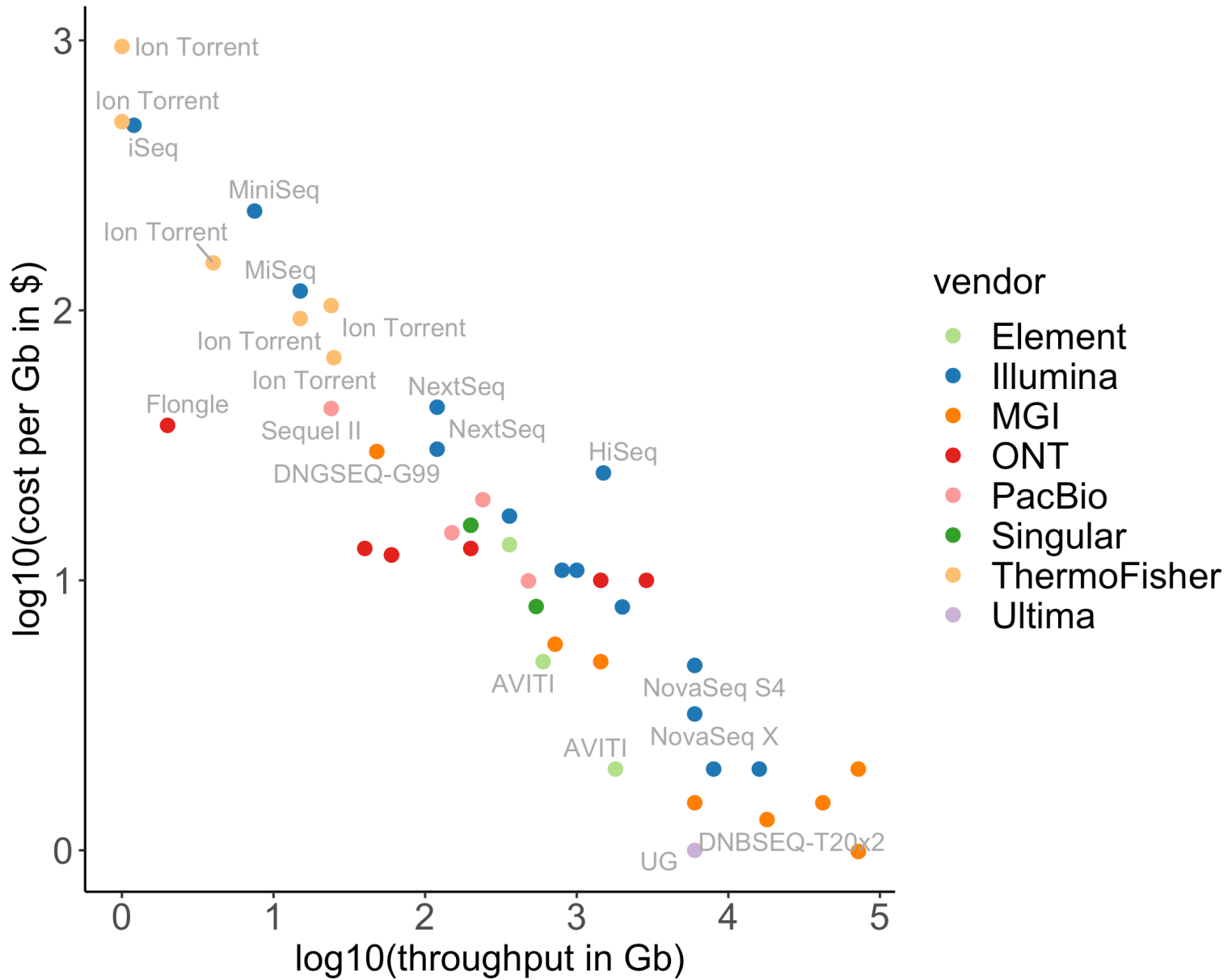


NGS - variant analysis

Sequencing and alignment



Sequencing-by-synthesis

- 2nd generation sequencing
- Massive throughput: up to 500×10^9 bases/run
- Illumina still most used platform today



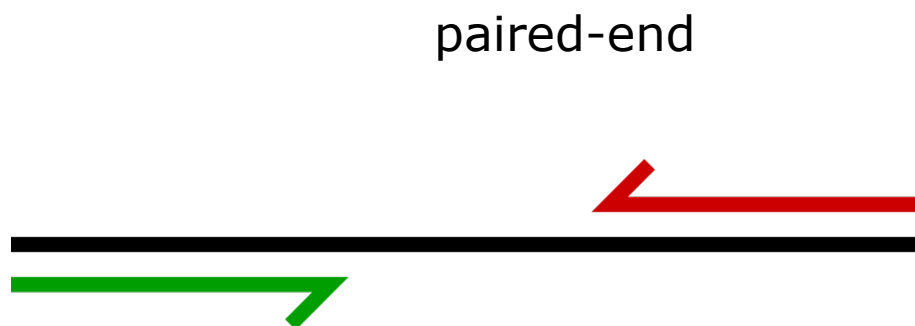
S I N G U L A R
G E N O M I C S

Sequencing-by-synthesis

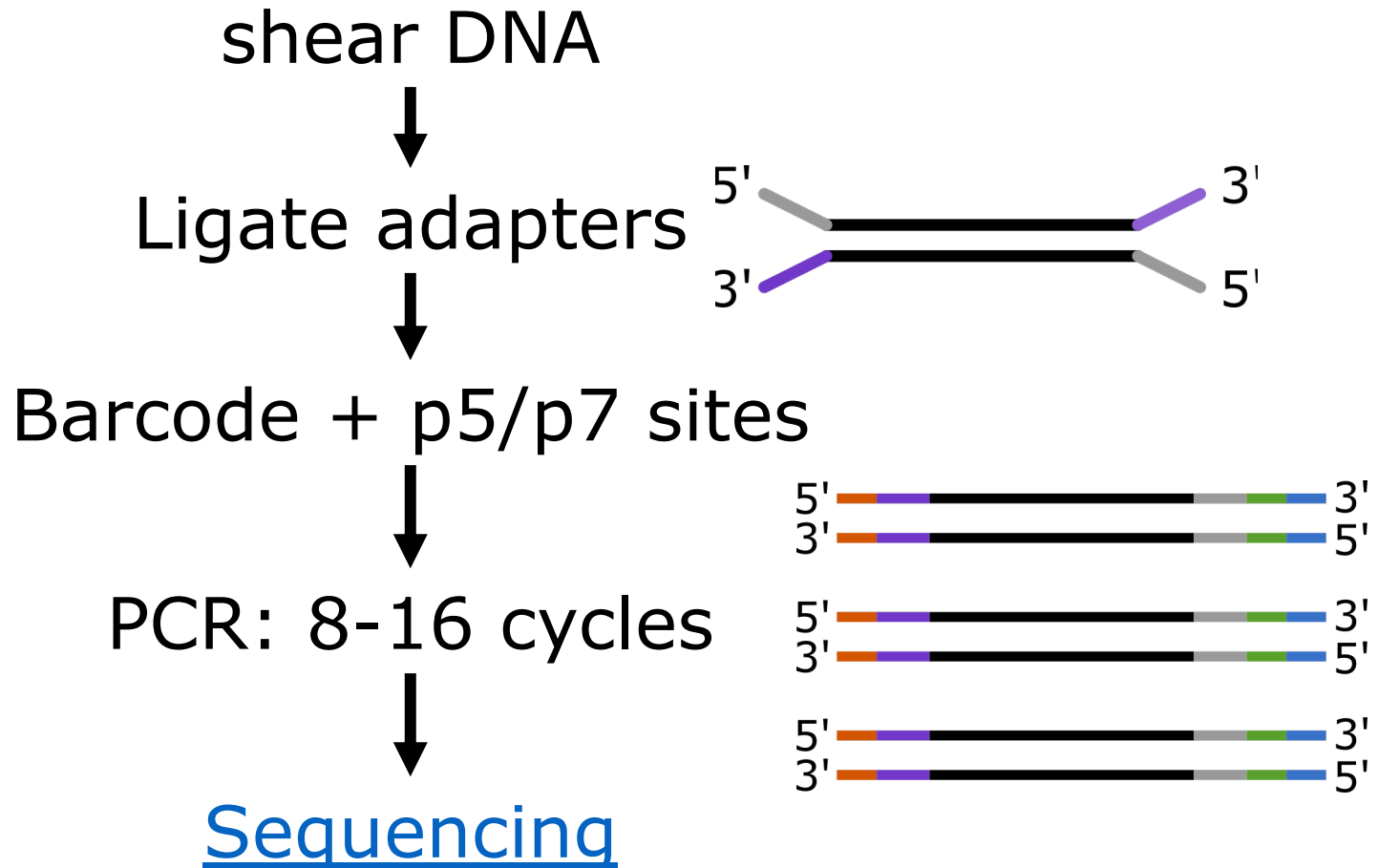
- 2nd generation sequencing
- Massive throughput: up to 500×10^9 bases/run
- Most used platform today

Sequencing-by-synthesis

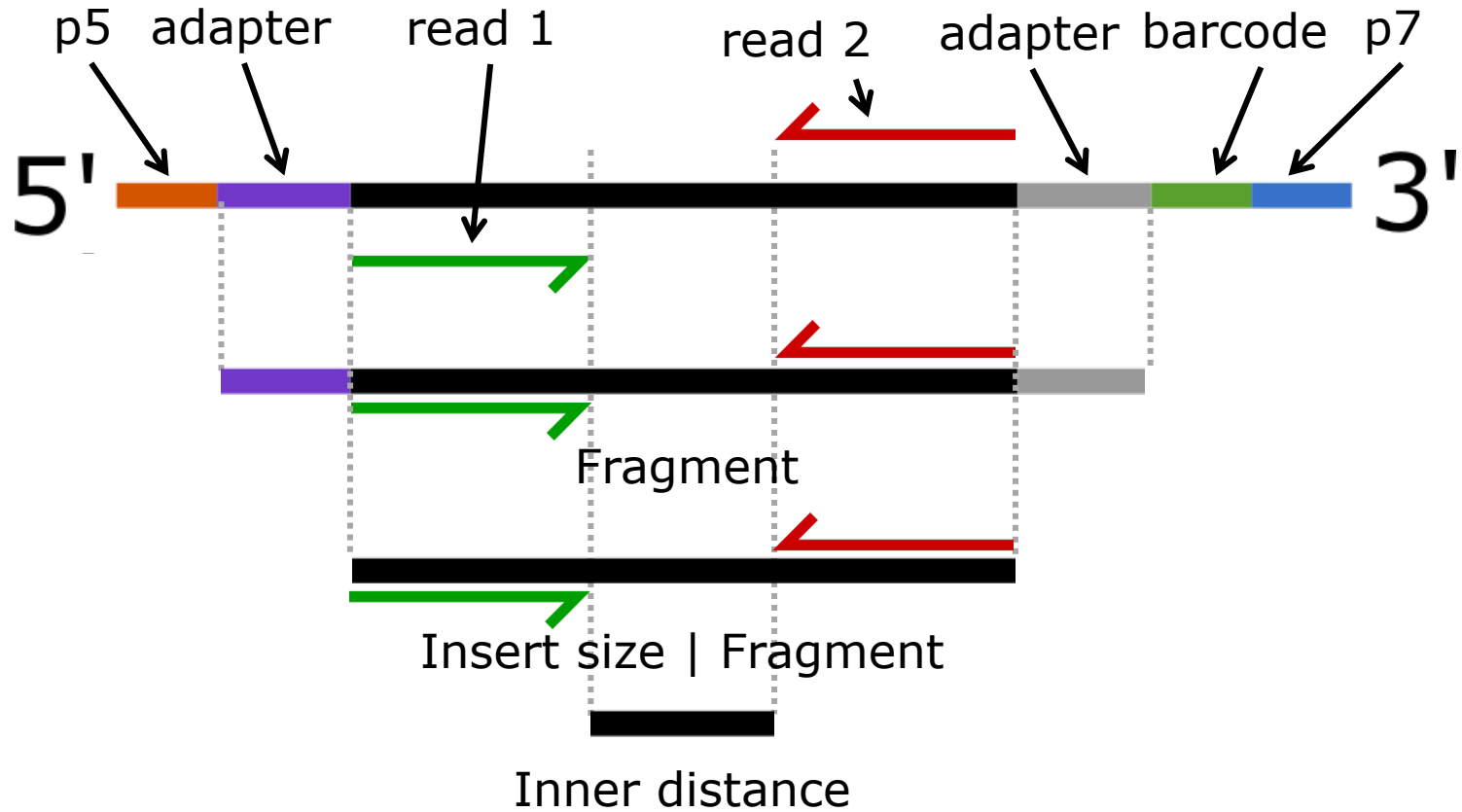
- 50 – 300 bp
- Paired-end (or single-end)



Illumina library prep

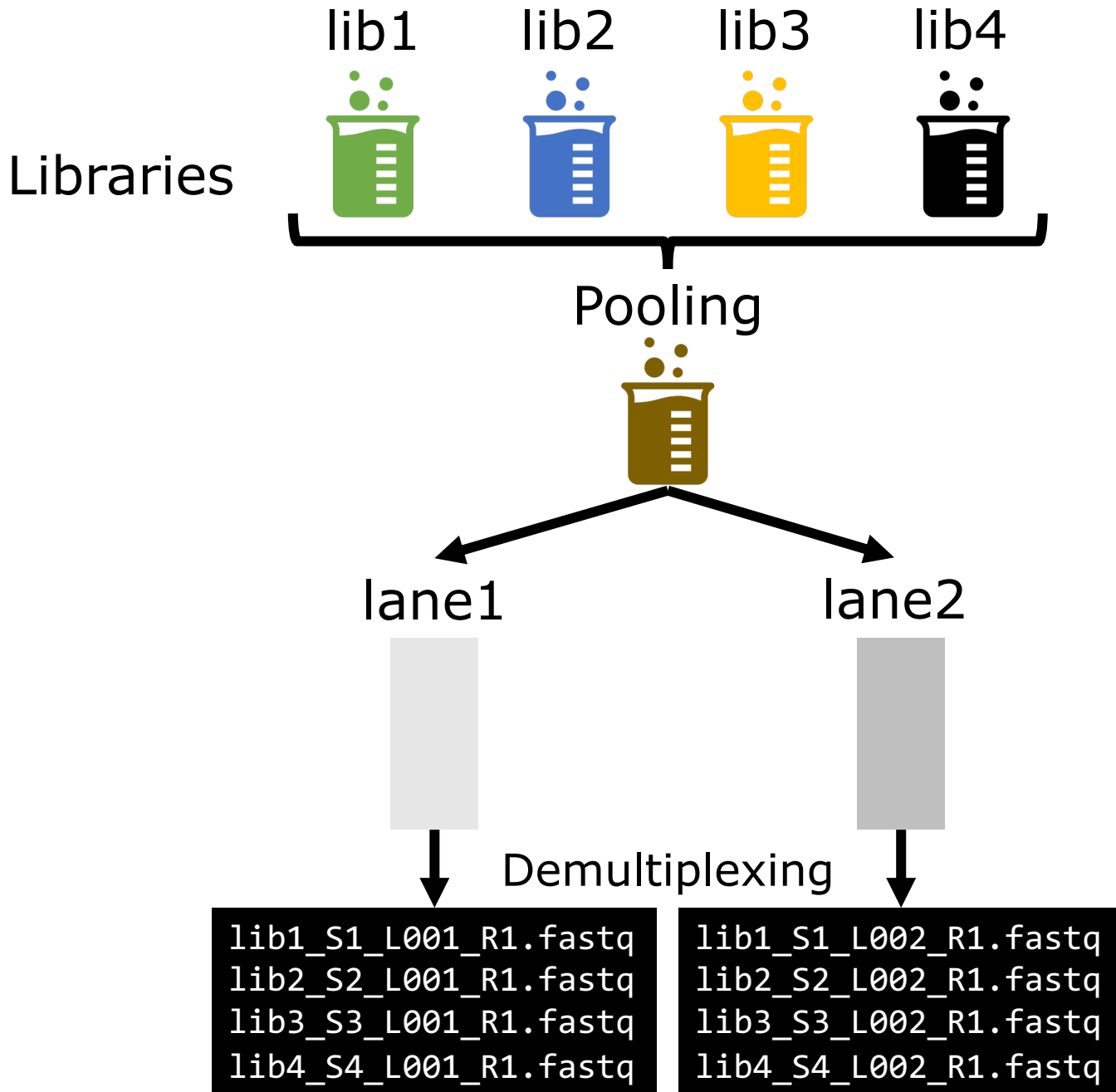


Some definitions



Some more definitions..

- **Library:** fragments from one (c)DNA sample that share a barcode
- **Sequencing run:** complete cycle of generating reads on a machine
- **Flow cell:** physical platform where sequencing reactions take place. Used once in a sequencing run.
- **Lane:** compartment within the flow cell. An Illumina flow cell often has multiple lanes (2 or 4)



fastq

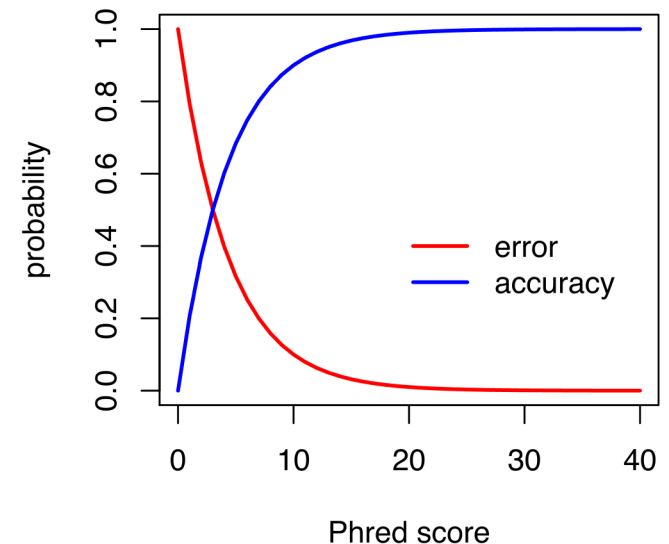
fasta + basequality (fasta + q = fastq)

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

$$-10 \log_{10} (0.01) = 20$$

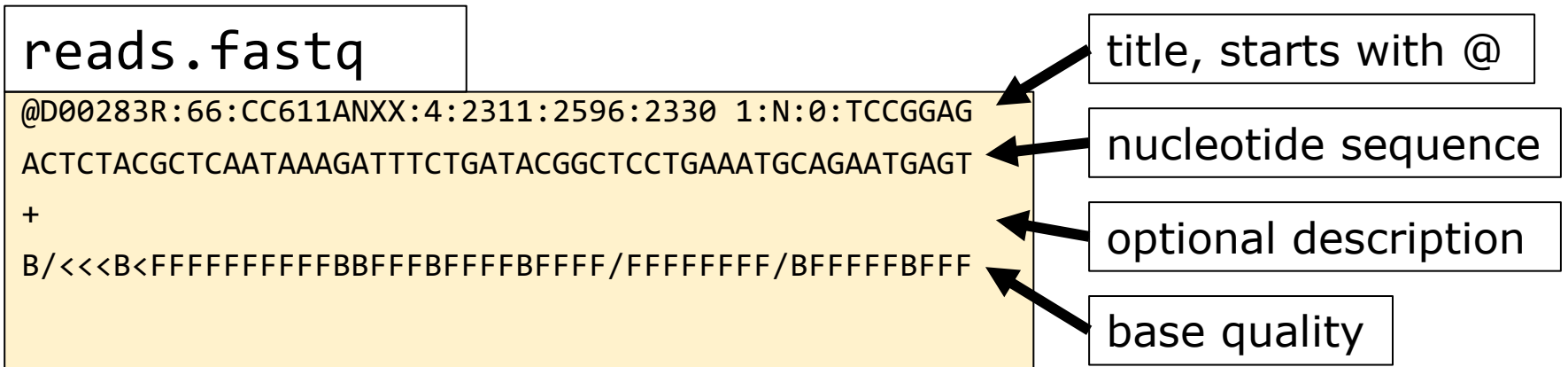
$$-10 \log_{10} (0.1) = 10$$

$$-10 \log_{10} (0.5) = 3$$

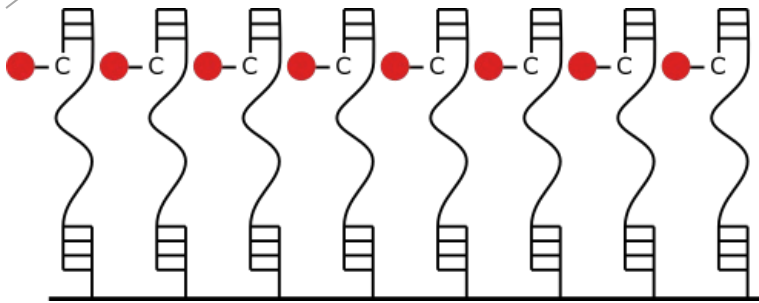
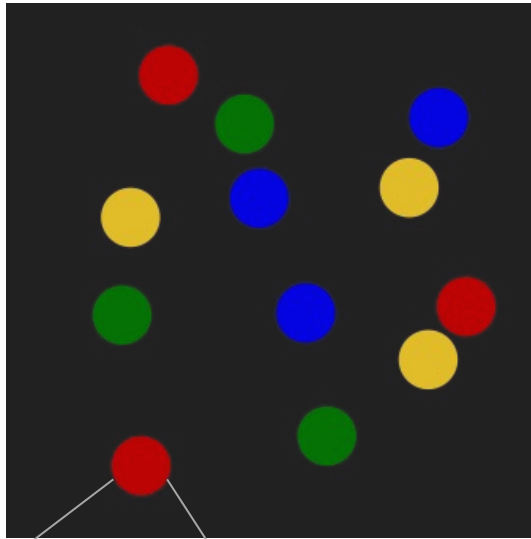


Question

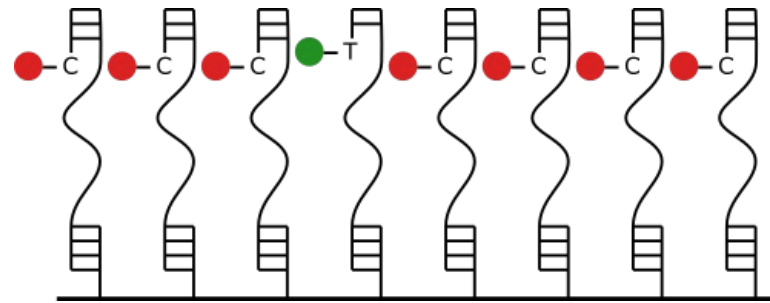
fastq files



!"#\$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
| | | | |
0.2.....26...31.....41



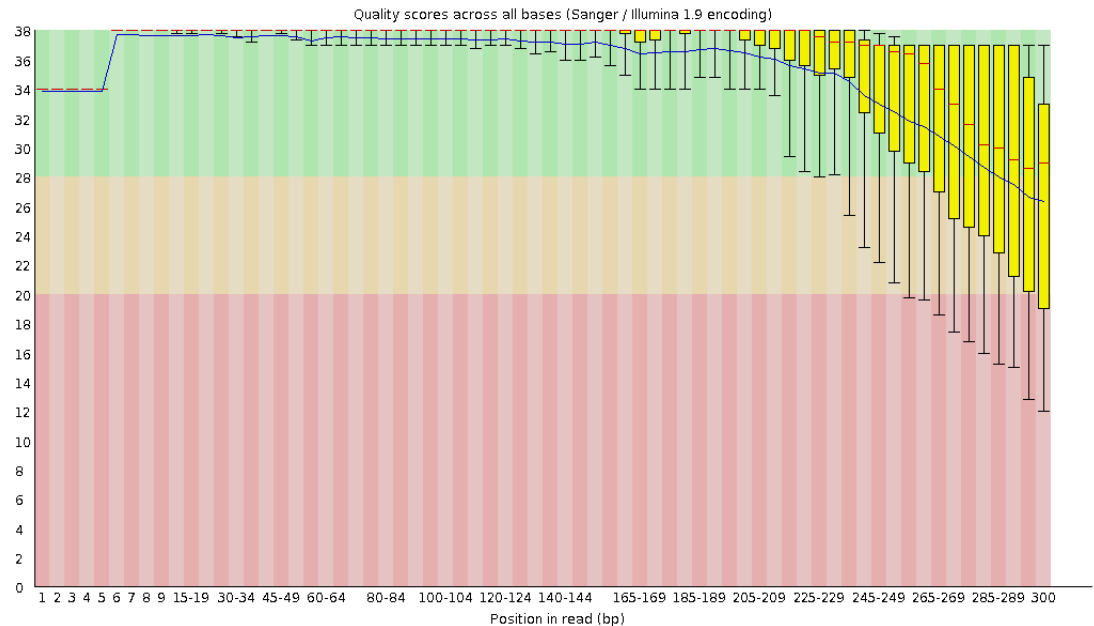
in phase



out of phase

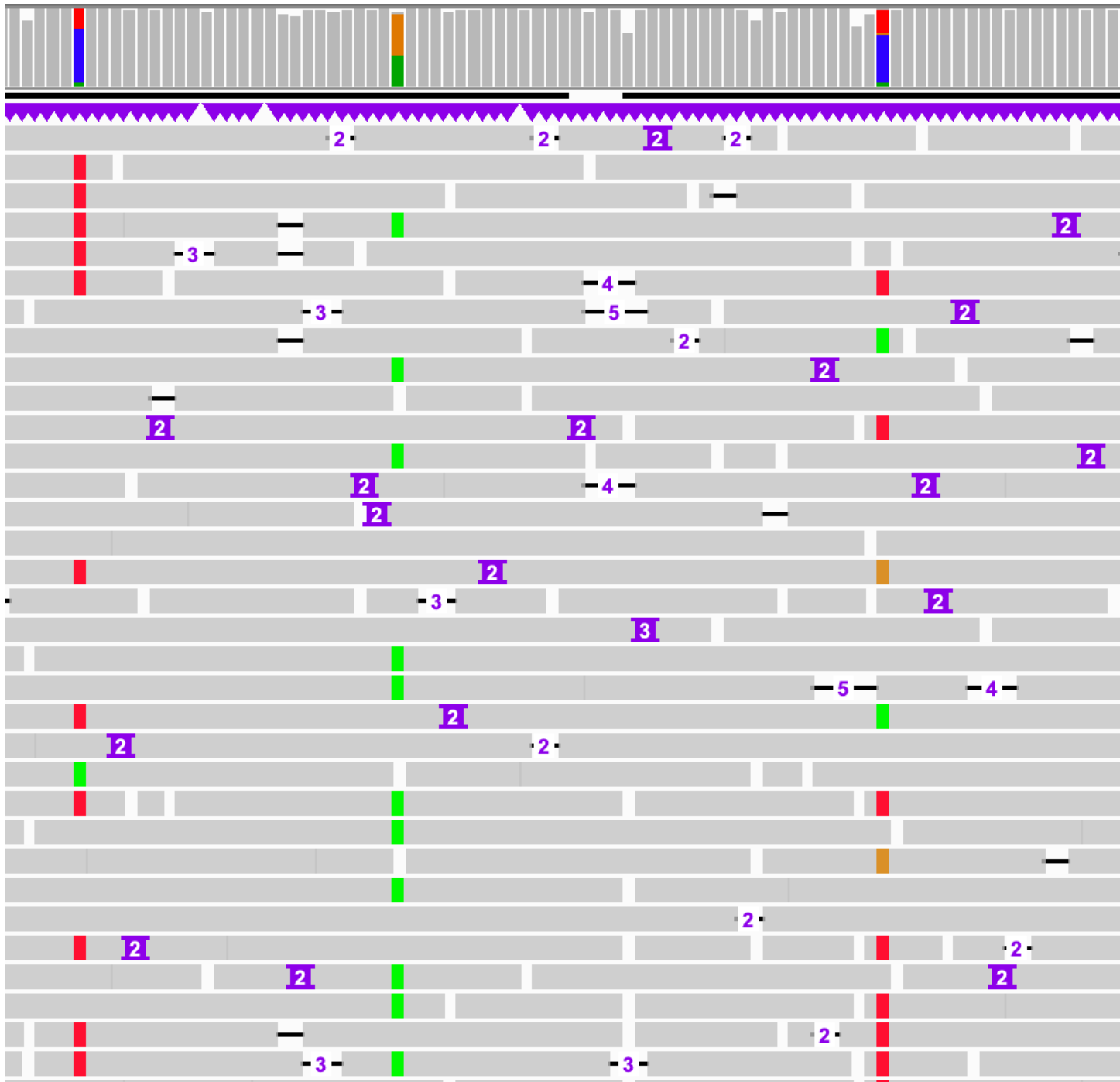
Illumina - limitations

- 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



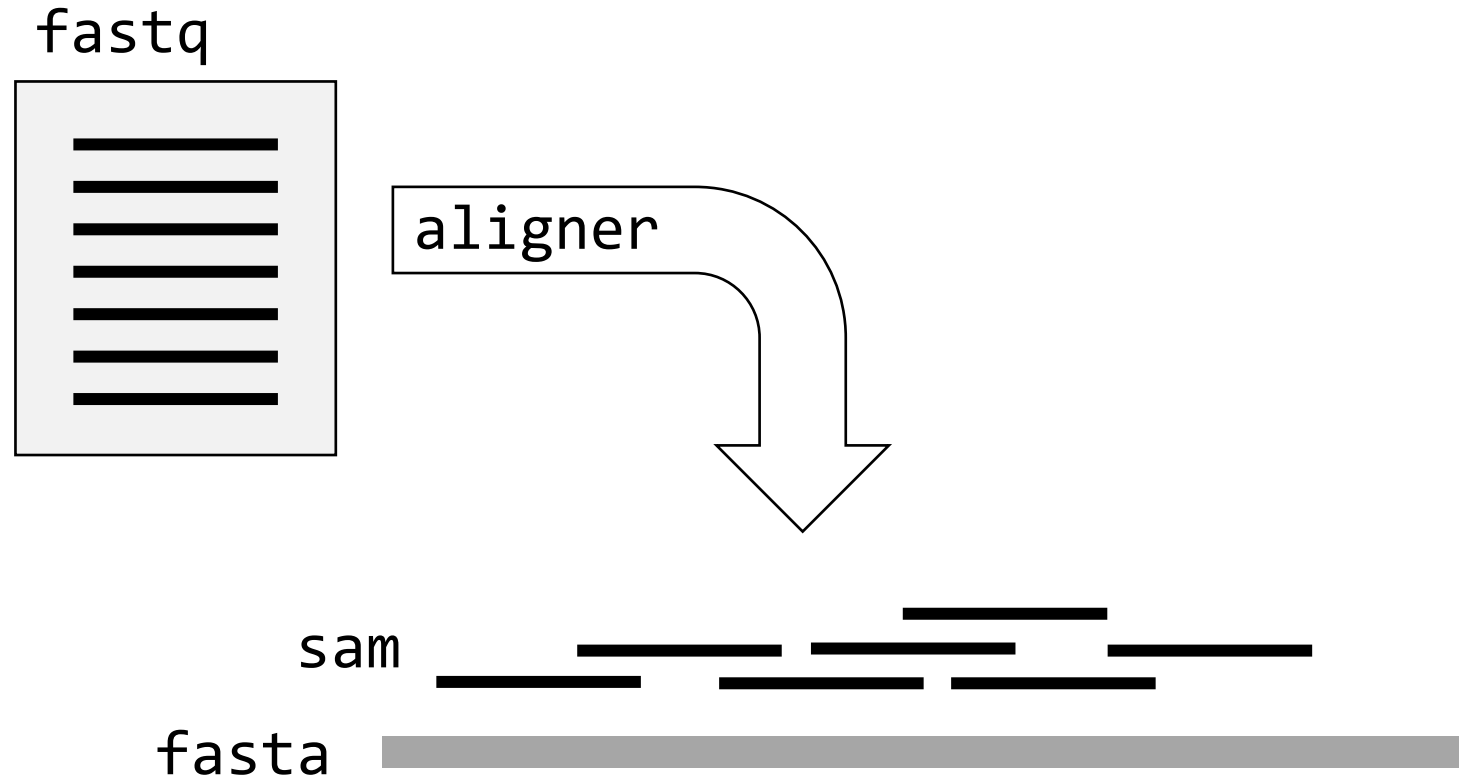
Long reads

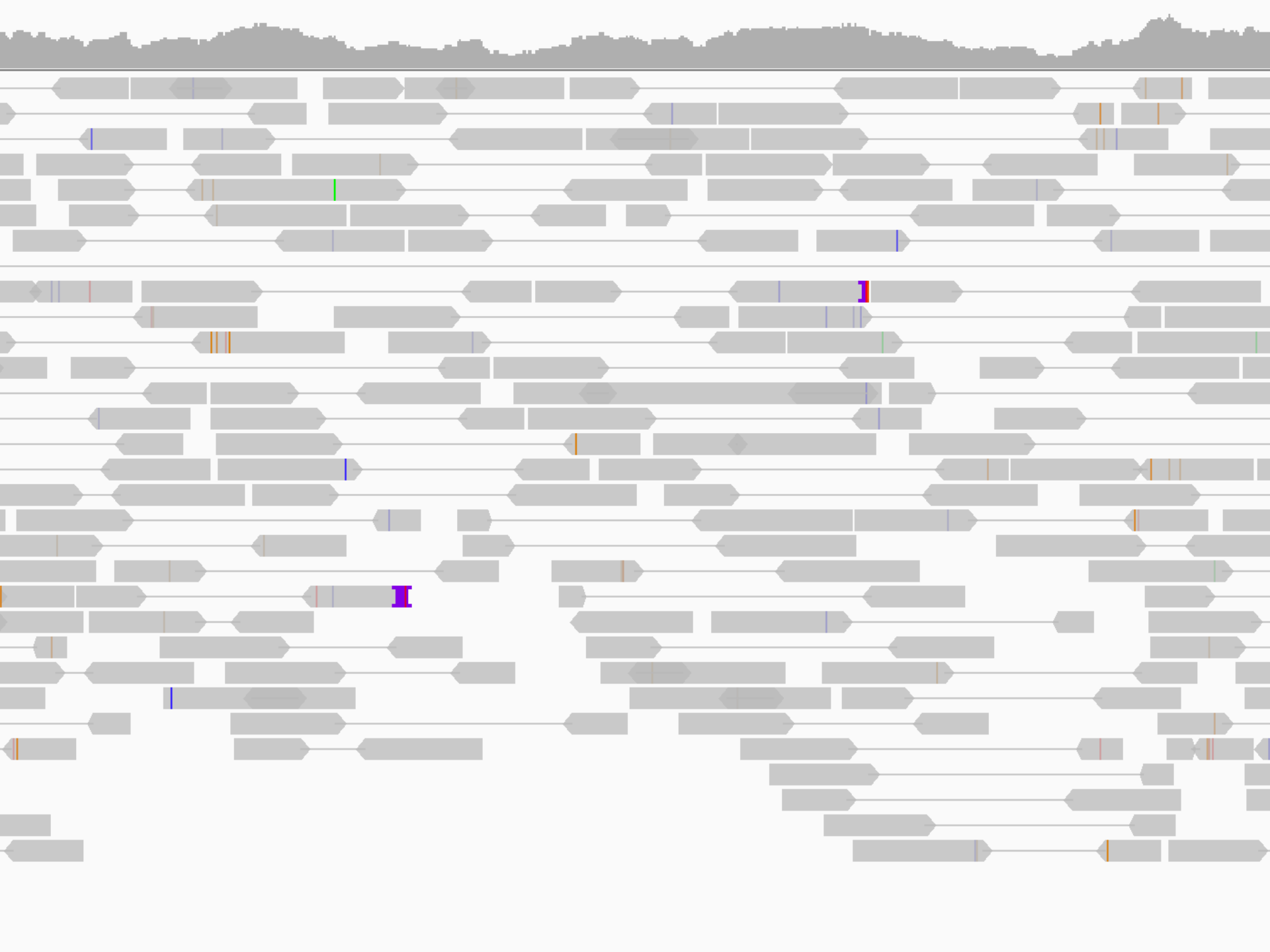
- More error -> difficulties for variant analysis
- But:
 - PacBio HiFi: high baseQ + no bias
 - Long reads can have higher mapping qualities
 - Long reads improve haplotyping

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)

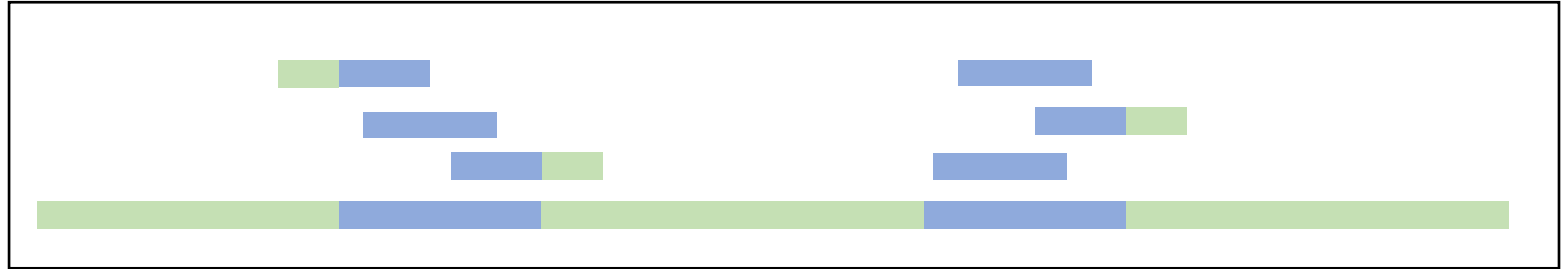




Software

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

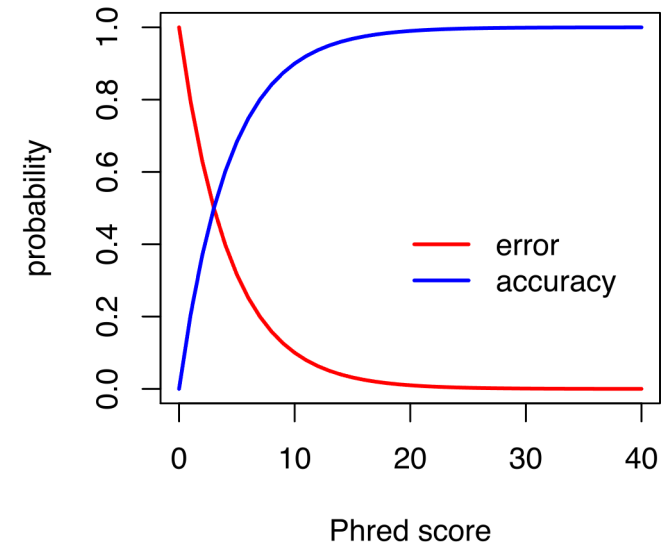
Mapping quality



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

$$-10 \log_{10} (0.01) = 20$$

$$-10 \log_{10} (0.5) = 3$$



Question

sam

sequence alignment format



sam header

```
@HD      VN:1.0  S0: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	RG:Z:rg1

Question

Read groups

- Have multiple groups of reads in a bam file
- Add metadata to alignments:
 - Samples
 - Libraries
 - Lanes
 - ..

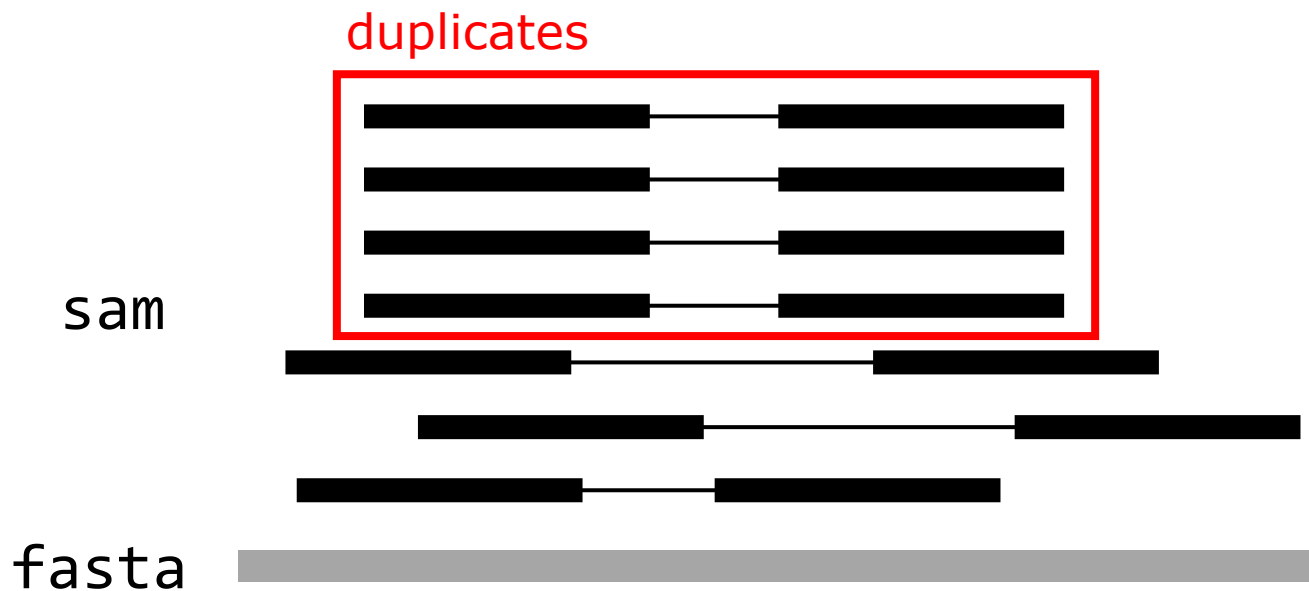
Read groups

@RG ID:rg1 LB:lib1 SM:sampleA

@RG ID:rg2 LB:lib2 SM:sampleA

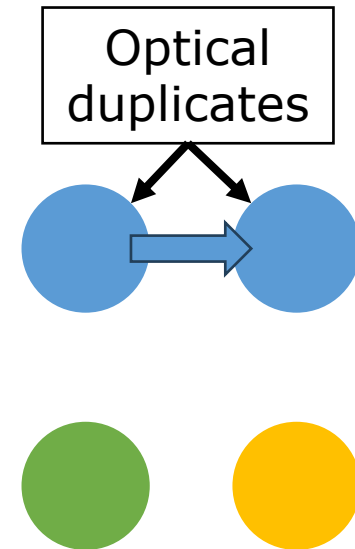
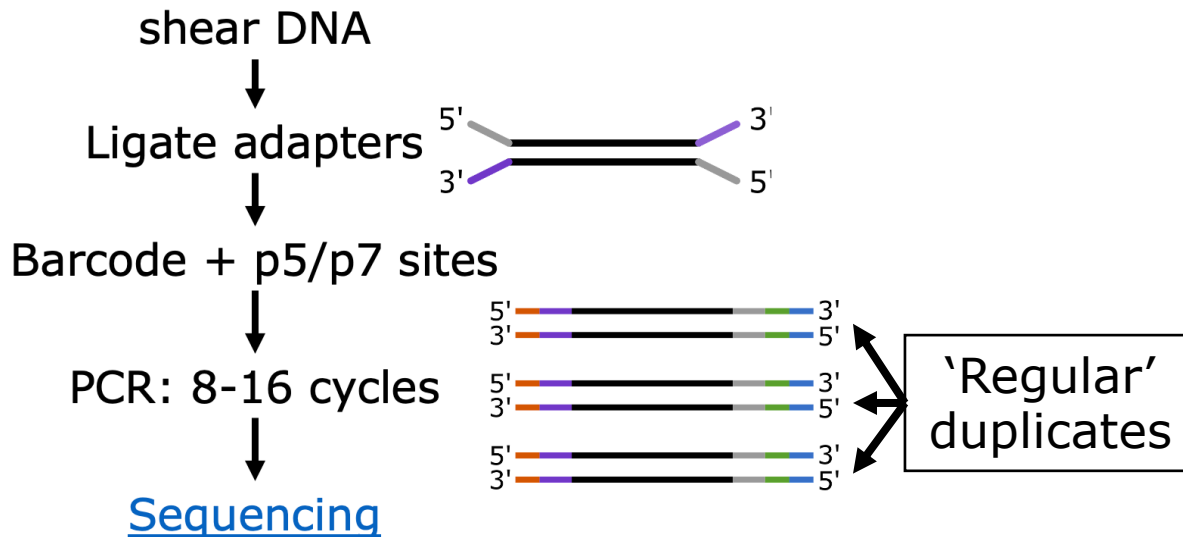
read1	456345	chr20	RG:Z:rg1
read2	456348	chr20	RG:Z:rg2
read3	456357	chr20	RG:Z:rg2
read4	456359	chr20	RG:Z:rg1

Marking duplicates



Marking duplicates

- 'Regular' duplicate: originates from PCR during library prep
- Optical duplicate: originates from bridge amplification



Marking duplicates

- Variant calling: each read is an independent observation of the genome
- Duplicates (can) have the same molecular origin -> not independent
- In a high-quality library, 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

