# NGS - quality control, alignment, visualisation

Quality control + database retrieval

# Why Quality control?

1. How is the base quality?
2. What is the read length?
3. Are there adapters/barcodes in my sequences?
4. Are there overrepresented sequences?

# Dedicated software

- Manufacturers' software
- Illumina: fastQC
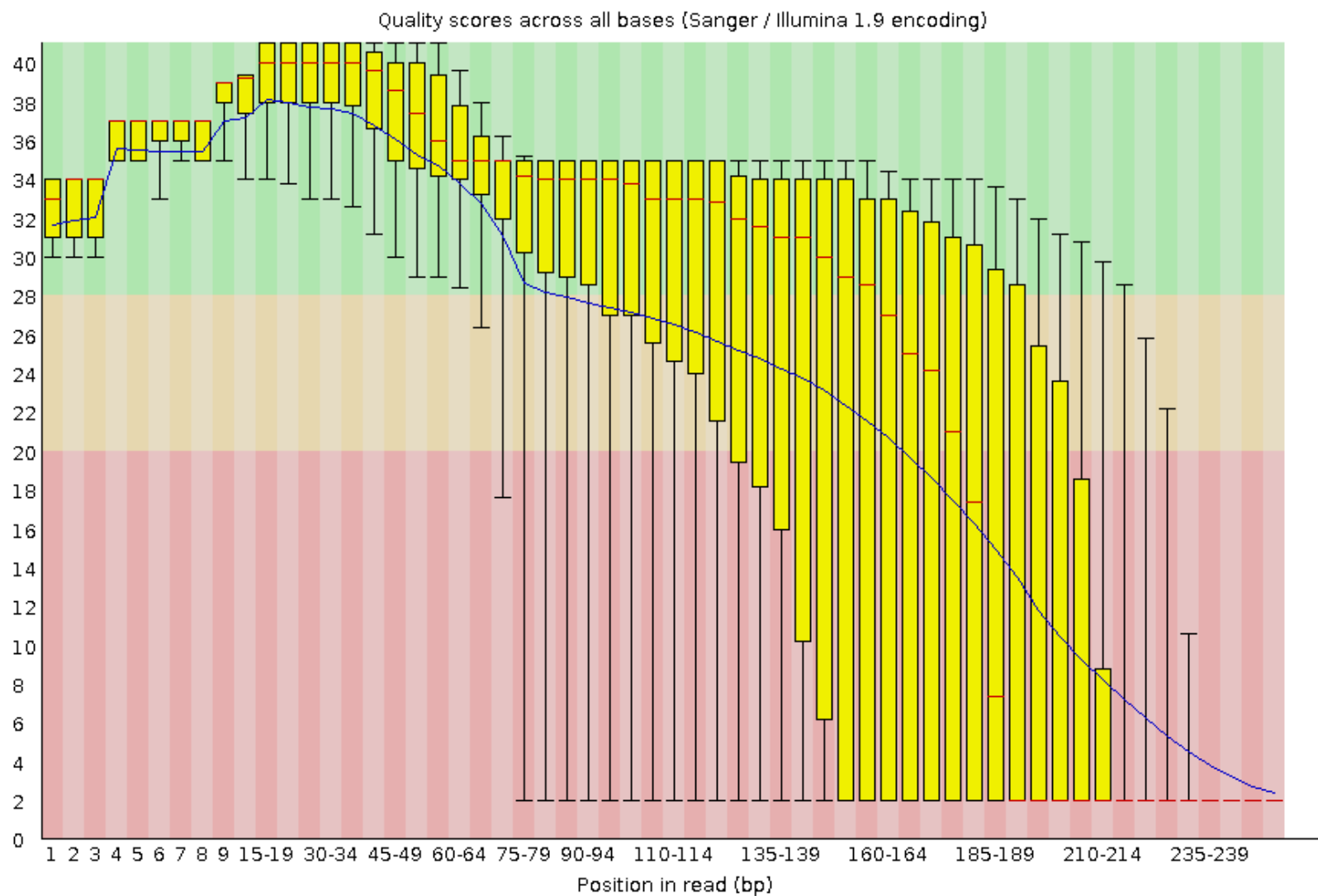- ONT: pycoQC
- ONT + PacBio: NanoPlot
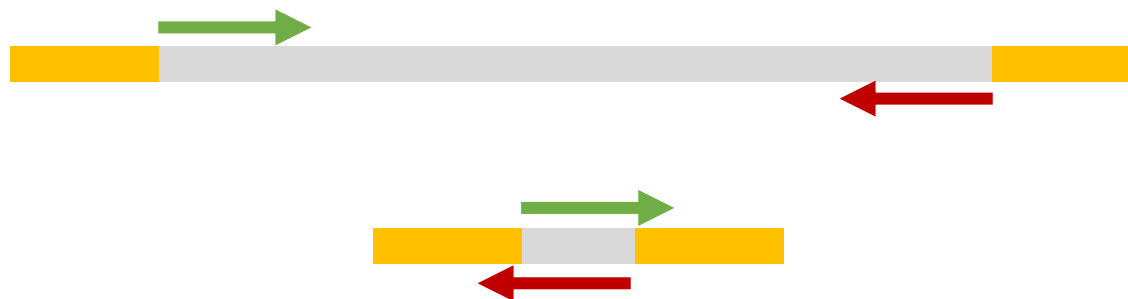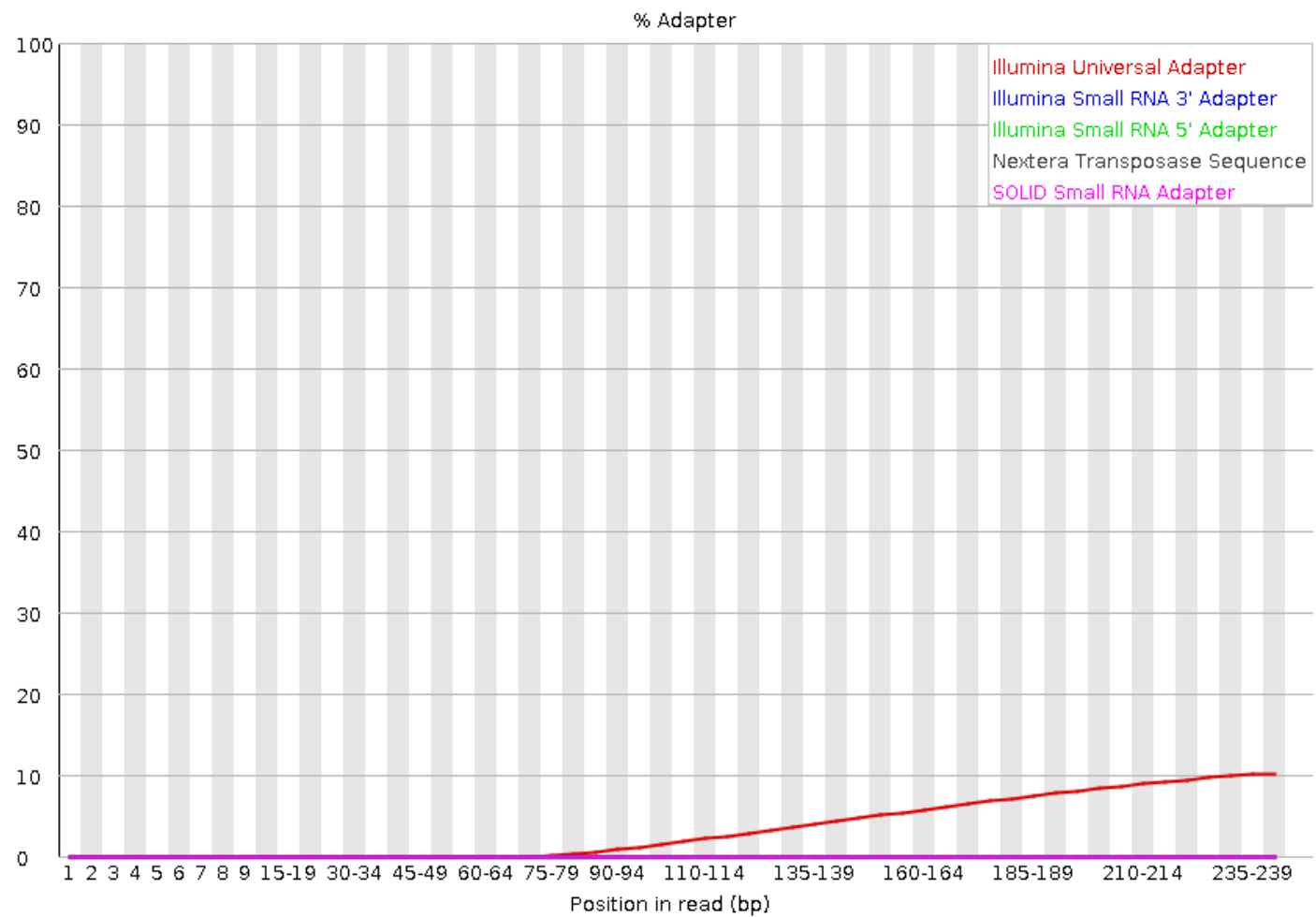
# fastq

fasta + basequality (fasta + q = fastq)

$$BASEQ = -10 log_{10} \Pr\{base\ is\ wrong\}$$
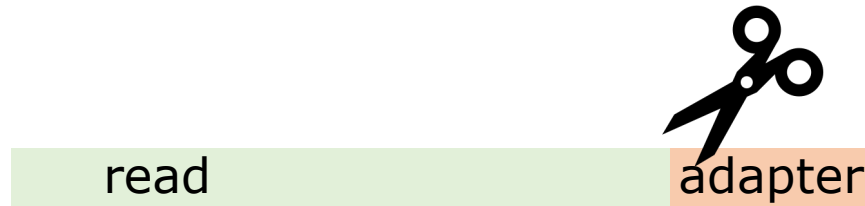
$$-10 log_{10} (0.01) = 20$$
$$-10 log_{10} (0.1) = 10$$
$$-10 log_{10} (0.5) = 3$$

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Position in read (bp)

5

# Trimming

- Find and remove:
  - Regions or reads with low base quality
  - Adapter sequences
- Software: trimmomatic



read adapter

# Databases

**BioProject** (Former DRA Study)

BioProject　　PRJD
- Project description
- Grants
- Publications

**BioSample** (Former DRA Sample)

BioSample　SAMD

BioSample　SAMD

BioSample　SAMD
- Sample description
- Taxonomy ID

**Sequence Read Archive**

Experiment　　DRX
- Library layout
- Sequencing platform

Run　　DRR

Run　　DRR

Run　　DRR
- Data files

DDBJ

Sequence data files (fastq, BAM)

Prefix of accession number

9

# Command line tools

- Retrieve raw data: SRA-tools
  - `prefetch`
  - `fastq-dump`

- Retrieve sequences: Entrez Direct
  - `esearch`
  - `efetch`