NGS – variant analysis

Sequencing and alignment

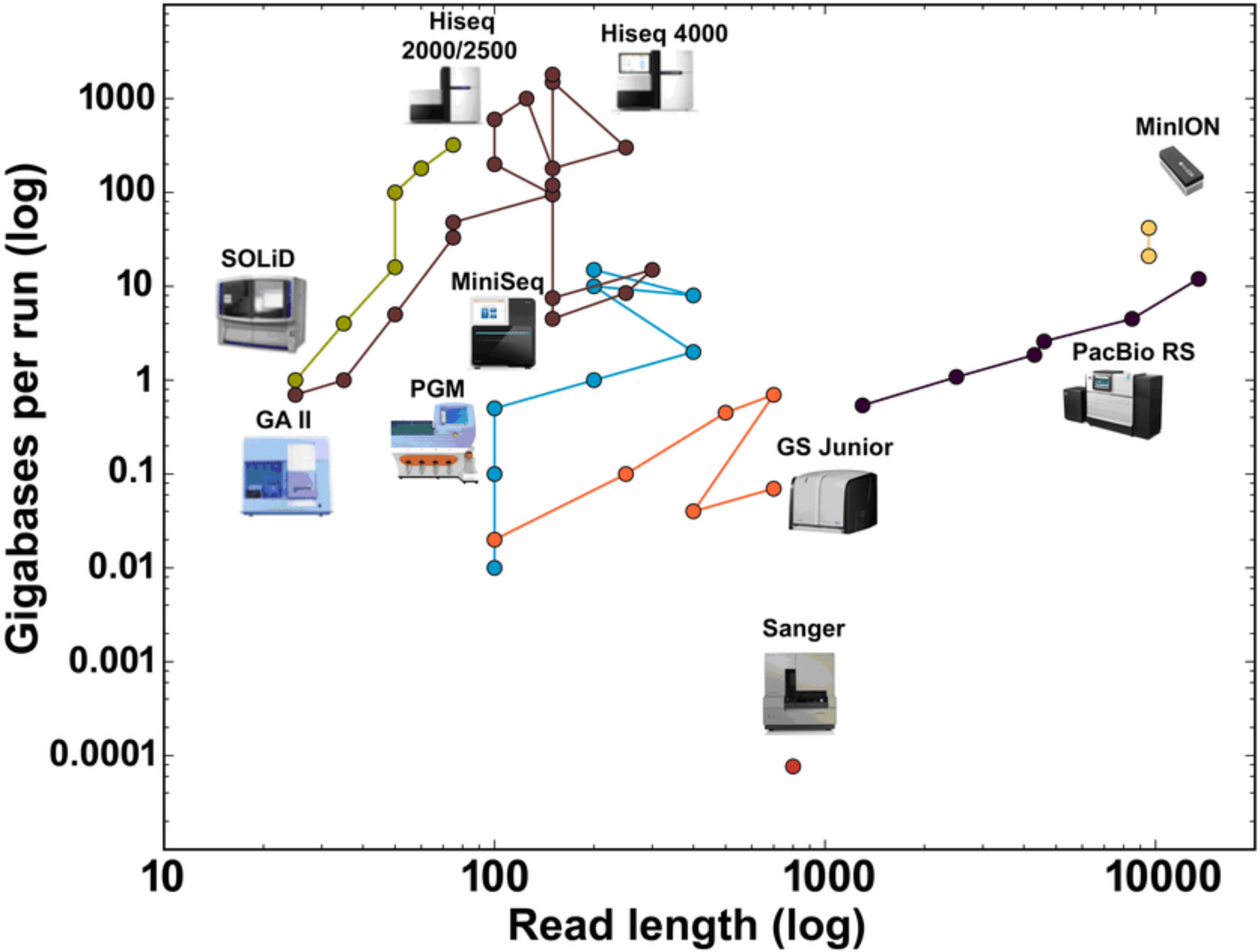


Image from: G. Silva (2016)

Illumina sequencing

- Sequencing-by-synthesis: 2nd generation sequencing
- Massive throughput: up to 500x10^9 bases/run
- Most used platform today

illumina®

Illumina sequencing

- 50 – 300 bp
- Paired-end (or single-end)

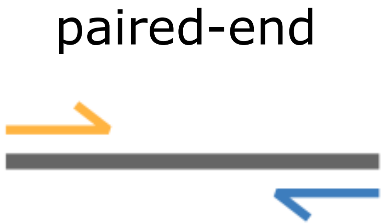


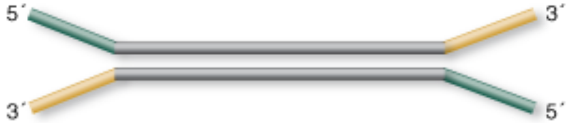
Image from: Illumina (2020)

Illumina libray prep

shear DNA

↓

Ligate adapters

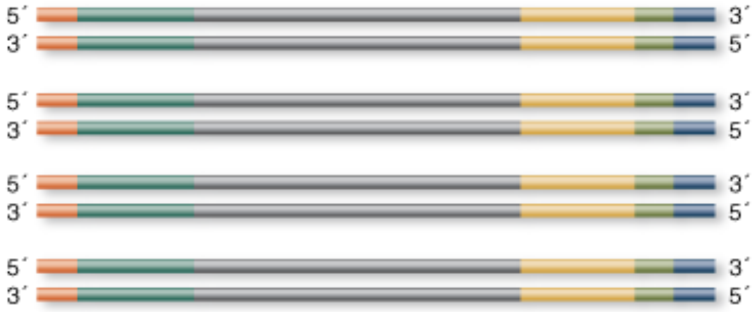


↓

Barcode + p5/p7 sites

↓

PCR: 8-16 cycles



↓

Sequencing

Images from: New England Biolabs Inc. (2020)

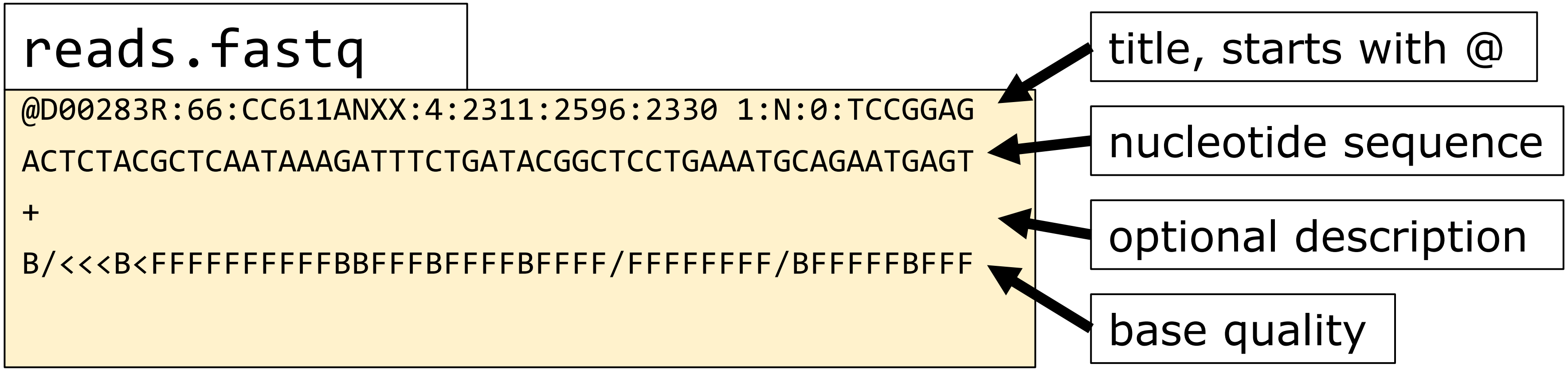
fastq

fasta + basequality (fasta + q = fastq)

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

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

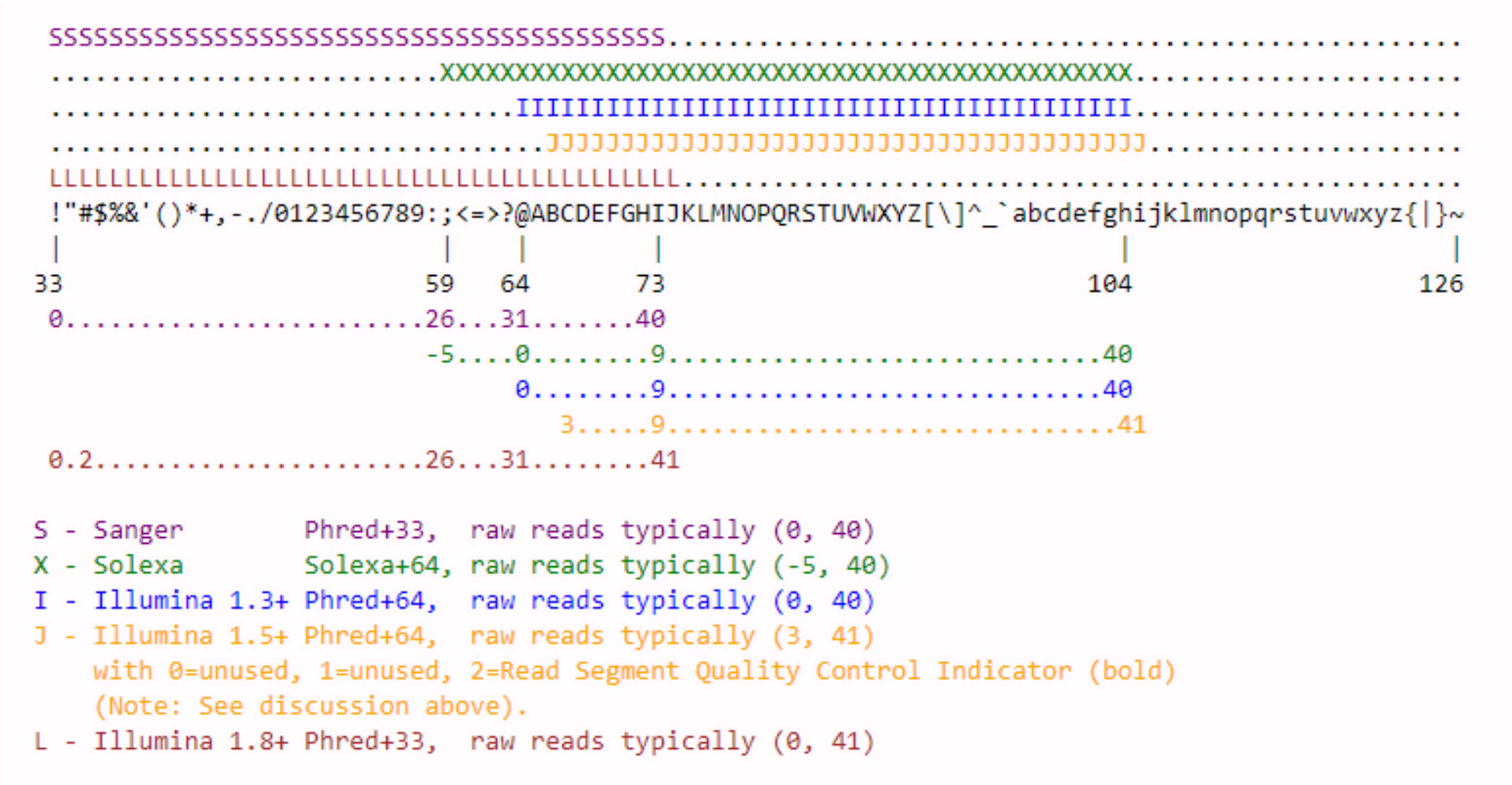
fastq



reads.fastq

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

Base quality (phred)



```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126
  0.........................26...31.......40
                           -5....0........9..............................40
                                 0........9..............................40
                                    3.....9...............................41
  0.2.......................26...31........41

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Image from: Wikipedia (https://en.wikipedia.org/wiki/FASTQ_format)

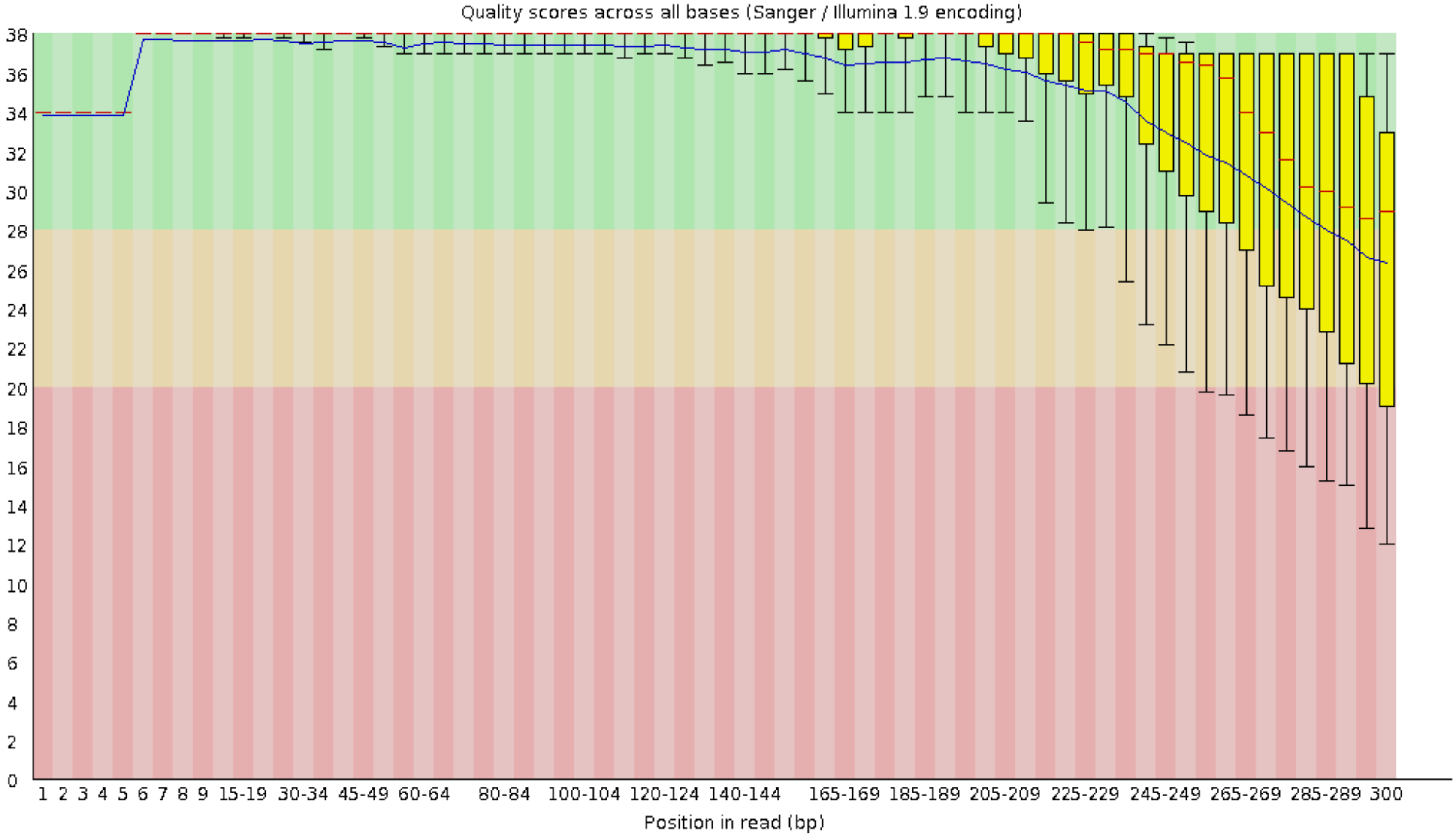
Quiz Question 4

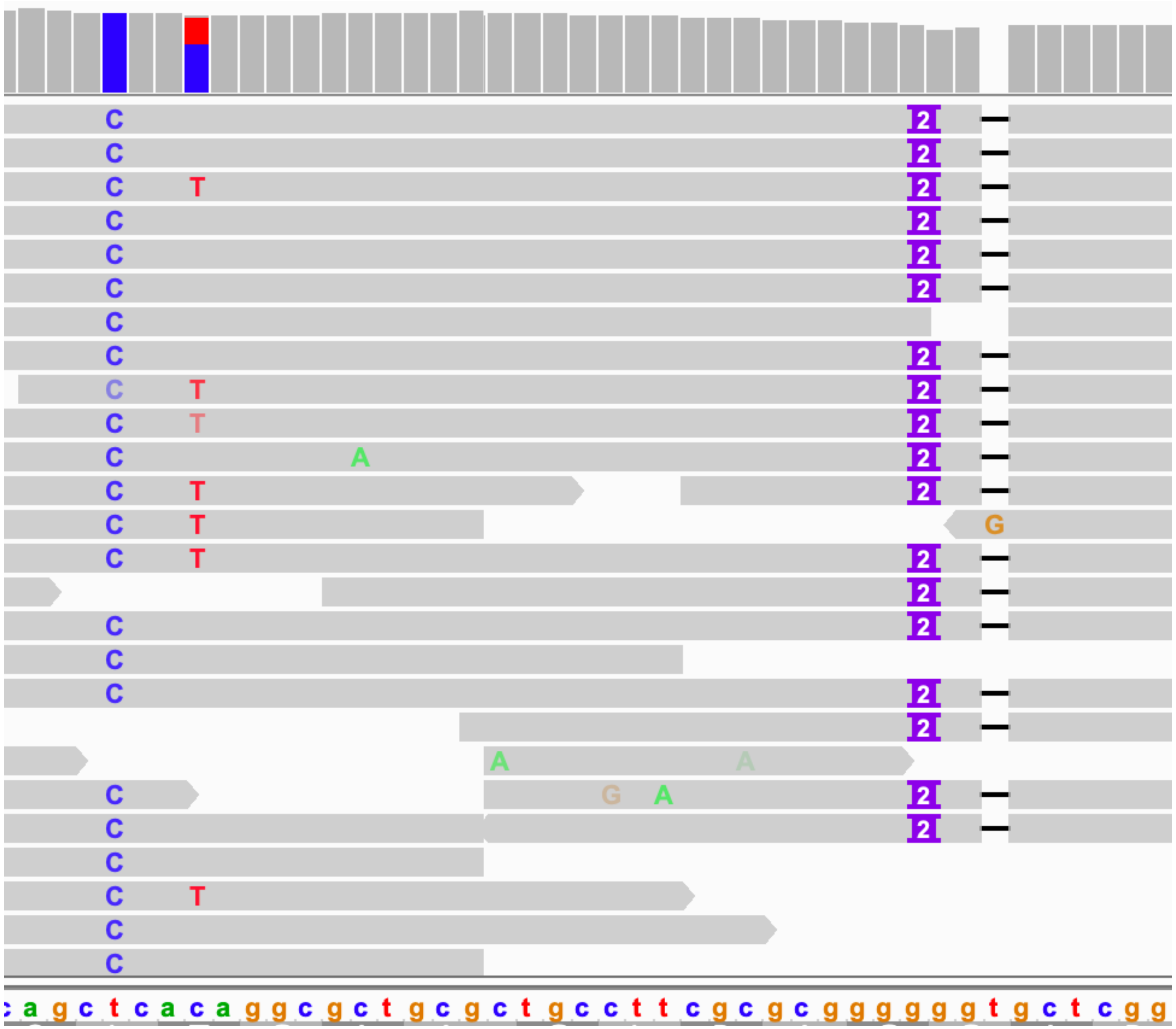
With `grep -c "^>"` it is easy to count the number of sequences in a fasta file. Why doesn’t `grep -c "^@"` work for fastq sequences?

A. @ is a special character for regular expressions

B. The @ can also occur elsewhere in a fastq file

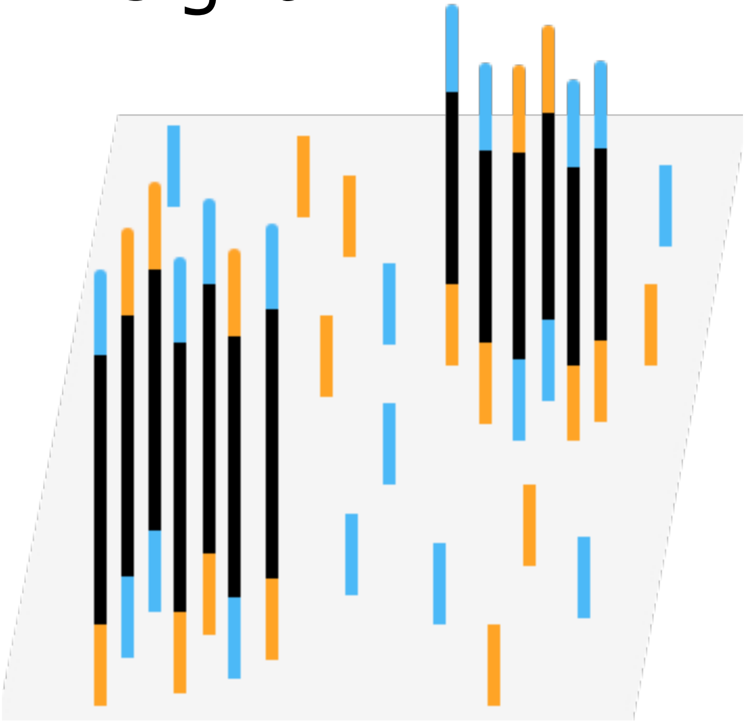
C. There is no specific title start symbol for fastq files

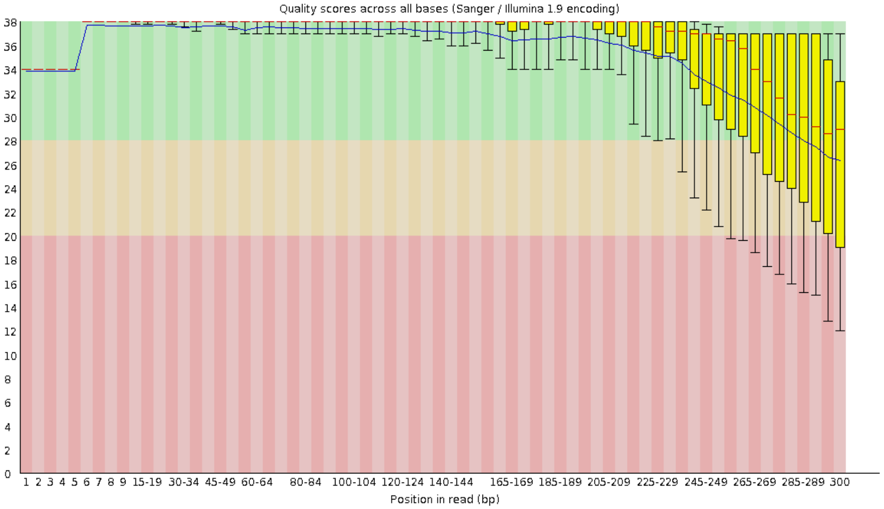

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 80-84 100-104 120-124 140-144 165-169 185-189 205-209 225-229 245-249 265-269 285-289 300
Position in read (bp)


C
T
A
G
2
c a g c t c a c a g g c g c t g c g c t g c c t t c g c g c g g g g g g t g c t c g g

Illumina - limitations

- Bridge amplification
- Lengths are limited by out-of-phase of signal



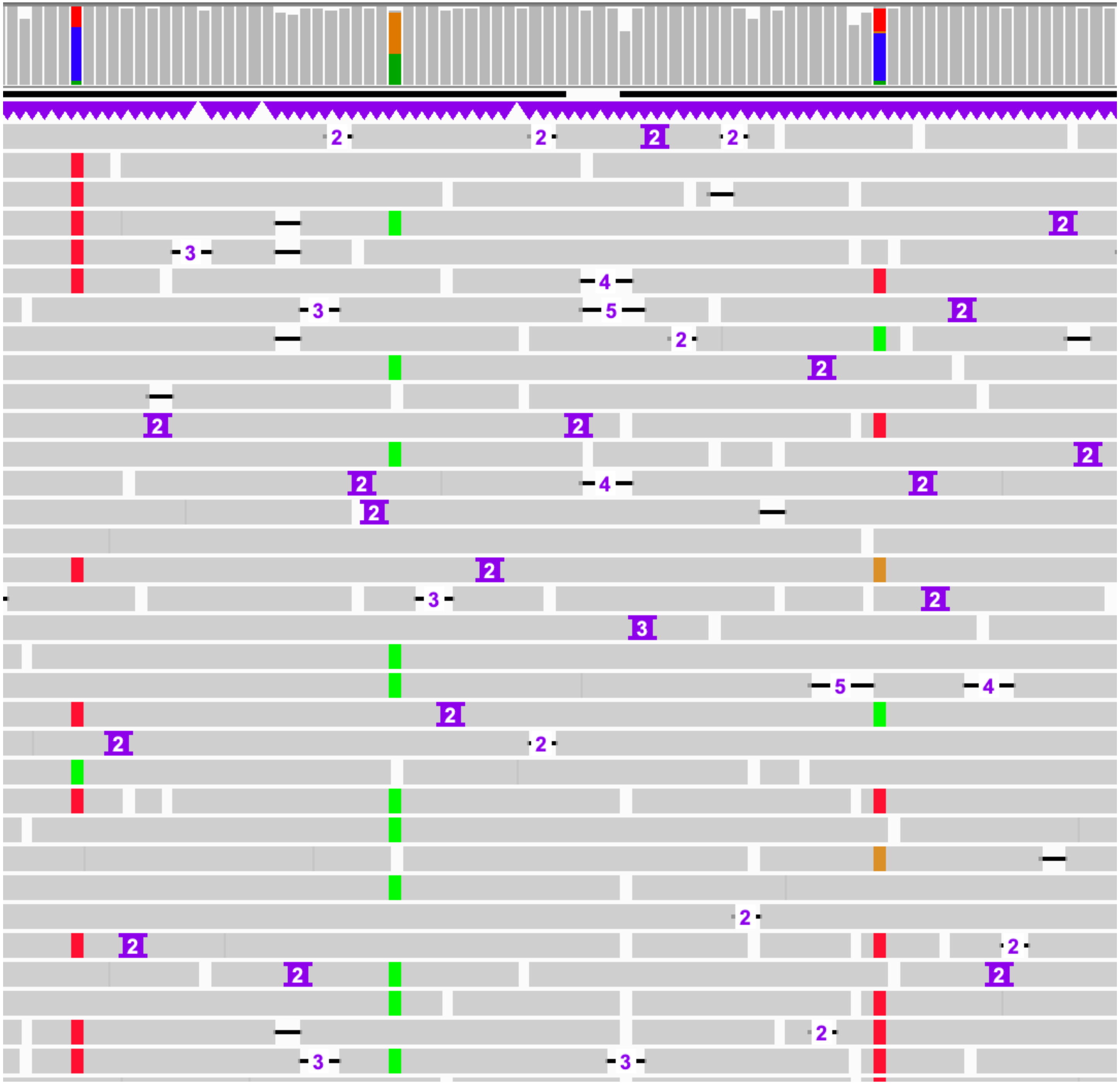


Long reads (3rd generation)

- Crux: maximizing signal from a single-molecule base read-out
- Single molecule, so no out-of-phase signal
- Two frequently used platforms:
  - PacBio SMRT sequencing
  - Oxford Nanopore Technology

PACBIO®

Oxford NANOPORE Technologies


2
2
2
2
3
4
3
5
2
2
2
2
2
2
4
2
2
2
2
3
2
3
5
4
2
2
2
2
2
2
2
2
3
3

Question 4

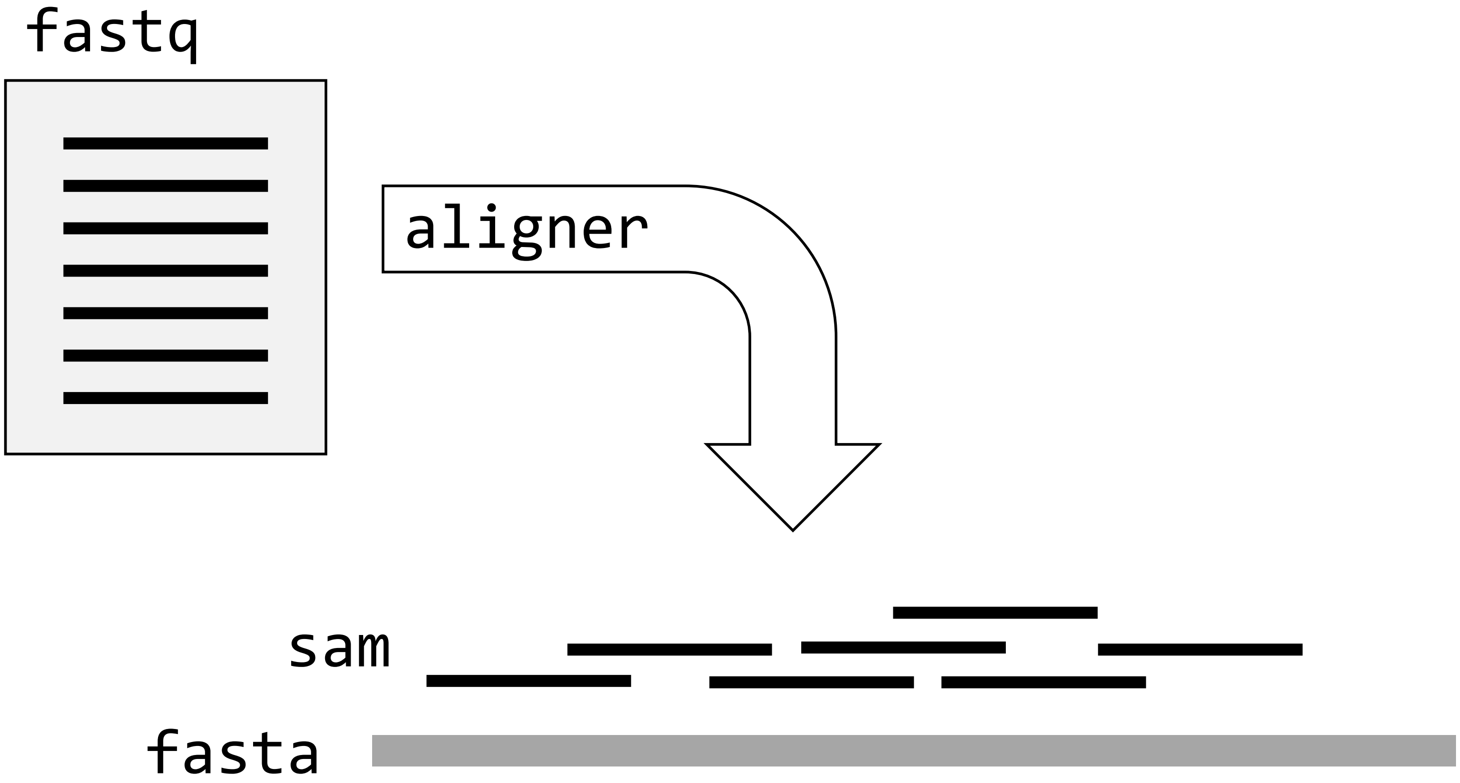
Why are INDELs more difficult to detect from alignments compared to SNPs?

A. Because there is no base quality
B. Because it is difficult to know the correct local alignment

Long reads

- More error -> difficulties for variant analysis
- But:
  - PacBio CCS: high baseQ + no bias
  - Long reads can have higher mapping qualities
  - Long reads improve haplotyping

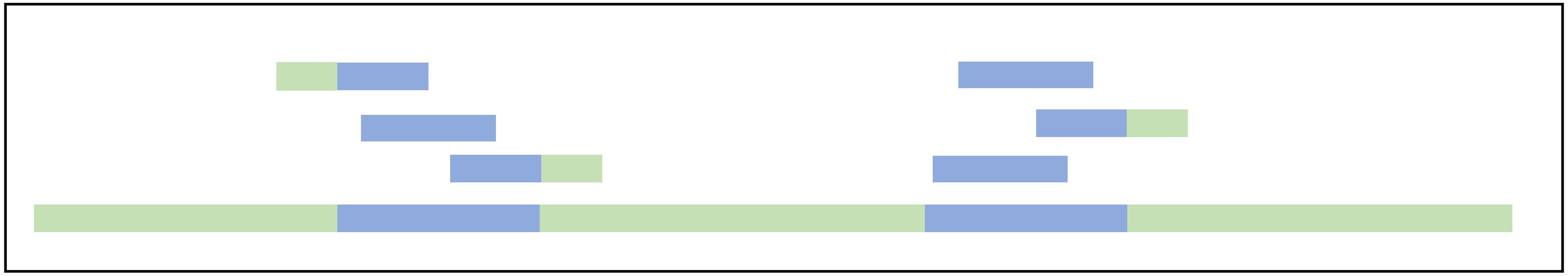
Read alignment (phred)



Software

- Basic alignment:
    - bowtie2
    - bwa-mem
- Long reads:
    - minimap2

Mapping quality

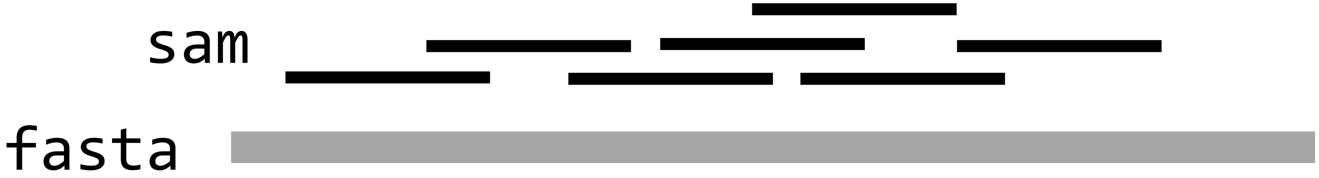


$$MAPQ = -10 log_{10} \Pr\{mapping\ position\ is\ wrong\}$$

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

sam

sequence alignment format



sam header

```
@HD	VN:1.0	SO:coordinate
@SQ	SN:U00096.3	LN:4641652
@PG	ID:bowtie2	PN:bowtie2	VN:2.4.1
CL: "/opt/miniconda3/envs/ngs/bin/bowtie2-align-s \
--wrapper basic-0 \
-x /home/ubuntu/ecoli/ref_genome//ecoli-strK12-MG1655.fasta \
-1 /home/ubuntu/ecoli/trimmed_data/paired_trimmed_SRR519926_1.fastq \
-2 / home/ubuntu/ecoli/trimmed_data/paired_trimmed_SRR519926_2.fastq"
```

SAM column	**example**
read name	SRR519926.5
flag	89
reference	U00096.3
start position	61
mapping quality	42
CIGAR string	214M
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 5

Can you technically regenerate the fastq file out of the SAM file? And can you regenerate the reference sequence (fasta) file from the SAM file?

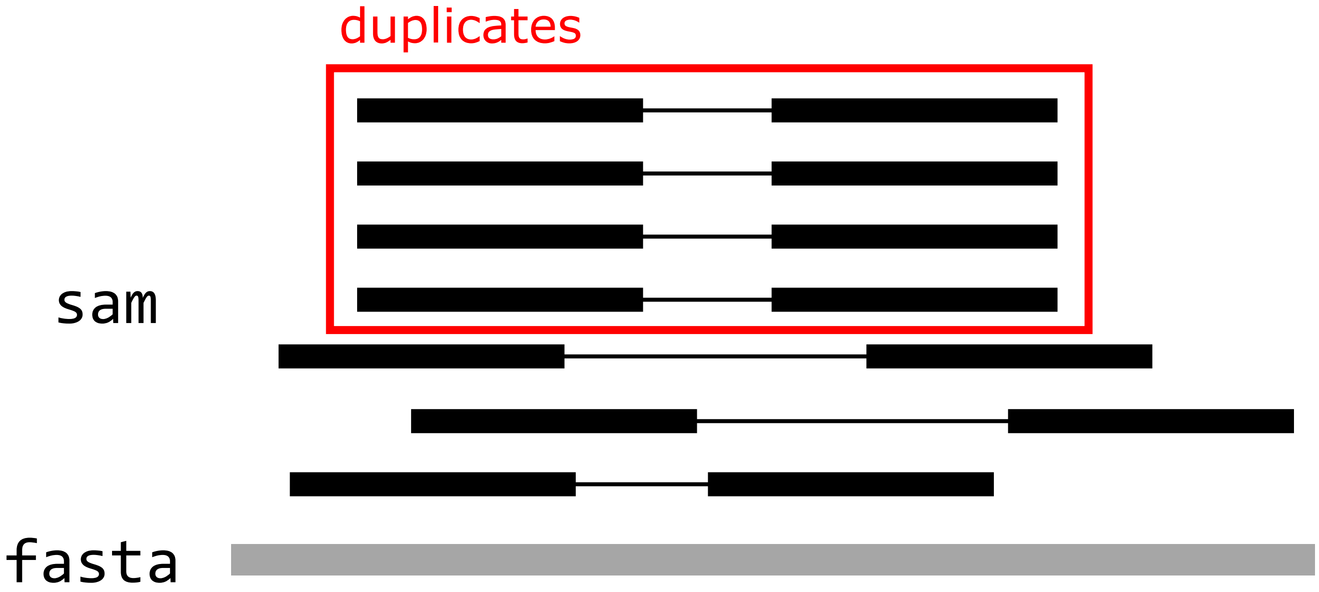
A. Only the fastq file

B. Only the fasta file

C. Both files

D. None of those

Marking duplicates



Marking duplicates

- Variant calling: each read is an independent observation of the genome
- Duplicates (can) have the same molecular origin -> not independent
- Removing duplicates probably doesn’t have a big effect on variant analysis

Ebbert MTW et al. (2016) Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics.

Unique Molecular Identifiers

- UMI added before PCR reaction
- Detect PCR duplicates and PCR errors