

NGS - quality control, alignment, visualisation

Quality control + database retrieval

Why Quality control?

- How is the base quality?
- What is the read length?
- Are there adapters/barcodes in my sequences?
- Are there overrepresented sequences?

Dedicated software

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

fastq

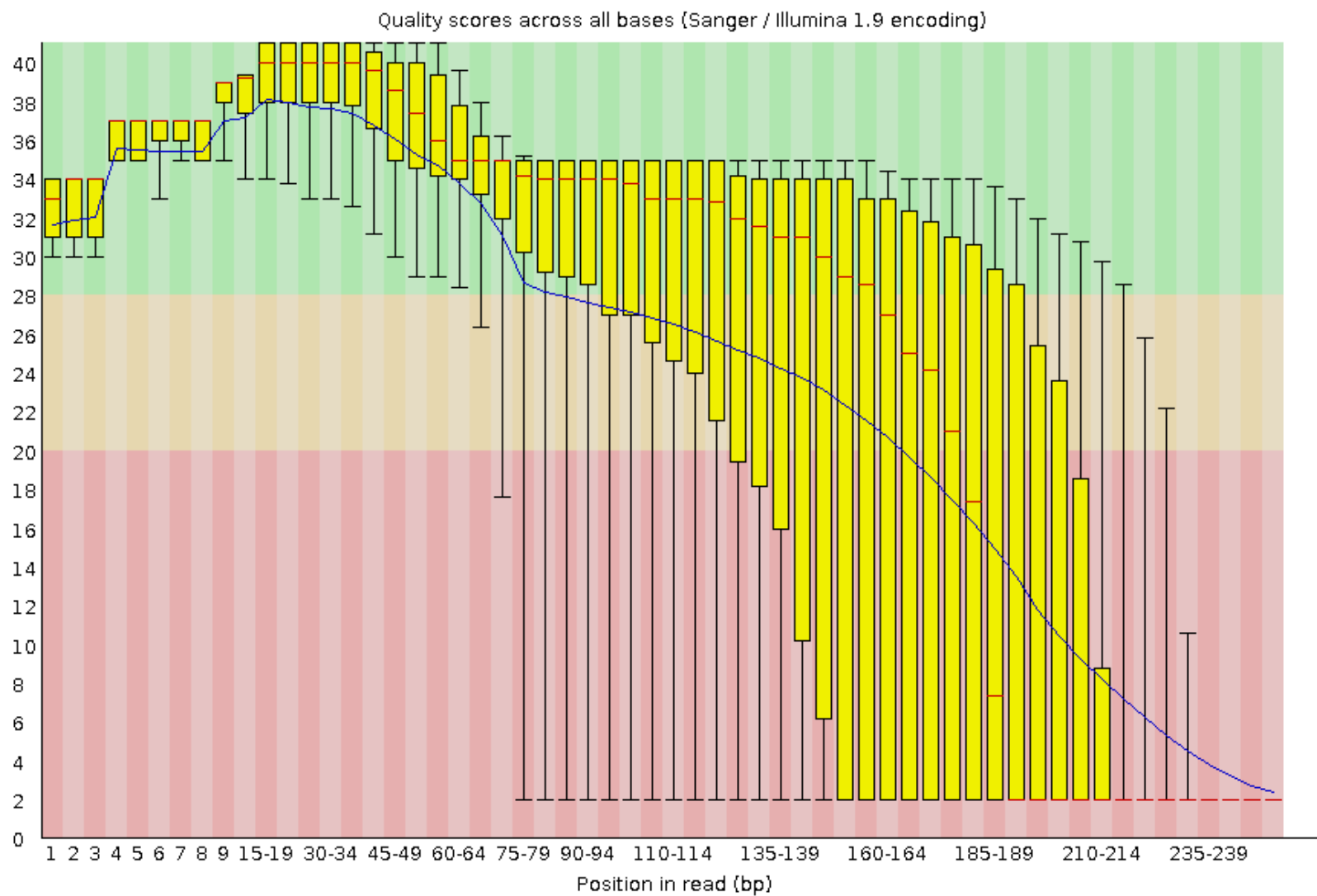
fasta + basequality (fasta + q = fastq)

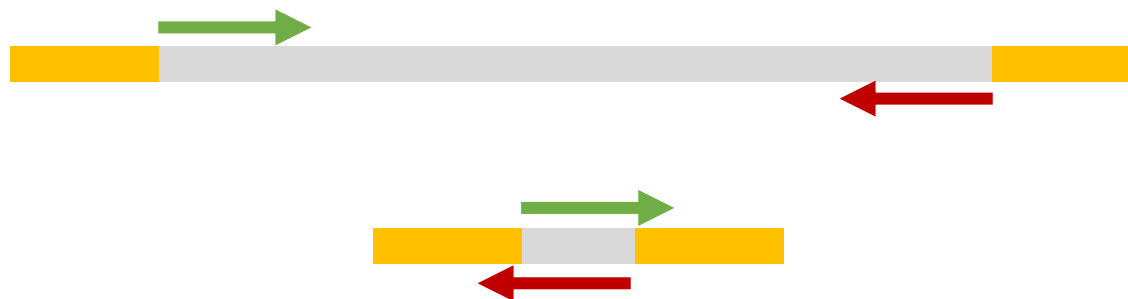
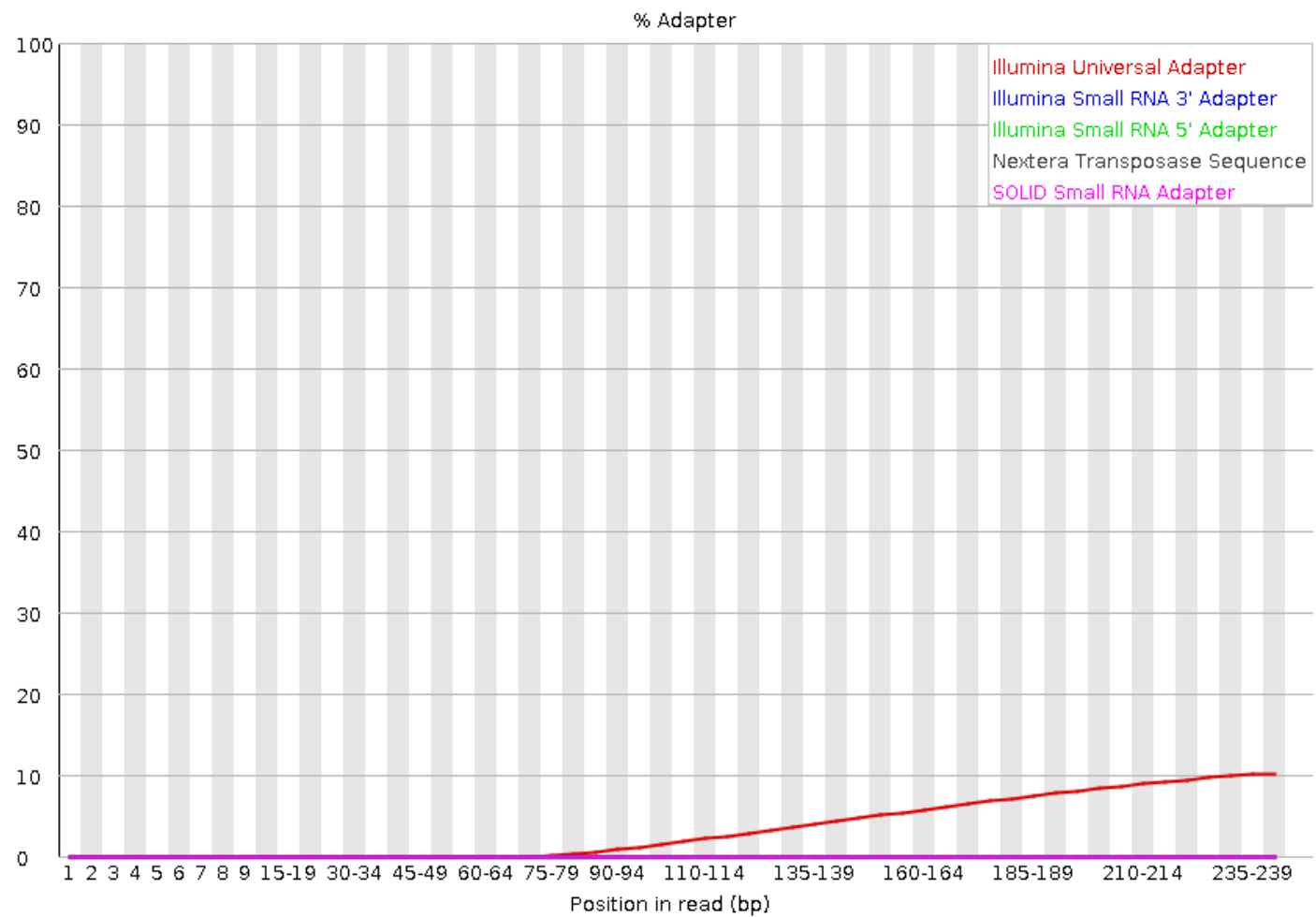
$$BASEQ = -10\log_{10} \Pr\{base\ is\ wrong\}$$

$$-10\log_{10} (0.01) = 20$$

$$-10\log_{10} (0.1) = 10$$

$$-10\log_{10} (0.5) = 3$$





Trimming

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

Databases



BioProject (Former DRA Study)

BioProject PRJD

- Project description
- Grants
- Publications

BioSample (Former DRA Sample)

BioSample SAMP

BioSample SAMP

BioSample SAMP

- Sample description
- Taxonomy ID

Sequence Read Archive

Experiment DRX

- Library layout
- Sequencing platform

Run DRR

Run DRR

Run DRR

- Data files

Sequence data files (fastq, BAM)



Prefix of accession number

Command line tools

- Retrieve raw data: SRA-tools
 - prefetch
 - fastq-dump
- Retrieve sequences: Entrez Direct
 - esearch
 - efetch