

Quality control of sequencing data

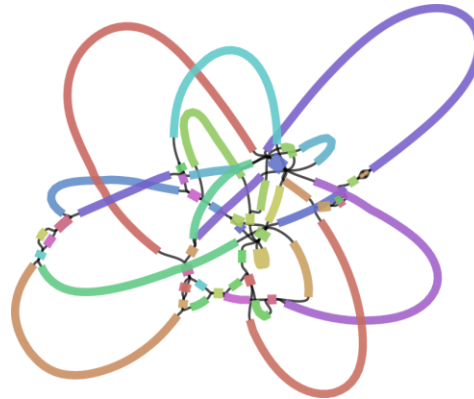
Illumina and ONT

Vanni Benvenga
Bioinformatician
Applied Microbiology Lab
PI: Prof. Dr. Dr. Adrian Egli
Institute of Medical Microbiology
University of Zurich

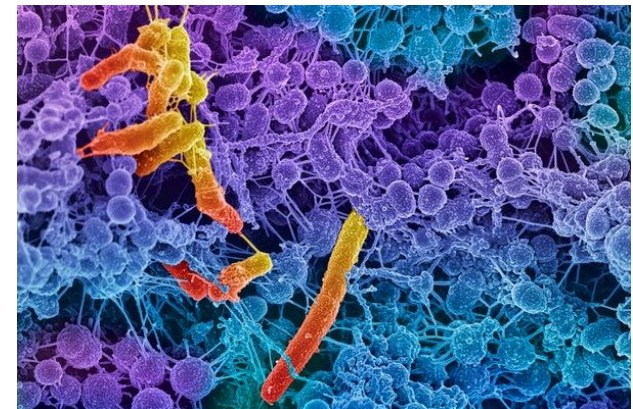
Learning outcomes

- You understand the importance of quality control (QC) in Illumina and ONT sequencing
- You can describe key sequencing quality metrics
- You know common sequencing errors for both technologies
- You know which software you can use to pinpoint QC issues

- Pictures? Squiggles? Phasing? QC?
- Reads? Base quality? Homopolymers? QC?
- Assembly algorithms? Tandem repeats? QC?



- What is it?
- How does it compare to others?
- How can I best use the data?



Key quality metrics

Illumina  ONT

Phred scores

Read length distribution

GC content variability

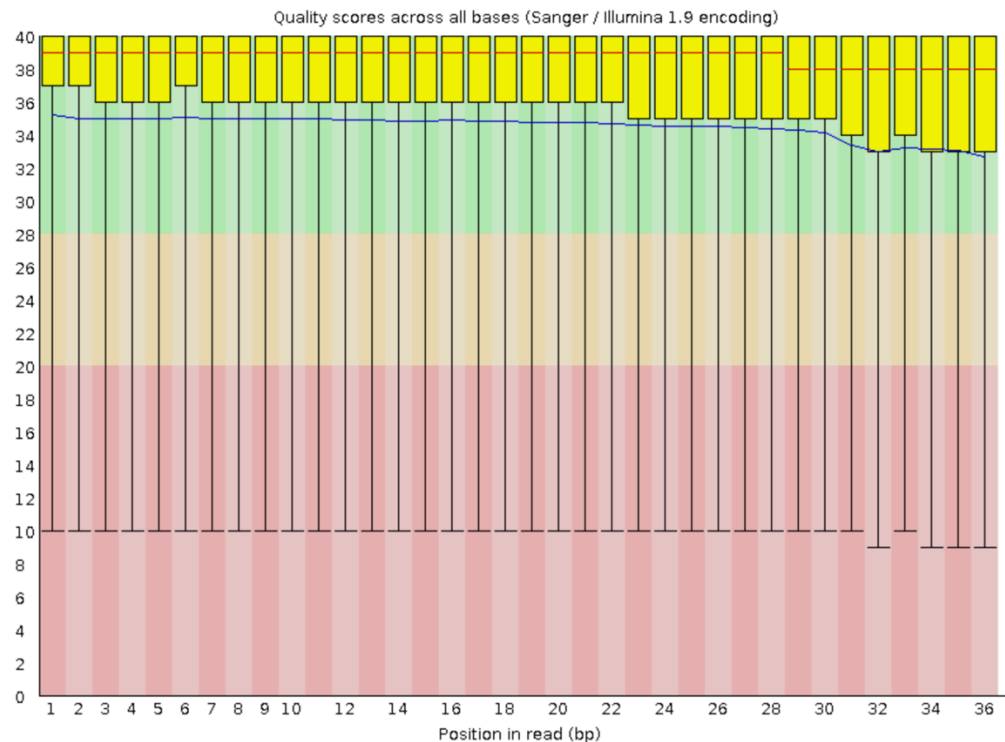
Adapter/contaminant detection

Phred score, a common measure for quality

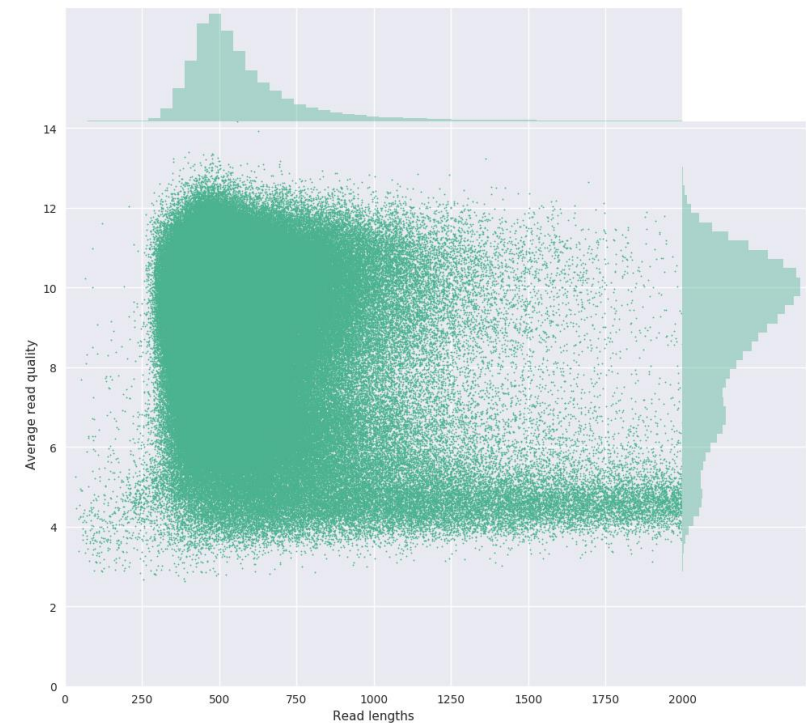
$$Q = -10\log_{10}P$$

Where Q is the Phred score and P is the probability of a sequencing error

✓ Per base sequence quality



Read lengths vs Average read quality plot



Common sequencing errors

Illumina

- substitution errors/phasing
- GC and PCR bias
- adapter contamination

ONT

- homopolymer errors
- indels
- adapter contamination

How can we deal with these errors

- fastQC is a simple software that displays all these metrics at once and with thresholds
- [good_sequence_short.txt FastQC Report.html](#)
- [bad_sequence.txt FastQC Report.html](#)
- multiQC is a great option when dealing with a lot of samples