

# Next-gen Sequencing Technologies

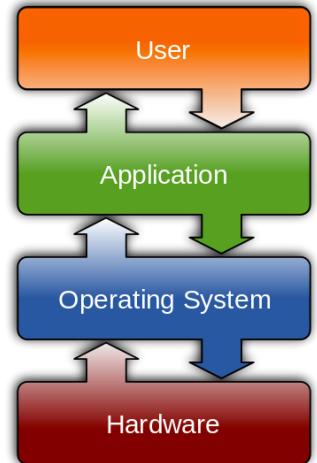
Bioinformatics Applications (PLPTH813)

Sanzhen Liu

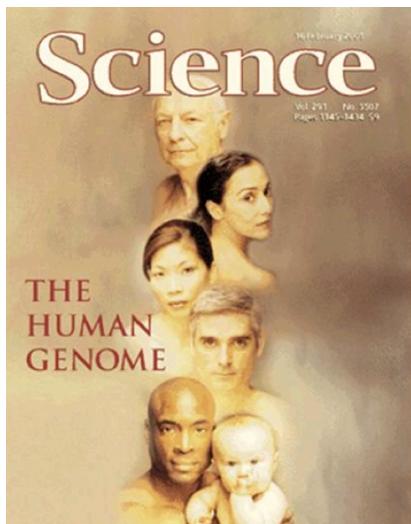
2/4/2021

# 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

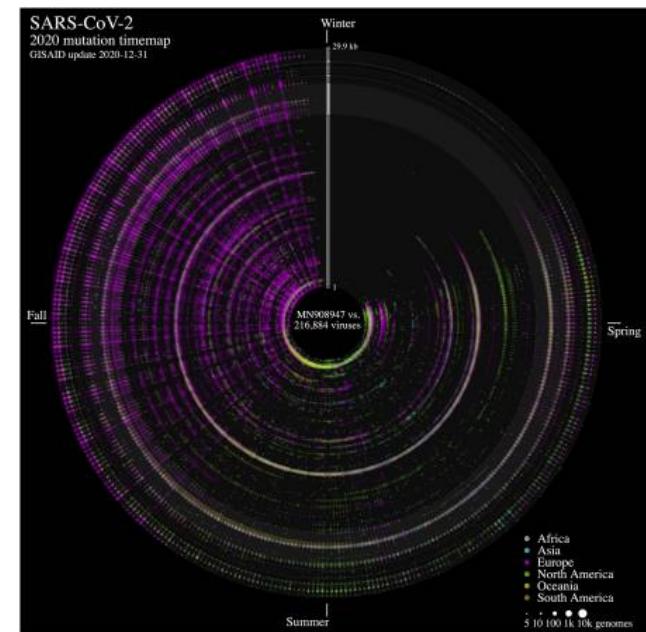
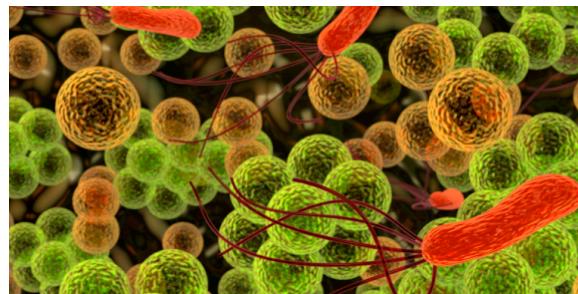
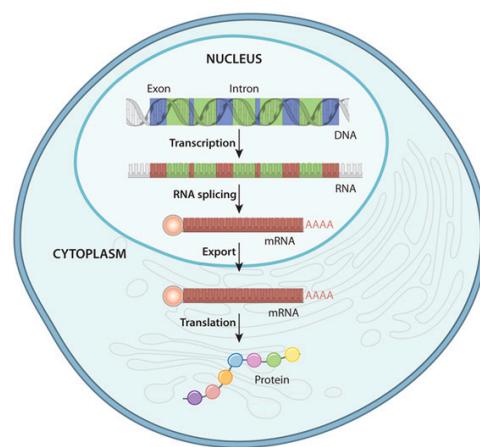
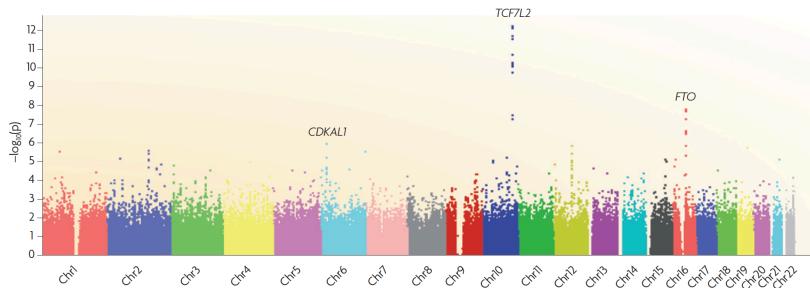


The tragic ripples of  
an epic fraud p. 638 | Insect pest profiles from  
male mouse defenses p. 642 & 644 | Photoredox activation of methane p. 642 & 644

**Science** 27 AUGUST 2018  
Volume 351 Issue 6274  
DOI: 10.1126/science.aar6000  
ISSN 0036-8075



ROAD MAP FOR  
**WHEAT**  
Ordered sequence will  
speed research p. 625, 630, & 632

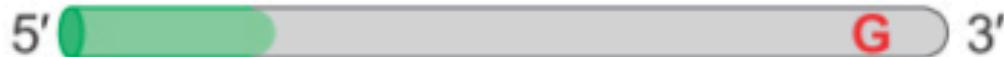


# Sanger sequencing technology - I

a



primer



DNA synthesis →

substrates:

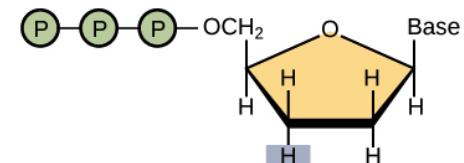
dATP dGTP

ddGTP

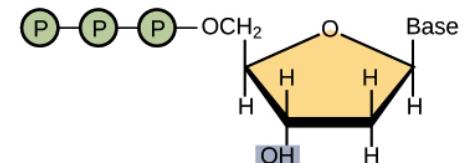
dCTP dTTP



Frederick Sanger



Dideoxynucleotide (ddNTP)



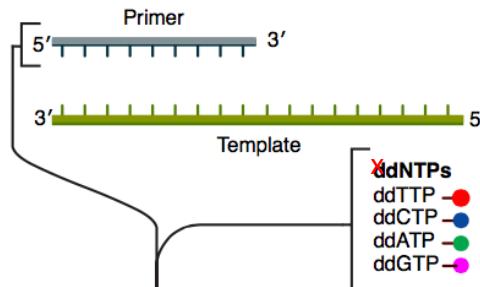
Deoxynucleotide (dNTP)

**Key innovation**

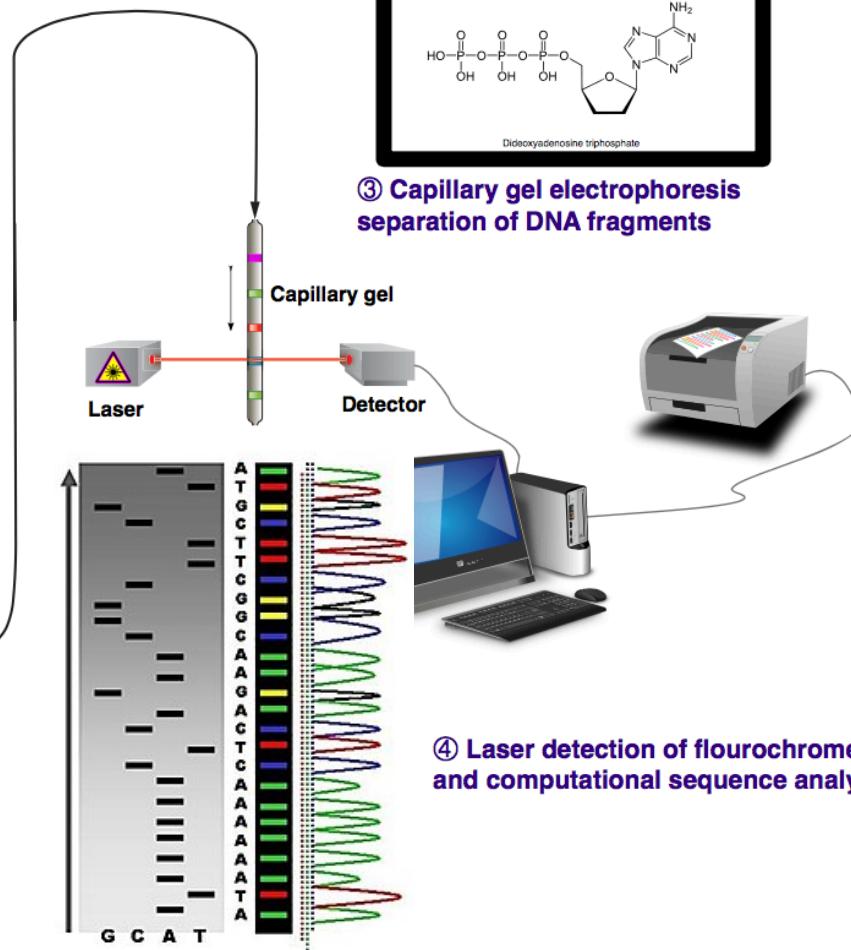
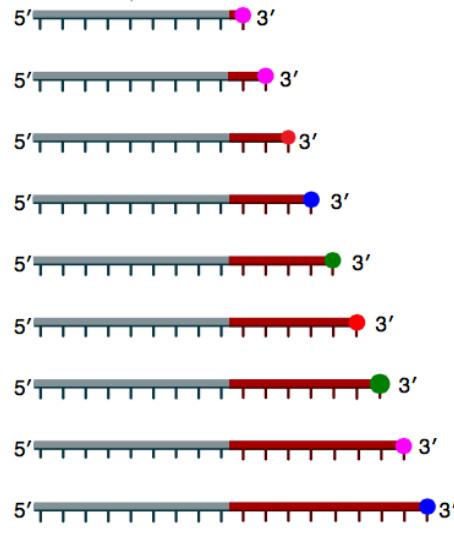
# Sanger sequencing technology - II

## ① Reaction mixture

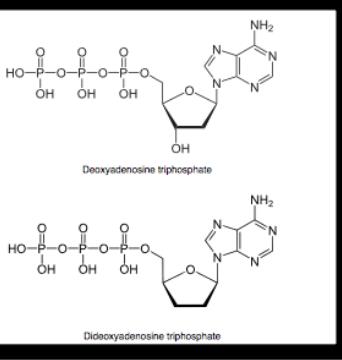
- Primer and DNA template → DNA polymerase
- ddNTPs with flourophores → dNTPs (dATP, dCTP, dGTP, and dTTP)



## ② Primer elongation and chain termination



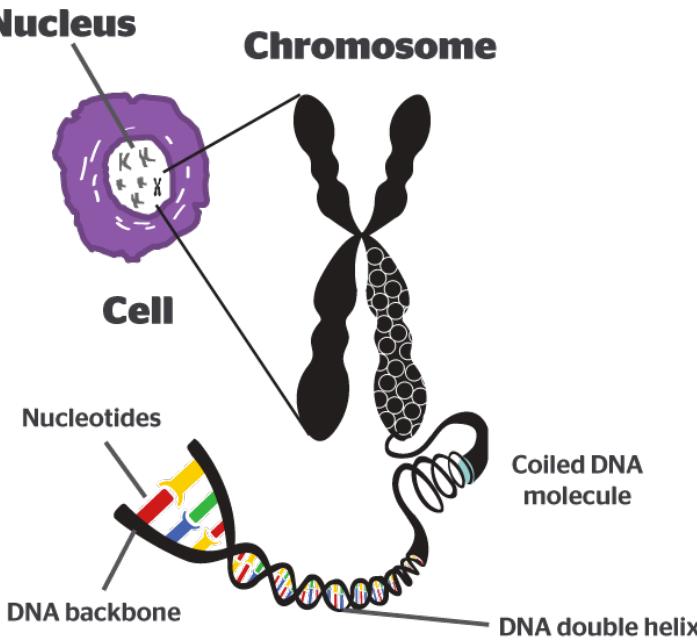
## ③ Capillary gel electrophoresis separation of DNA fragments



## ④ Laser detection of flourophores and computational sequence analysis

# Major Next-gen sequencing (NGS) technologies

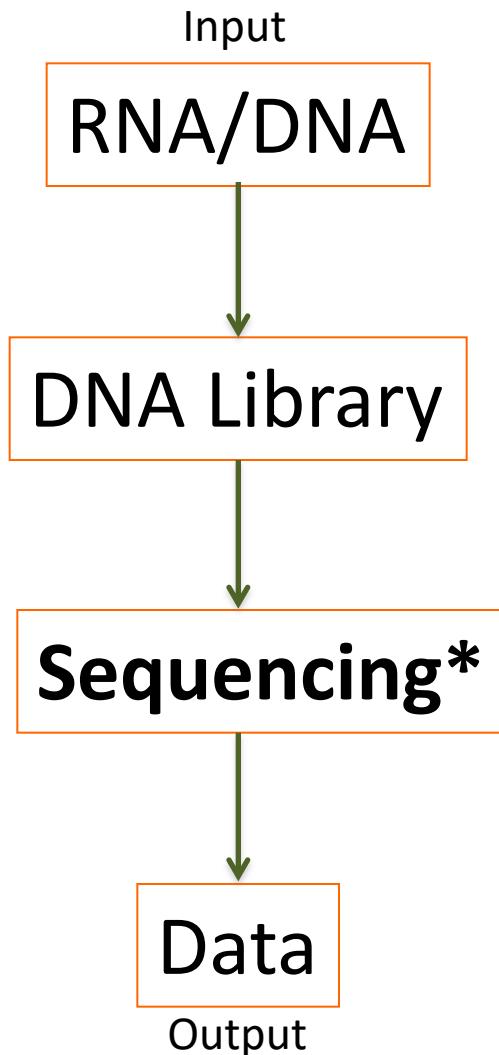




sequencing sensitivity and read length

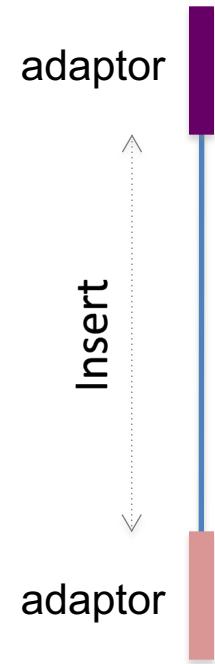
Before single molecular & "super long" sequencing technologies, **fragmentation** and **amplification/cloning** of a single nucleotide molecule are needed for sequencing.

# 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**

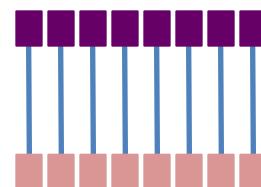


Nucleotide detector:  
**VERY** sensitive

Directly read sequence  
single-molecule



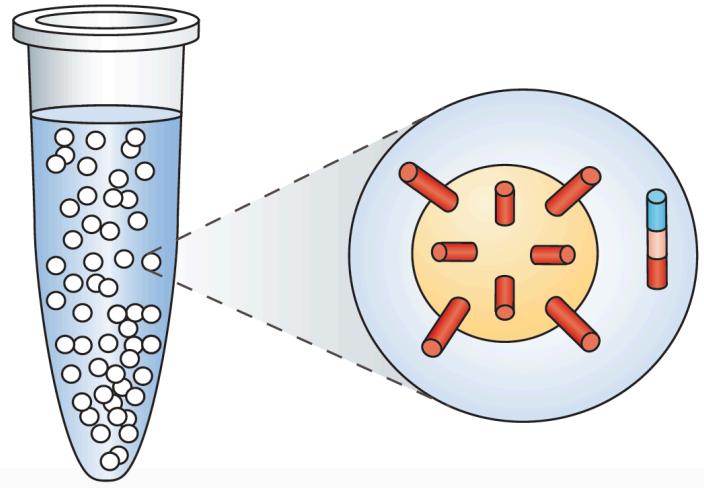
Nucleotide detector:  
Not sensitive at the  
single molecular level



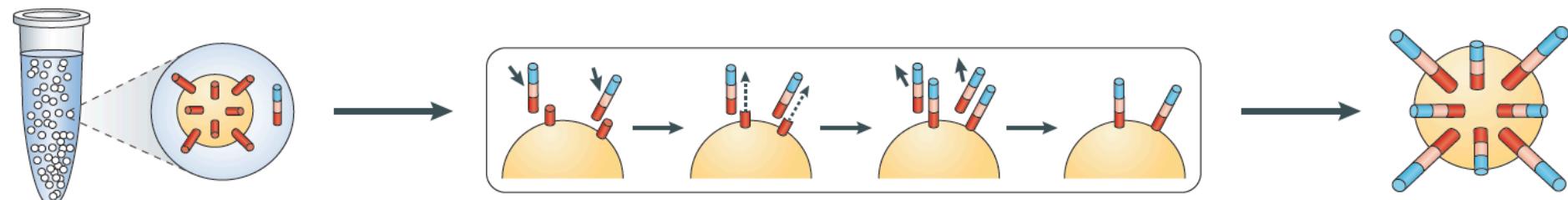
amplify and then  
read sequence

"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



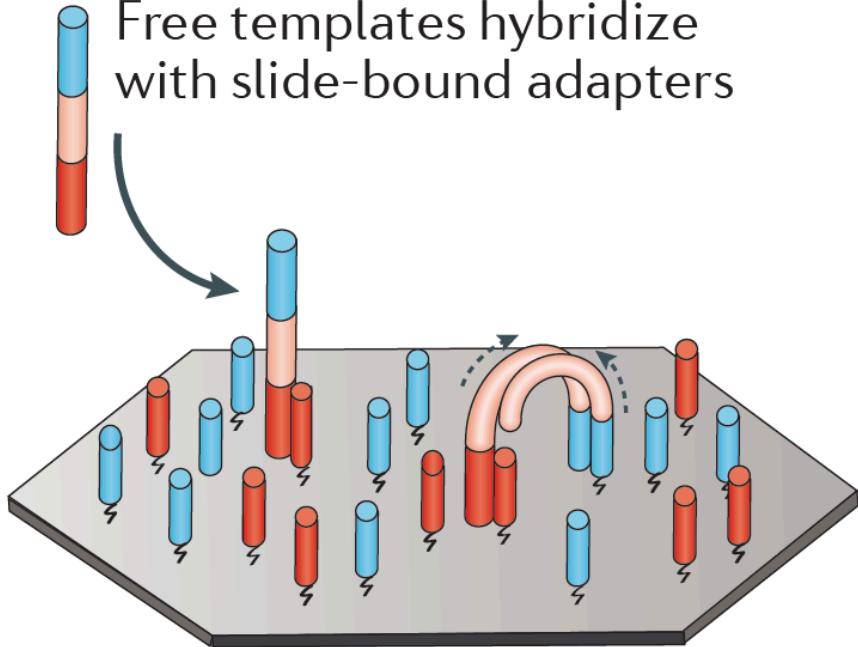
Water-in-oil emulsion PCR  
(454 and Ion Torrent)



# Massive independent amplifications – bridge PCR

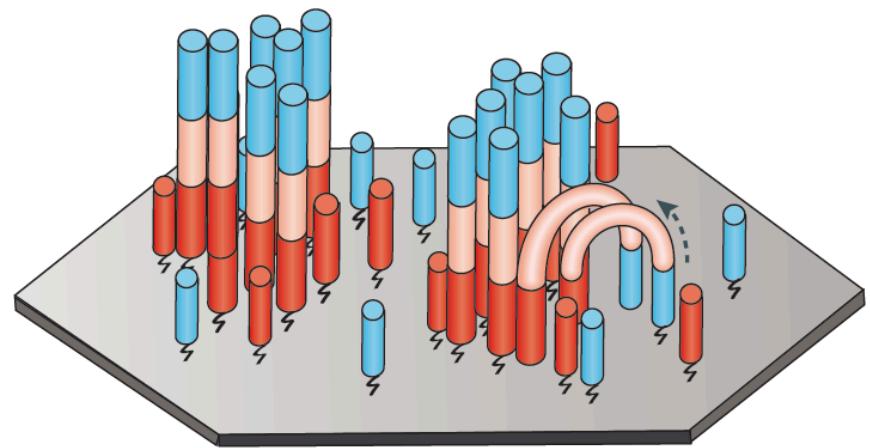
## Template binding

Free templates hybridize with slide-bound adapters



## Bridge amplification

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

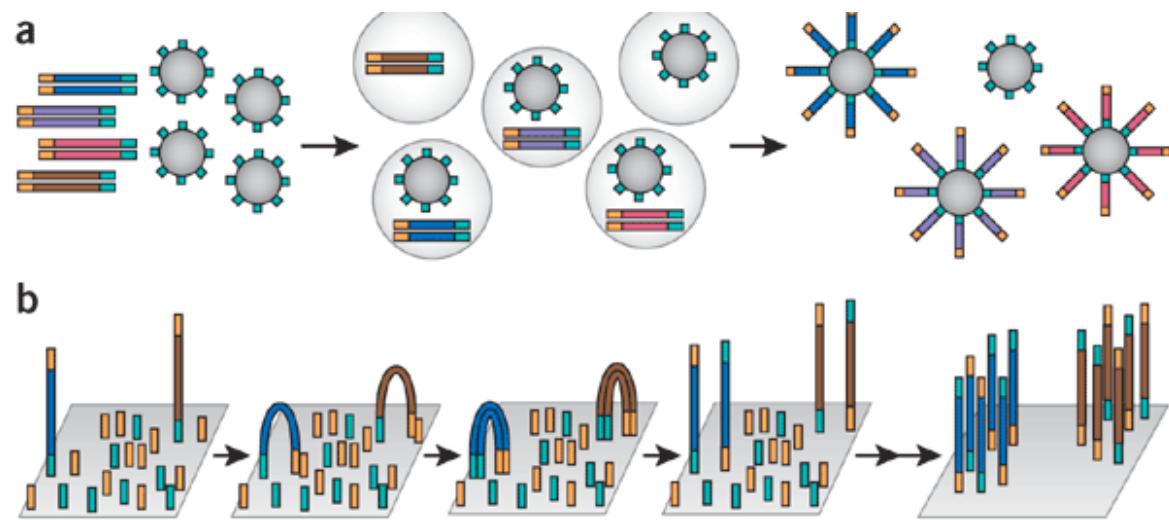
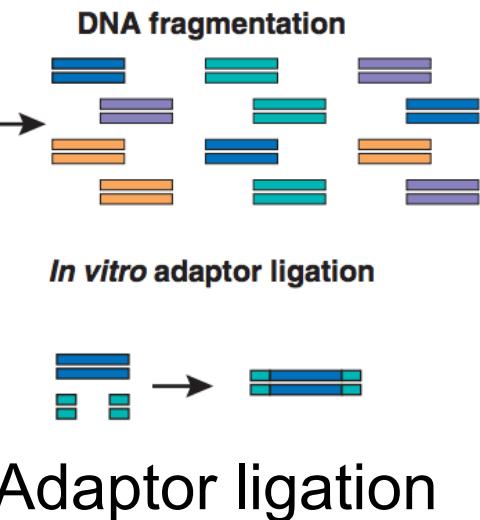


## Cluster generation

After several rounds of amplification, 100–200 million clonal clusters are formed

# DNA amplification

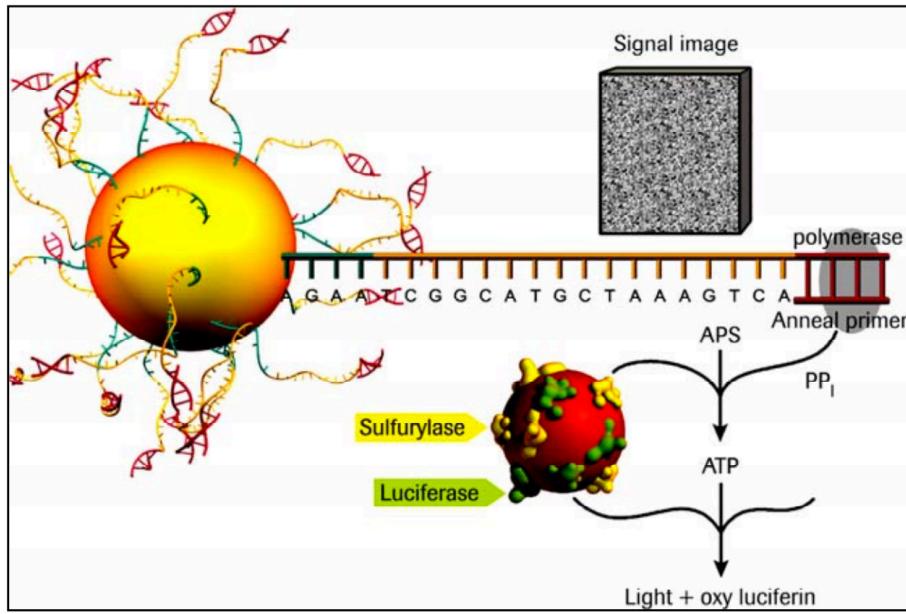
## Water-in-oil emulsion PCR (454 and Ion Torrent)



## Bridge PCR on slides (Illumina)

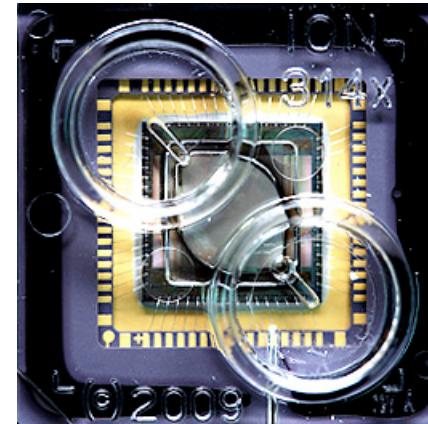
Nature Biotechnology, 2008, 26: 1135-45

# 454 and Ion Torrent signal detectors



454 technology, Nature 2005, 437: 376-380

1. Sequencing by synthesis
2. Pyrosequencing (454)



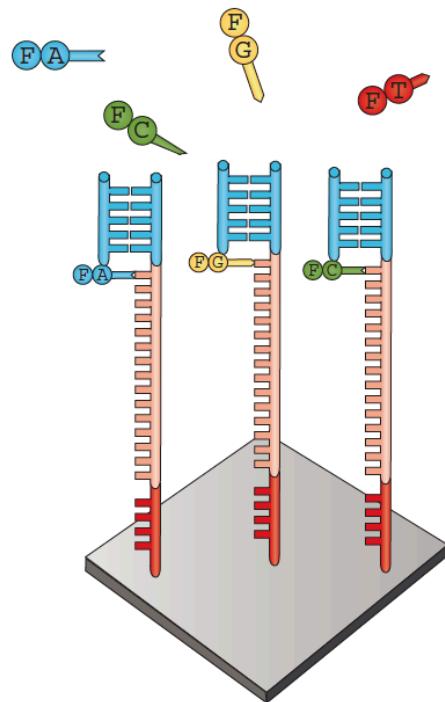
[Ion Torrent video](#)

1. Ion Torrent technology is similar to 454 technology
2. The signal is  $H^+$  rather than pyrophosphate

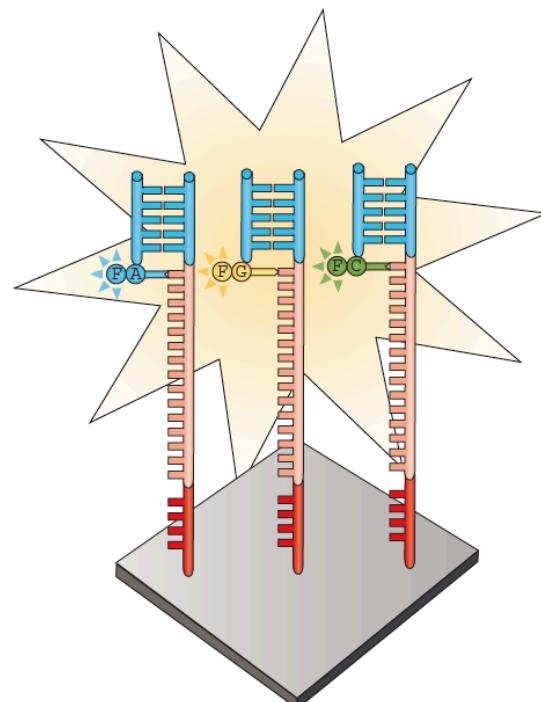
# Illumina

Two key technologies:

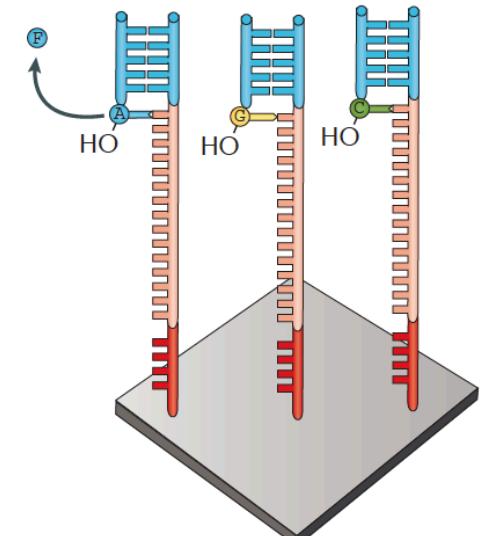
1. Bridge PCR
2. Reversible terminator chemistry



Nucleotide addition



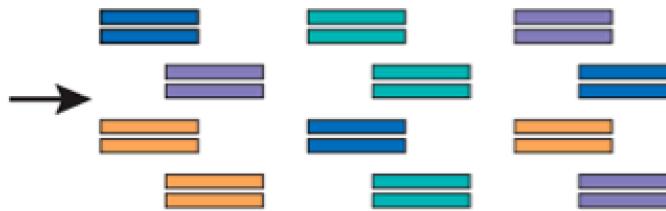
Imaging



Cleavage

# Illumina sequencing

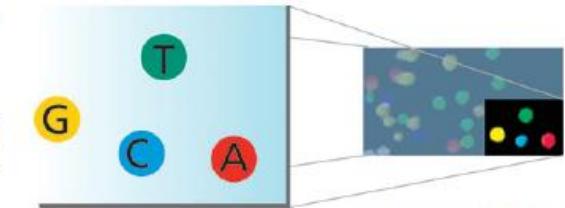
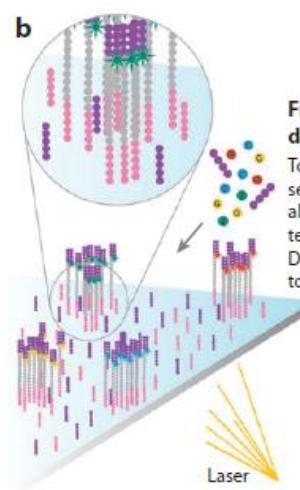
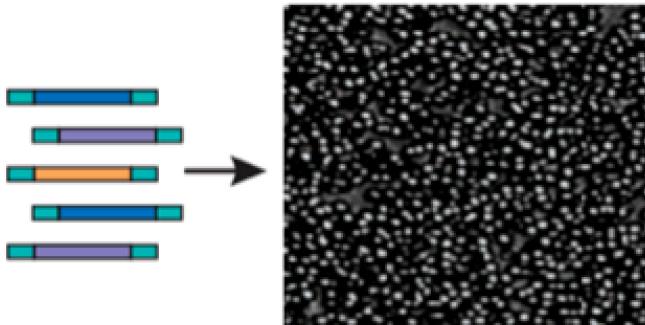
## DNA fragmentation



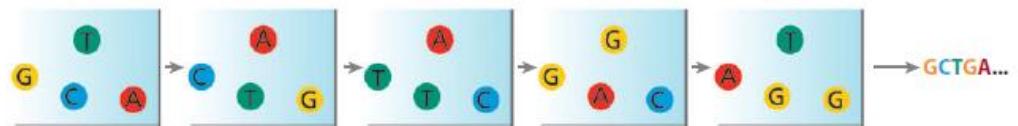
## In vitro adaptor ligation



## Generation of polony array



Before initiating the next chemistry cycle  
The blocked 3' terminus and the fluorophore from each incorporated base are removed.



## Sequence read over multiple chemistry cycles

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

## Two key technologies:

1. Bridge PCR
2. Reversible terminator chemistry

# Illumina Sequencers



The MiSeq Series sequencer is a compact, benchtop instrument designed for high-throughput sequencing. It features a sleek design with a black control module on the right and a white sequencing unit on the left. A large touchscreen display on the control module shows the 'Welcome to MiSeq Control Software' interface.

**MiSeq Series**

<b>MAX OUTPUT</b>	<b>15 Gb</b>
<b>MAX READ NUMBER</b>	<b>25 million</b>
<b>MAX READ LENGTH</b>	<b>2x300 bp</b>



The HiSeq Series sequencer is a larger, more complex instrument than the MiSeq. It consists of a central white sequencing unit connected by a blue cable to a black control module on the left. Both units feature illuminated displays showing the 'Welcome to HiSeq Control Software' interface.

**HiSeq Series**

<b>MAX OUTPUT</b>	<b>1500 Gb</b>
<b>MAX READ NUMBER</b>	<b>5 billion</b>
<b>MAX READ LENGTH</b>	<b>2x150 bp</b>



# 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

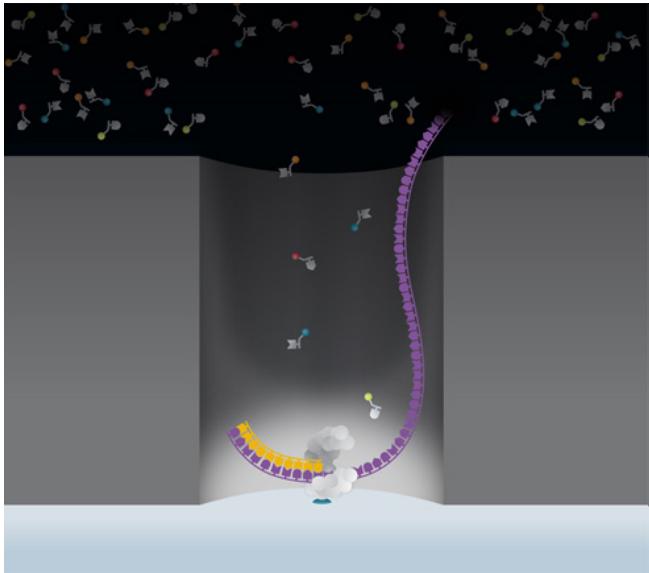
Record signal per **nucleotide type**:

A T G C A A A A

Sequencing errors at homopolymers

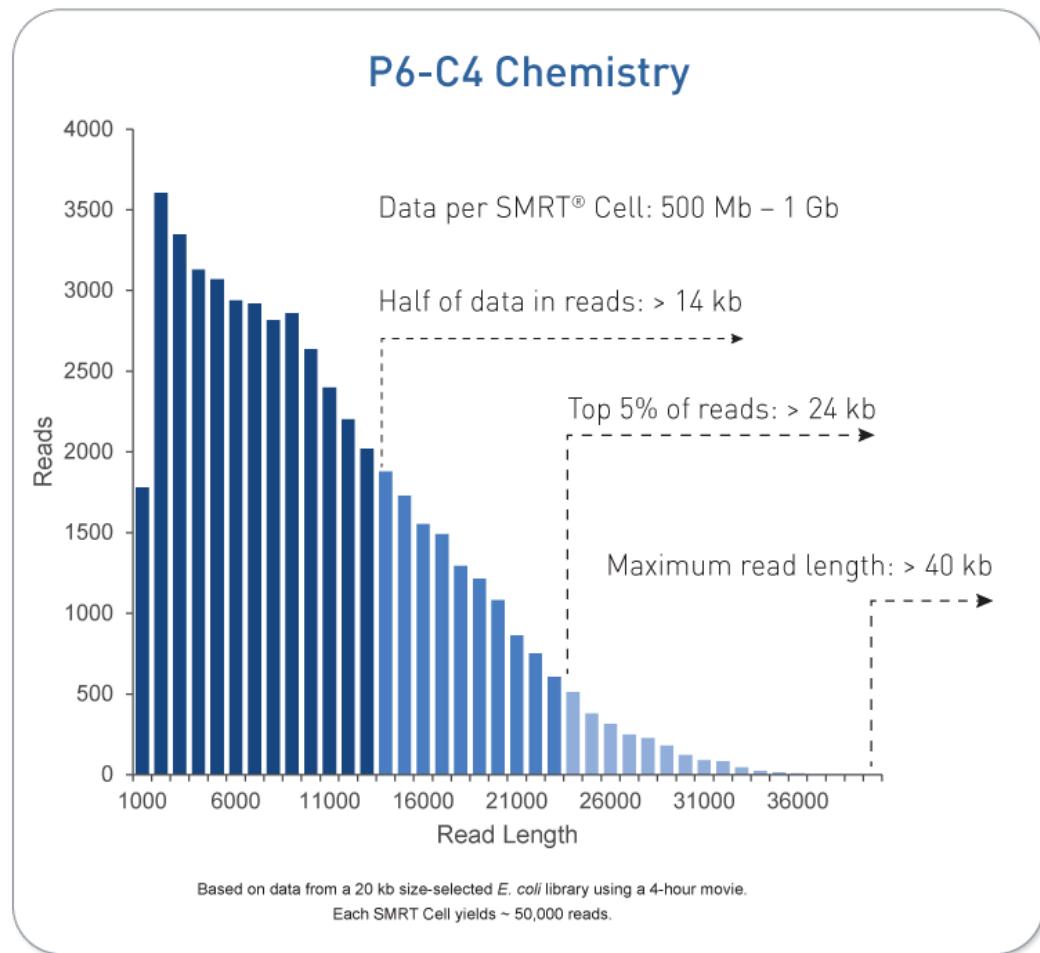
When the single molecular sequencing technology is ready, **amplification or cloning** is not necessary.

# PacBio – Single Molecule Real Time (SMRT)

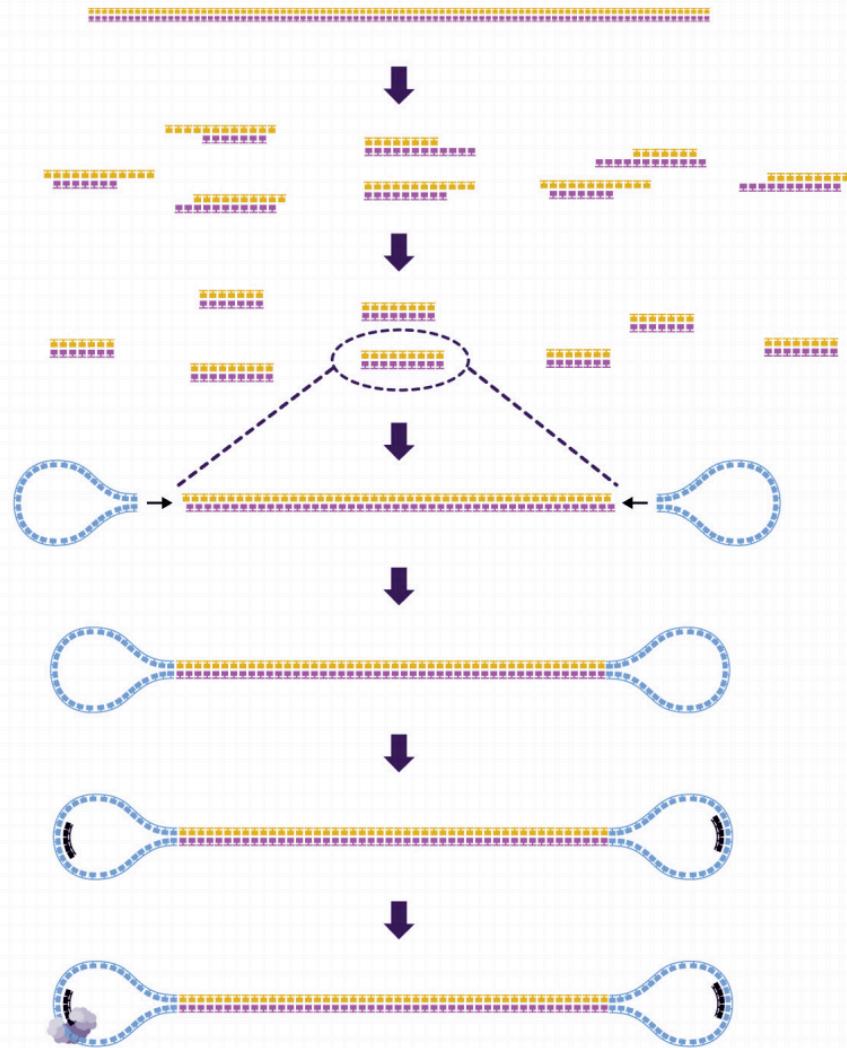


[PacBio tech video](#)

- 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)

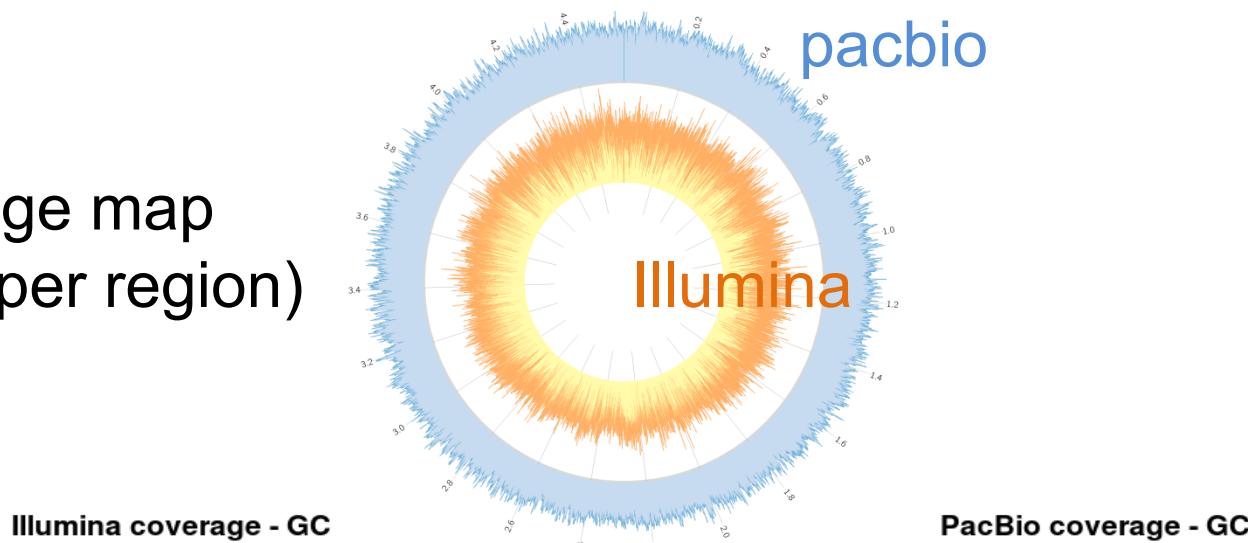


# PacBio library prep workflow



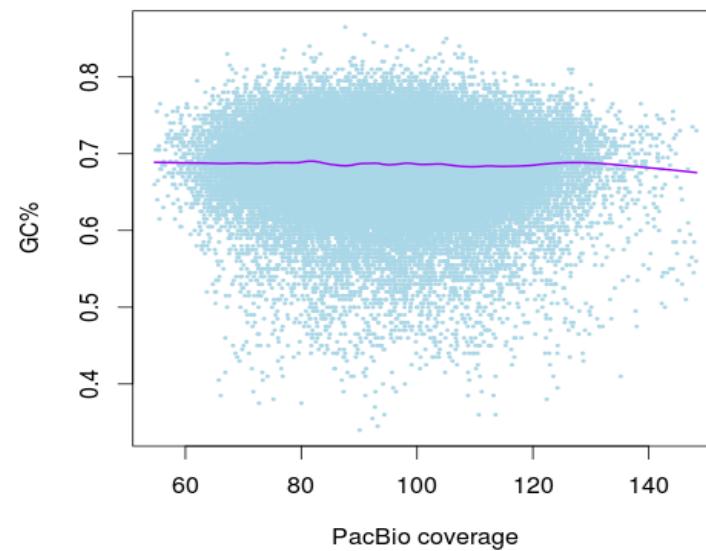
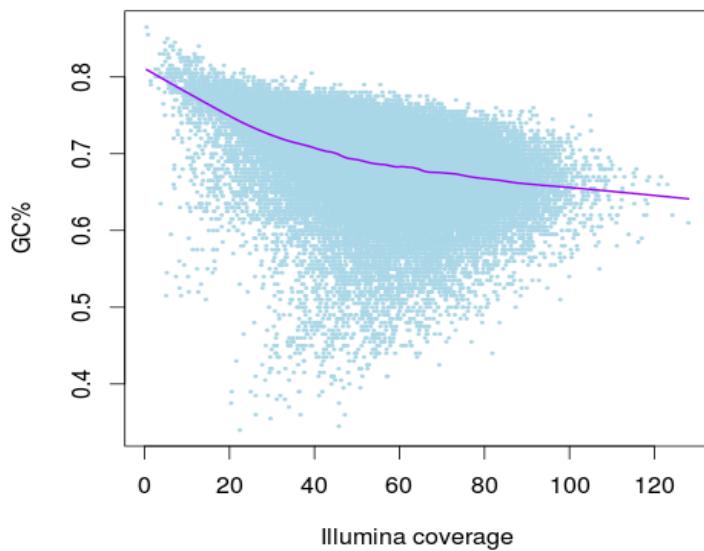
Less biases (e.g., GC)

Coverage map  
(depth per region)

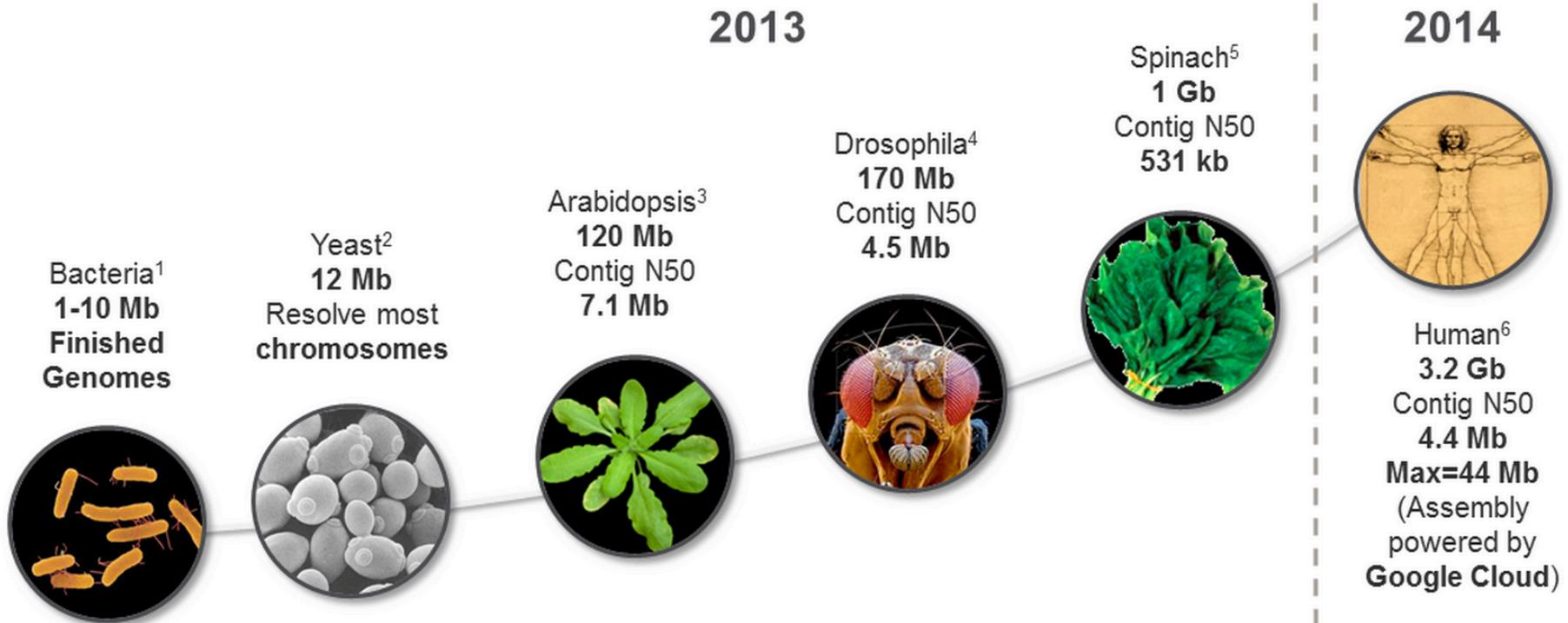


Illumina coverage - GC

PacBio coverage - GC

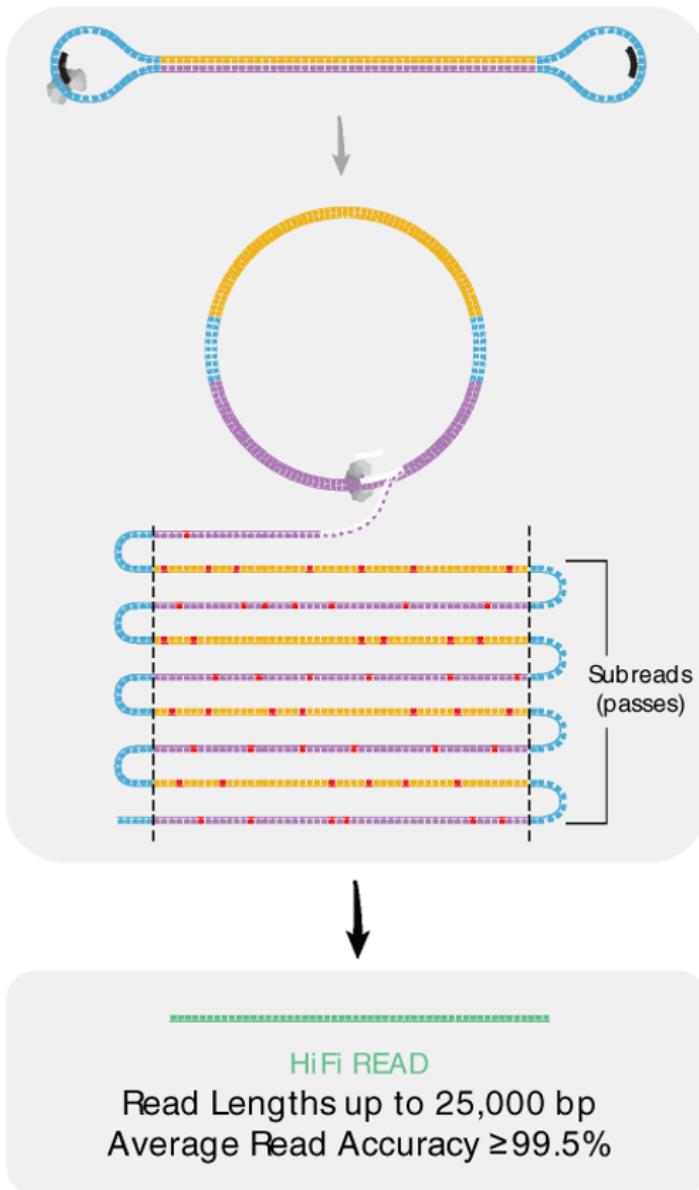


# PacBio for genome assembly



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.

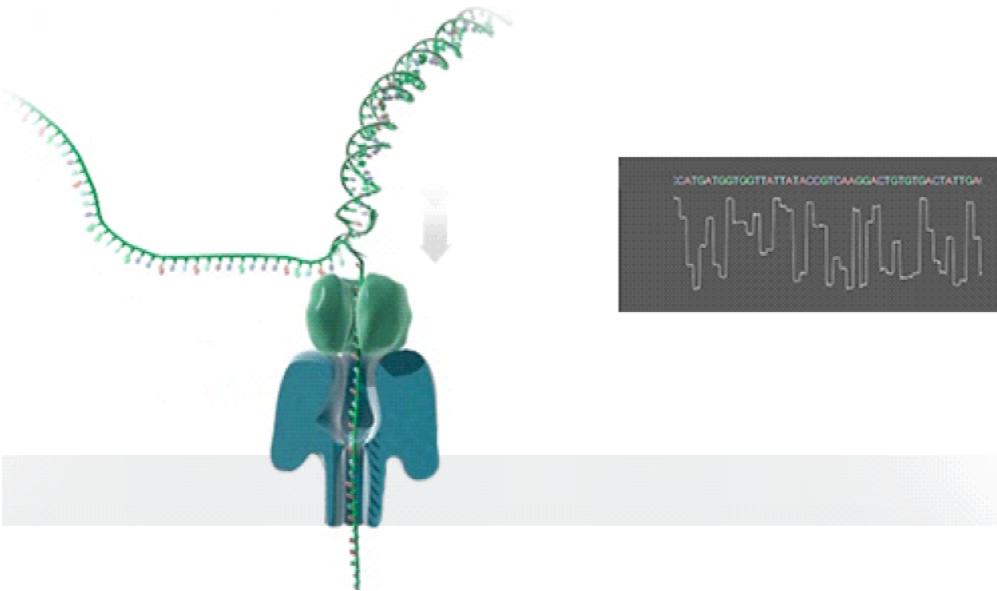
# HiFi PacBio data



Up to 25 kb with  
~99% accuracy

# 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



Flongle

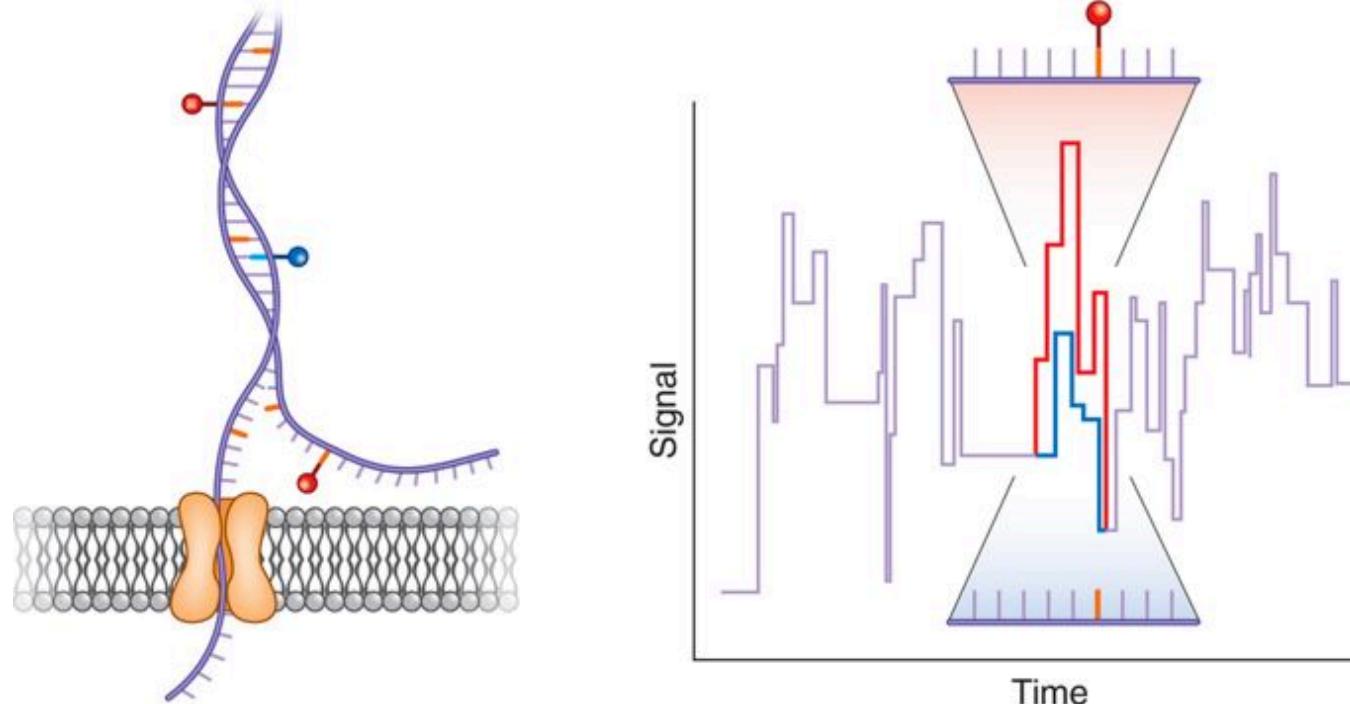
MinION

GridION<sub>X5</sub>

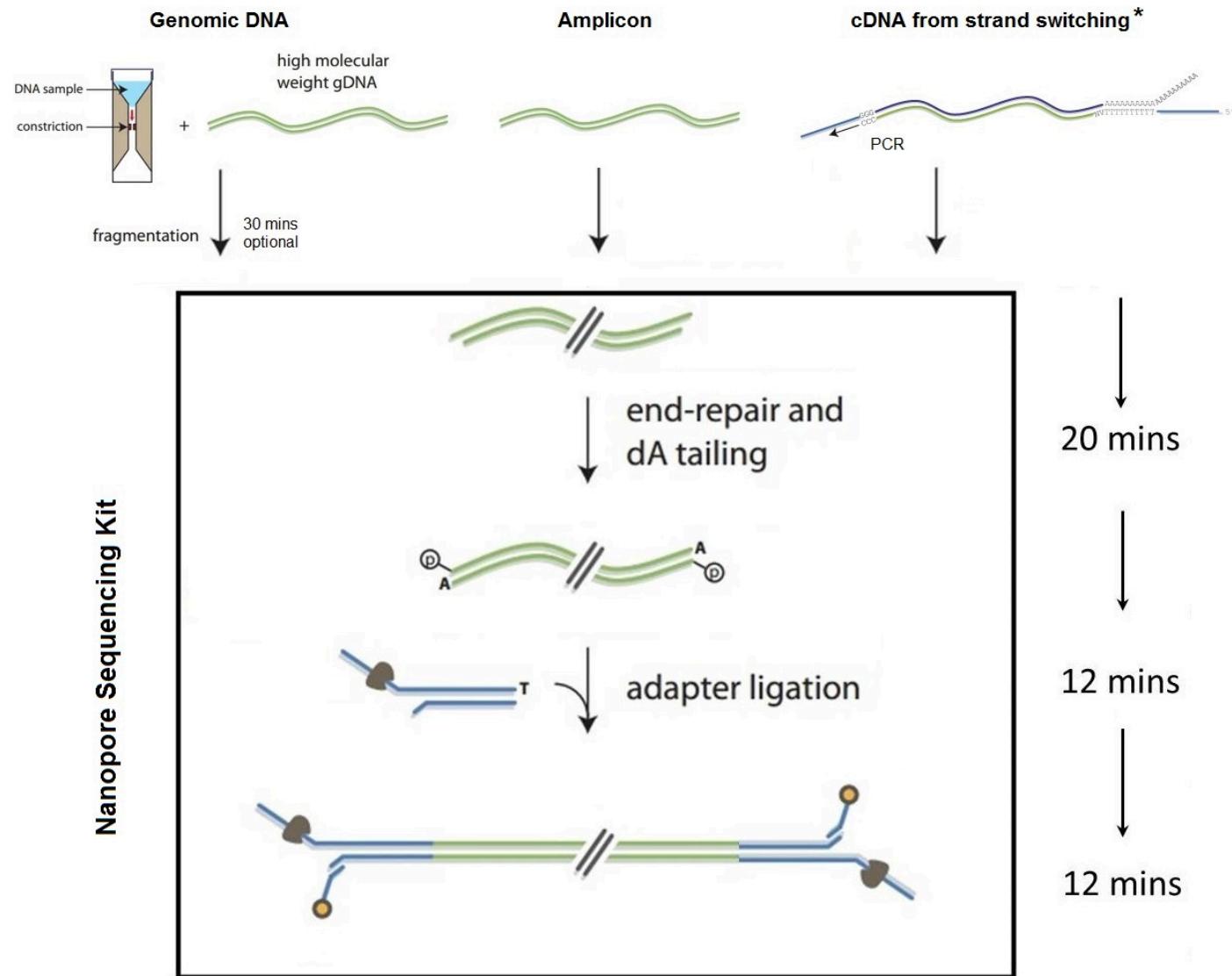
PromethION

# Applications of Nanopore sequencing

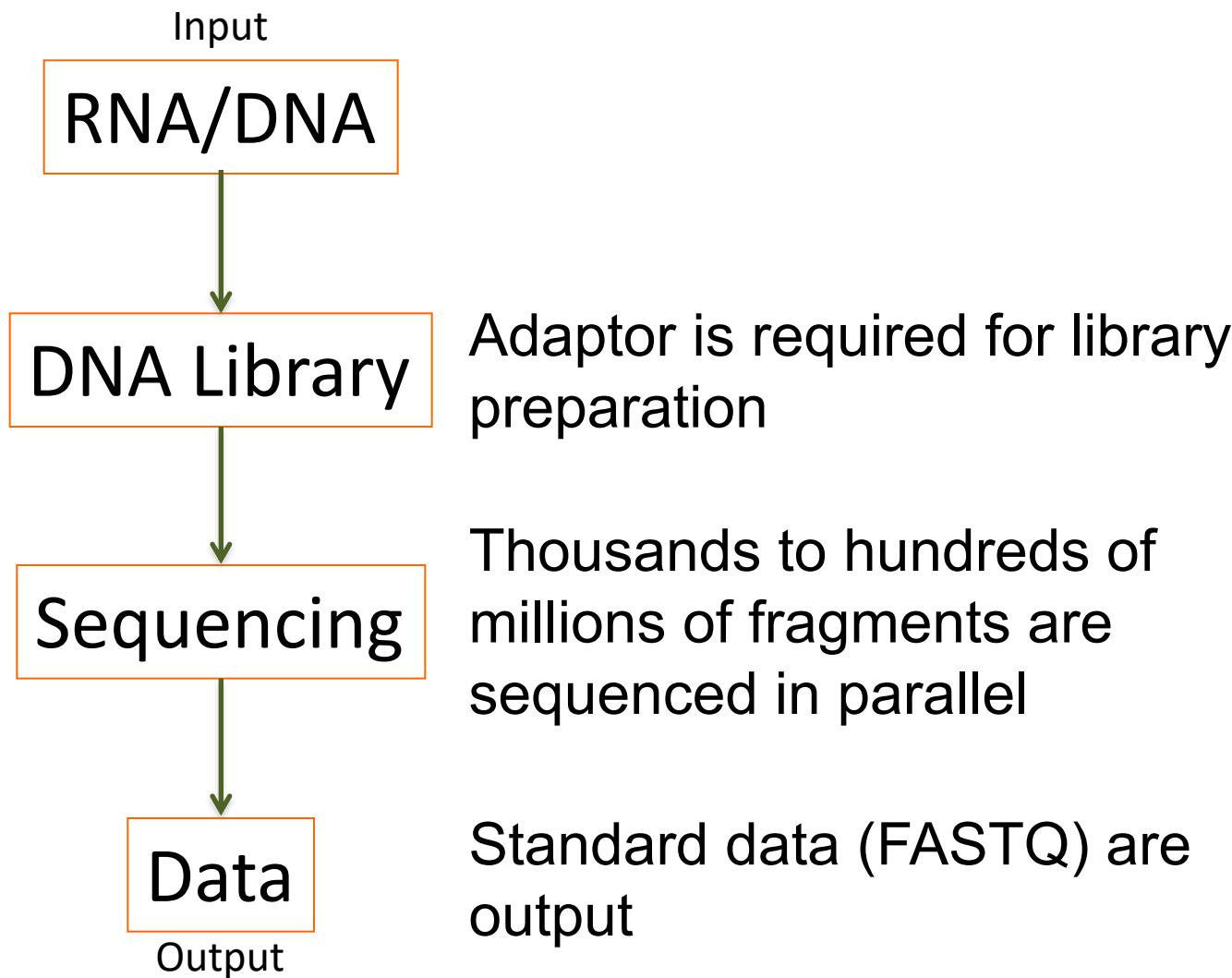
1. Genomic DNA sequencing
2. RNA sequencing (direct RNA or cDNA)
3. DNA methylation and other modifications



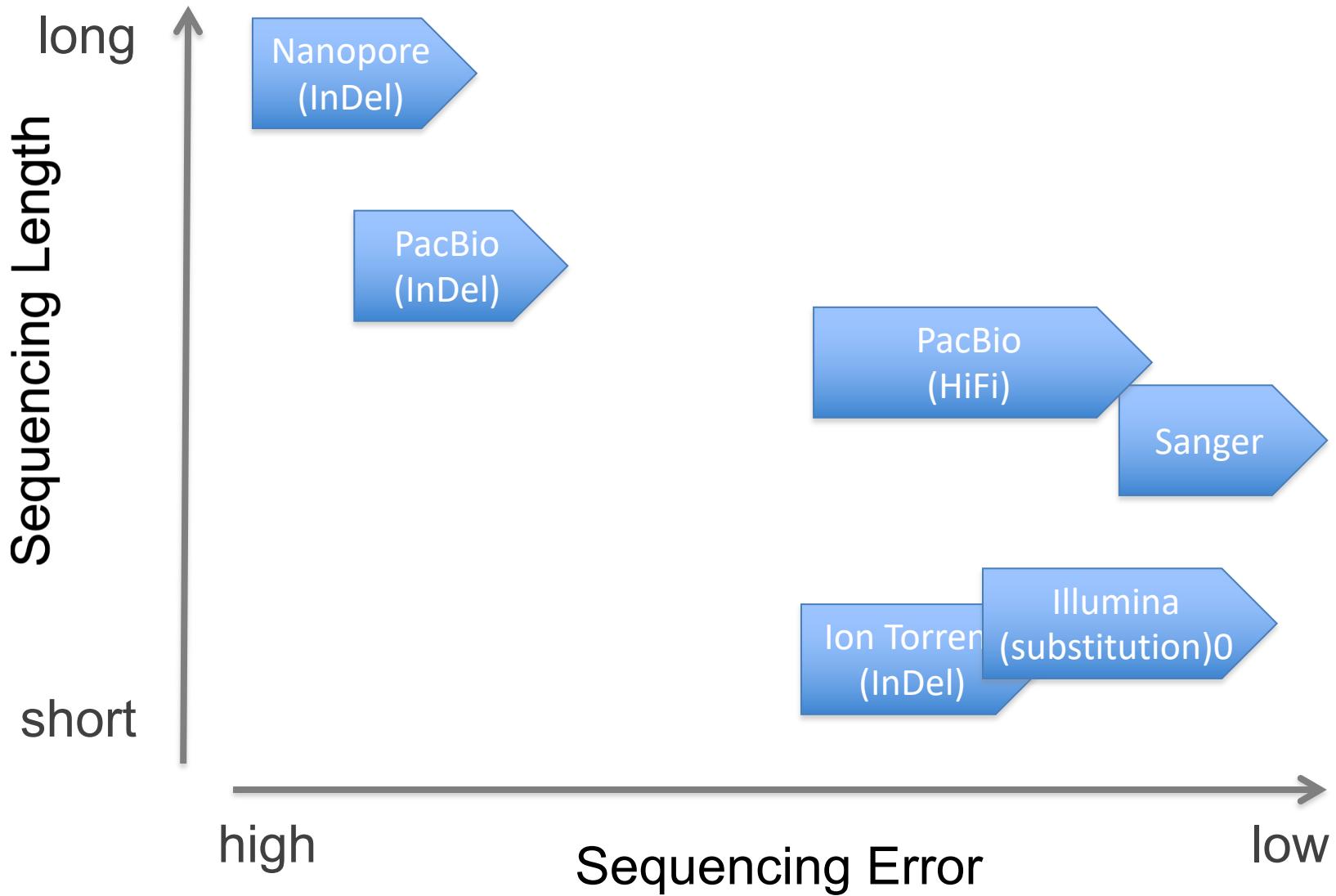
# Nanopore library preparation



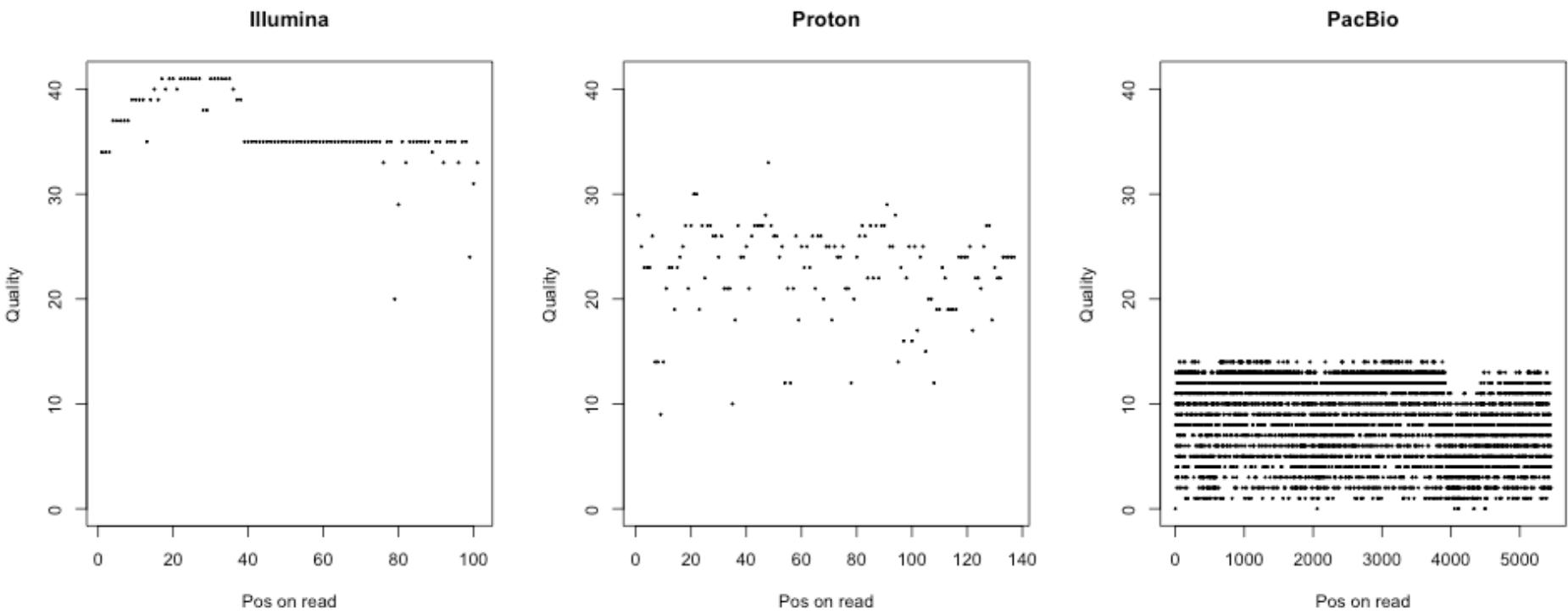
## COMMON in all NGS platforms



# 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%)



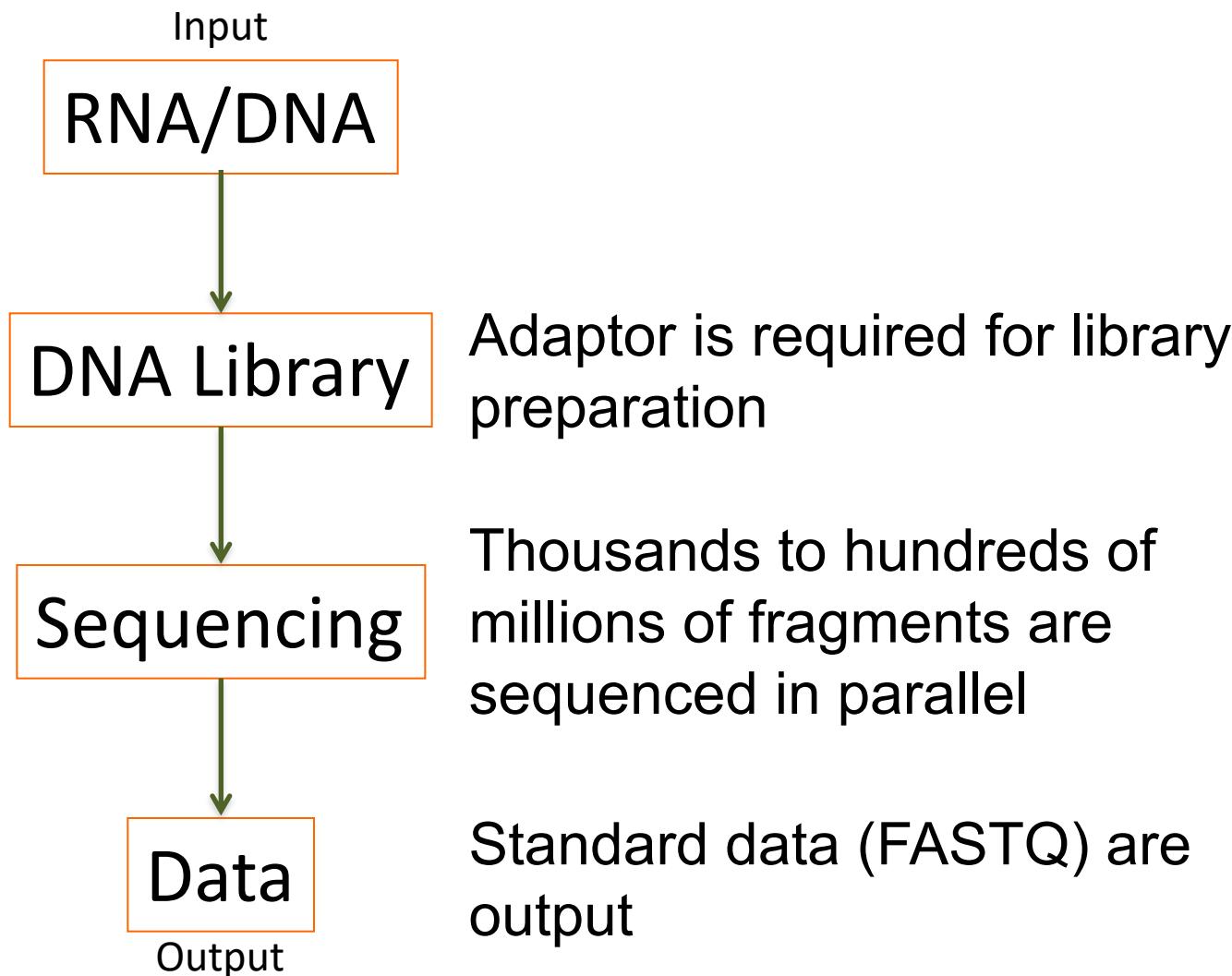
@anne\_churchland (twitter)

# Experimental design

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

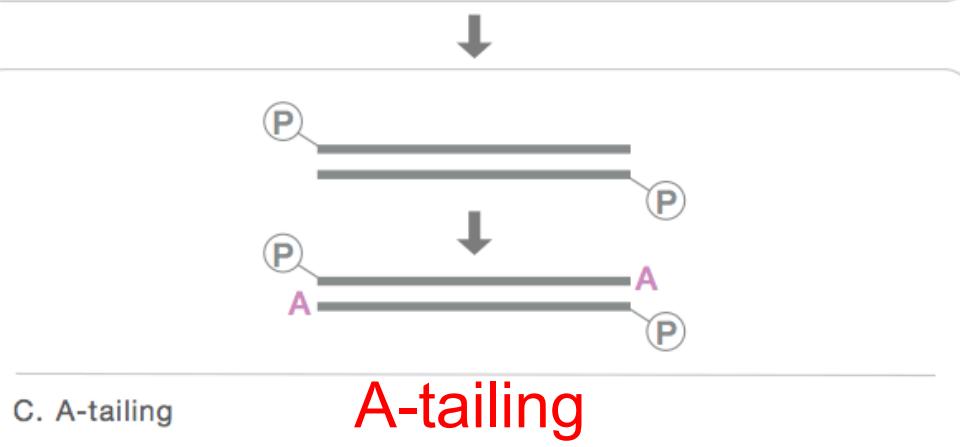
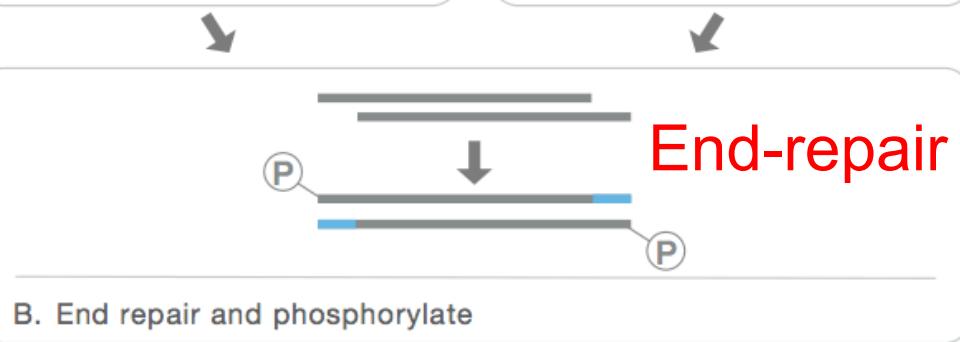
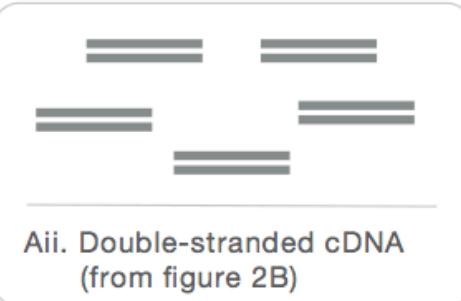
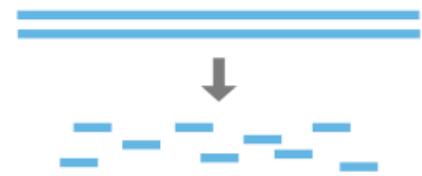
Platform	Templates	Signal	Read length	Run time	reads per run	Error type	Error rate
Illumina Miseq	PCR or PCR-free	fluorescent	up to 2x300	1-2 days	Up to 10 Gb	substitutions	~0.1-1%
Illumina Hiseq	PCR or PCR-free	fluorescent	up to 2x250	days	Hundreds of Gb	substitutions	~0.1-1%
Ion Torrent	PCR	H+	300-500	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

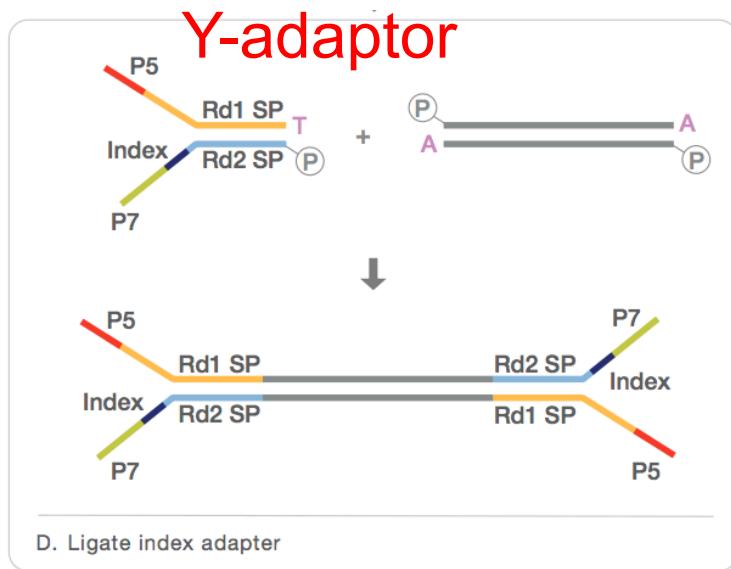


# Library preparation – Y-adaptor method

## a. Fragmentation



b.



PCR or PCR-free  
Final product

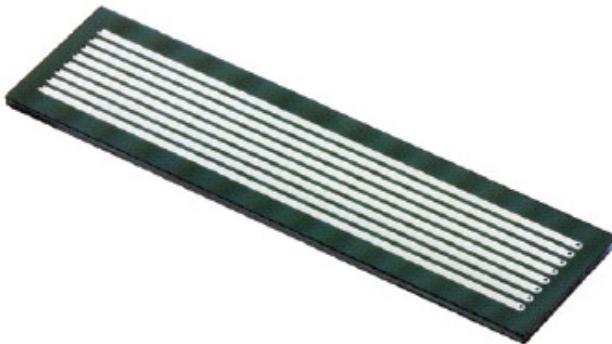


E. Denature and amplify for final product

From TruSeq Manual

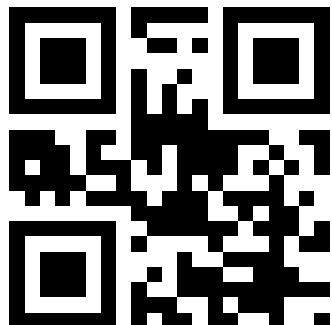
# Multiplexing (DNA barcode/Index)

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

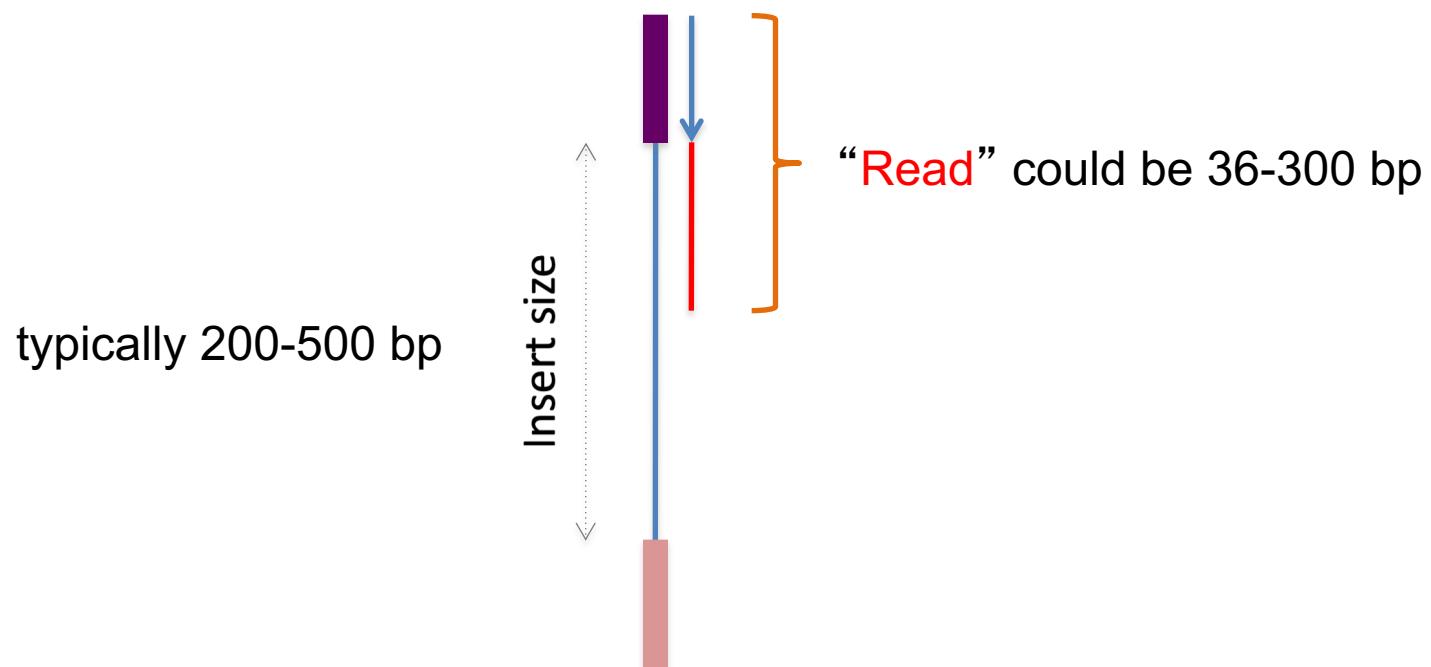
# Barcode / Index



	AGTGCAxxxxxxxxxxxx
sample 1	AGTGCAxxxxxxxxxxxx
	AGTGCAxxxxxxxxxxxx
	CATGTGxxxxxxxxxxxx
sample 2	CATGTGxxxxxxxxxxxx
	CATGTGxxxxxxxxxxxx

# Single-end sequencing

A single read is generated for each template/cluster



# Paired-end sequencing

Two reads are generated for each template cluster;  
the 1<sup>st</sup> is from one end with one primer;  
the 2<sup>nd</sup> 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