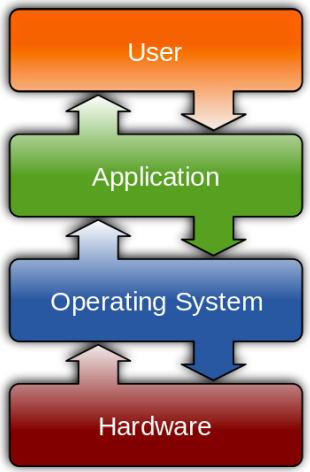
Next-gen Sequencing Technologies

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

2/4/2021

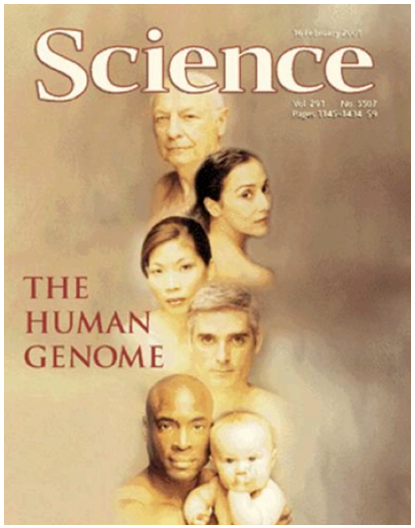
Unix commands

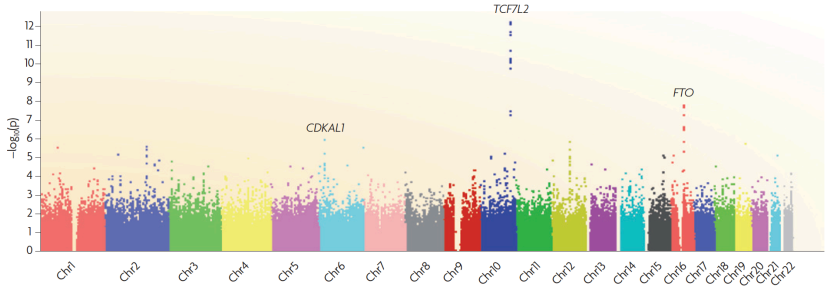


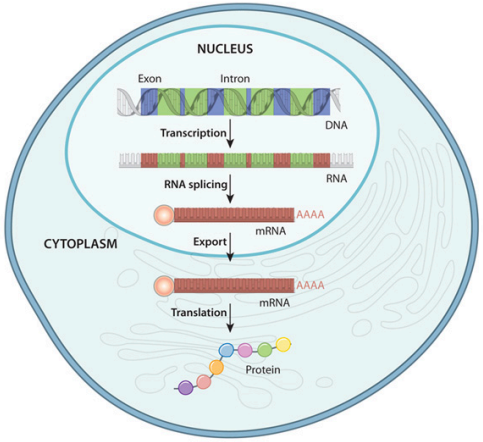
- **cd** - change the working directory
- **mkdir** - make directories
- **pwd** - print name of current working directory
- **ls** – list directory contents
- **chmod** - change the access permissions to files and directories

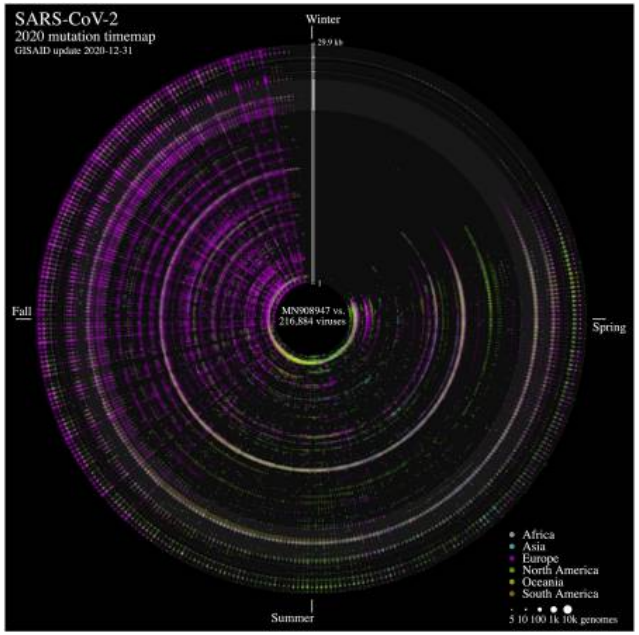
- **head** - output the first part of files
- **tail** - output the last part of files
- **more** and **less** display contents of large files page by page or scroll line by line up and down
- **cat** - concatenate files
- **paste** - merge lines of files
- **wc** - print line, word, and bytes for each file
- **grep** - print lines matching a pattern

The sequencing technology is key for a wide range of biological researches

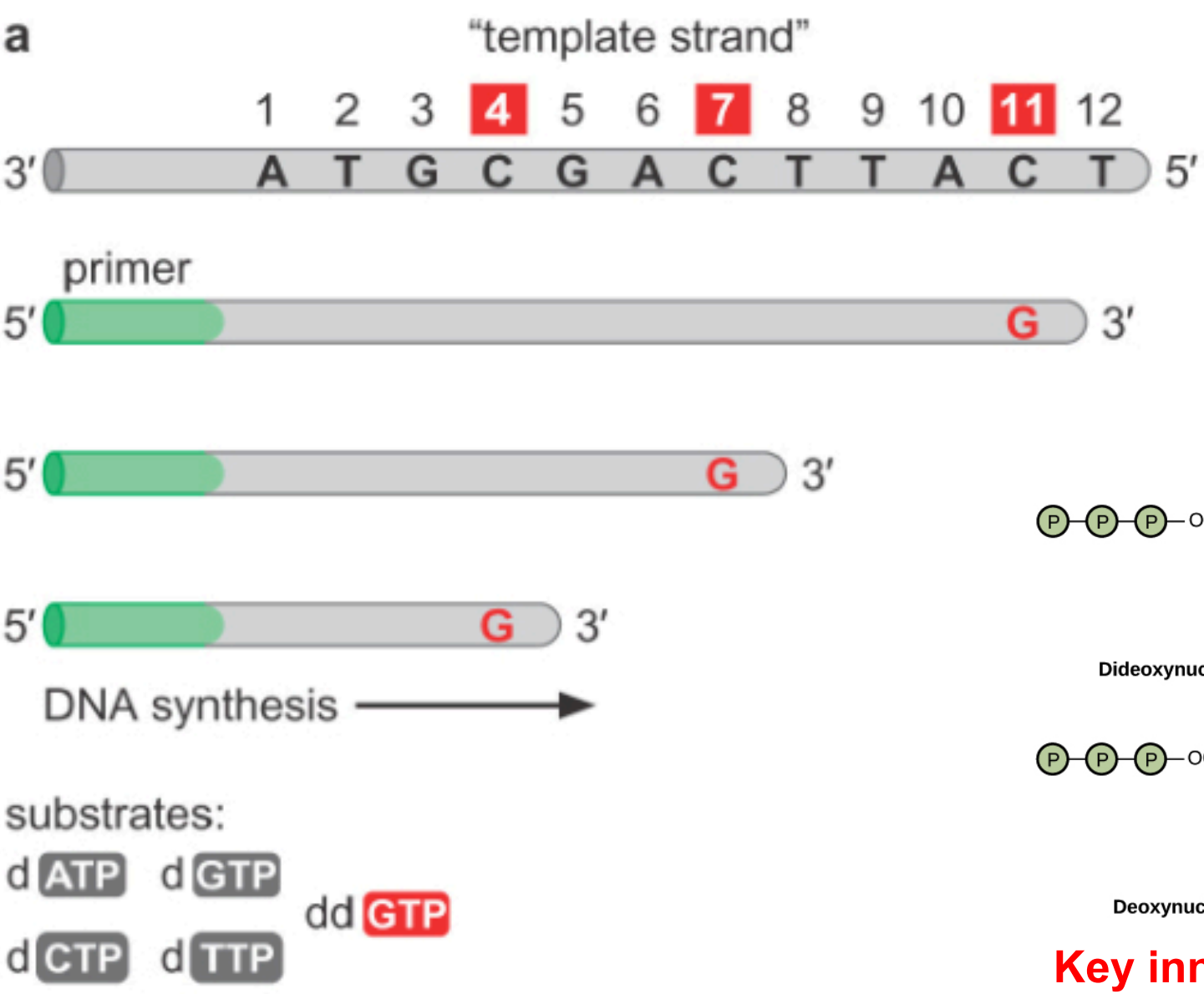


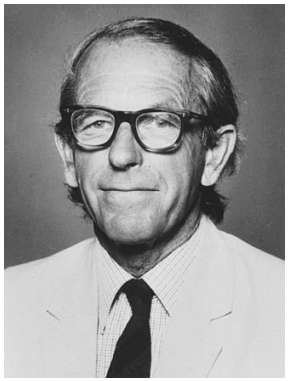




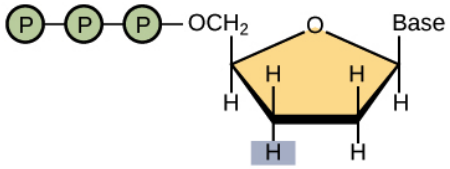


Sanger sequencing technology - I

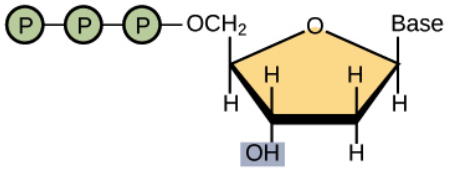




Frederick Sanger



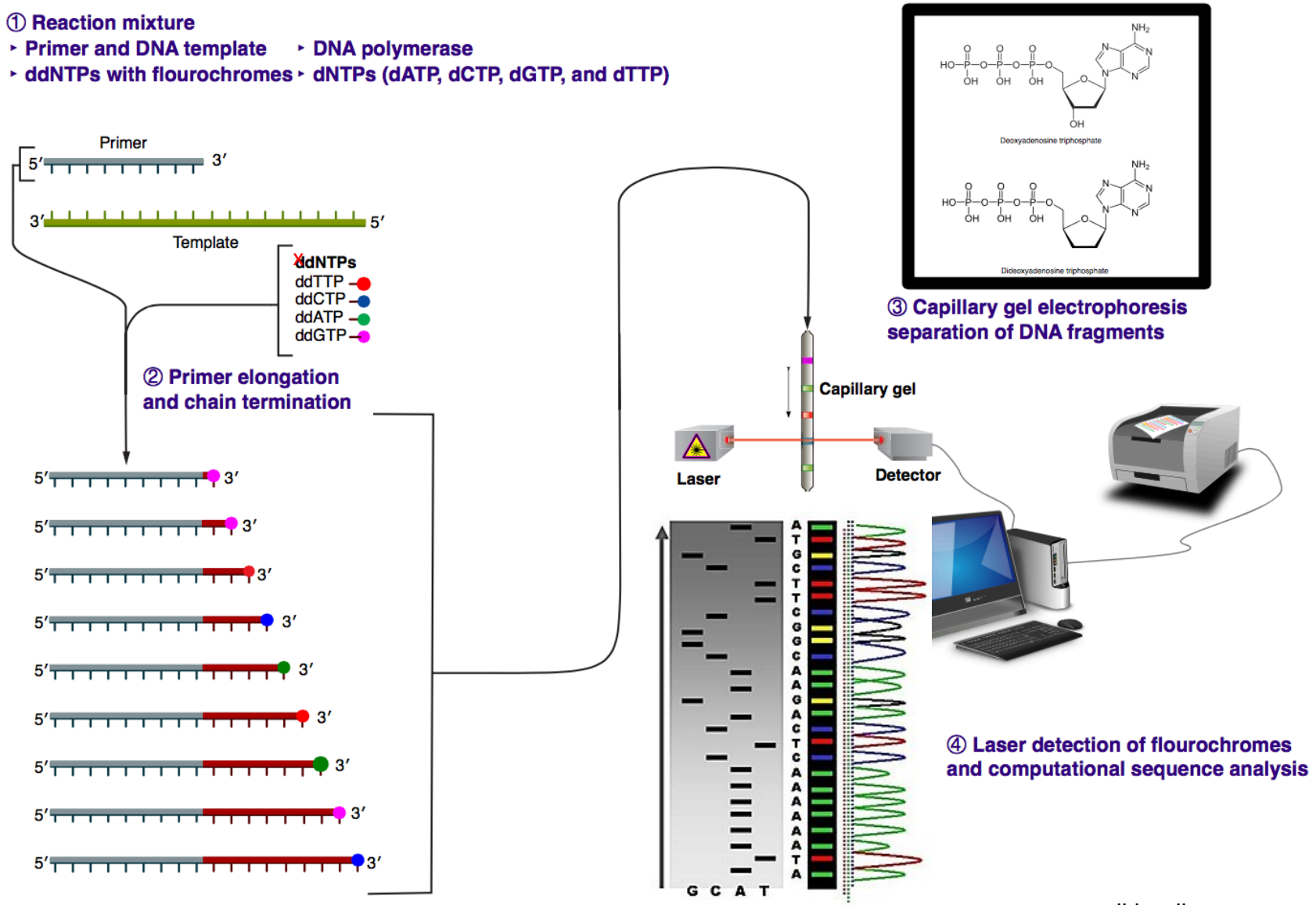
Dideoxynucleotide (ddNTP)



Deoxynucleotide (dNTP)

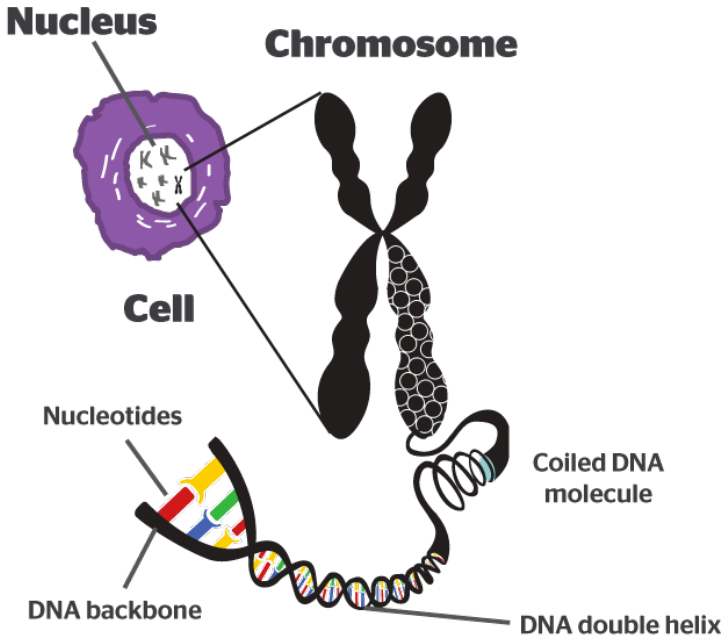
Key innovation

Sanger sequencing technology - II



wikipedia

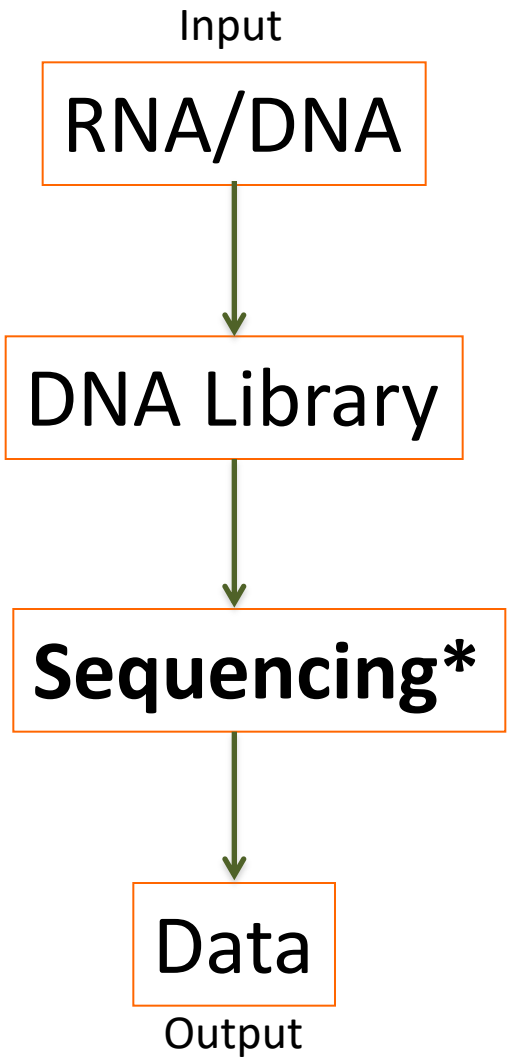
Major Next-gen sequencing (NGS) technologies



sequencing sensitivity and read length

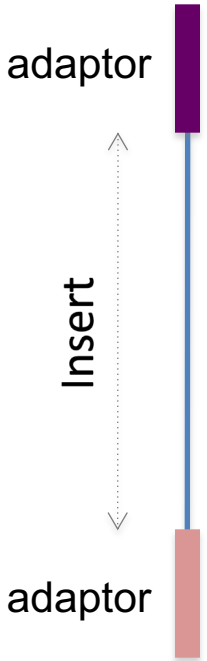
Before single molecular & "super long" sequencing technologies, **fragmentation** and **amplification/cloning** of a single nucleotide molecule are needed for sequencing.

COMMON in all NGS platforms



The adaptor is required for library preparation

Hundreds to thousands of millions of fragments are sequenced ***in parallel***

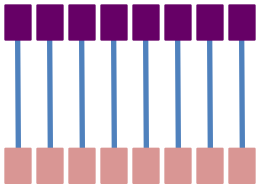


Single-molecule and ***amplification-based*** approaches

Nucleotide detector:
VERY sensitive

Directly read sequence
single-molecule

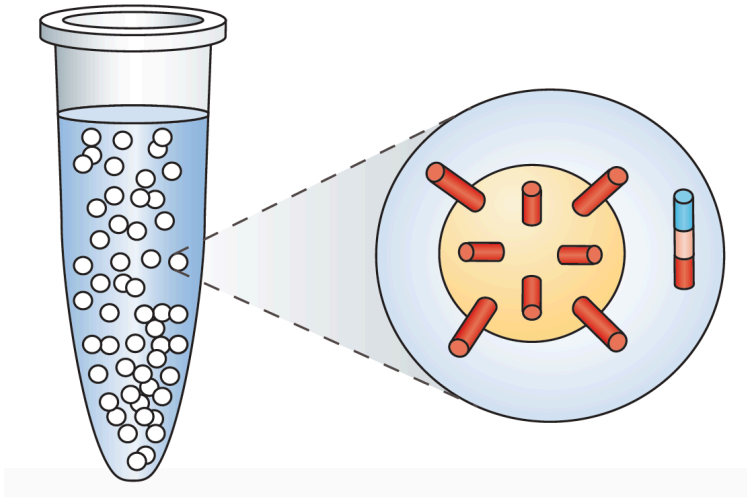
Nucleotide detector:
Not sensitive at the
single molecular level



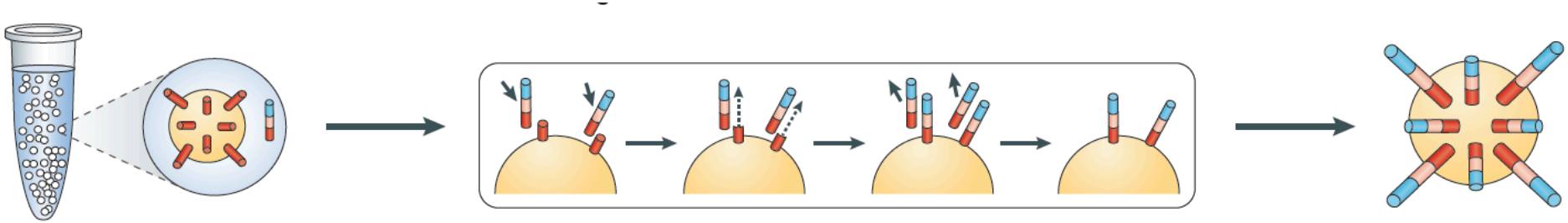
amplify and then
read sequence

"Having many thousands of identical copies of a DNA fragment ***in a defined area*** ensures that the signal can be distinguished from background noise."

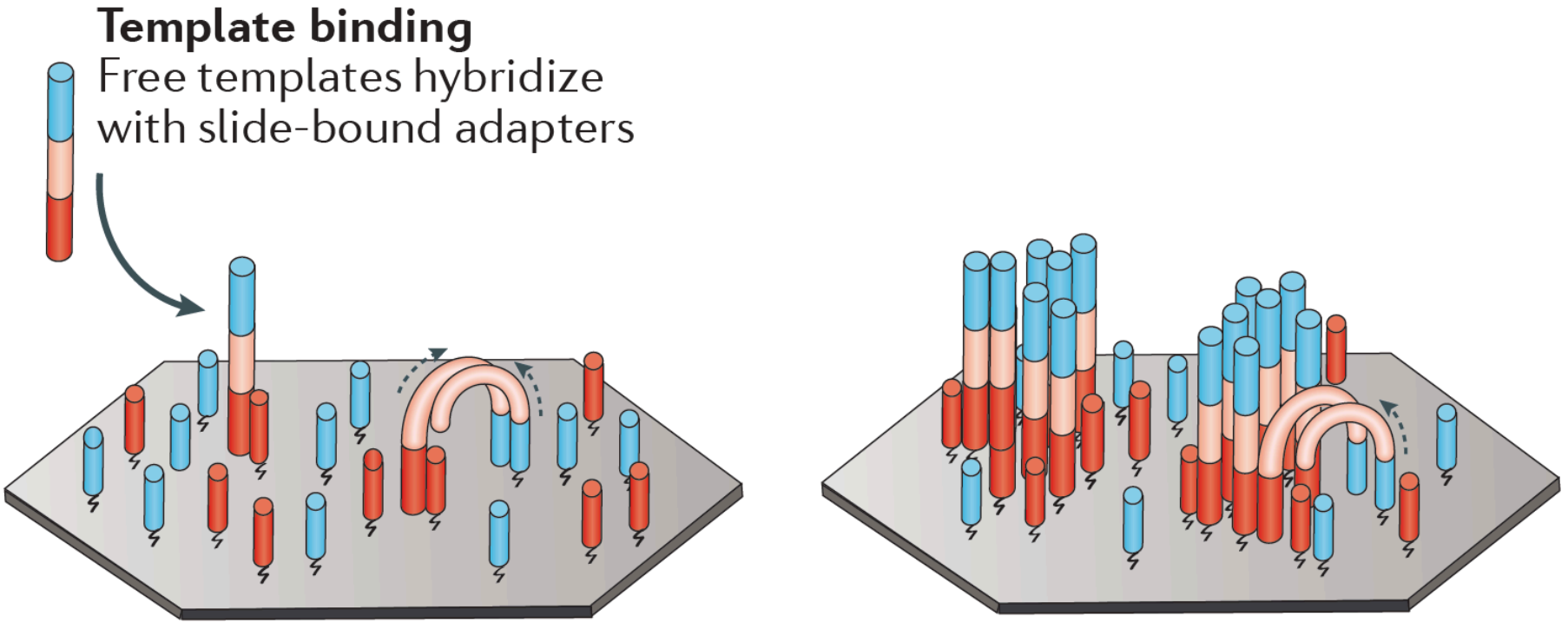
Massive independent amplifications – emulsion PCR



Water-in-oil emulsion PCR
(454 and Ion Torrent)

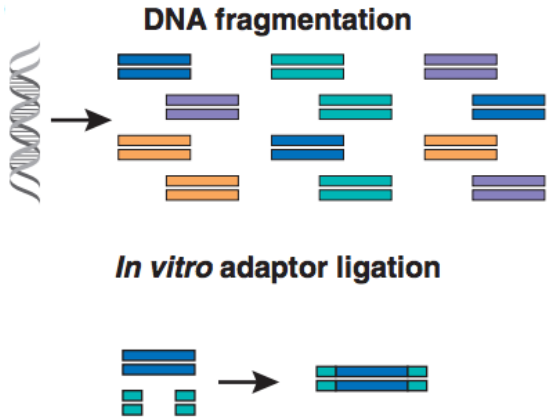


Massive independent amplifications – bridge PCR

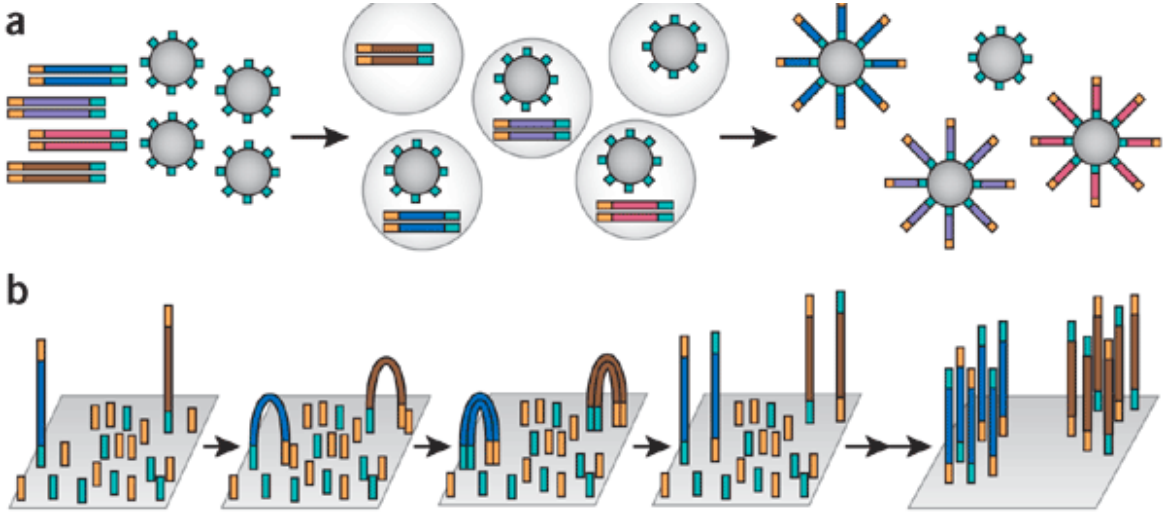


DNA amplification

Water-in-oil emulsion PCR
(454 and Ion Torrent)



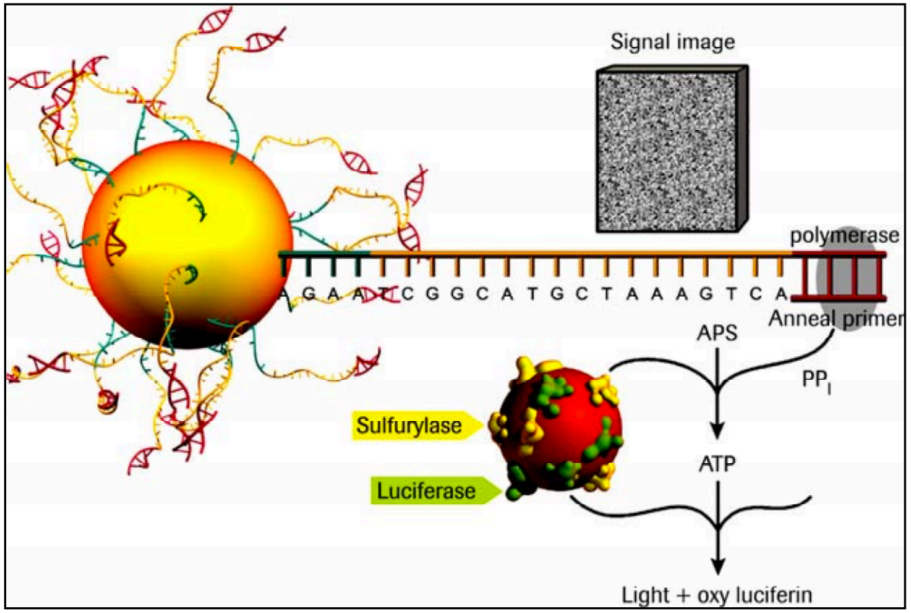
Adaptor ligation



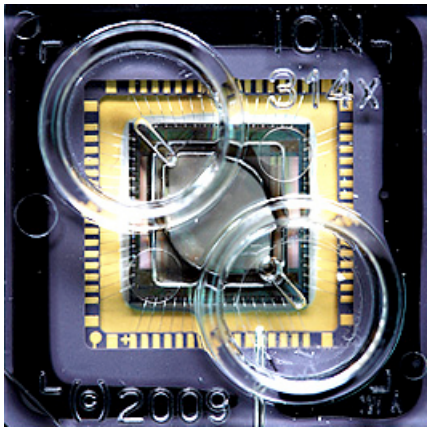
Bridge PCR on slides
(Illumina)

Nature Biotechnology, 2008, 26: 1135-45

454 and Ion Torrent signal detectors



454 technology, Nature 2005, 437: 376-380



Ion Torrent video

1. Sequencing by synthesis
2. Pyrosequencing (454)

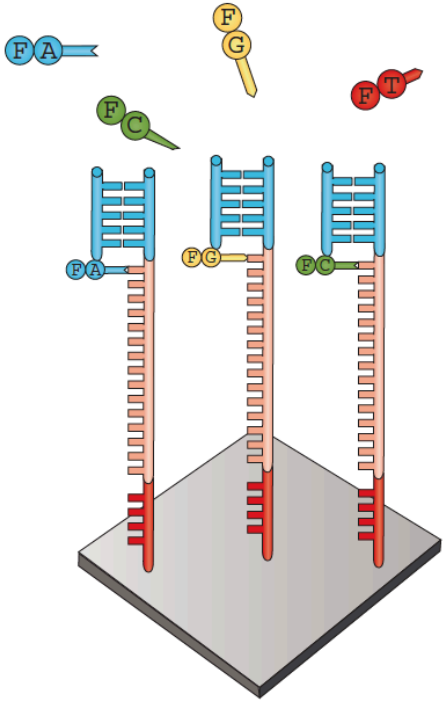
1. Ion Torrent technology is similar to 454 technology

2. The signal is H^+ rather than pyrophosphate

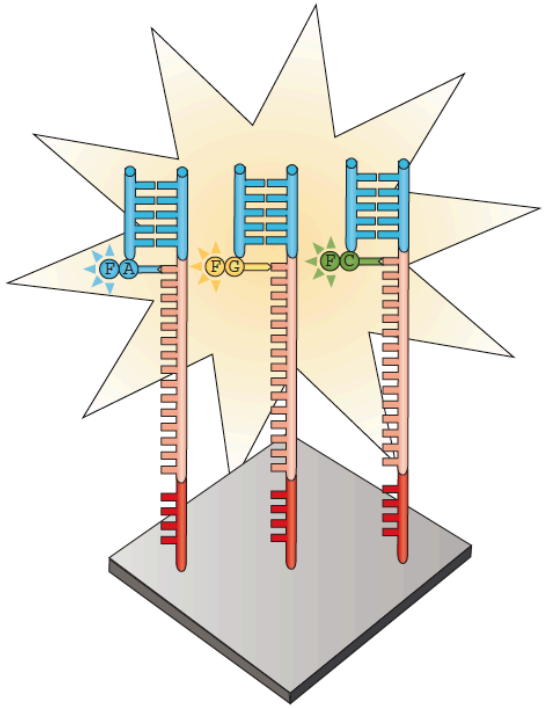
Illumina

Two key technologies:

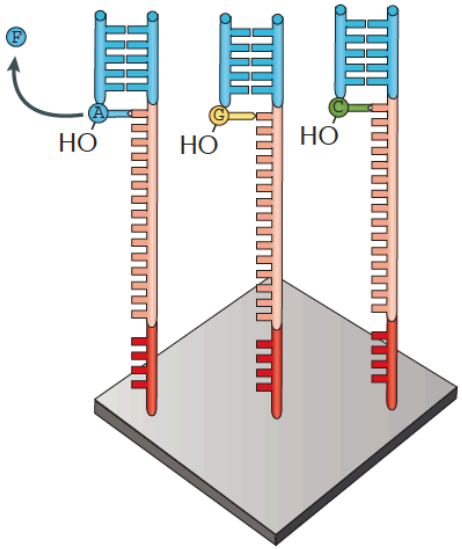
1. Bridge PCR
2. Reversible terminator chemistry



Nucleotide addition

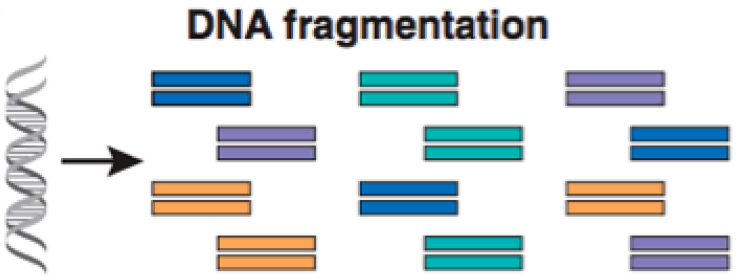


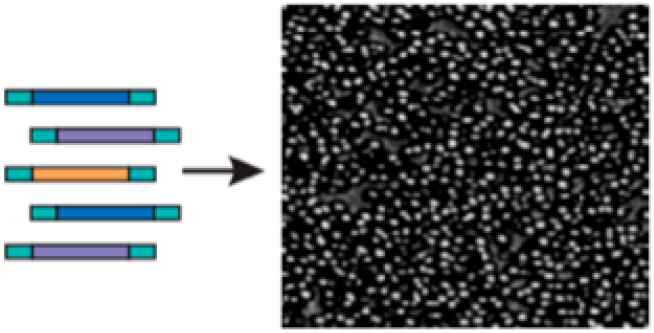
Imaging

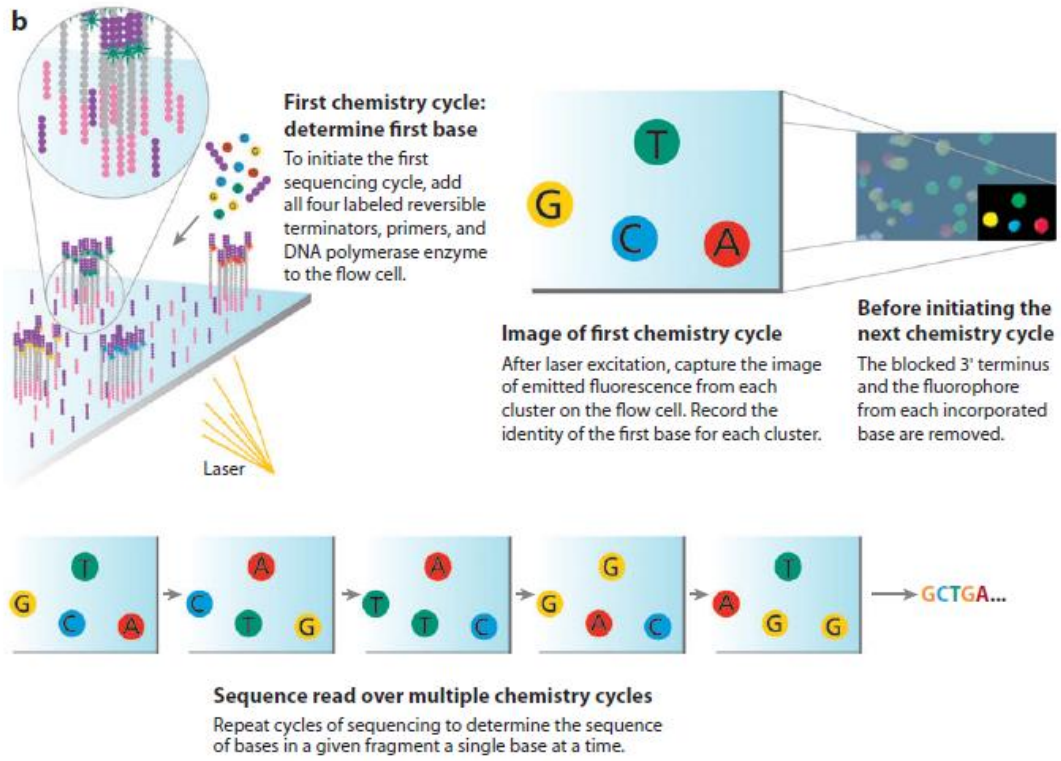


Cleavage

Illumina sequencing







Two key technologies:

1. Bridge PCR
2. Reversible terminator chemistry

Illumina Sequencers

Illumina *versus* Ion Torrent & 454

Illumina
Record signal per **nucleotide position**:

A T G C A A A A

A T G C A A A A

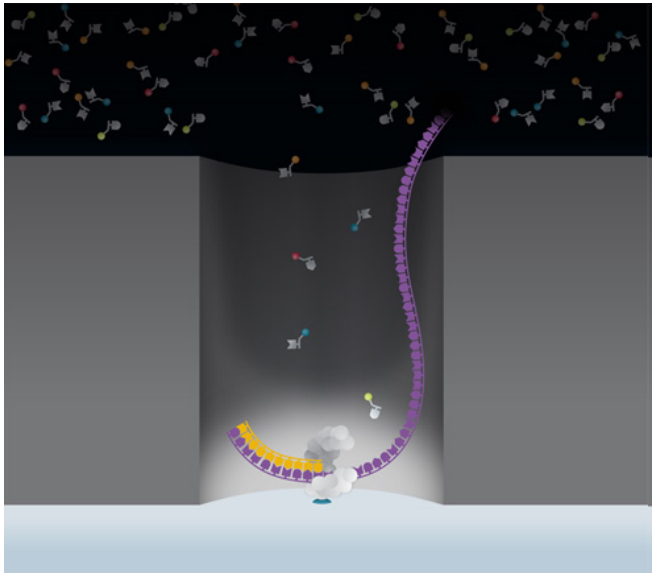
Life technology Ion Torrent & Roche 454
Record signal per **nucleotide type**:

A T G C A A A A

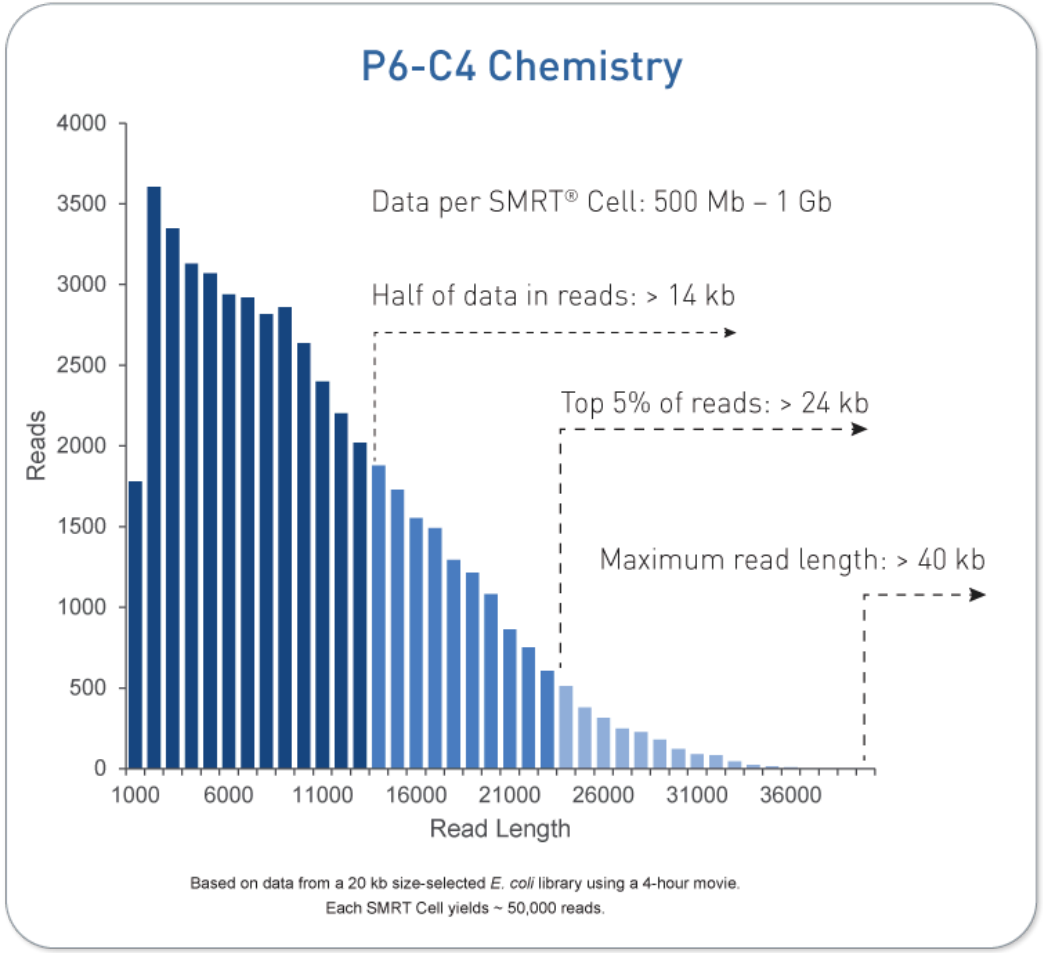
Sequencing errors at homopolymers

When the single molecular sequencing technology is ready, **amplification or cloning** is not necessary.

PacBio – Single Molecule Real Time (SMRT)



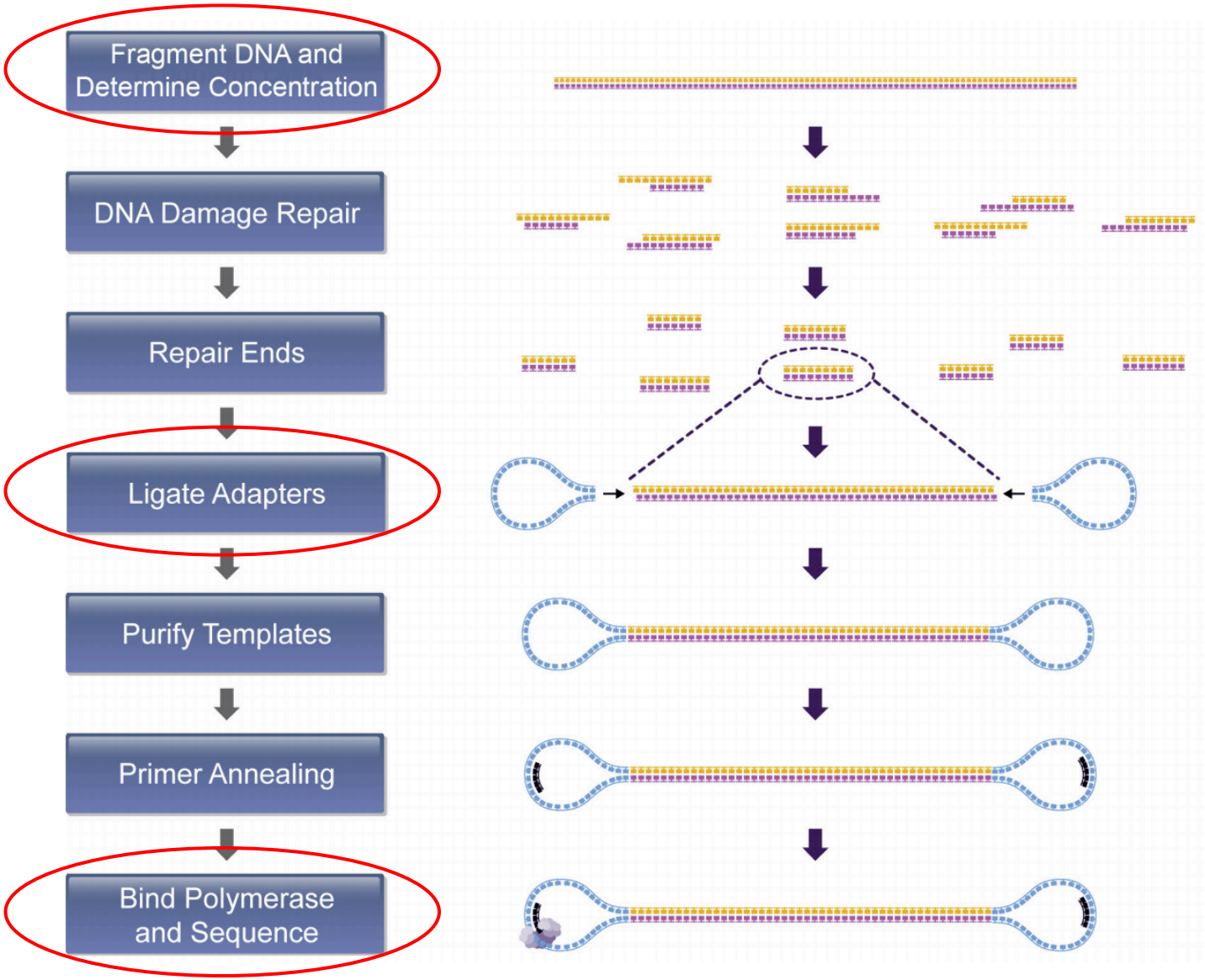
PacBio tech video



- Single molecule sequencing
- no amplifications required
- up to 70+ kbp sequencing
- Moderate sequencing throughput
- high sequencing error rate (5-10%, random, no-context-specific errors)

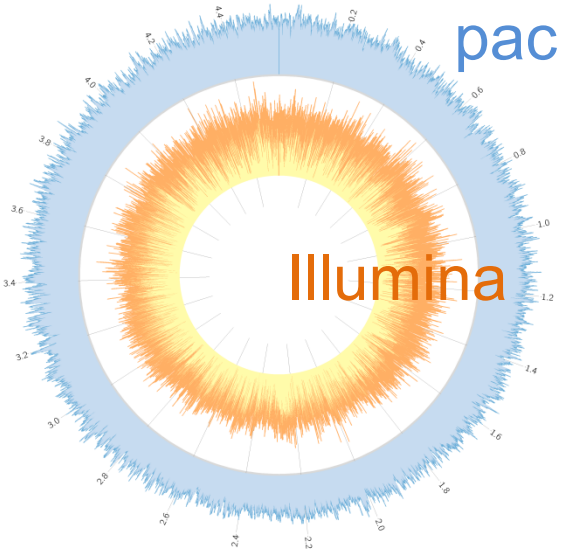
pacbio.com

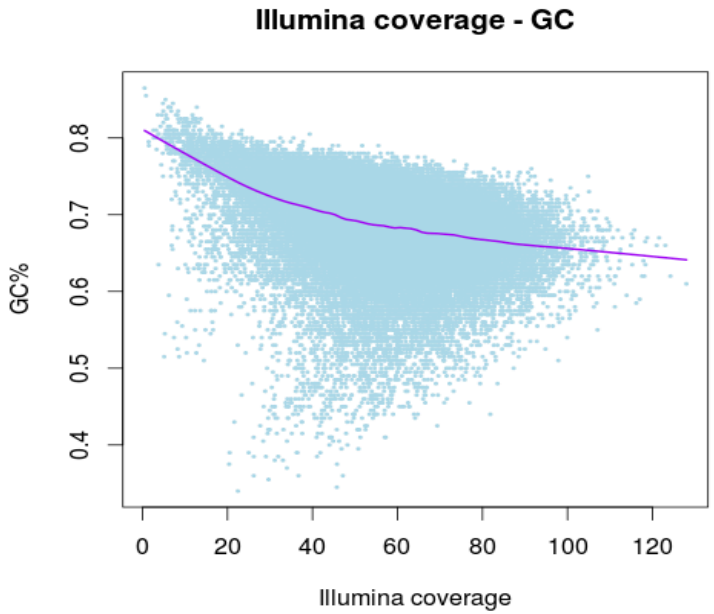
PacBio library prep workflow

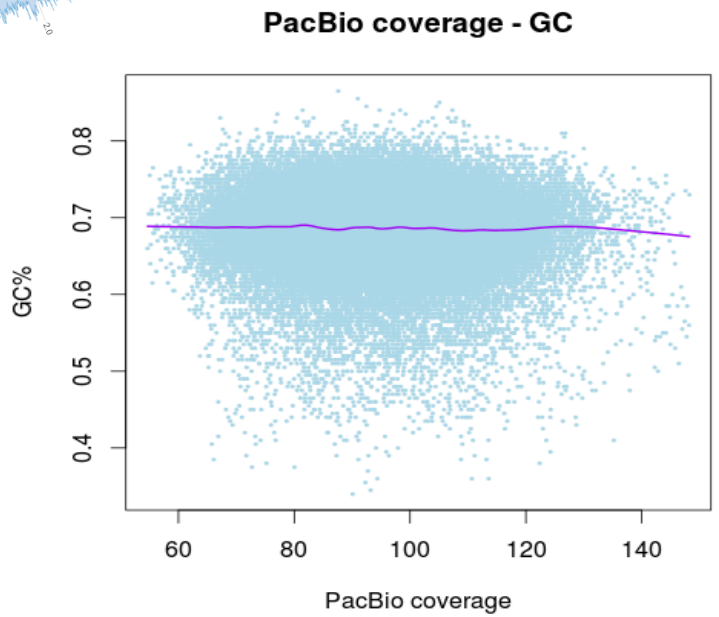


Less biases (e.g., GC)

Coverage map
(depth per region)

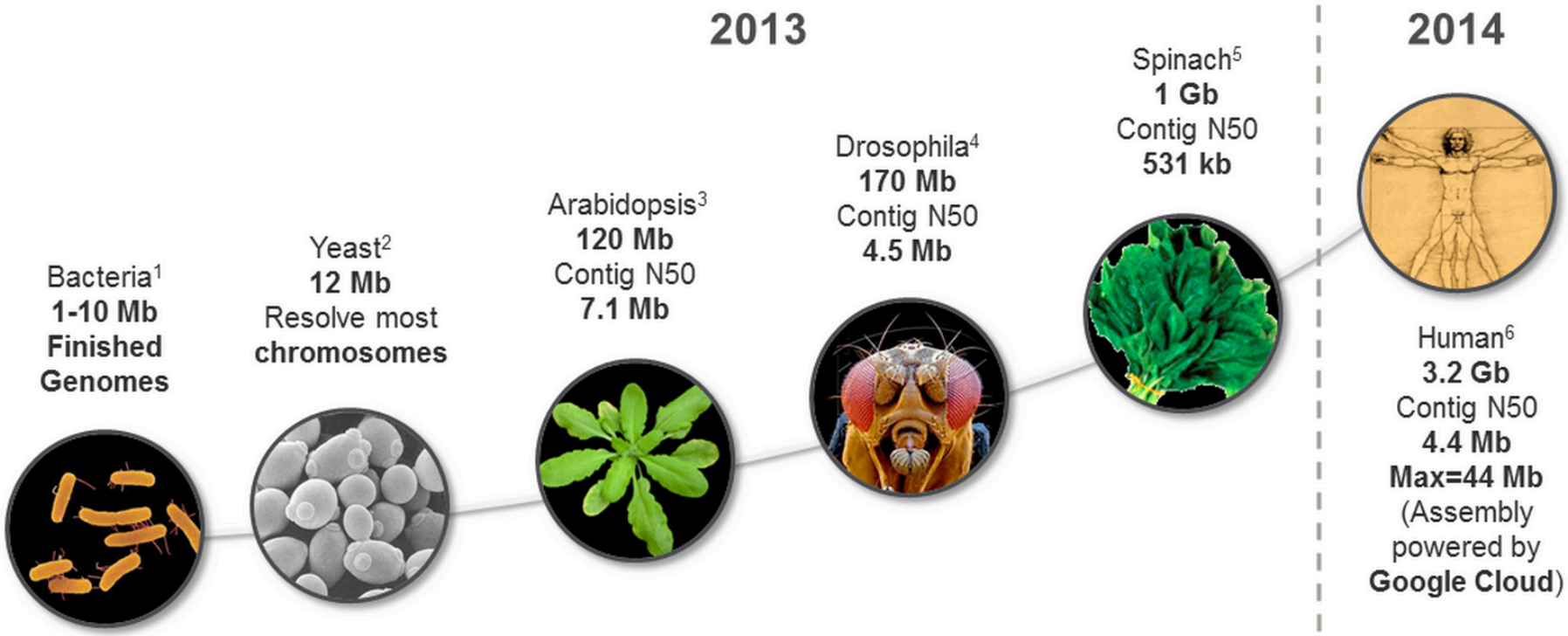






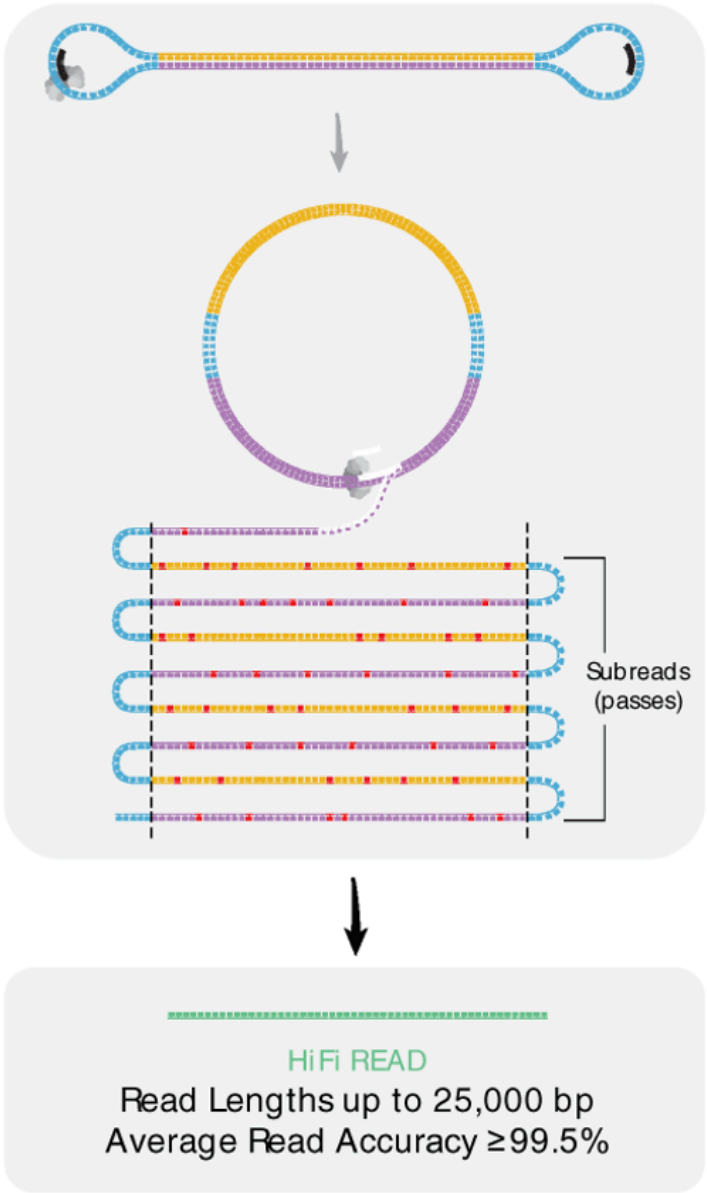
Peng *et al*., BMC Genomics, 2016, 17:21

PacBio for genome assembly



PacBio has solved *de novo* assemblies of most bacterial genomes and it will solve assemblies of small “simple” genomes (e.g., <500 Mbp) with increasing read length and improved sequencing quality.

HiFi PacBio data

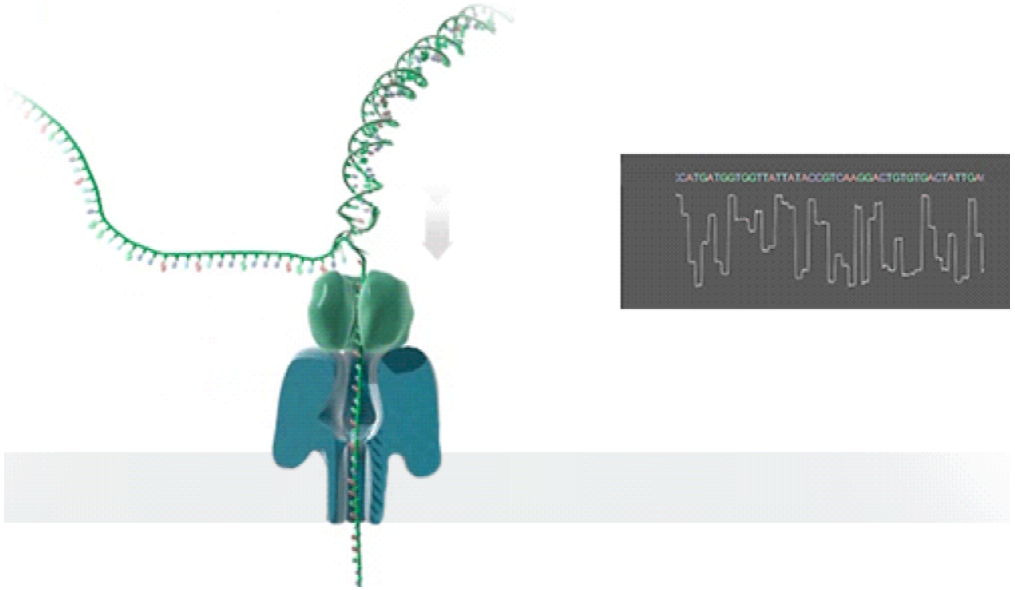


Up to 25 kb with
~99% accuracy

Scientific Data, 2020, 7:399

Oxford Nanopore

A promising technology



As each nucleobase passes through the pore the current is affected and this change allows sequence to be read out.

- Single molecular sequencing
- No amplifications
- **Long reads (typically 10-200kb)**
- **Error rate is high (~5-15%)**

Nanopore devices

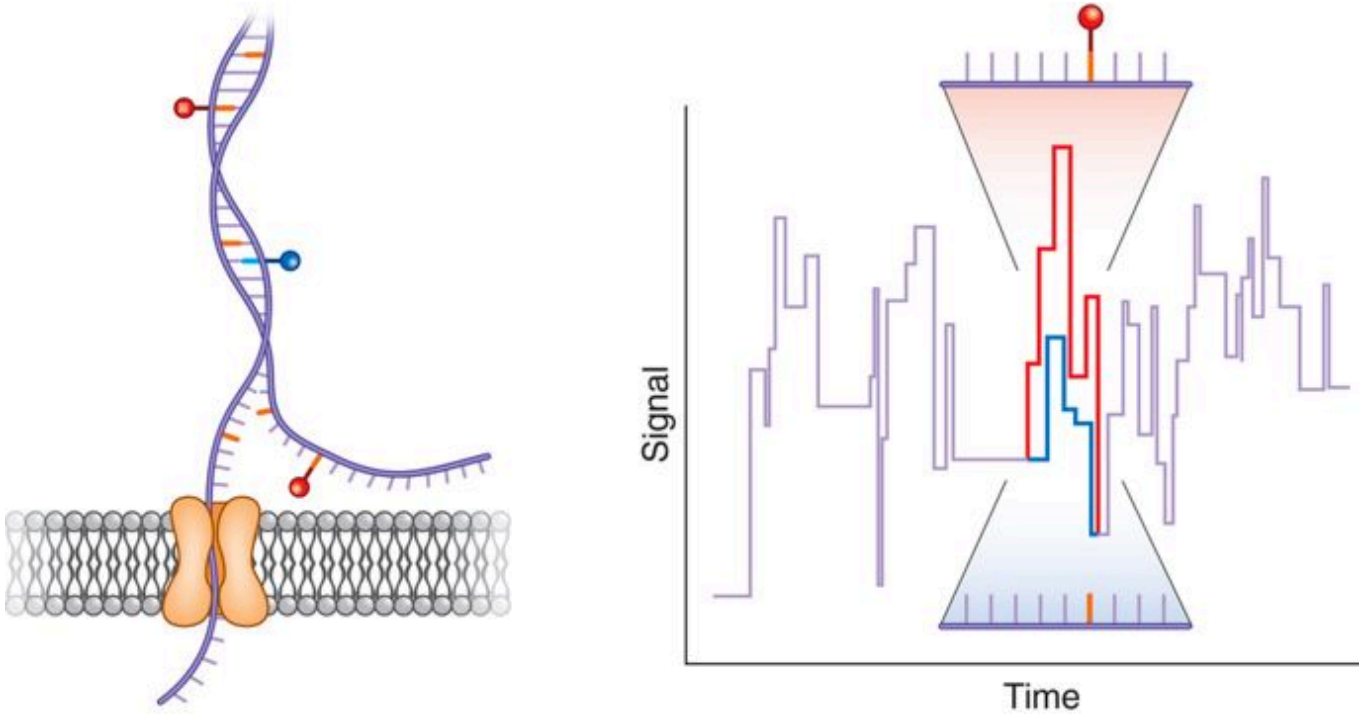
MinION
1. USB disposable sequencer
2. ~10Gb in about two days

PromethION
1. High-throughput
2. lower cost (<$1000 per human genome)

portable device:
MinION Mk1C

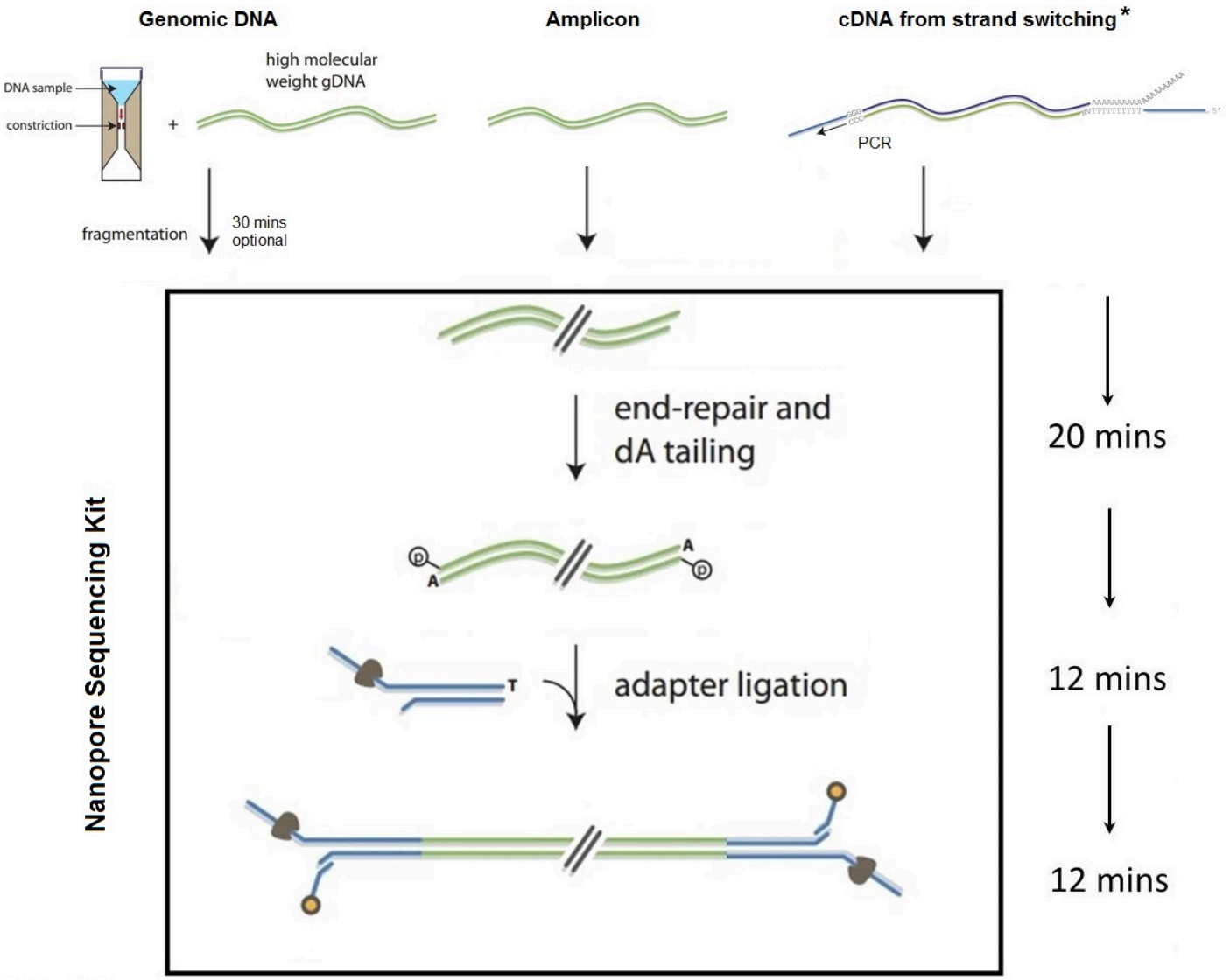
Applications of Nanopore sequencing

1. Genomic DNA sequencing
2. RNA sequencing (direct RNA or cDNA)
3. DNA methylation and other modifications



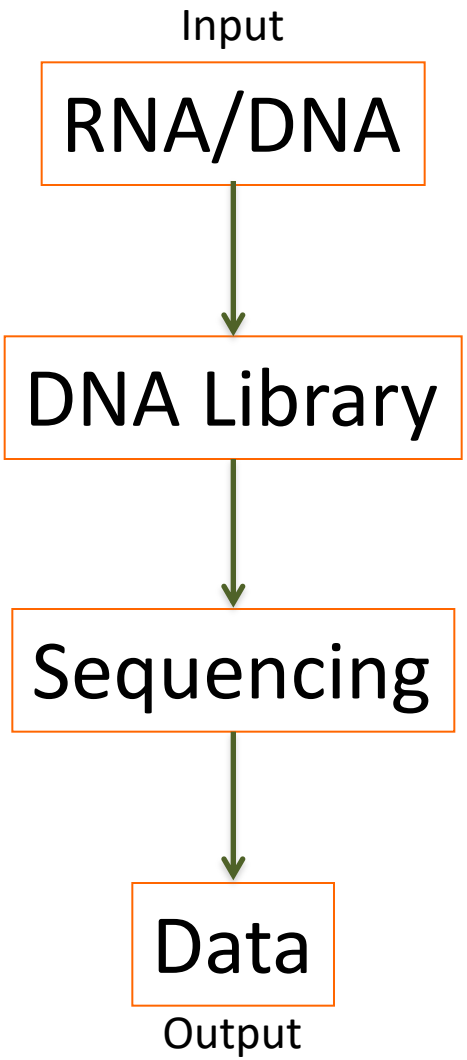
Nature Methods, 2017, 14:347–348

Nanopore library preparation



nanopore.com

COMMON in all NGS platforms

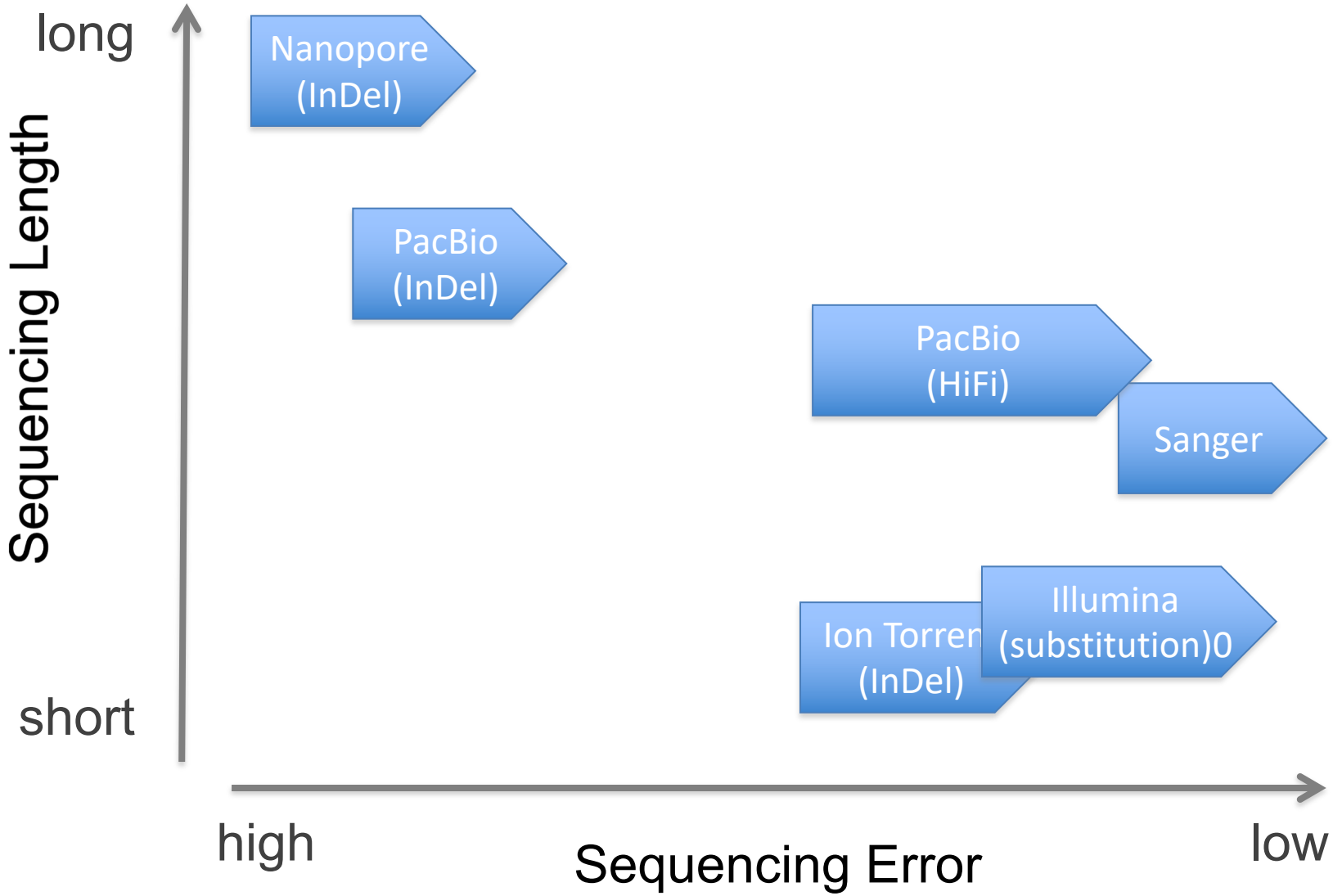


Adaptor is required for library preparation

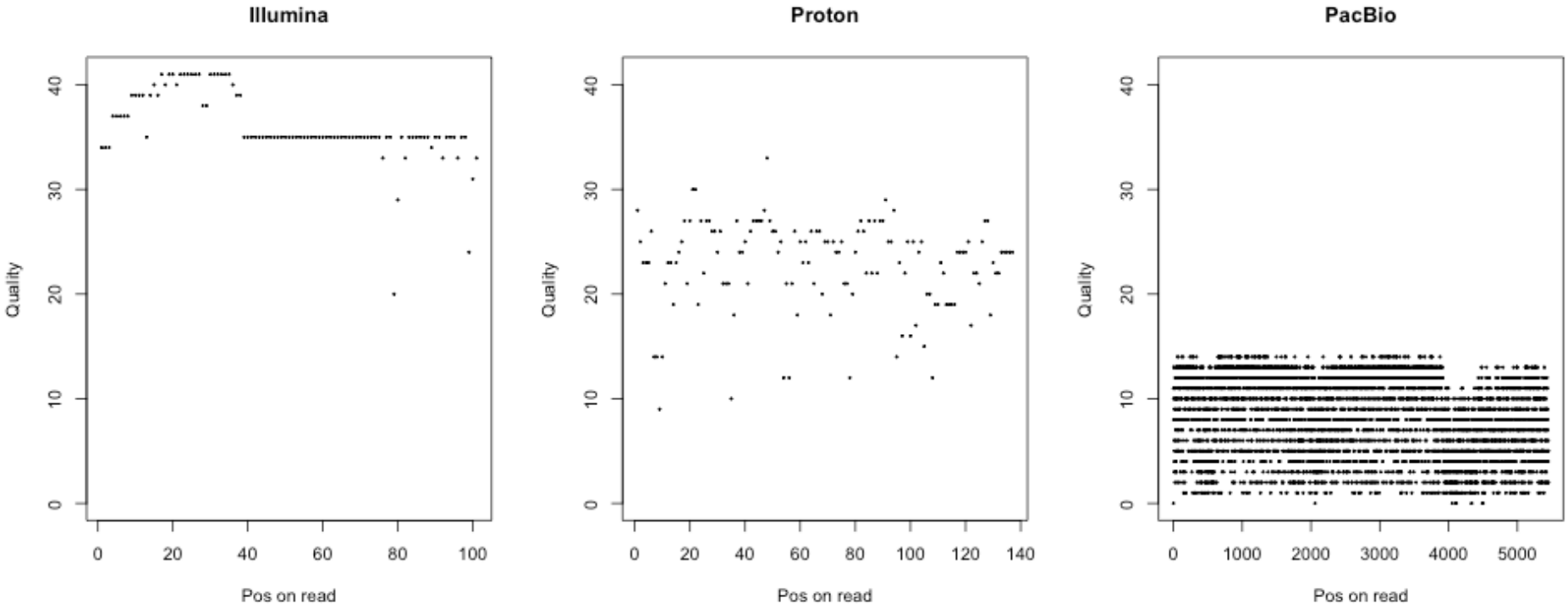
Thousands to hundreds of millions of fragments are sequenced in parallel

Standard data (FASTQ) are output

Sequencing error rates and lengths



Typical reads in different platforms



Read length
Read quality

Applications of NGS

1. Whole-genome sequencing/re-sequencing / target-region sequencing (Assembly, Variant discovery)
2. Genome-reduction sequencing (GBS, RAD-Seq)
3. RNA-Seq: differential expression, alternative splicing and variant discovery
4. Small RNA-Seq
5. ChIP-Seq: Elucidate DNA-protein interaction
6. Metagenomics
7. Others

Case study

1. *De novo* assembly of a strain of *E.coli*
2. Human whole genome sequencing for SNP discovery

Which platform(s)?

Sequencing depth?

Sequence platforms

Illumina (MiSeq, NextSeq, HiSeq)

very high throughput, up to 2x300 bp, and high accuracy (<1%)

Proton (Ion Torrent)

high throughput, up to 300-500 bp, but high errors at homopolymer regions

PacBio

Moderate sequencing throughput, very long (>70kb+), but high errors (5-10%)

PacBio HiFi

~15 kb, ~1% error

Nanopore

Moderate sequencing throughput, very long (> 1 Mb), but high errors (5-15%)

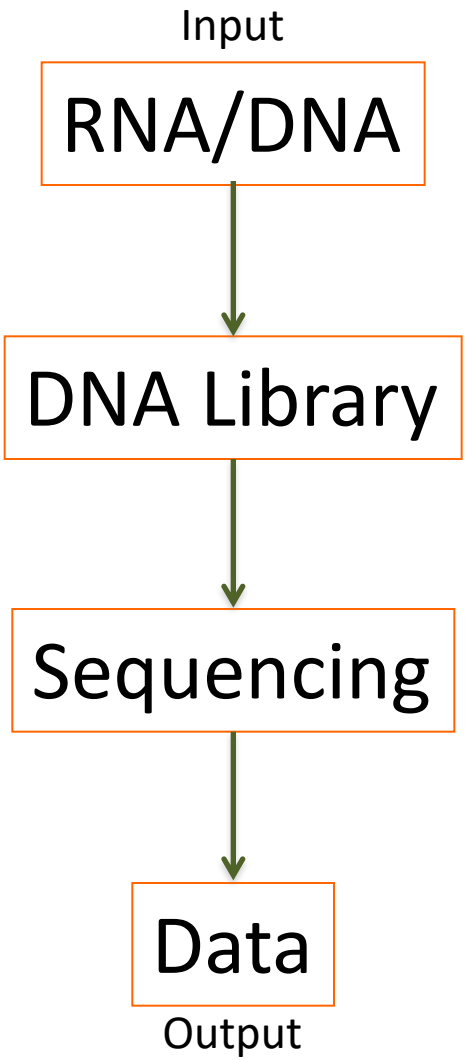
@anne_churchland (twitter)

Experimental design

- Goal
- Platform
- Read length
- Rate and type of sequence errors
- Sequencing depth
- Replication
- Control
- Budget

Platform	Templates	Signal	Read length	Run time	reads per run	Error type	Error rate
Illumina Miseq	PCR or PCR-free	fluorescent	up to 2x300	1-2 days	Up to 10 Gb	substitutions	~0.1-1%
Illumina Hiseq	PCR or PCR-free	fluorescent	up to 2x250	days	Hundreds of Gb	substitutions	~0.1-1%
Ion Torrent	PCR	H+	300-500	2 hours	10 Gb?	InDel	>1%
PacBio	Amplification not required	fluorescent	Average >5,000	30min	500 Mb – 1 Gb	InDel	5-10%
Nanopore	Amplification not required	Electronic flow change	>1,000	hours	>5Gb per MinION	InDel	5-15%

COMMON in all NGS platforms



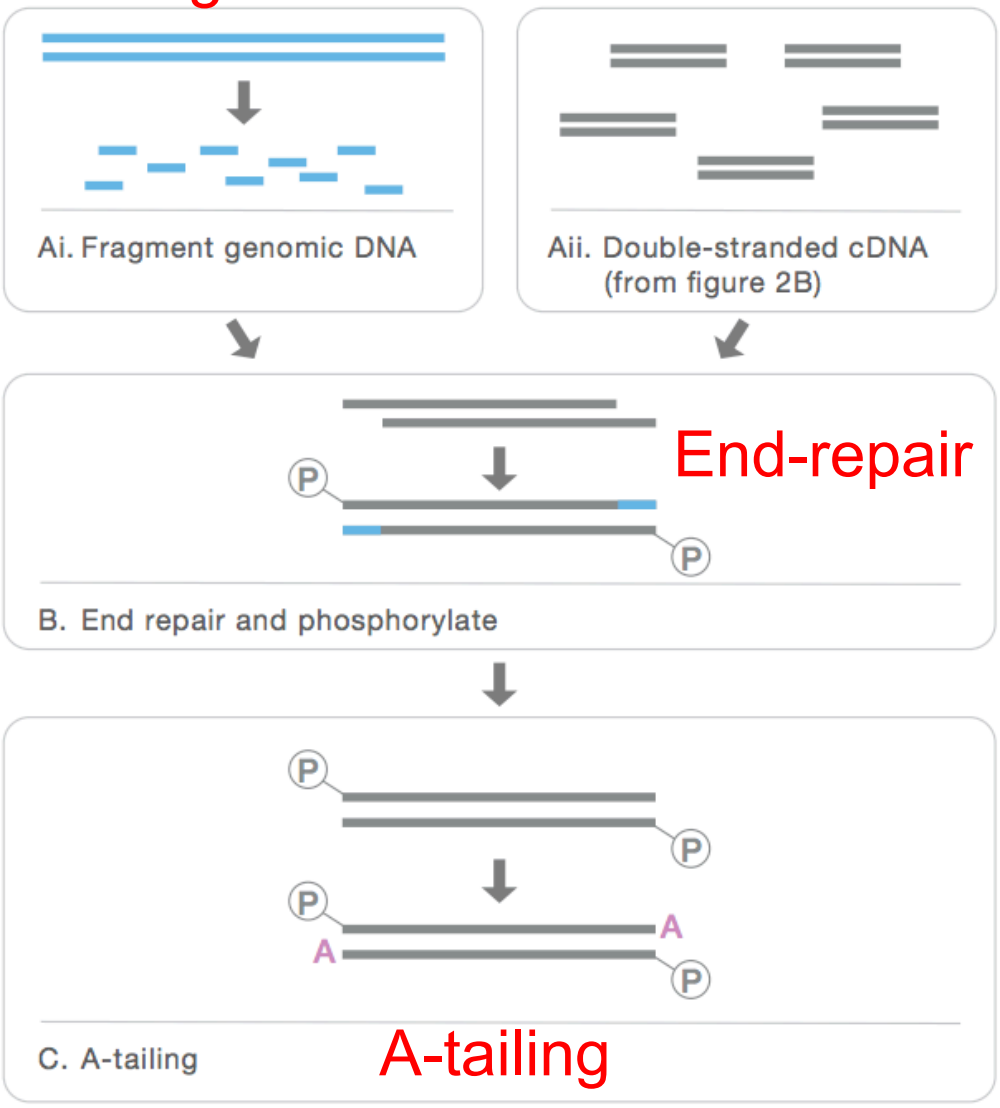
Adaptor is required for library preparation

Thousands to hundreds of millions of fragments are sequenced in parallel

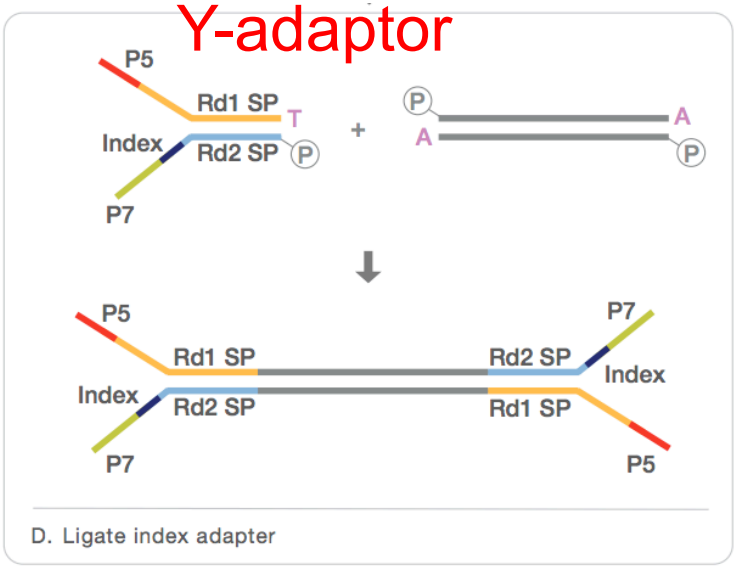
Standard data (FASTQ) are output

Library preparation – Y-adaptor method

a. Fragmentation

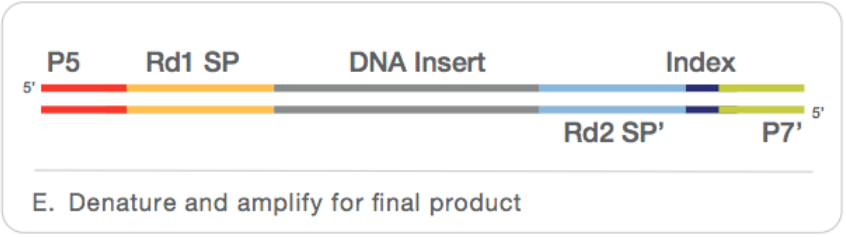


b.



PCR or PCR-free

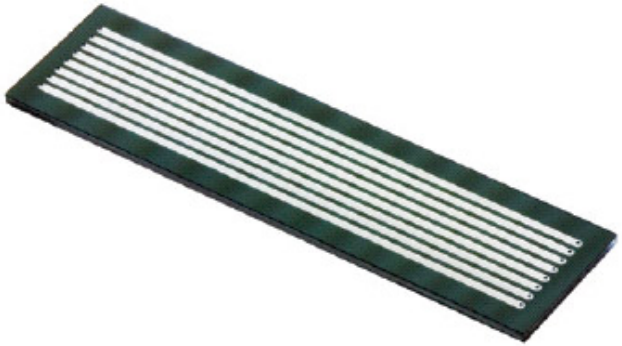
Final product



From TruSeq Manual

Multiplexing (DNA barcode/Index)

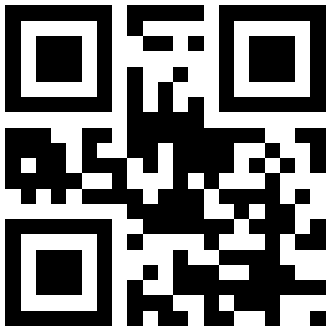
flowcell
lane



- per lane's data are more than needed in many cases

- Multiplexing: To put multiple samples in a lane via using **DNA barcodes** to distinguish samples

Barcode / Index

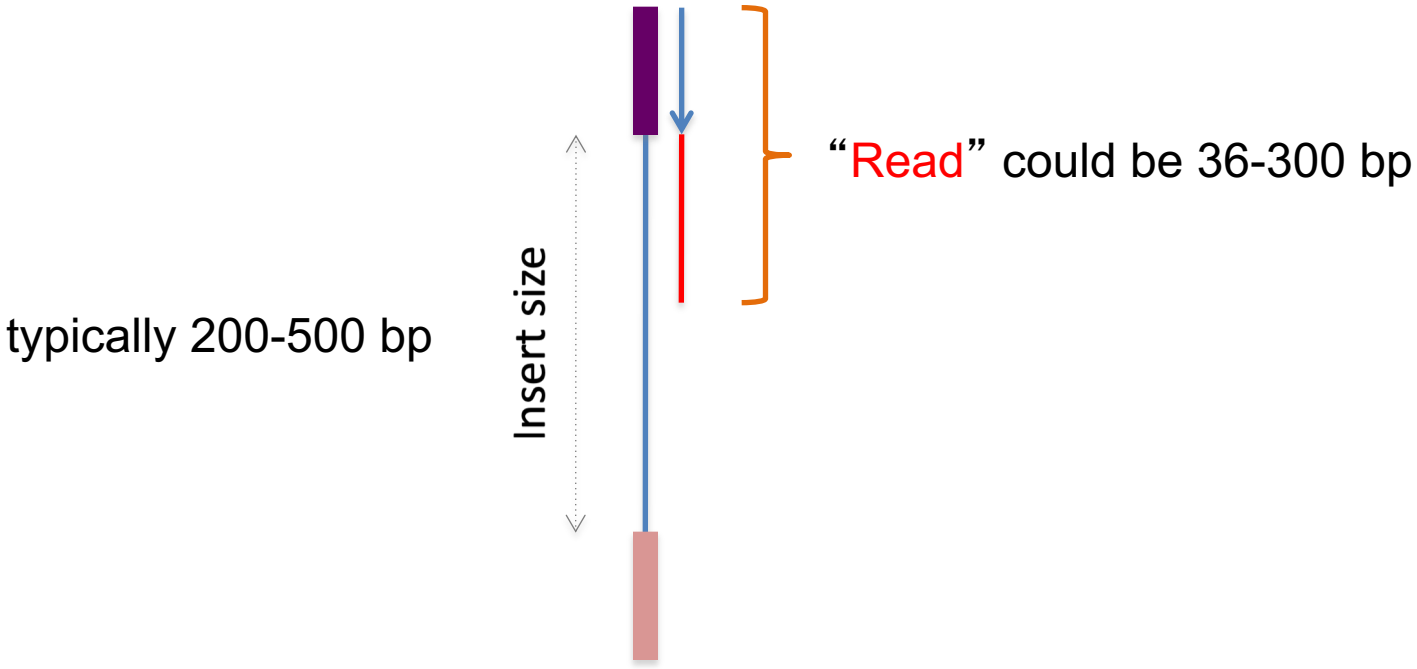


```
           AGTGCAxxxxxxxxxxxxx
sample 1   AGTGCAxxxxxxxxxxxxx
           AGTGCAxxxxxxxxxxxxx

           CATGTCxxxxxxxxxxxxx
sample 2   CATGTCxxxxxxxxxxxxx
           CATGTCxxxxxxxxxxxxx
```

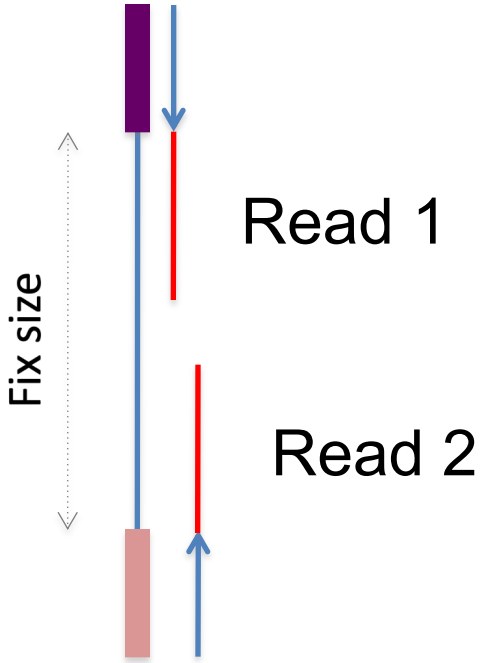
Single-end sequencing

A single read is generated for each template/cluster



Paired-end sequencing

Two reads are generated for each template cluster;
the 1st is from one end with one primer;
the 2nd is for the other end with the other primer.



Illumina platforms and terminologies

Illumina video

1. Library preparation
2. Sequencing procedure
3. Single-ends and paired ends

Summary

1. NGS platforms
2. Pro and con of each platform
3. Approaches for library preparation
4. Applications of various NGS tech