Introduction to Next-Generation Sequencing
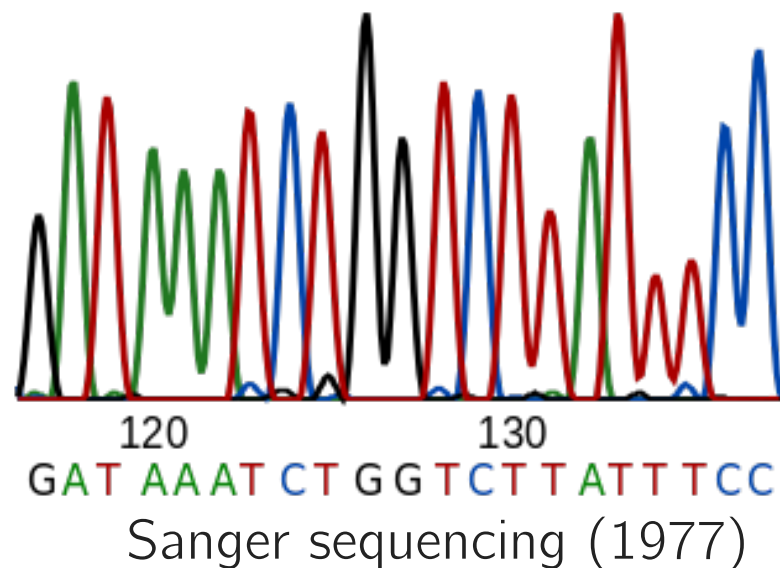
Joanna Krupka

CRUK Summer School in Bioinformatics

Cambridge, July 2019

CANCER RESEARCH UK

MRC | Cancer Unit
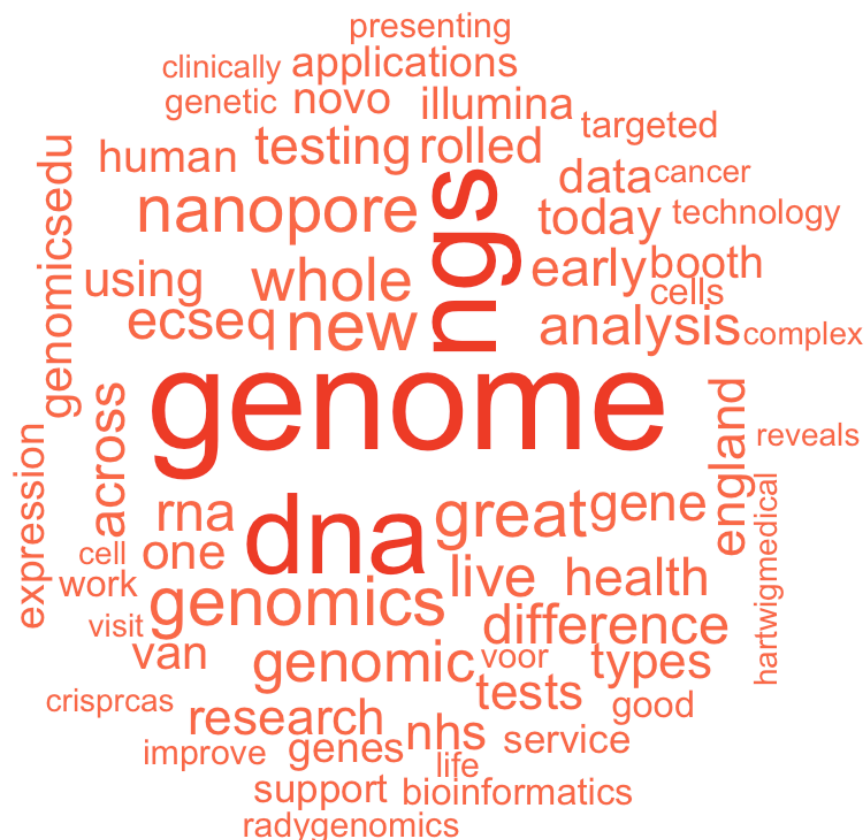
UNIVERSITY OF CAMBRIDGE

Sanger sequencing (1977)

Human Genome Project

1990 - 2006

DNA Sequencing Technologies Key to the Human Genome Project

By: Heidi Chial, Ph.D. (*Write Science Right*) © 2008 Nature Education

Citation: Chial, H. (2008) DNA sequencing technologies key to the Human Genome Project. *Nature Education* 1(1):219
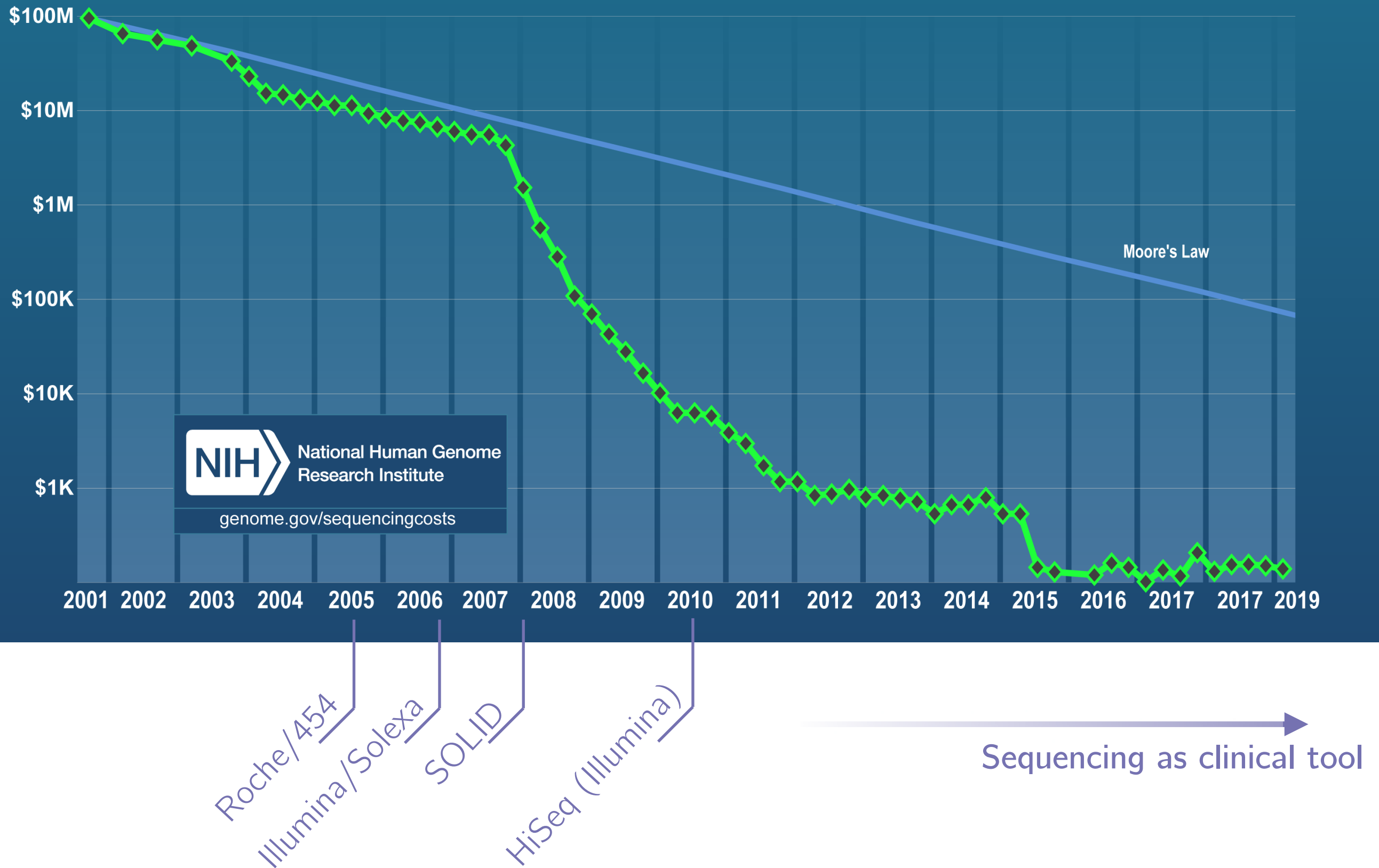

Next Generation Sequencing

mid 2000–present

= high-throughput sequencing

quicker and cheaper parallel sequencing of DNA and RNA

Next generation sequencing

Short-read NGS · · · · · · · · · · · · · · · Long-read NGS

"Second-generation sequencing" · · · · · "Third-generation sequencing"
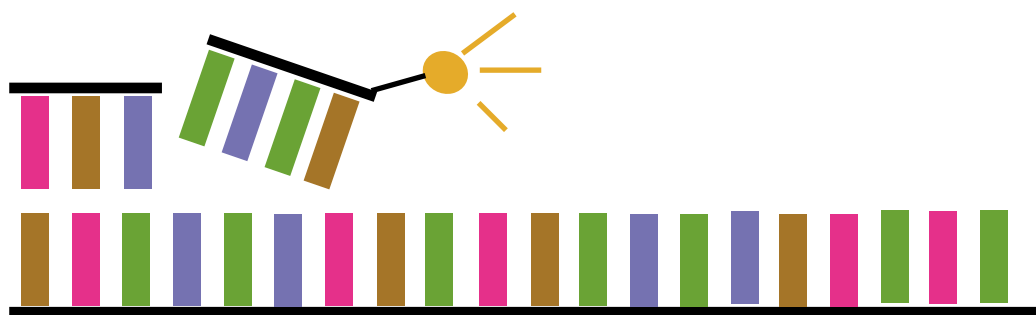- error rates (0.1–15%)
- read lengths (35–700 bp)

Sequencing by ligation
SOLiD

Sequencing by synthesis
Illumina/Solexa
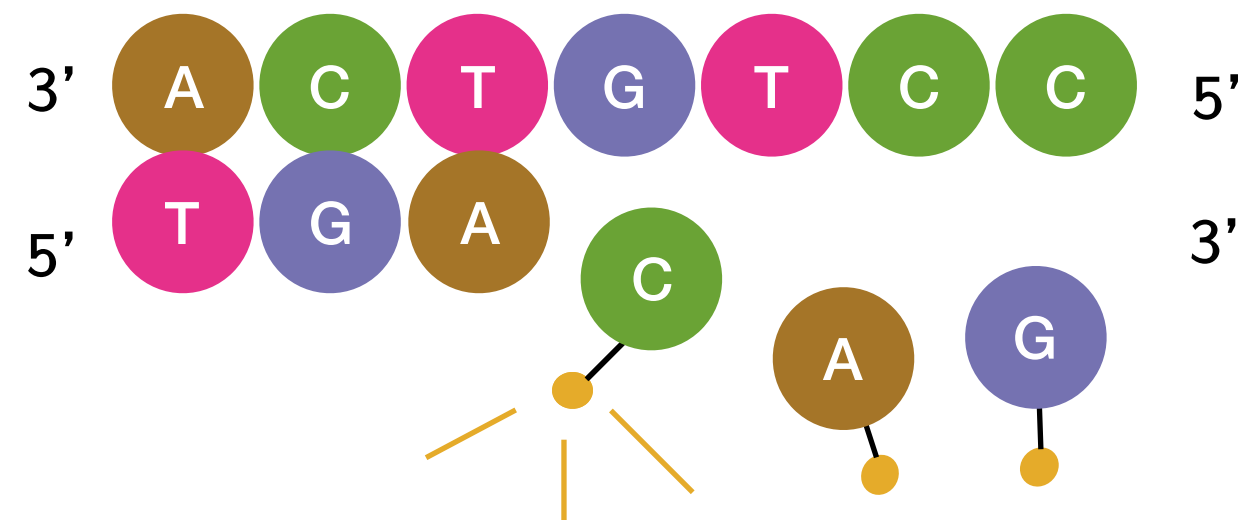
3' A C T G T C C 5'

5' T G A 3'

C

A G

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.

Next generation sequencing

Short-read NGS

"Second-generation sequencing"

Long-read NGS

"Third-generation sequencing"

Real-time long read sequencing

Pacific Biosciences
Oxford Nanopore Technologies

Synthetic long-read sequencing

Illumina
10X Genomics

Single cell focus
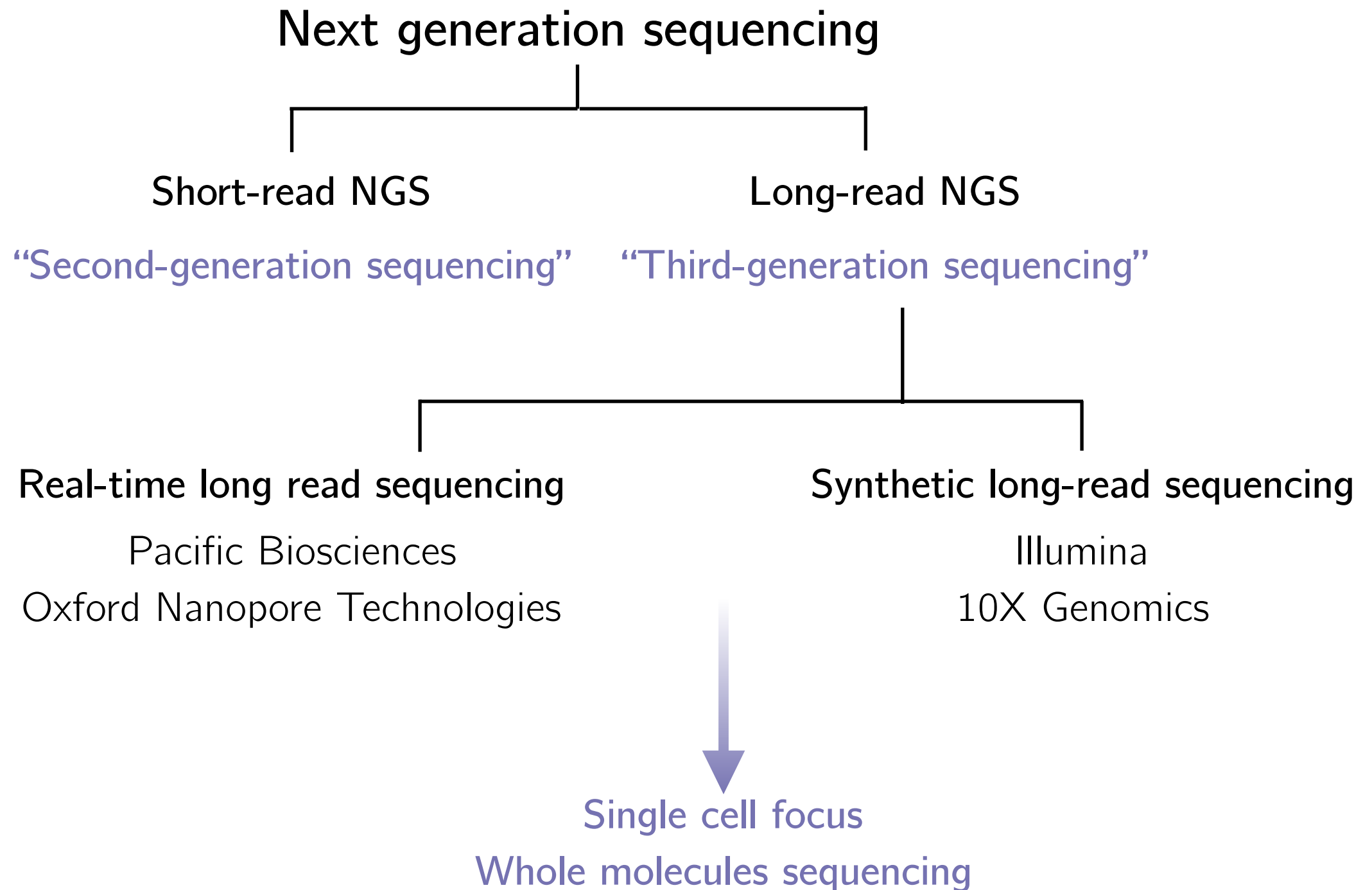Whole molecules sequencing

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.
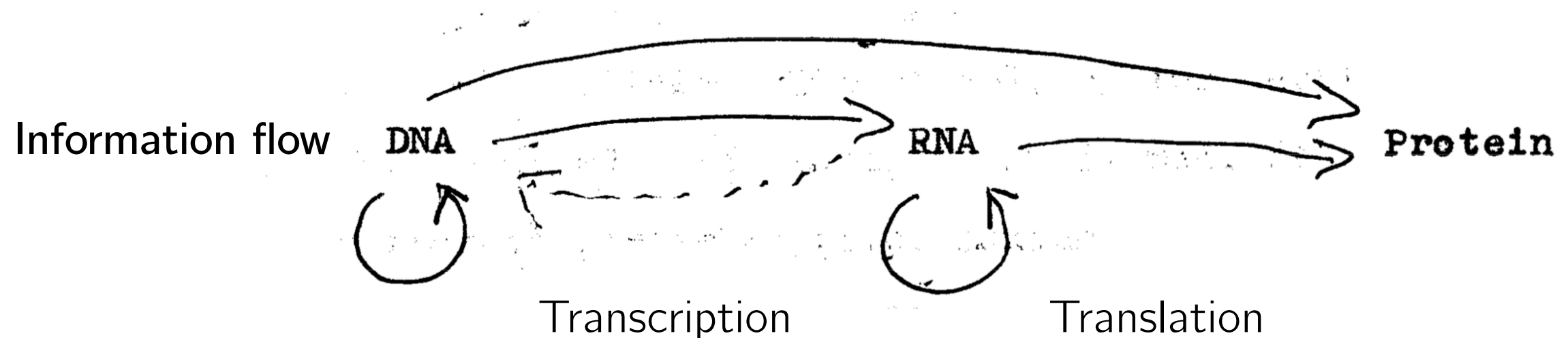
Sequencing techniques

Central dogma of molecular biology (Crick F. 1958)

Information flow — DNA → RNA → Protein

Transcription · Translation

Whole genome sequencing

Whole exome sequencing

HiC-Seq · ChIP-Seq · ATAC-Seq · ...

RNA-Seq · scRNA-Seq · Ribo-Seq · SLAM-Seq · ...

DNA · RNA

Based on the Solexa technology developed by **Shankar Balasubramanian** and **David Klenerman** at the University of Cambridge (1998)

1 **Library preparation**

b Solid-phase bridge amplification (Illumina)

Template binding
Free templates hybridize with slide-bound adapters

Sequence

2 **Bridge amplification**
Distal ends of hybridized templates interact with nearby primers where amplification can take place

3 **Cluster generation**
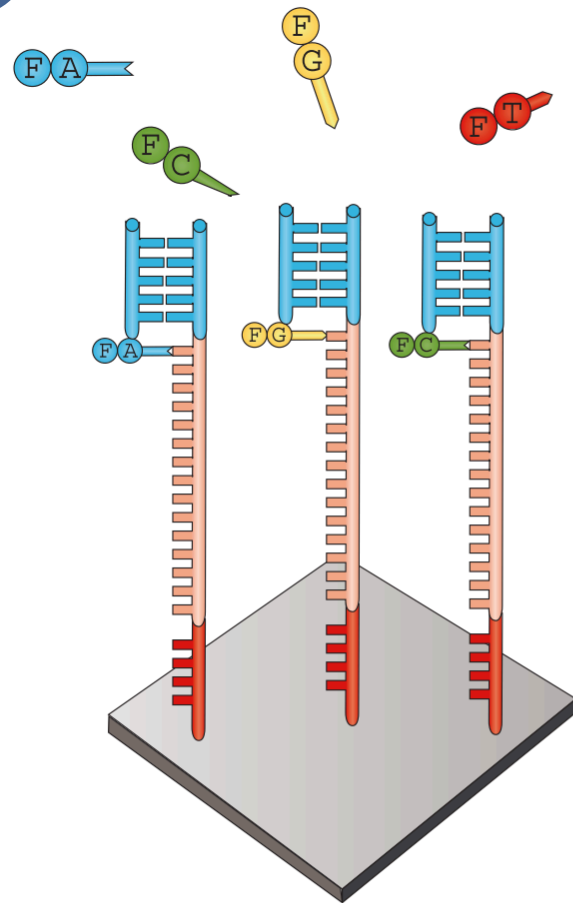After several rounds of amplification, 100–200 million clonal clusters are formed

Flow cell

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.
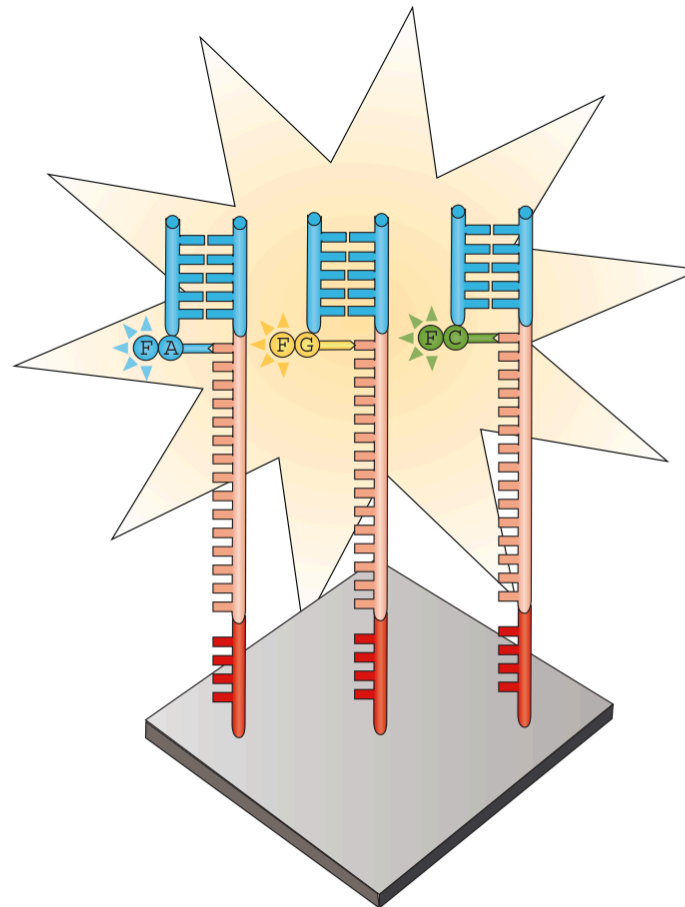
4 Sequencing using reversible terminators

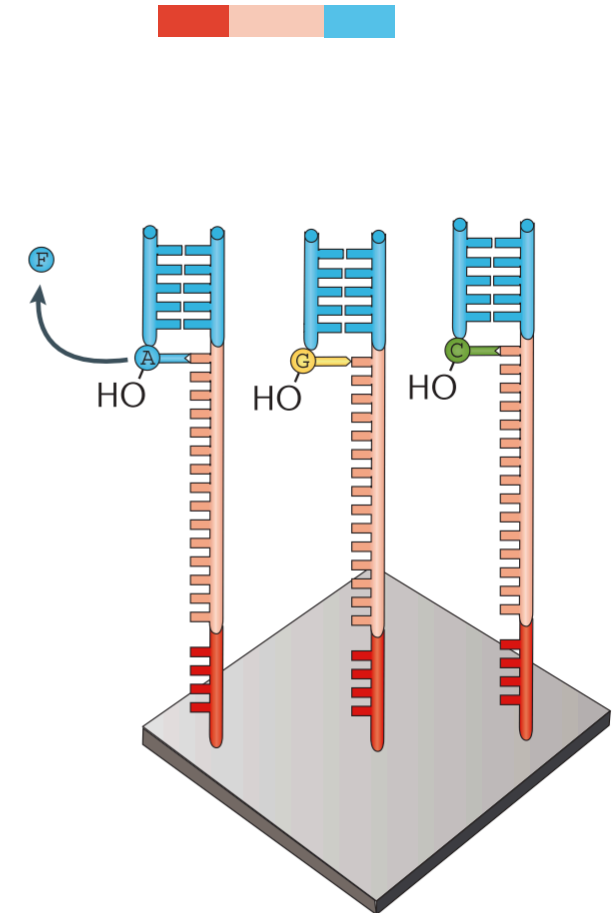**Nucleotide addition**
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.

Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.

Cleavage
Fluorophores are cleaved and washed from flow cells and the 3'-OH group is regenerated. A new cycle begins with the addition of new nucleotides.
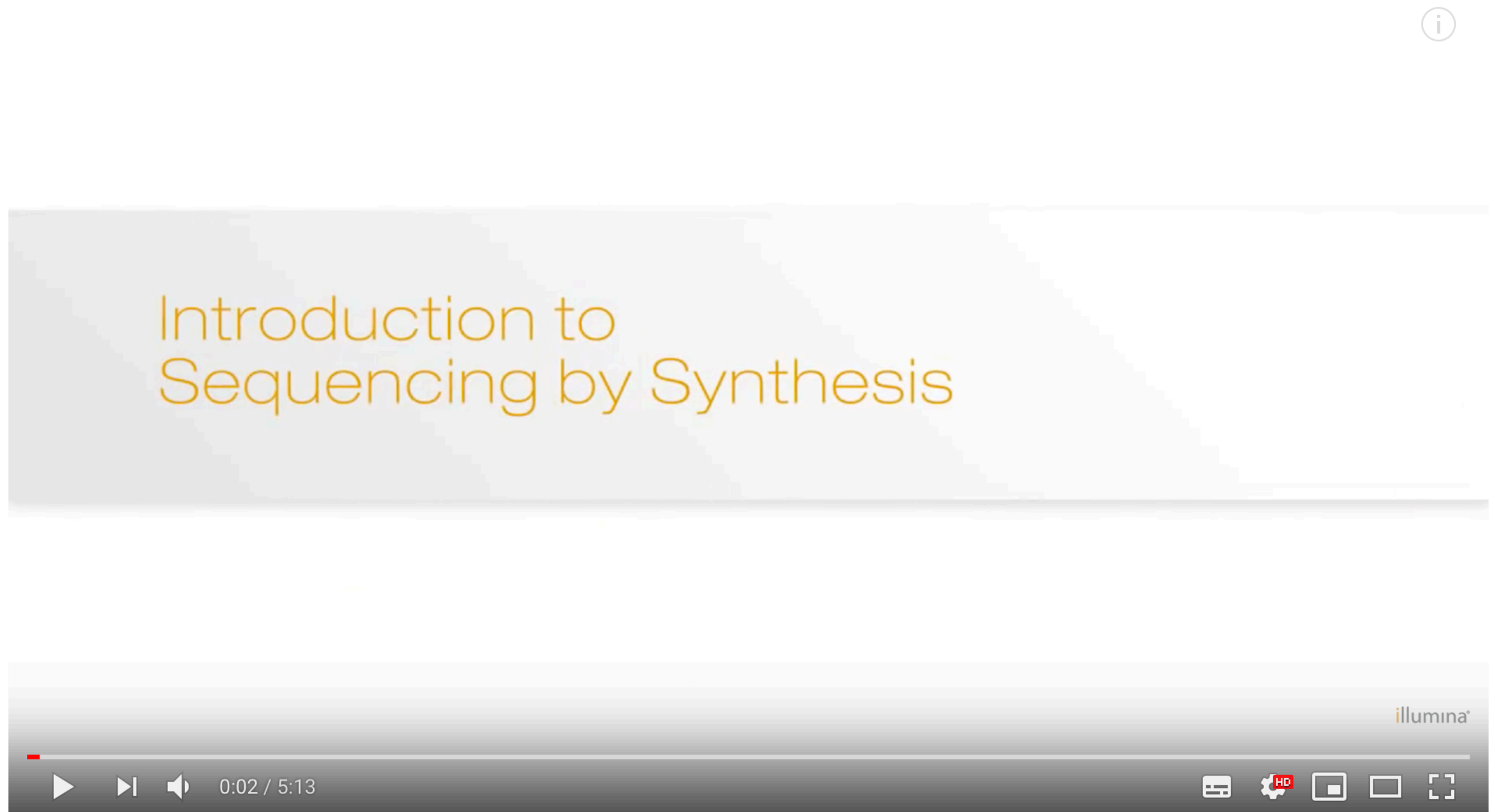
5 Output: sequence saved in FASTQ format

6 Bioinformatic analysis: quality check, alignment and data analysis

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.
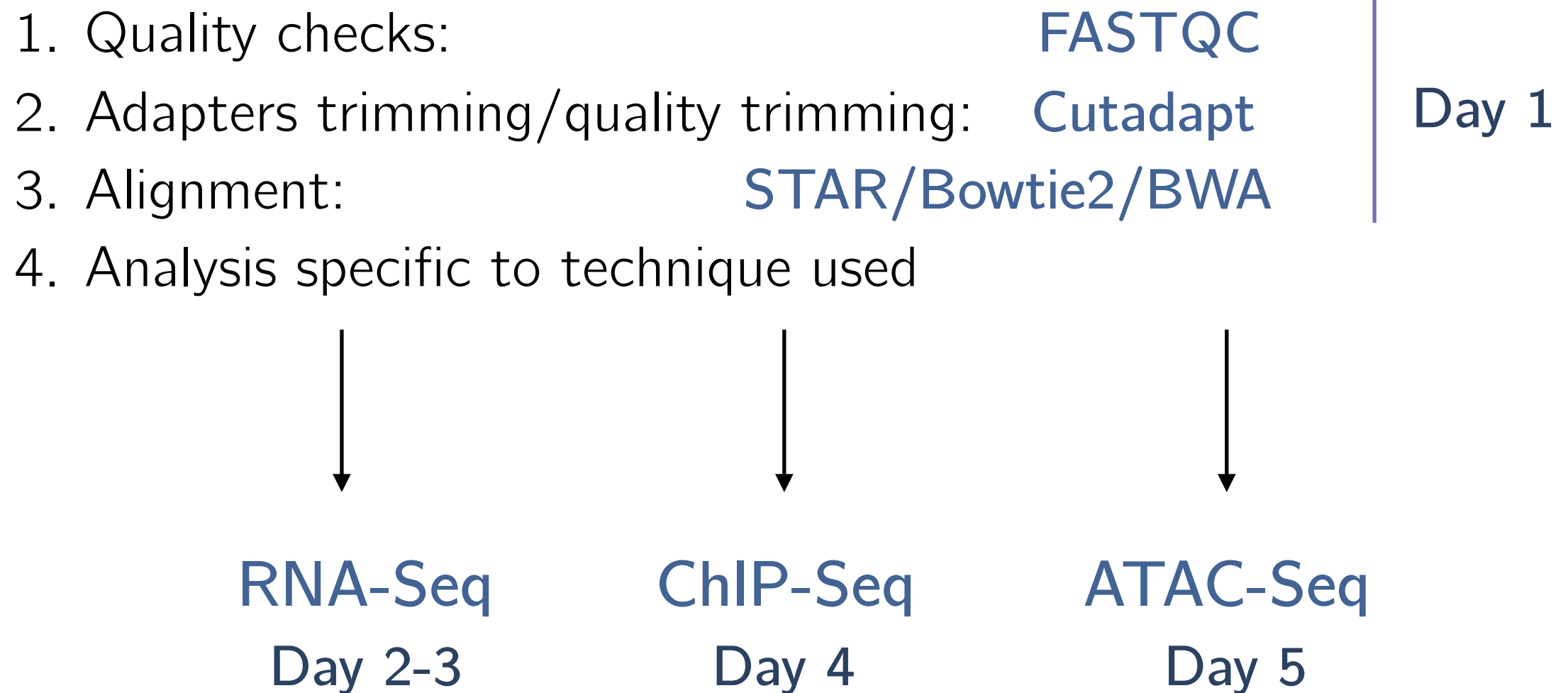
https://www.youtube.com/watch?v=fCd6B5HRaZ8

1. Quality checks: FASTQC
2. Adapters trimming/quality trimming: Cutadapt   Day 1
3. Alignment: STAR/Bowtie2/BWA
4. Analysis specific to technique used

RNA-Seq	ChIP-Seq	ATAC-Seq
Day 2-3	Day 4	Day 5

- Multiplexing gives the ability to sequence multiple samples at the same time.
- Useful when sequencing small genomes or specific genomic regions.

Different barcode adaptors are ligated to different samples.

Reads de-multiplexed after sequencing.

Source: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/multiplex-sequencing.html
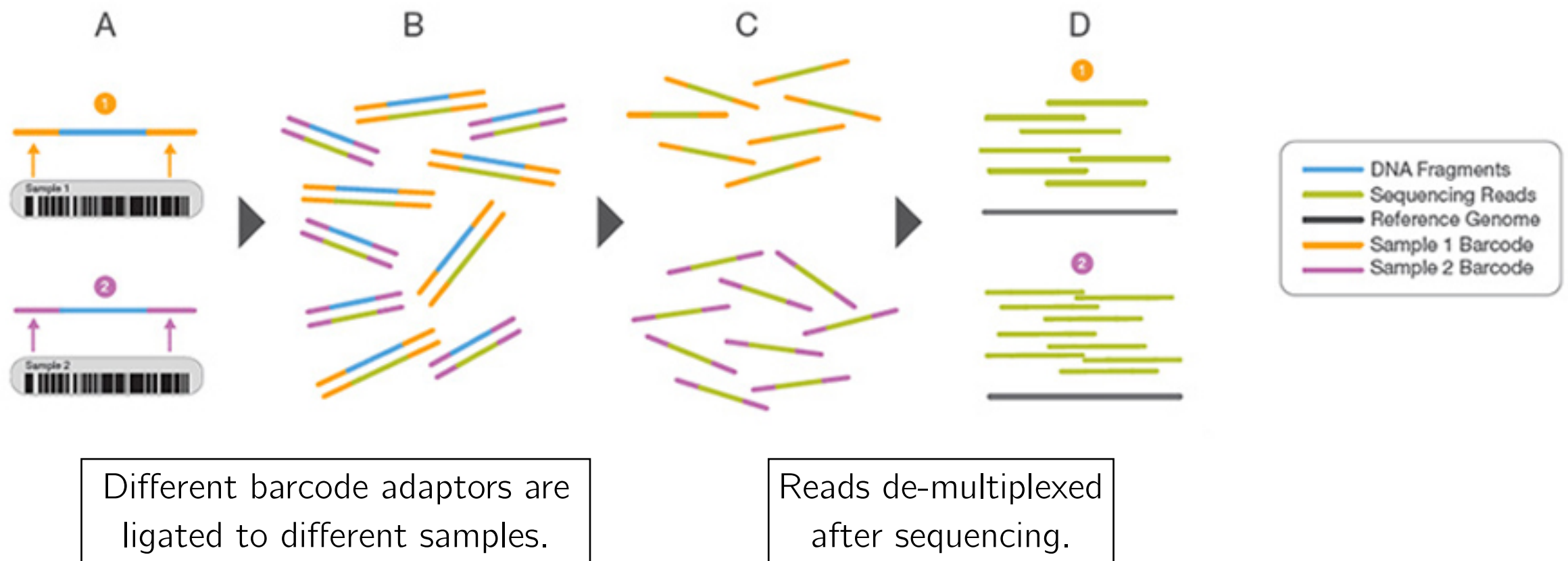
Sequencing data repositories

Example data sets

Study type	Recommended submissions route(s)	Data repository/ies	Recommended retrieval route(s)
Array-based mouse genotyping	MAGE-Tab	ArrayExpress	ArrayExpress
Small-scale sequence-based mouse genotyping	MAGE-Tab SRA-Webin	SRA	ArrayExpress SRA
Human (restricted access) genotyping	EGA	EGA	EGA

More about recommended data repositories: https://www.nature.com/sdata/policies/repositories

Data downloading: https://www.ebi.ac.uk/ena/browse/read-download

https://sites.psu.edu/yuka/2016/04/07/how-to-use-sra-toolkit/
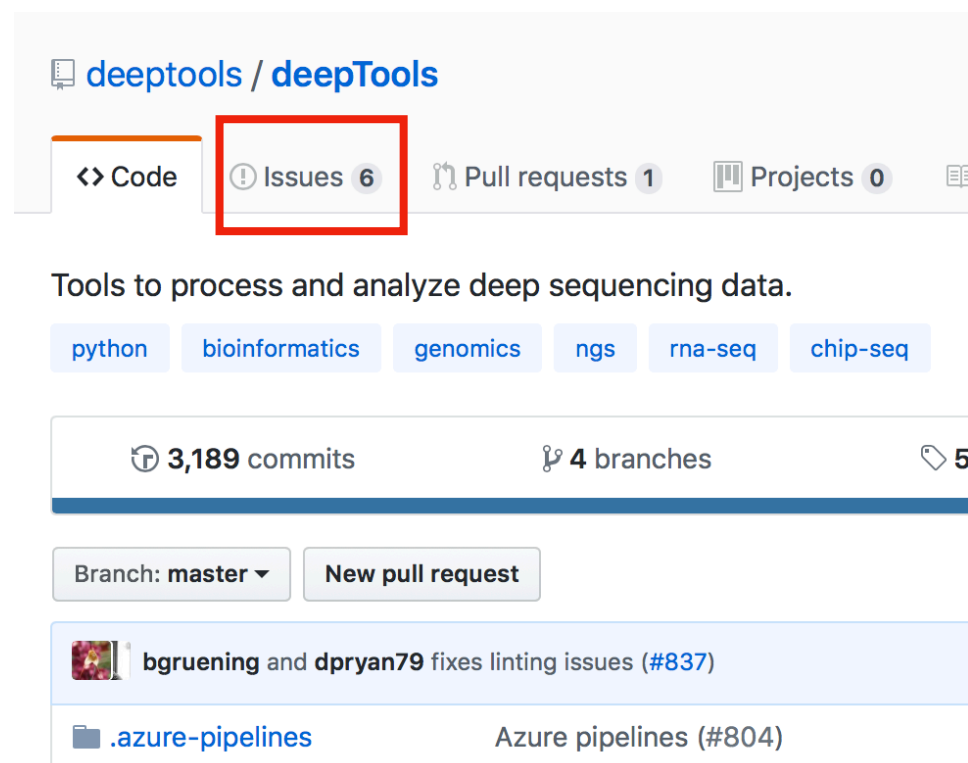
Still lost?

Google!

Package manual, GitHub

Bioinformatics forums and discussion groups:

https://www.biostars.org

https://support.bioconductor.org

http://seqanswers.com