Whole-Genome Sequencing: Technology and Data

Bioinformatics Workshop for *M. tuberculosis*
Genomics and Phylogenomics

July 9-14, 2018 @The Philippine Genome Center

Ulas Karaoz, PhD

Ecology Department, Berkeley Lab

University of California San Francisco
advancing health worldwide

PHILIPPINE GENOME CENTER

BERKELEY LAB
Lawrence Berkeley National Laboratory
Bringing Science Solutions to the World

https://eesa.lbl.gov/profiles/ulas-karaoz, Email: ukaraoz@lbl.gov, Twitter: @ukaraoz

DNA Sequencing Technologies: Past and Present

Illumina MiSeq

Illumina HiSeq

454 Sequencer

SOLiD system

PacBio

Ion Proton

Nanopore

DNA Sequencing: Why do we care?

"... [A] knowledge of sequences could contribute much to our understanding of living matter."
Frederick Sanger, 1980.

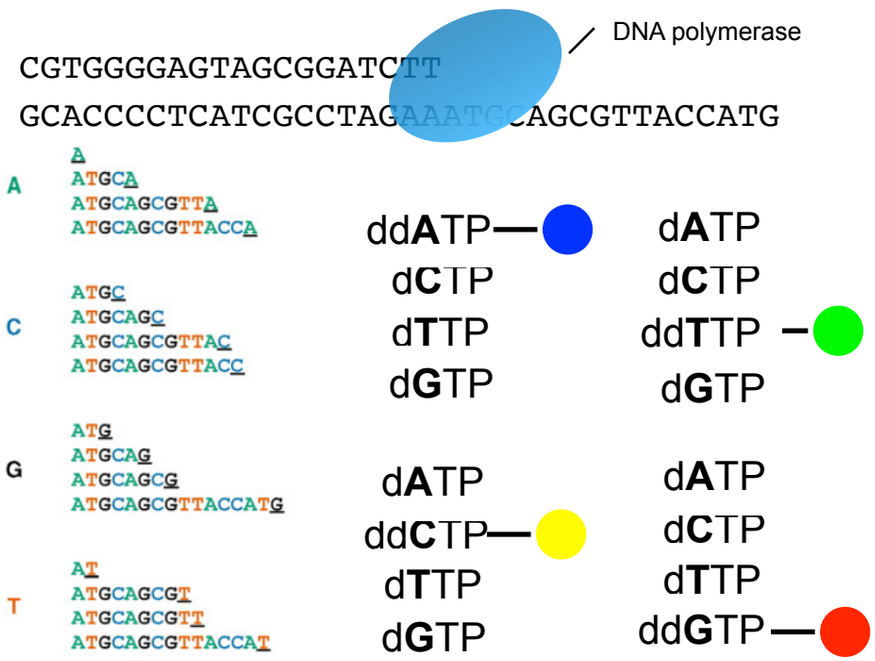
First Generation DNA Sequencing: Sanger dideoxy sequencing (~1975)

DNA Sequencing: Why do we care?

"... [A] knowledge of sequences could contribute much to our understanding of living matter."
Frederick Sanger, 1980.

First Generation DNA Sequencing: Sanger dideoxy sequencing (~1975)

I. DNA Synthesis with dideoxynucleotides



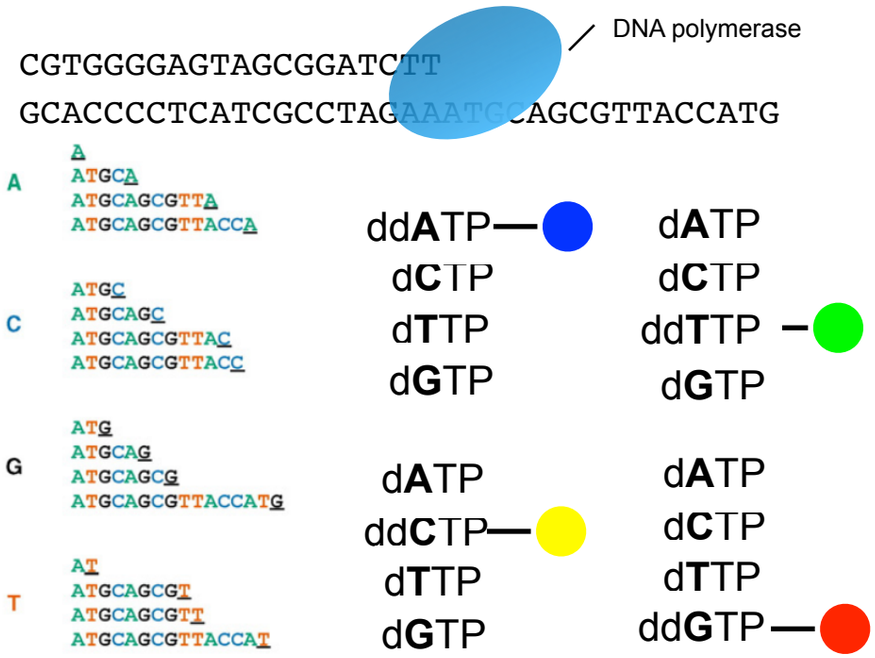
Heather JM. The sequence of sequencers: The history of sequencing DNA. Genomics 2016.

DNA Sequencing: Why do we care?

"... [A] knowledge of sequences could contribute much to our understanding of living matter."
Frederick Sanger, 1980.

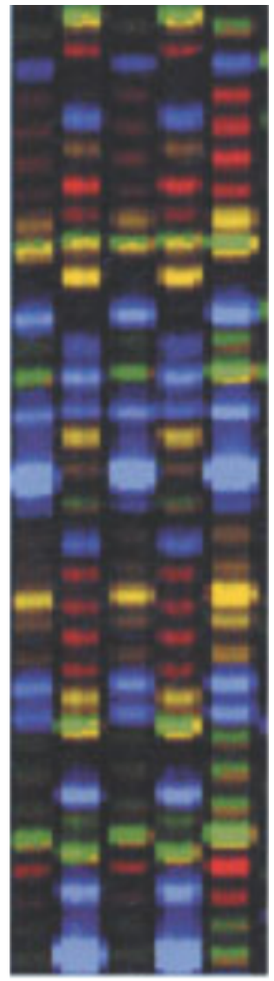
First Generation DNA Sequencing: Sanger dideoxy sequencing (~1975)

I. DNA Synthesis with dideoxynucleotides



Heather JM. The sequence of sequencers: The history of sequencing DNA. Genomics 2016.

II. Electrophoresis

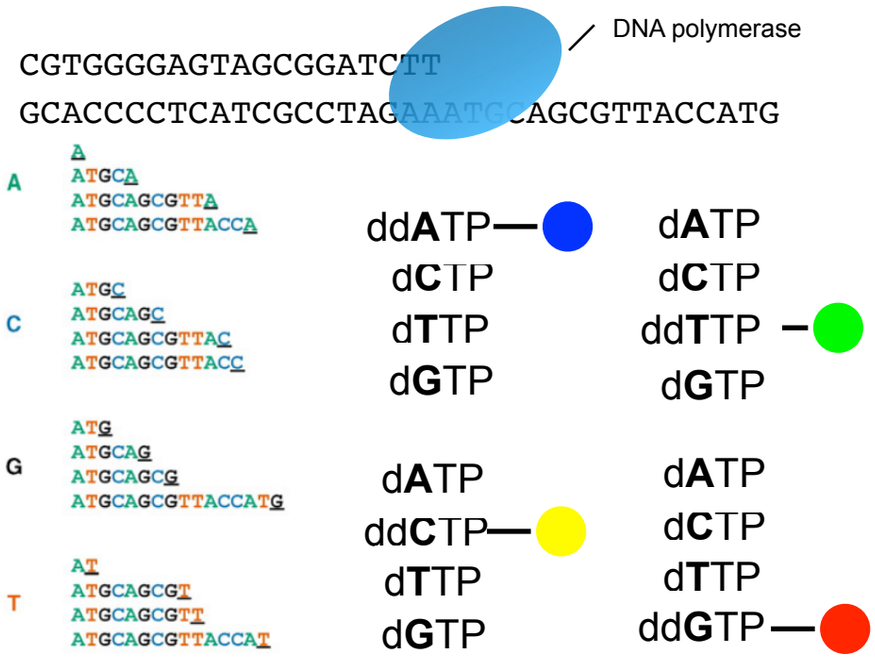


DNA Sequencing: Why do we care?

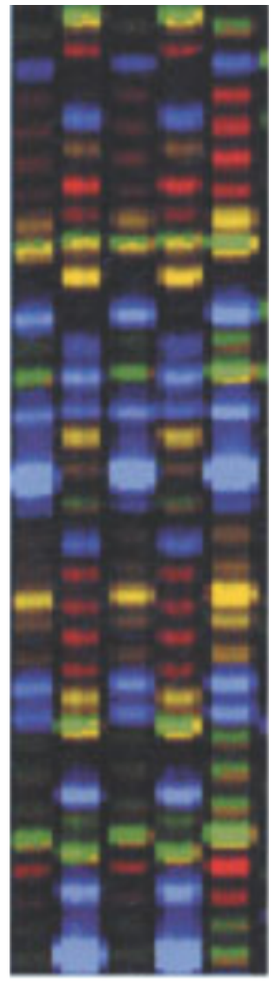
"... [A] knowledge of sequences could contribute much to our understanding of living matter."
Frederick Sanger, 1980.

First Generation DNA Sequencing: Sanger dideoxy sequencing (~1975)

I. DNA Synthesis with dideoxynucleotides

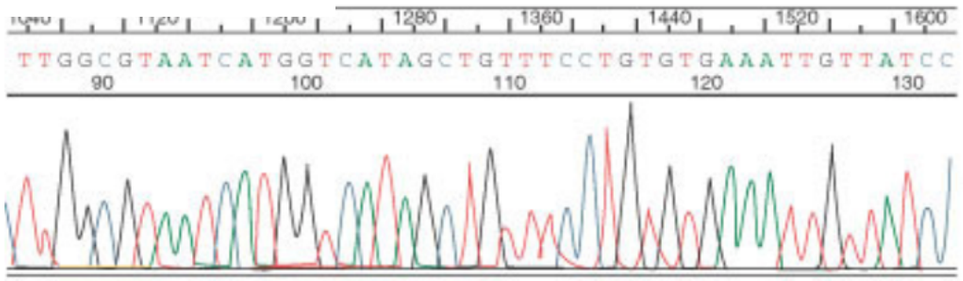


II. Electrophoresis



Heather JM. The sequence of sequencers: The history of sequencing DNA. Genomics 2016.

III. Electropherogram



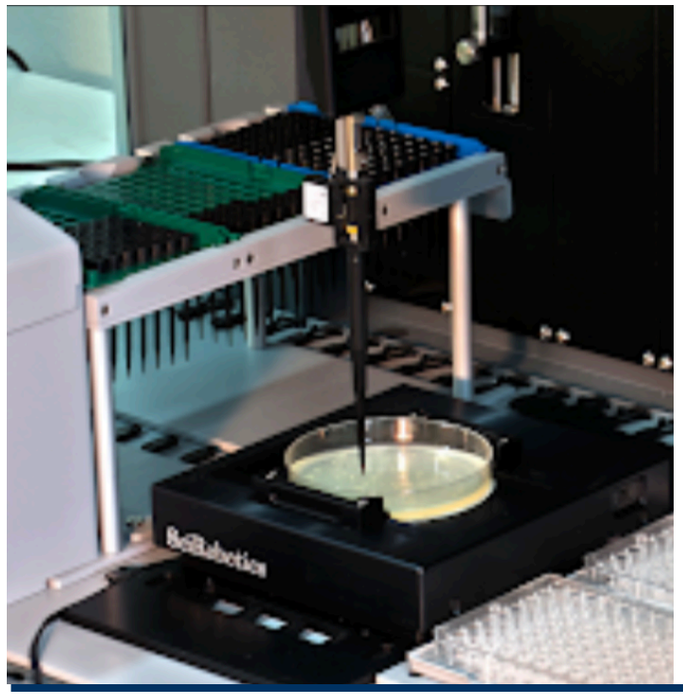
Automation of Sanger Sequencing

ABI 3730xl: 96/384 well capillary system

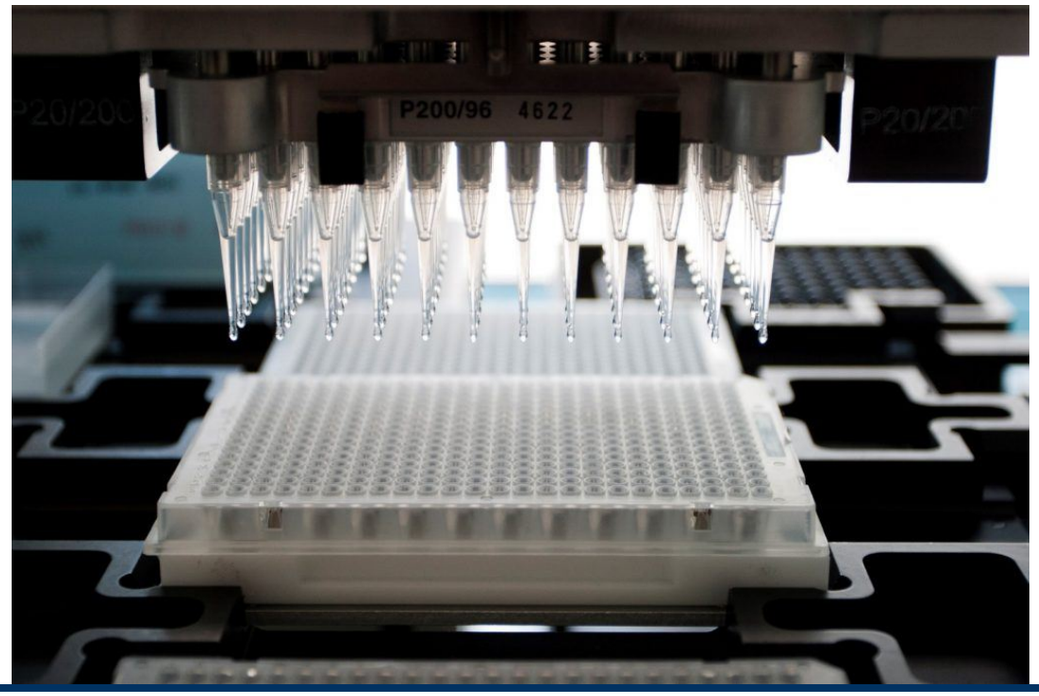
2001

ABI Sequencers, Venter Institute

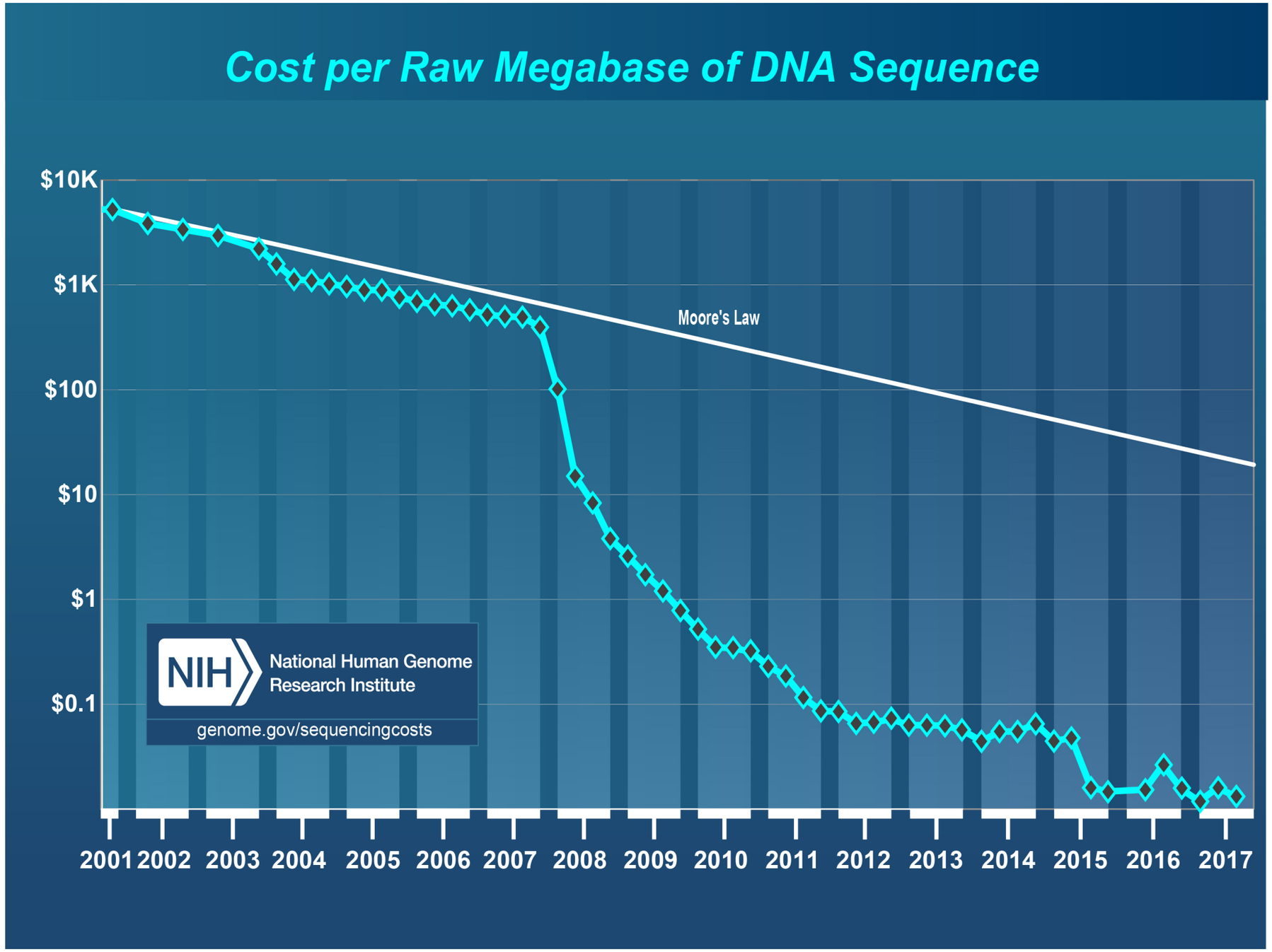
Automated colony picking



Automated plating



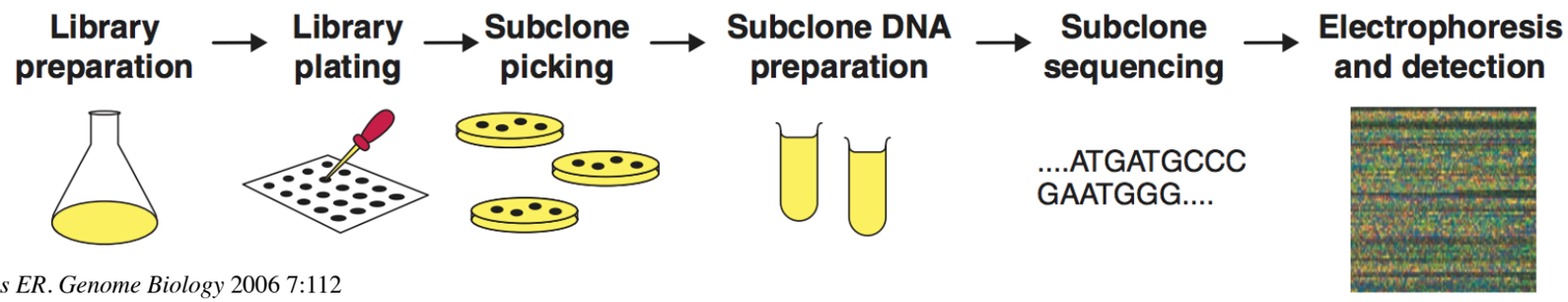
2004-2007: from *colonies* to *clusters*



Next-Generation Sequencing (NGS)

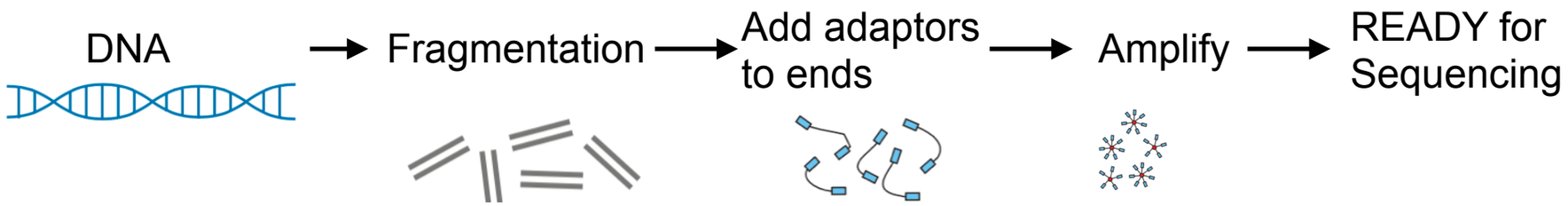
Next-generation Sequencing = **Second-generation Sequencing** = **Massively Parallel Sequencing**

Conventional Sequencing



Mardis ER. Genome Biology 2006 7:112

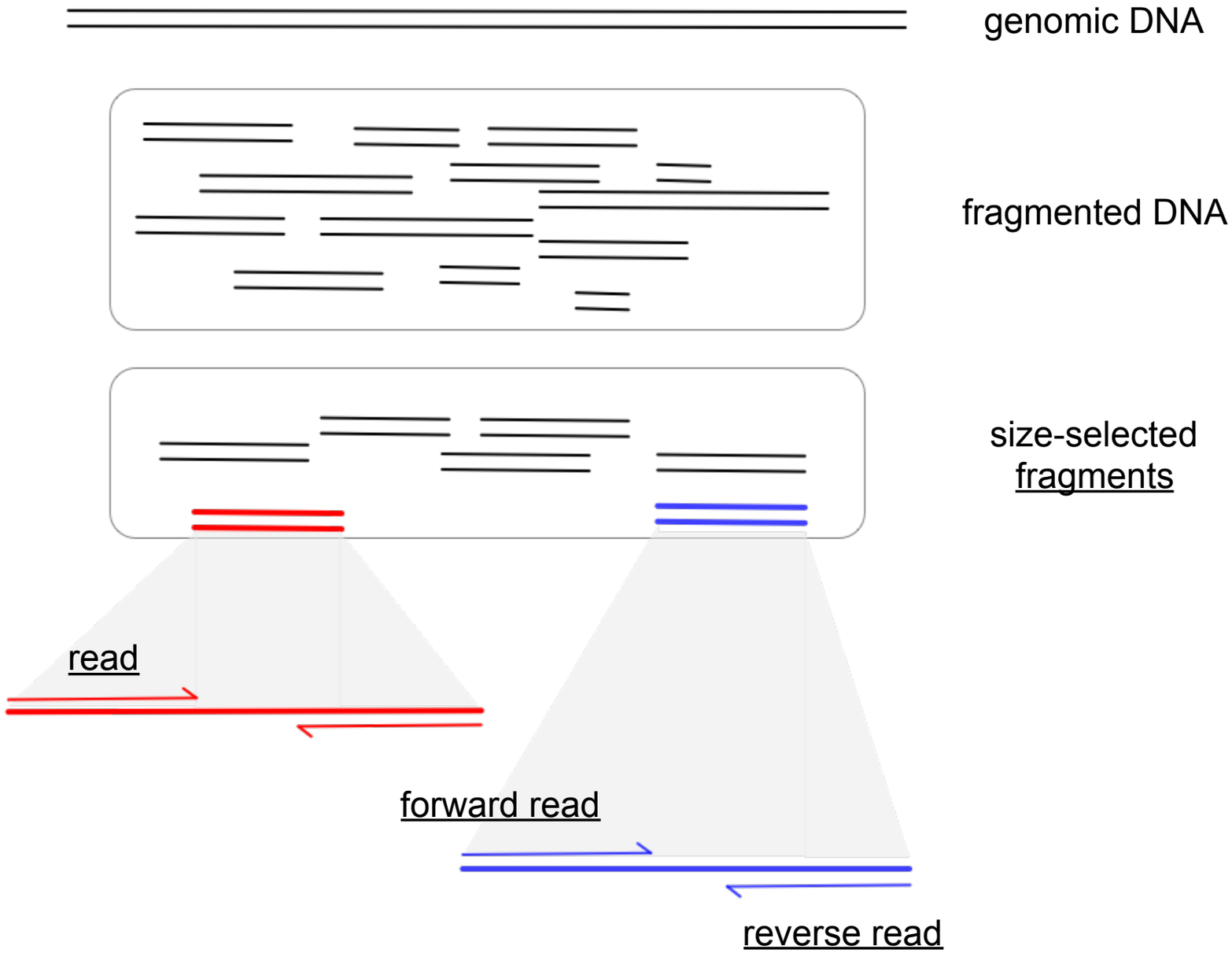
NGS



NGS Commonalities

- Sequencing by synthesis: coupling of molecular biology and detection
- Library construction: easier, faster, cheaper
- randomly fragmented DNA + "adapter" sequences (platform-specific)
- Amplification needed before sequencing

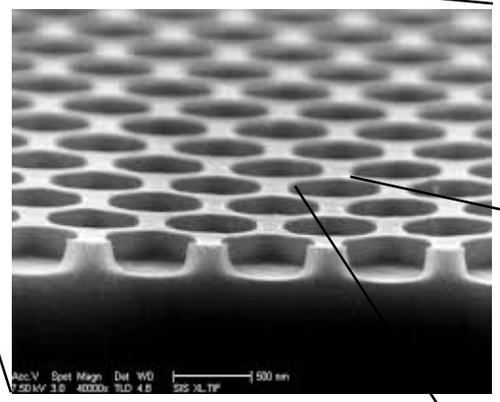
Next-Generation Sequencing (NGS)



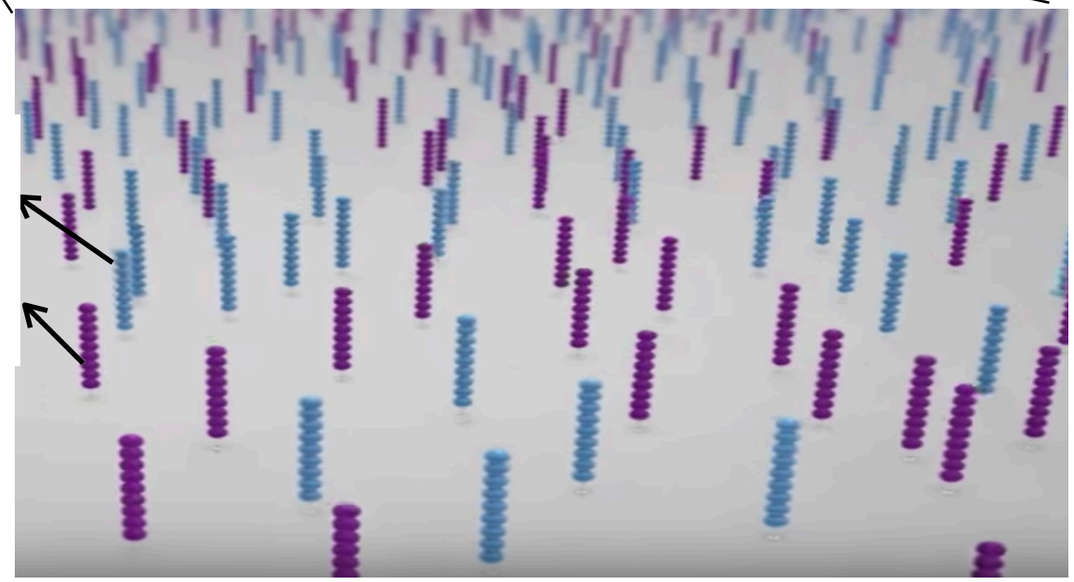
Illumina Sequencing: Flow cell

Illumina flow cell: glass slide where sequencing chemistry occurs

lane

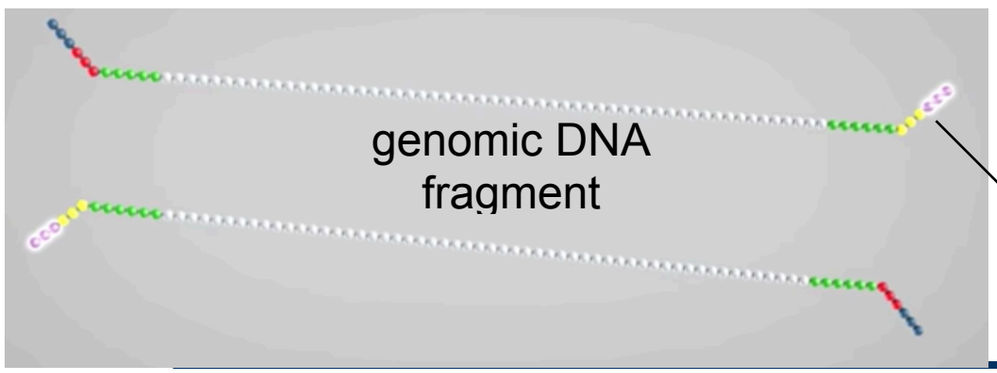


"Lawn" of oligos on the flow cell surface



2 types of flow cell oligos

DNA fragments with adapters complementary to these oligos at ends



region complementary to flow cell oligo

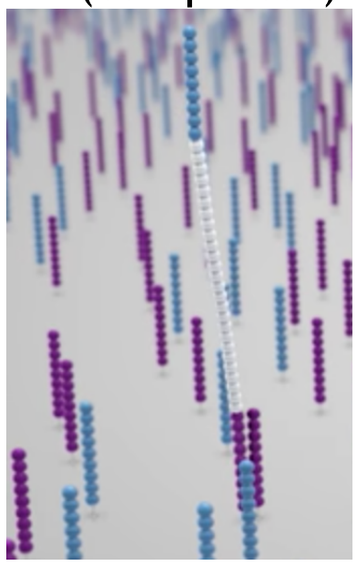
Illumina Sequencing: Amplification

Clustering: Isothermal amplification of the DNA fragments on the solid surface

Illumina Sequencing: Amplification

Clustering: Isothermal amplification of the DNA fragments on the solid surface

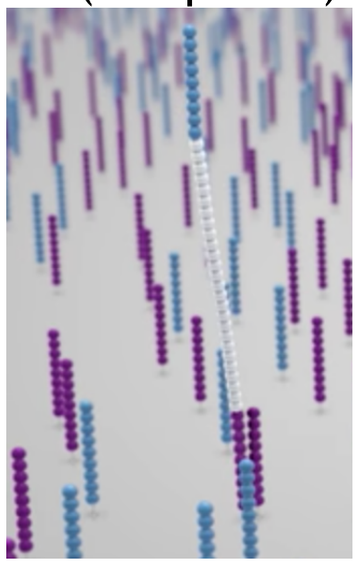
Hybridization of fragments (templates)



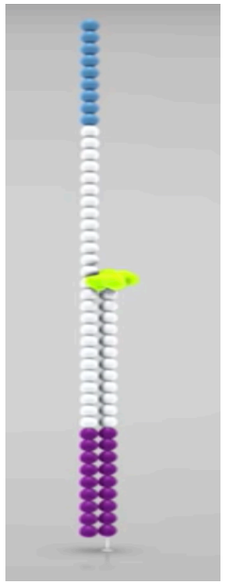
Illumina Sequencing: Amplification

Clustering: Isothermal amplification of the DNA fragments on the solid surface

Hybridization of fragments (templates)



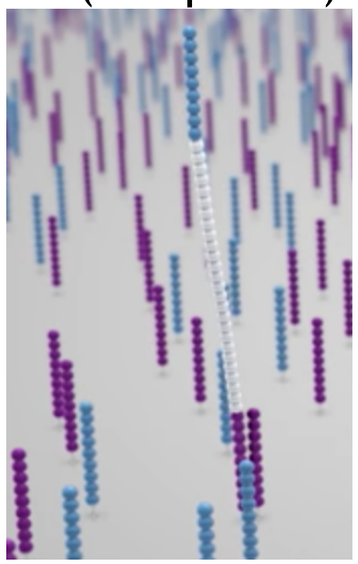
Synthesis of dsDNA



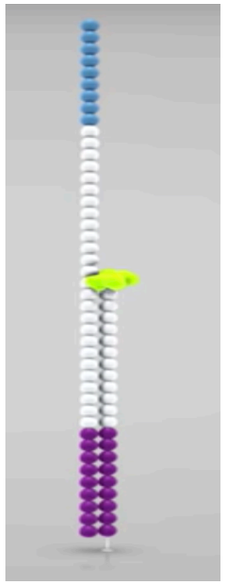
Illumina Sequencing: Amplification

Clustering: Isothermal amplification of the DNA fragments on the solid surface

Hybridization of fragments (templates)



Synthesis of dsDNA

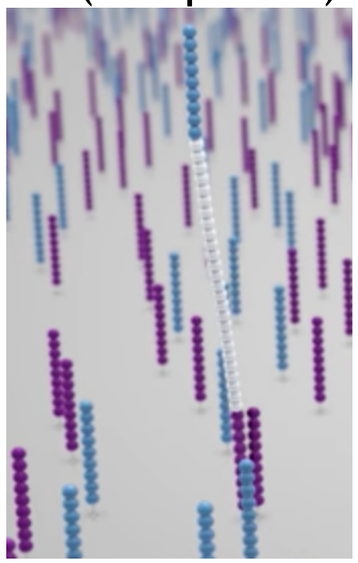


Original strand denatured & washed away

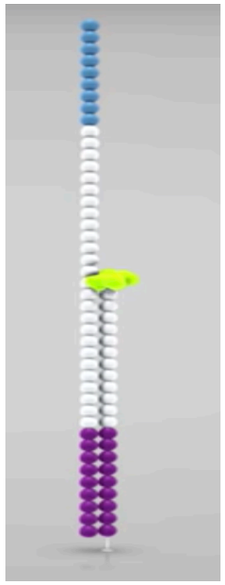
Illumina Sequencing: Amplification

Clustering: Isothermal amplification of the DNA fragments on the solid surface

Hybridization of fragments (templates)



Synthesis of dsDNA



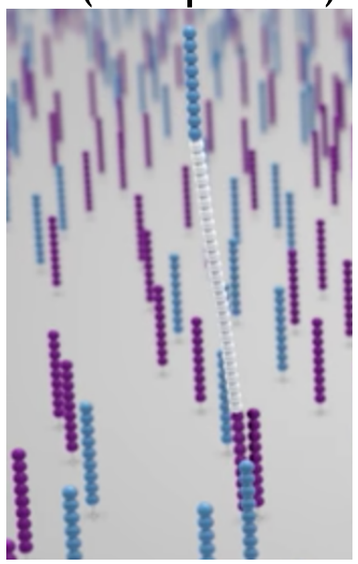
Original strand denatured & washed away

Strand folds over & hybridizes

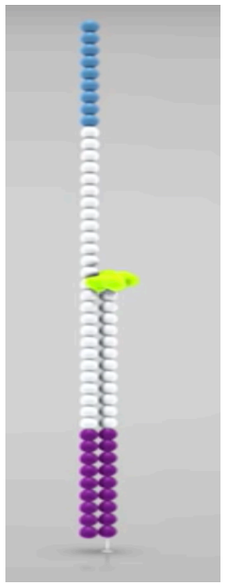
Illumina Sequencing: Amplification

Clustering: Isothermal amplification of the DNA fragments on the solid surface

Hybridization of fragments (templates)



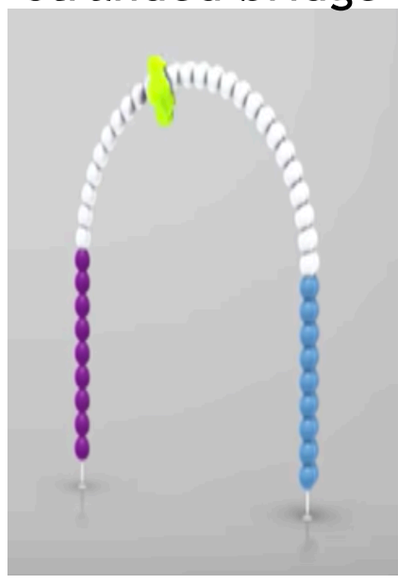
Synthesis of dsDNA



Original strand denatured & washed away

Strand folds over & hybridizes

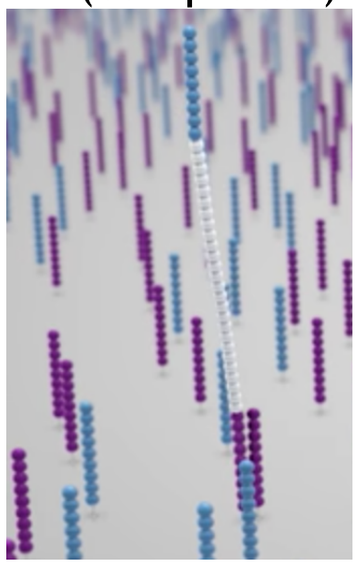
Synthesis of double stranded bridge



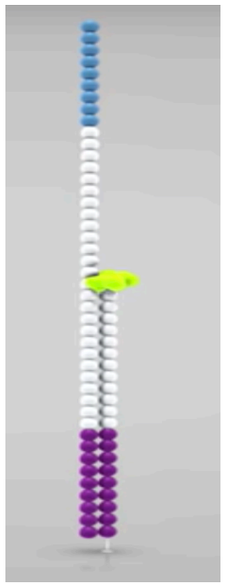
Illumina Sequencing: Amplification

Clustering: Isothermal amplification of the DNA fragments on the solid surface

Hybridization of fragments (templates)



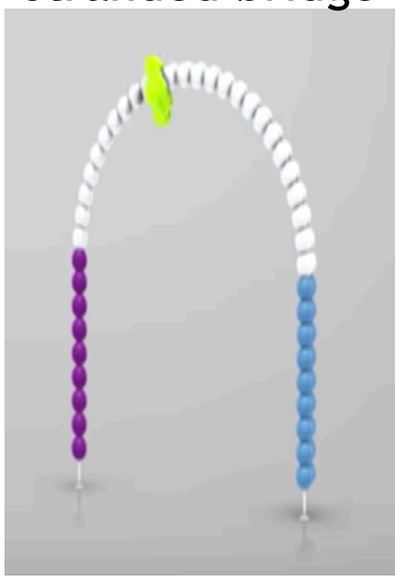
Synthesis of dsDNA



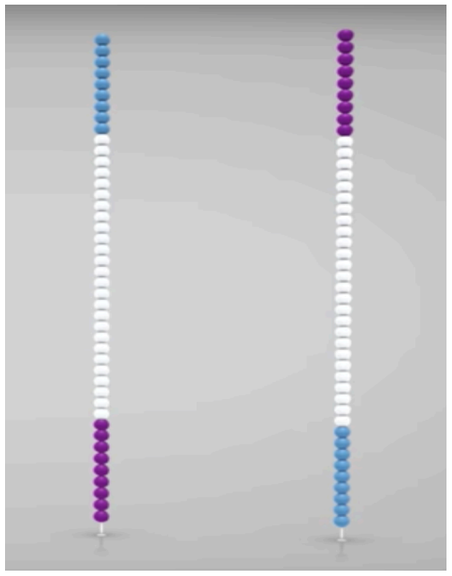
Original strand denatured & washed away

Strand folds over & hybridizes

Synthesis of double stranded bridge



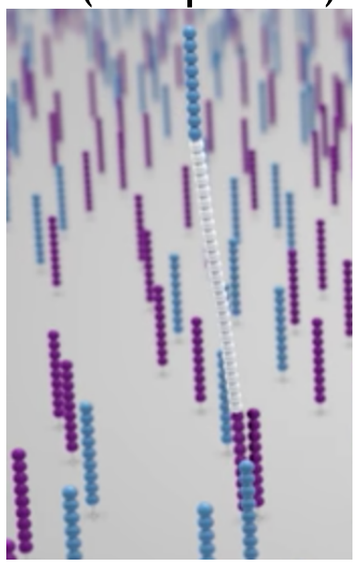
Single stranded DNA molecules tethered to the flow cell



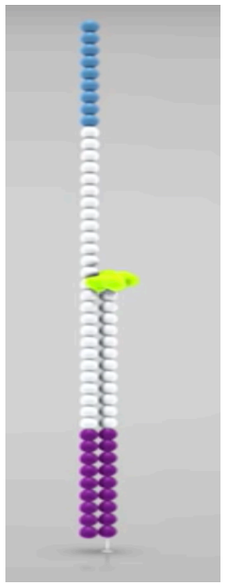
Illumina Sequencing: Amplification

Clustering: Isothermal amplification of the DNA fragments on the solid surface

Hybridization of fragments (templates)



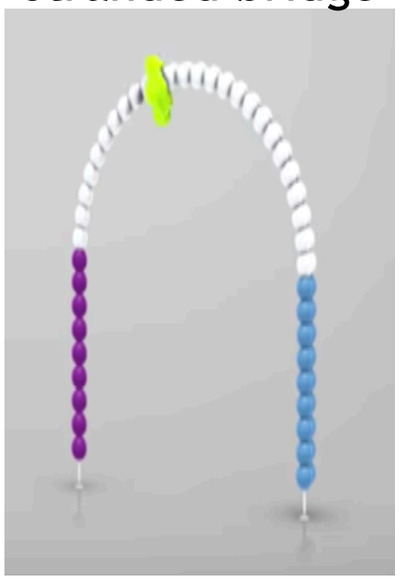
Synthesis of dsDNA



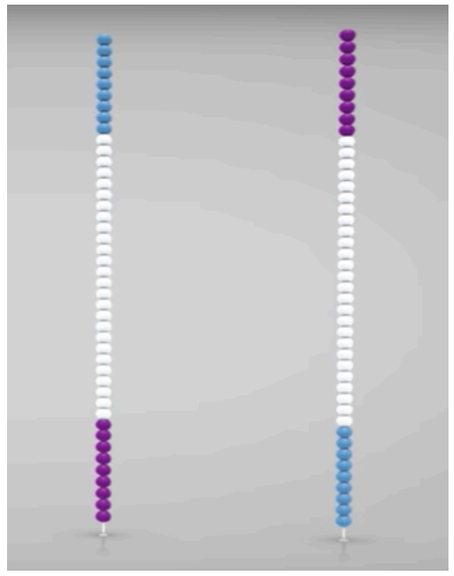
Original strand denatured & washed away

Strand folds over & hybridizes

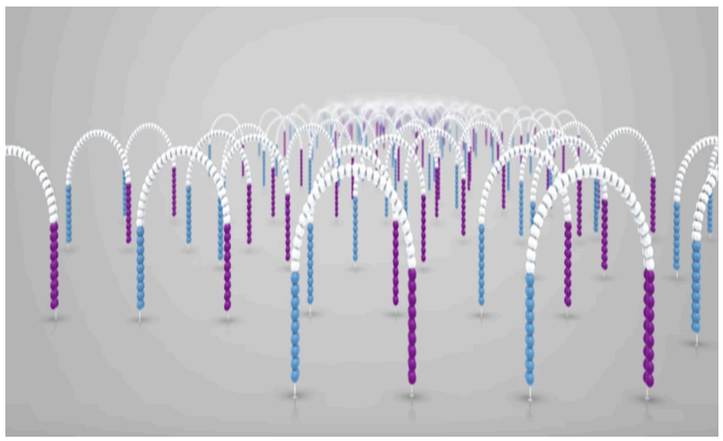
Synthesis of double stranded bridge



Single stranded DNA molecules tethered to the flow cell



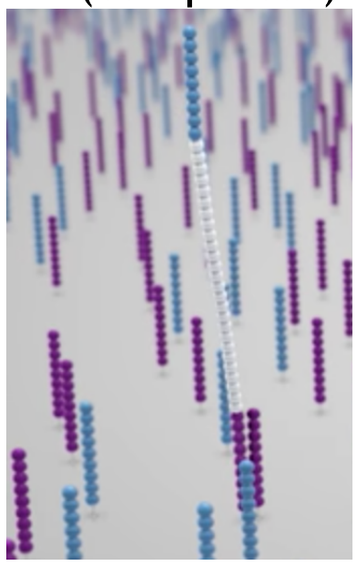
Repeat in parallel



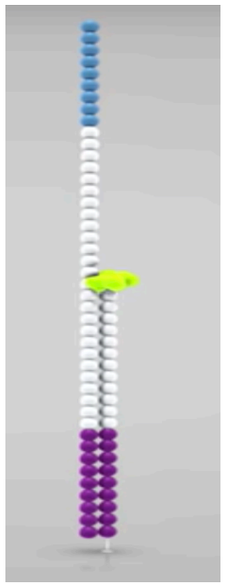
Illumina Sequencing: Amplification

Clustering: Isothermal amplification of the DNA fragments on the solid surface

Hybridization of fragments (templates)



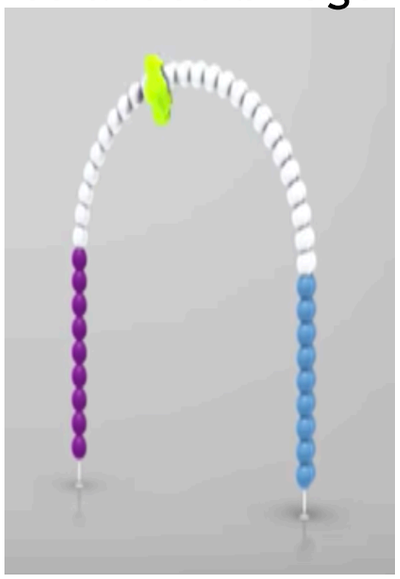
Synthesis of dsDNA



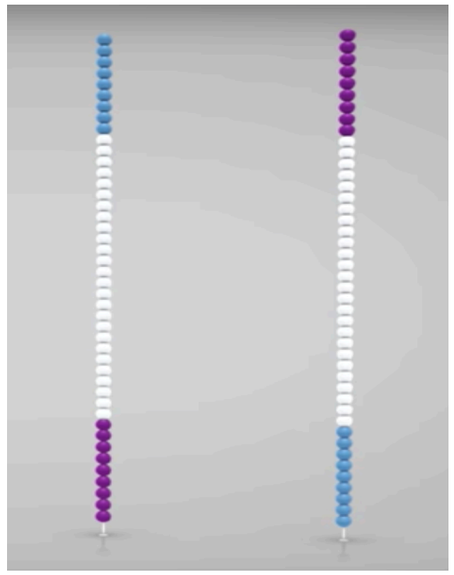
Original strand denatured & washed away

Strand folds over & hybridizes

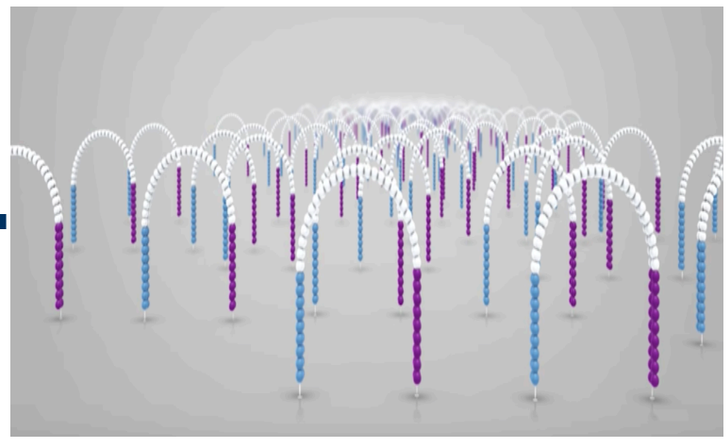
Synthesis of double stranded bridge



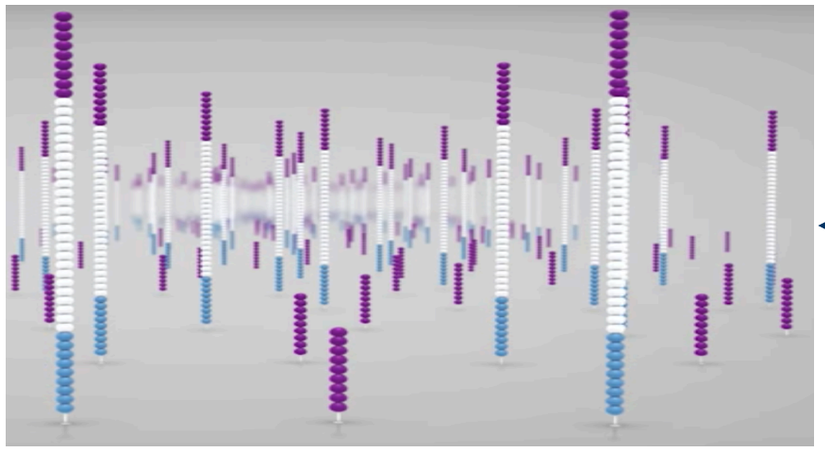
Single stranded DNA molecules tethered to the flow cell



Repeat in parallel

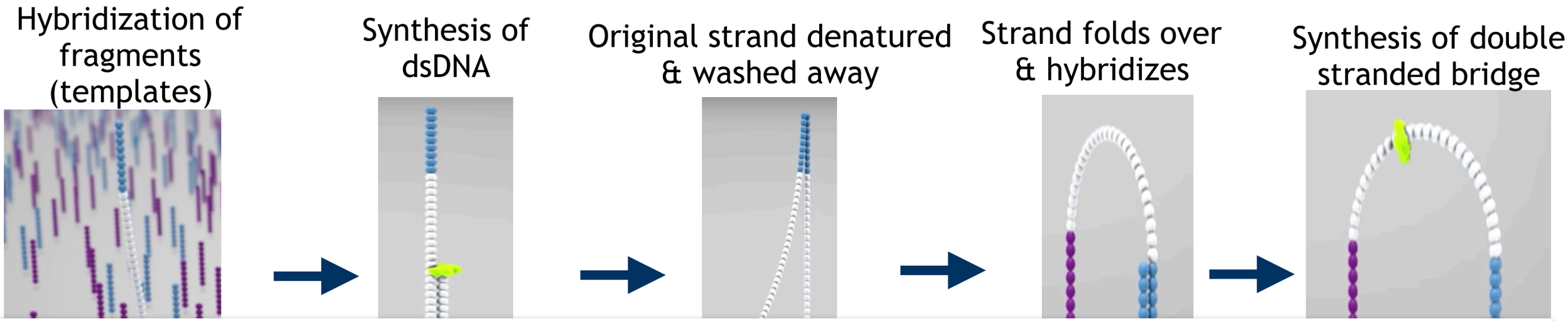


Millions of copies of forward Strand DNA tethered to the flow cell



Illumina Sequencing: Amplification

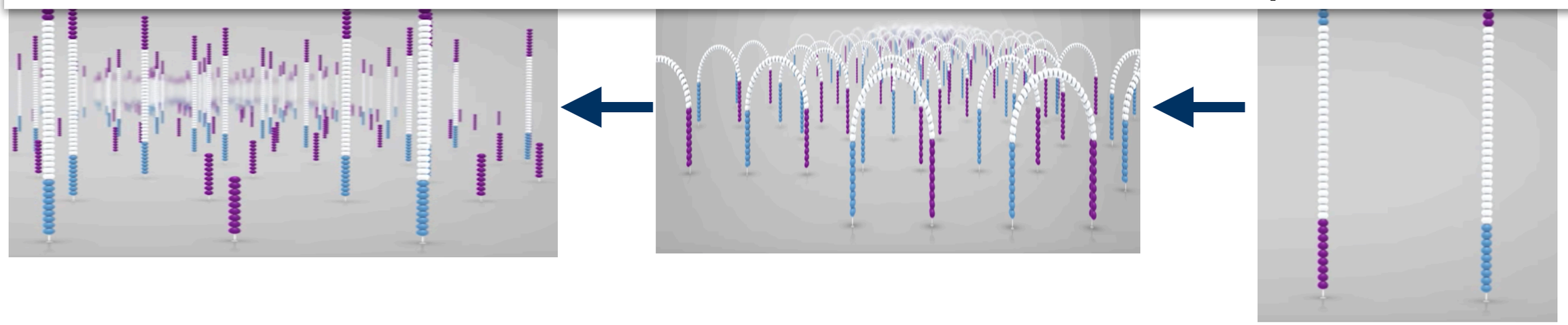
Clustering: Isothermal amplification of the DNA fragments on the solid surface



Cluster formation is a type of PCR ("bridge amplification")

PCR can introduce preferential amplification of some fragments

PCR can introduce artifacts, which will lead to false positive variants



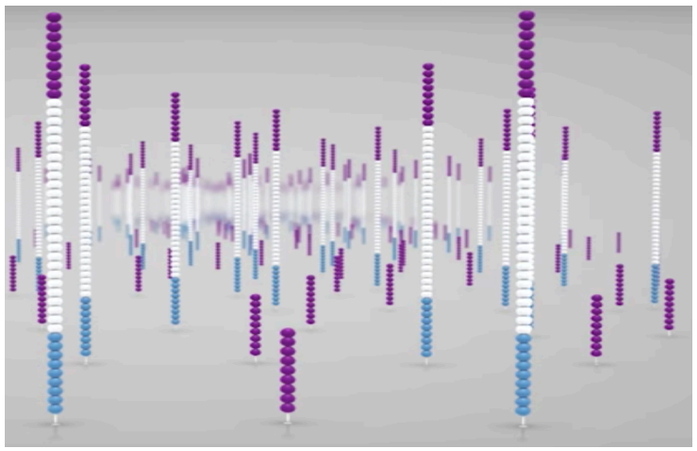
Illumina Sequencing: Sequencing by synthesis

number of cycles => read length

Illumina Sequencing: Sequencing by synthesis

number of cycles => read length

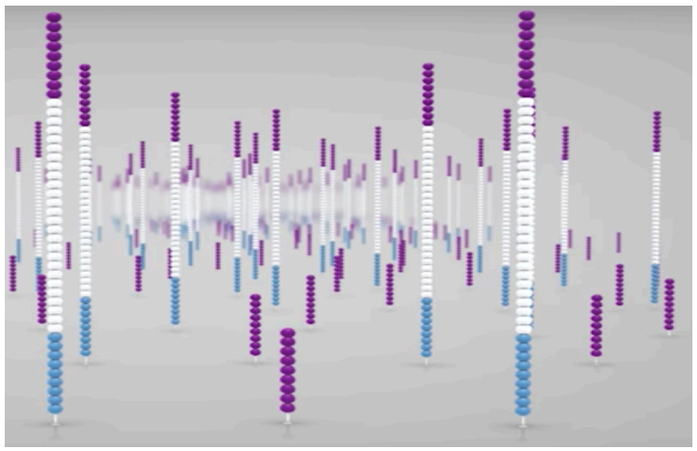
Millions of copies of forward Strand DNA tethered to the flow cell



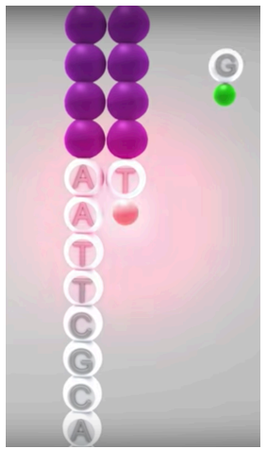
Illumina Sequencing: Sequencing by synthesis

number of cycles => read length

Millions of copies of forward Strand DNA tethered to the flow cell



Sequencing primer extended, sequencing begins



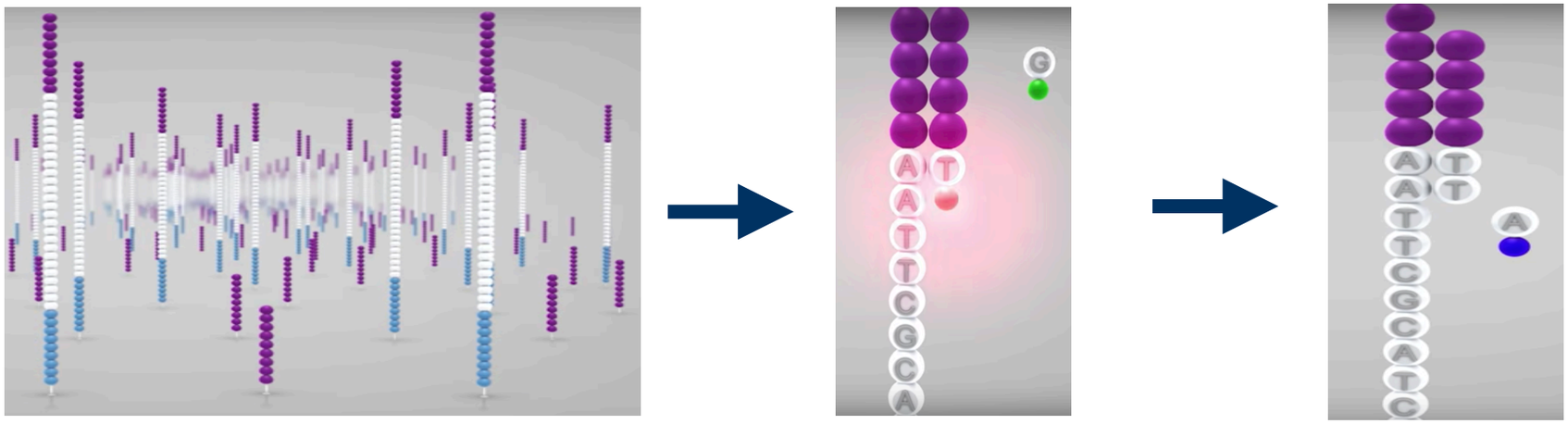
Illumina Sequencing: Sequencing by synthesis

number of cycles => read length

Millions of copies of forward Strand DNA tethered to the flow cell

Sequencing primer extended, sequencing begins

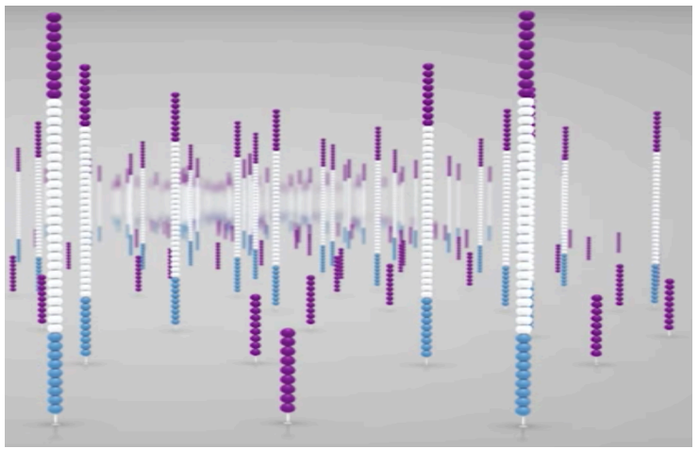
In each cycle, only one fluorescently tagged nucleotide is incorporated



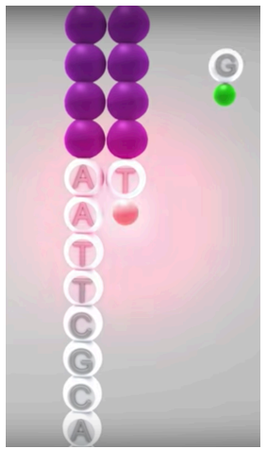
Illumina Sequencing: Sequencing by synthesis

number of cycles => read length

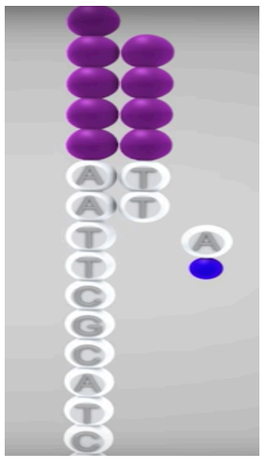
Millions of copies of forward Strand DNA tethered to the flow cell



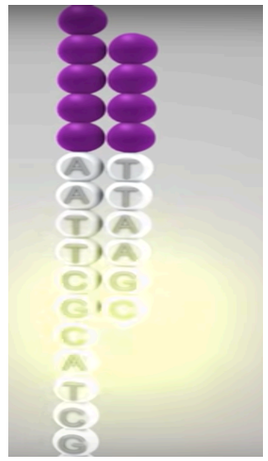
Sequencing primer extended, sequencing begins



In each cycle, only one fluorescently tagged nucleotide is incorporated



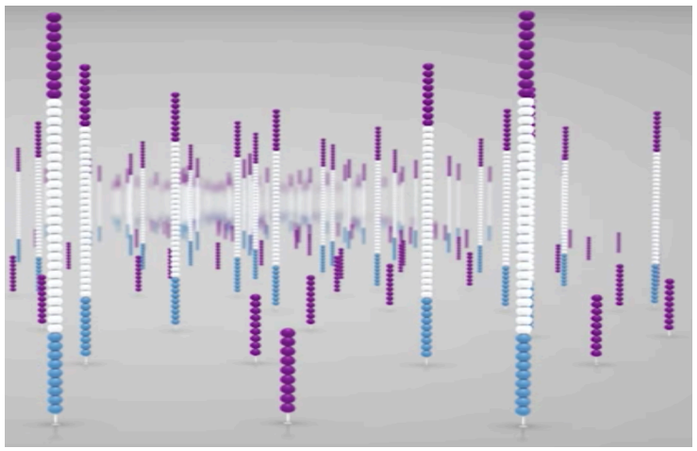
Excitation & Emission: After nucleotide addition, clusters are excited by a light source



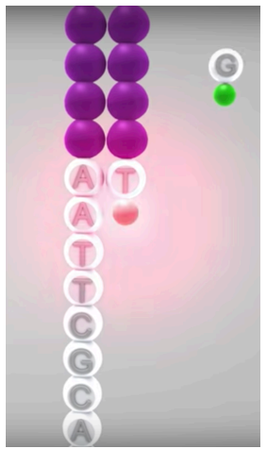
Illumina Sequencing: Sequencing by synthesis

number of cycles => read length

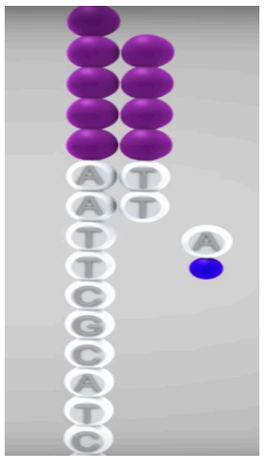
Millions of copies of forward Strand DNA tethered to the flow cell



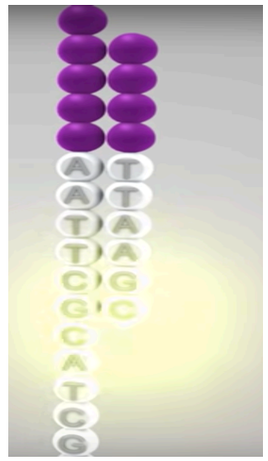
Sequencing primer extended, sequencing begins



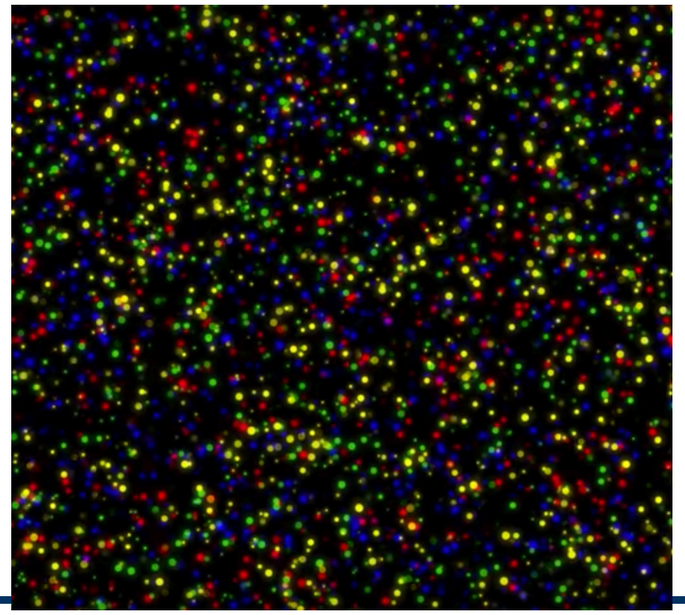
In each cycle, only one fluorescently tagged nucleotide is incorporated



Excitation & Emission: After nucleotide addition, clusters are excited by a light source



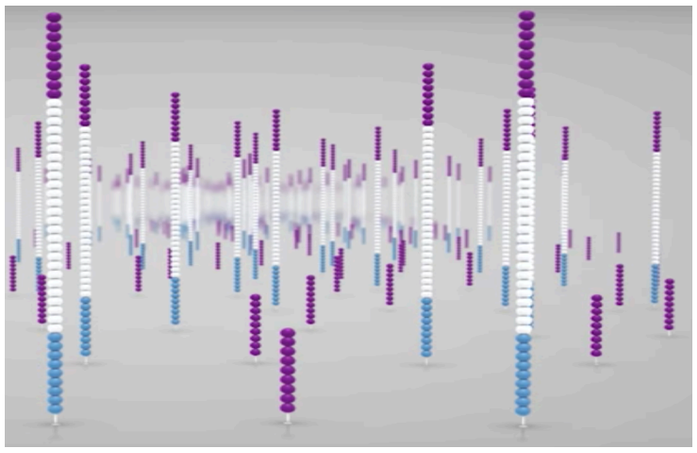
Four images taken per sequencing cycle



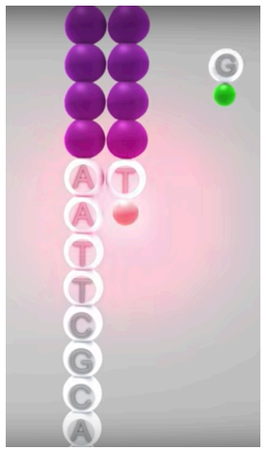
Illumina Sequencing: Sequencing by synthesis

number of cycles => read length

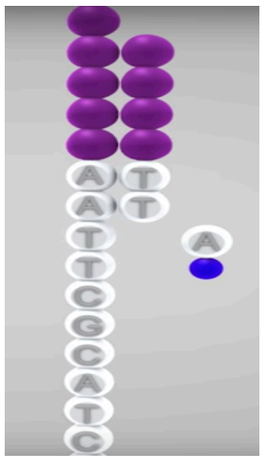
Millions of copies of forward Strand DNA tethered to the flow cell



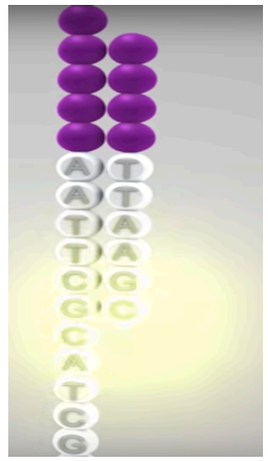
Sequencing primer extended, sequencing begins



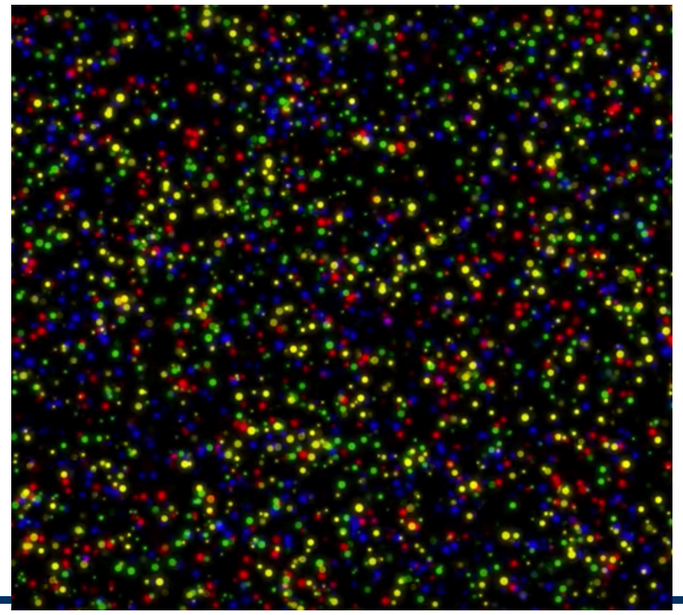
In each cycle, only one fluorescently tagged nucleotide is incorporated



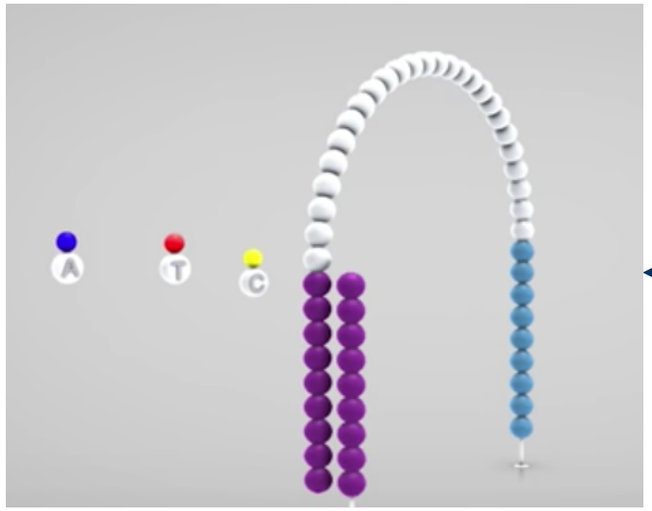
Excitation & Emission: After nucleotide addition, clusters are excited by a light source



Four images taken per sequencing cycle



Generate Read 2, repeating for the reverse strand



...

Illumina Platforms

Miniseq

- small,benchtop
- 2x150 bp
- Two output modes:
  - high: ~6 Gb
  - mid: ~2 Gb

MiSeq

- benchtop
- v3 chemistry offers 2x300 bp reads
- Reverse read quality drops after ~200th cycle
- Throughput: 25 million reads/lane

<u>Illumina HiSeq</u>
~3 billion paired 100bp reads
~600Gb, $10K, 8 days
(or “rapid run” ~90Gb in 1-2 days)

<u>Illumina X Ten</u>
~6 billion paired 150bp reads
1.8Tb, <3 days, ~1000 / genome($$)
(or “rapid run” ~90Gb in 1-2 days)

<u>Illumina NextSeq</u>
One human genome in **<u><30 hours</u>**

Illumina Platforms

Illumina HiSeq
~3 billion paired 100bp reads
~600Gb, $10K, 8 days
(or “rapid run” ~90Gb in 1-2 days)

Illumina X Ten
~6 billion paired 150bp reads
1.8Tb, <3 days, ~1000 / genome($$)
(or “rapid run” ~90Gb in 1-2 days)

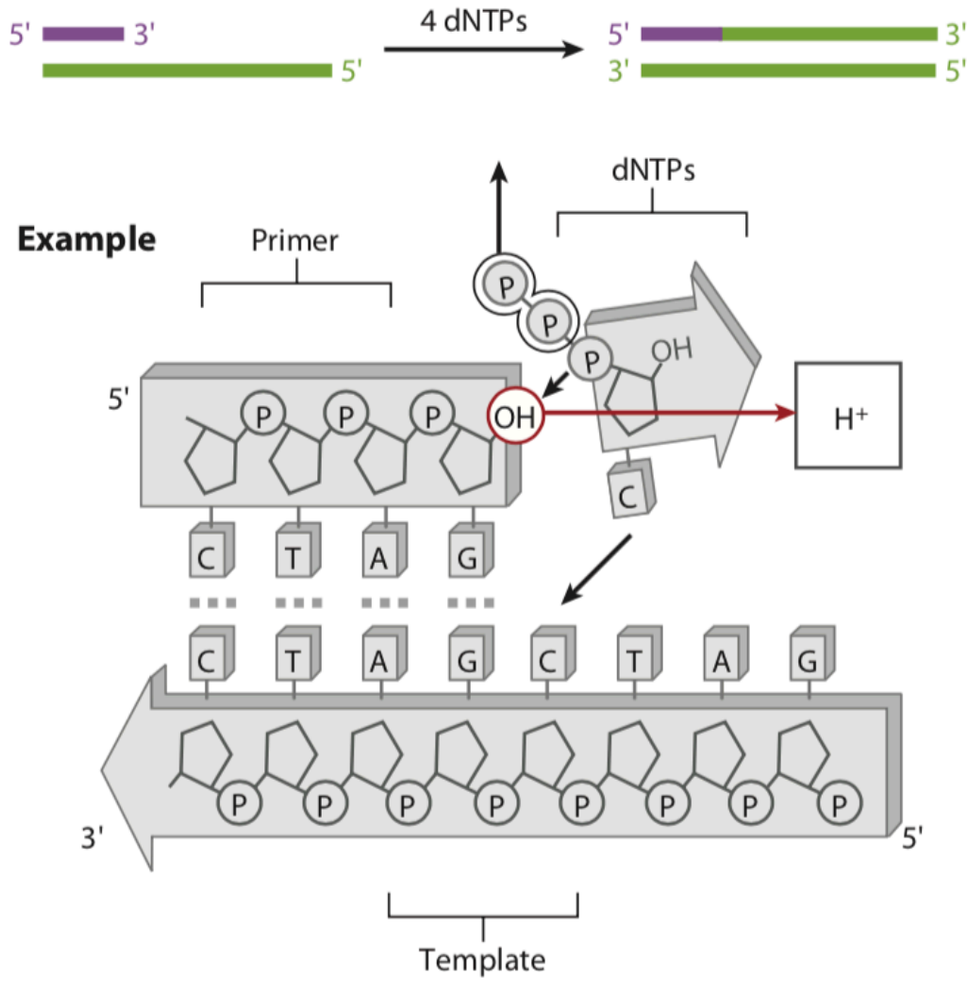
Illumina NextSeq
One human genome in **<30 hours**

Ion Semiconductor Sequencing

Detection is electrochemical, no optics

What is detected?

Hydrogen ions (H^+) released during nucleotide incorporation

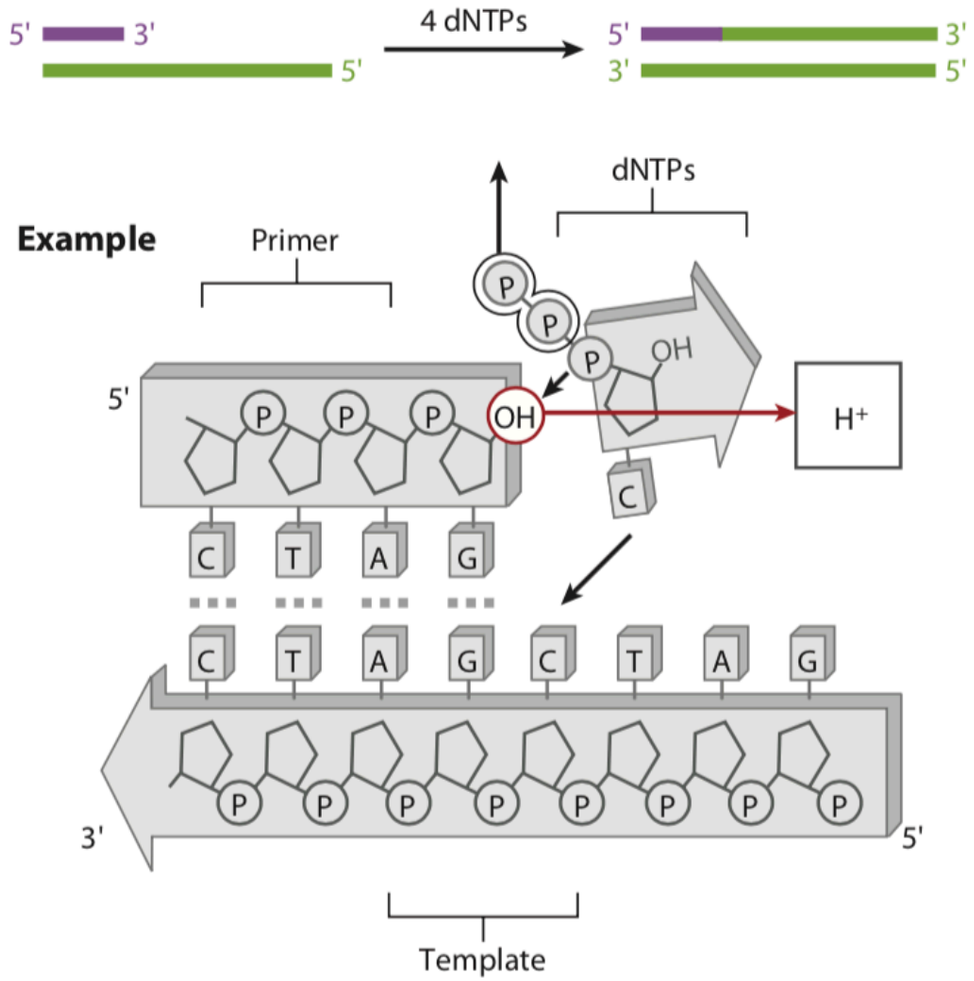


Ion Semiconductor Sequencing

Detection is electrochemical, no optics

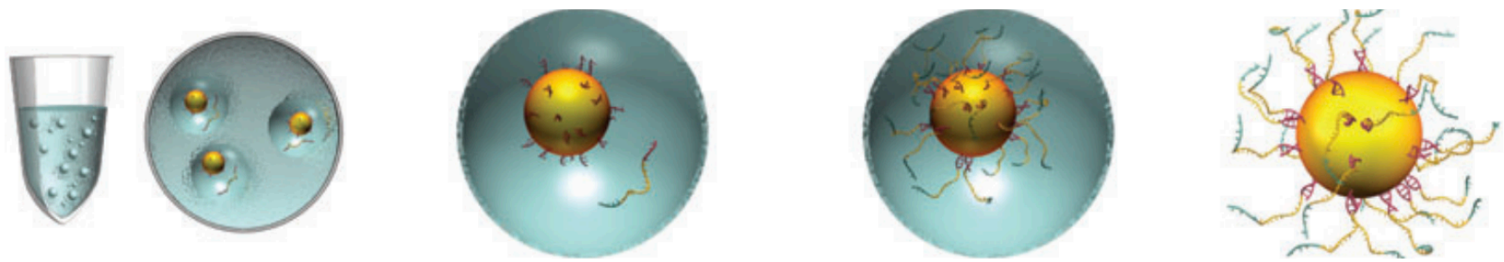
What is detected?

Hydrogen ions (H^+) released during nucleotide incorporation



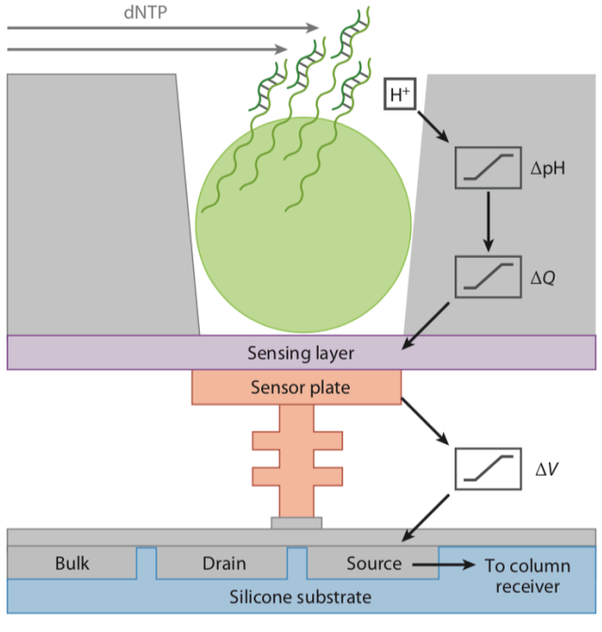
Amplification:

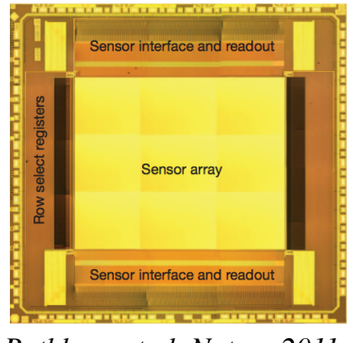
bead-based, emulsion-PCR

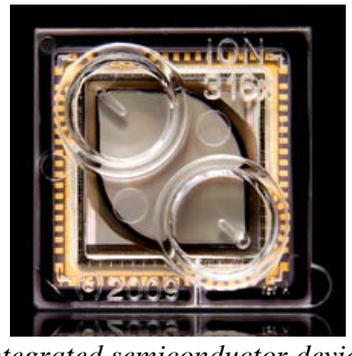


adapted from Mardis 2008. Annu. Rev. Genomics Hum. Genet.

Ion Semiconductor Sequencing







Rothberg et al. Nature 2011. An integrated semiconductor device enabling non-optical genome sequencing.

1-11 million wells

ion torrent
by Thermo Fisher Scientific

Ion Proton

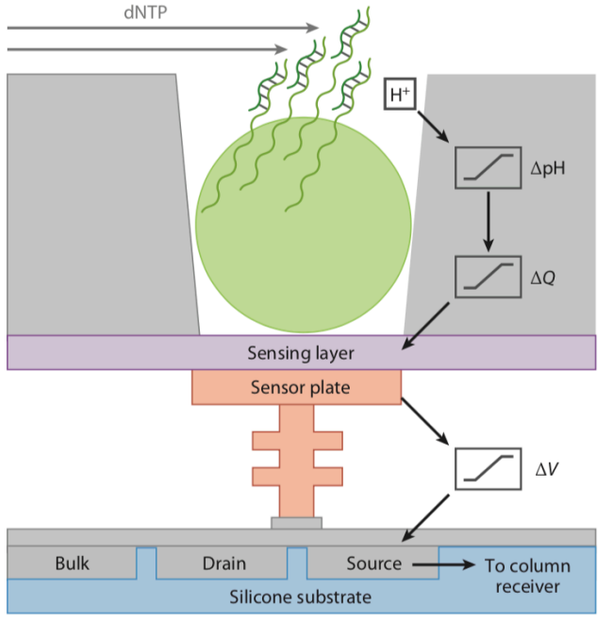
Ion PGM

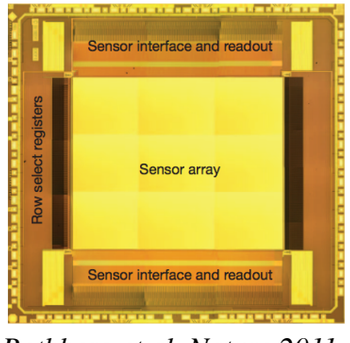
- linear dynamic range
- no substitution errors
- low startup cost
- short run time

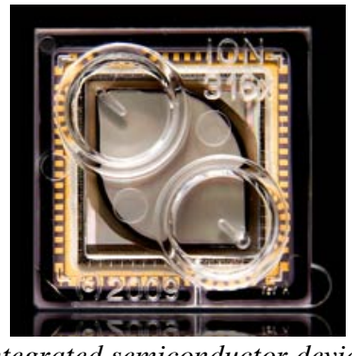
- paired-end reads not supported
- higher error rates
- cost/base high
- short read lengths (200 bp)

Ion Semiconductor Sequencing

Semiconductor based pH-meter







Rothberg et al. Nature 2011. An integrated semiconductor device enabling non-optical genome sequencing.

1-11 million wells

ion torrent
by Thermo Fisher Scientific

Ion Proton

Ion PGM

- linear dynamic range
- no substitution errors
- low startup cost
- short run time

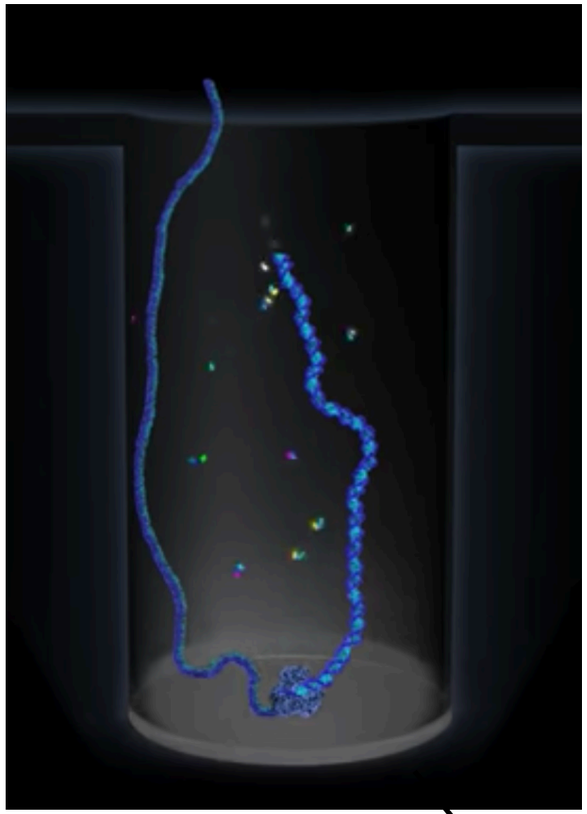
- paired-end reads not supported
- higher error rates
- cost/base high
- short read lengths (200 bp)

Third-generation Sequencing

Single Molecule Sequencing

"Real-time" Sequencing

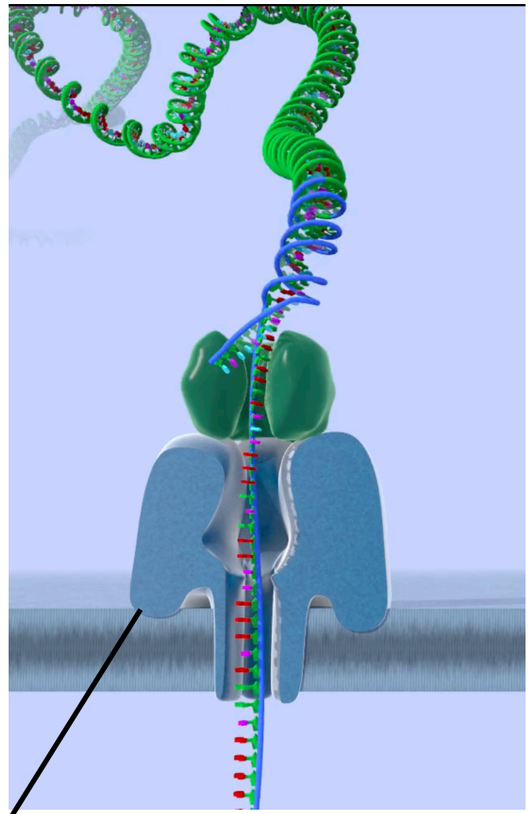
Third-generation Sequencing



nano-well

Single Molecule Sequencing

"Real-time" Sequencing



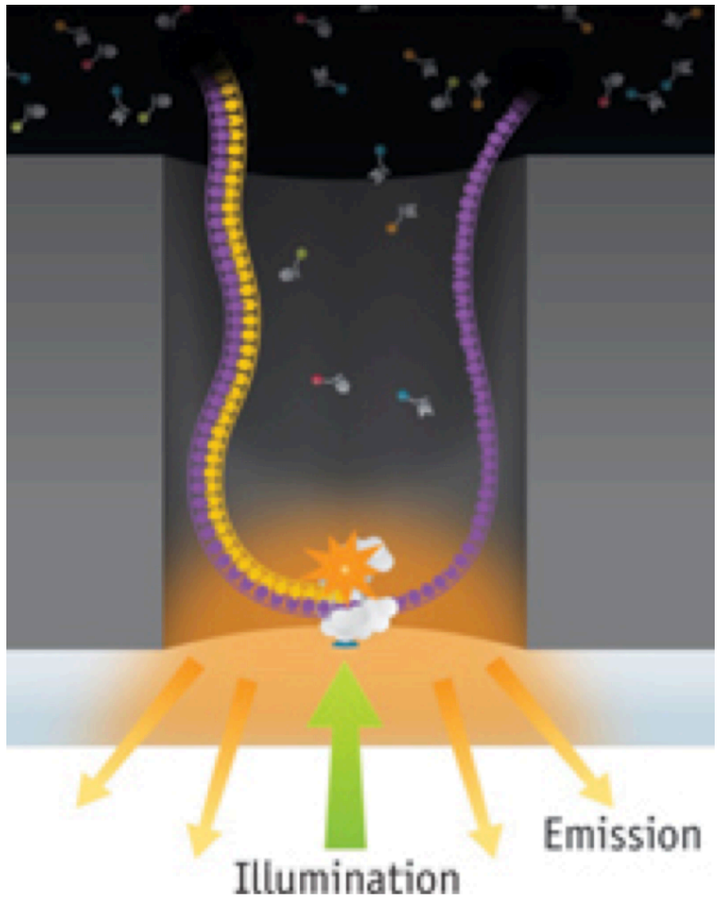
nanopore

PacBio Single Molecule Sequencing using *Zero-mode Waveguides*

Zero-mode Waveguide (ZMW):

provides *zeptoliter* (10^{-21}) scale detection volume

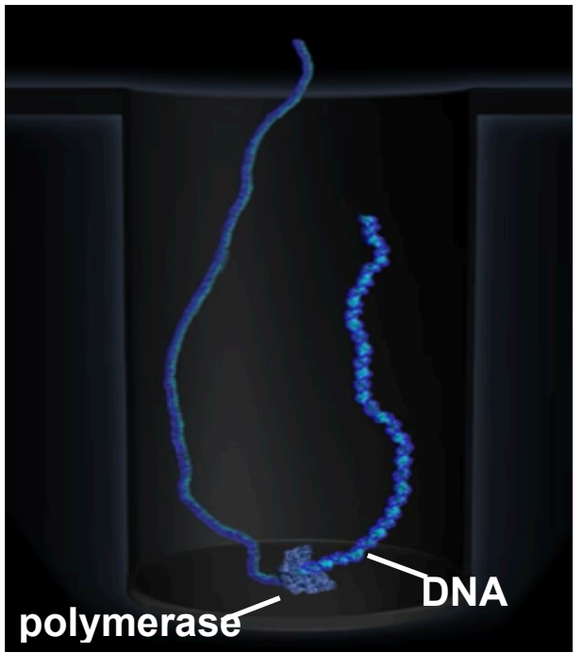
SMRT Cell:
nanofabricated array of ZMWs



Zero-mode Waveguide (ZMW)

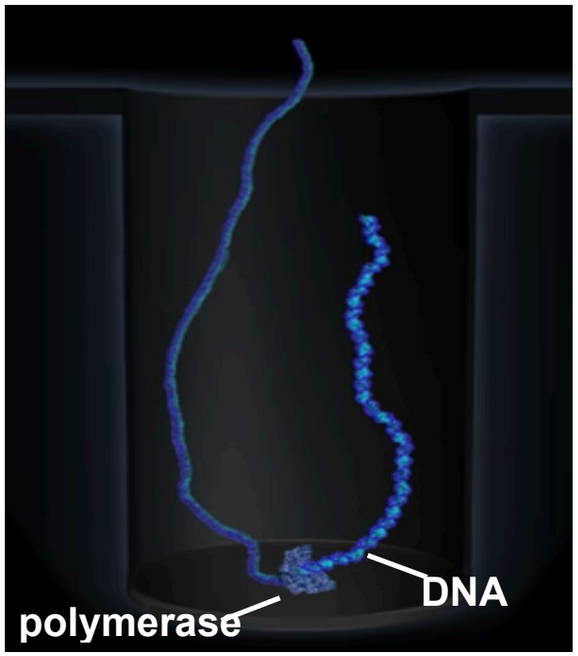
PacBio Sequencing: Single-molecule, real time sequencing

I. DNA:polymerase complex, immobilized at the bottom of ZMW

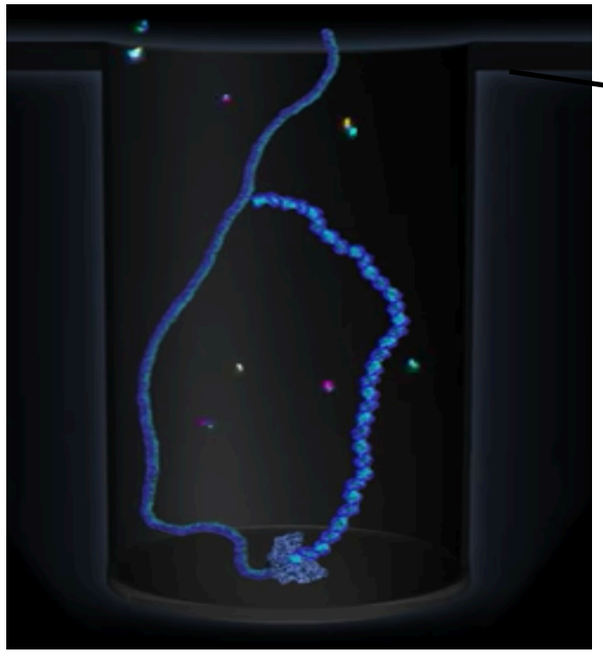


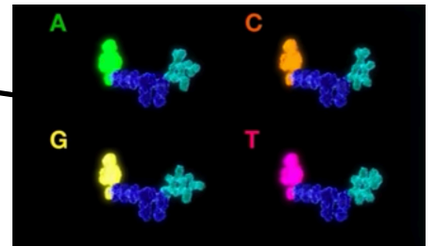
PacBio Sequencing: Single-molecule, real time sequencing

I. DNA:polymerase complex, immobilized at the bottom of ZMW



II. Flow in fluorescently labeled nucleotides

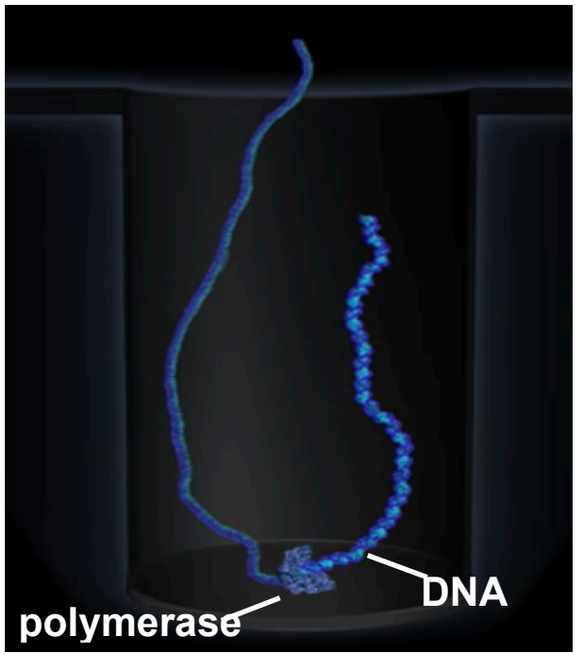




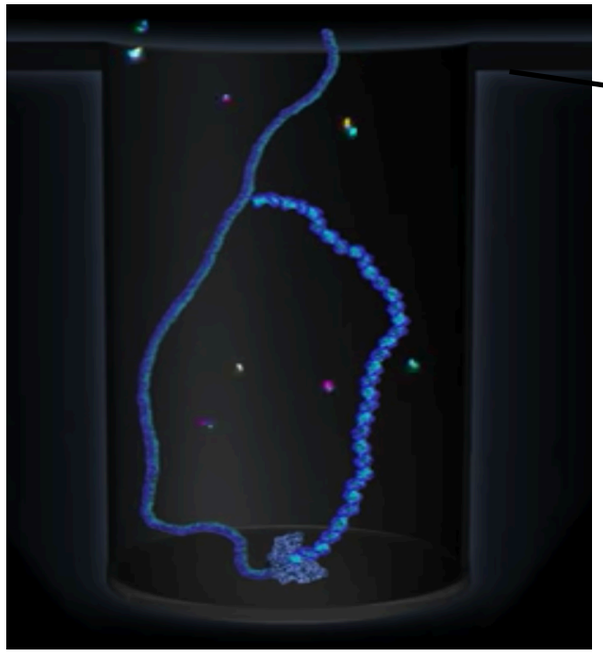
phospholinked nucleotides

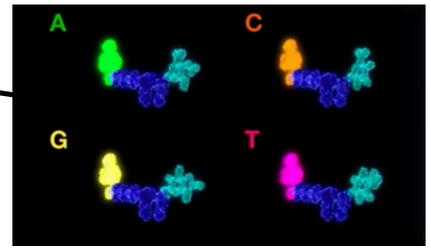
PacBio Sequencing: Single-molecule, real time sequencing

I. DNA:polymerase complex, immobilized at the bottom of ZMW



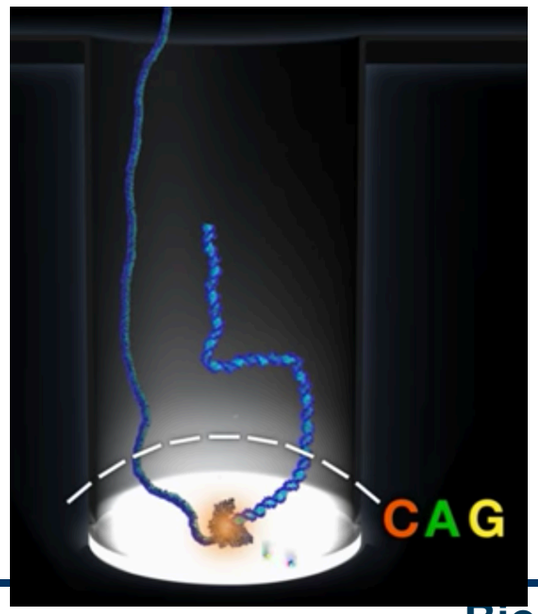
II. Flow in fluorescently labeled nucleotides





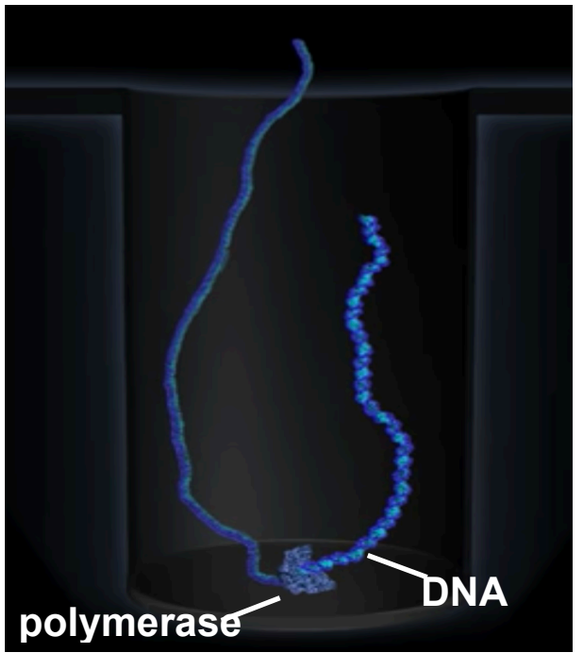
phospholinked nucleotides

III. Fluorescent nucleotide in the active site, a light pulse is produced

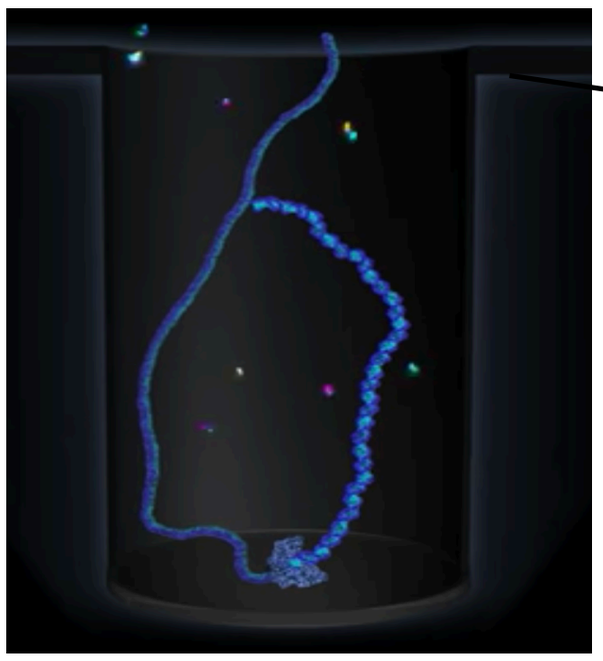


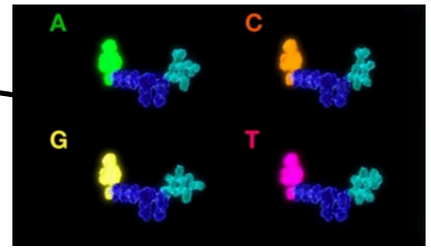
PacBio Sequencing: Single-molecule, real time sequencing

I. DNA:polymerase complex, immobilized at the bottom of ZMW



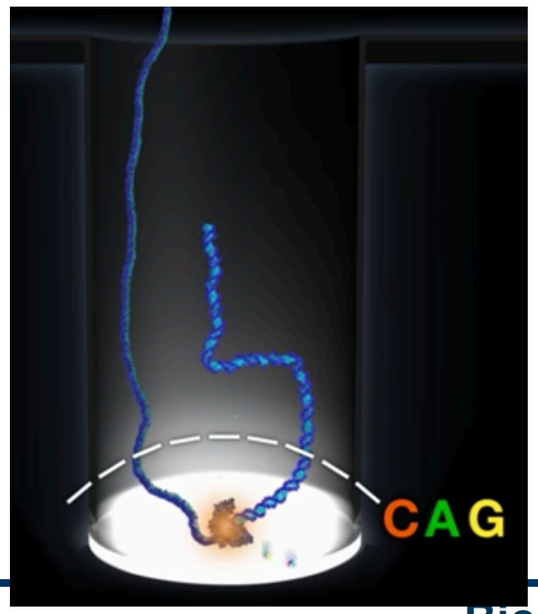
II. Flow in fluorescently labeled nucleotides

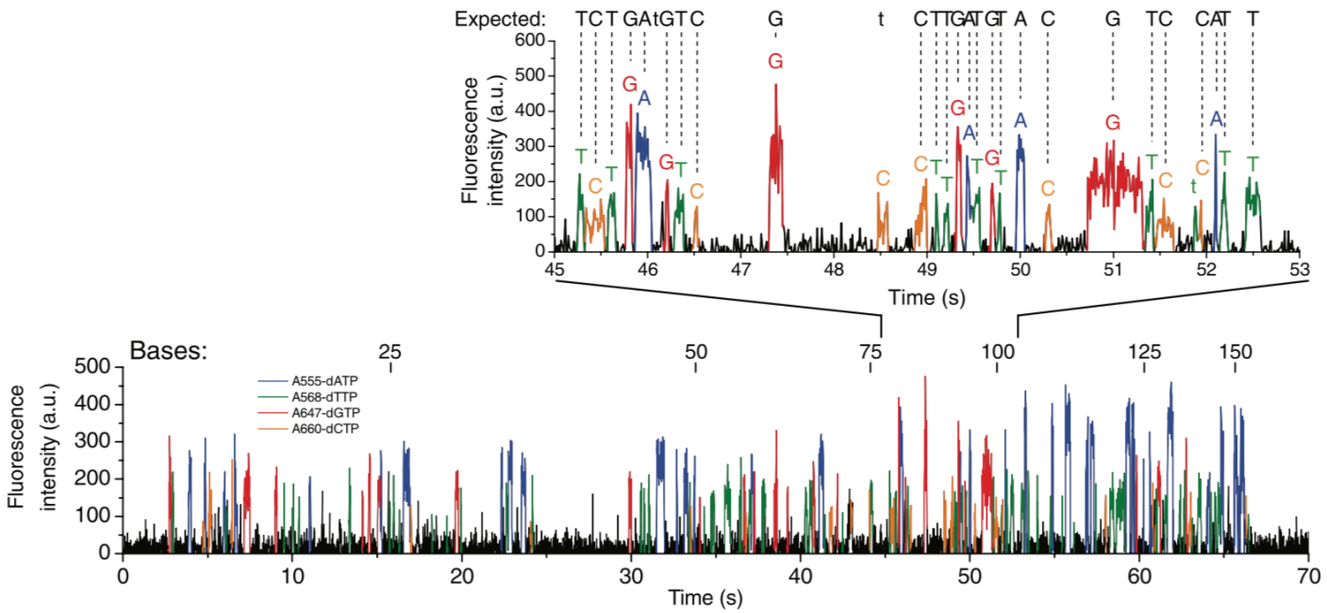




phospholinked nucleotides

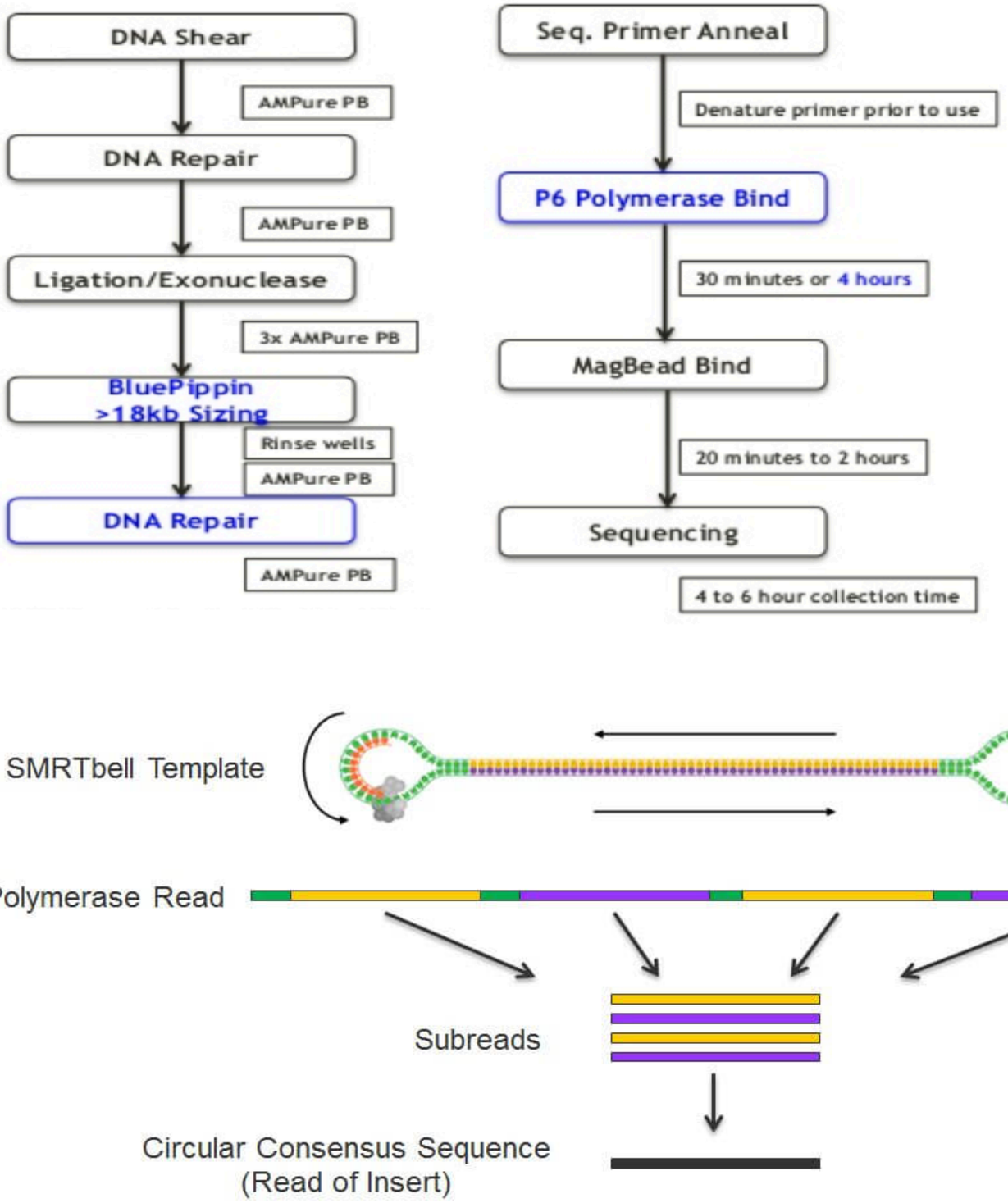
III. Fluorescent nucleotide in the active site, a light pulse is produced





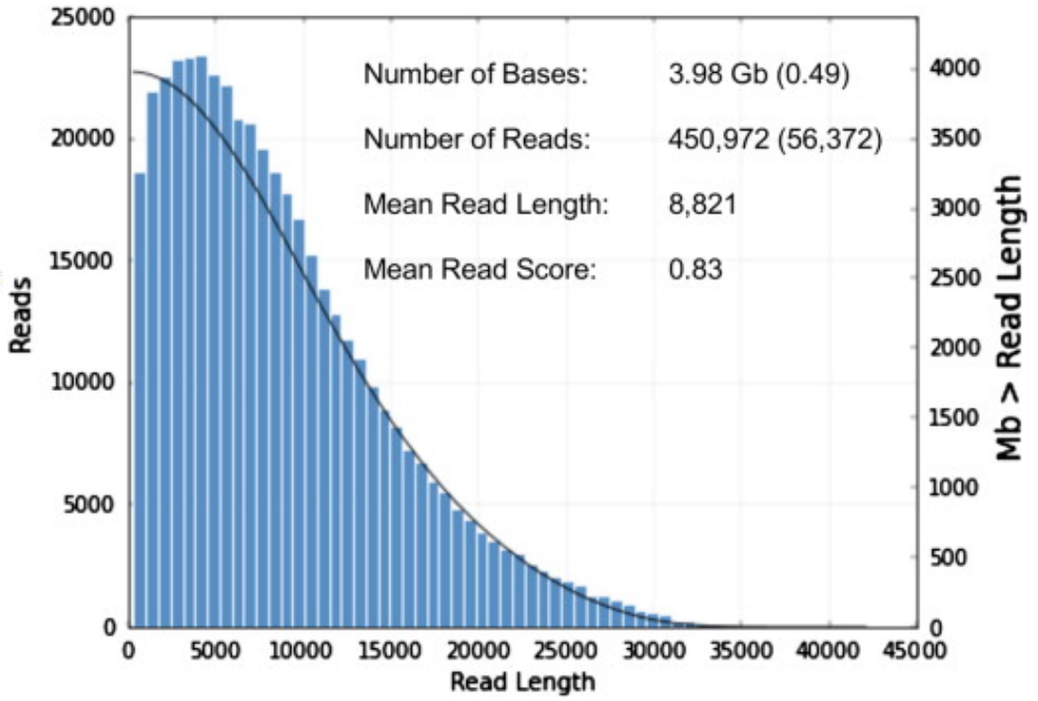
Eid et al. 2009. Science. Real-Time DNA Sequencing from Single Polymerase Molecules

PacBio Sequencing: Library Workflows

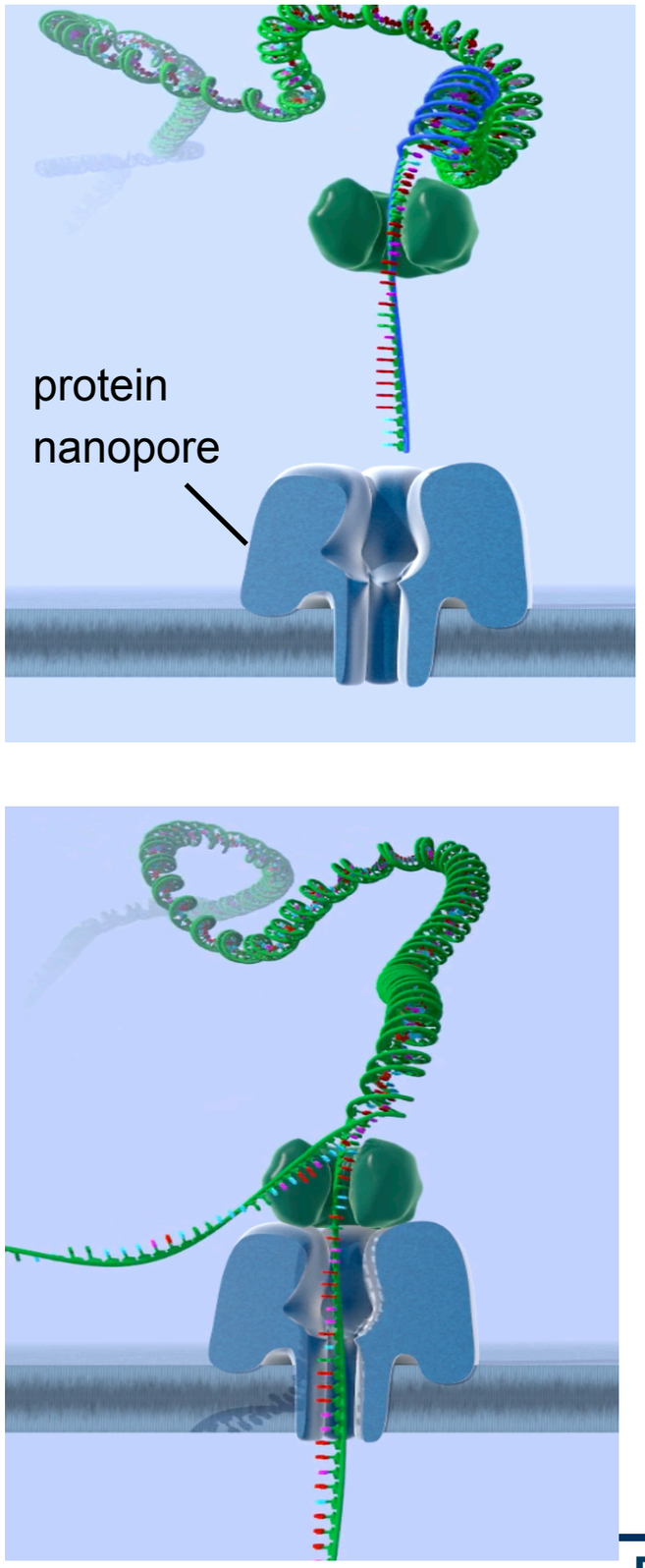


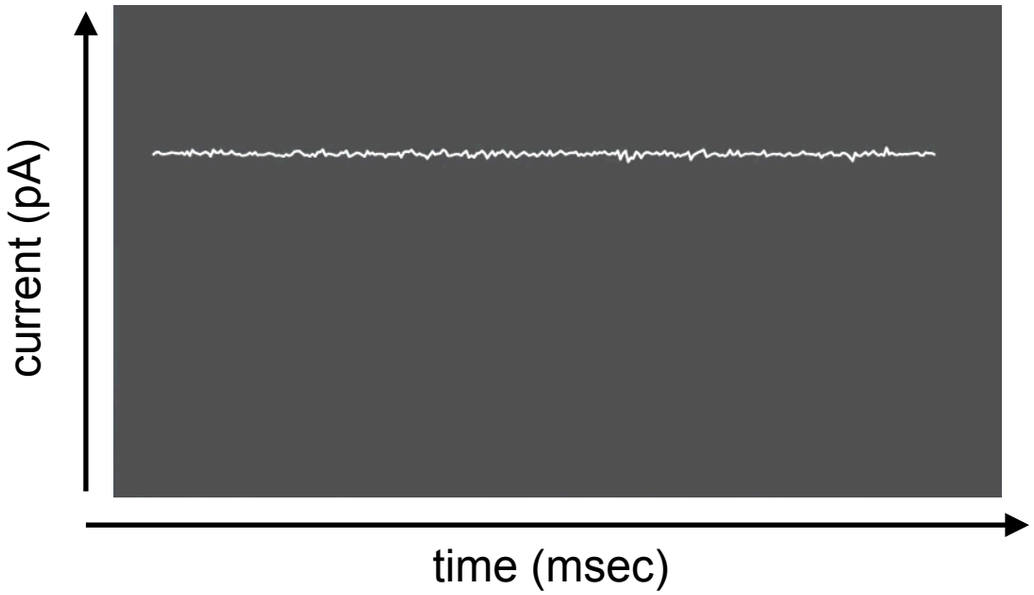
Considerations

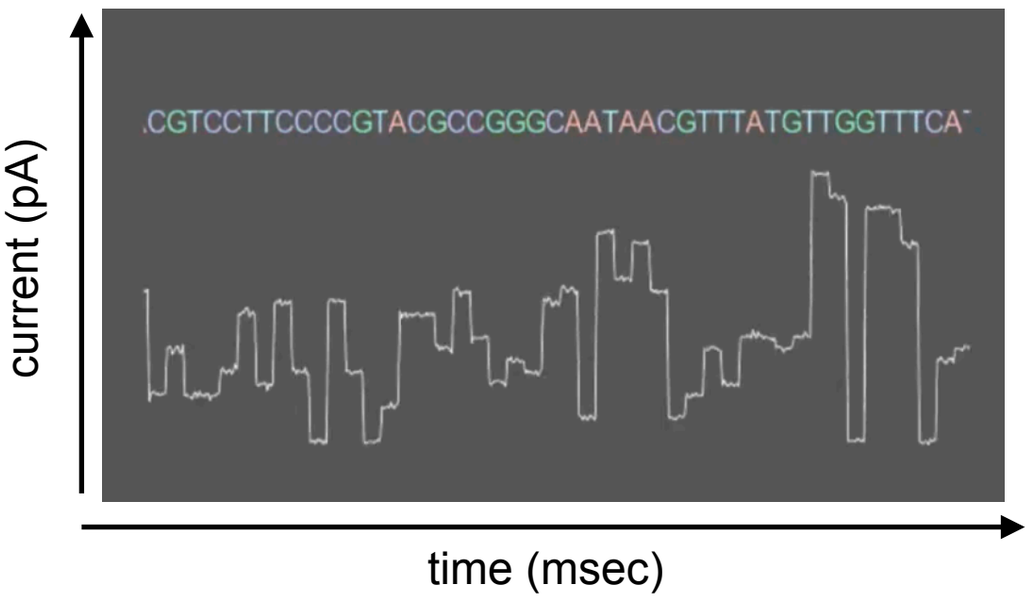
- High-quality, high molecular weight DNA
- Consistent shearing > 30 kb



Nanopore Sequencing

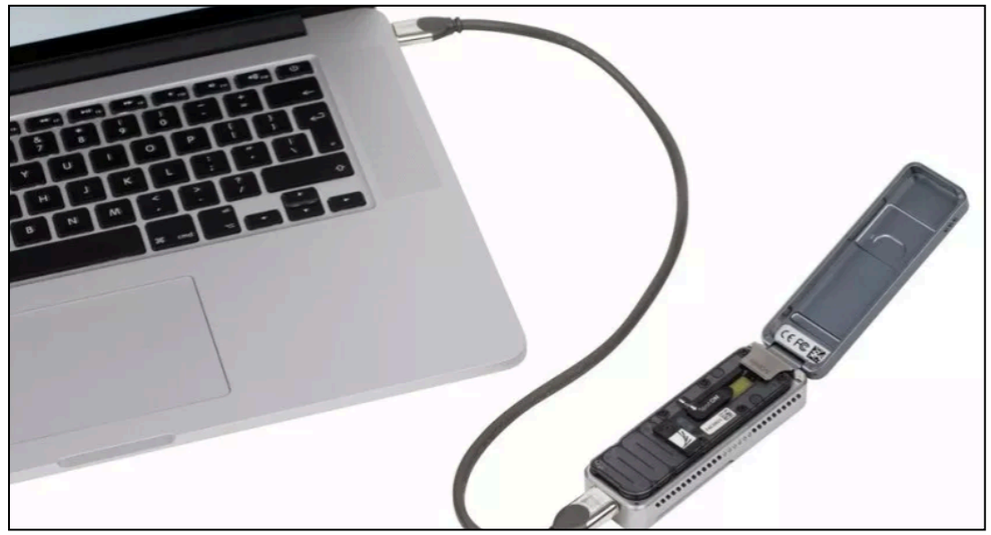






- Electrical current based detection of nucleotides in the pore
- Variable read lengths
- Error rate: 5-15%

Oxford Nanopore Sequencing Device



- Handheld
- Low power
- Low capital cost ($1000)

NGS vs. Single-Molecule Sequencing

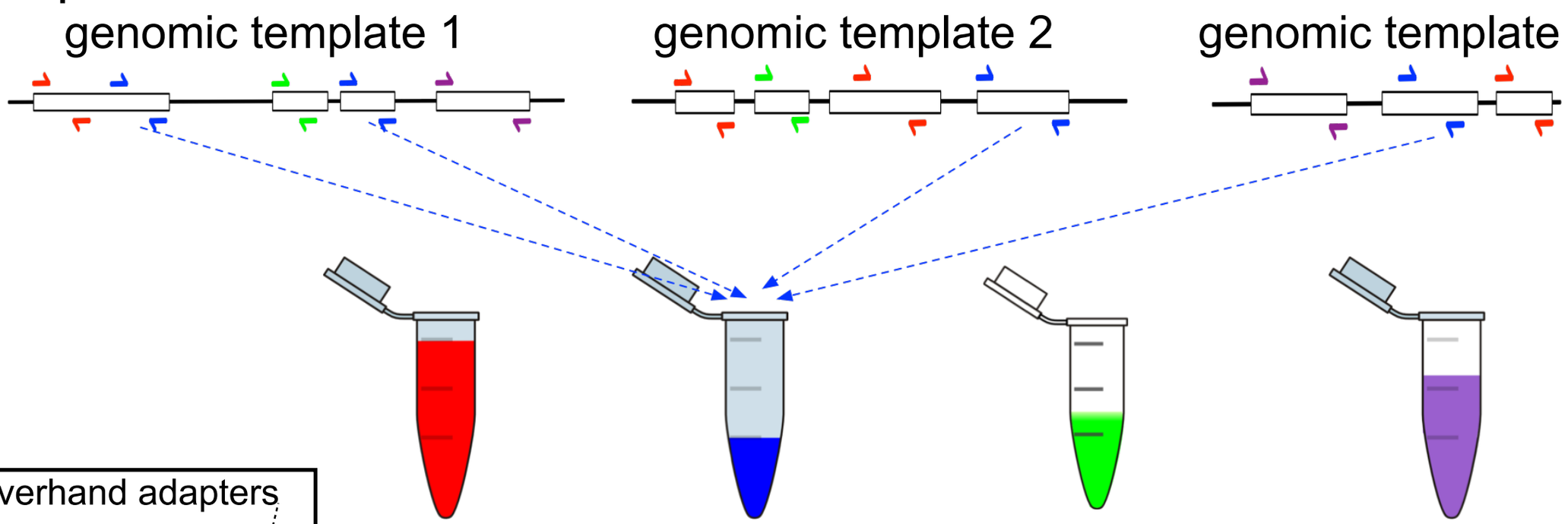
	Next-generation	Single Molecule
Amplification	needed	none
Cost (startup)	high	low
Cost (per bp)	low	high
Run Time	hours (IonTorrent)-days (Illumina)	hours
Read Length	short (<400bp)	long
Error Rate	low	high

Targeted Enrichment and Sequencing

- **Hybridization-based**
- **PCR-based**

Multiplex PCR Panels

12 primer pairs

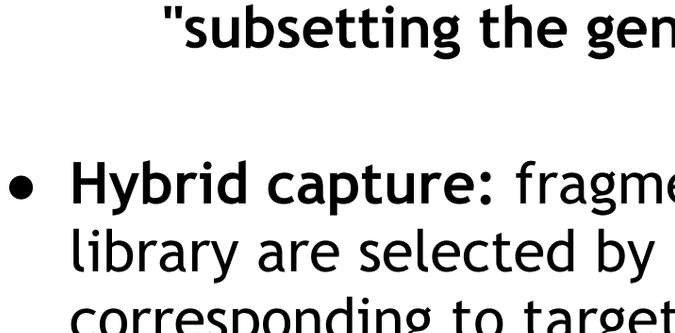


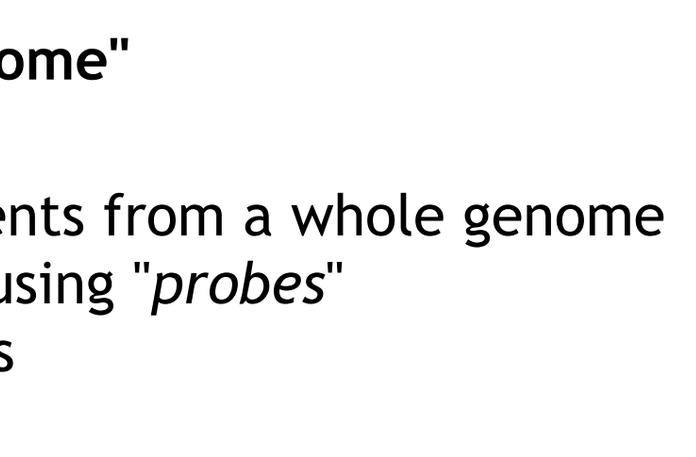
- Design primer pairs for targets, longer regions might need tiled primers
- Group primer pairs according to GC content, Tm and reaction condition specifics
- Amplify genomic DNA to generate multiple products from each primer set, pool products
- Create sequencing library by ligation or tail platform specific adaptors on the primer ends
- SEQUENCE

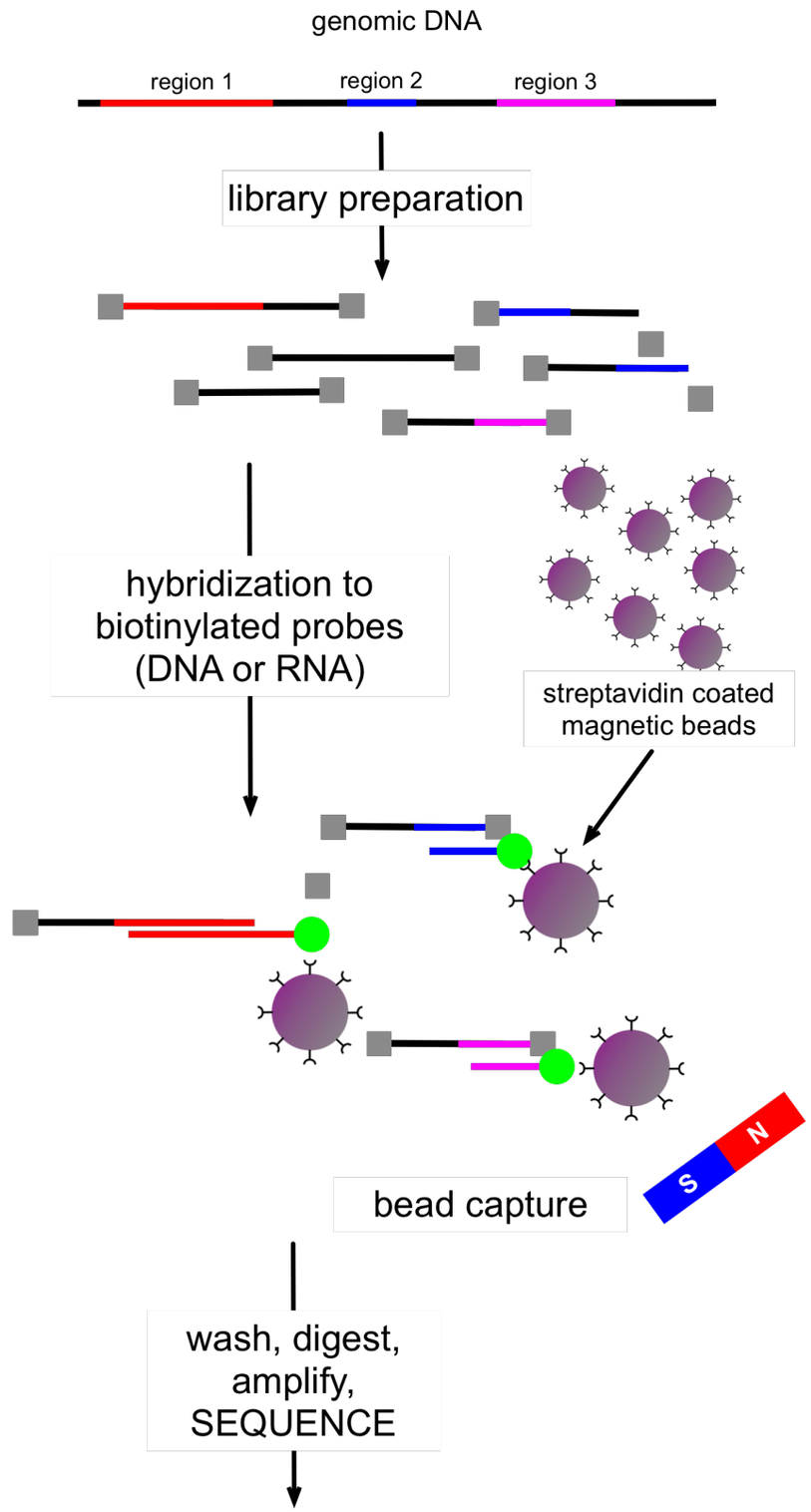
Hybrid Capture (since ~2010)

"subsetting the genome"

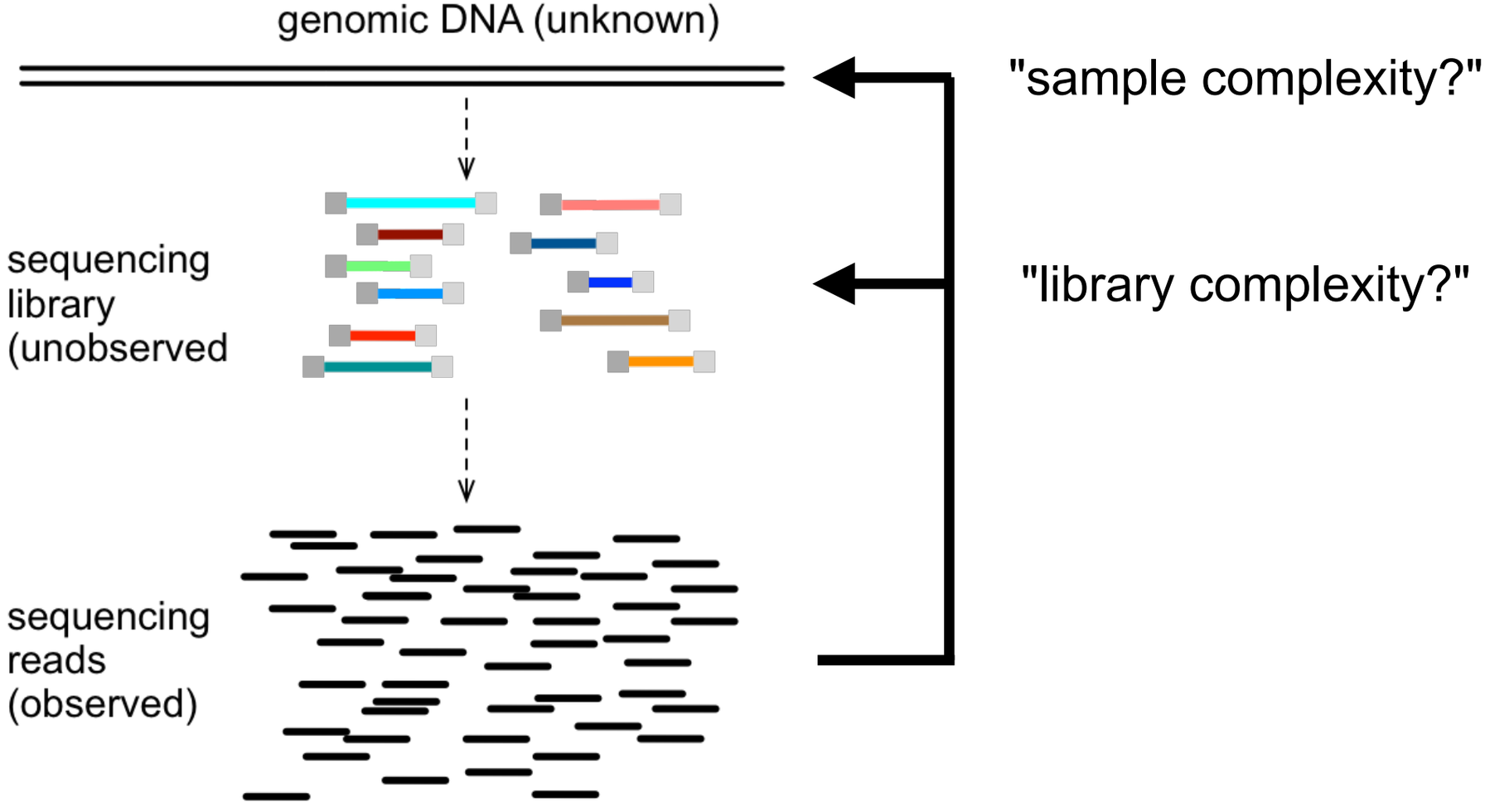
- **Hybrid capture:** fragments from a whole genome library are selected by using "*probes*" corresponding to targets
- DNA-DNA or DNA-RNA
- Probes are biotinylated, enabling selection from solution with streptavidin magnetic beads
- below 3-4 Mb of target sequence, target capture sequencing is not efficient, off-target effects etc.



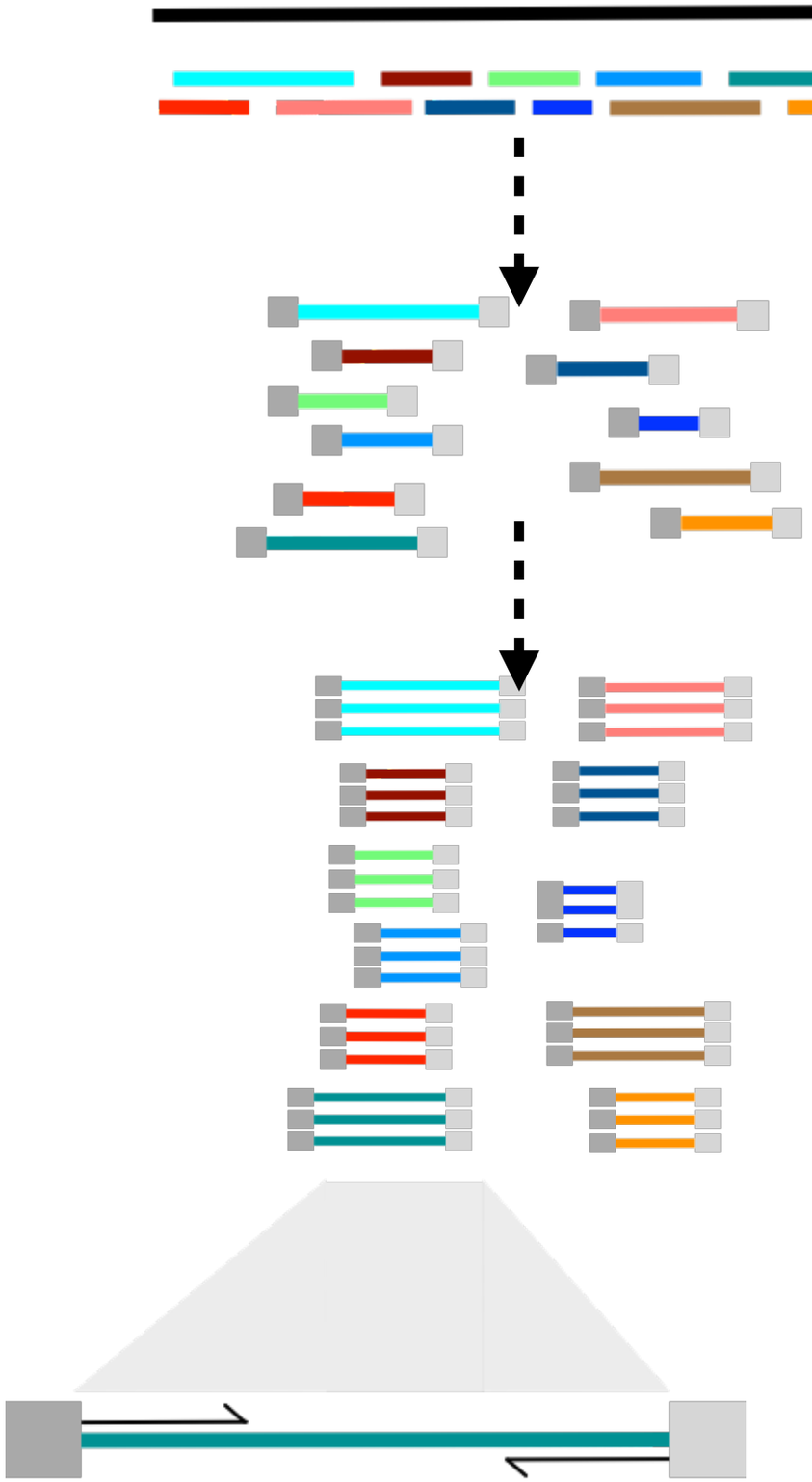




Library Complexity



Library Complexity



library complexity=11

Modeling Library Complexity: Poisson Model

Sequencing a library is *sampling* from it:estimate complexity from sequencing data

C: library complexity, sequence diversity of molecules, #types of molecules

N: number of reads we have

Modeling Library Complexity: Poisson Model

Sequencing a library is *sampling* from it:estimate complexity from sequencing data

C: library complexity, sequence diversity of molecules, #types of molecules

N: number of reads we have

- Let X denote the R.V. for the number of times we sequence/sample a specific molecule, then we have $X \sim Binomial(N, p = \frac{1}{C})$

Modeling Library Complexity: Poisson Model

Sequencing a library is *sampling* from it:estimate complexity from sequencing data

C: library complexity, sequence diversity of molecules, #types of molecules

N: number of reads we have

- Let X denote the R.V. for the number of times we sequence/sample a specific molecule, then we have $X \sim Binomial(N, p = \frac{1}{C})$
- Can use Poisson approximation to Binomial: $X \sim Poisson(\lambda = \frac{N}{C})$

Modeling Library Complexity: Poisson Model

Sequencing a library is *sampling* from it:estimate complexity from sequencing data

C: library complexity, sequence diversity of molecules, #types of molecules

N: number of reads we have

- Let X denote the R.V. for the number of times we sequence/sample a specific molecule, then we have $X \sim Binomial(N, p = \frac{1}{C})$
- Can use Poisson approximation to Binomial: $X \sim Poisson(\lambda = \frac{N}{C})$
- We don't see what we donot sequence: use a truncated Poisson-we only observe events that happened between a and b times

Truncated Poisson distribution:

$$Poisson(x_i \mid \lambda) = \frac{1}{K_{a,b}(\lambda)} \cdot \frac{e^{-\lambda} \cdot \lambda^{x_i}}{x_i!}; K_{a,b}(\lambda) = \sum_{x=a}^{b} P(x_i \mid \lambda)$$

Cohen et al, Estimating parameters in a conditional Poisson distribution. JASA 1960

Modeling Library Complexity: Poisson Model

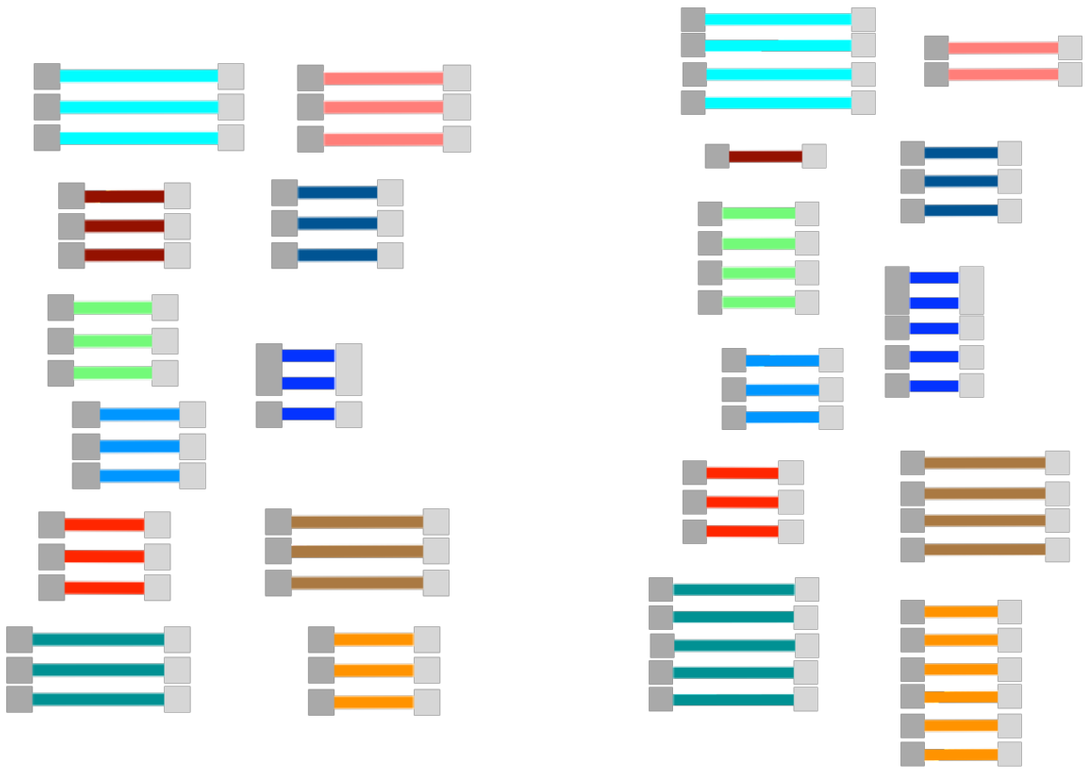
- Maximum Likelihood estimate of the library size:

$$\hat{C} = \frac{M}{K_{a,b}(\lambda)} \approx \frac{M}{1 - Poisson(0,\lambda)}$$

where M is the number of unique sequences

- Poisson model underestimates library complexity: non-uniformity in the original population
  - PCR bias
  - repeats...
  - Poisson:

$$mean = variance = \lambda$$



Modeling Library Complexity: Negative Binomial Model

$$Poisson(x;\lambda) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$ ← Poisson sampling models sequencing

$$Gamma(x;\alpha,\beta) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$$ ← models the entire population (library preparation)

Gamma distribution is a conjugate prior for Poisson:

$$NegBinomial(y;\alpha,\beta) = \int_{0}^{\infty} Poisson(y;x) Gamma(x;\alpha,\beta) \cdot dx$$

Complexity Estimate using Negative Binomial Model

$$P(x_i \mid \lambda, k) = NegBinomial(x_i \mid \lambda, k)$$

dispersion, sampling rate variance (latent variable)

$$= NegBinomial(x_i \mid n, p)$$

$$n = \frac{1}{k} \qquad p = \frac{\lambda}{(\lambda + 1/k)}$$

C: library complexity

N: number of reads we have

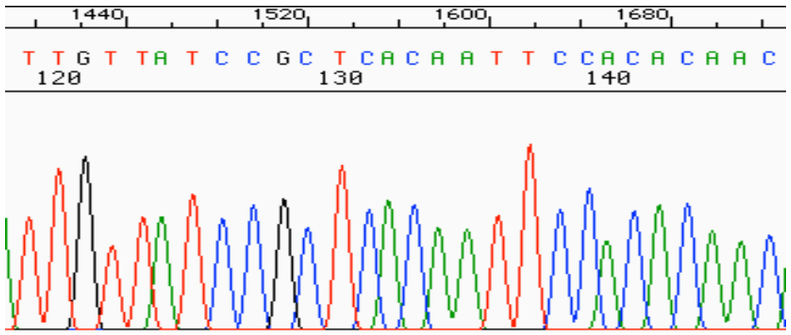
M: number of unique sequences

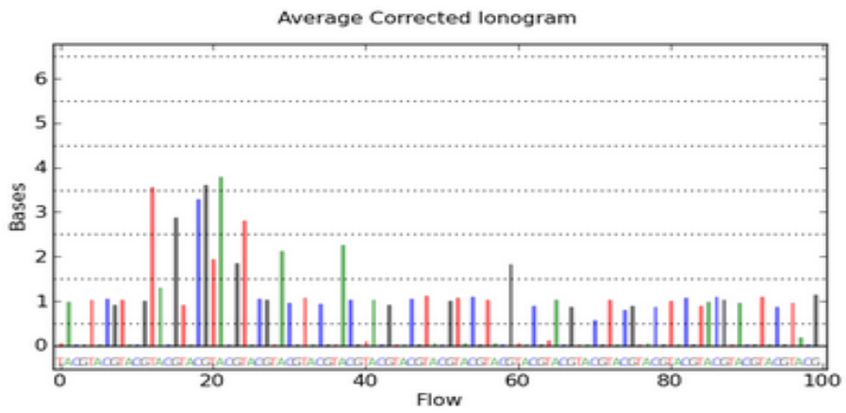
$$\text{M: } (1 - NegBinomial(0 \mid \lambda, k)) * C$$

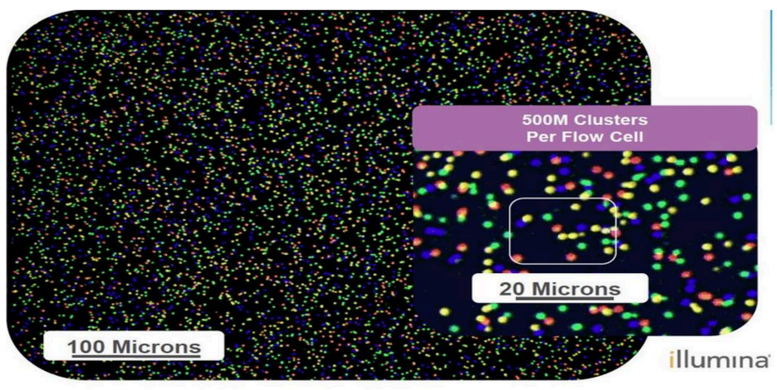
Sequencing Data Aspects

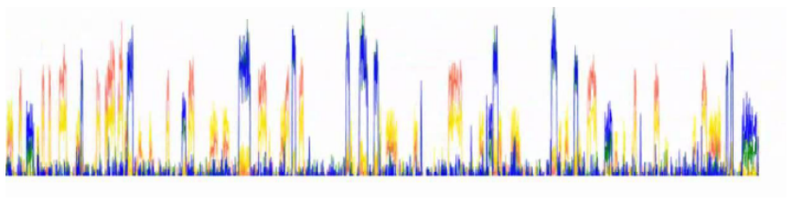
Basecalling

How do we translate the machine output to base calls?









...TGTAGCAGAGAAGACGCCTACTGAATT...

How do we quantify base quality?

FASTQ format & base qualities

```
@J00113:349:HMJHHBBXX:2:1101:12165:4110 1:N:0:TCACTCGA+TCGAGTGA
CTGTCGACCCGCGGAAAGCTTTGGAGCTACACCGAAAACCGGTACGCCCCGCCACCGCCGTAC
+
?AF<FJFJJJJFJJJJJJJJJJJJJAFJJJFJJJJFJJJJFJJJJJJJJAJJJJJJJFJJJJJJ
```

? = ASCII code 63

Base quality = ASCII code - 33 = 30

base quality	P_{error}
3	50%
5	32%
10	10%
20	1%
30	0.1%
40	0.01%

$$Q_{phred} = -10 \log_{10}(P_{error})$$

Error rates

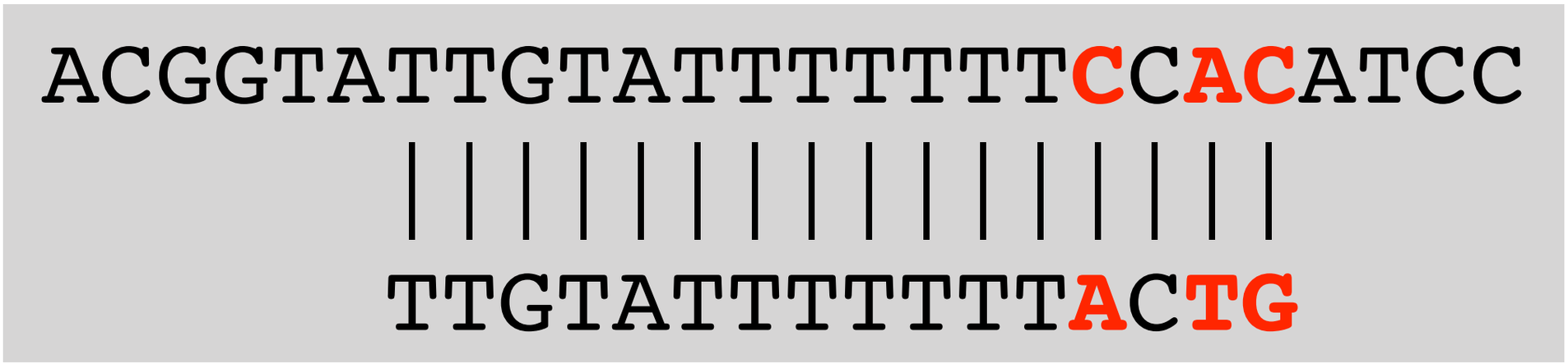
	Primary Errors	Single-pass Error Rate (%)	Final Error Rate (%)
3730xl (capillary)	substitution	0.1-1	0.1-1
Illumina	substitution	~0.1	~0.1
PacBio RS	indel	~13	<=1
Oxford Nanopore	deletions	>=4	4
Ion Torrent	indel	~1	~1

http://www.molecularecologist.com/next-gen-table-3c-2014/

Challenges: Sequencing by Synthesis

illumina®

Sequencing errors tend to be more prominent at the end of the reads



sequencing errors

Challenges: Single Molecule

PACBIO®

dominated by *indel* errors

reference

```
ACGGTATTGTATTTTTTTCCACATCC
|||||||||||||||||||||||||
TTGTAT-TT-TTCC-C
```

deletion errors

Challenges: Ion Semiconductor Sequencing

ion torrent
by Thermo Fisher Scientific

difficult to estimate the length of long homopolymers

homopolymer run

reference

```
ACGGTATTGTATTTTTTTTCCACATCC
||||||||||||||||||||||||||
TT--ATTTTT--CCAC
```

gap1 gap2

Take Home: Planning Experiments

Considerations

- **Number of biological replicates needed**
  - biological variability & technical "noise"
  - sequencing depth (effect size: will I see differences at this coverage?)
  - sample heterogeneity
- **Sequencing decisions (every company will tell you they have the greatest technology)**
  - coverage, coverage, coverage!: number of reads/sample (->sequencing depth)
  - read length
  - base-level quality: get your money's worth
  - paired end vs single end
  - be attentive about batch effects
  - consider library complexity

Sequencing experiment starts before sequencing:
PLAN, THINK, REVIEW and PLAN again