NGS – variant analysis

Sequencing and alignment

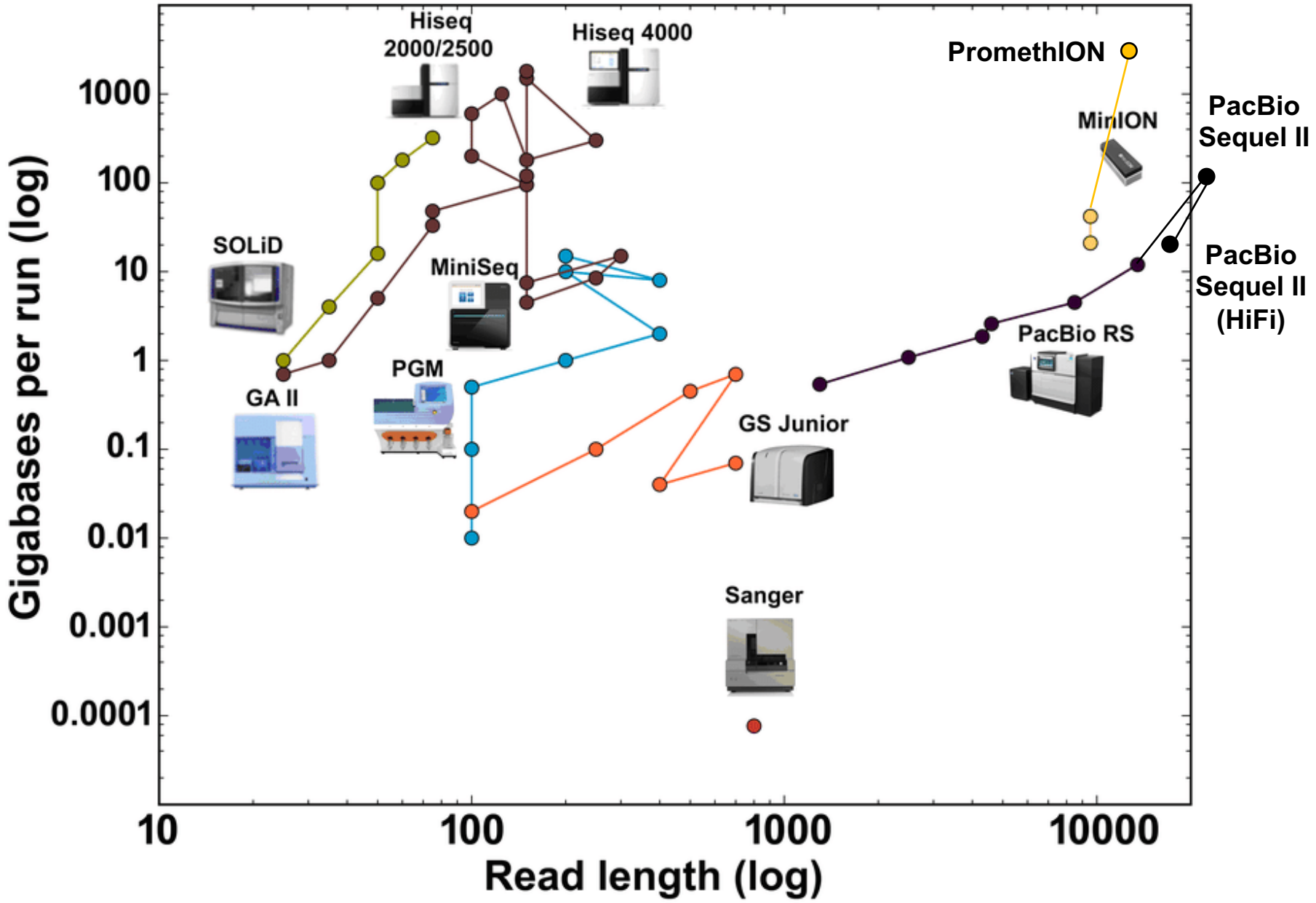


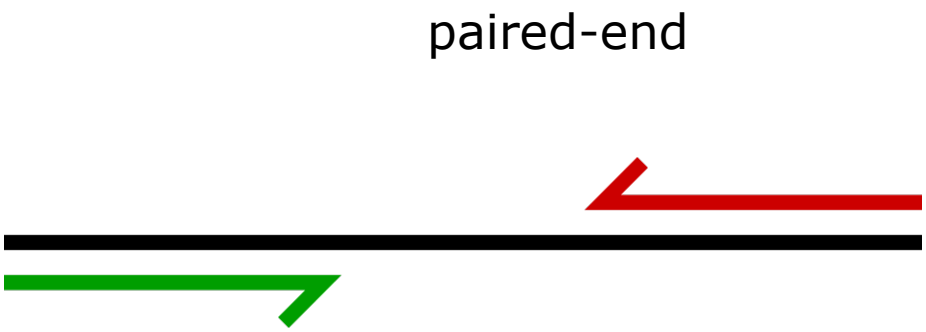
Image from: G. Silva (2016)

Illumina sequencing

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

Illumina sequencing

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

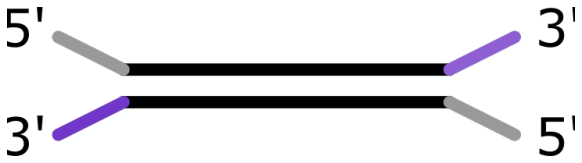


Illumina libray prep

shear DNA

↓

Ligate adapters



↓

Barcode + p5/p7 sites

↓

PCR: 8-16 cycles

↓

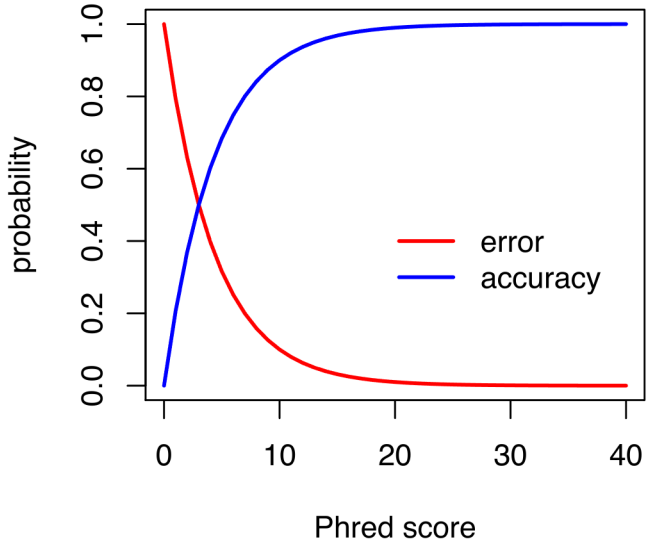
Sequencing

fastq

fasta + basequality (fasta + q = fastq)

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

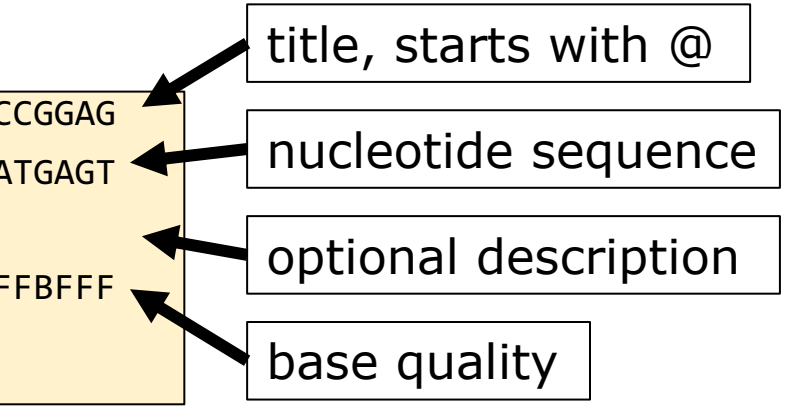
$$-10log_{10}(0.01) = 20$$
$$-10log_{10}(0.1) = 10$$
$$-10log_{10}(0.5) = 3$$



fastq

reads.fastq

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



Base quality (phred)

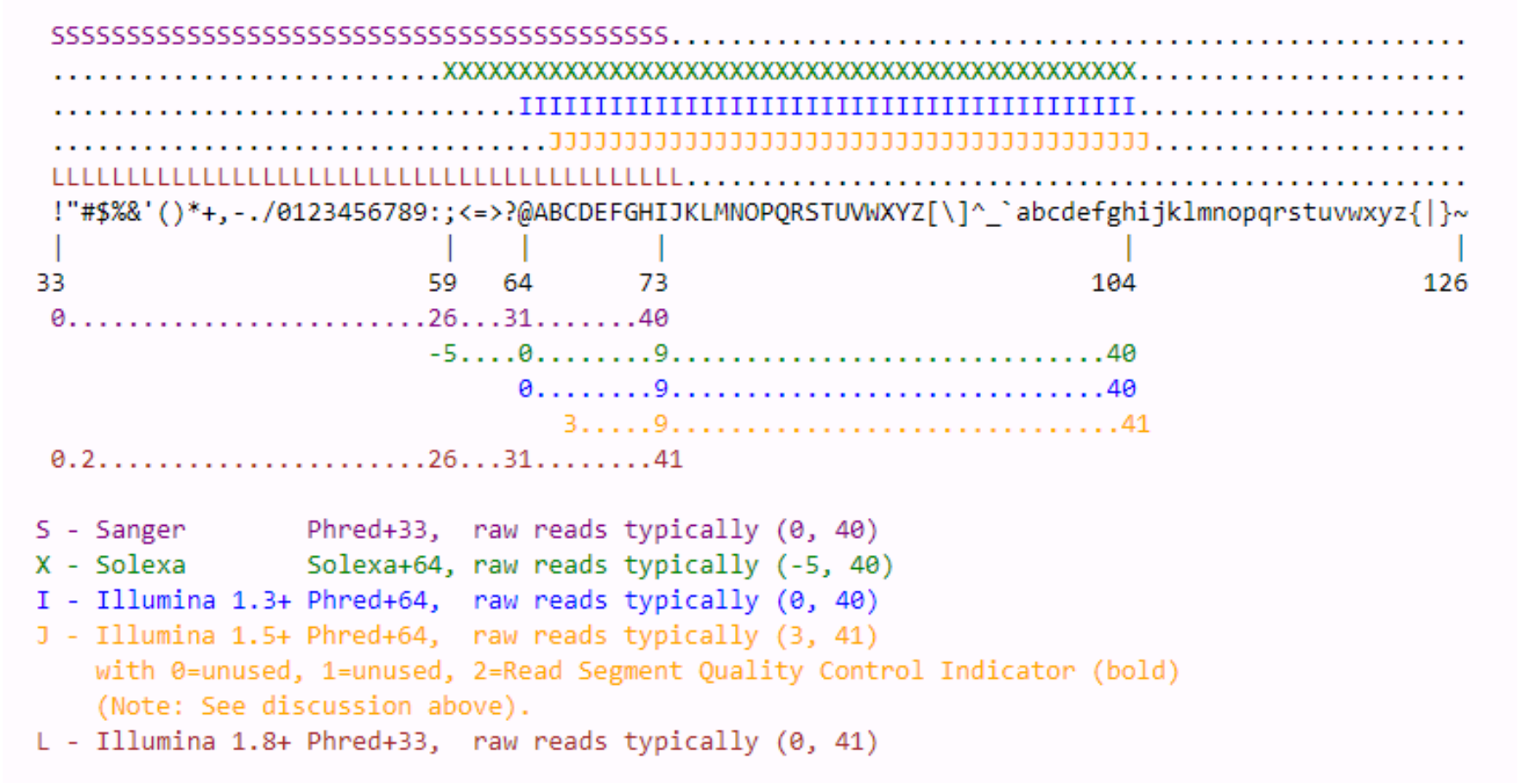
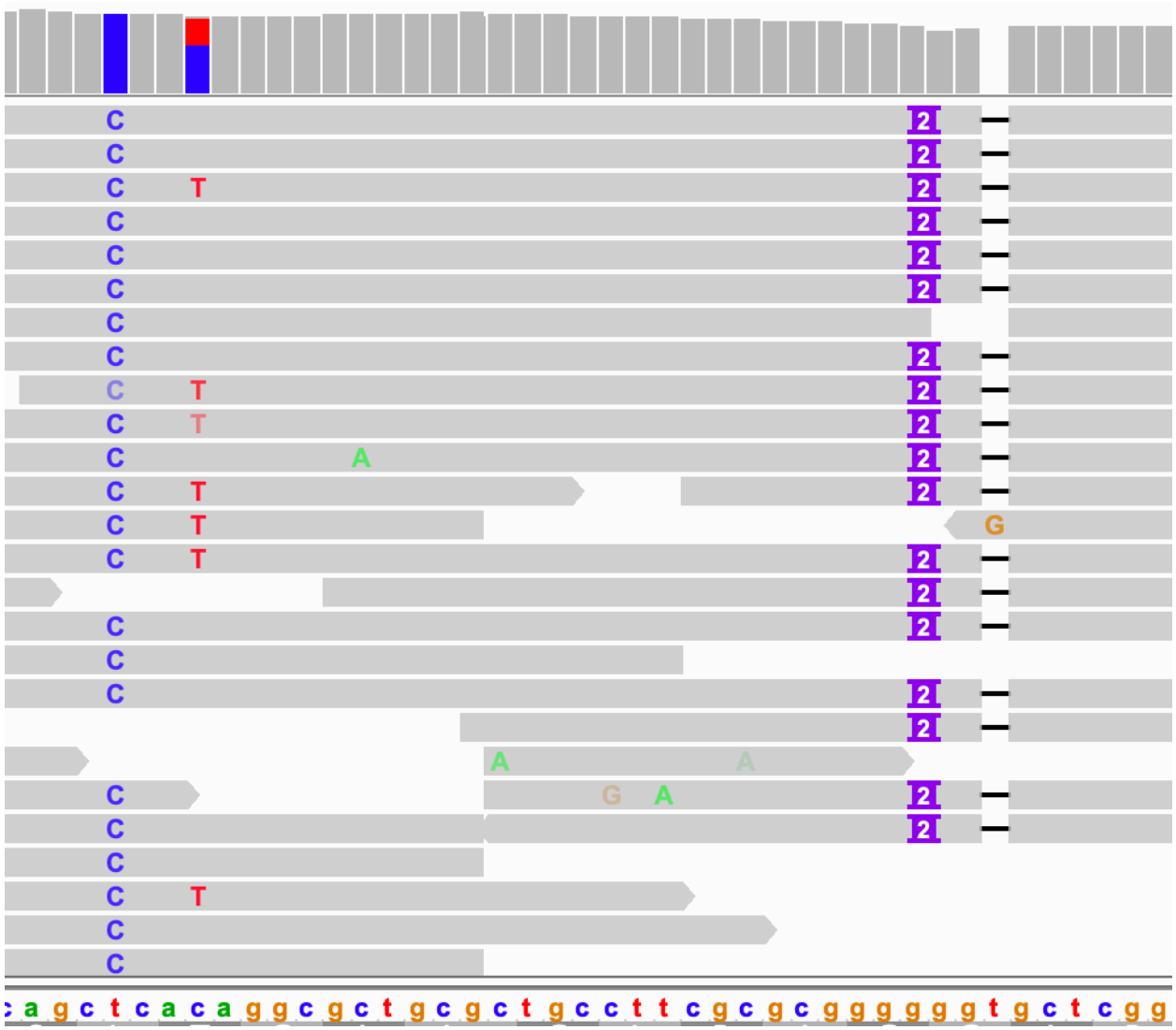
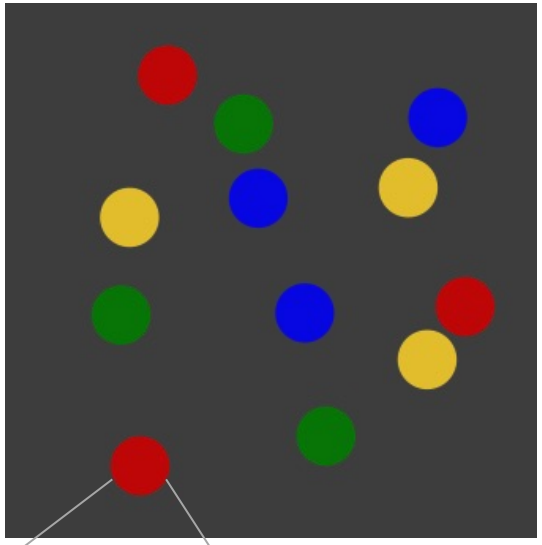
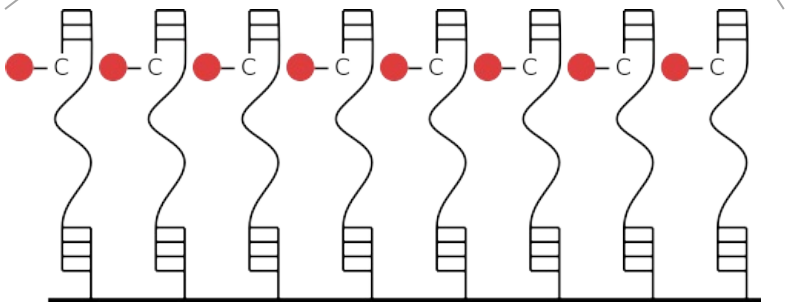


Image from: Wikipedia (https://en.wikipedia.org/wiki/FASTQ_format)

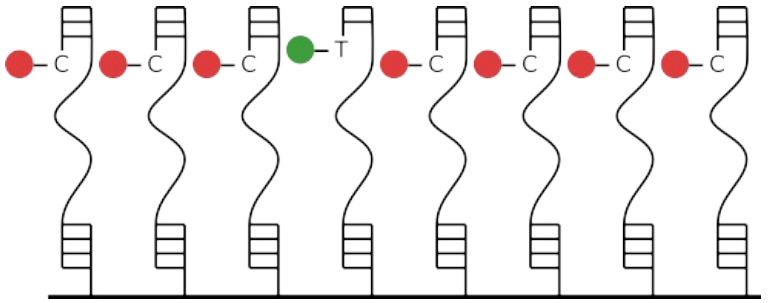
Quiz Question 6







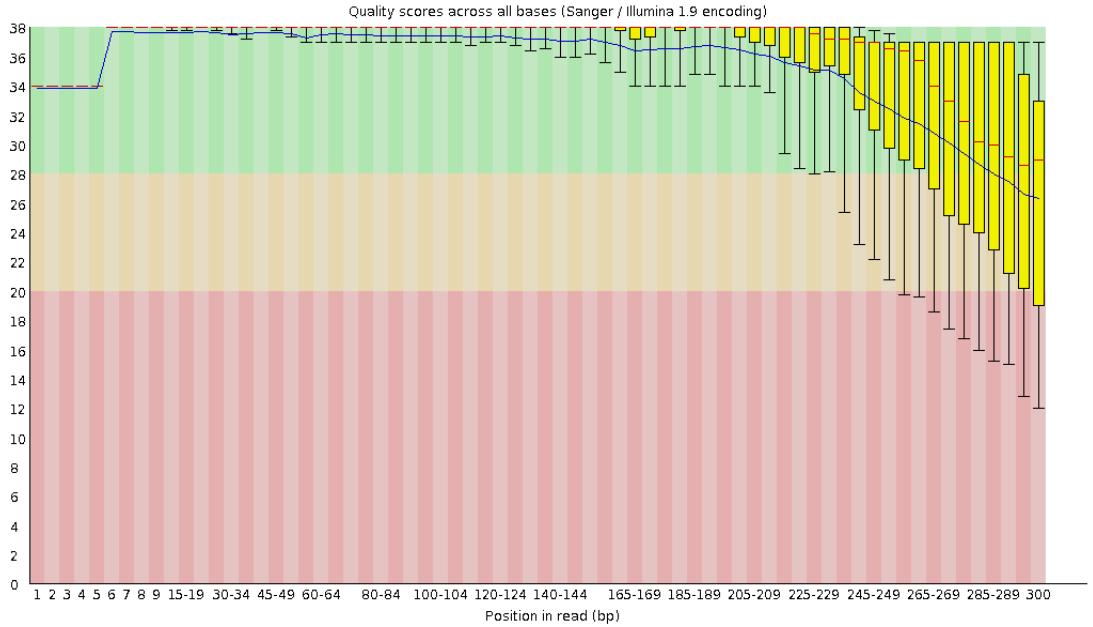
in phase



out of phase

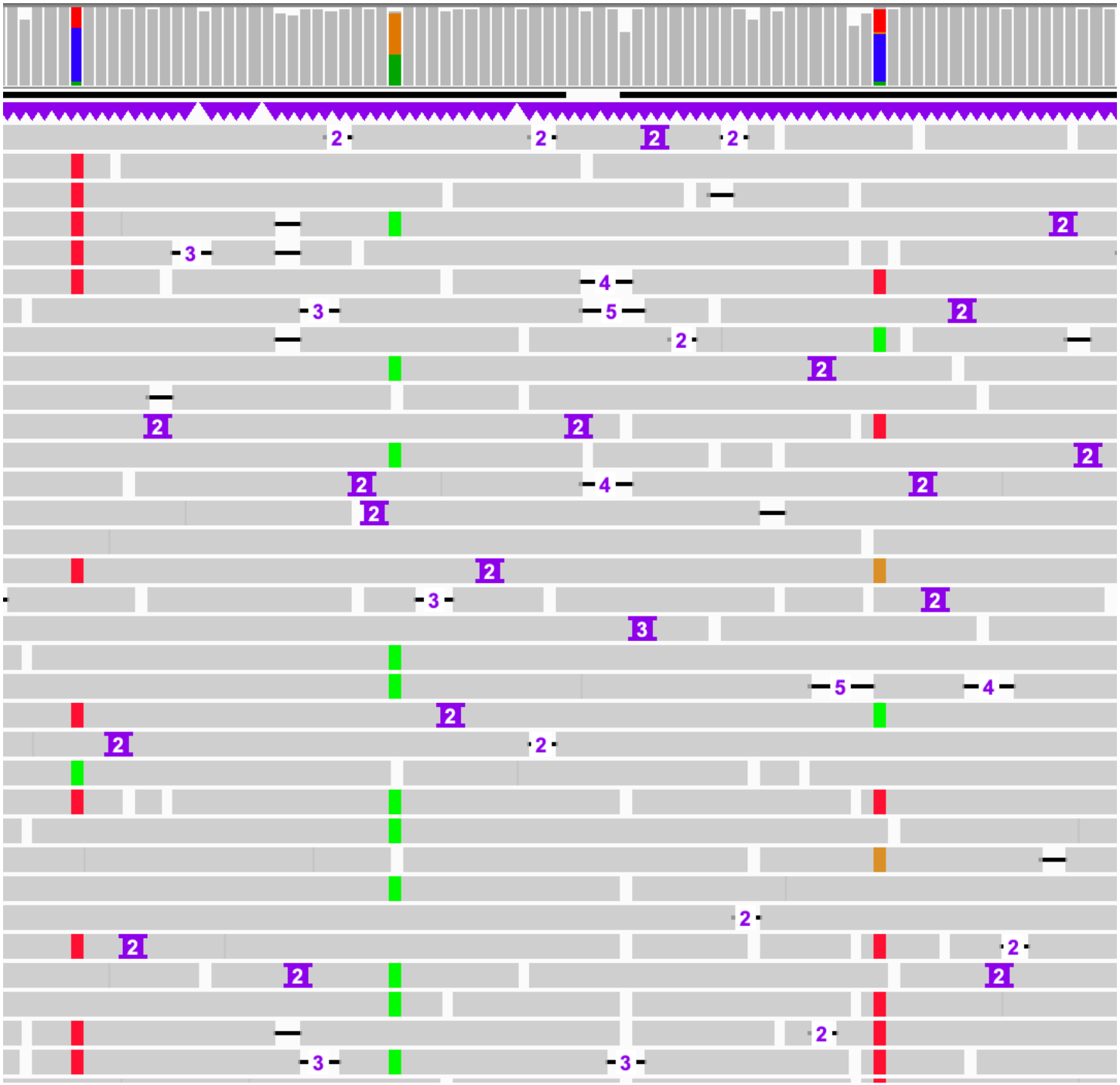
Illumina - limitations

- 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



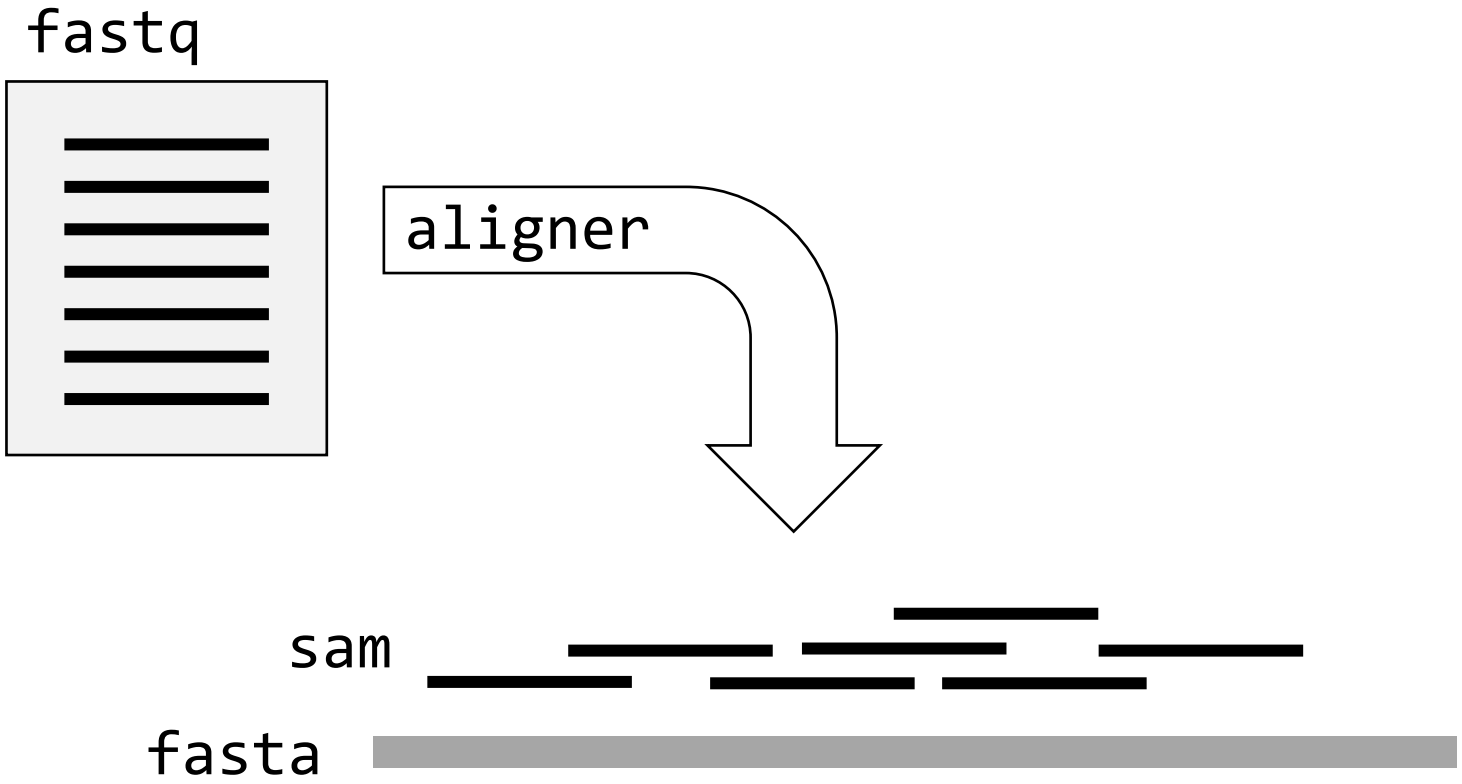
Long reads

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

What to sequence?

- Whole genome/metagenome
- Reduced representation:
  - Bait capture/whole exome sequencing
  - Restriction Enzyme based (e.g. RAD seq)
  - Amplicon sequencing
  - RNA-seq

Read alignment (phred)



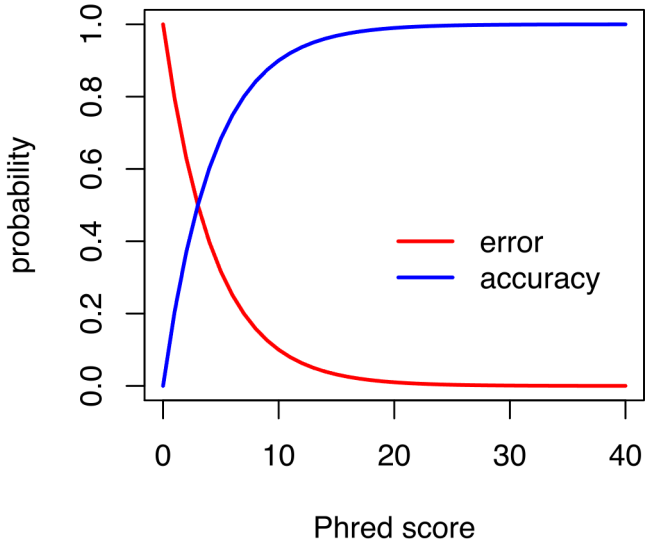
Software

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

Mapping quality

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

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



Quiz Question 7

sam

sequence alignment format

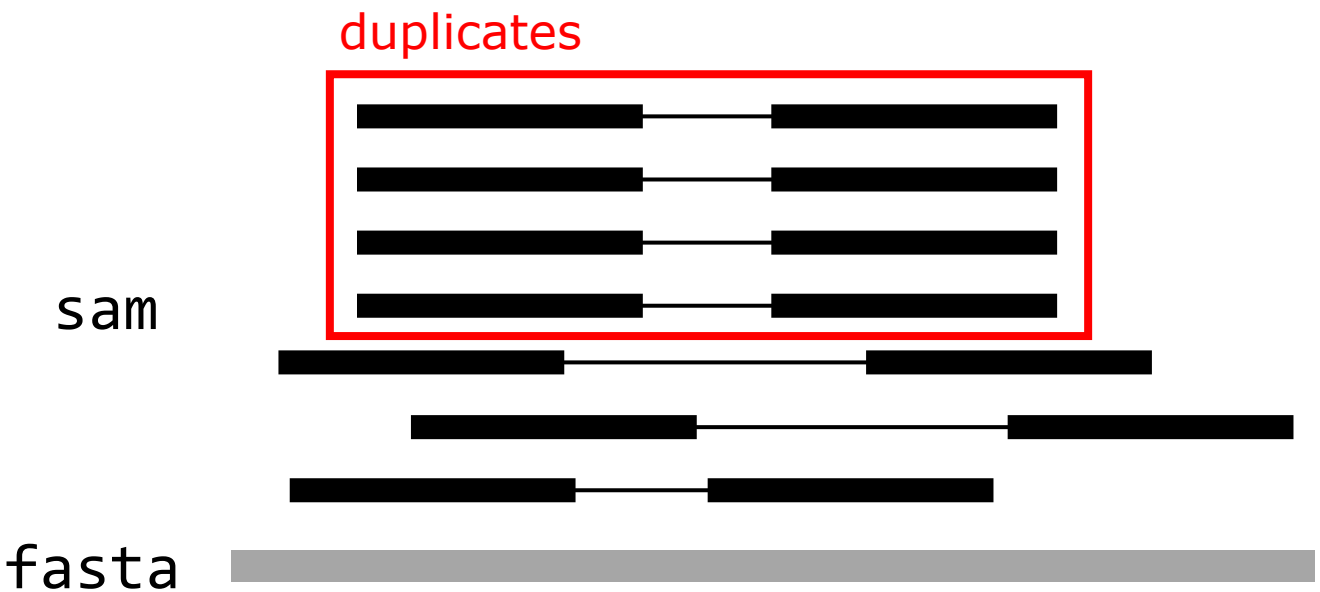
sam header

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:U00096.3     LN:4641652
@PG     ID:bowtie2      PN:bowtie2      VN:2.4.1        CL: 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

Quiz Question 8

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