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

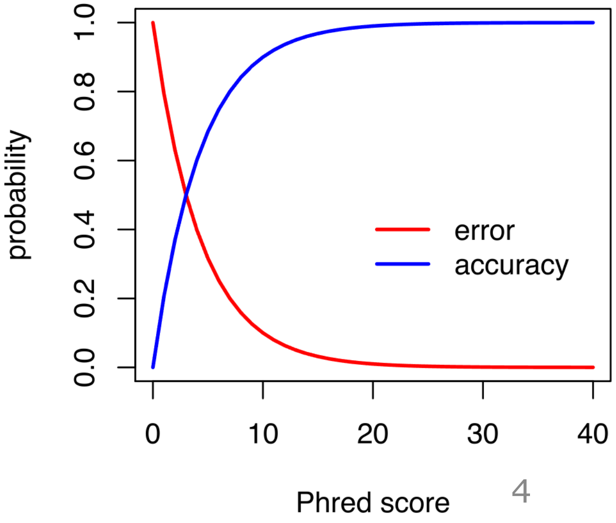
- Manufacturers’ software
- Illumina: fastQC
- ONT: pycoQC
- ONT + PacBio: NanoPlot

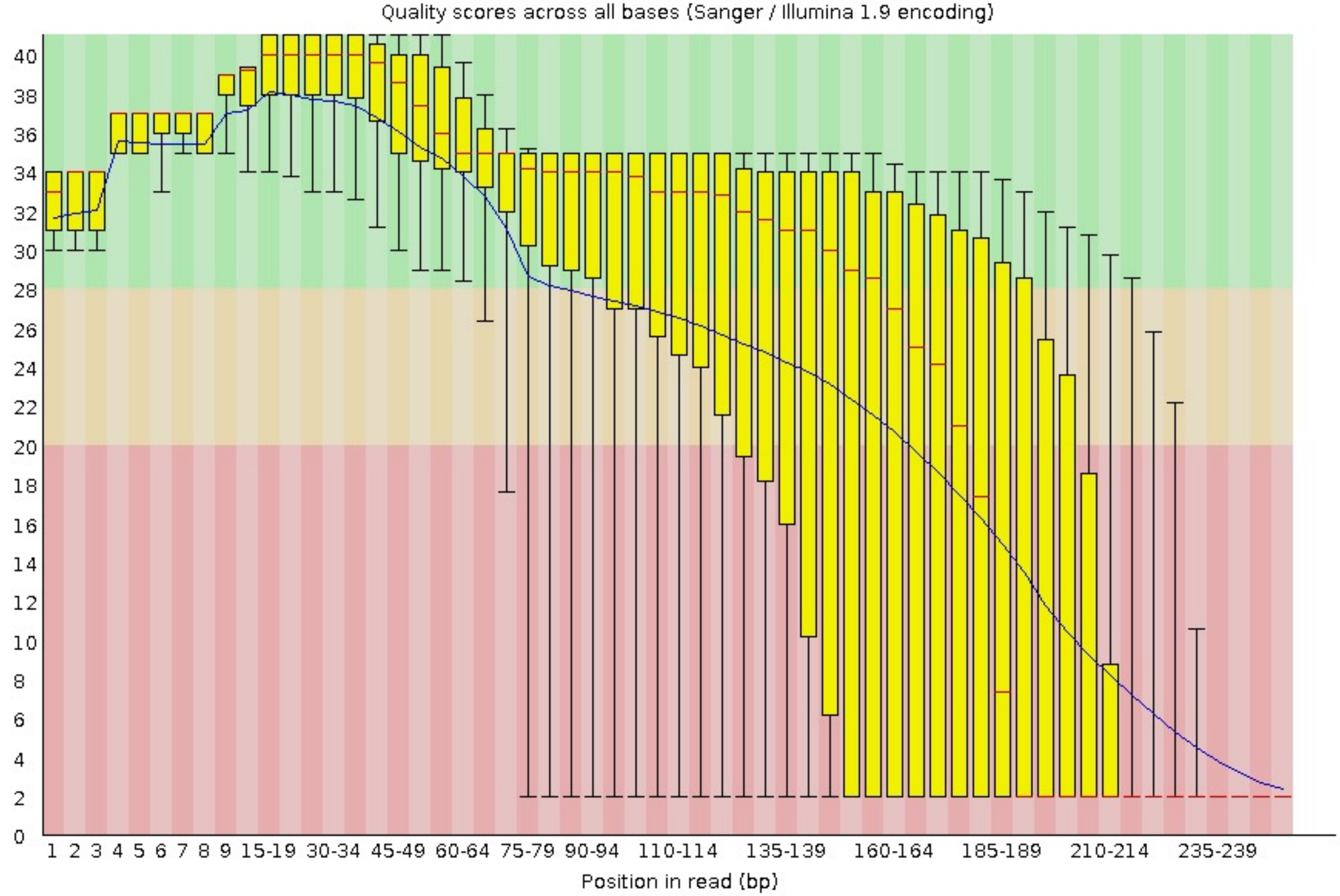
fastq

fasta + basequality (fasta + q = fastq)

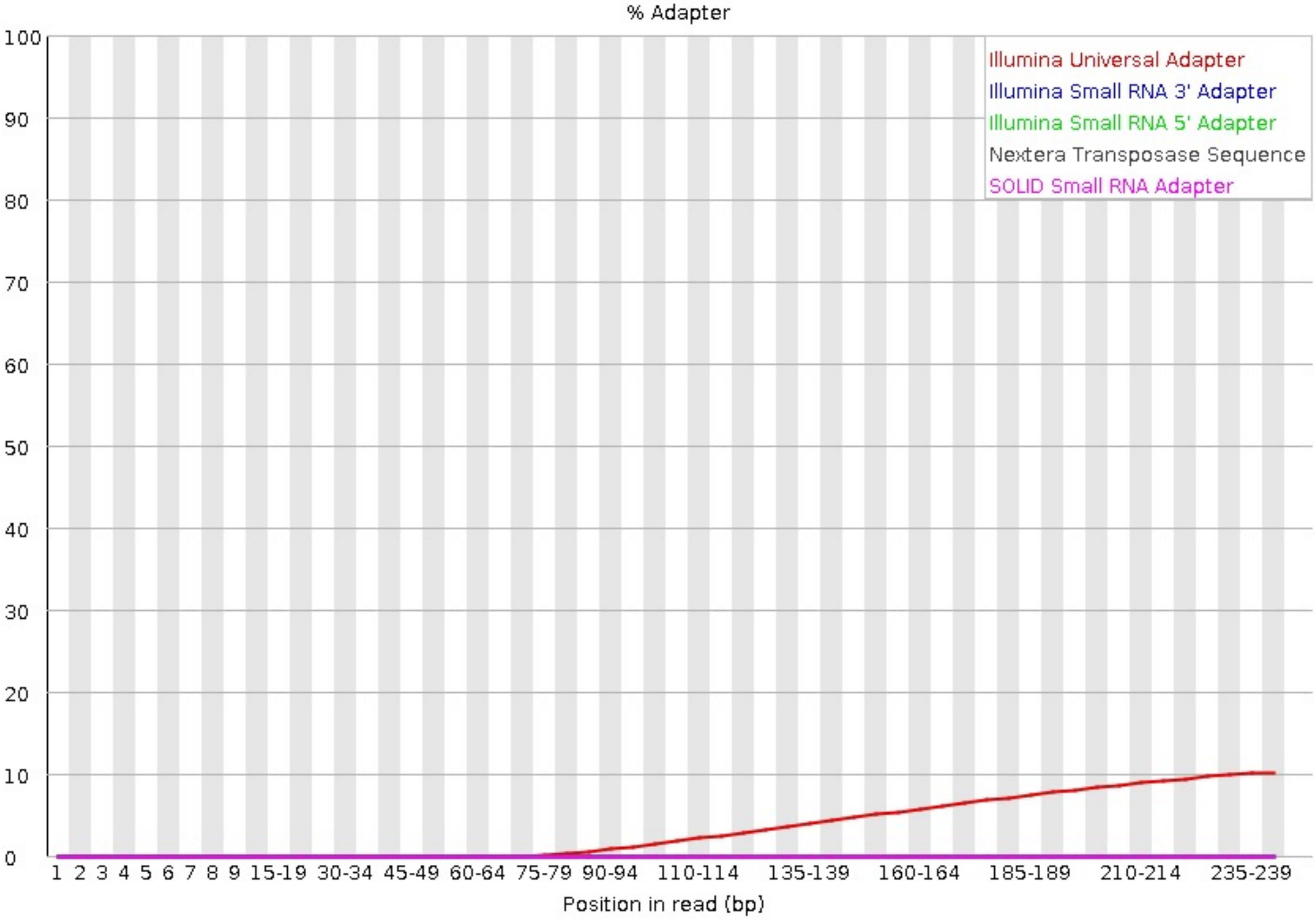
$$BASEQ = -10log_{10}\,\Pr\{base\ is\ wrong\}$$

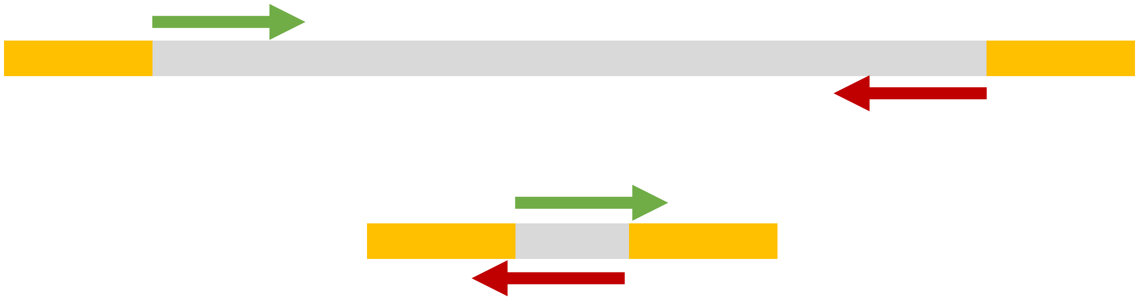
$$-10log_{10}\,(0.01) = 20$$
$$-10log_{10}\,(0.1) = 10$$
$$-10log_{10}\,(0.5) = 3$$





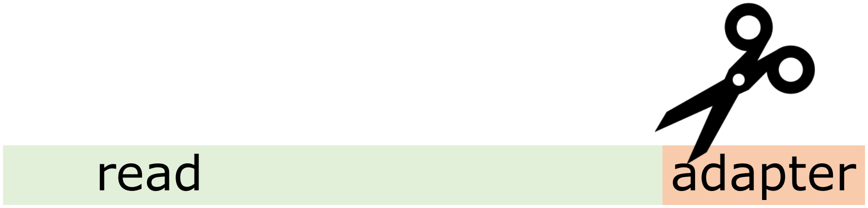
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
1
2
3
4
5
6
7
8
9
15-19
30-34
45-49
60-64
75-79
90-94
110-114
135-139
160-164
185-189
210-214
235-239
Position in read (bp)


% Adapter
100
90
80
70
60
50
40
30
20
10
0
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 110-114 135-139 160-164 185-189 210-214 235-239
Position in read (bp)

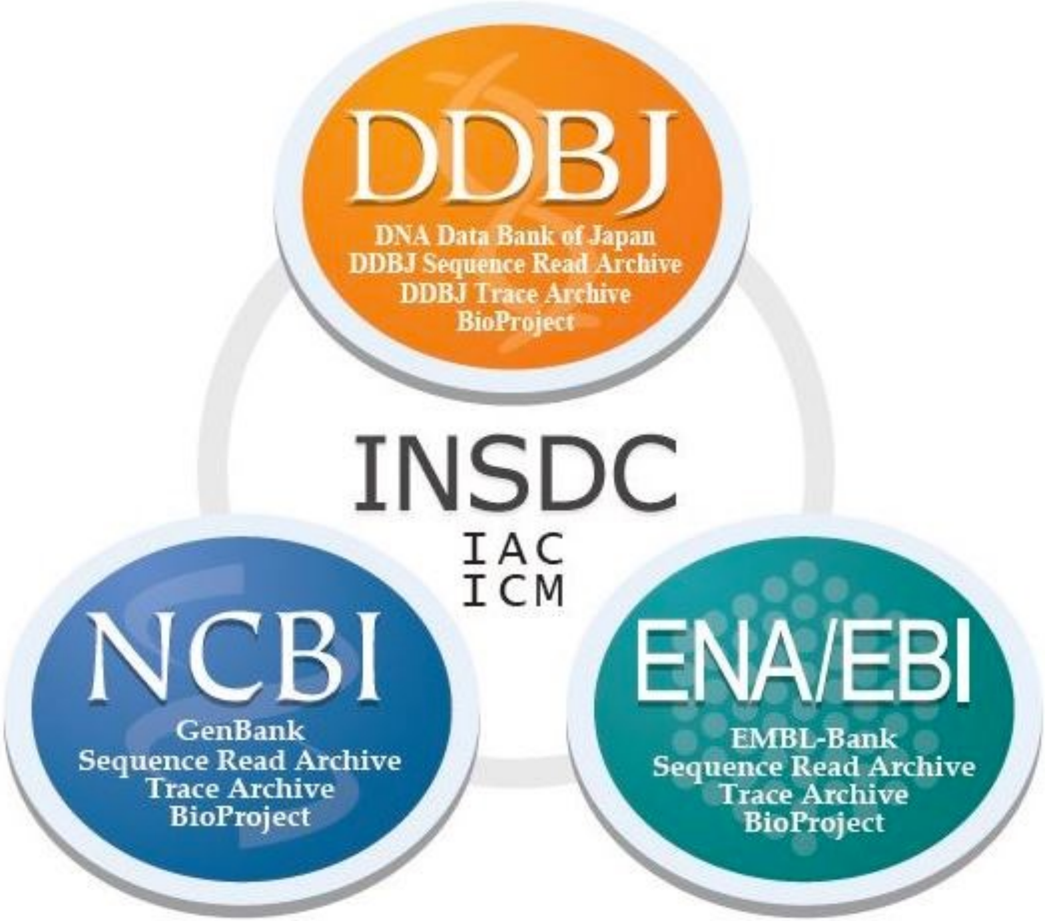


Trimming

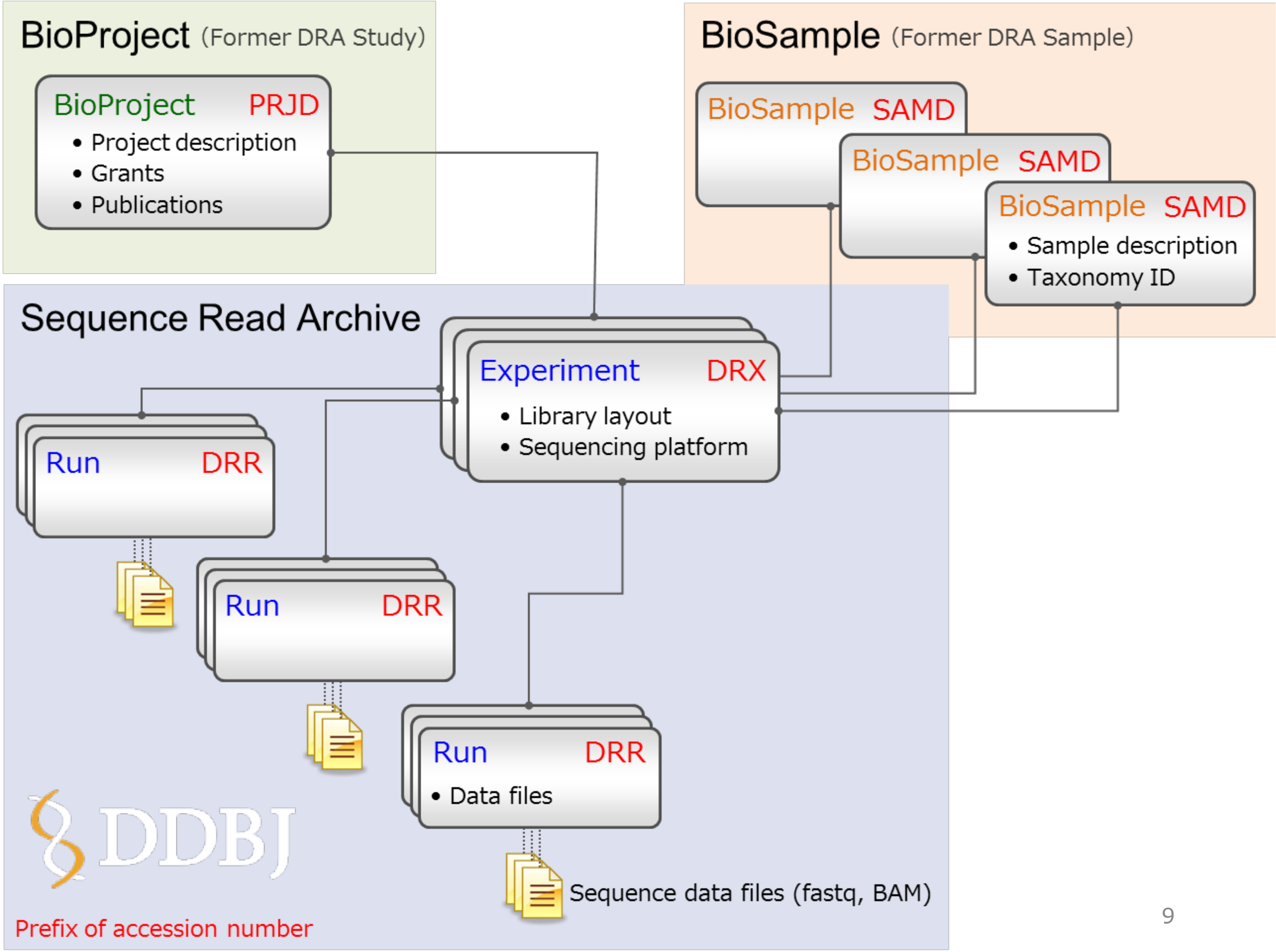
- Find and remove:
  - Regions or reads with low base quality
  - Adapter sequences
- Software: `cutadapt` (or `trimmomatic`, `trim_galore`, `bbduk ..`)



Databases



INSDC: **I**nternational **N**ucleotide **S**equence **D**atabase **C**ollaboration



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)

DDBJ

Prefix of accession number

Command line tools

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