

# Next-gen Sequencing Technologies

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

2/2/2017

# R

- Data structure

Vector: numeric, character, logical, ...

Data frame: 2-dimension table

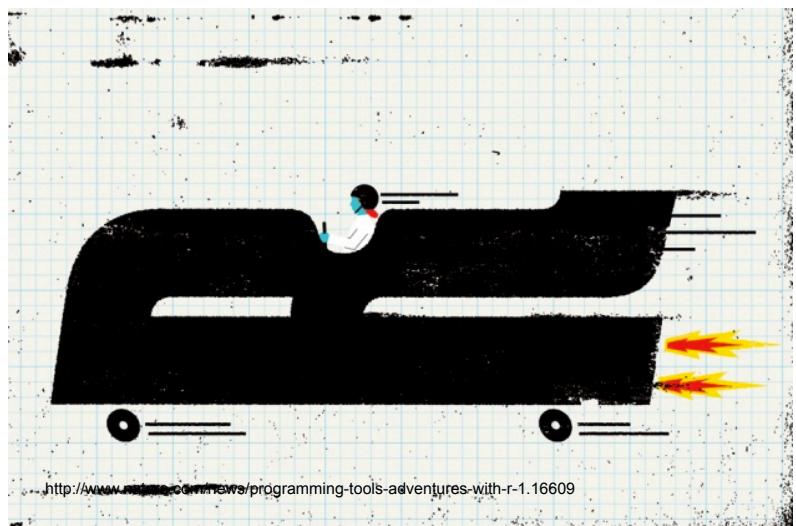
List: flexible for different types of data

- Data importing and exporting

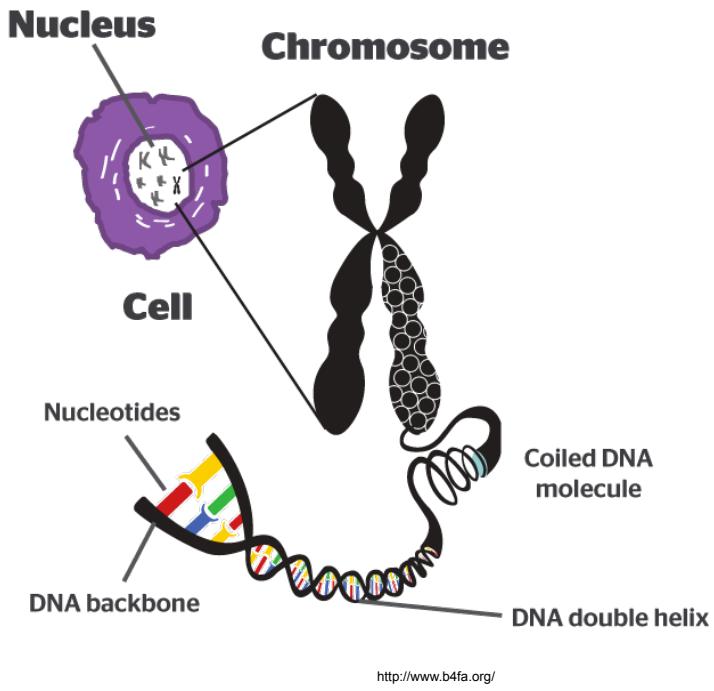
read.table, write.table

- Plotting:

plot, points, lines, abline



# Genome sequencing



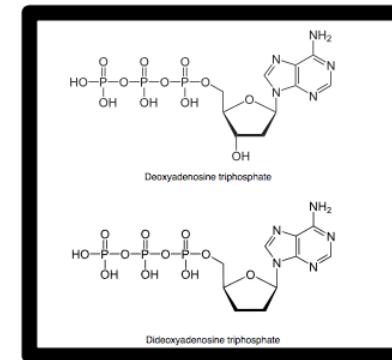
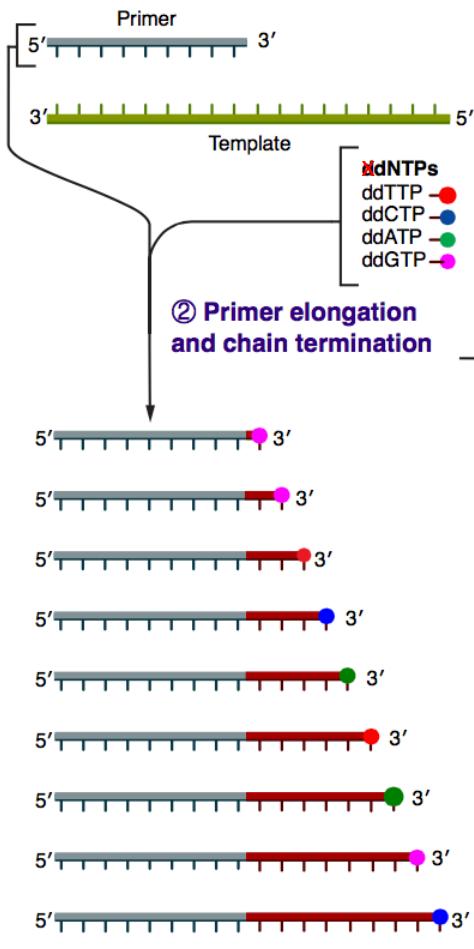
## DNA sequence

1 ccctgaggct tttcgagcg agctcctcaa atcgcatcca gatttcggg tccgaggaa  
61 ggaggaccct gcgaaagctg cgacgactat ctccccctgg ggccatggac tcggacgcca  
121 gcctgggtgc cagccgccc tcgtcgccag agcccgatga ccttttctg cccgccccgga  
181 gtaagggcag cagcggcagc gccttactg ggggcaccgt gtcctcggtcc accccgagtg  
241 actgcccggcc gagctgagc gcccagctgc gggcgctat gggctctgctg ggcgcgcac  
301 ctggggacaa cttaggaggc agtggctca agtcgtcctc gtccagcacc tcgtcgctca  
361 cgtcgctggc ggctgcgtcg tccaccaaga aggacaagaa gcaaatgaca gagccggagc  
421 tgcagcagct gcgtctcaag atcaacagcc gcgagcgc当地 ggcgcac gacctaaca  
481 tcgcatggc tggcctccgc gaggtcatgc cgtaacgcaca cggcccttcg gtgcgaagg  
541 ttccaaagat cgccacgctg ctgtggcgc gcaactacat cctcatgctc accaactcgc  
601 tggaggagat gaagcgactg gtgagcgaga tctacggggg ccaccacgct ggcttccacc  
661 cgtcgccctg cggcggccctg ggcactccg cggcccttcg cggccgc当地 ggcacccgg  
721 cagcagcagc gcaacgc当地 catcaccccg cggtgacca ccccatctc cggccggccg  
781 cccgacggc tgctgccgc gctgcagccg cggctgtgc cagcgcctct ctgcccggat  
841 ccgggctgcc gtcggcggc tccatccgtc caccgcacgg cctactcaag tctccgtctg  
901 ctgcgcggc cggccggcgt ggggggggg ggggggggg ggggggggg gggggcttcc  
961 agcaactgggg cggcatgccc tgccccgtca gcatgtgcca ggtggccccc cggcaccacc  
1021 acgtgtcgcc tatgggcgcc ggcagctgc cgcgcctcac ctccgacgccc aagtgagccg  
1081 actggcgccc ggcgttctg ggcacagggg agccagggg cggggggaaag cgaggactgg  
1141 cctgcgtgg gtcgggagc tctgtcgca ggagggggc当地 aggaccatgg actgggggt  
1201 gggcatgggtg gggattccag catctgc当地 cccaaagcaat gggggccccc acagagc  
1261 gggggatgtgag gggatgttot ctccggacc tgatcgagcg ctgtctggct ttaacctgag  
1321 ctgggtccagt agacatcgat ttataaaaag gtaccgtgt gtgcattccct cactagaact

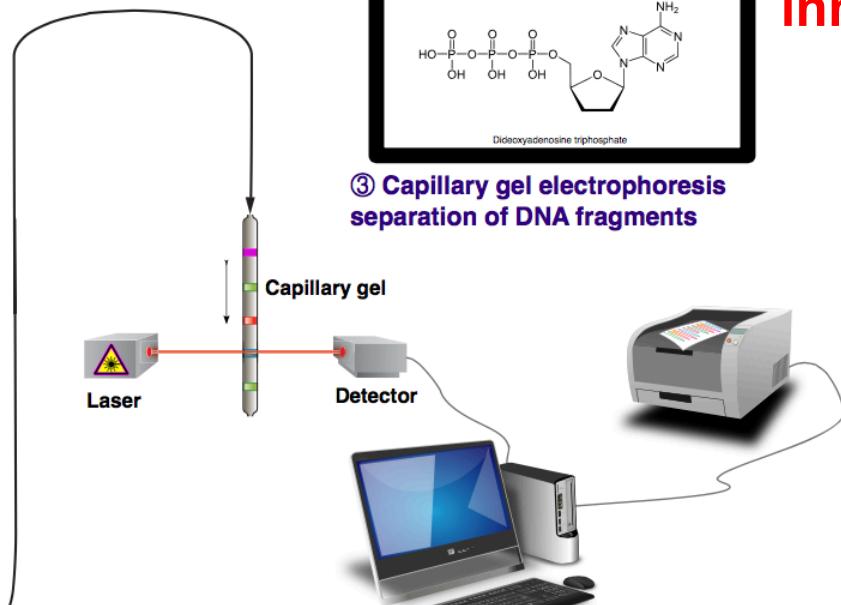
# Sanger sequencing technology

## ① Reaction mixture

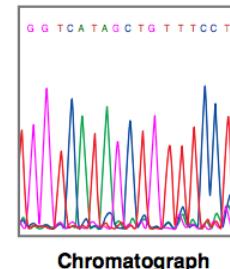
- Primer and DNA template → DNA polymerase
- ddNTPs with flourochromes → dNTPs (dATP, dCTP, dGTP, and dTTP)



Key innovation



## ③ Capillary gel electrophoresis separation of DNA fragments



## ④ Laser detection of flourochromes and computational sequence analysis

wikipedia

# DNA sequencing technology

1st-gen sequence  
(Sanger)

1980 Nobel Prize

```
>sample
TGCAGGCTACTAACCGGTTCTGAGAGTTCTGAGATG
AGAGAATGCCACTAACCGGTTCTGAGAAATGCCA
CTAACCGATGCCACTAACCGGTTCTGAGAGTTCTG
AGCTGAGATGCCACTAACCGGTTCTGAGAAATGCC
ACTAACCGATGCCACTAACCGGTTCTGAGAGTTCTG
AGATGCCACTAACCGGTTCTGAGAGTTCTGAGAA
TGCCTACTAACCGATGCCACTAACCGGTTCTGAGA
GTTCTGAGCTATGCCACTAACCGGTTCTGAGAAATG
CCACTAACCGATGCCACTAACCGGTTCTGAGAGT
TCTGAGATGAGAGAAATGCCACTAACCGGTTCTGA
GAATGCCACTAACCGCTACTAACCGGTTCTGAGA
ATGCCCTACTAACCGATGCCACTAACCGGTTCTGAG
AGTCTGAGATGAGAGAAATGCCACTAACCGGTTCTG
GAGAAATGCCACTAACCGATGCCACTAACCGGTTCTG
AGAGGTTCTGAGCTCCGATGCCACTAACCGGTTCTG
TGAGAGTTCTGAGCTATAACTAACCGGTTCTGCTA
CTAACCGGTTCTGAGAAATGCCACTAACCGGTTCTG
AGAAATGCCACTAACCGATGCCACTAACCGGTTCTG
GAGAGTTCTGAGATGAGAGAAATGCCACTAACCGG
TTCTGAGAAATGCCACTAACCGATGCCACTAACCG
GTTCTGAGAGTTCTGAGCTGAGAA
```

800 letters

next-gen sequence (NGS)



billions of letters

# Major NGS technologies in market

**454** LIFE SCIENCES

life  
technologies™

ion torrent  
△ \* ○ × □ + ≈



PACIFIC  
BIOSCIENCES®



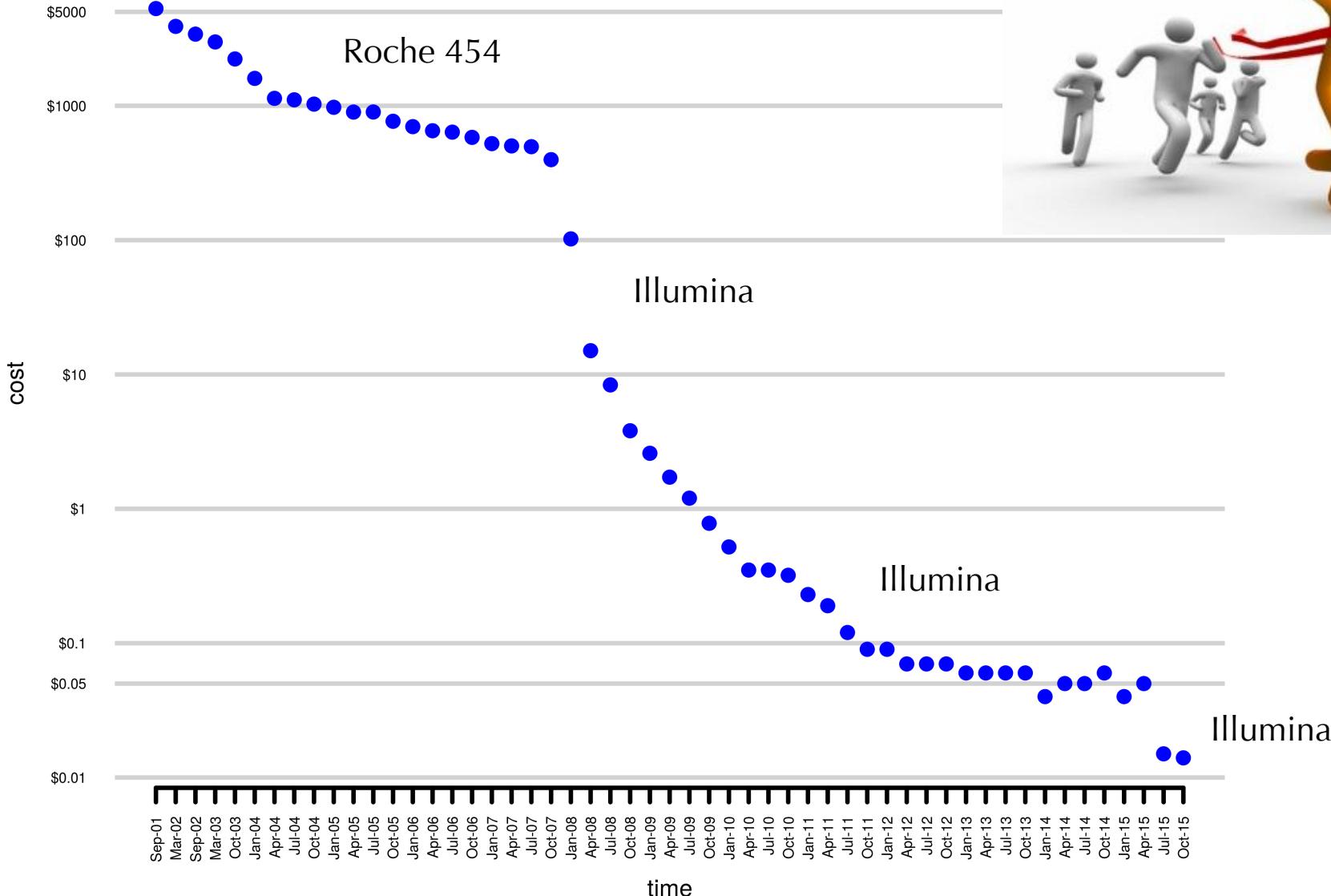
# Sequencing cost

cost per megabase

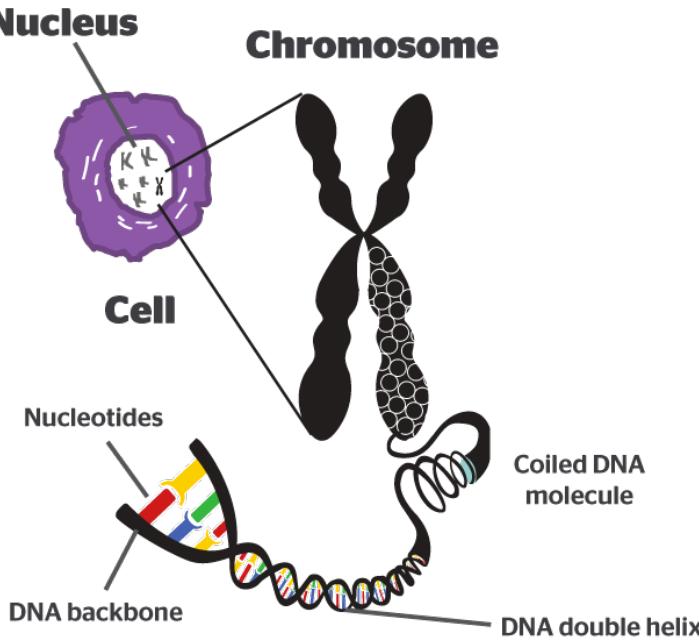
1970's Sanger sequencing

Roche 454

Illumina



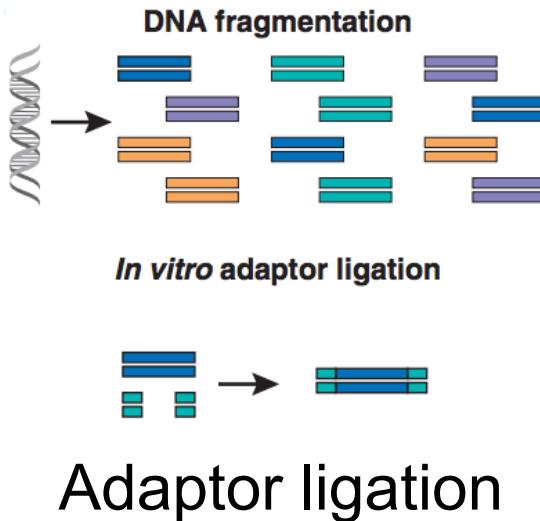
Data source: genome.gov/sequencingcosts



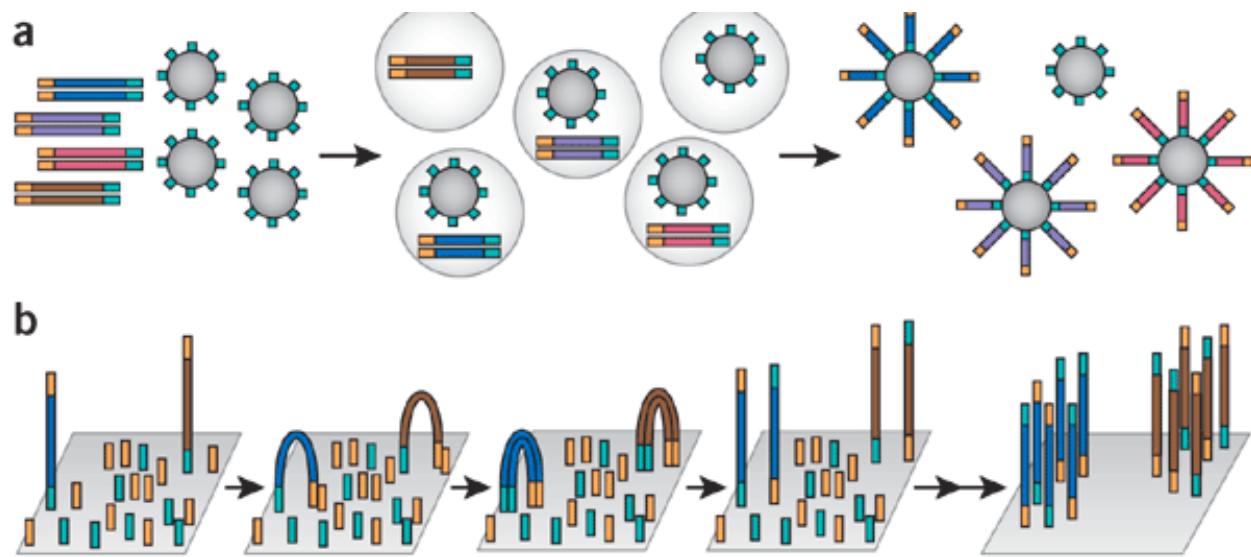
## sequencing sensitivity and read length

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

# DNA amplification/cloning

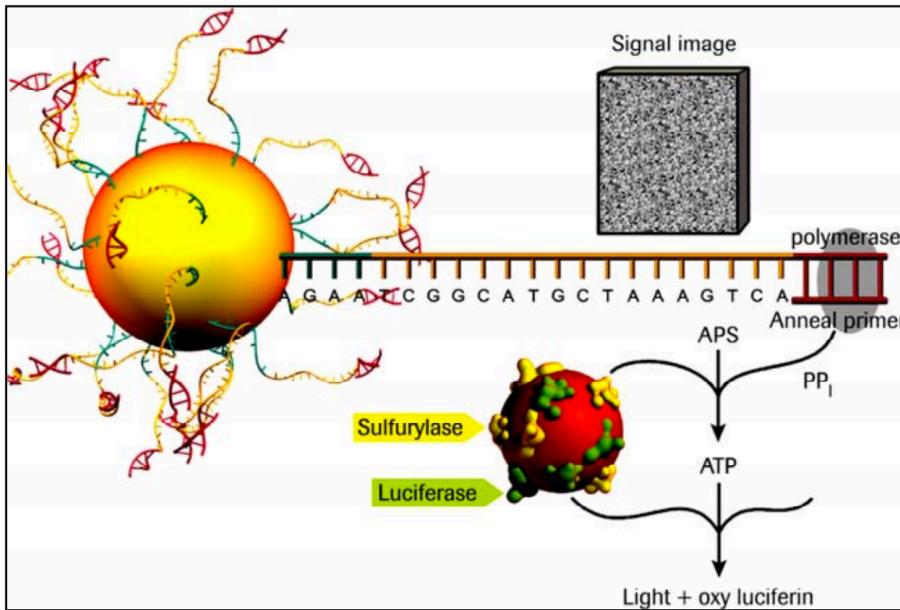


Water-in-oil emulsion PCR  
(454 and Ion Torrent)



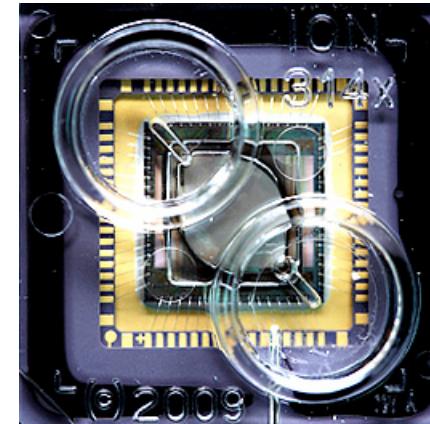
Bridge PCR on slides  
(Illumina)

# 454 and Ion Torrent



454 technology, Nature 2005, 437: 376-380

1. Sequencing by synthesis
2. Pyrosequencing (454)



[Ion Torrent video](#)

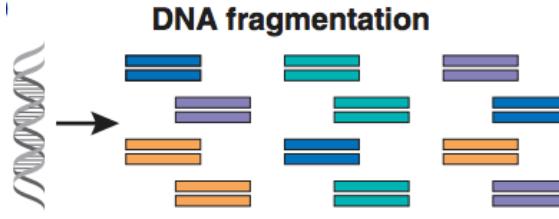
1. Ion Torrent technology is similar to 454 technology
2. The signal is H<sup>+</sup> rather than pyrophosphate

# Ion Torrent & 454

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

A T G C A A A A

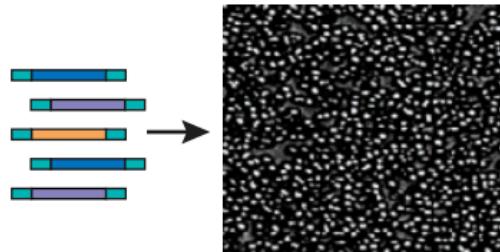
A T G C A A A A



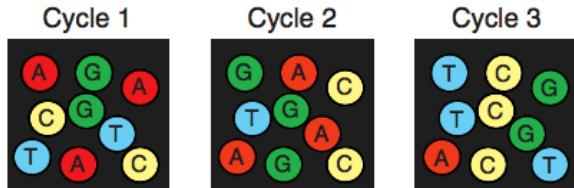
### *In vitro* adaptor ligation



### Generation of polony array



### Cyclic array sequencing ( $>10^6$ reads/array)



# Illumina sequencing

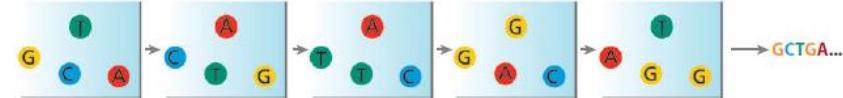
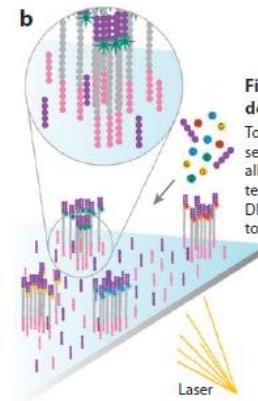
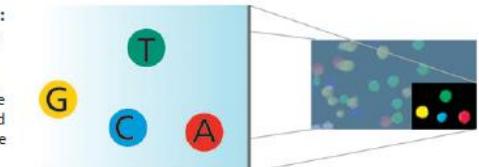


Image of first chemistry cycle  
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.



[philos.biol.mun.ca](http://philos.biol.mun.ca)

## Two key technologies:

1. Bridge PCR
2. Reversible terminator chemistry

# 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

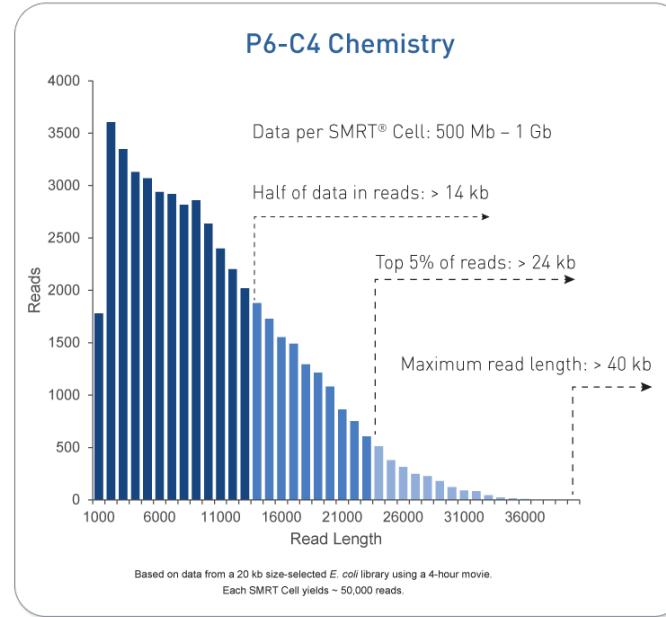
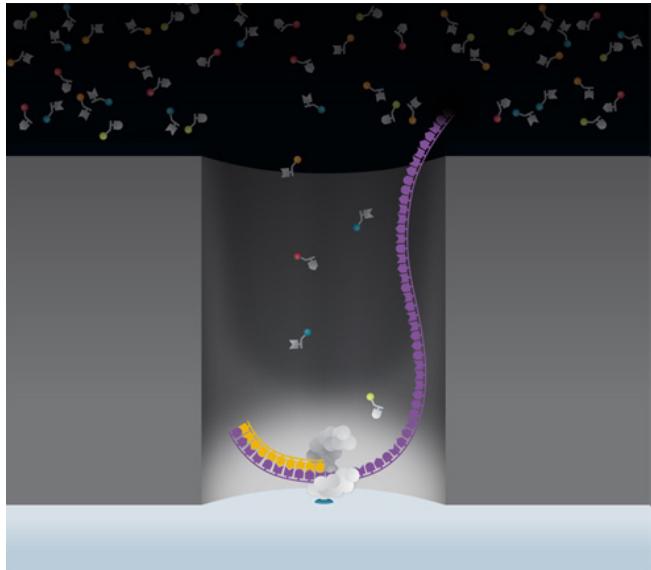
Life technology Ion Torrent & Roche 454:  
Record signal per **nucleotide type**:

A	T	G	C	A	A	A	A
---	---	---	---	---	---	---	---

**Sequencing errors at homopolymers**

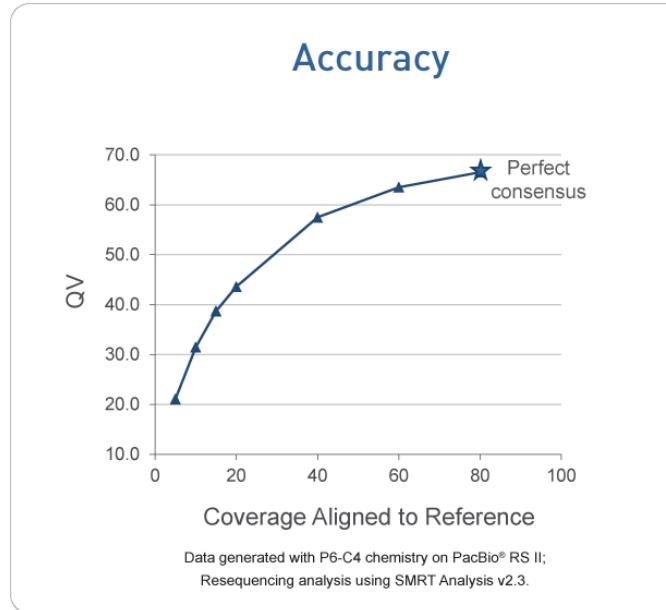
When the single molecular sequencing technology is ready, **amplification or cloning** is not necessary.

# PacBio – Single Molecule Real Time (SMRT)

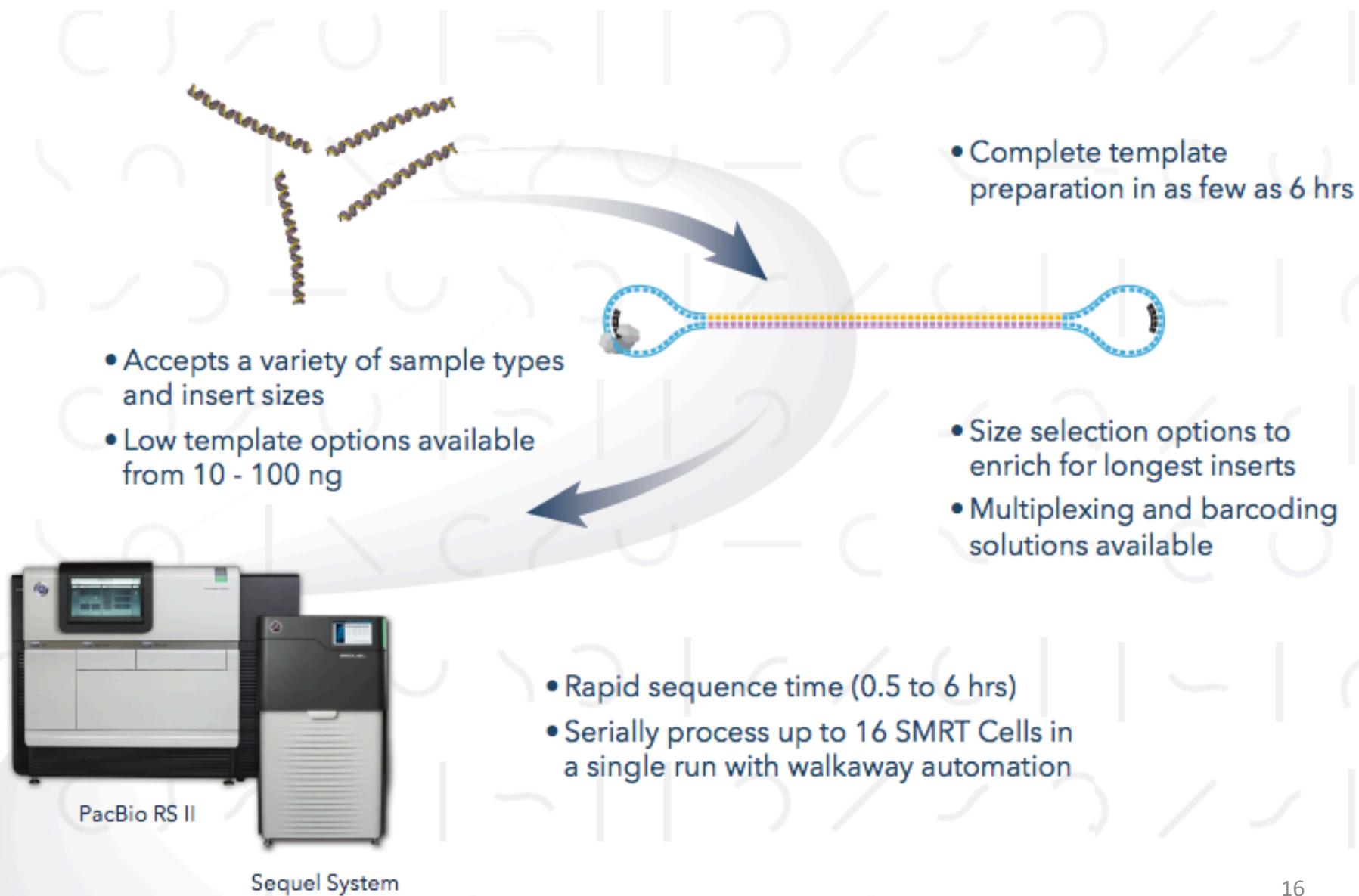


## PacBio tech video

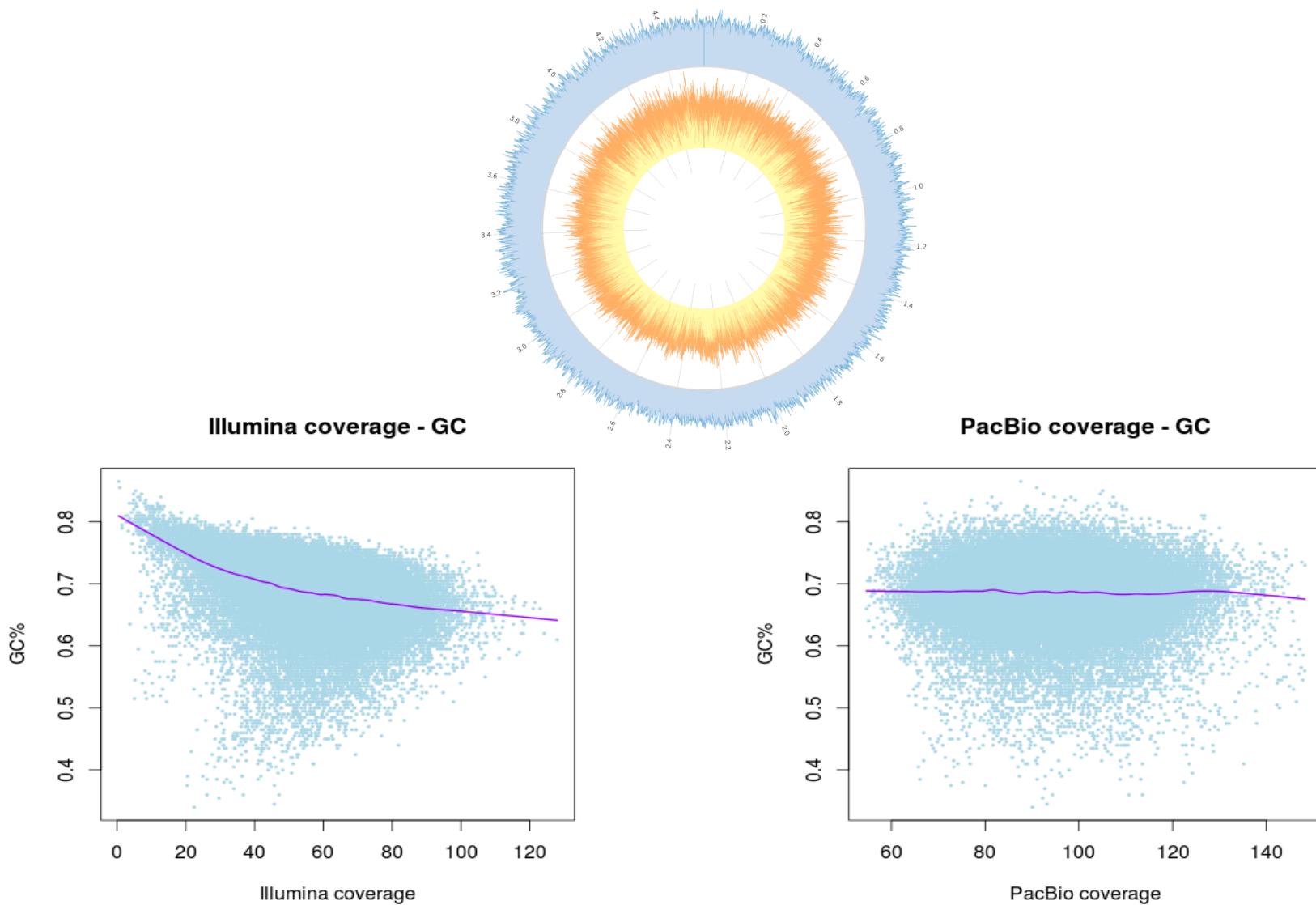
- Single molecule sequencing
- no amplifications required
- up to 70+ kbp sequencing
- Moderate sequencing throughput
- high sequencing error rate (~15%, random, no-context-specific errors)



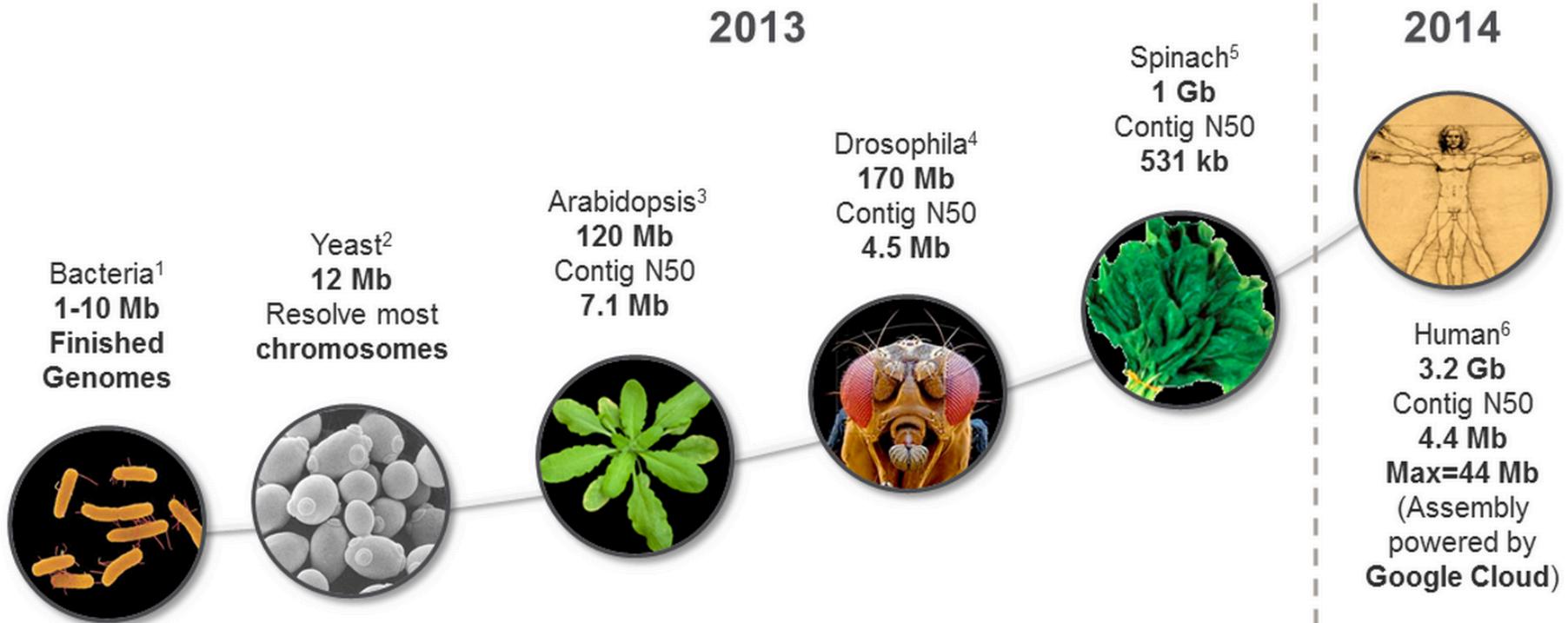
# PacBio Sequencing procedure



# Less GC-related biases



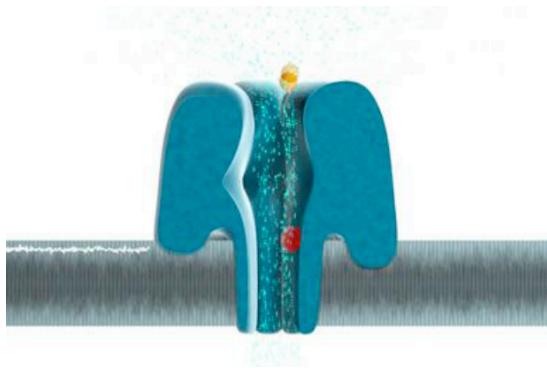
# PacBio for genome assembly



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.

# New Platform – 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 (kbp)**
- **Error rate is high (~30%)**

## MinION

1. USB disposable sequencer
2. Hundreds of Mb in several hours



## PromethION (NEW)

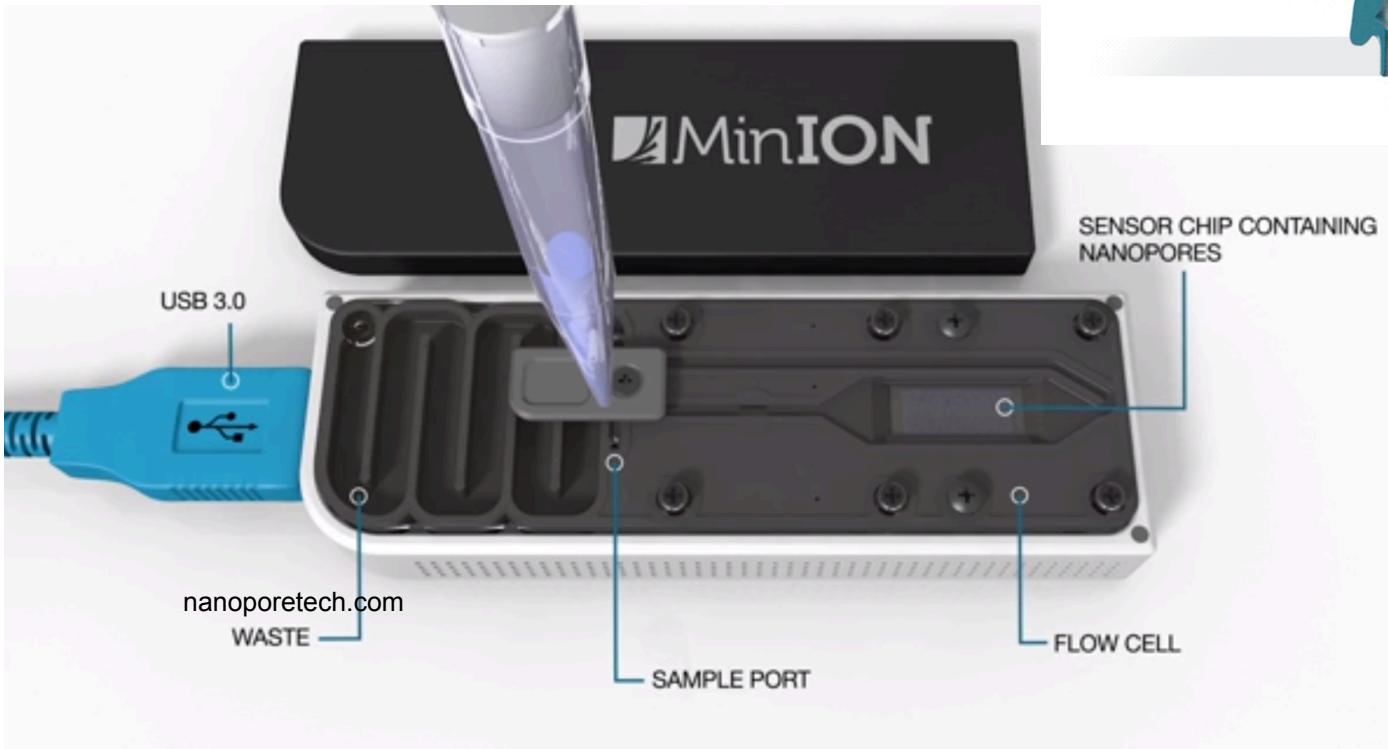
1. “MinION cluster”?
2. High-throughput (1Tb output)



# MinION specification (1/31/2017)

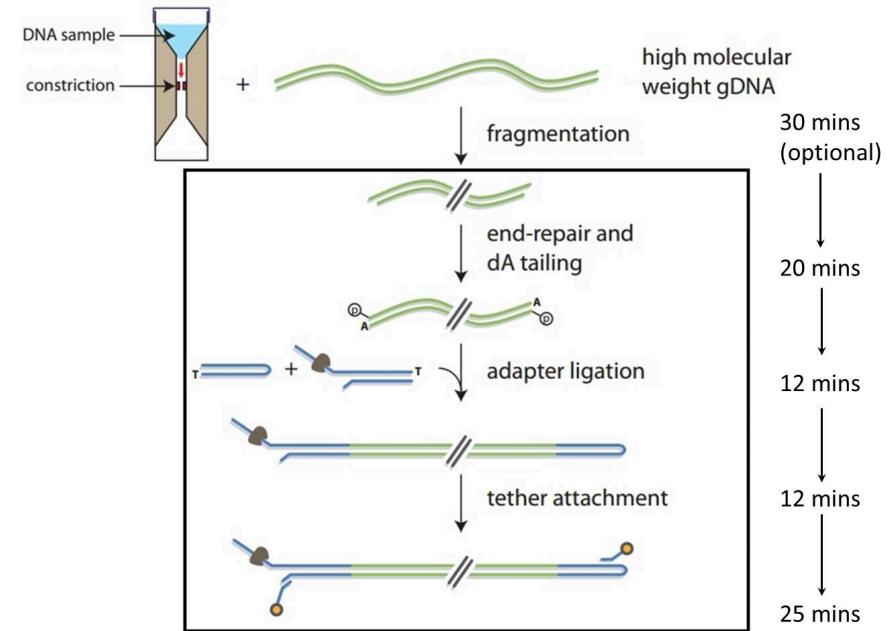
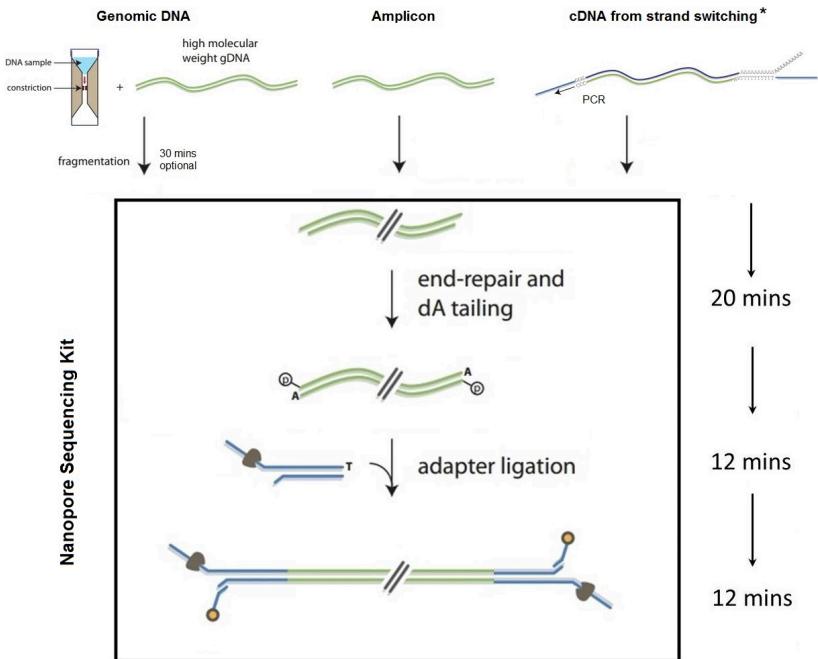
Item	MinION
Number of channels available for sequencing	Up to 512
Sample input Requirement PCR Free	10pg - 1µg
Sample preparation time 1D	10 minutes
Sample preparation time 2D	90 minutes
Run time	1 minute - 48 hours
Number of reads at 10Kb at standard speed (250bps)	Up to 2.2M
Read length	up to hundreds of kb
Flow Cell Cost (depending on order type and volume)	\$500 - \$900

# Potentials for Nanopore



smartphone  
+  
Nanopore  
(SmidgION)

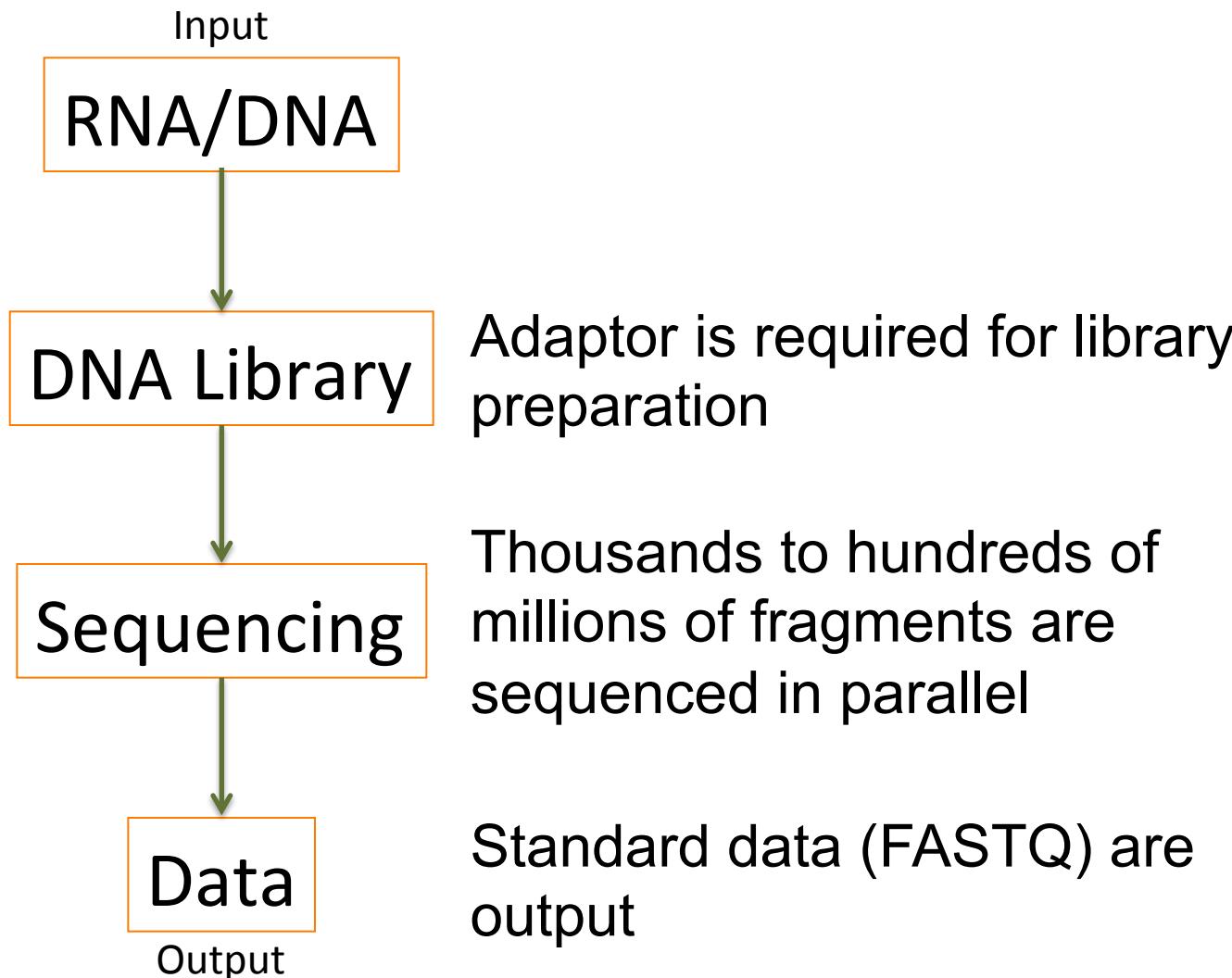
# Nanopore library preparation



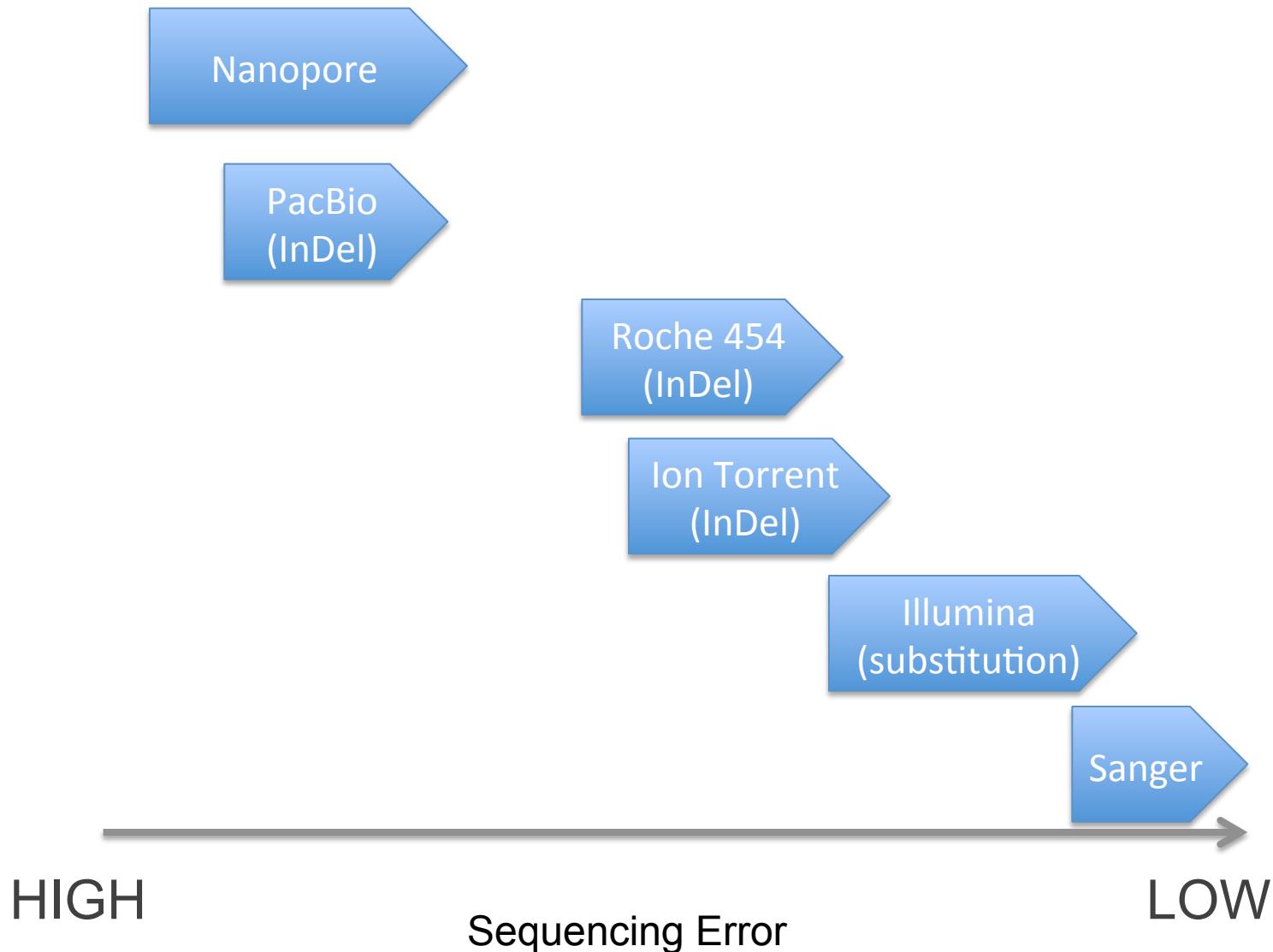
1D

2D

# COMMON in all NGS platforms



# Sequencing error rates



# 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 (up to 70kb+), but high errors (15%)

## Nanopore

Moderate sequencing throughput, very long (up to 300kb), but high errors (10-20%)



@anne\_churchland (twitter)

# Experimental design

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

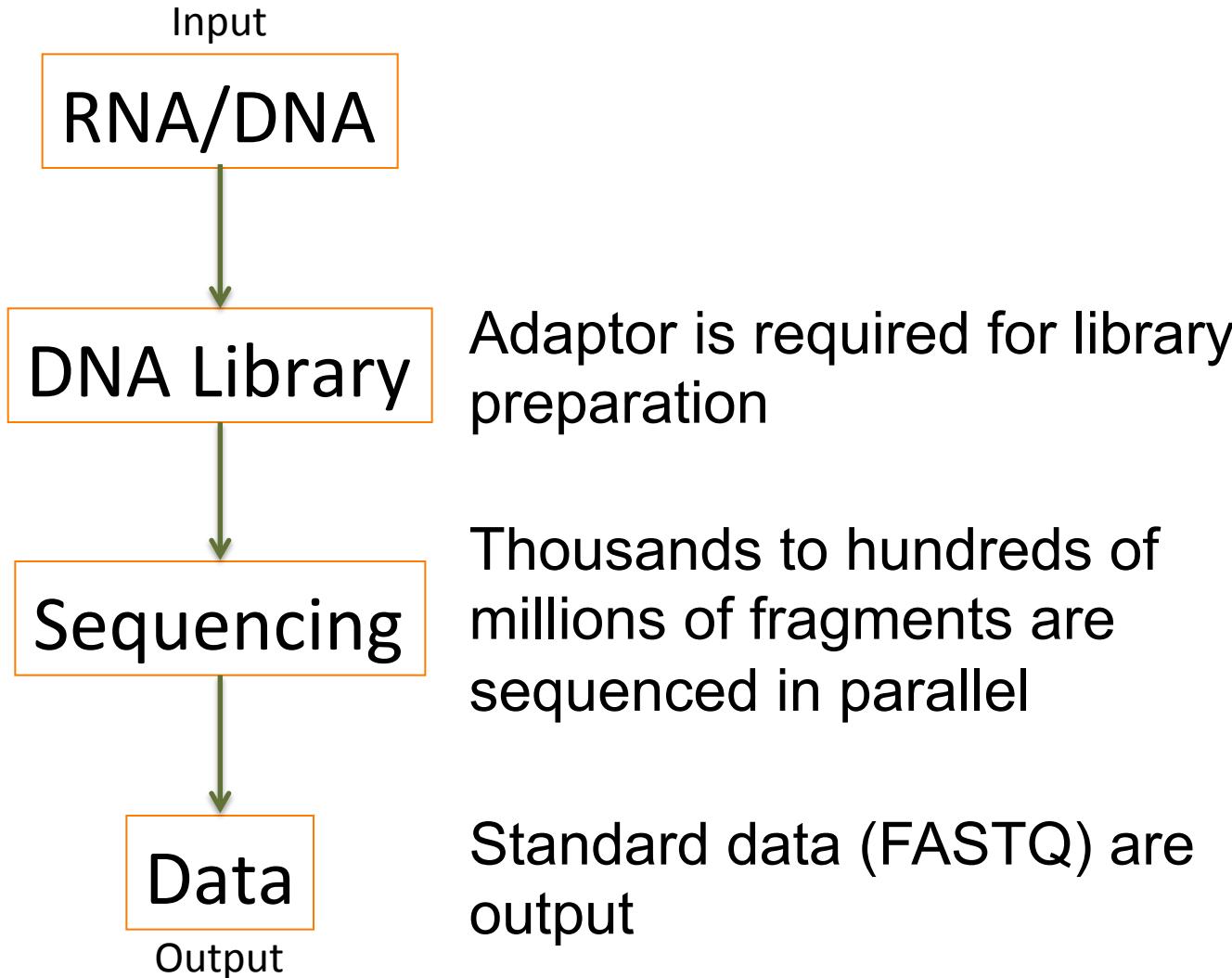
Platform	Templates	Signal	Read length	Run time	reads per run	Error type	Error rate
Illumina Miseq	PCR or PCR-free	fluorescent	up to 2x300	1-2 days	Up to 10 Gb	substitutions	~0.1-1%
Illumina Hiseq	PCR or PCR-free	fluorescent	up to 2x250	days	Hundreds of Gb	substitutions	~0.1-1%
Ion Torrent	PCR	H+	300-500	2 hours	10 Gb?	InDel	>1%
PacBio	Amplification not required	fluorescent	Average >5,000	30min	500 Mb – 1 Gb	InDel	~15%
Nanopore	Amplification not required	Electronic flow change	>1,000	hours	? Mb per MinION	Del?	~10-20%

# Illumina platforms and terminologies

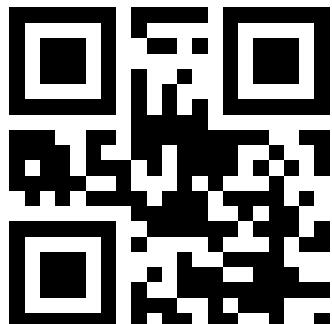
[Illumina video](#)

1. Library preparation
2. Single-ends and paired ends
3. Reads
4. Instruments

# COMMON in all NGS platforms



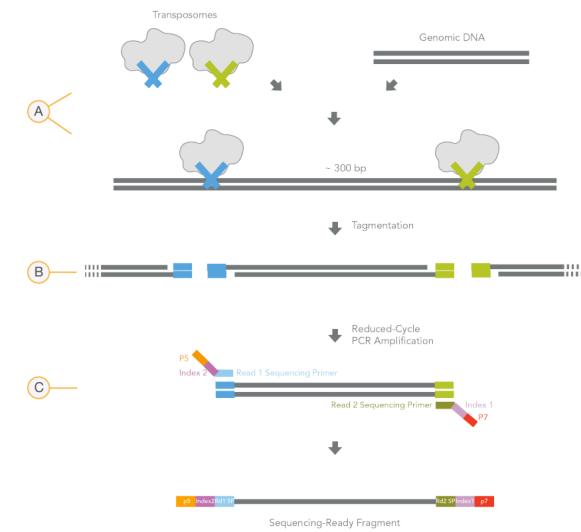
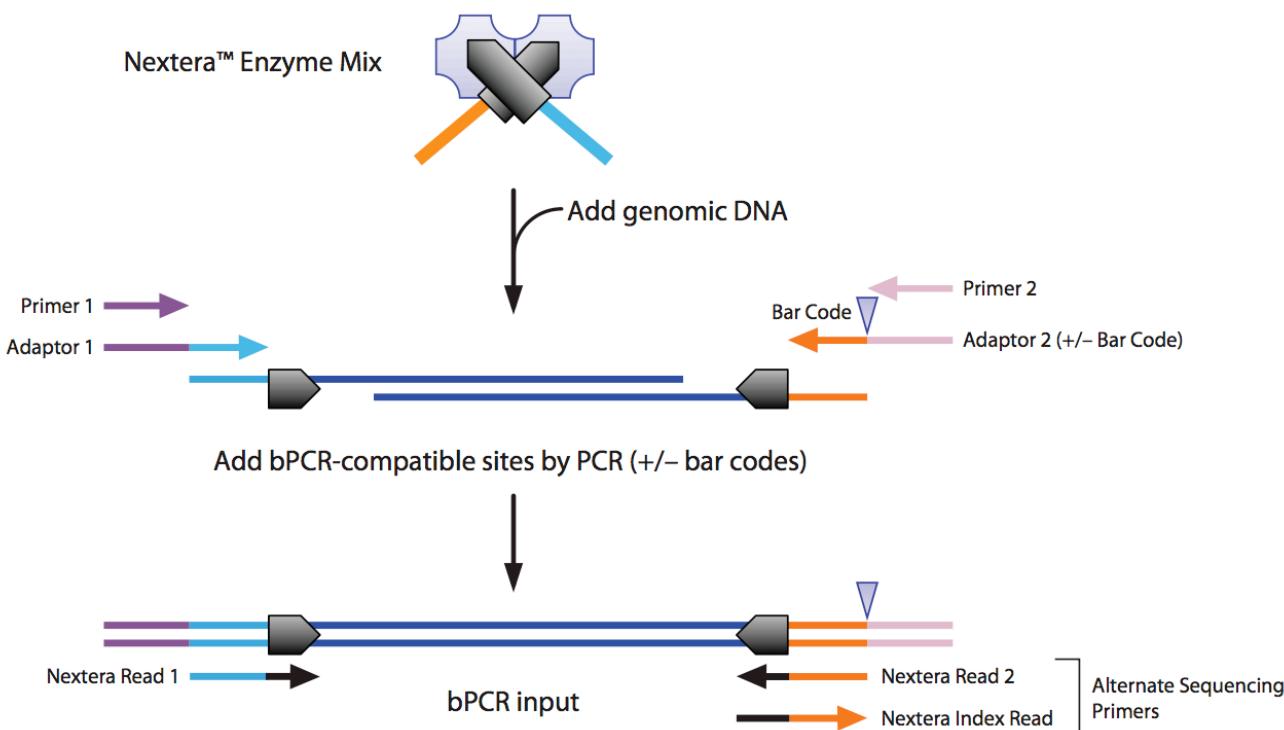
# Barcode / Index



	AGTGCAxxxxxxxxxxxx
sample 1	AGTGCAxxxxxxxxxxxx
	AGTGCAxxxxxxxxxxxx
	CATGTGCAxxxxxxxxxxxx
sample 2	CATGTGCAxxxxxxxxxxxx
	CATGTGCAxxxxxxxxxxxx

# Library preparation - Nextera

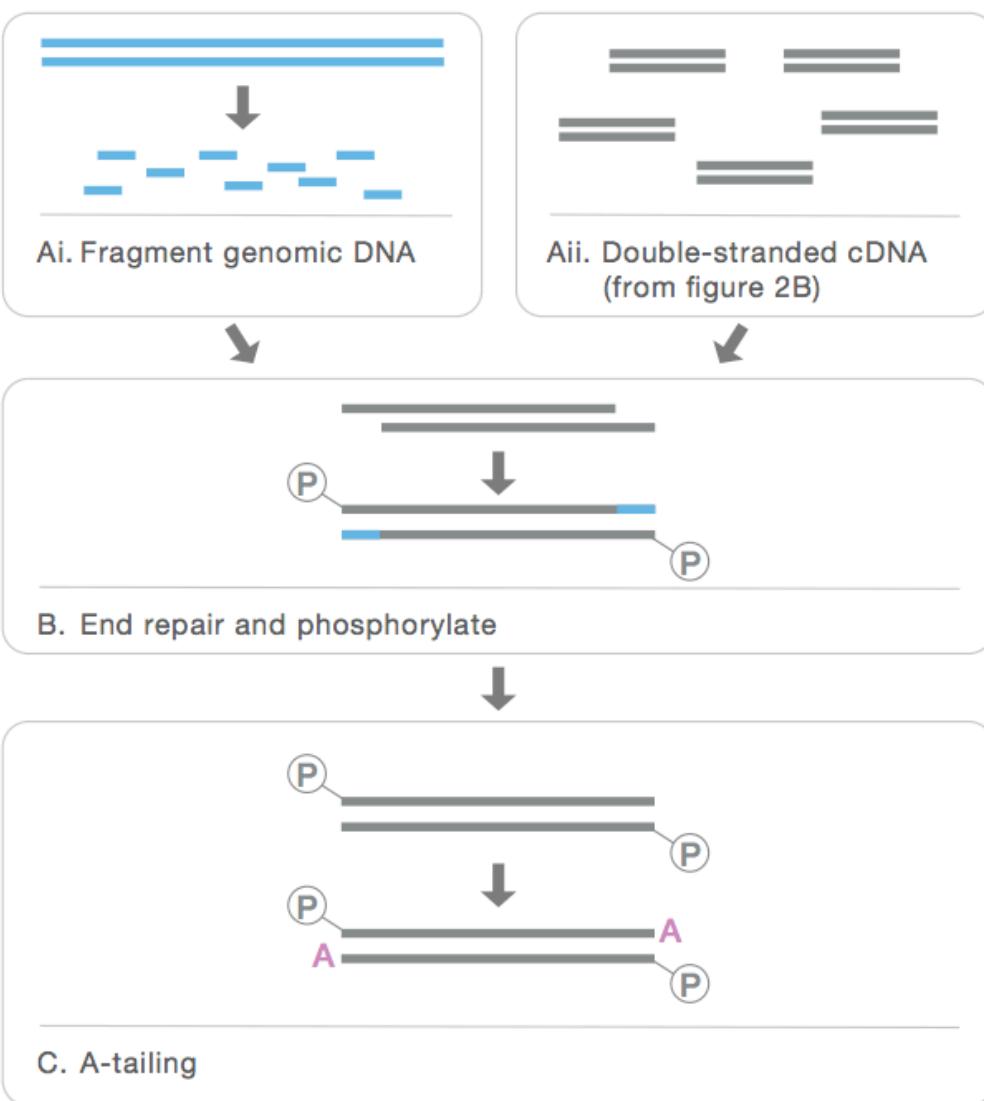
Nextera technology employs in vitro transposition to simultaneously fragment and tag DNA in a single-tube reaction, and prepare sequencer-ready libraries in under 2 hours.



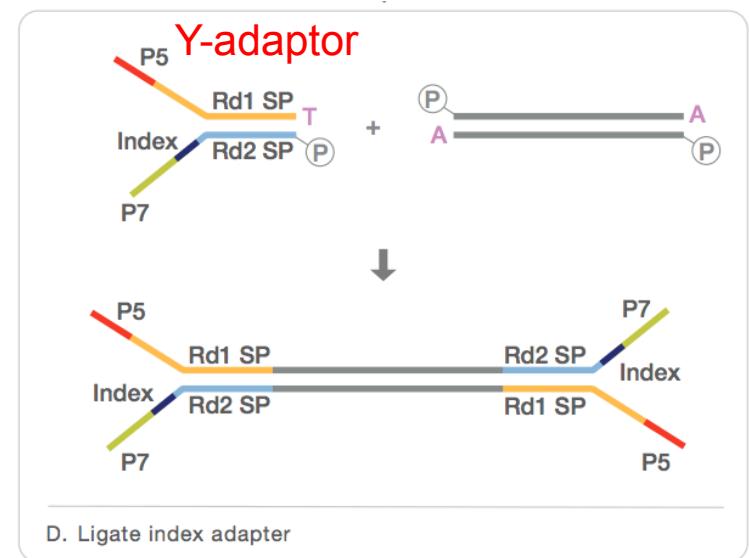
- A Nextera DNA transposome with adapters combined with template DNA  
B Tagmentation to fragment and add adapters  
C Limited cycle PCR to add index adapter sequences

# Library preparation – Y-adaptor method

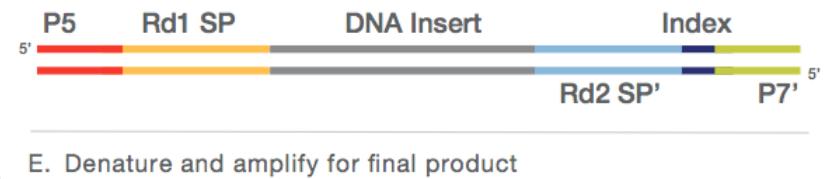
a.



b.

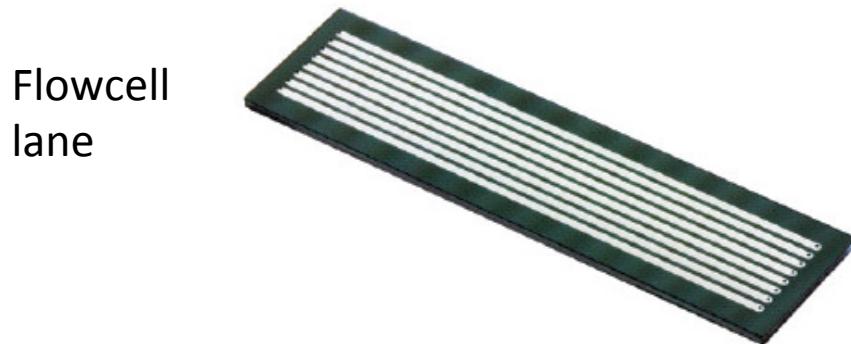


Final product



From TruSeq Manual

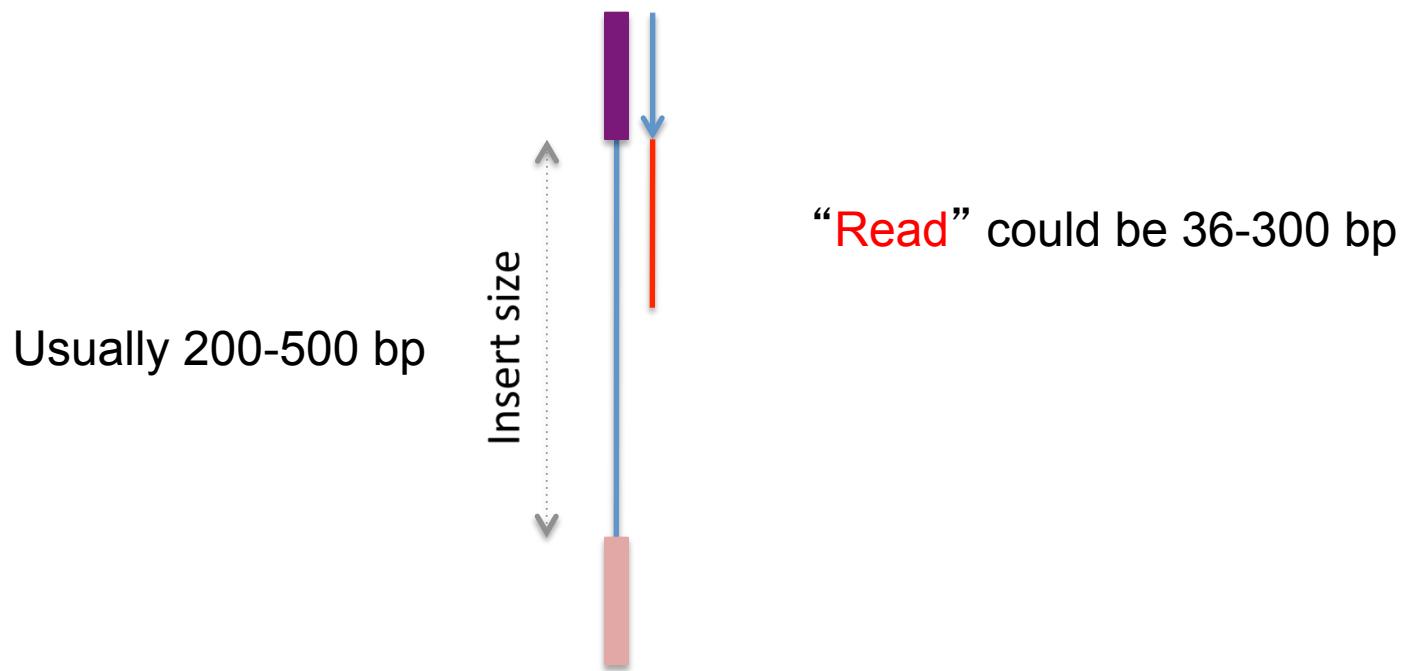
# Multiplexing (DNA barcode/Index)



- 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

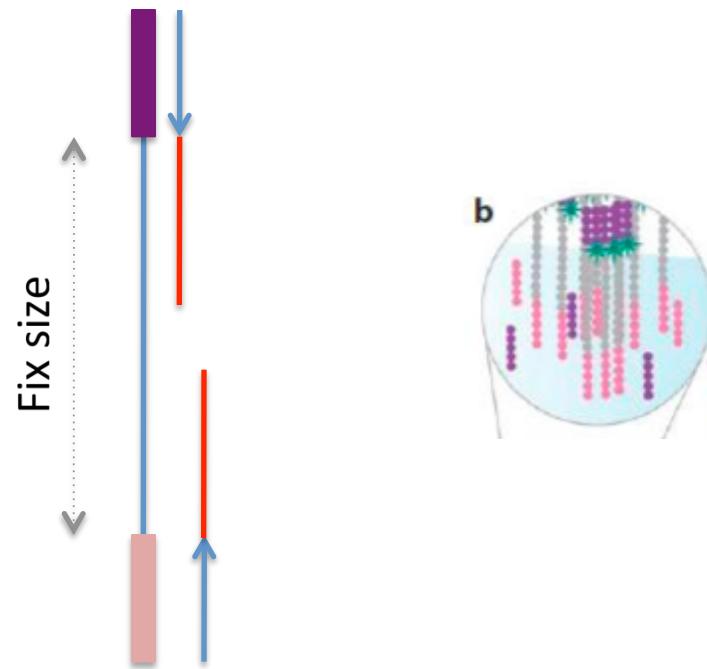
# Single-end sequencing

A single read is generated for each template/cluster

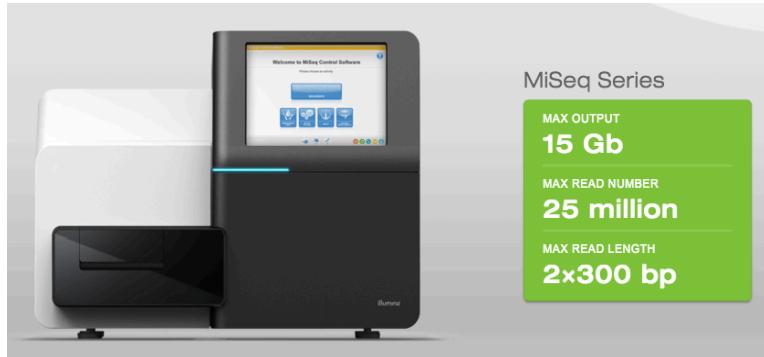


# Paired-end sequencing

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.

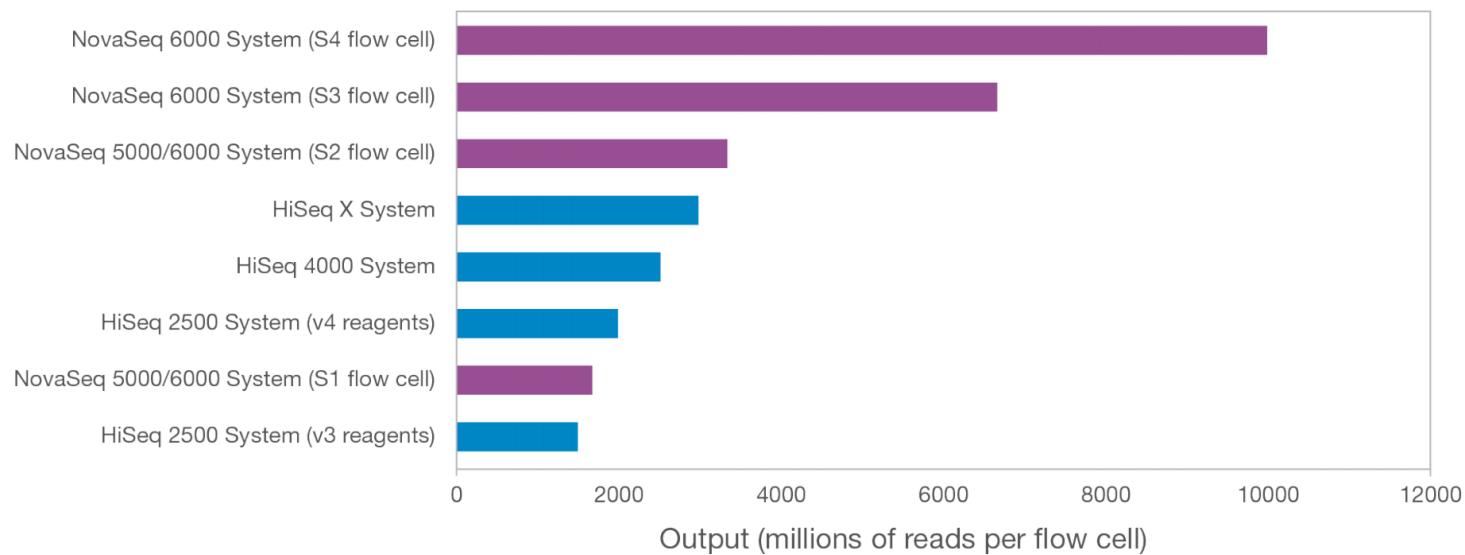


# Illumina Sequencers



**NovaSeq**

# NovaSeq



# NovaSeq5000 & 600 output (as of 1/31/2017)

## Sequencing Output per Flow Cell

	NovaSeq 5000 and 6000 Systems		NovaSeq 6000 System	
Flow Cell Type	S1*	S2	S3*	S4*
2 x 50 bp	up to 167 Gb	280–333 Gb	NA**	NA**
2 x 100 bp	up to 333 Gb	560–667 Gb	NA**	NA**
2 x 150 bp	up to 500 Gb	850–1000 Gb	up to 2000 Gb	up to 3000 Gb

Specifications based on Illumina PhiX control library at supported cluster densities.

\* The NovaSeq 5000 System, NovaSeq 5000 System Upgrade, and NovaSeq Reagent Kits with S1, S3, or S4 flow cells will be available later in 2017.

\*\* NA: not applicable

## Reads Passing Filter

	NovaSeq 5000 and 6000 Systems		NovaSeq 6000 System	
Flow Cell Type	S1*	S2	S3*	S4*
	up to 1.6 B	2.8–3.3 B	up to 6.6 B	up to 10 B