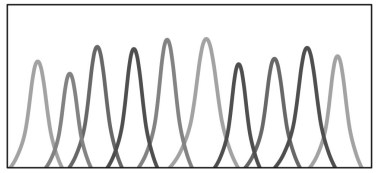
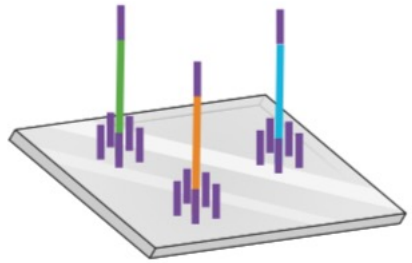
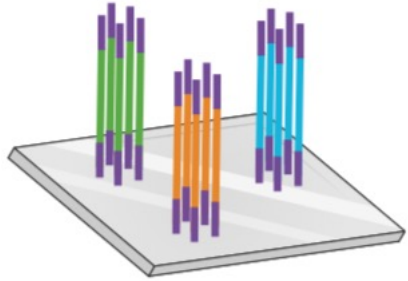




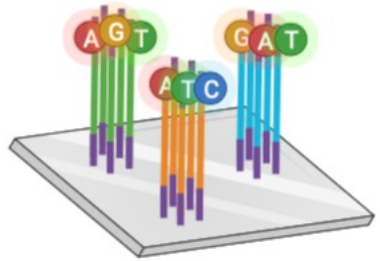
Introduction to Next-Generation Sequencing



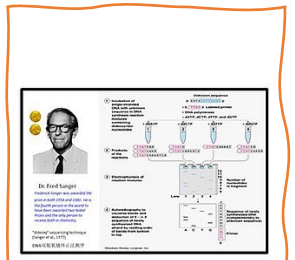
Junfan Huang

MRC Cancer Unit
University of Cambridge

CRUK Bioinformatics Summer School 2021
21th July 2021

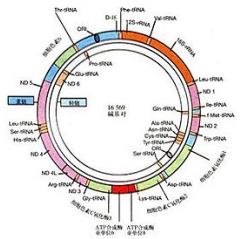


DNA sequencing



1977 Sanger method

1981 Human mitochondrial genome sequence



1990 Human Genome Project

1995 Complete cell genome

1996 Complete eukaryotic genome

2001 Complete the Human Genome Project

2005 Second generation sequencer：454 GS20

2007 Second generation sequencer：Genetic Analyzer 2

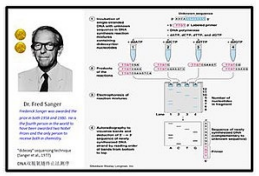
2008 Research Human Microbiome Project

2011 Third generation sequencer：PacBio RS

2014 Nanospace sequencing

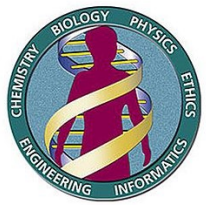
2019 The third stage Human microbiome project

DNA sequencing



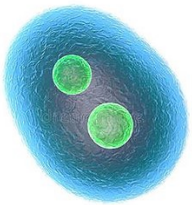
Sanger method

1977



Human Genome Project

1990



Complete eukaryotic genome

1996

Second generation sequencer：454 GS20

2005

Research Human Microbiome Project

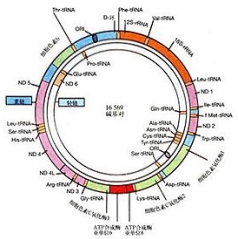
2008

Nanospace sequencing

2014

1981

Human mitochondrial genome sequence



1995

Complete cell genome

2001

Complete the Human Genome Project

2007

Second generation sequencer：Genetic Analyzer 2

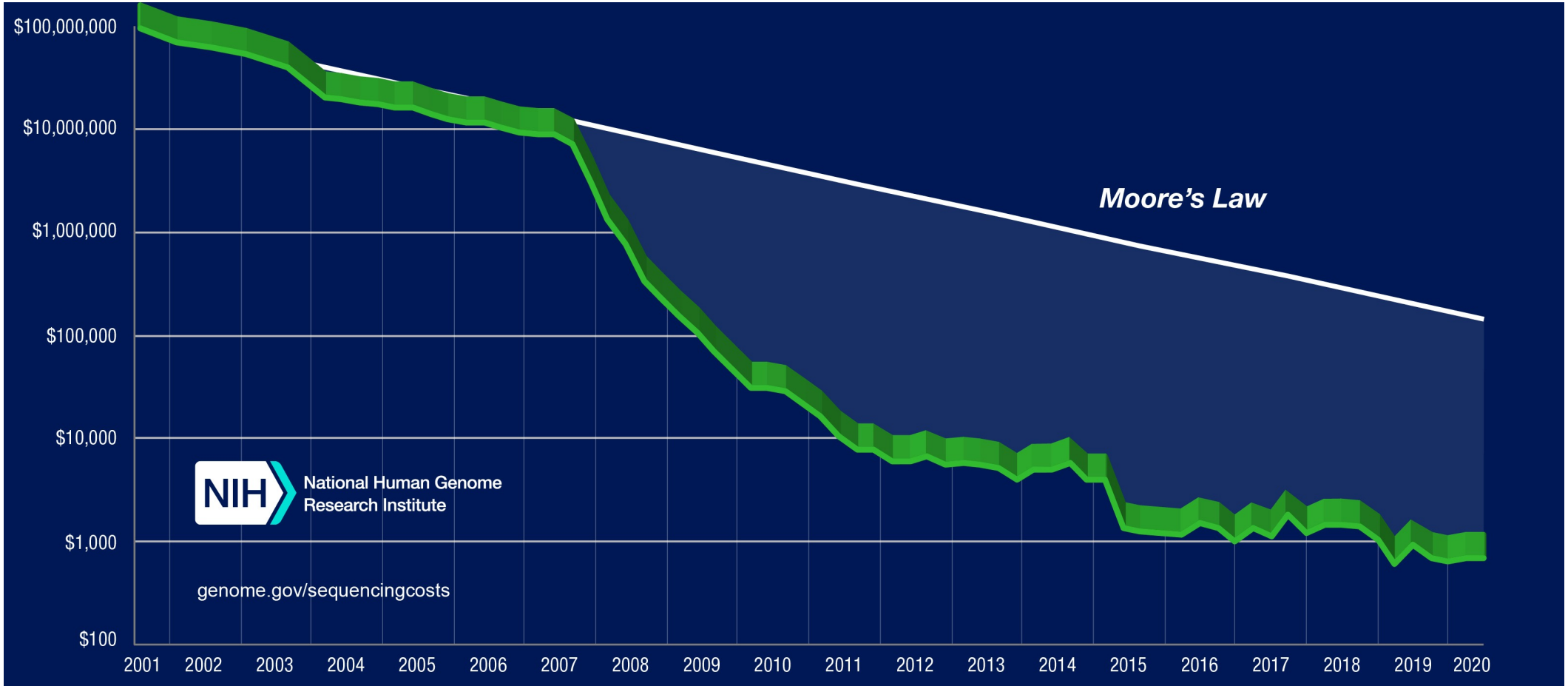
2011

Third generation sequencer：PacBio RS

2019

The third stage Human microbiome project

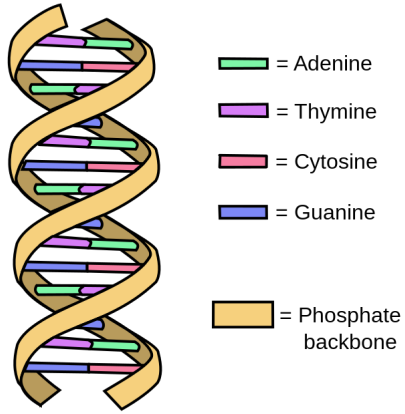
Cost per genome



Illumina sequencers

	MiSeq	HiSeq	NovaSeq	Sanger
Reads (millions)	30	3,000	13,000	0.0004
Gigabases/day	7	500	4000	0.001

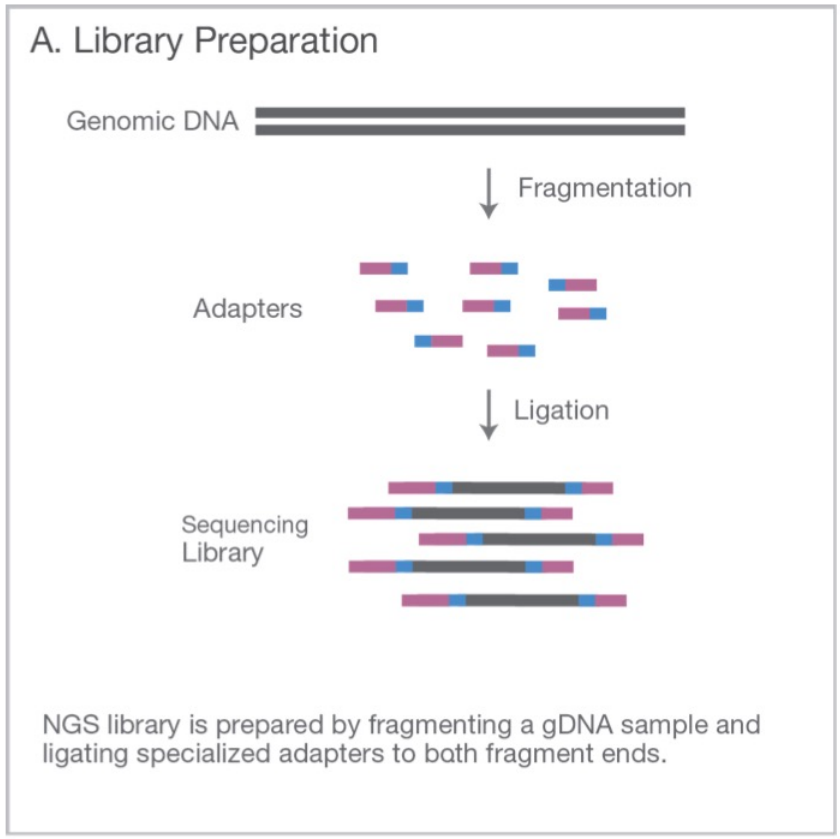
Illumina sequencers

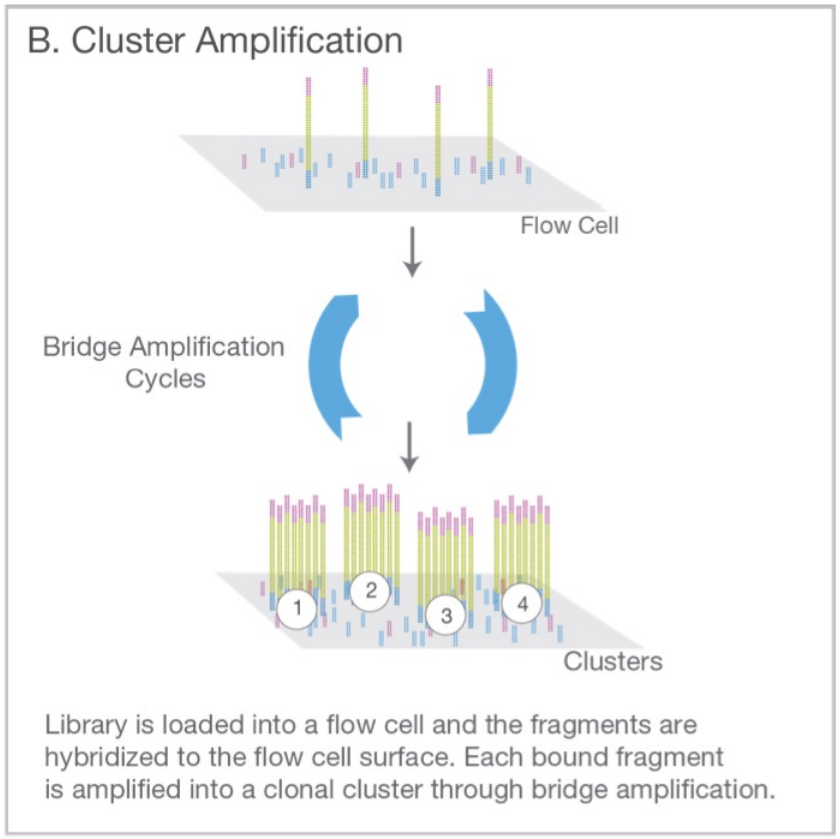


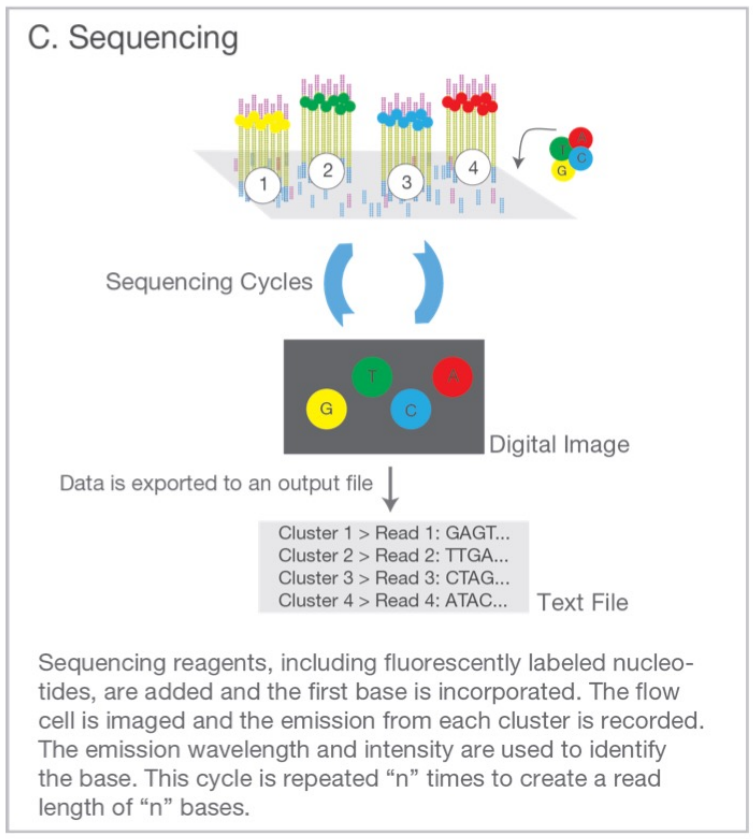
DNA

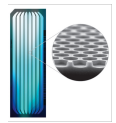
	MiSeq	HiSeq	NovaSeq	Sanger
Reads (millions)	30	3,000	13,000	0.0004
Gigabases/day	7	500	4000	0.001

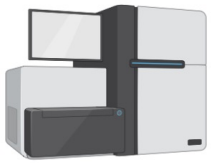
Illumina sequencing by synthesis









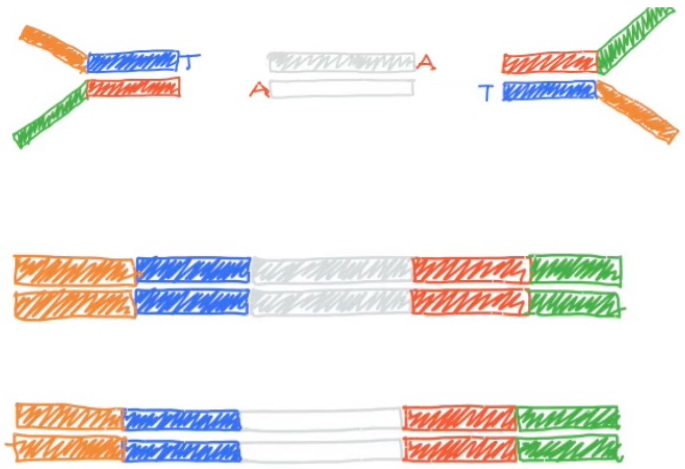


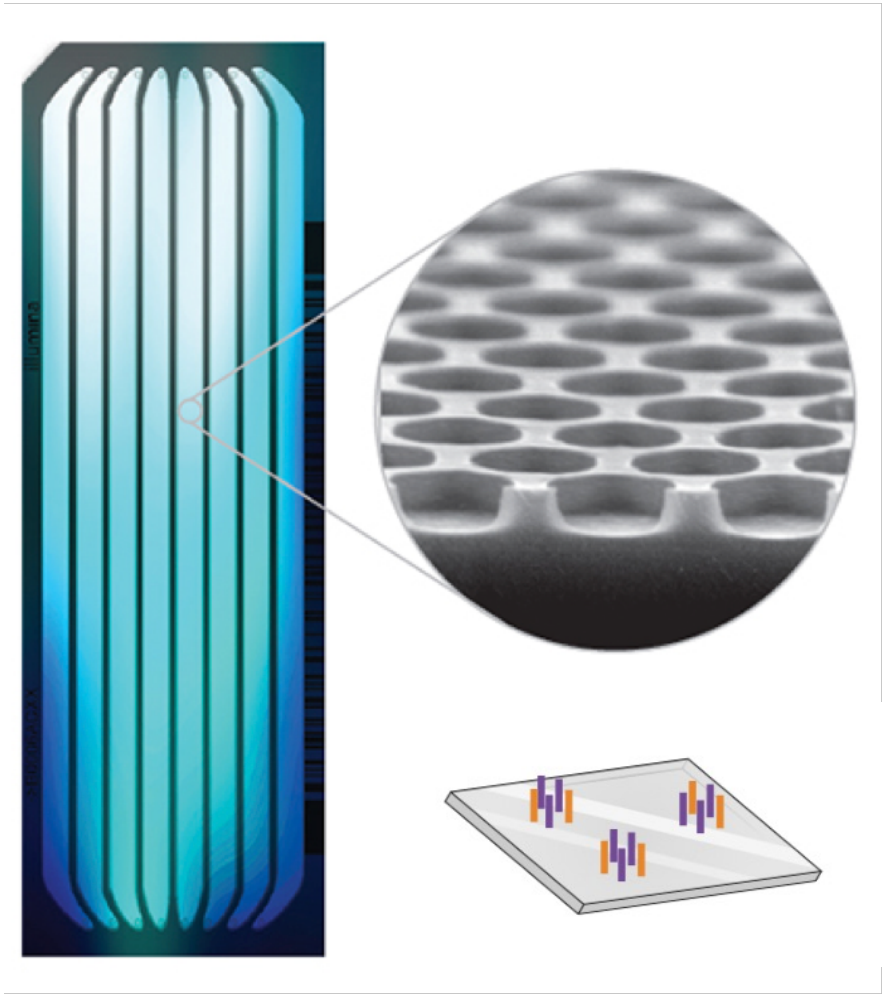
Library Preparation

Add sequencing adapters to DNA fragments

TruSeq-style

Adapter (20-40 bp)

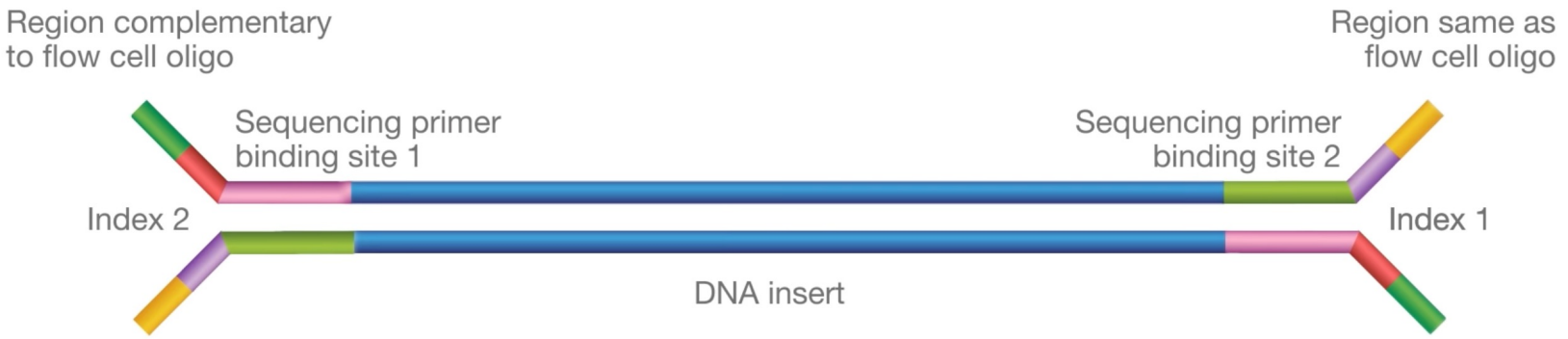


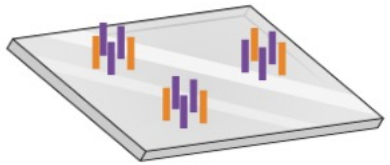


Flow cell

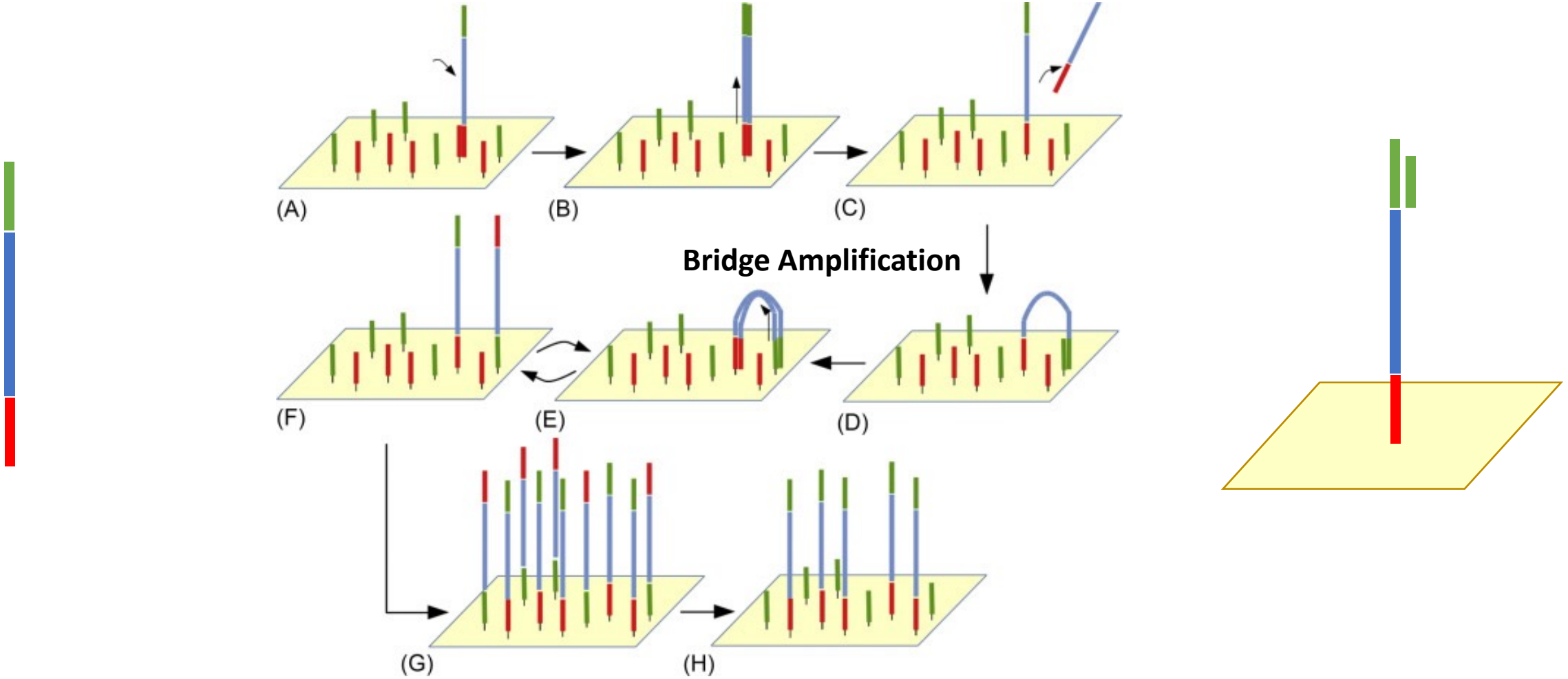
Library Preparation

Add sequencing adapters to DNA fragments



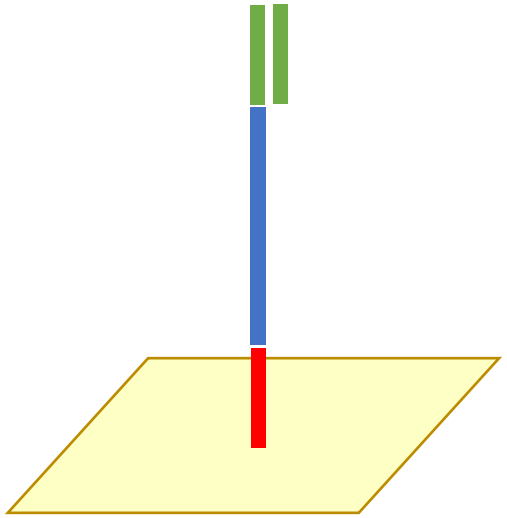


Cluster generation

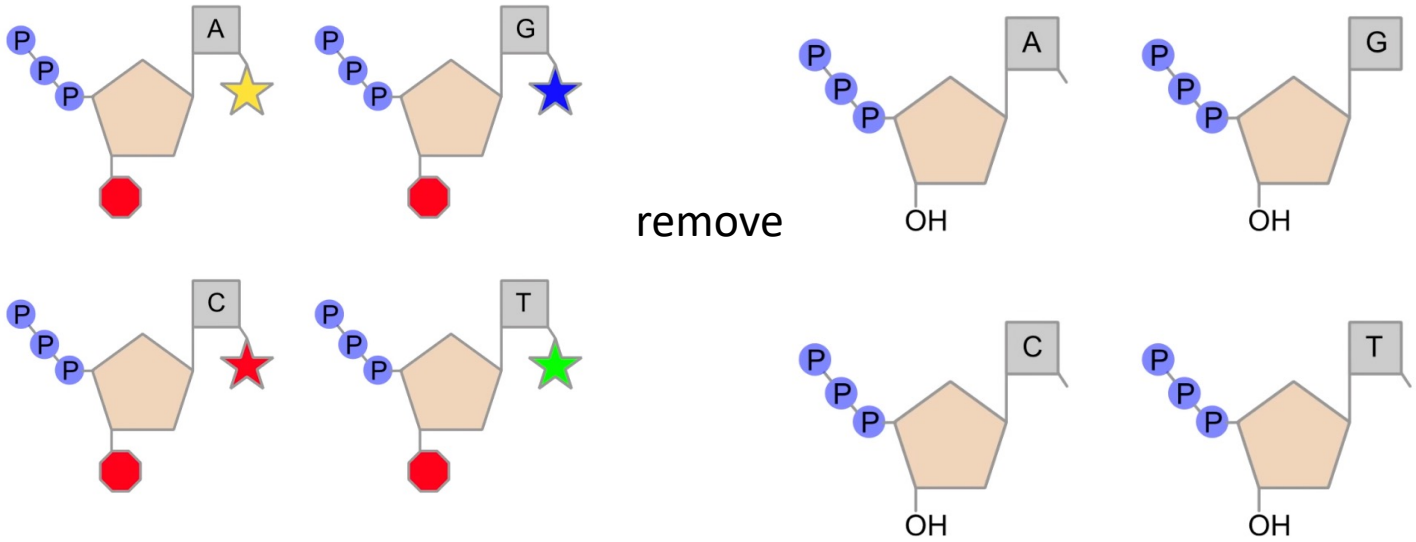


Rosario Michael Piro, *Chapter 7 - sequencing technologies for epigenetics: From basics to applications*, Epigenetics of the Immune System (Dieter Kabelitz and Jaydeep Bhat, eds.), Translational Epigenetics, vol. 16, Academic Press, 2020, pp. 161–183.

Sequencing



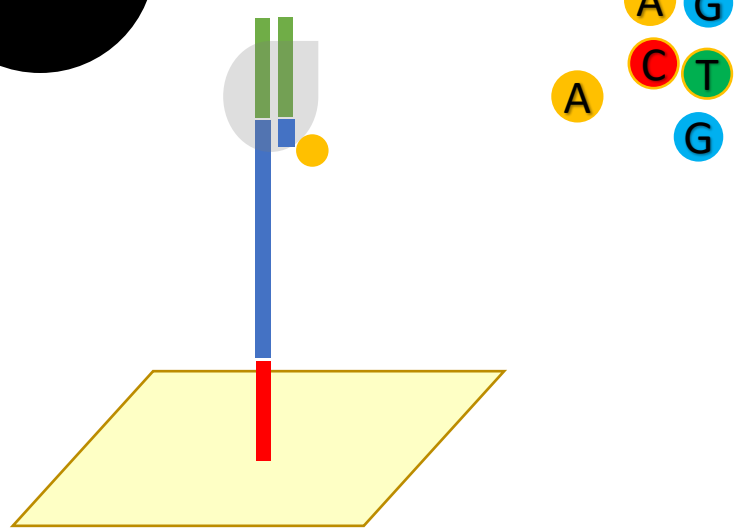
Chemistry for sequencing



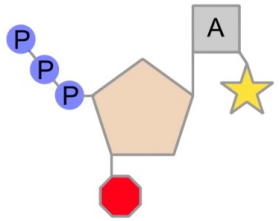
terminators

fluorescent

Sequencing



Chemistry for sequencing

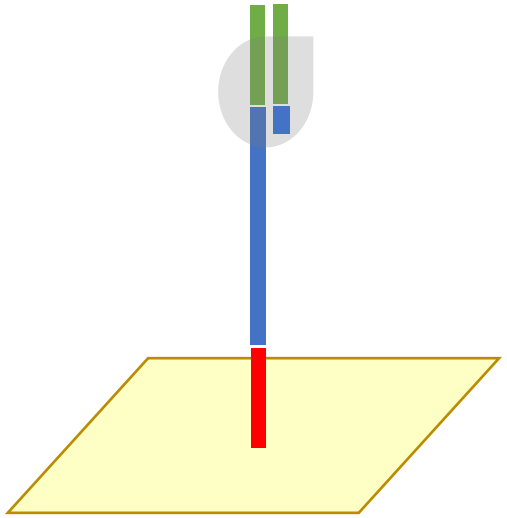


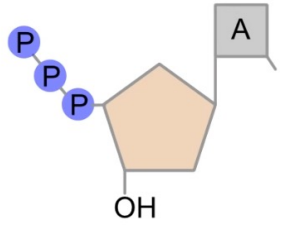
terminators

fluorescent

Sequencing

Chemistry for sequencing



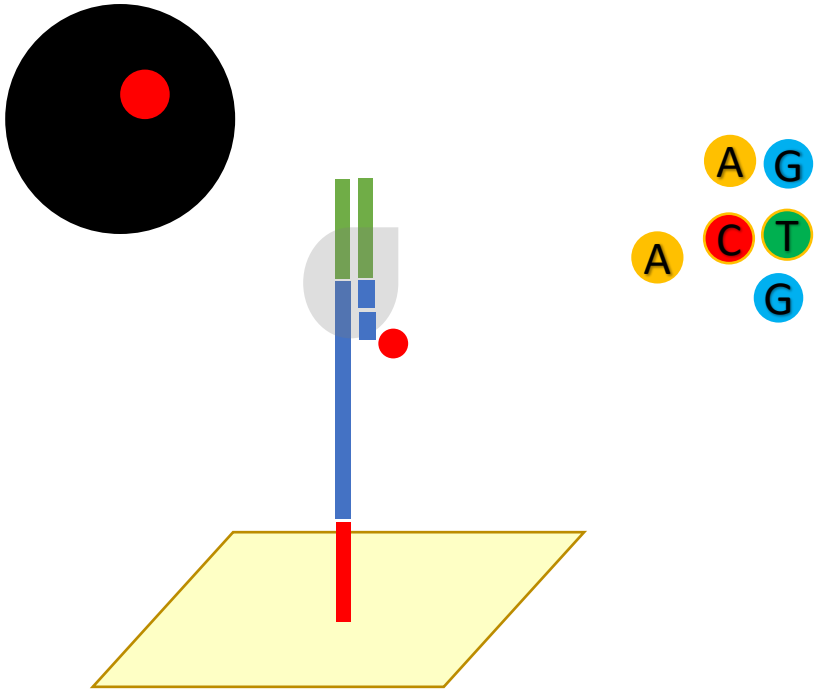


remove

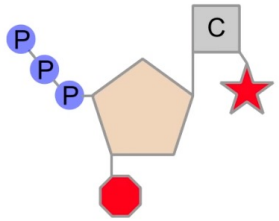
terminators

fluorescent

Sequencing



Chemistry for sequencing

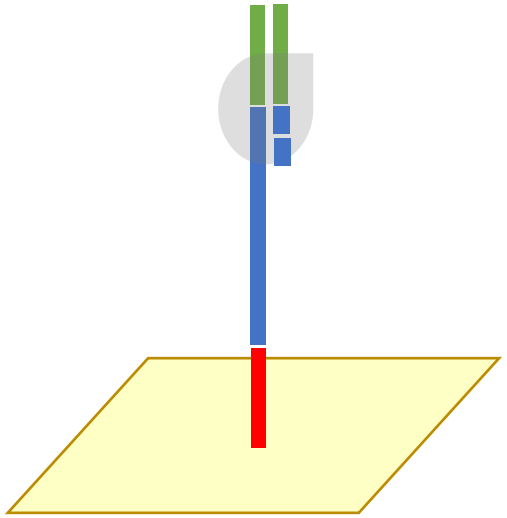


terminators

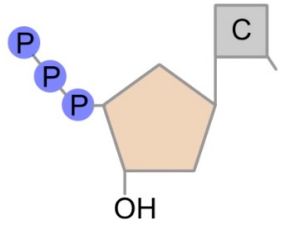
fluorescent

Sequencing

Chemistry for sequencing



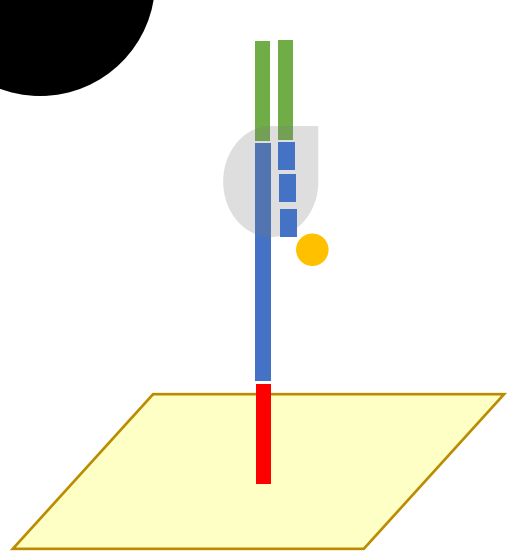
remove

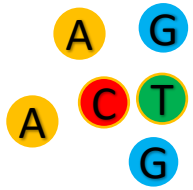


terminators

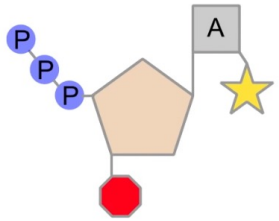
fluorescent

Sequencing





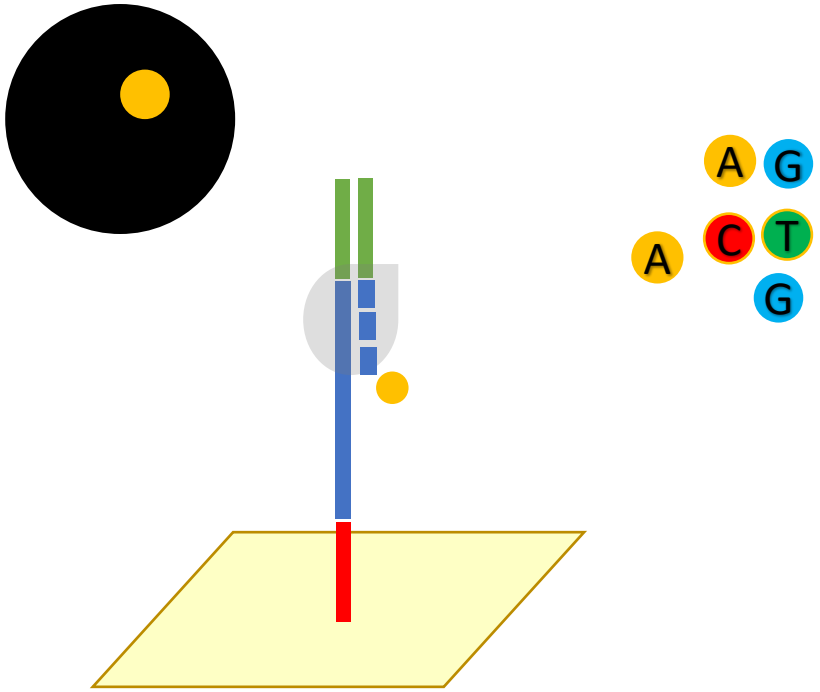
Chemistry for sequencing

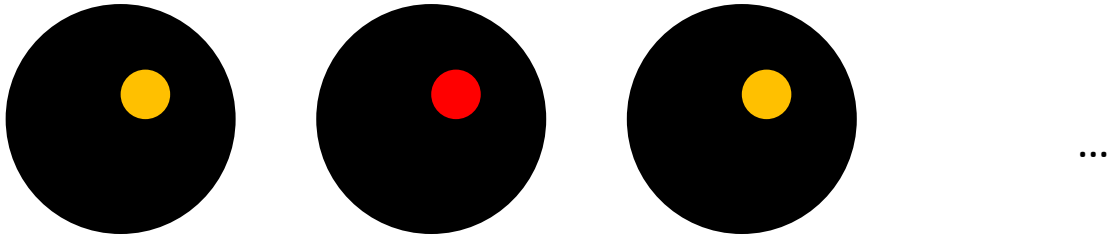


terminators

fluorescent

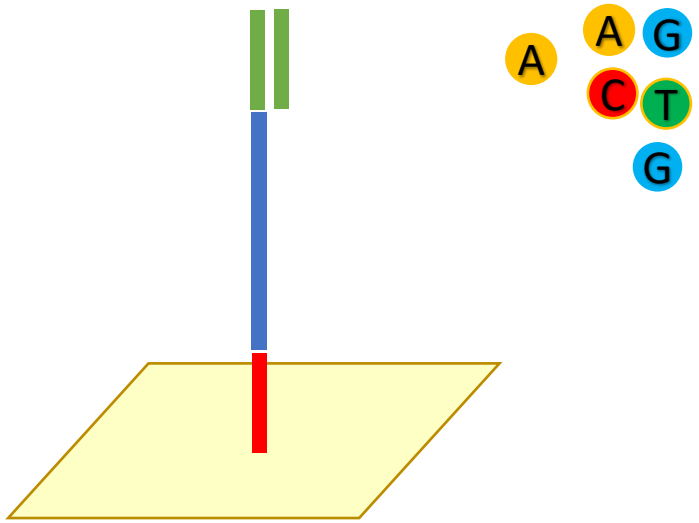
Sequencing

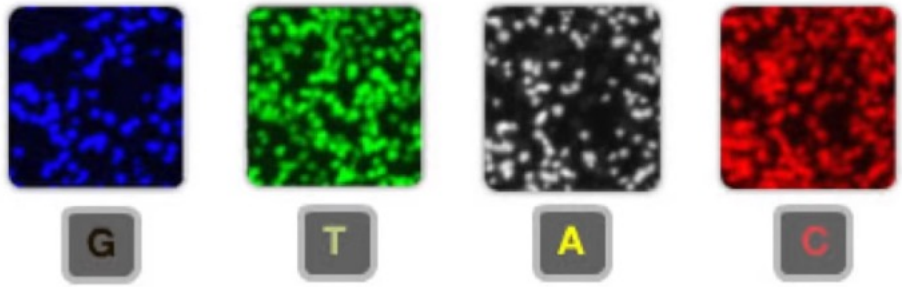




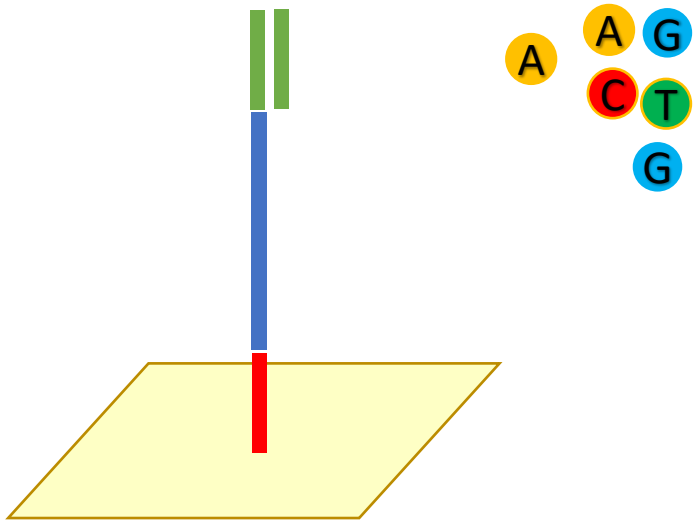
A C A …

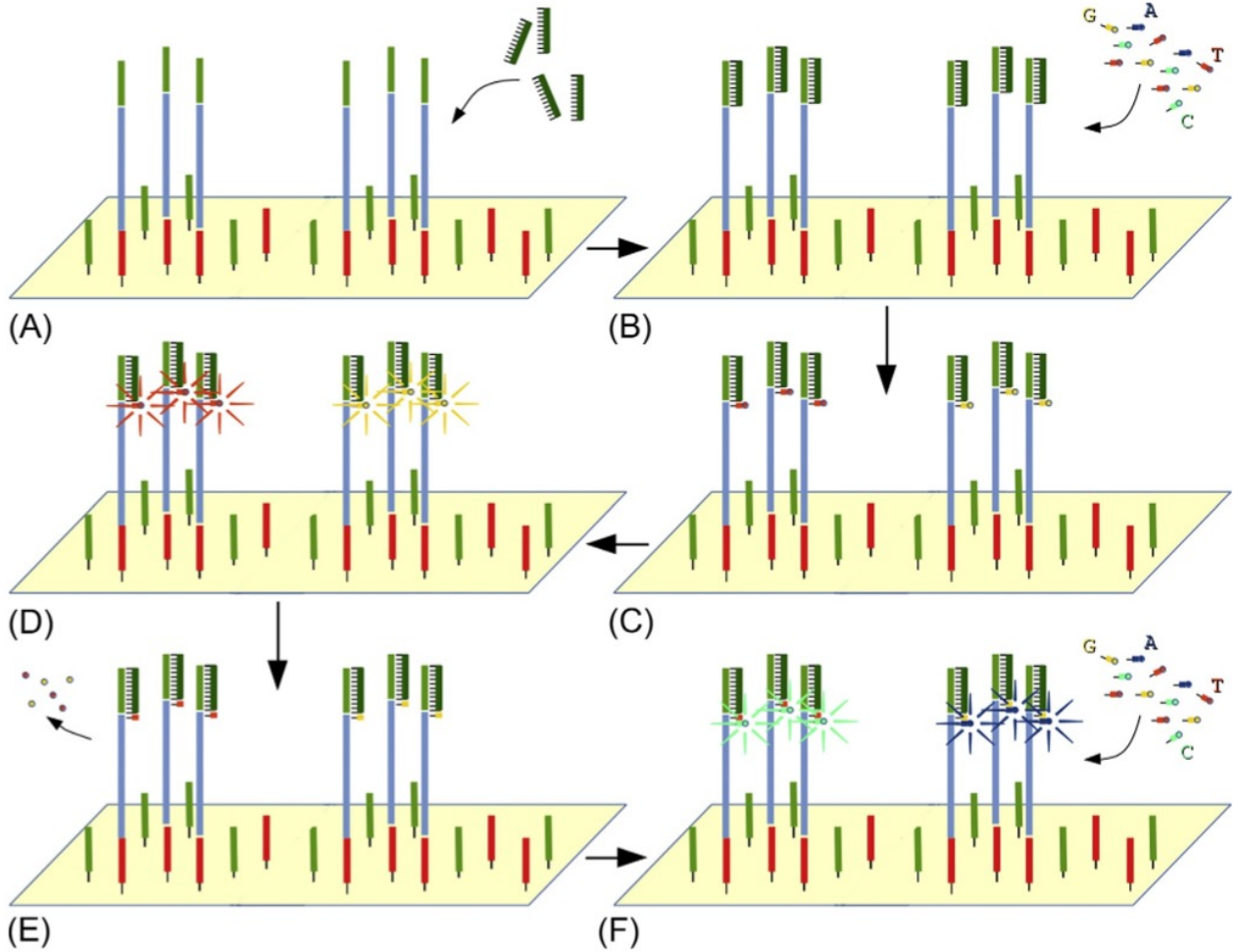
Sequencing





Sequencing





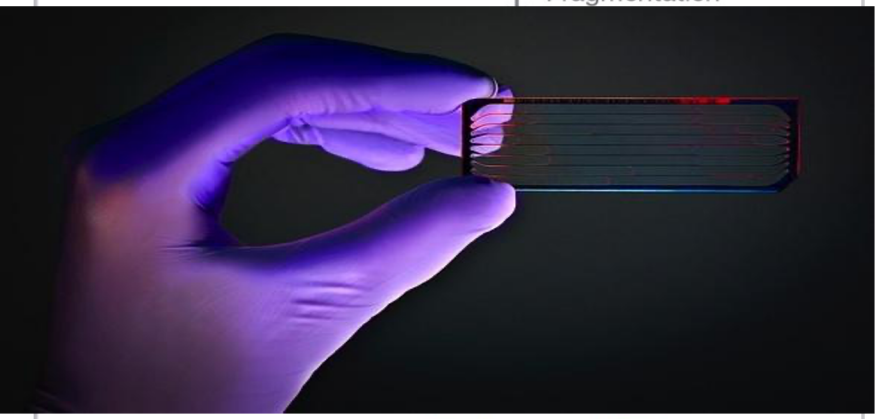
Rosario Michael Piro, *Chapter 7 - sequencing technologies for epigenetics: From basics to applications*, Epigenetics of the Immune System (Dieter Kabelitz and Jaydeep Bhat, eds.), Translational Epigenetics, vol. 16, Academic Press, 2020, pp. 161–183.

Illumina sequencing by synthesis

A. Library Preparation

Genomic DNA

Fragmentation



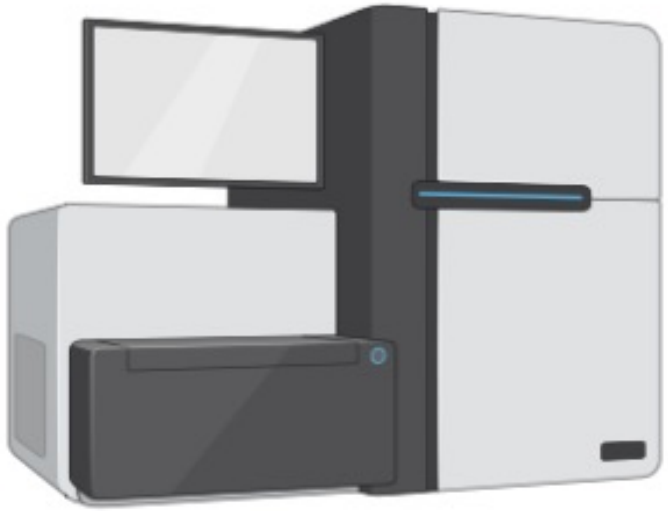
NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

B. Cluster Amplification

Brid

Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

C. Sequencing

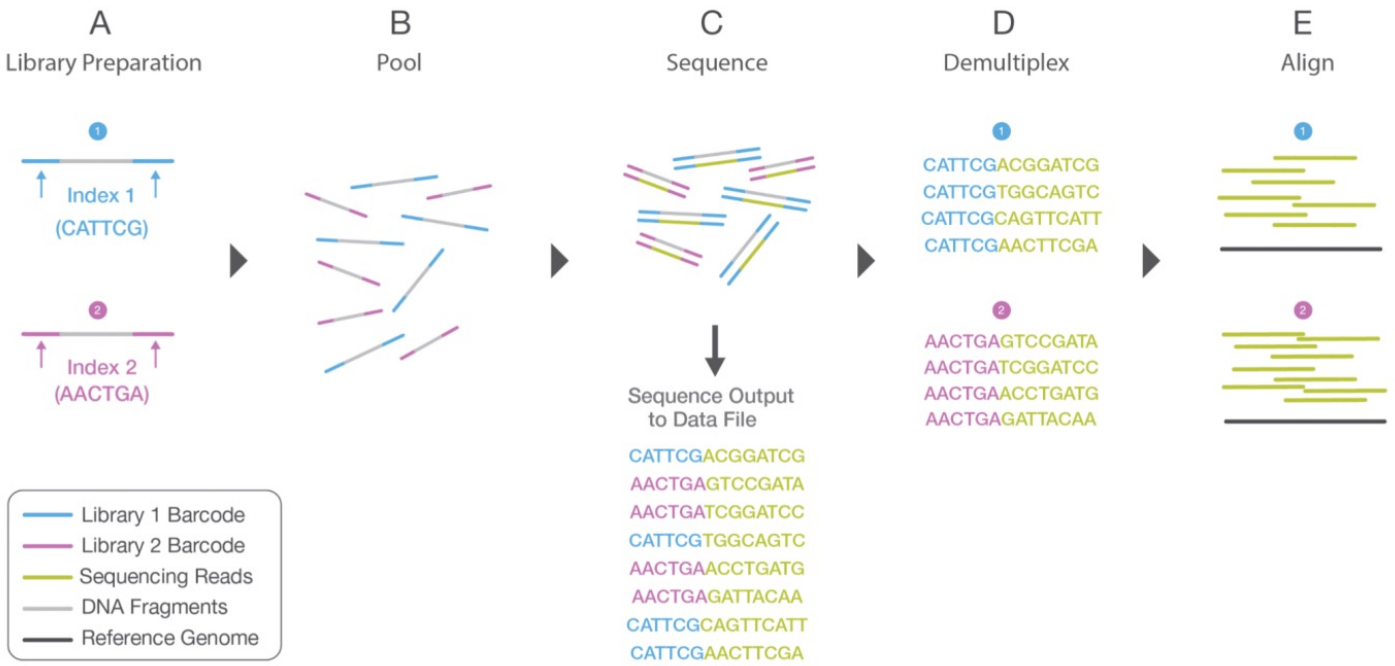


The emission wavelength and intensity are used to identify the base. This cycle is repeated “n” times to create a read length of “n” bases.

FASTQ

Multiplexing

- Sequences multiple samples at the same time
- Blocks against possible technical bias caused by differences between flow cell lanes
- Sequences small genomes or specific genomic regions.



Unaligned sequence: FASTQ

FASTQ header decoded (Illumina example):

Label `@K00359:71:HJJL7BBXX:3:1101:1996:1508 1:N:0:ATCACG`

Sequence `AAAATTCCAAGCTGGTTTCAACAGTACTTTGTTTCCAGAACAAAGAAATG`

Label `+`

Quality score `AAAFFJJJJJFJJ<J<FJJJJJJJJJJJJJJFJJFJJJJFFJFJJJJJJ<`

ASCII code

Worst quality

Best quality

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

Alignment

Trimmed DNA sequences

?

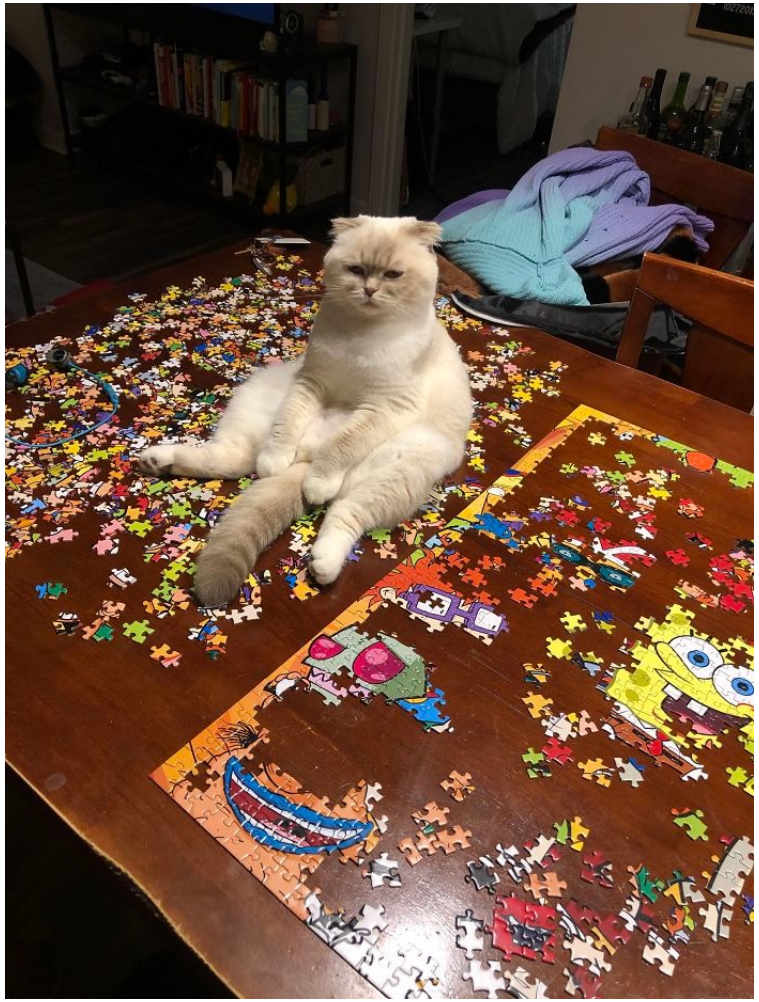
GRCh38

Human genome

Alignment

?

GRCh38



10 min break!