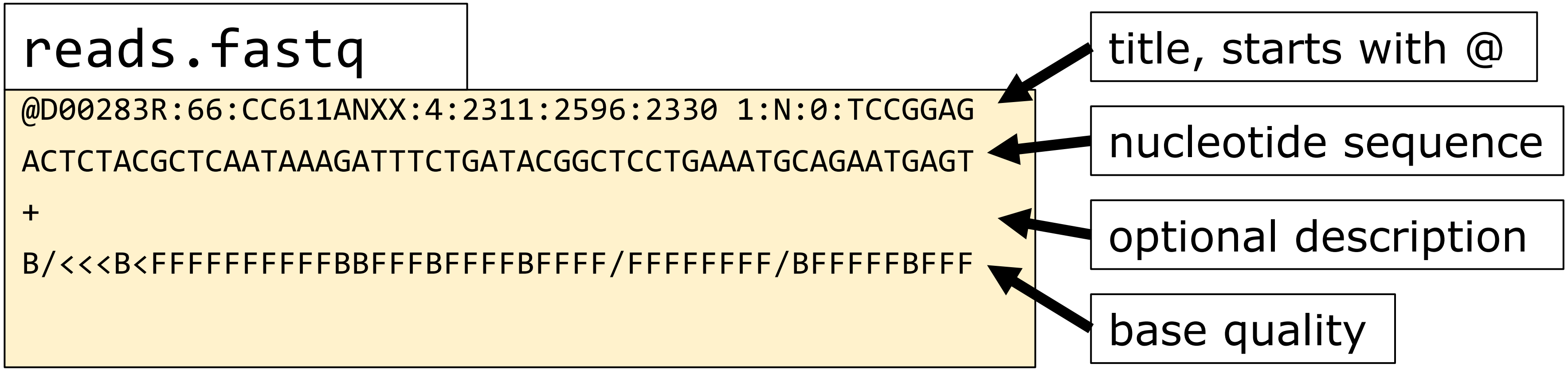
Long-read sequence analysis

File formats and QC

Raw file formats

- ONT:
  - POD5 (new, apache arrow)
  - FAST5 (HDF5)
  - Base calling: MinKNOW (guppy)/dorado/third party
- PacBio:
  - unaligned BAM (binary sequence alignment format SAM)

fastq



reads.fastq

```
@D00283R:66:CC611ANXX:4:2311:2596:2330 1:N:0:TCCGGAG
ACTCTACGCTCAATAAAGATTTCTGATACGGCTCCTGAAATGCAGAATGAGT
+
B/<<<B<FFFFFFFFFFBBFFFBFFFFBFFFF/FFFFFFFF/BFFFFFBFFF
```

title, starts with @

nucleotide sequence

optional description

base quality

fastq

fasta + basequality (fasta + q = fastq)

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

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

$$Acurracy = 1\ - \Pr\{base\ is\ wrong\}$$

$$-10log_{10}\,(0.01) = 20$$
$$-10log_{10}\,(0.05) = 13$$
$$-10log_{10}\,(0.5) = 3$$

Question 9

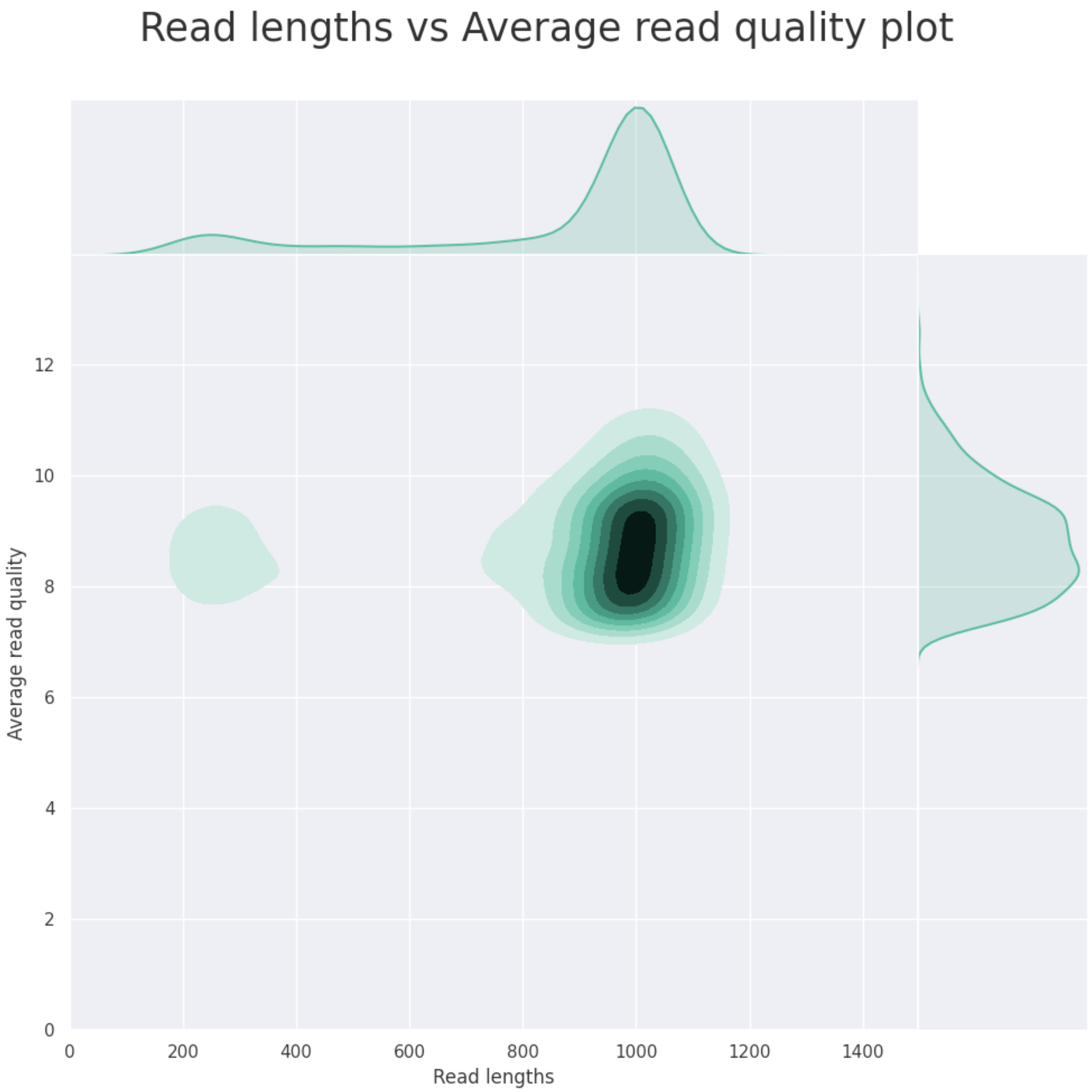
Read quality control

- Number of reads
- Read length (mean and spread)
- Base quality
- GC content
- Demultiplexing statistics
- Run duration/location dependency
- Others?

Question 10

Read quality software

- Software of manufacturer: SMRT Link; MinKNOW
- NanoPlot (https://github.com/wdecoster/NanoPlot)
  - Takes many input formats
  - Basic statistics
- PycoQC (https://github.com/a-slide/pycoQC)
  - Specific for ONT
  - Requires so-called `sequencing_summary` file
- FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
  - Works also for long reads
  - Familiar output to most people



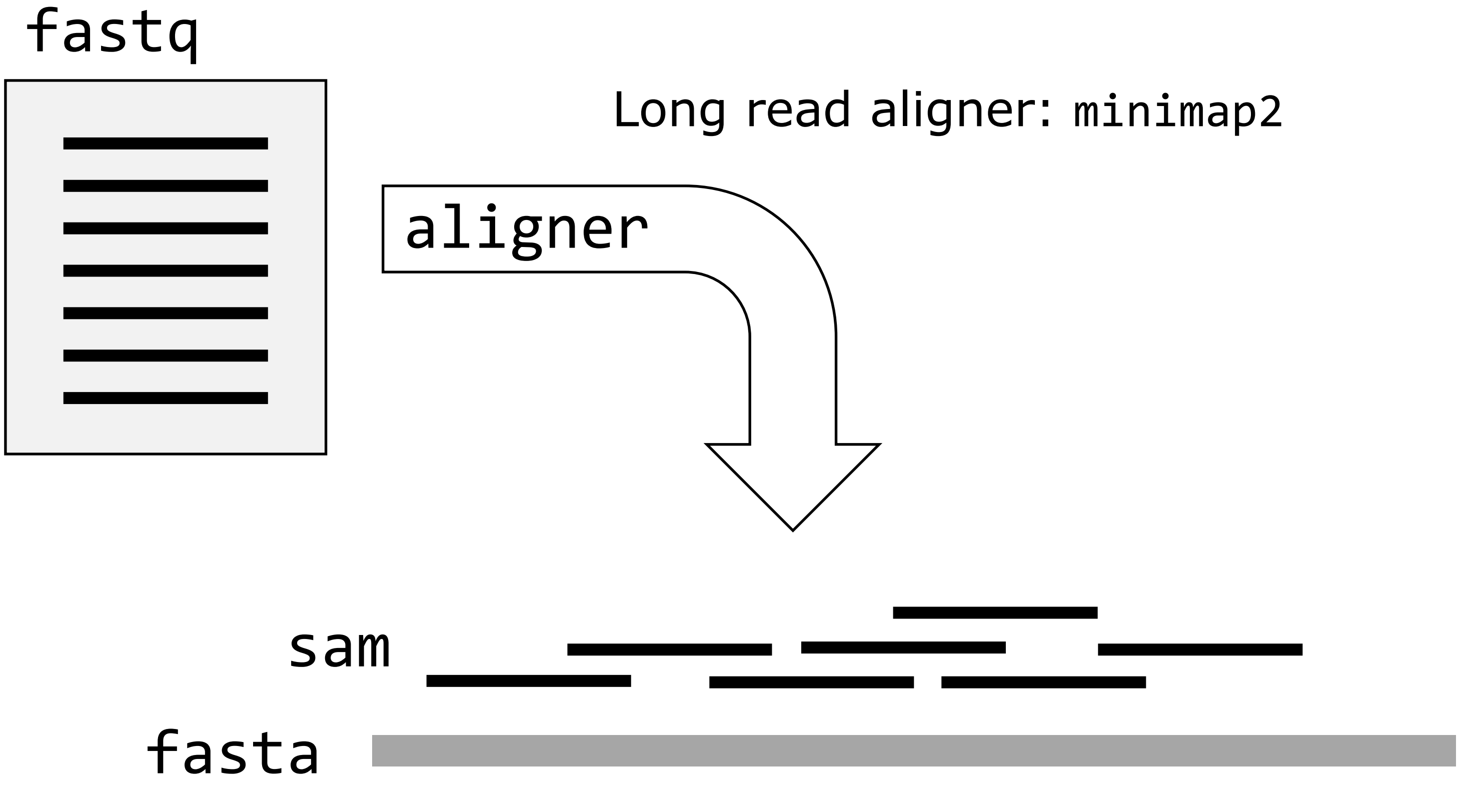
output of NanoPlot (https://github.com/wdecoster/NanoPlot)

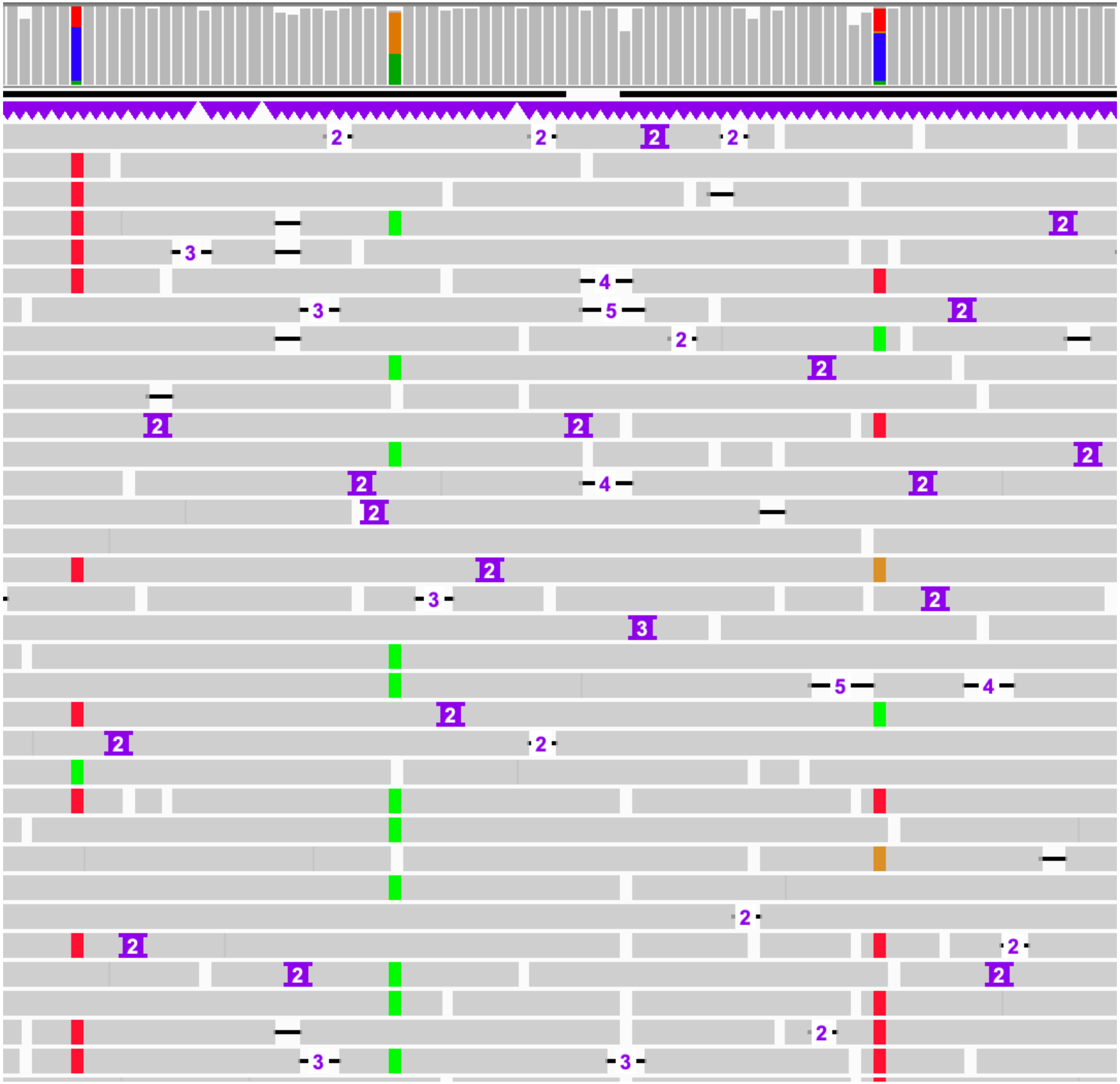
Quality trimming

- Removal of:
  - Low quality sequences
  - Adapters/barcodes
- Oxford nanopore: On-instrument (`guppy`)
- PacBio:
  - On-instrument
  - During CCS generation (`pbccs`)

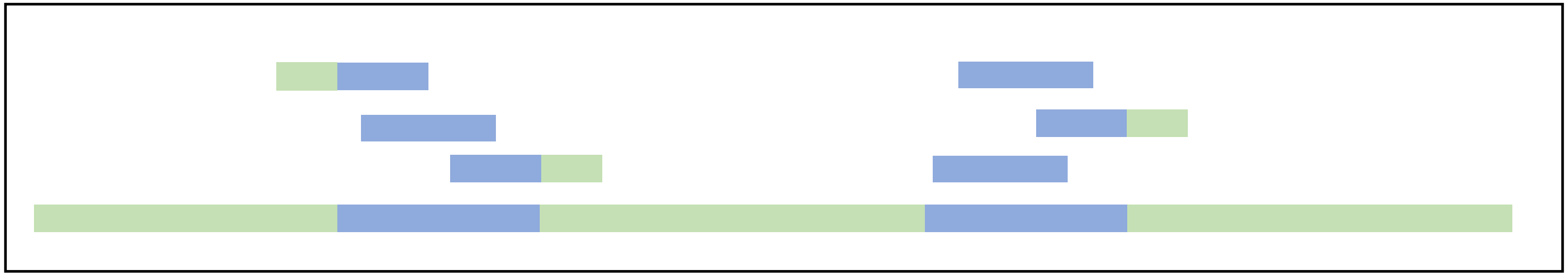
Read alignment

Long read aligner: minimap2





Mapping quality



$$MAPQ = -10log_{10}\Pr\{mapping\ position\ is\ wrong\}$$

$$\Pr\{mapping\ position\ is\ wrong\} = 10^{\frac{-MAPQ}{10}}$$

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

sam header

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:U00096.3     LN:4641652
@PG     ID:bowtie2      PN:bowtie2      VN:2.4.1        CL: bowtie2-
align-s --wrapper basic-0 -x ref.fasta -1 reads_1.fastq -2
reads_2.fastq"
```

SAM column	**example**
read name	SRR519926.5
flag	89
reference	chr20
start position	61
mapping quality	42
CIGAR string	150M
reference name mate is mapped	=
start position mate	476
fragment length	515
sequence	CATCACCATTCCCAC
base quality	@>4:4C@89+&9CC@
optional	AS:i:-2
optional	XN:i:0

Question 11

samtools

- Convert .sam files into (a.o.)
    - .bam (compressed .sam)
    - .fastq
- Subset alignments based on:
    - flag
    - region
- Ordering
- Mark alignment duplicates
- And many other things

Long-reads & fastq

- fastq format is limited to:
    - base
    - base-quality
- Long-read technologies -> need to store more information:
    - PacBio: (unaligned) bam
    - ONT: fast5/pod5/bam/rich fastq

Methylation calling

- PacBio – always done
- ONT Remora
  - https://github.com/nanoporetech/remora
  - https://nanoporetech.com/sites/default/files/s3/literature/epigenetics-workflow.pdf
- Stored in bam file (MM and ML tags)

Group work preference

Fill out the google form:

https://forms.gle/YXV5DwBe5DeD3Yx76