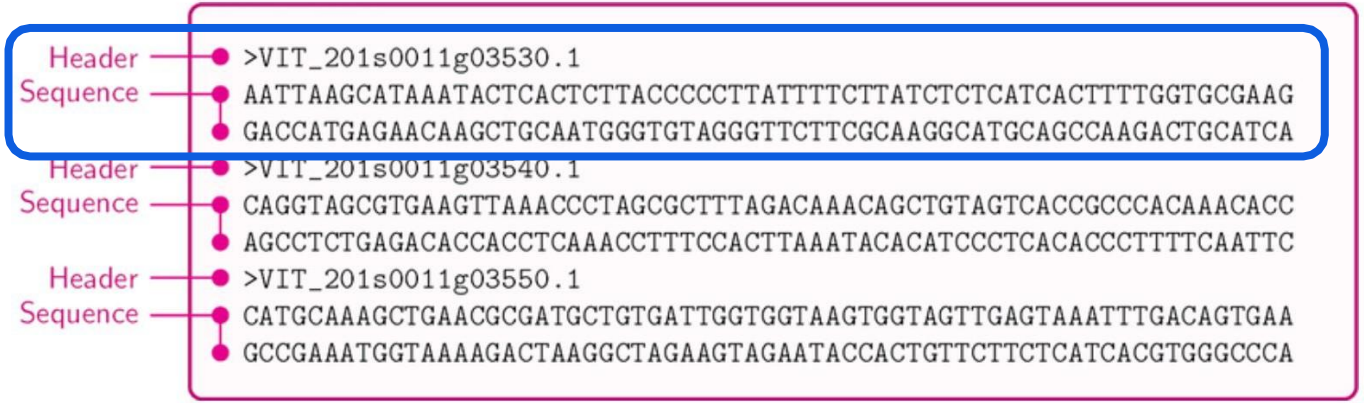
- Divergent expectations/standards across:
  - Organisms: RNA viruses, DNA viruses, bacteria, microbial eukaryotes
  - Sequencing approach: metagenome, single cell, amplicon, WGS
  - Technology: long-read, short-read, hybrid
  - Use-case: clinical, outbreak, surveillance, diagnostics
- Consequent heterogeneity in QC criteria naming/determination
- QC reporting can be complex

FASTQ file input



FASTQ file - a standardized format representing unprocessed sequencing fragments, each starting with a unique identifier followed by sequence data and associated quality scores

FASTA file input



FASTA file - a standardized format representing genetic sequences, each starting with a unique identifier followed by sequence data

Read QC

```
+SEQ_ID
!''*((((***+))%%%++)(%%%%).1**
```

A quality value Q is an integer representation of the probability *p* that the corresponding base call is incorrect.

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

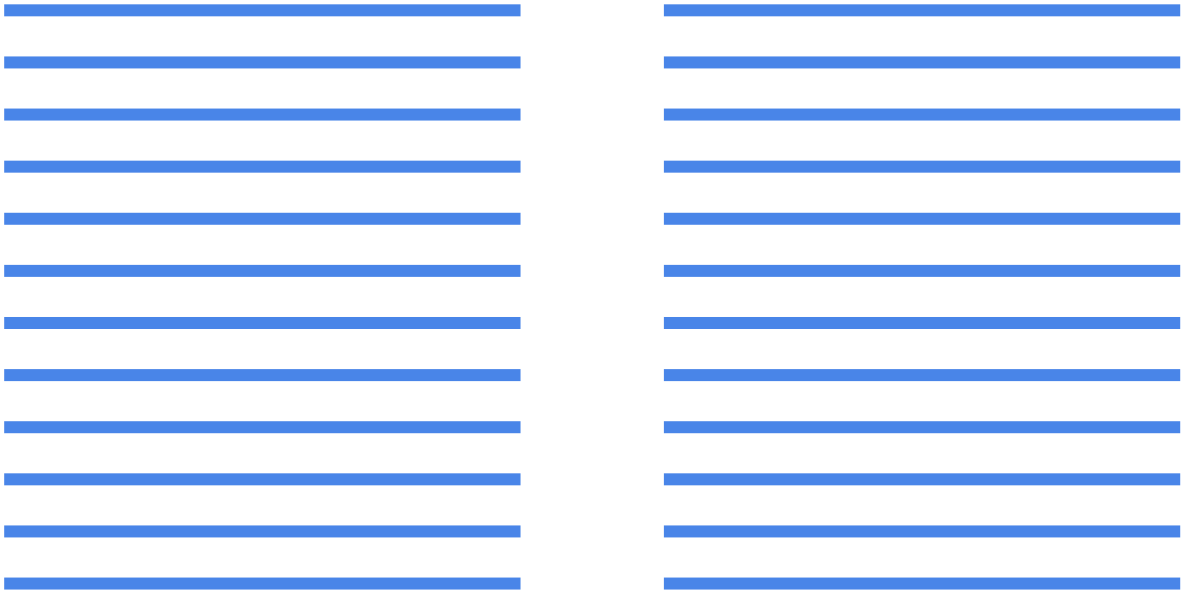
Theiagen®

Things to bear in mind…

- Read quality usually decreases towards the end of the read
- Reverse read is usually poorer quality than forward read

FastQC

Read trimming: Trimmomatic and BBDuk



Remove:

- Low quality reads
- Low quality ends of reads
- Sequencing adapters

Read trimming: Trimmomatic and BBDuk

Remove:

- Low quality reads
- Low quality ends of reads
- Sequencing adapters

Interpreting read quality outputs

- **Number of reads**: Too low = Too much multiplexing, not enough input library; Too high = too little multiplexing, downsample
- **Average Read Quality:** Too low = error with library preparation or sequencing; data is not accurate
- **Average read length**: Too low = fragmented library preparation

Assembly QC

Assembly QC

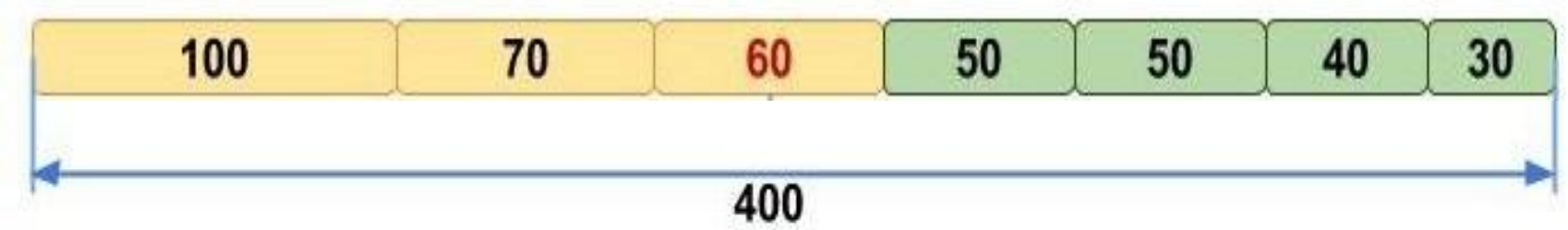
Number of contigs

n=7

Fewer is better

Assembly QC

Assembly length (total length of all contigs)

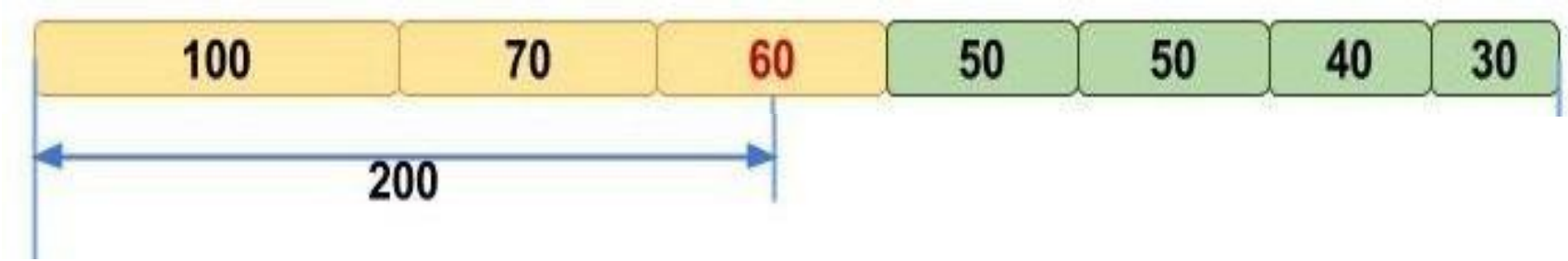


length=400

Should be close to the expected genome length

Assembly QC

N50 value



N50=60

Bigger is better

Length of the shortest contig needed when 50% of the total genome length is covered using the fewest (longest) contigs

Assembly QC

GC content

ATGCCAACAGTTCTGACTGA

9/20 = 45%

Close to expected GC % for taxon

Different bacterial species have different average GC %

V. cholerae GC content ~47.5%

Assembly quality control outputs

- Number of contigs
- Assembly length
- N50 value
- GC percent

QC Summary: The 4Cs

The 4 C's of Quality control

01 **Completeness**

Do you have the whole genome represented in the sequence data and assembly?

Assembly length: sum of all contigs from a single assembly

The 4 C’s of Quality control

02 **Contiguity**

How broken is the genome assembly?

Number of Contigs

N50

Number of Ns

The 4 C's of Quality control

03	**Correctness**	Is the assembly correct on a per-base basis, and are the reads correctly assembled?

Average Read Quality

Average read depth throughout the genome (sometimes referred to as *depth of coverage***)**

The 4 C’s of Quality control

04 **Contamination**

Are (too many) reads from non-target taxa or multiple clones?

Kraken taxon

Solutions for QC failures

01	**Completeness**	Do you have the whole genome represented in the sequence data and assembly?	Sequence deeper
02	**Contiguity**	How fragmented is the assembly?	Use longer read lengths
03	**Correctness**	Is the assembly correct on a per-base basis, and are the reads correctly assembled?	Sequence deeper/downsample reads/use higher accuracy sequencing
04	**Contamination**	Are (enough of) the reads from the target organism?	Re-isolate and re-sequence

Bad Data, Is Bad Data!