

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

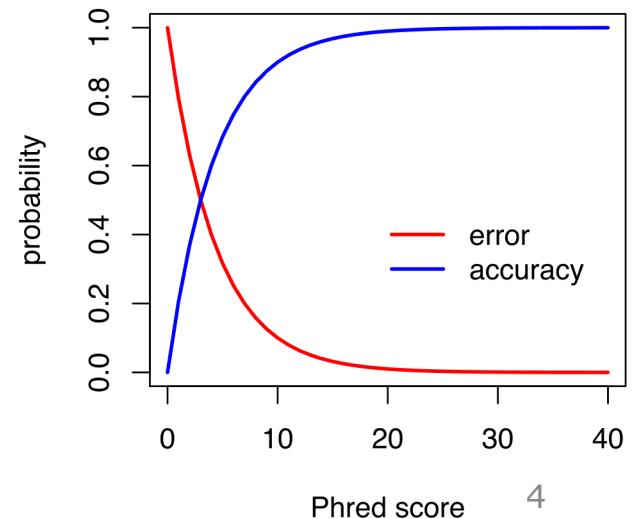
fasta + basequality (fasta + q = fastq)

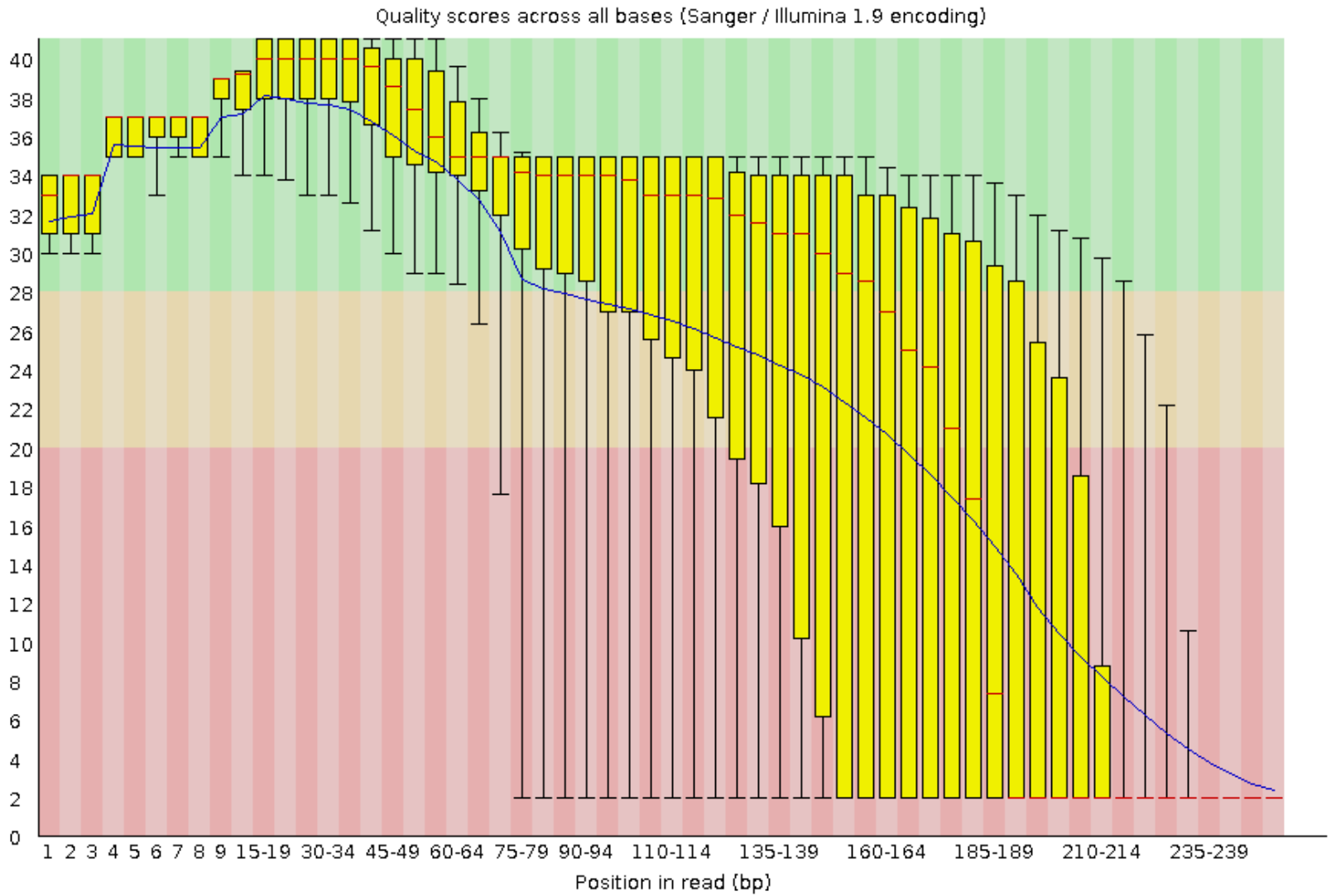
$$BASEQ = -10 \log_{10} \Pr\{base\ is\ wrong\}$$

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

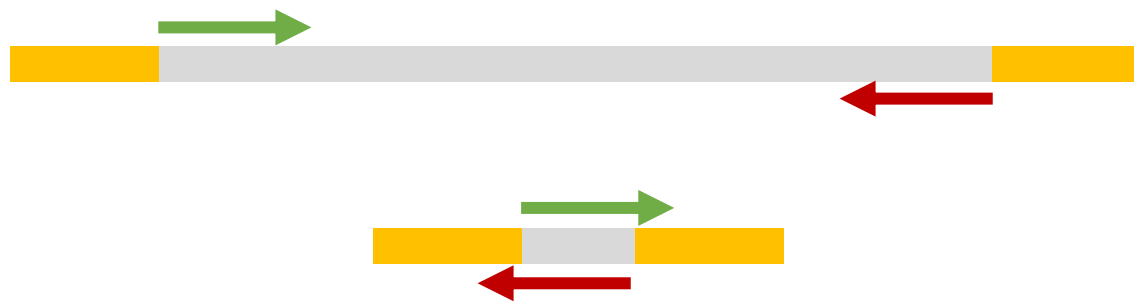
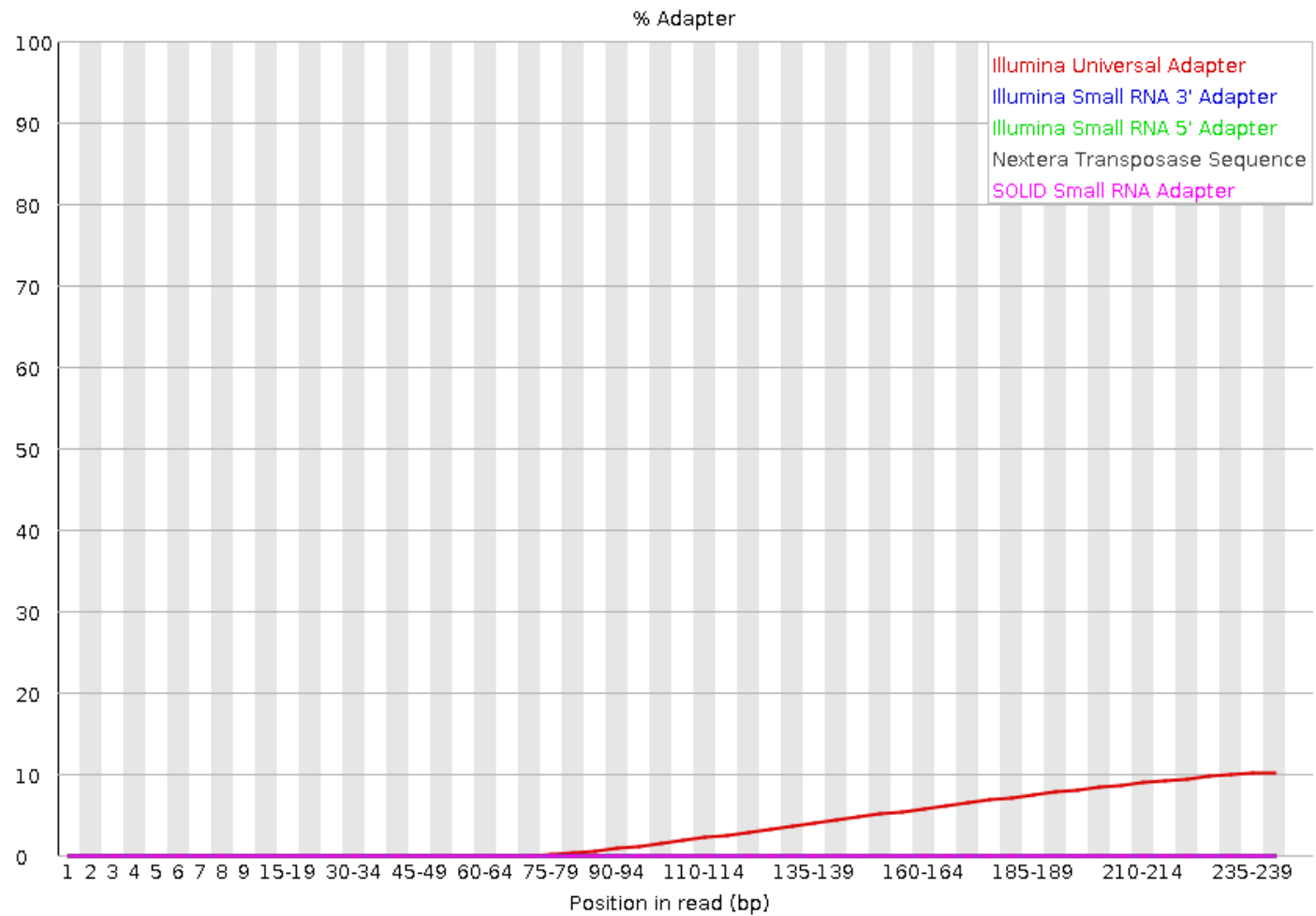
$$-10 \log_{10} (0.1) = 10$$

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



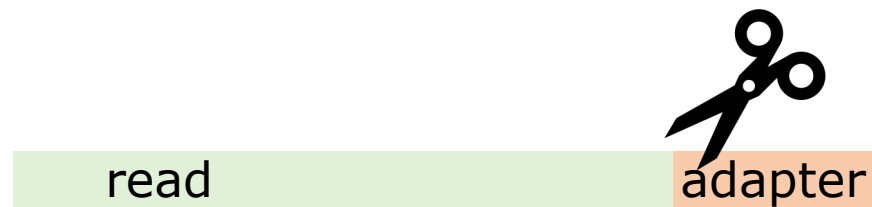






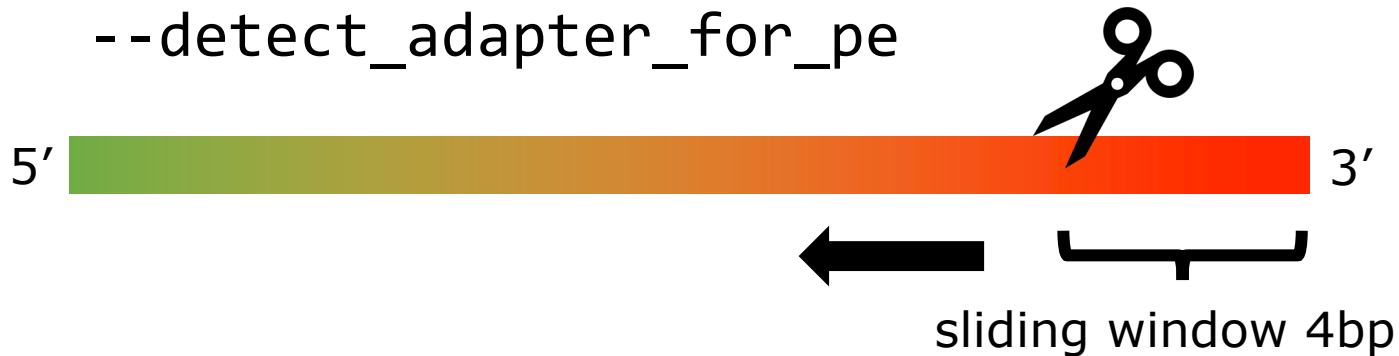
Trimming

- Find and remove:
 - Regions or reads with low base quality
 - Adapter sequences
 - poly G sequences (e.g. with NovaSeq 6000)
- Software: fastp (or cutadapt, trimmomatic, trim_galore, bbduk ..)
- Articles on frequently occurring issues: sequencing.qcfail.com



Quality trimming with fastp

- Default:
 - Remove reads with $>40\%$ bases $<Q15$
 - Trim poly N (and poly G)
 - Autodetect adapters in R1, for both:
`--detect_adapter_for_pe`



- 'Classical' trimming: sliding window
 - options `--cut_front` and `--cut_tail`

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

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