## Introduction to Bioinformatics

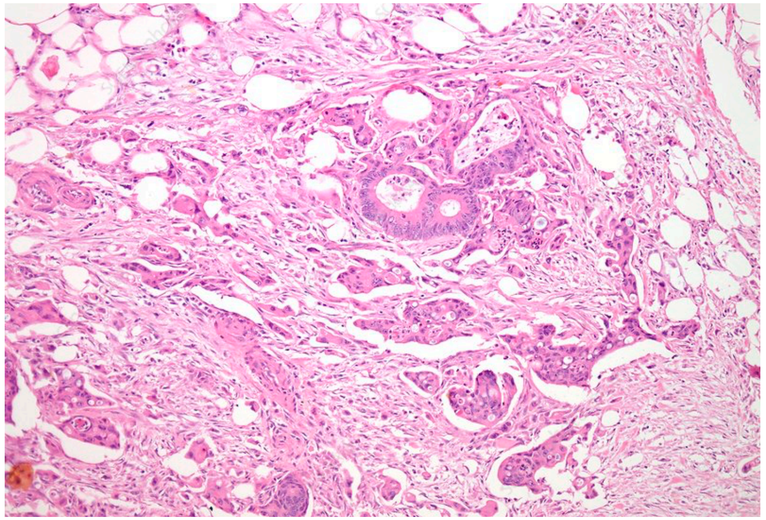
# Sample preparation & sequencing

Matej Hrnčiar

24. 02. 2026

# First, we need to collect a sample

- Tissue
- Blood
- Saliva
- Water
- Hair follicle
- Plant cutting
- …



# We have a sample, now what?

- ~~Sequence it!~~
- Prepare the sample in wet lab

**Aren't we missing something?**

- Storage
- Transport

# Short-term storage

**Fresh**

- 20 to 4 °C
- immediate processing
- plants, fungi, water

**Fresh frozen**

- -20 to -80 °C
- slows down degradation (but does not halt it completely)
- blood, saliva, tissue

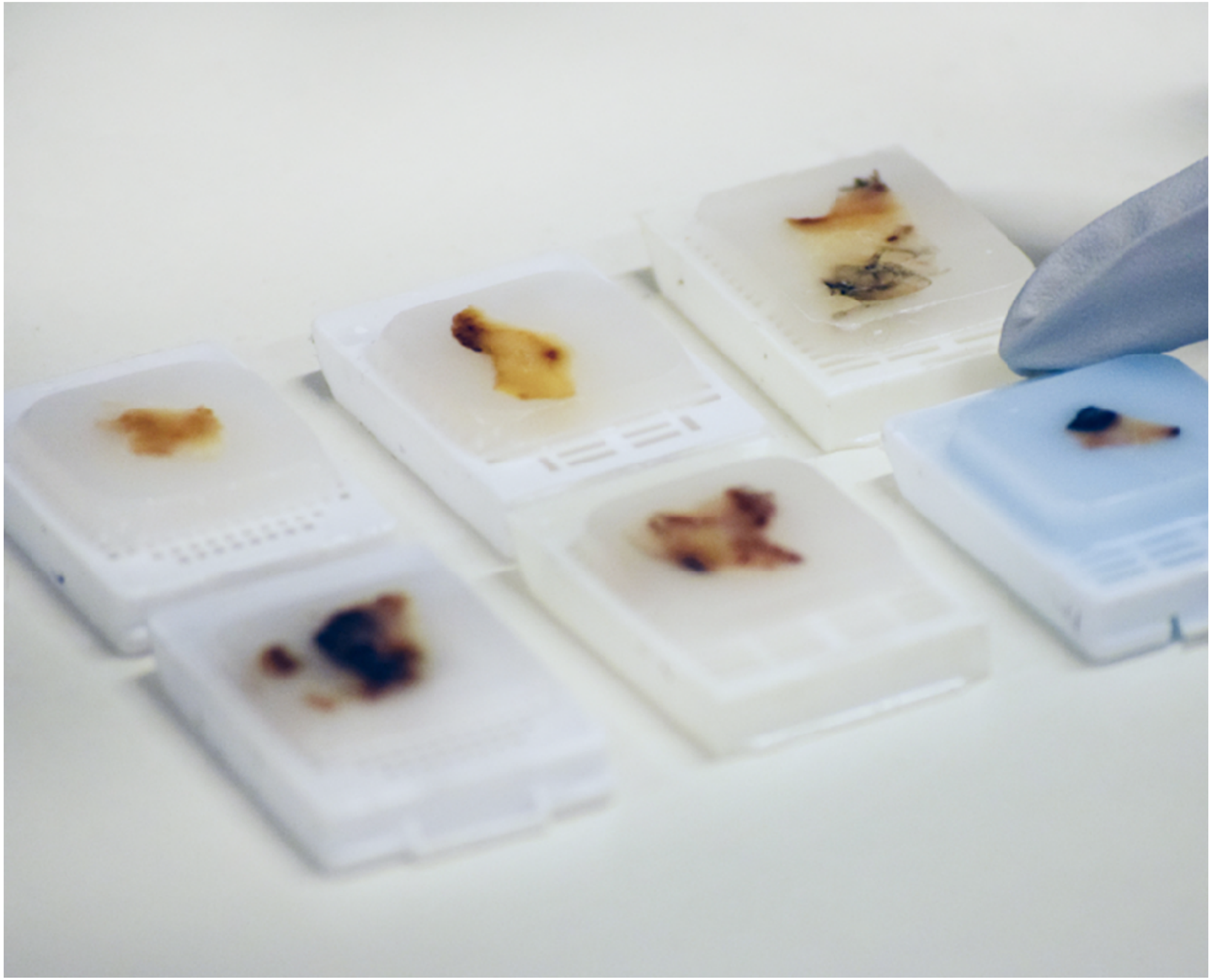
# Long-term storage

**Flash frozen**

- some samples need to be frozen in liquid nitrogen to -196 °C
- complicated transport - usually done by surrounding the samples with dry ice (solid form of $CO_2$)

**FFPE: Formalin-Fixed, Paraffin-Embedded**

- long-term storage
- can be kept at room temperatures - better for transport
- causes DNA fragmentation



# Terminology, part 1

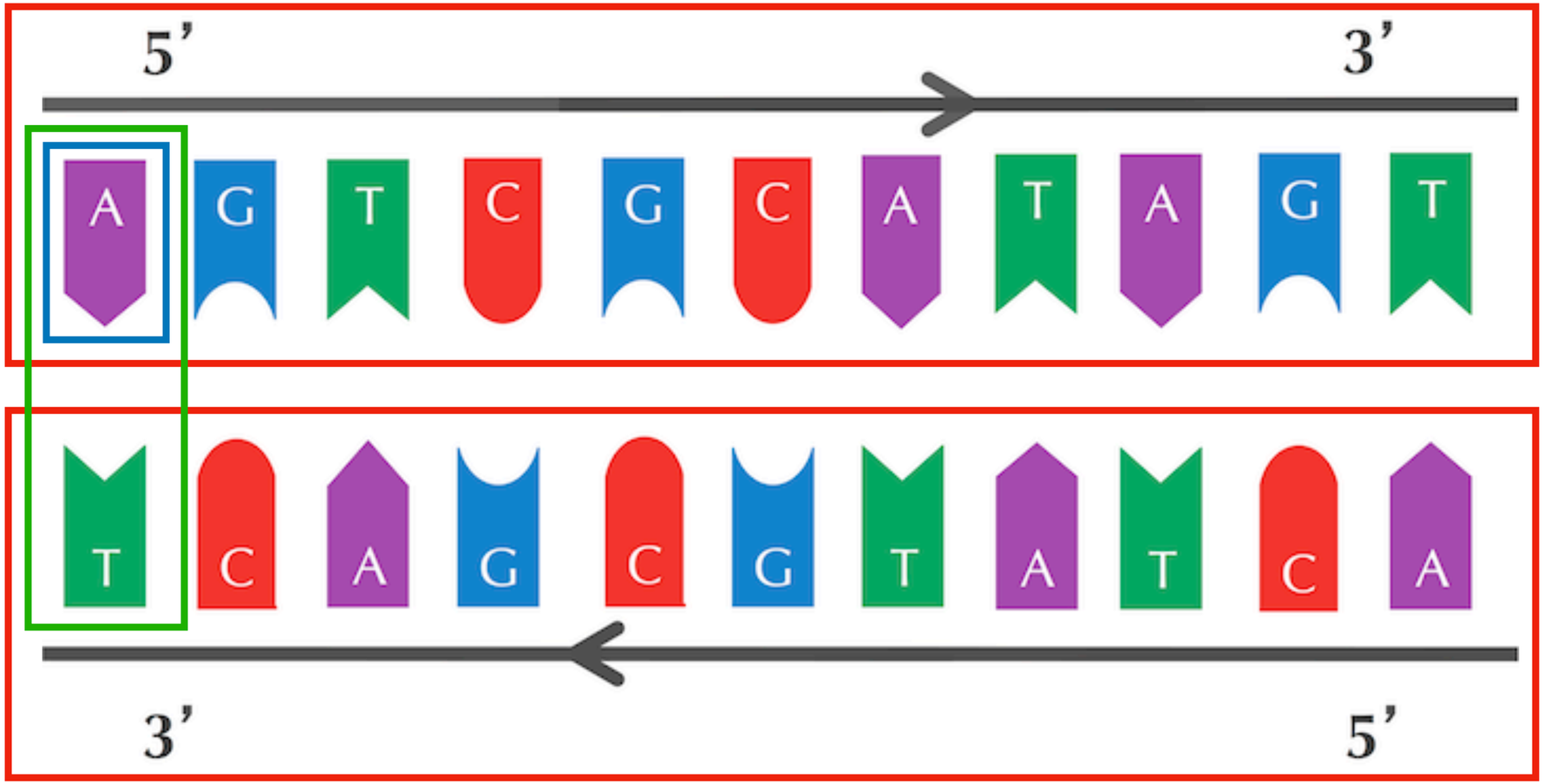
- **Fragment** (**insert**): physical piece of DNA / RNA, **insert** usually represents only the genomic sequence without adapters, indexes and other identifiers
- **DNA complementarity**: **A** pairs only to **T**, **C** pairs only to **G**
- **Strand**: DNA consists of 2 strands: original and reverse complement
- **Reverse complement**: strand of opposite polarity to the original strand following the rule of DNA complementarity
- **Base pair** (**bp**): one base and its complement on the opposite strand

# Base pairs and strands

Original strand: AGTCGCATAGT

Base

Base pair

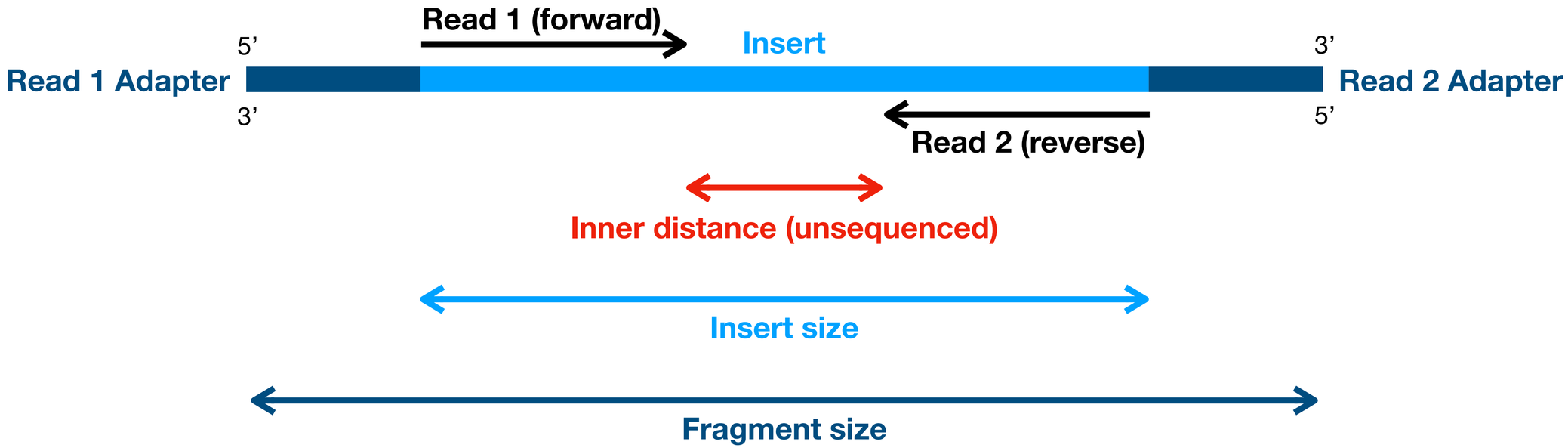



Reverse complement strand: ACTATGCGACT

# Terminology, part 2

- **Shotgun sequencing**: randomly breaking up the genome into small DNA fragments that are sequenced individually — genome is reconstructed later in the computer
- **Read**: digitised sequence generated from a DNA / RNA fragment
- **Library**: collection of reads, usually representing a single sample
- **Single-read** (**single-end**) **sequencing**: DNA fragment is read from only one end
- **Paired-end sequencing**: DNA fragment is read from both ends
  - usually, it means that the 2$^{nd}$ strand is reverse complement of the first one
  - however, the reads **don't have to overlap**

# Fragment, insert & read




**Insert size**: length of fragment from the beginning of the the forward read to the end of the reverse read, **including the unsequenced middle part**

# Metrics

- **Read length**: b (base), bp (base pair), kb (1,000 bases), Mb (million bases), Gb (billion bases), Tb (trillion bases)

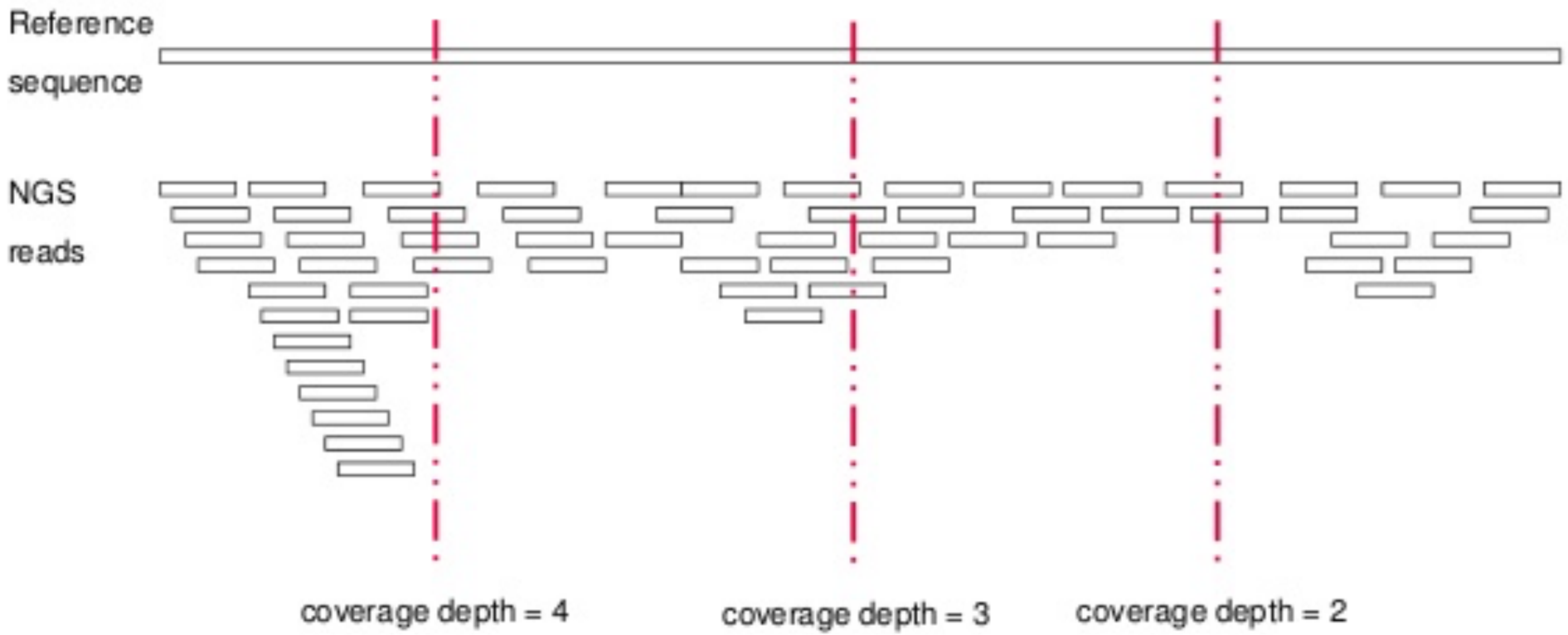
- **Coverage**: how many reads on average overlap the same base

$$C = \frac{\text{Number of Reads } (N) \cdot \text{Read Length } (L)}{\text{Genome size } (G)}$$

- **Q Score**: base call accuracy

$$Q = -10 \cdot log_{10}(Pe)$$

# Reads and coverage



# Q Score

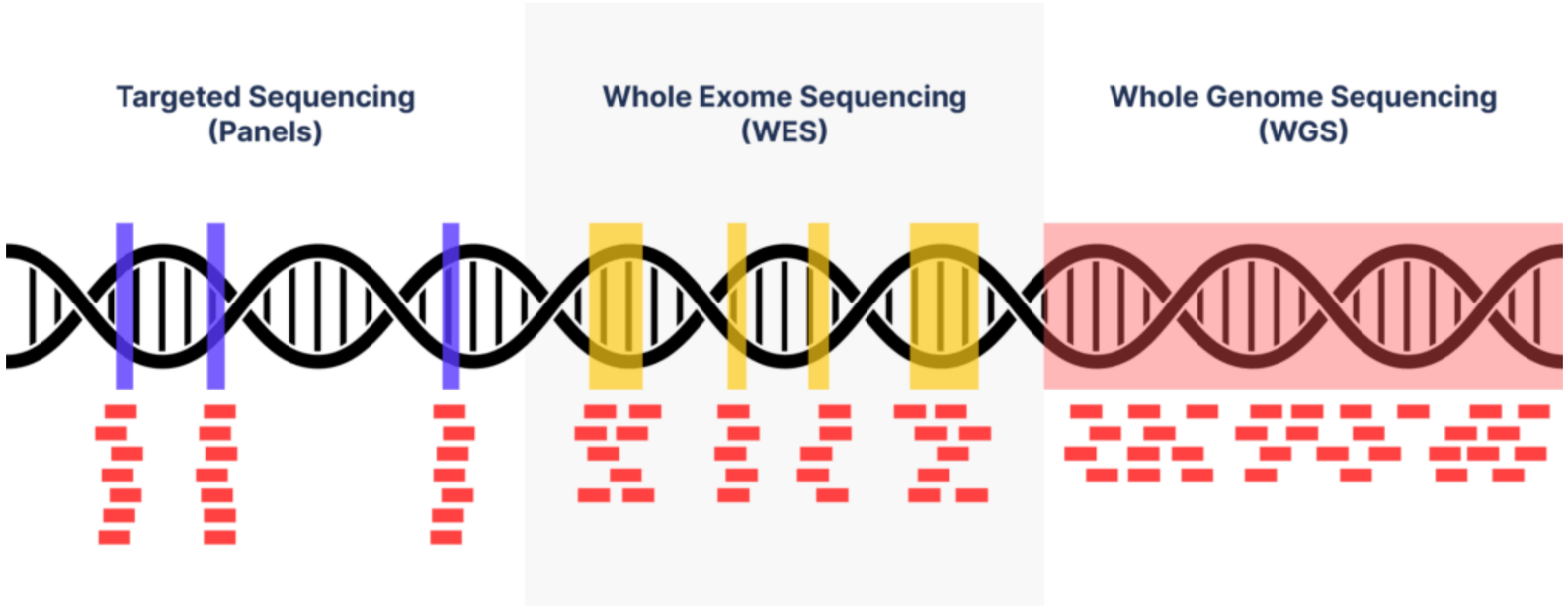
| Q Score | Error rate | Inferred base call accuracy |
|---|---|---|
| **Q10** | 1 in 10 bp | 90 % |
| **Q20** | 1 in 100 bp | 99 % |
| **Q30** | 1 in 1 000 bp | 99.9% |
| **Q40** | 1 in 10 000 bp | 99.99% |
| **Q50** | 1 in 100 000 bp | 99.999% |

# Library preparation

*What happens in the wet lab*

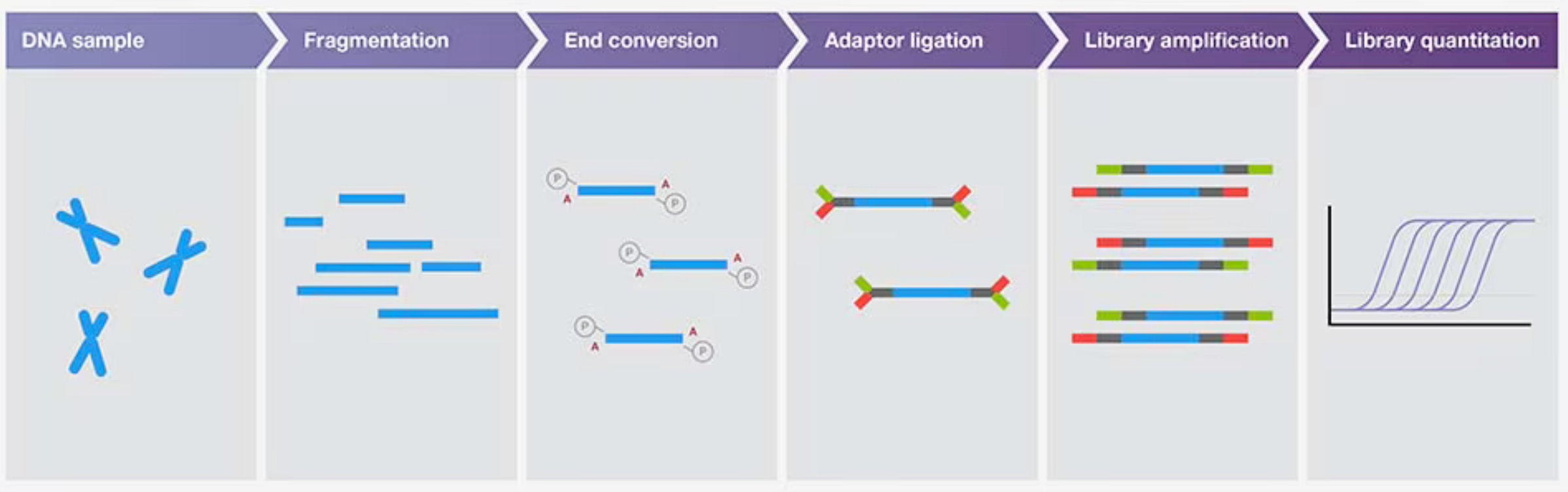
# We first need to answer question: “What do we want to sequence?”

- DNA? RNA? Poly-A? 16S? Methylation arrays? Single cell? CNV arrays?
- Humans? Mice? Dogs? Invertebrate? Plants? Bacteria? Viruses?
- Whole genomes? Exomes? Panels?



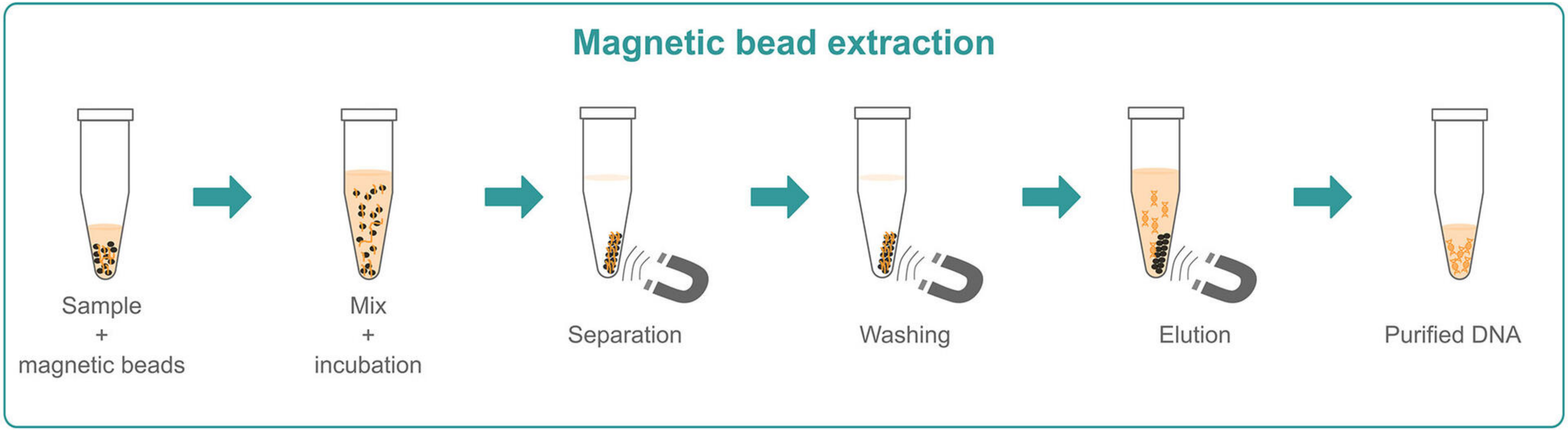

https://3billion.io/blog/which-is-the-best-ngs-approach-for-rare-disease-diagnosis-panels-wes-or-wgs

# 6 main steps of library preparation



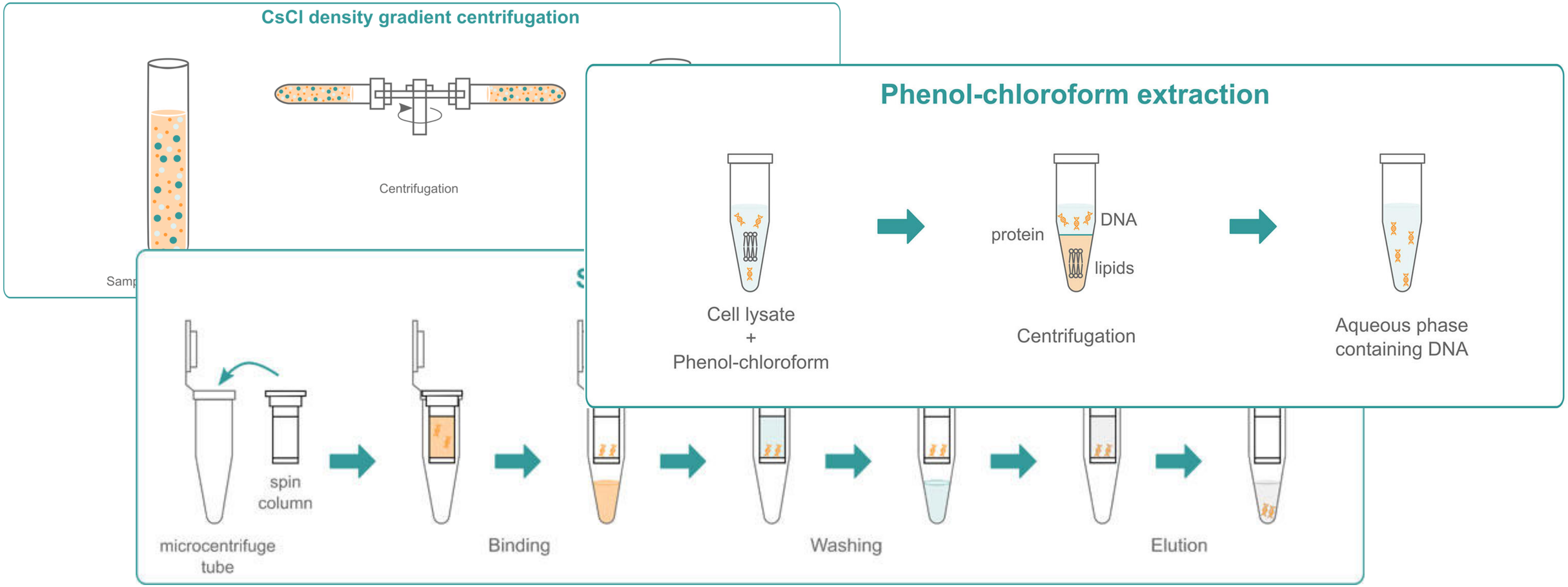

https://www.thermofisher.com/sk/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/next-generation-sequencing/dna-sequencing-preparation-illumina.html

# Extraction



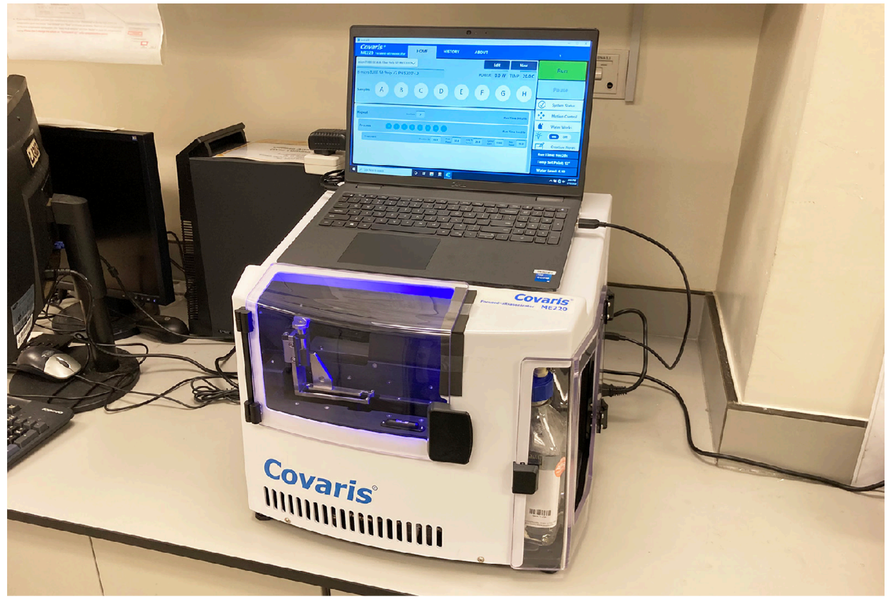

https://www.integra-biosciences.com/global/en/blog/article/comprehensive-introduction-dna-extraction

# Extraction



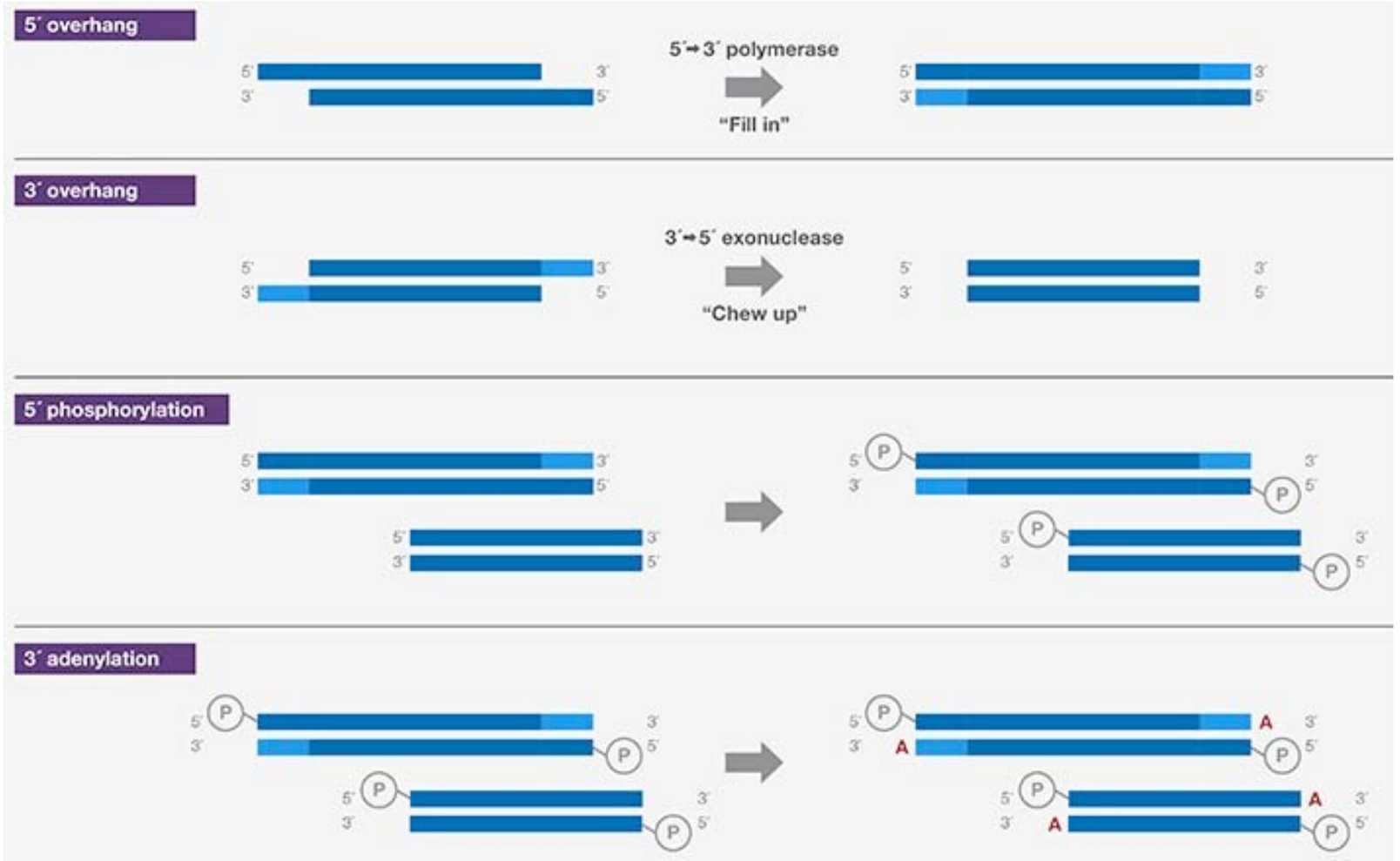

https://www.integra-biosciences.com/global/en/blog/article/comprehensive-introduction-dna-extraction

# Fragmentation




- Fragmentation of DNA / RNA into desirable range
- 100 - 300 bp, 300 - 600 bp
- Three approaches:
    1. **Mechanical shearing**: more consistent fragment sizes
    2. **Enzymatic digestion**: requires lower DNA input, less precise
    3. **Transposon-based fragmentation**: allows skipping of end conversion and adapter ligation

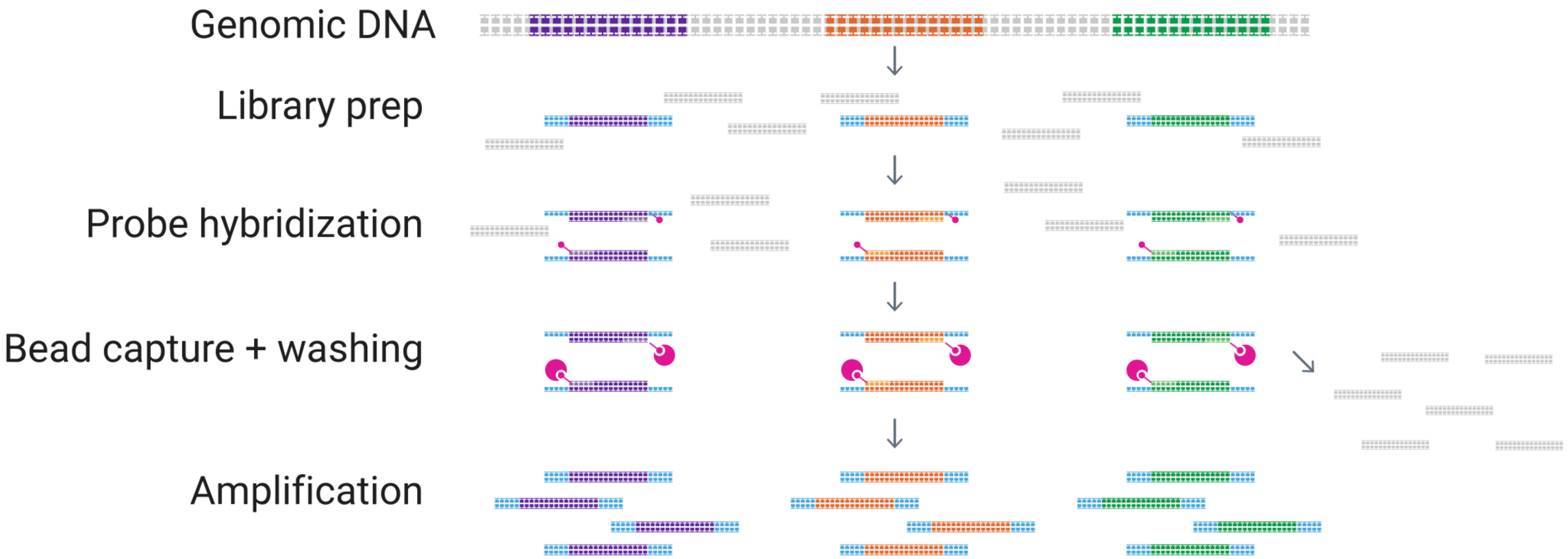
# End repair (end conversion)



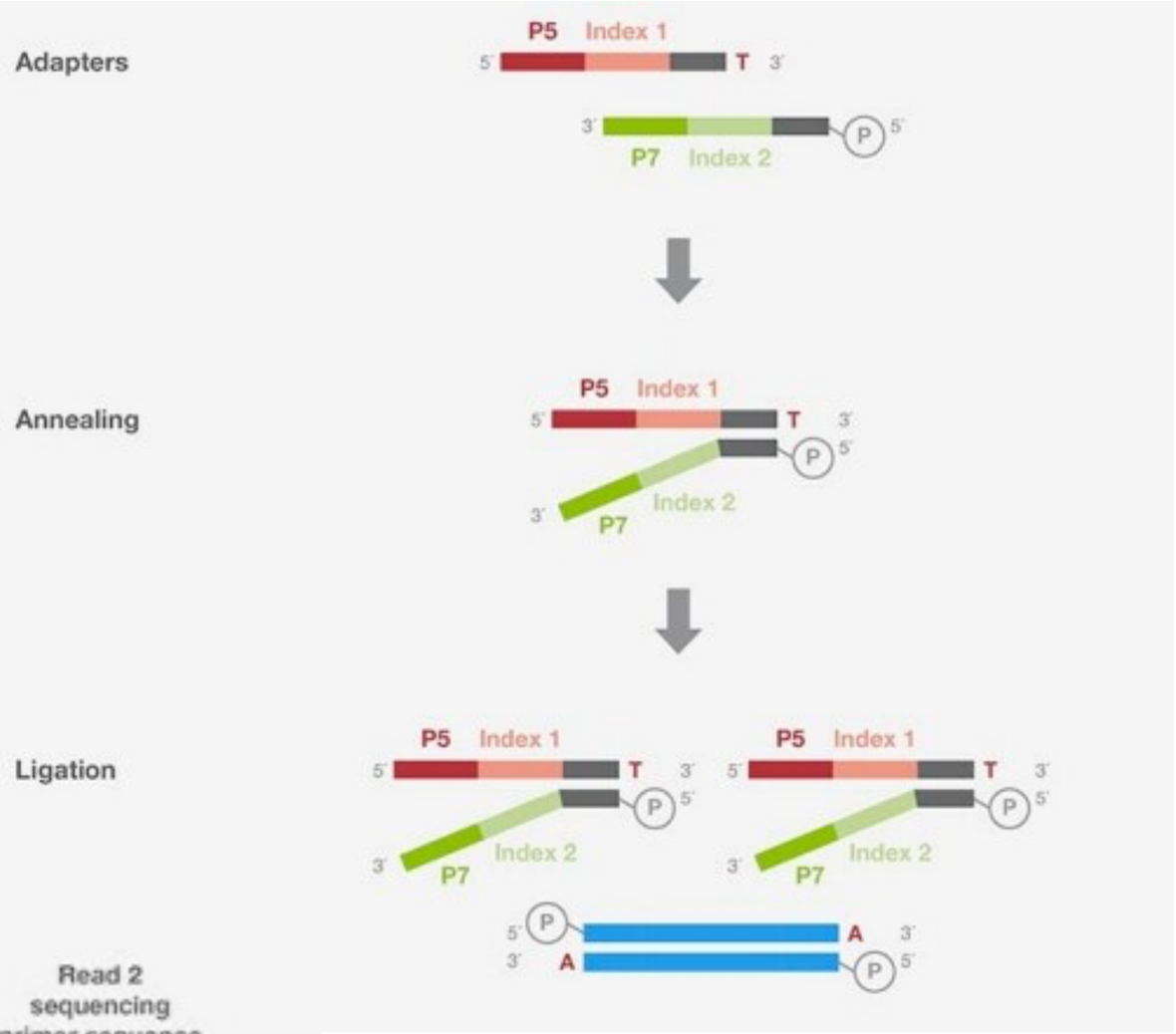
# Extra step: Target enrichment & Probe hybridisation

- If we are performing WES or TXS (panels), we need to select only the fragments from regions of interest (ROIs)
- **Target enrichment**: The method of selectively capturing and retaining DNA fragments from specific genomic ROIs while removing non-target sequences
- **Probes**: Short, synthetic, single-stranded fragments designed to match ROIs
- **Probe hybridisation**: The process where a single-stranded probe binds specifically to its complementary DNA or RNA sequence through base pairing
- **Tiling**: For longer sequences (exons), multiple overlapping probes ensure “full” coverage

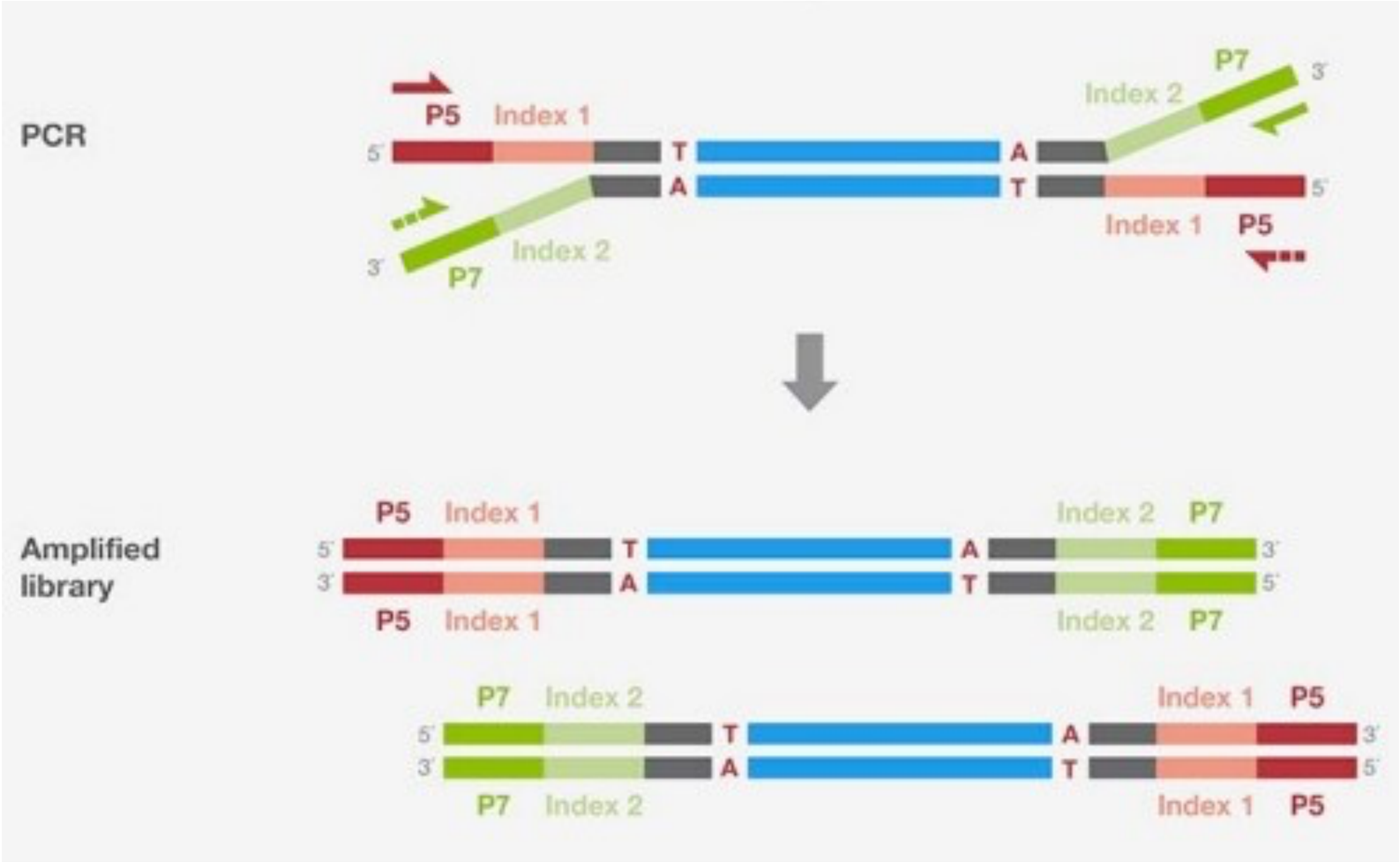
# HYBRIDIZATION CAPTURE



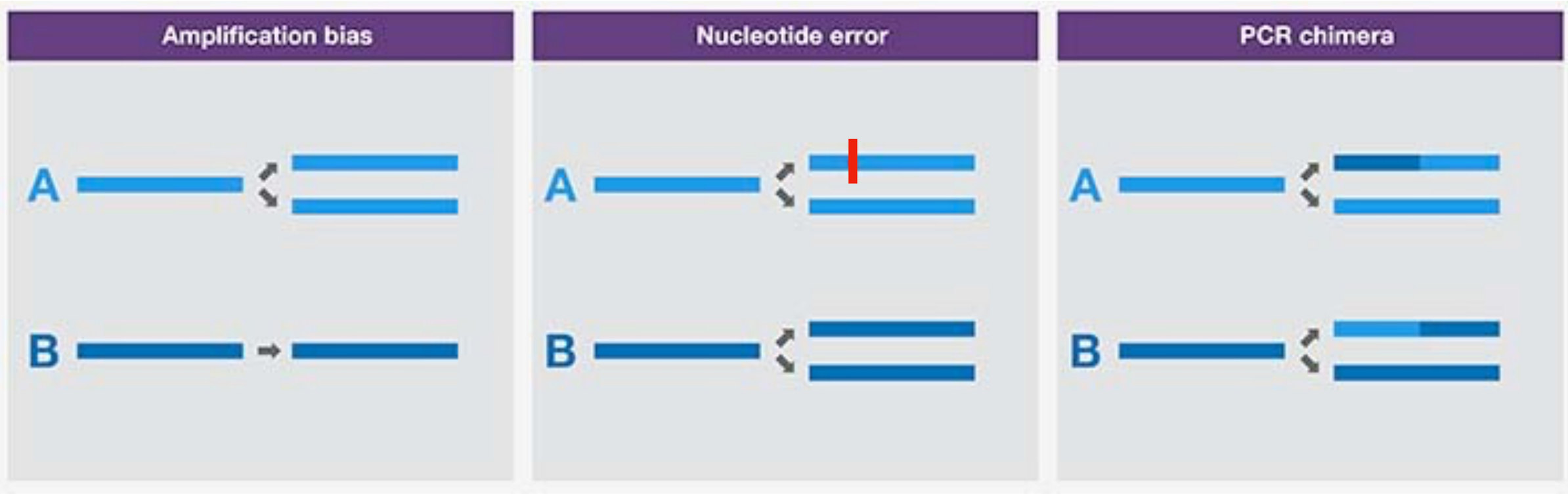
# Adapter ligation



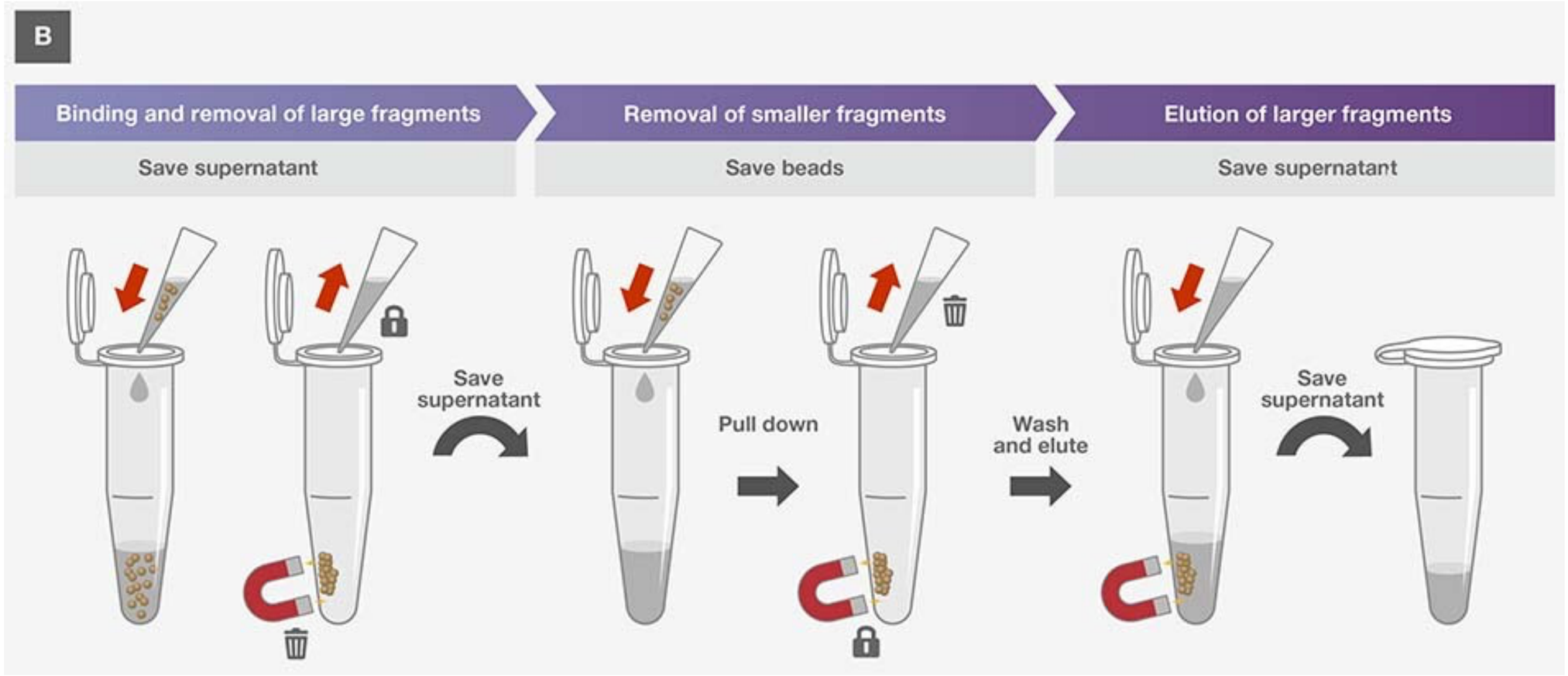

# Library amplification



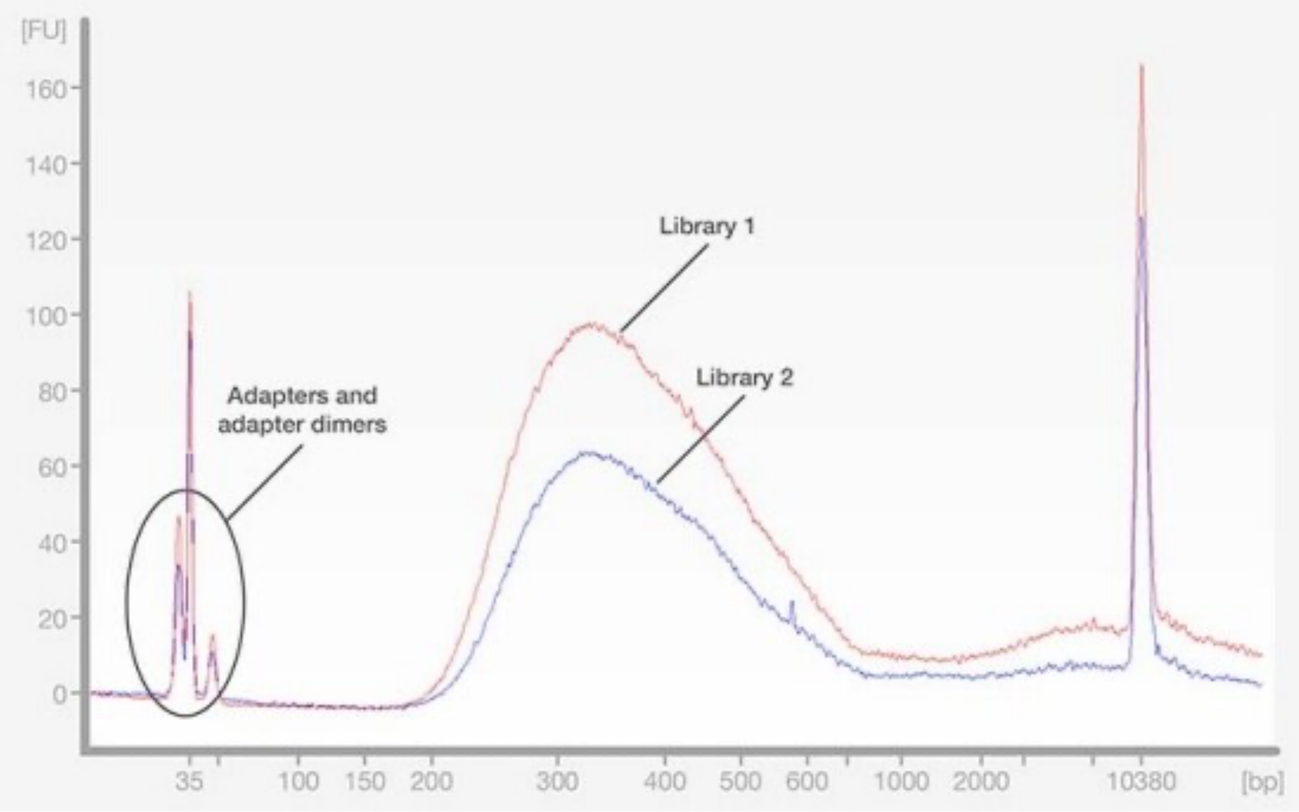
# PCR bias



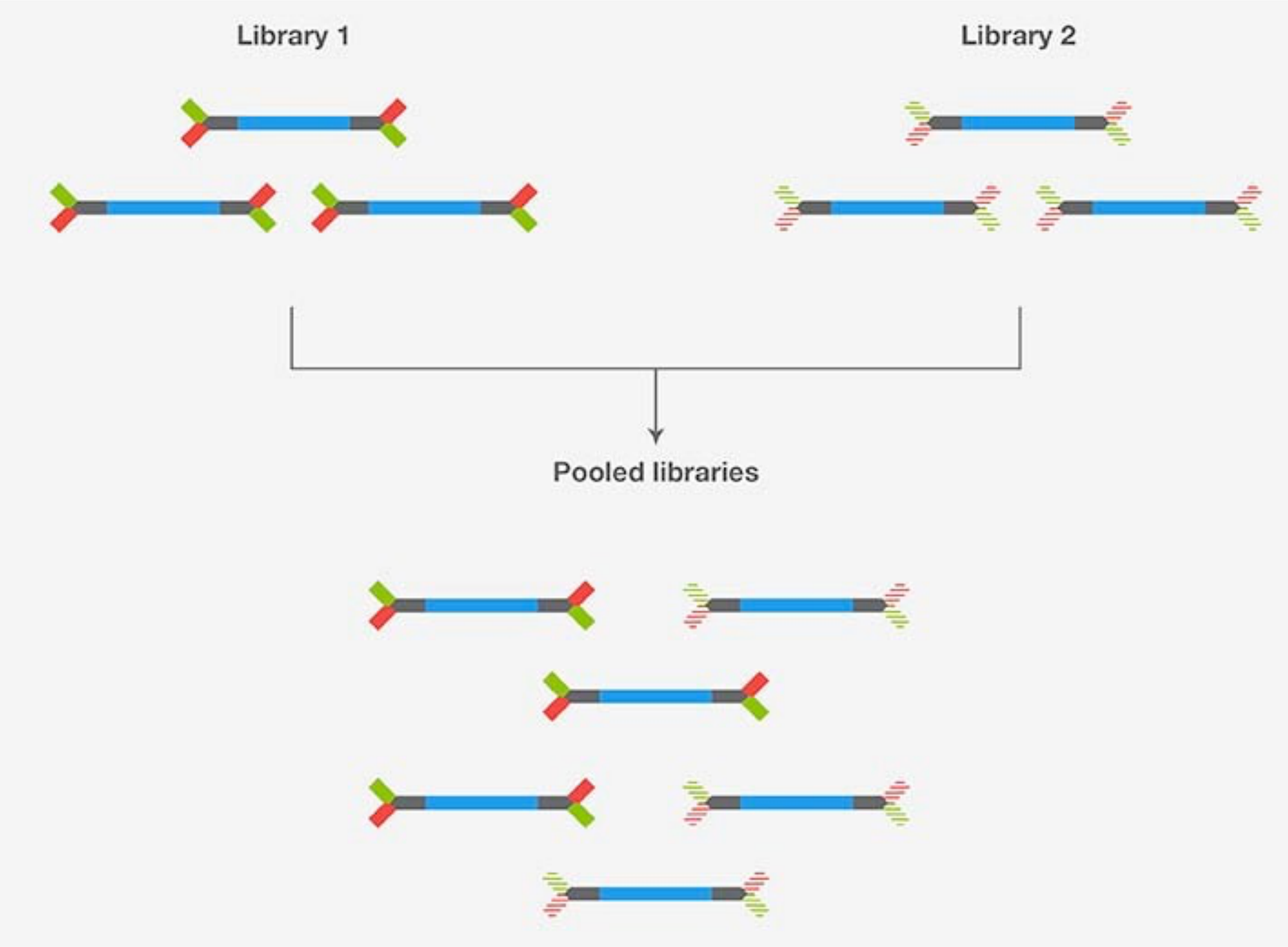
# Size selection



# Quality control




# Pooling



# Sequencing

*3 revolutions in genetics*

# Sequencing

*DNA sequencing is a laboratory technique for determining the exact sequence of nucleotides (bases) in a DNA molecule.*

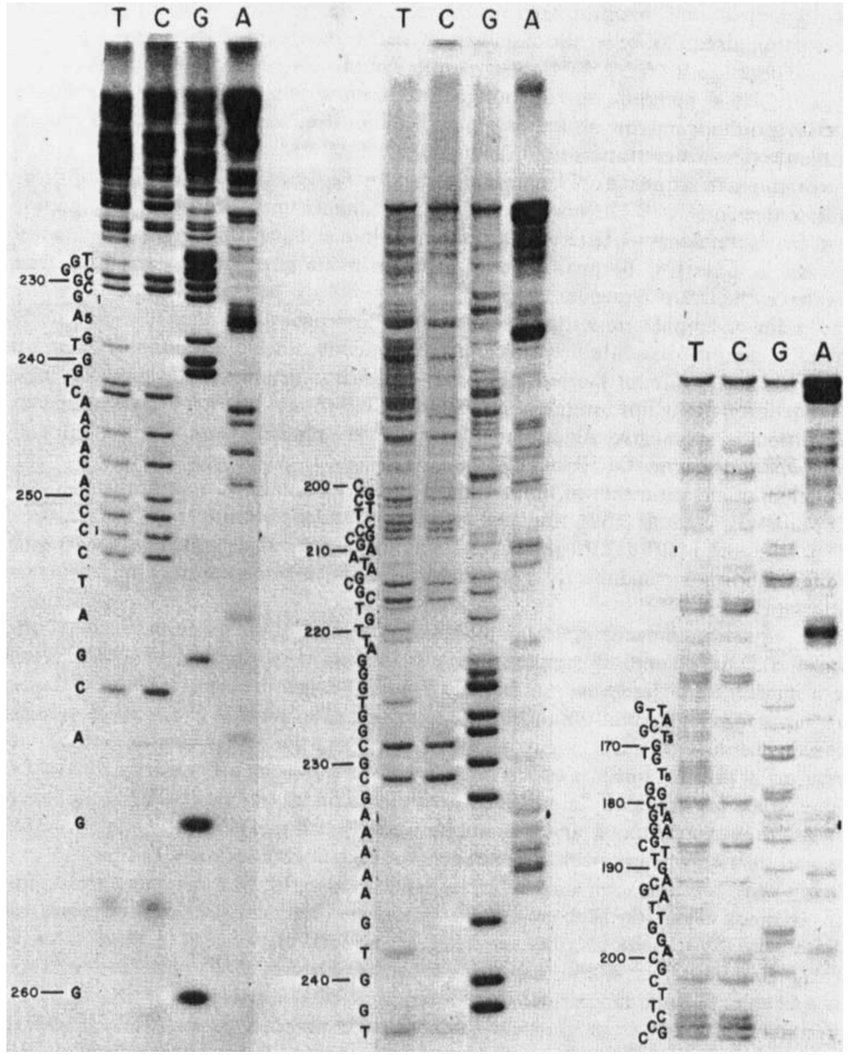
- **Primary analysis**
- Varying approaches, speed, accuracy, read length and price
- Used library preparation protocol depends on sequencing technology

# 1st generation

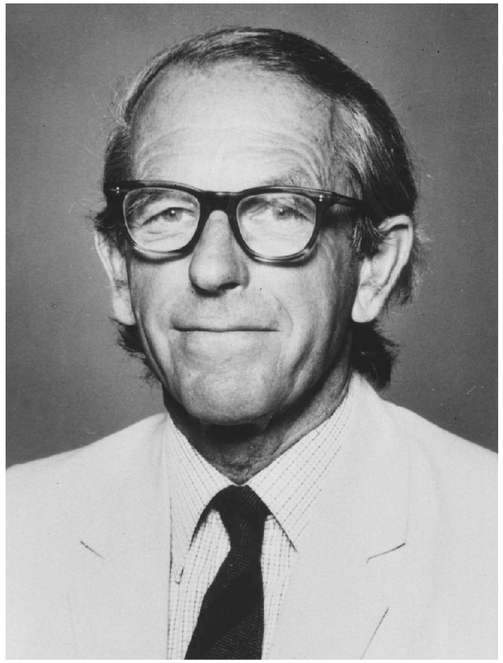
*Sanger sequencing*

# Short detour: Maxam-Gilbert sequencing

- **Chemical sequencing**
- Allan Maxam & Walter Gilbert
- Developed in 1977
- Quickly replaced by Sanger sequencing

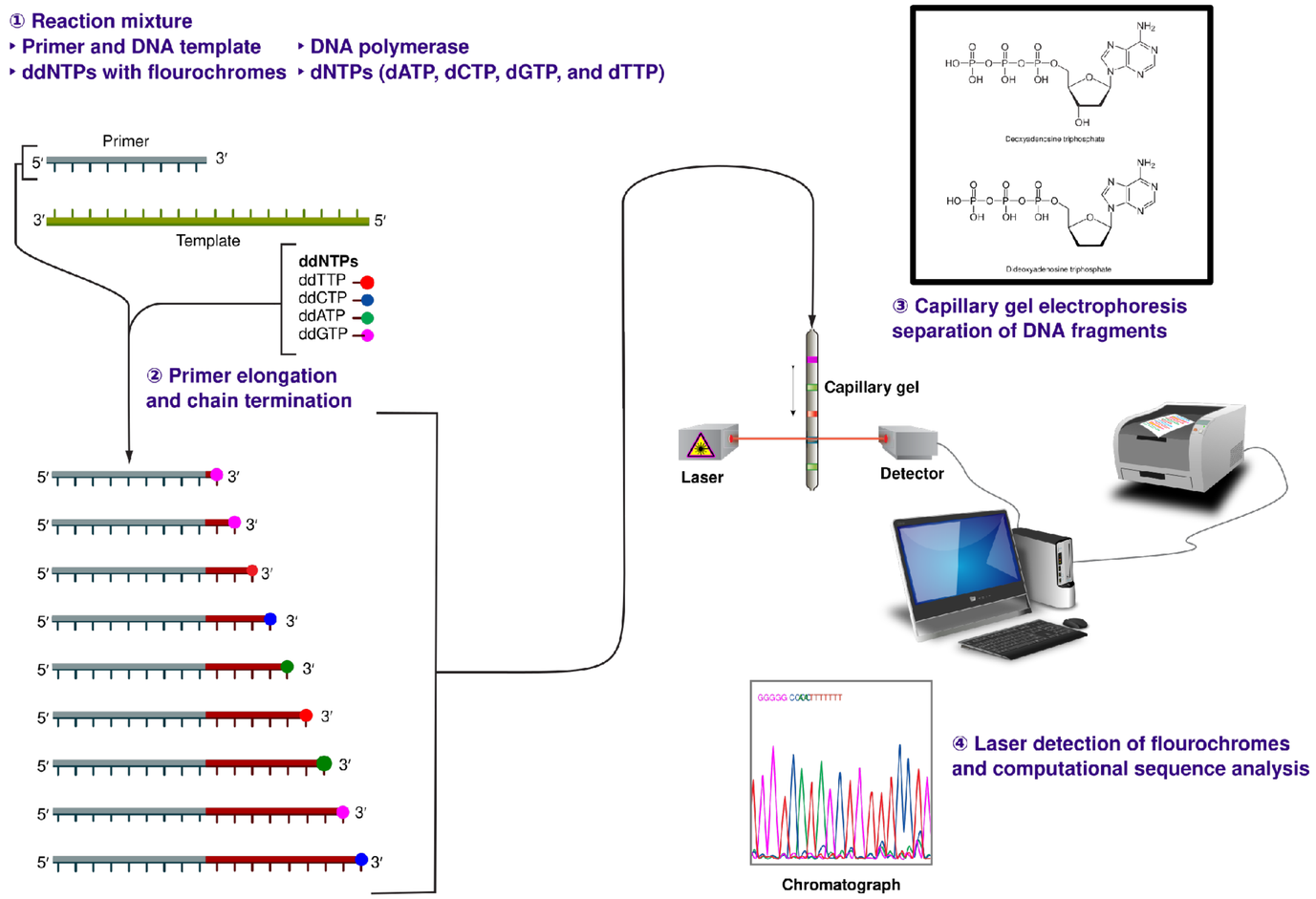


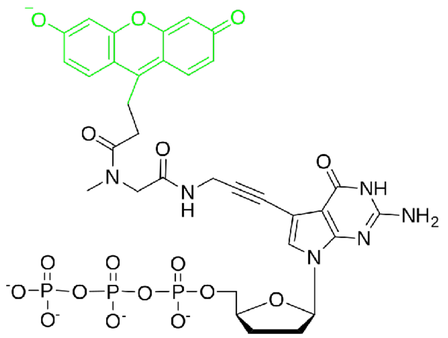
# Sanger sequencing



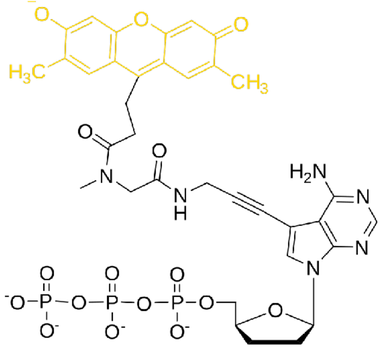
- **Chain-termination method**
- Frederick Sanger
- Developed in 1977
- Uses less toxic chemicals and lower amounts of radioactivity than Maxam-Gilbert sequencing
- Used to construct first human genome in 2001
- **Very expensive**: $100 million in 2001, $10 000 in 2011 per one human genome

# Different bases emit different “colour”

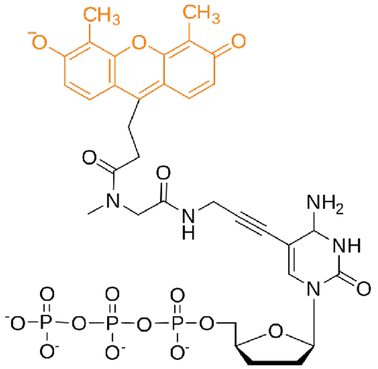


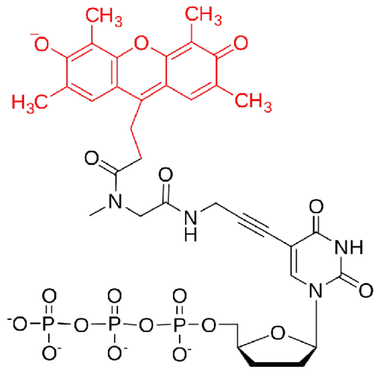
G-505



A-512

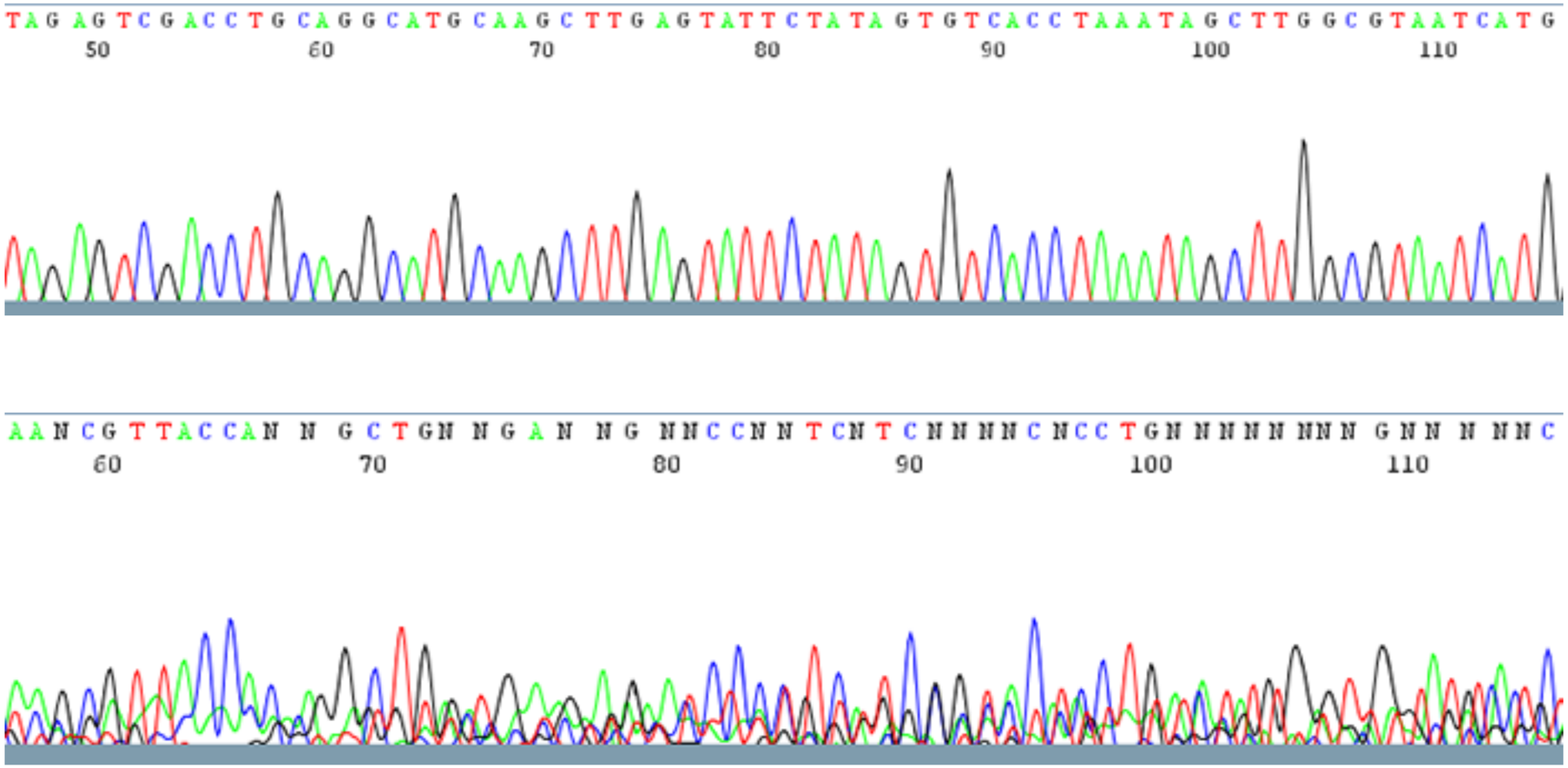


C-519



T-526

# Expectation vs. reality



# 2nd generation

*Next generation sequencing (NGS)*

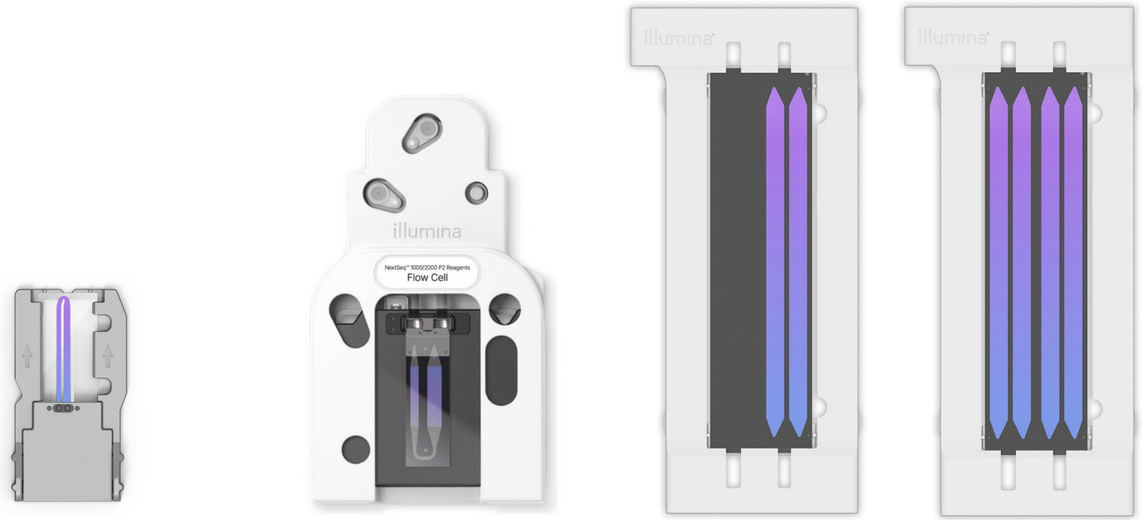
*Massive parallel sequencing (MPS)*

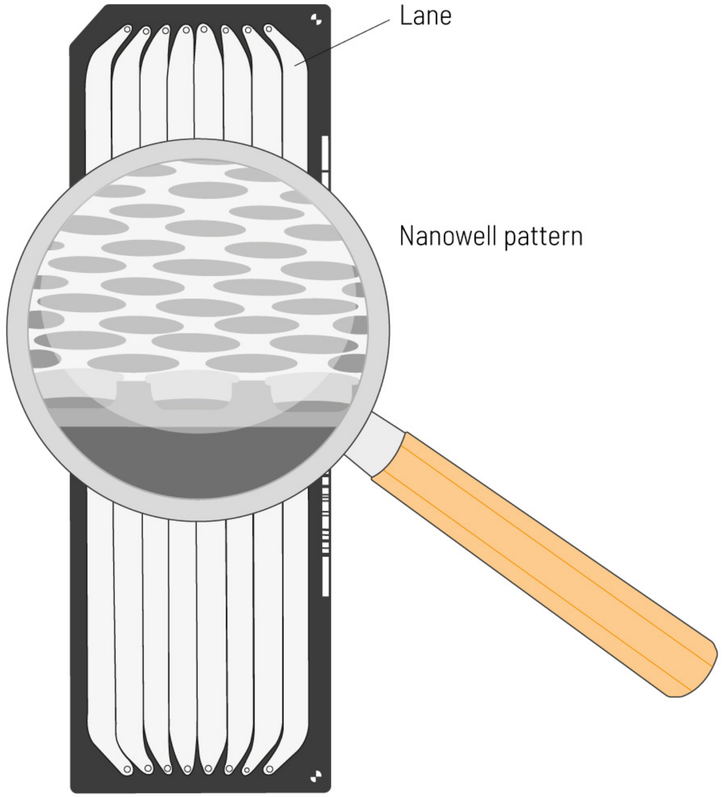
# Next generation sequencing

- Aimed to reduce the price per genome below $1 000
- One sequencing run produces between 1 million and 50 billion reads, depending on platform and sample type
- **Capable of sequencing only short reads (< 400 bp)**
- **Takes advantage of multiplexing to sequence multiple samples at once**

# Sequencing happens on a flow cell

- Reagent cartridge + flow cell
- Cartridge contains chemicals necessary for synthesis
- Flow cell is where the reaction happens

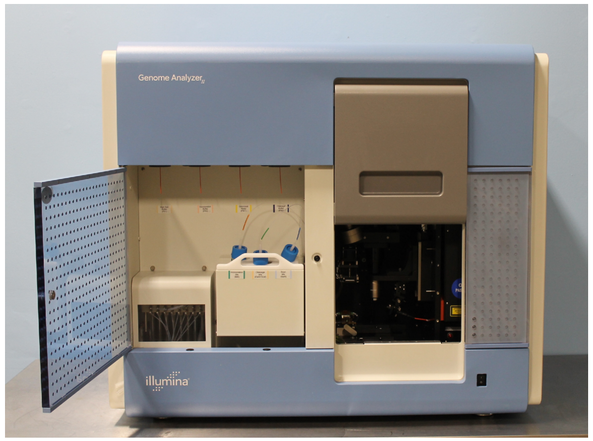



300 Cycles
P4

# Quick overview of NGS

- **2003**: 454 Life Sciences - GS20 / GS FLX
- **2006**: Solexa - Genome Analyzer
- **2007**: Applied Biosciences - SOLiD
- **2011**: Ion Torrent Systems - Ion Proton
- **2019**: MGI / BGI - DNBSEQ-T7
- **2022**: Element Biosciences - AVITI
- Usually acquired by Illumina, Roche or other biotech giants




# Illumina

- **Sequencing-by-Synthesis** based on technology acquired from **Solexa**
- **Largest supplier of NGS technology in the world**
- Sequencing platforms, chemistry, secondary analysis software DRAGEN
- Multiple sequencing platforms:
  - Small genomes (bacteria and viruses): iSeq, MiniSeq, MiSeq
  - Medium genomes (low coverage human and other mammals): HiSeq, NextSeq
  - Large genomes: NextSeq, NovaSeq

# Most recent Illumina sequencers

- Small genomes: 2 × 150 - 2 × 300 bp reads, 1 - 30 Gb per flow cell
- Medium genomes: 2 × 150 - 2 × 300 bp reads, 30 - 540 Gb per flow cell
- Large genomes: 2 × 150 - 2 × 250 bp reads, 540 Gb - 8 Tb per flow cell


iSeq (2018)


NextSeq 2000 (2020)


NovaSeq X (2022)

# 3rd generation

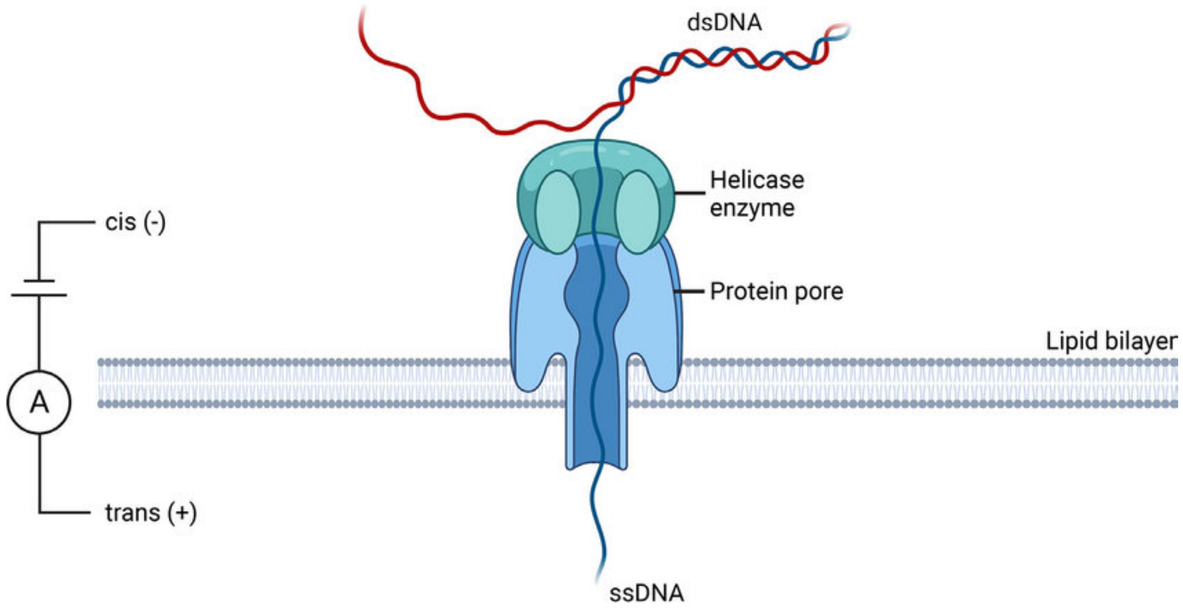
*Long-read sequencing*

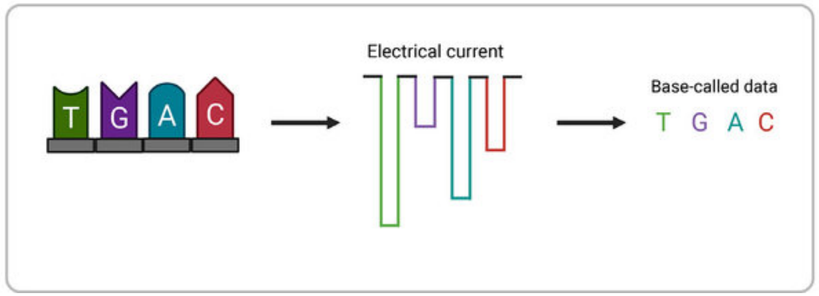
# Long-read sequencing

- **Main downside of NGS are short reads**:
  - Useless for identification of large structural variants
  - Poor accuracy in highly repetitive regions
  - Still quite expensive
- Technology emerged in 2008
- Initially had low accuracy, but improved over time
- **10 kbp to more than 4 Mbp reads**

# Oxford Nanopore Technology (ONT)

- **Nanopore sequencing**
- Run can be stopped and resumed later
- Ultra long reads - 100 kb to > 4 Mb
- Lower accuracy than NGS



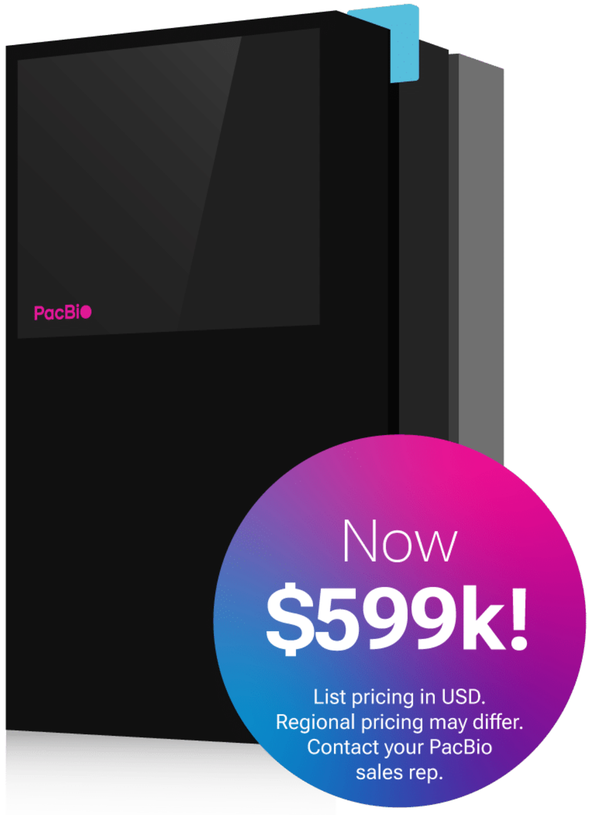



# ONT sequencers


- **MinION** (2015):
  - 48 Gb per flow cell
  - Small and portable - was successfully used on ISS
  - **GridION** (2017) - parallel processing of up to 5 MinIONs
- **PromethION** (2018 - ):
  - Can process 1 - 48 flow cells at once
  - Up to 290 Gb per flow cell, 13.3 Tb in total

# Pacific Biosciences (PacBio)

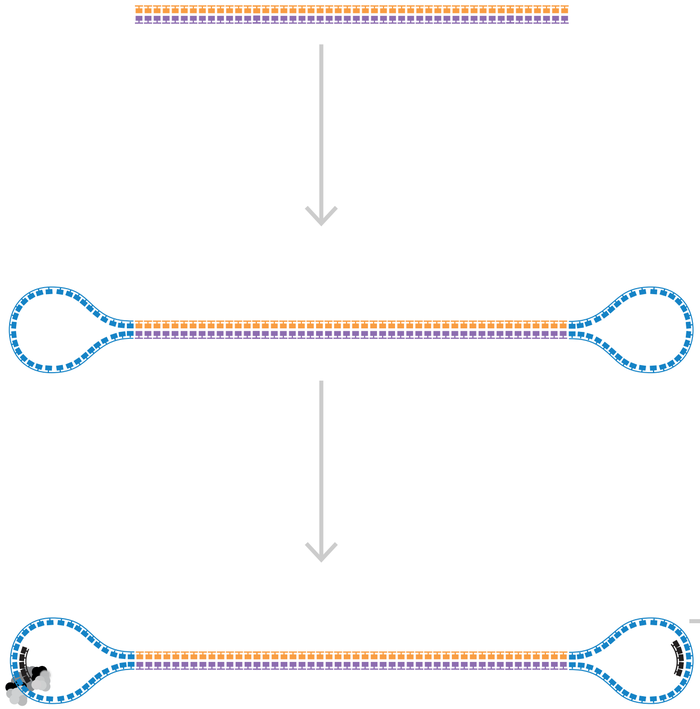
- **High Fidelity (HiFi) sequencing**
- Very accurate, but only thanks to a "trick" - more than 90% of bases are Q30+
- Shorter reads than ONT - up to 25 kb
- **PacBio Revio** (2022):
  - 480 Gb per flow cell
  - **Cheap**: $995 per whole genome with 30x coverage



Start with high-quality double stranded DNA

Prepare SMRTbell libraries

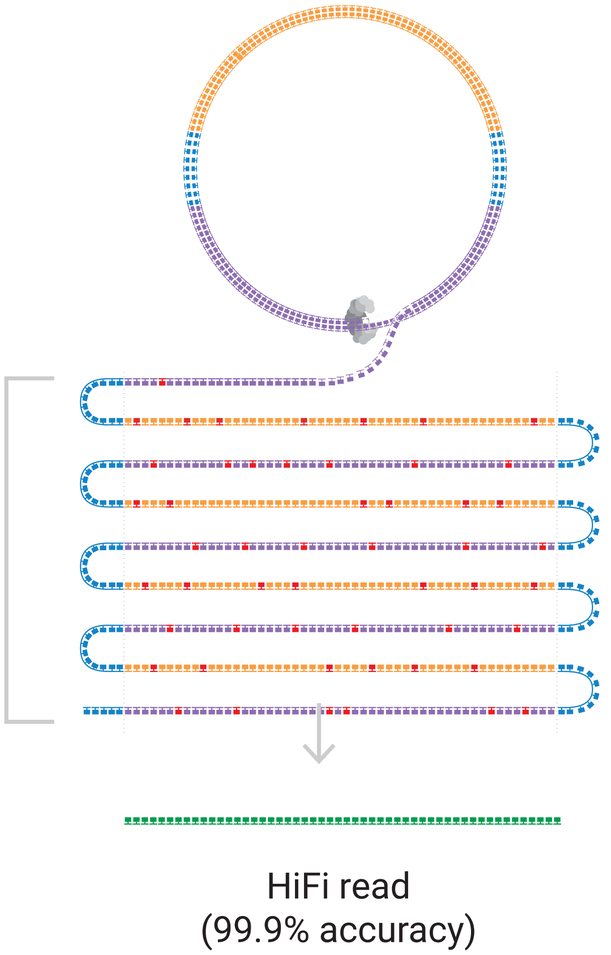
Anneal primers and bind DNA polymerase



Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads

Consensus and methylation status are called from subreads



HiFi read
(99.9% accuracy)

# Base calling & FASTQ

*What remains to do after sequencing*

# What are the sequencing outputs?

- Raw output is an array of light / electric current measurements

  + quality

  + identifiers

- Multiple file types:

  - **Binary Alignment Map (BAM)** - PacBio
  - **FASTQ** - BGI / MGI
  - **Binary Base Call (BCL)** - Illumina
  - **FAST5, POD5** - ONT

**We can work with them**

**But not with these**

# Base calling can help with that

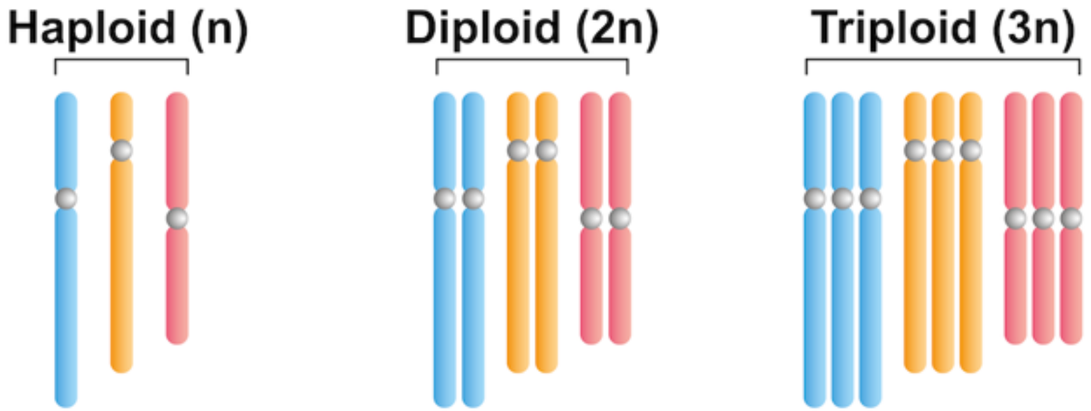
- **Base calling**: conversion of signal into bases [A, C, G, T (, U)]
- Transformation of raw files into unified format - **FASTQ**
- **How important is base caller accuracy?**
  - Extremely
  - Small mistake during base calling can cause big problems later
- **Demultiplex**: use index to split pooled libraries into individual samples

# Base calling can help with that

- Is there a universal base caller?
  - Of course not :)
  - Each technology has its own, sometimes more than one
- **Sanger**: *KB Basecaller* (2007), *Smart Deep Basecaller* (2022)
- **Illumina**: *CASAVA* (2009), ***bcl2fastq*** (2013)
- **ONT**: *Guppy* (2017), ***Dorado*** (2022), ***Bonito*** (2020 - uses RNN to improve quality)

# But wait! Humans are diploid organisms!

- **Ploidy**: number of full sets of chromosomes in a single cell
- **Haploid**: one set of chromosomes - male bees, wasps and ants, sperm and ova, plants and fungi alternate between haploid and diploid
- **Diploid**: two sets, one from father and the other one from mother - humans and other mammals
- **Polyploid**: three and more sets - plains viscacha rat - 4x, wheat - 6x, strawberry - 10x, black mulberry - 44x



# So, how do we know which set the read came from?

- **We don't**
- Because sequencing is a stochastic process, we assume symmetrical binomial distribution of read counts between alleles - **the sets *should* be balanced**
- Biases (GC, PCR, mapping, ...) will be solved during secondary analysis
- **However, there is one more problem**:
  - One sample can be more reactive (sequencing compatible) than another
  - Fragments from this sample will have higher coverage than other samples

# Repooling

- Second run with altered proportion of libraries can balance the final coverage
- If a sample exceeds the required coverage after first run, it is **not included in the second pool**

| Sample | Proportion in 1st pool | Coverage after 1st run | Proportion in 2nd pool | Coverage after 2nd run | Final coverage |
|---|---|---|---|---|---|
| **A** | 33 % | 5x | 62 % | 43x | **48x** |
| **B** | 33 % | 60x | **REMOVED FROM POOL** | **0x** | **60x** |
| **C** | 33 % | 22x | 38 % | 40x | **62x** |

# Sequencing quality

- **Phred score**: base qualities encoded as ASCII characters ranging from **!** (lowest, 100% chance of incorrect call) to **K** (highest, 0.006% chance of incorrect call)

- **Phred is a text representation of Q score**

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJK
```

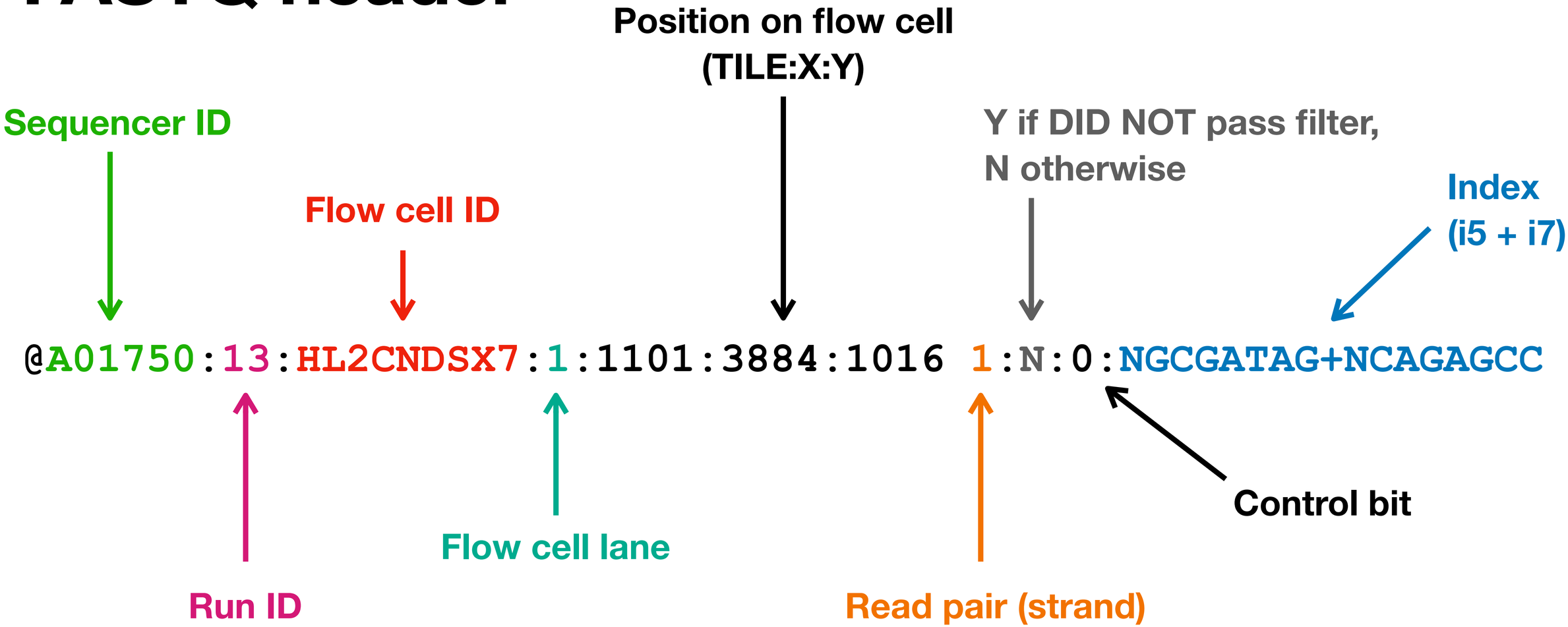
# FASTQ file

- **Text-based**: easy to read and manipulate, not space efficient
- Paired-end: two files with **_R1 and _R2 suffixes** (one file for each strand)
- 1 read = 4 lines:
    1. Header - starts with @
    2. Sequence of bases
    3. Separator
    4. Sequence of phred qualities

# FASTQ file

```
@A01750:13:HL2CNDSX7:1:1101:3884:1016 1:N:0:NGCGATAG+NCAGAGCC
ANTTTGGGCATCATGGAGGGAAGCAAACCCCCAGTAACTGGGCAGCTCAGTTCTGCTGGACCCTTCAGGAGGCA
+
F#F:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A01750:13:HL2CNDSX7:1:1101:4481:1016 1:N:0:NGCGATAG+NCAGAGCC
GNAATCGTGTTTCCCCTATTTCTTTGCCTTCTCTATTTTCTCTTTTTTCTTTTAGCCTACTGTAGGAGGCATAT
+
F#FFFFFFFFFFFFFFFFFFFFFFFFF:FFF:FF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FF
@A01750:13:HL2CNDSX7:1:1101:4553:1016 1:N:0:NGCGATAG+NCAGAGCC
GNTTGCTTTAACAGCCCTAAGCAGTTCCCAGCATGACTTATTCCTTTAGCTAAGTGATTTTGGGGGCCCCCAAG
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFF,,FFFFFFFFFFFFFFFFFFFF
```

# FASTQ header



# What can we do with FASTQ?

- Reads are not aligned - we don't know which part of genome they came from
- **Alignment** - *BWA, Bowtie2, STAR, Minimap2, vg giraffe*, **...** :
  - Maps reads to reference genome
  - Creates SAM / BAM file
- **Pre-processing:**
  - Trimming & filtering - *Trimmomatic, Cutadapt*
  - Quality control - *FastQC*
  - Expression quantification - *Salmon* (**RNA-Seq only!**)

# Questions?