Long-read sequence analysis

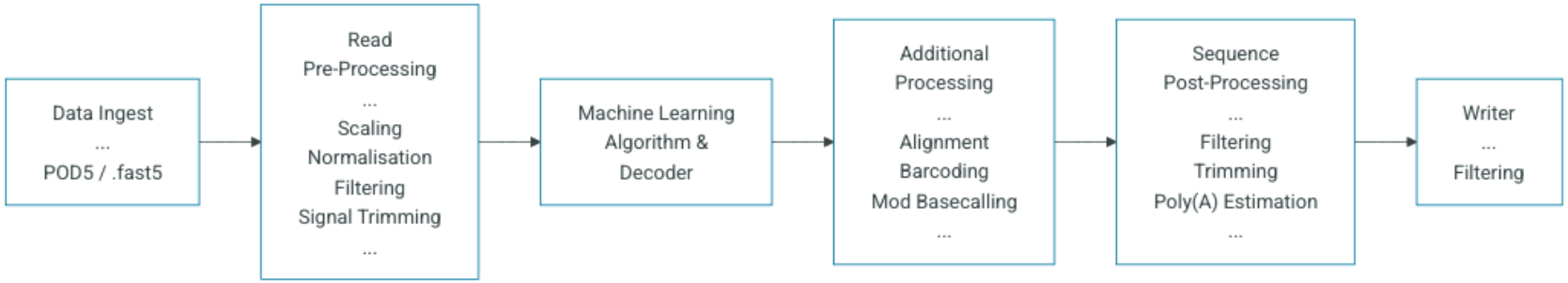
File formats and QC

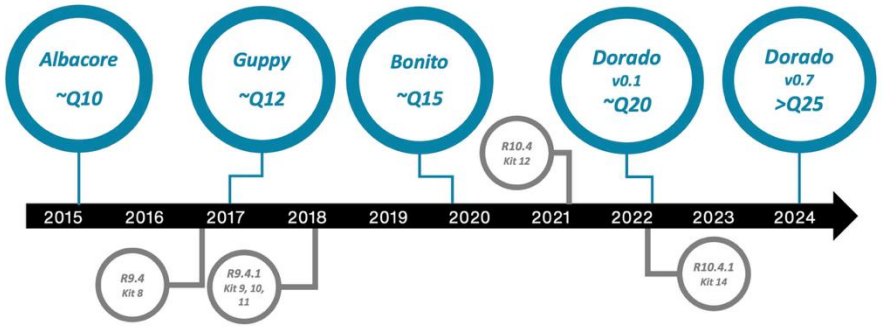
Raw file formats

- PacBio:
  - unaligned BAM/fastq (outdated)
- ONT:
  - POD5 (storing nanopore sequencing data)
  - FAST5 (HDF5, outdated)
  - Base calling: MinKNOW (guppy)/dorado/third party → unaligned BAM/fastq (outdated)

ONT basecalling pipeline

- DORADO (oxford Nanopore basecaller)



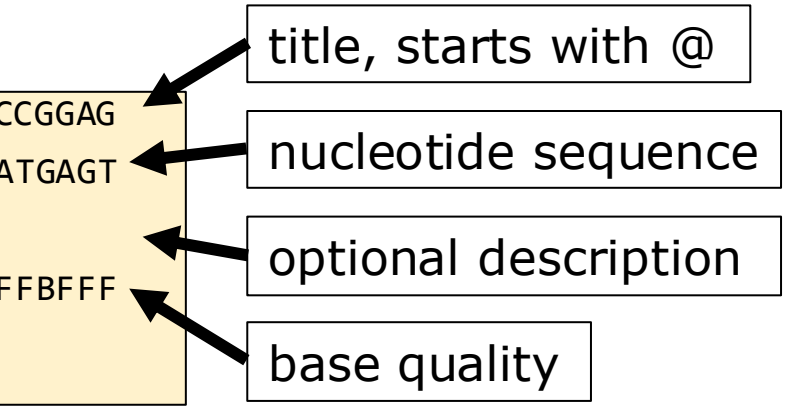


- Other basecallers (based on neural networks)

fastq

reads.fastq

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



fastq

fasta + basequality (fasta + q = fastq)

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

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

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

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

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 Dorado
  - https://nanoporetech.com/sites/default/files/s3/literature/epigenetics-workflow.pdf
- Stored in bam file (MM and ML tags)

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

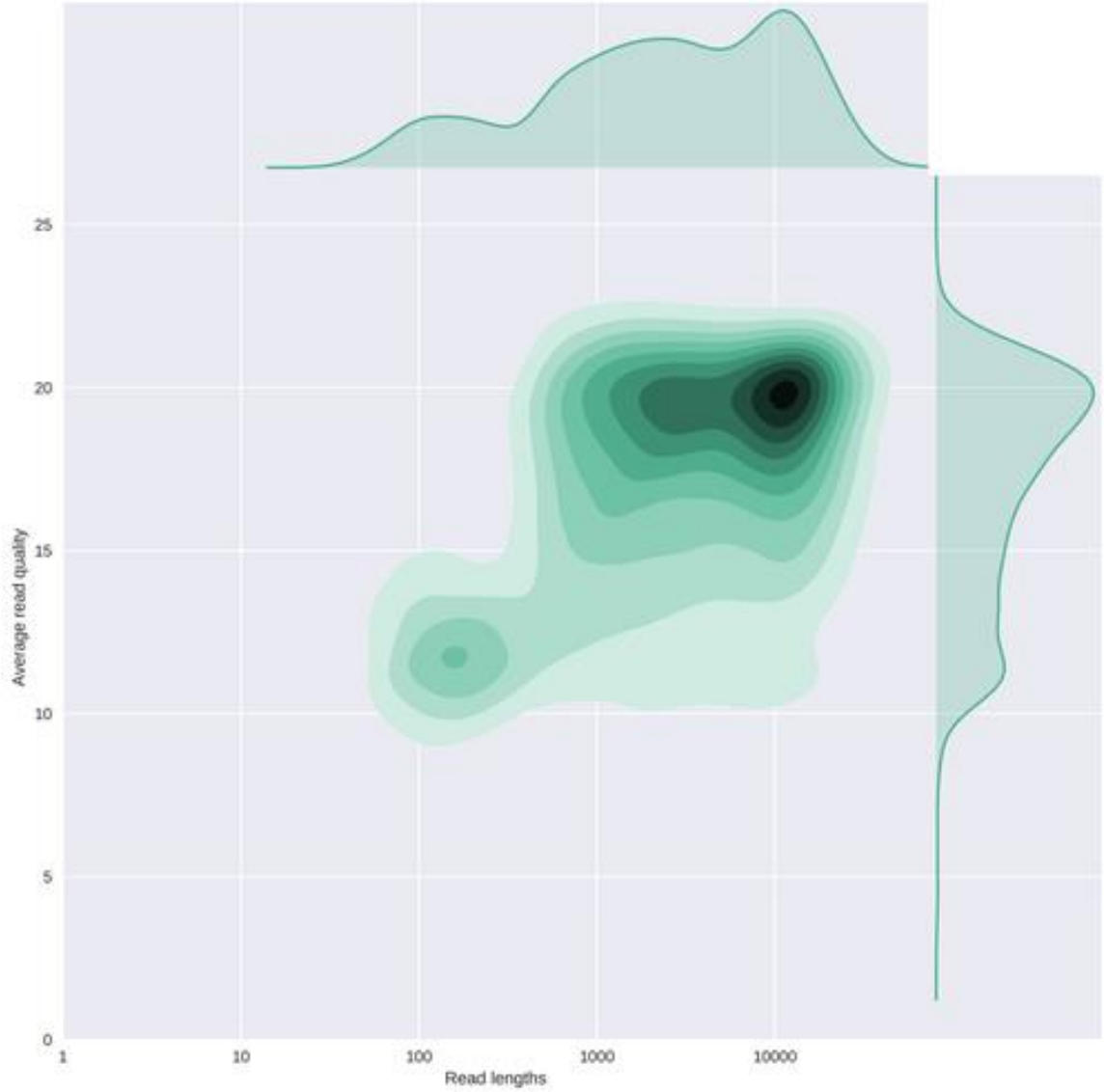
- NanoPlot (https://github.com/wdecoster/NanoPlot)
  - Takes many input formats
  - Basic statistics

- PycoQC (https://github.com/a-slide/pycoQC)
  - Specific for ONT
  - Requires so-called `sequencing_summary` file

- NanoPack (https://github.com/wdecoster/nanopack)
  - Set of tools for visualisation and processing

- FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
  - Works also for long reads
  - Familiar output to most people

Read lengths vs Average read quality plot

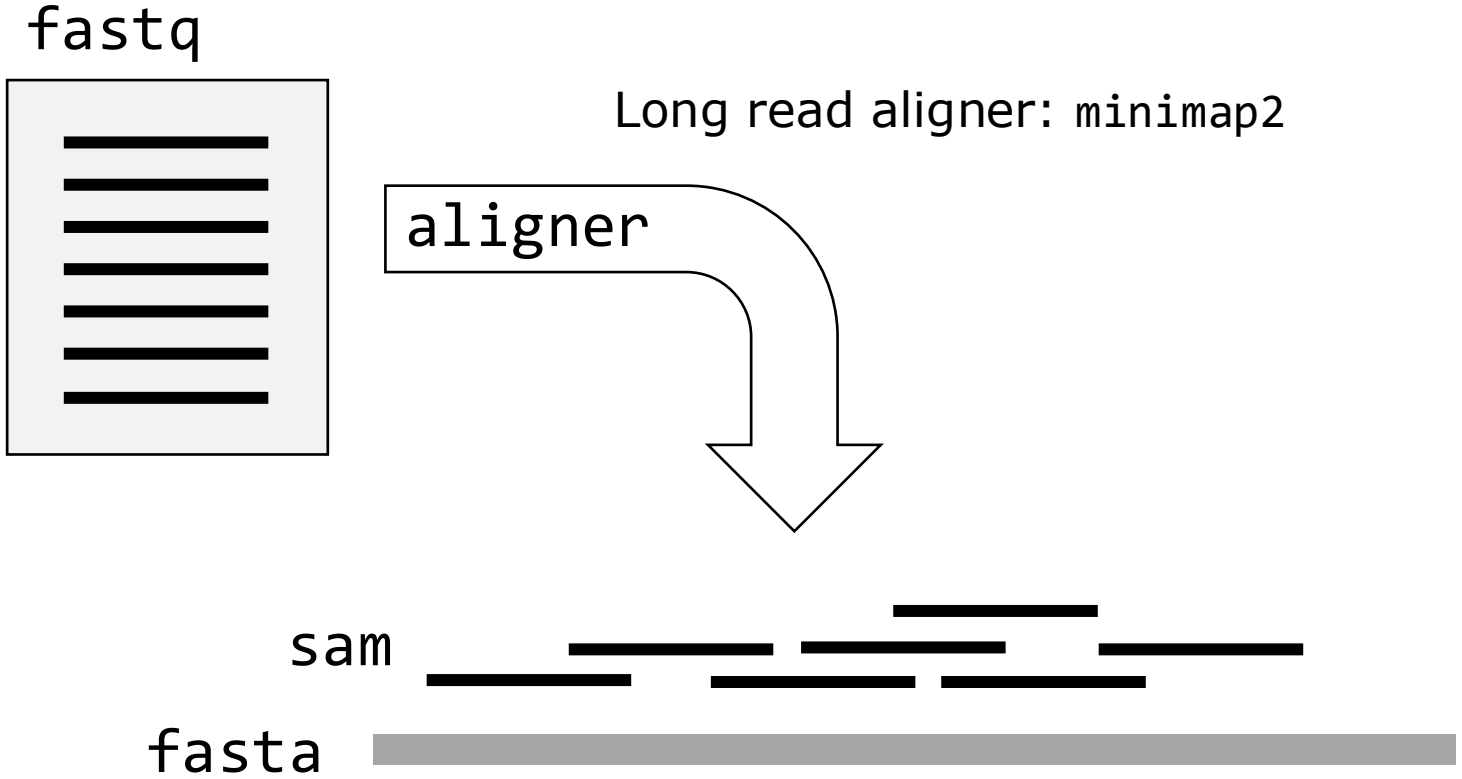


output of NanoPlot (https://github.com/wdecoster/NanoPlot)

Quality trimming

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

Read alignment



Read aligners

- Minimap2 (https://github.com/lh3/minimap2)
  - Widely used
  - Can struggle with repetitive and regions and rearrangements

Minimap2 Common Modes (Presets)

Mode	Usage Description
map-ont	For Oxford Nanopore long reads.
map-pb	For PacBio long reads (HiFi or CLR).
splice	For spliced alignment, typically for RNA-Seq reads to a genome (includes intron handling).
asm5 / asm10	For aligning assembled genomes to other genomes (low divergence).

Read aligners

- Pbbm2 / Dorado
  - minimap2 wrappers developed by PacBio and ONT
  - minimal flexibility in terms of alignment options, less stable than minimap (more frequent updates)
- FLAIR
  - Wrapper around minimap2
  - Uses specific options for spliced alignment
  - Requires FASTA/FASTQ read files
- Winnowmap2 and VACmap
  - Newer aligners
  - Improved accuracy in challenging genomic regions
- Vulcan
  - Use minimap2 for initial alignment
  - Reprocess alignments with other alignment tools

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

Mapping quality

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

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

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

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

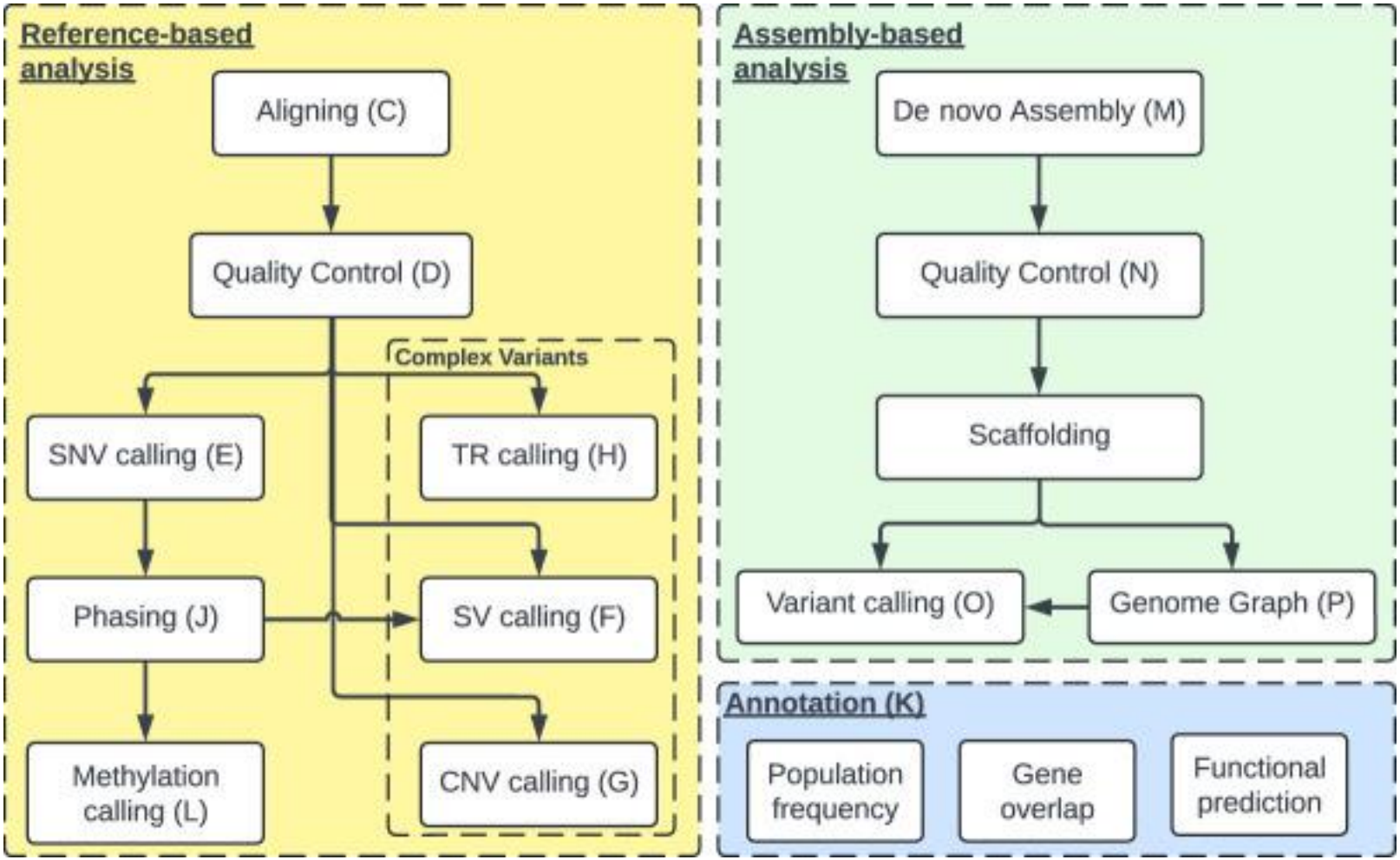


Image from: Mahmoud M., Agustinho DP., Sedlazeck FJ. 2025; Genome Research

Exercises

QC and alignment