

Welcome to Genome assembly and annotation workshop

August 6-9, 2018

Center for Agricultural Biotechnology
Kasetsart University, Kamphaeng Saen Campus

DNA Sequencing Technology

Pichahpuk Uthaipaisanwong, Ph. D

6.08.2018

Kasetsart University, Kamphaeng Saen Campus
Genome assembly and annotation workshop

Content

- DNA
 - Structure and function
- Sequencing technology time line
- First generation sequencing
 - Sanger method (The chain termination)
- Second generation sequencing
 - Roach (454 pyrosequencing)
 - Thermo Fisher Scientific (Ion Torrent)
 - Illumina
- Third generation sequencing
 - Pacific Biosciences (PacBio)
 - Oxford Nanopore Sequencing

DNA

- Deoxyribonucleic acid

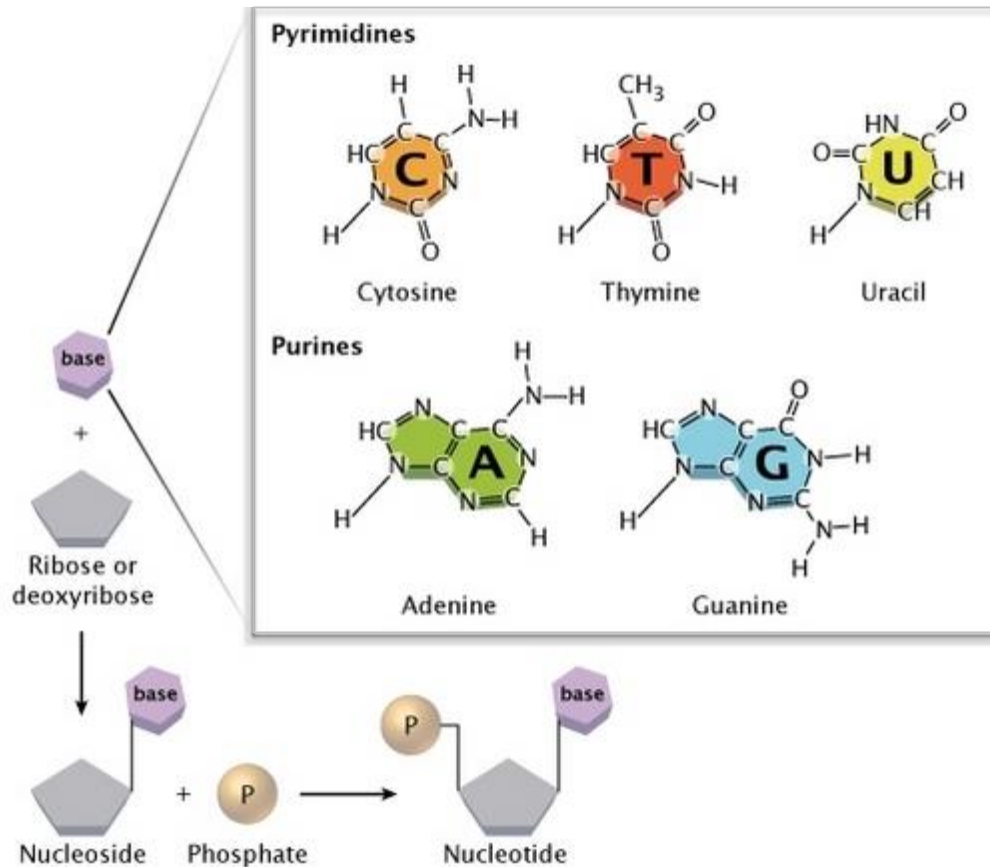


Figure1: DNA structure

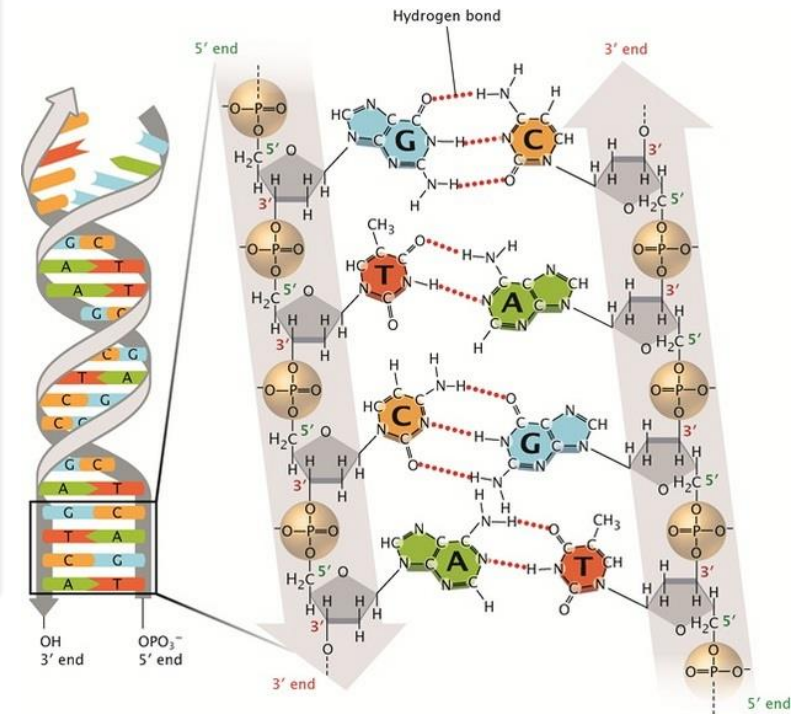


Figure2: Base pairing in DNA

Characteristics of DNA

- stable
- transmitted from parent to progeny without change (carry information and inherit to child)
- capable of being expressed (gene function)
- allowed for information to change (mutation)
- capable of accurate replication

Where is/are genetic material?

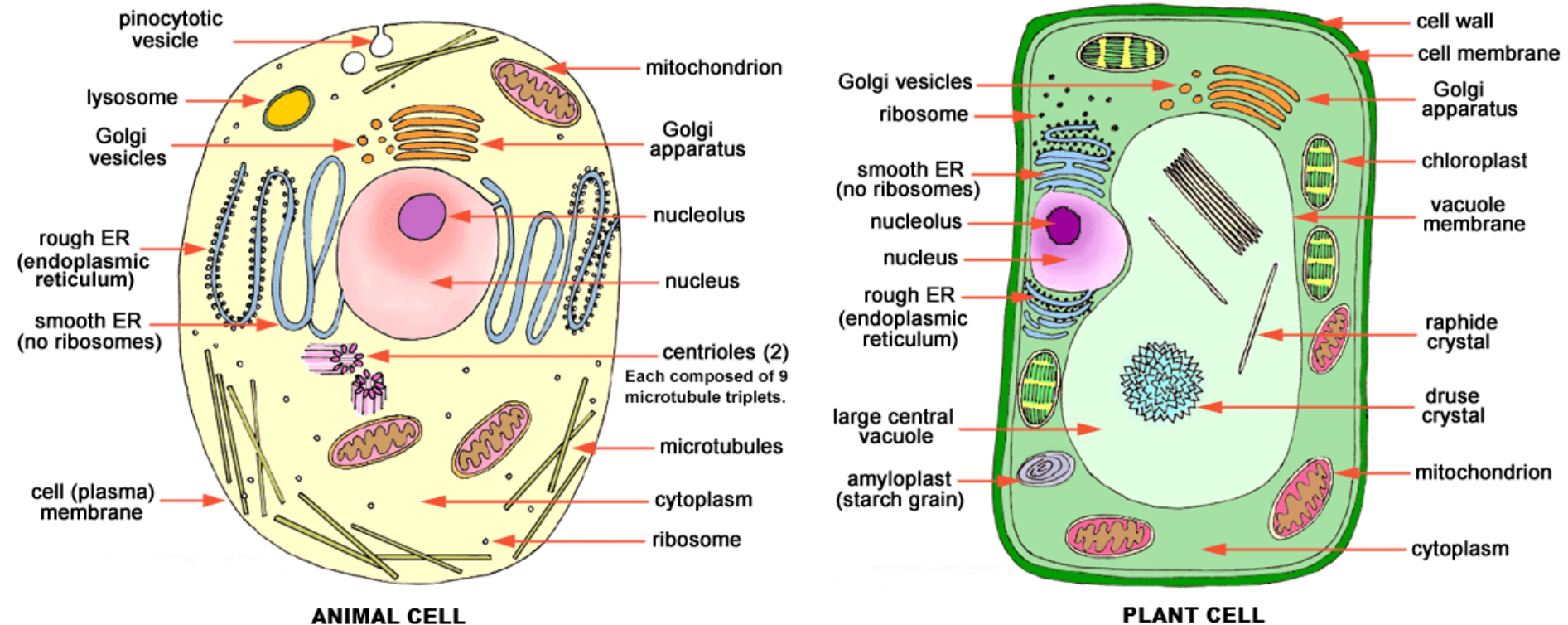


Figure3 : plant and animal cell

Genome

- A genome is the complete set of genetic information in an organism. It provides all of the information required by an organism to function.

DNA sequencing

- DNA sequencing is technology that allows researchers to determine the order of bases in a fragment of DNA sequence

Sequencing Technology Timeline

Approaching to Next Generation Sequencing (NGS)

1980

1990

1995

2000

2005

2010

2015

Gel-based sequencing

Capillary Sequencing

Next generation Sequencing

1977 Sanger sequencing method by F. Sanger

1983 PCR by K. Mullis

1953 Discover of DNA structure
by J. Watson and F. Crick

1990-2003

Nature, 409 (6822): 860-921, 2001
"Initial Sequencing and Analysis of the Human Genome"
Science, Vol 219(5507): 1304-1351, 2001
"The Sequence of the Human Genome"

Human genome Project, > 2 billion \$

1993 Development of pyrosequencing

Single molecule emulsion PCR

1998



2005

454 GS 20 Sequencer
(First NGS sequencer)



2006

Solexa genome analyzer
(First short-read NGS sequencer)

Genome Analyzer IIx

2006

Illumina acquires Solexa

ABI SOLID

2007



ABI SOLID V4

(Short read sequencer based on ligation)

GS FLX sequencer

2008

(NGS with 400-600 pb read length)

NGS Human genome sequencing

2008

HiSeq2000

(200 Gbp per Flow cell)



Illumina HiSeq 2000

Ion Torrent, PGM/Proton

2010



PGM

(ion semiconductor sequencing)

Pacific Biosciences, RS

2011

(single-molecule real-time (SMRT) sequencing)

Oxford Nanopore Technologies

2014

Nanopore sequencing

Personal genome,
100\$ within 6 hours

Proc Natl Acad Sci U S A. 1977 Dec; 74(12): 5463-5467.
DNA sequencing with chain-terminating inhibitors
(citation: 69,576)

Nature, 171 (4356): 737-738, 1953
Molecular structure of nucleic acid

Slide was modified from Jonathan Eisen

First-generation sequencing

- Sanger sequencing method (Frederick Sanger;1977)
- Chain Termination or Dideoxy method
 - Utilizes 2',3'–dideoxynucleotide triphosphate

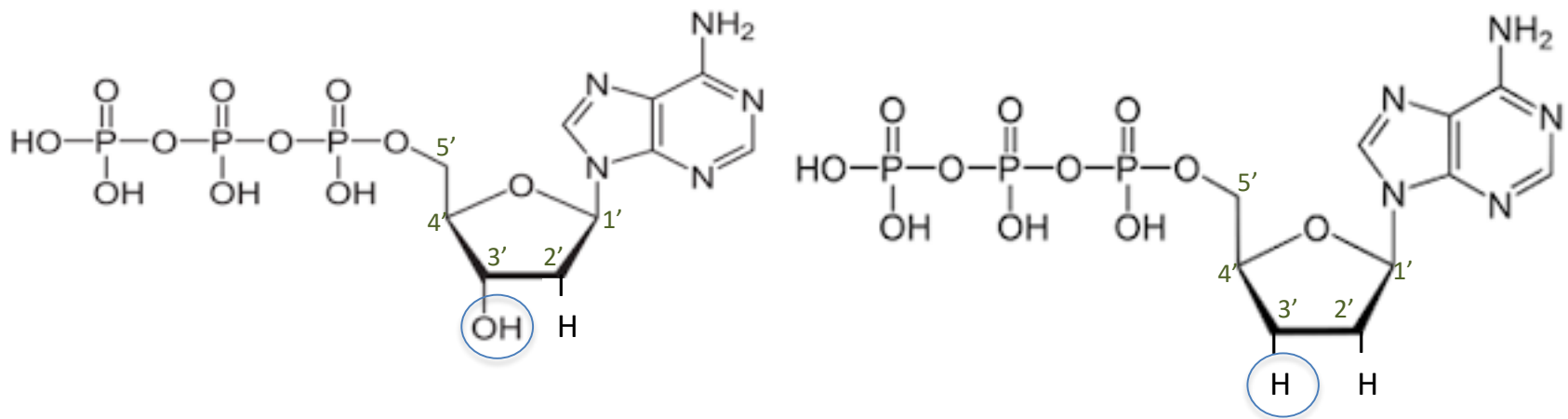
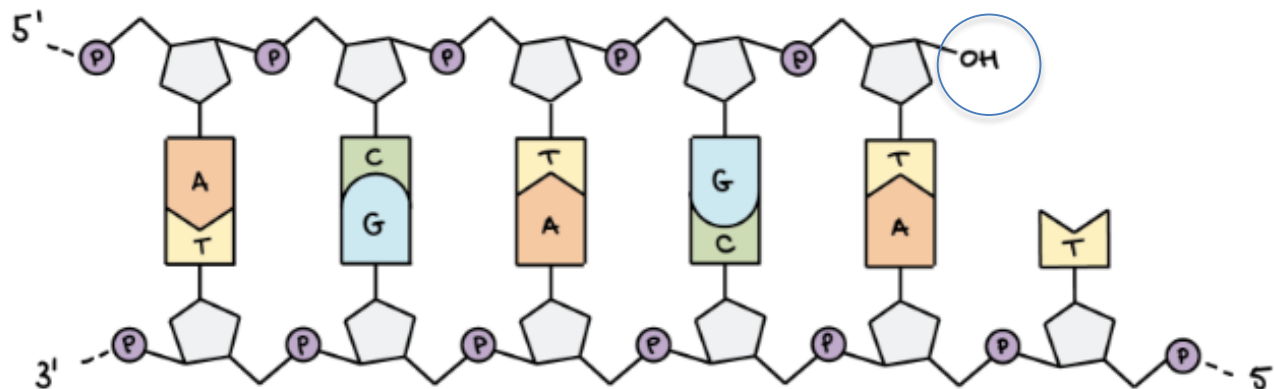
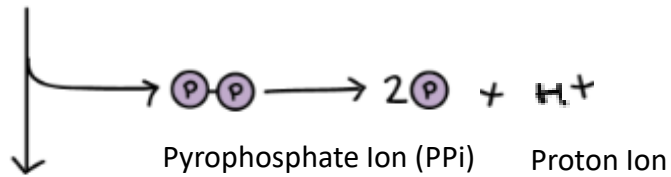
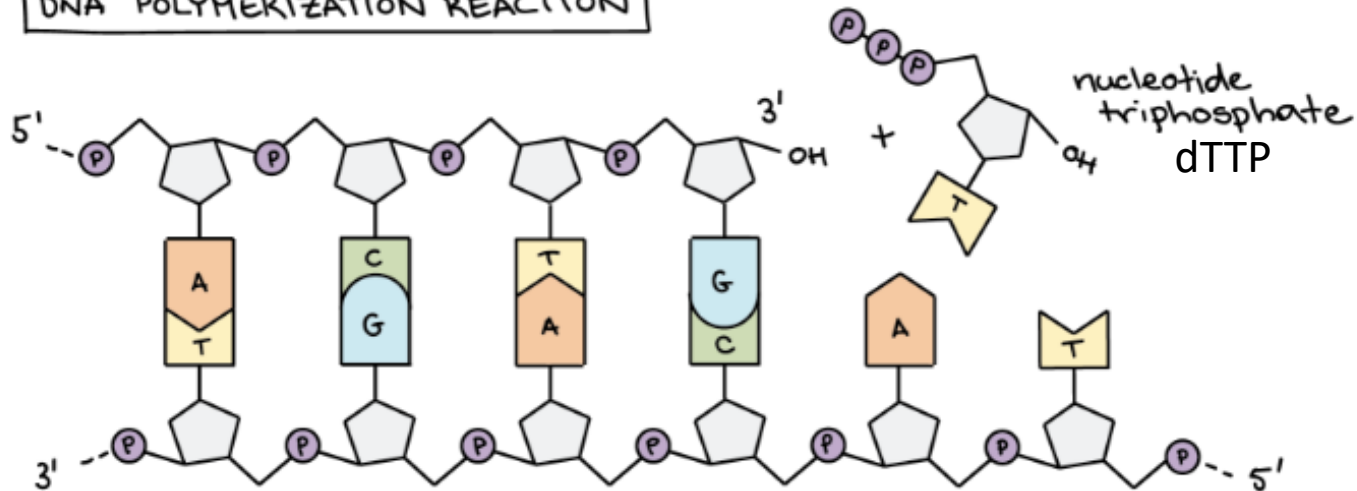
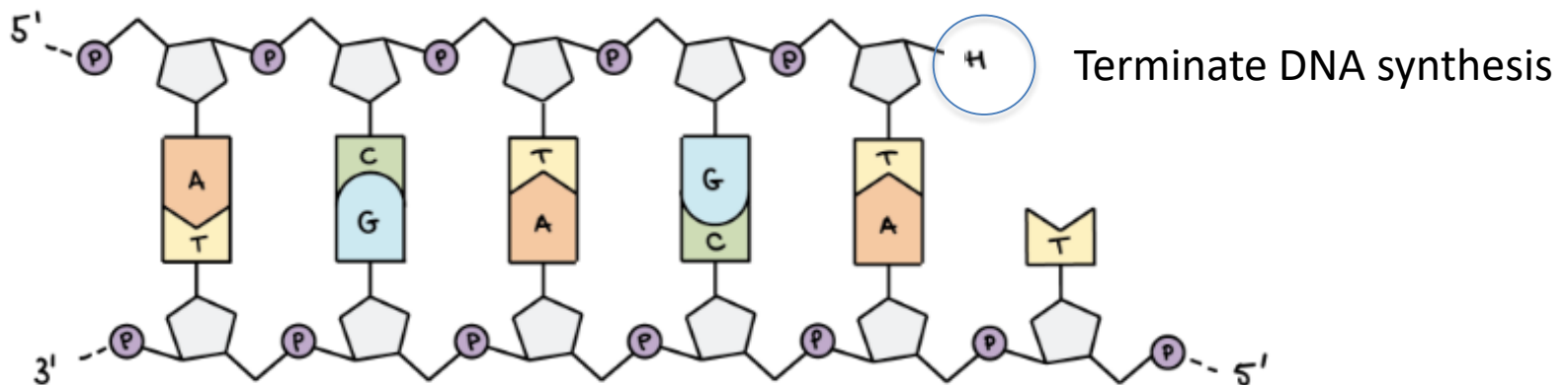
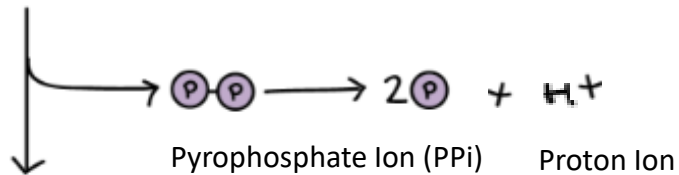
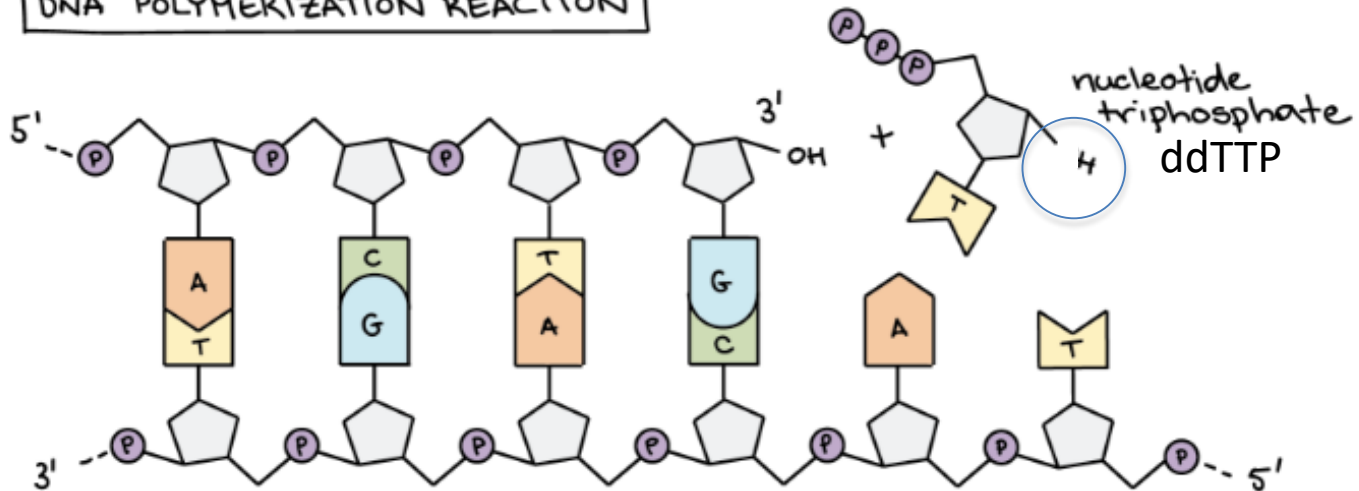


Figure: The structure of dNTP (left) and ddNTPs (right)

DNA POLYMERIZATION REACTION



DNA POLYMERIZATION REACTION



Sanger sequencing: gel-based method

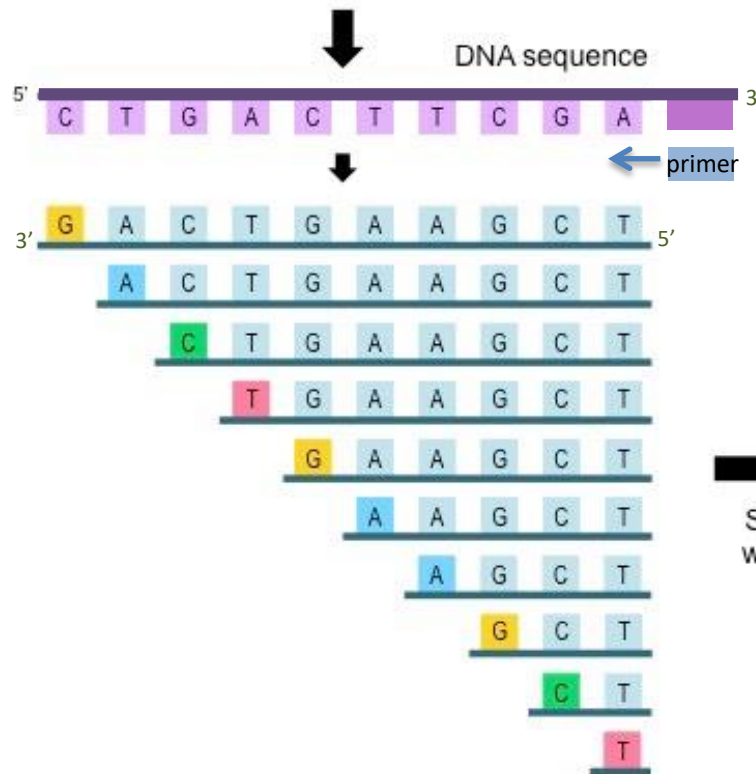
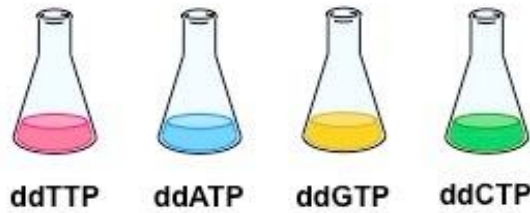
Sequencing Reaction mixture

-DNA template

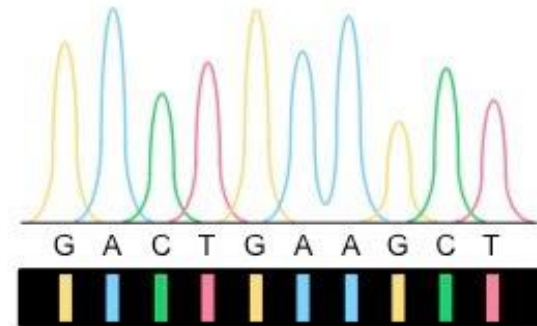
-Primer 4 x Mixture (+ one dideoxynucleotide)

-DNA polymerase

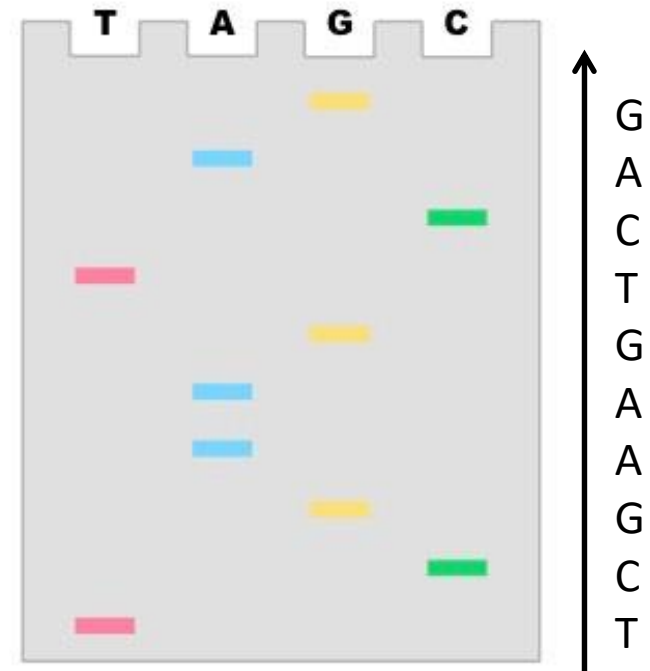
-dNTPs



Use a sequencing machine



Separate with a gel

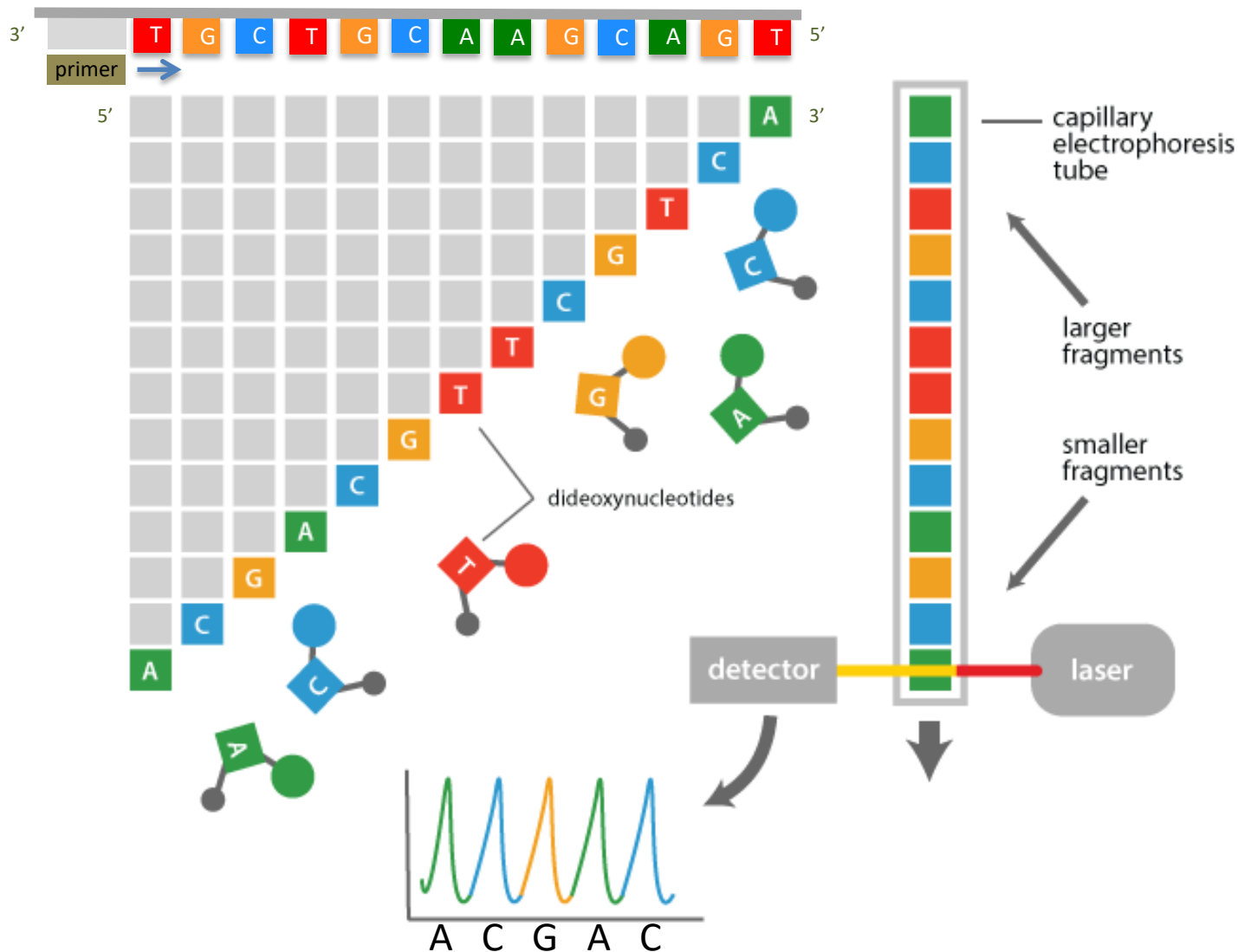


dNTPs = (dATP, dCTP, dGTP, dTTP)

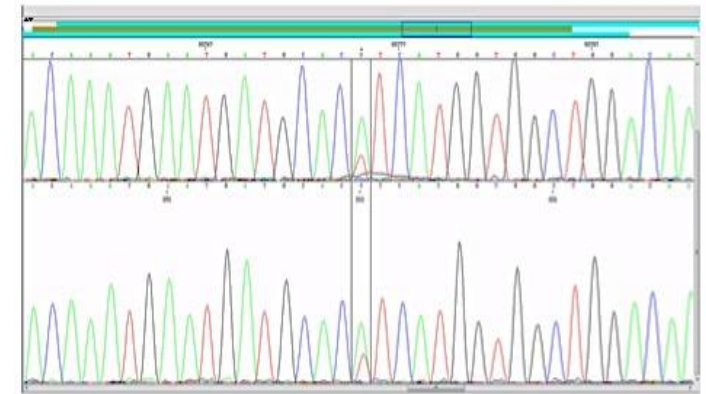
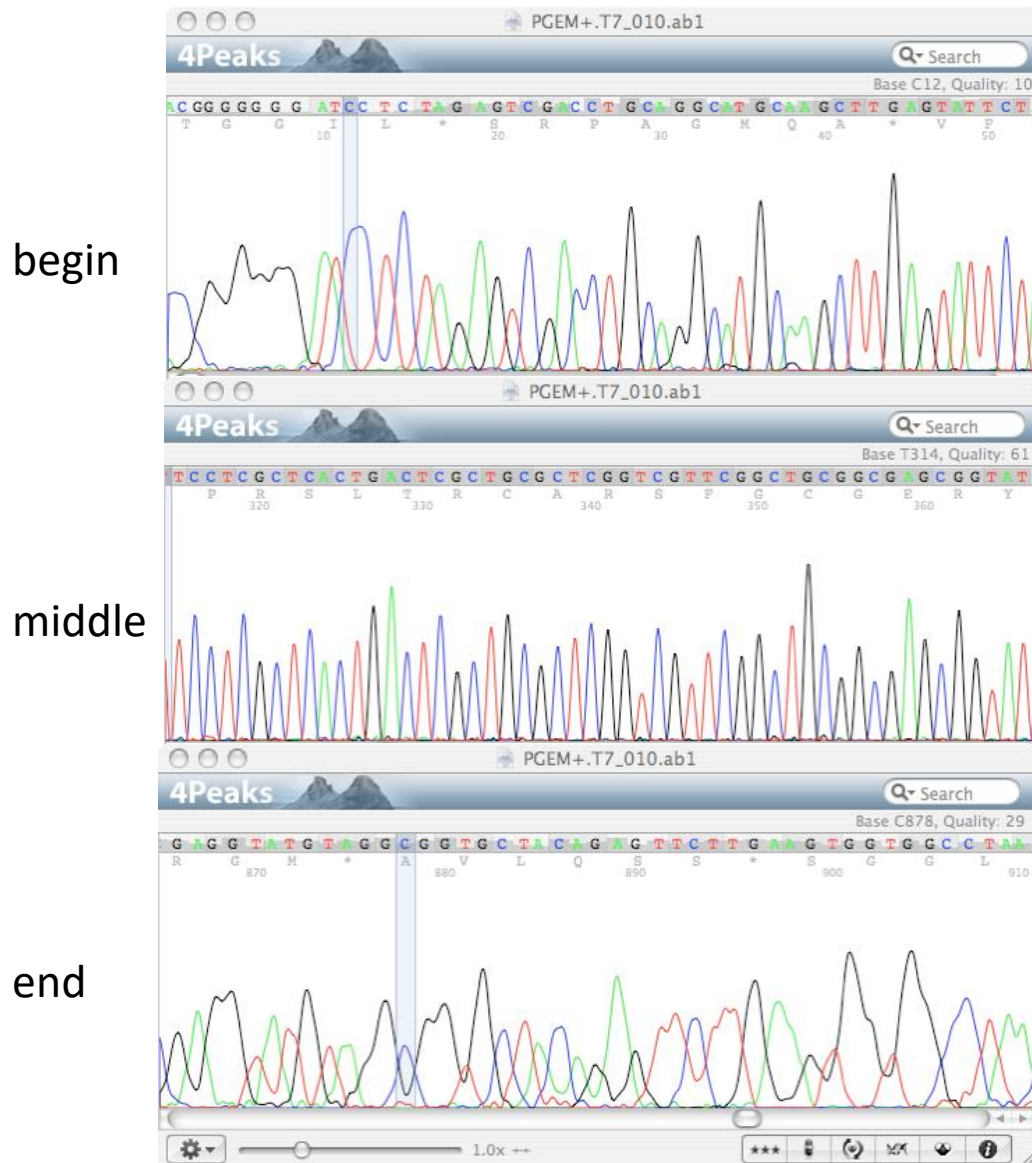
Figure was modified from http://ib.bioninja.com.au/_Media/dna-sequencing_med.jpeg

Sanger sequencing: capillary based method

Automation of sanger sequencing 1995-present



Sanger sequencing: output of the data



mutation

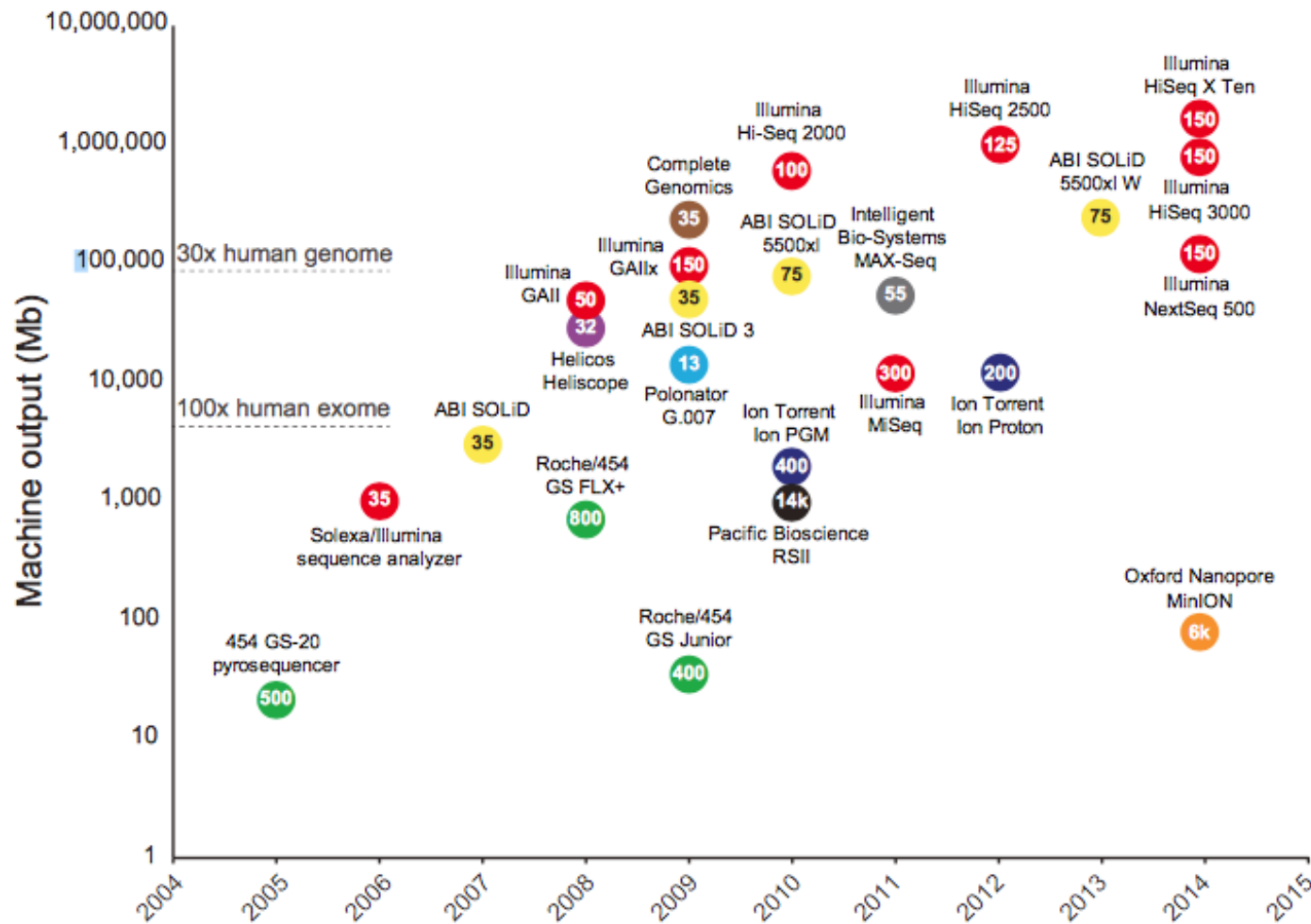
Figure: example of the output of the sanger sequencing

Sequencing Platform	Advantages	Disadvantages
Sanger sequencing	<ul style="list-style-type: none">- Lowest error rate- Long read length (up to 1000 bp)- Gold standard method	<ul style="list-style-type: none">- High cost per base- Long time to generate data- Need for cloning- Amount of data per run (1 seq / run)

Second generation sequencing

- Next generation sequencing (NGS)
- Deep sequencing
- High-throughput sequencing
- Massive parallel sequencing
- Whole genome sequencing
- Rapidly dropping price and time
- Ability to produce data quantity and quality

Timeline and comparison of commercial HTS Instrument



Plot of commercial release dates versus machine outputs per run

1,000Mb = 1Gb;

10,000Mb = 10Gb;

100,000Mb = 100Gb;

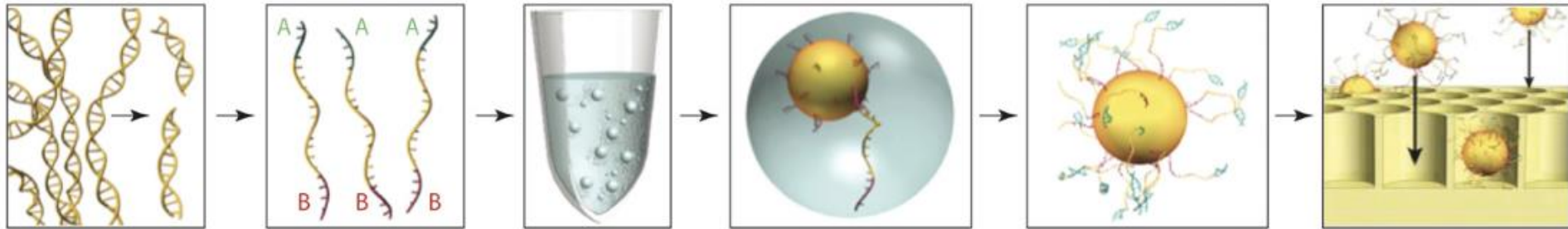
1,000,000Mb = 1,000Gb = 1Tb

454 pyrosequencing

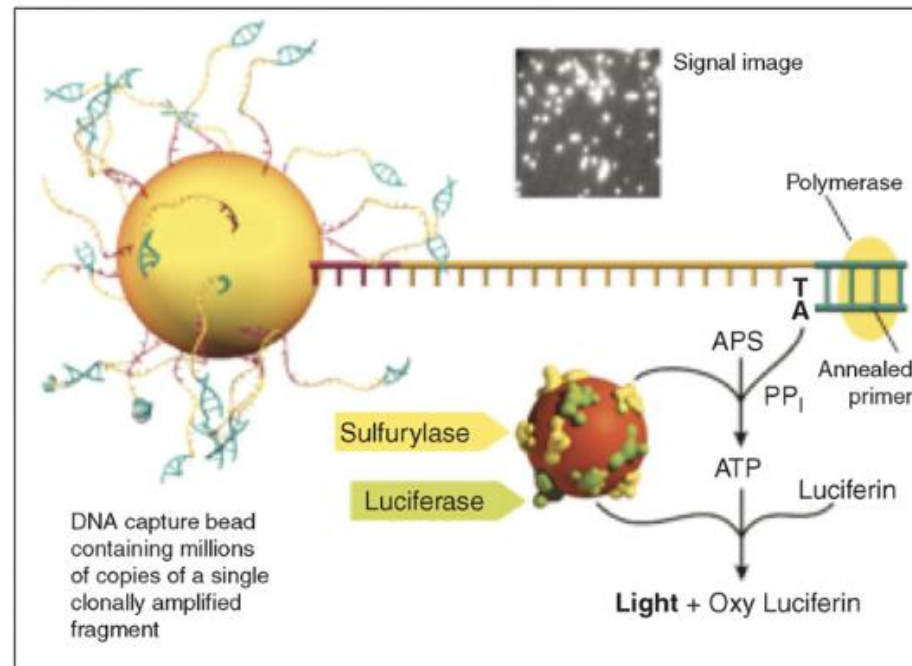
Library construction

Emulsion PCR

PTP loading



454 GS-FLX Titanium



Pyrosequencing reaction

454 pyrosequencing

DNA
Bead

dTTP
• Polymerase adds nucleotide (dNTP) (1)

Polymerase

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

Annealed Primer

APS

PP_i

(2) • Pyrophosphate is released (PP_i)

Sulfurylase
Luciferase

Enzyme Bead

ATP (3)

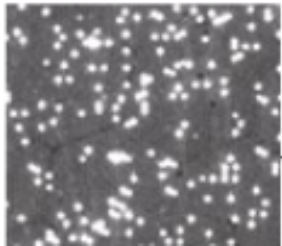
• Sulfurylase creates ATP from PP_i and APS

luciferin (4)

• Luciferase hydrolyses ATP to oxidize luciferin and produce light

(5) CCD camera detects bursts of light

Light + oxy luciferin



454 pyrosequencing

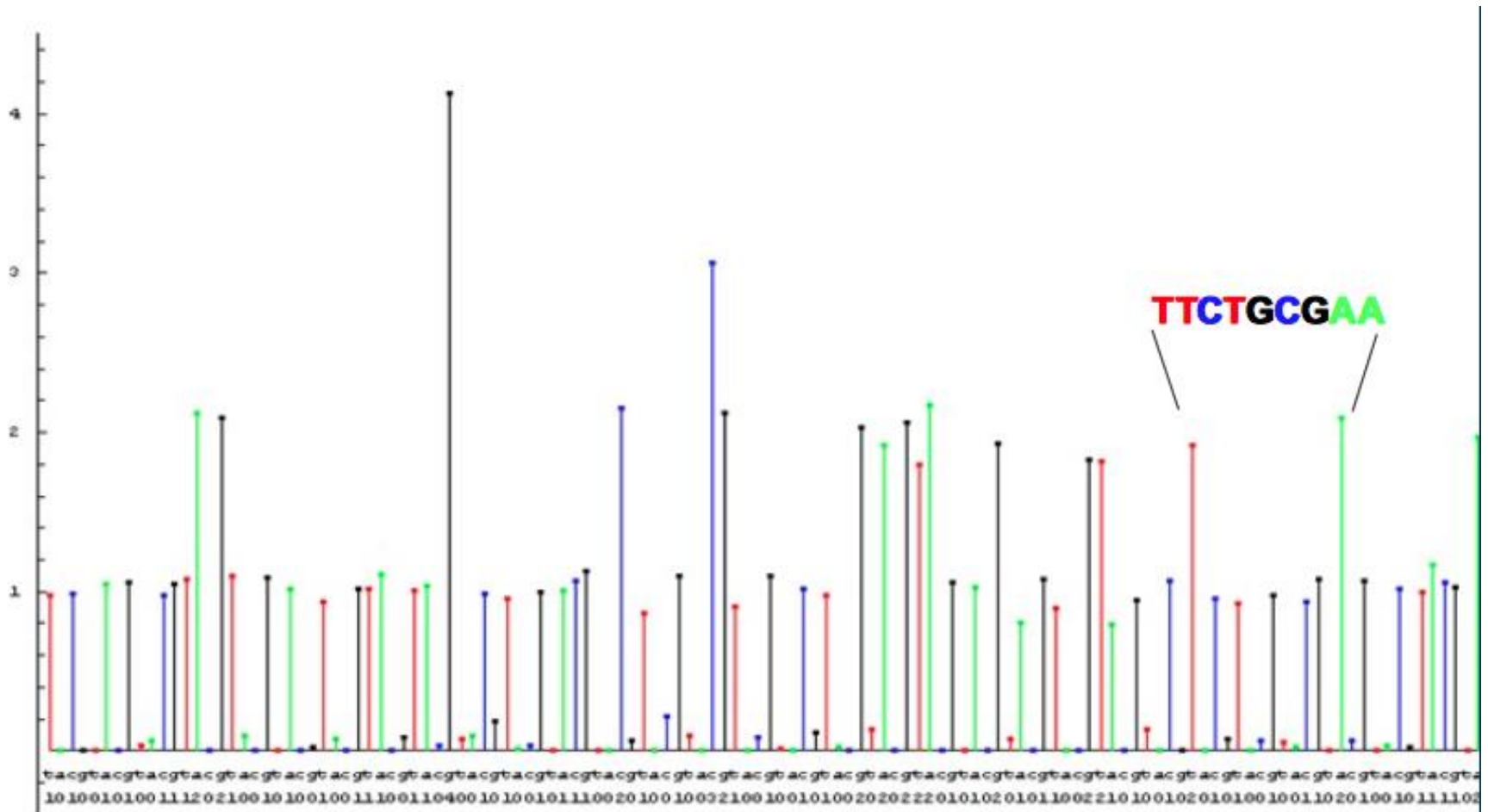
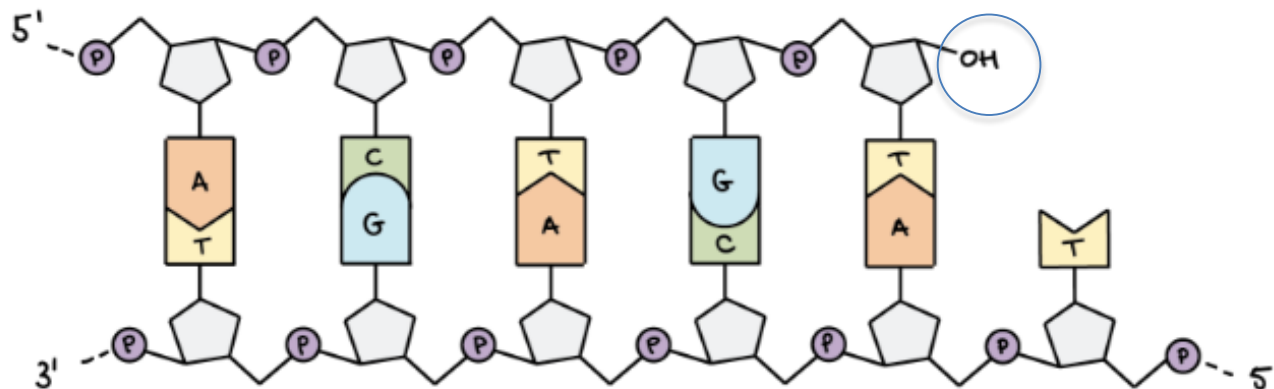
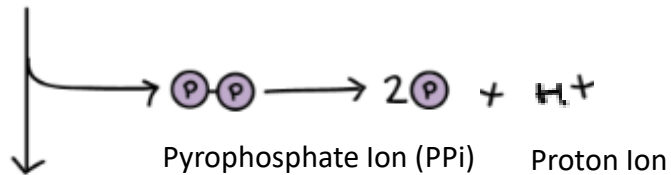
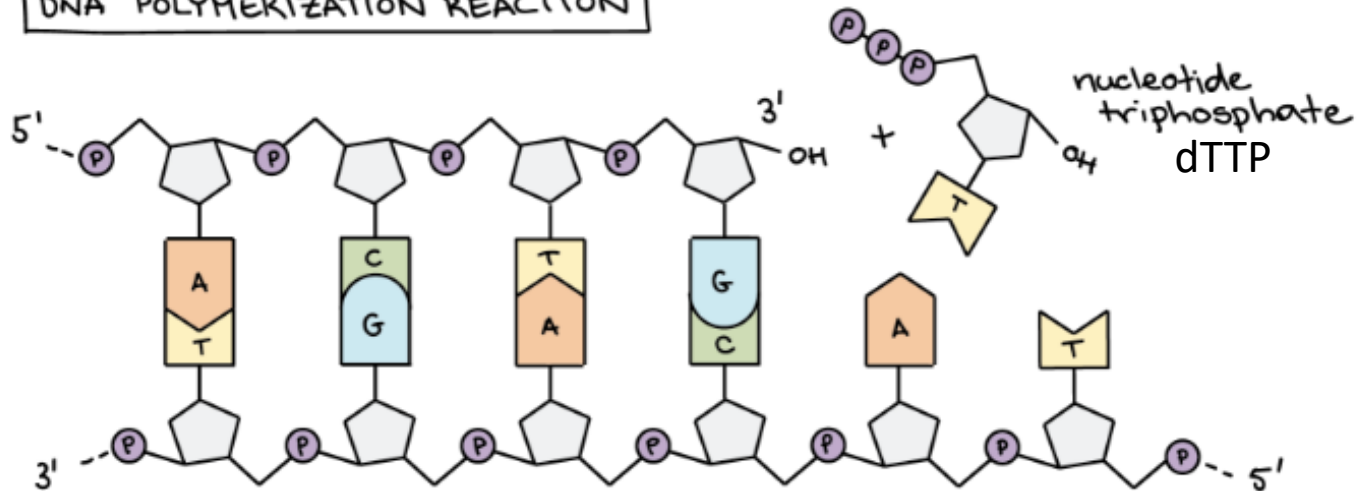
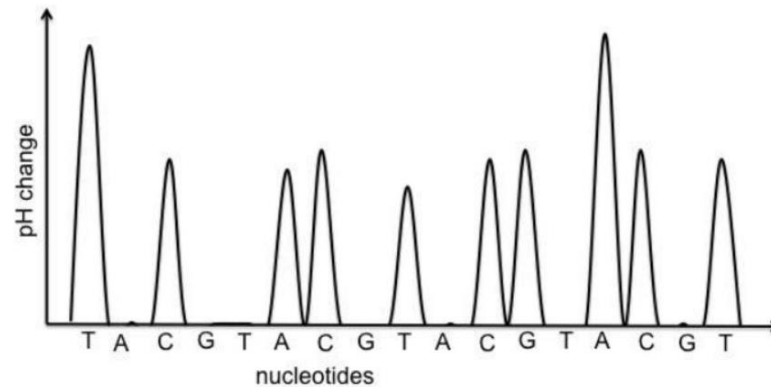
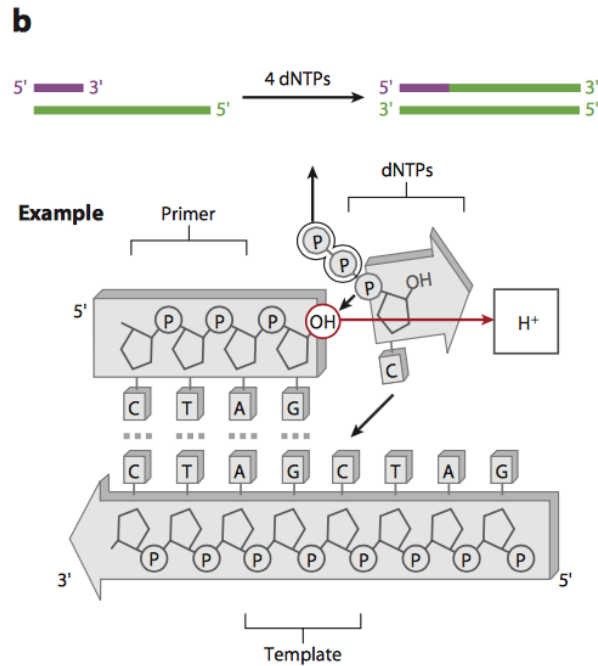


Figure: Base calling via flow gram

DNA POLYMERIZATION REACTION



- Platforms: PGM/Proton



Ion GeneStudio S5 Series | Flexible Portfolio Configurable to Your Needs



Ion GeneStudio™ S5



Fast

Ion GeneStudio™ S5 Plus



Flexible.

Ion GeneStudio™ S5 Prime



Powerful.



**Ion 510™
Chip**
2–3 M reads
Up to 400 bp



**Ion 520™
Chip**
3–6 M reads
Up to 600 bp



Ion 530™ Chip
15–20 M reads
Up to 600 bp



Ion 540™ Chip
60–80 M reads
Up to 200 bp

New



Ion 550™ Chip
100–130 M reads
Up to 200 bp

For Research Use Only. Not for use in diagnostic procedures. * Throughputs based on 200bp sequencing

Ion GeneStudio S5 series

Sequence up to 25 Gb in < 8.5 hours



Chip type	Number of reads	Read length (output*)	Ion GeneStudio™ S5 System	Ion GeneStudio™ S5 Plus System	Ion GeneStudio™ S5 Prime System
			Turnaround time (sequencing run** plus analysis time)		
Ion 510 Chip	2–3 million	200 bp (0.3–0.5 Gb)	4.5 hr	3 hr	3 hr
		400 bp (0.6–1 Gb)	10.5 hr	5 hr	5 hr
Ion 520 Chip	4–6 million	200 bp (0.6–1 Gb)	7.5 hr	3.5 hr	3 hr
		400 bp (1.2–2 Gb)	12 hr	5.5 hr	5.5 hr
	3–4 million	600 bp (0.5–1.5 Gb)	12 hr	5.5 hr	5.5 hr
Ion 530 Chip	15–20 million	200 bp (3–4 Gb)	10.5 hr	5 hr	4 hr
		400 bp (6–8 Gb)	21.5 hr	8 hr	6.5 hr
	9–12 million	600 bp (1.5–4.5 Gb)	21 hr	8 hr	7 hr
Ion 540 Chip	60–80 million	200 bp (10–15 Gb)	19 hr	10 hr	6.5 hr
		200 bp (20–30 Gb) 2 runs in 1 day	NA	20 hr	10 hr†
Ion 550 Chip	100–130 million	200 bp (20–25 Gb)	NA	11.5 hr	8.5 hr
		200 bp (40–50 Gb) 2 runs in 1 day	NA	NA	12 hr†

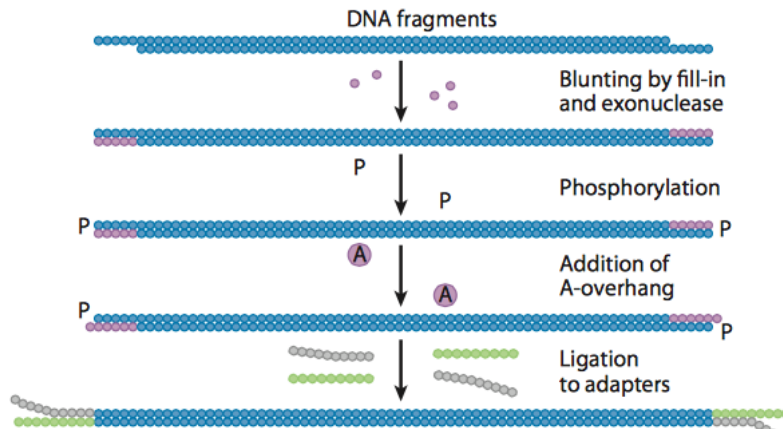
Ion Torrent

- Single end sequencing
- 200/ 400 / 600 bp
- Error at homopolymers size (≥ 7 bp)
- Bacterial whole genome sequencing
- Resequencing

Illumina

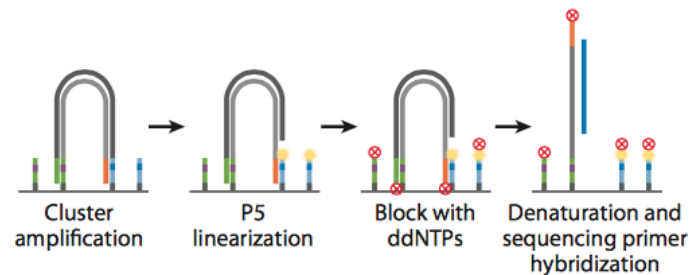
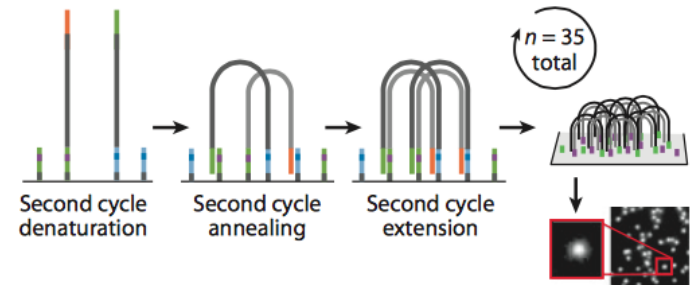
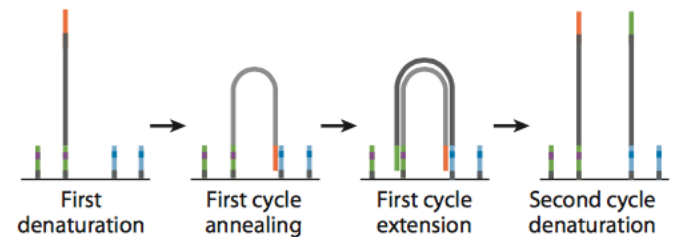
b Cluster generation

a Illumina's library-preparation work flow

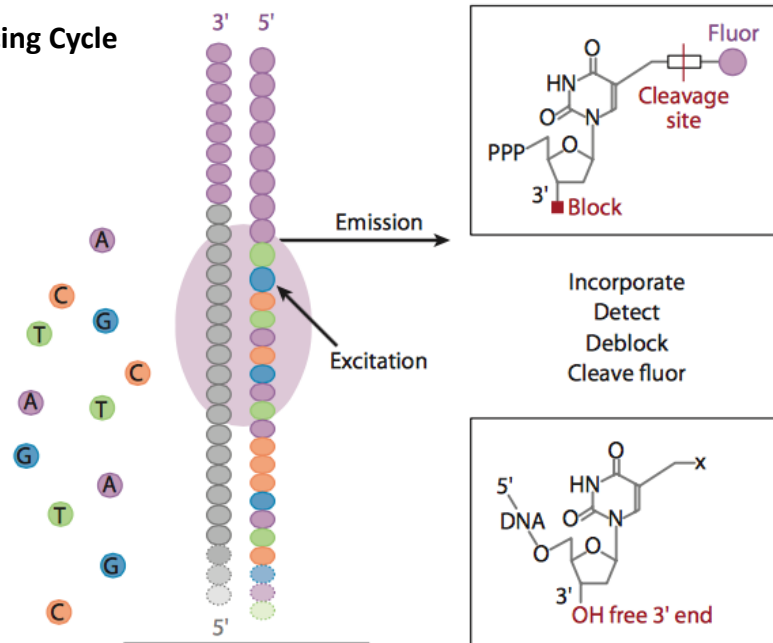


The diagram illustrates the four steps of the DNA microarray synthesis process:

- Grafted flow cell:** Shows a flow cell with two types of reactive groups, P7 (green) and P5 (blue), each with a hydroxyl (OH) group.
- Template hybridization:** Shows the hybridization of P7 and P5 templates to the flow cell.
- Initial extension:** Shows the initial extension of the P7 and P5 templates.
- Denaturation:** Shows the denaturation of the P7 and P5 templates, leaving the synthesized DNA on the flow cell.

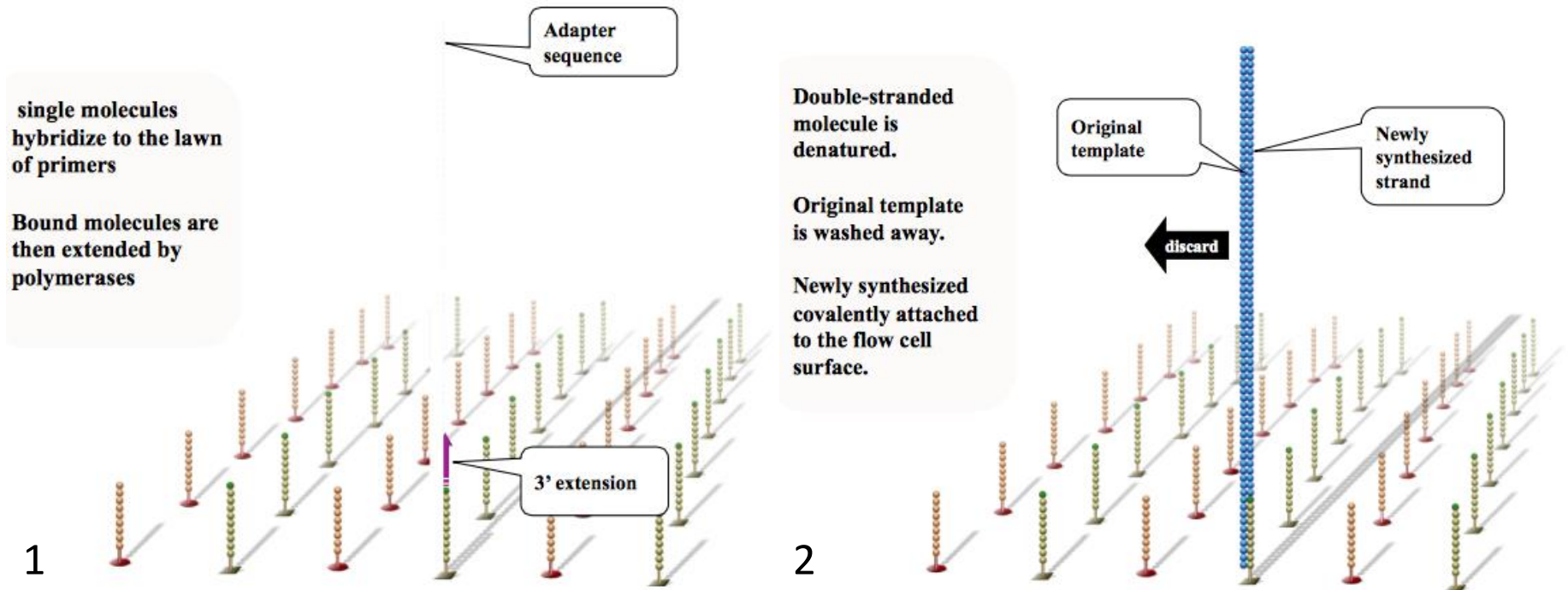


c Sequencing Cycle



Illumina

b Cluster generation



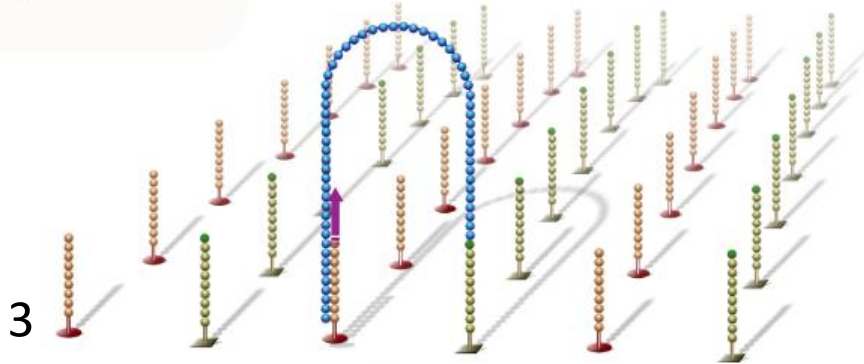
Illumin flow cell

Illumina

Bridge formation

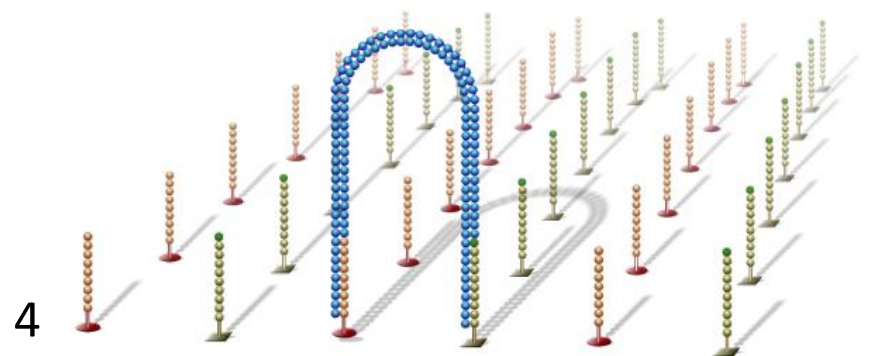
Single-strand flips over to hybridize to adjacent primers to form a bridge.

Hybridized primer is extended by polymerases.



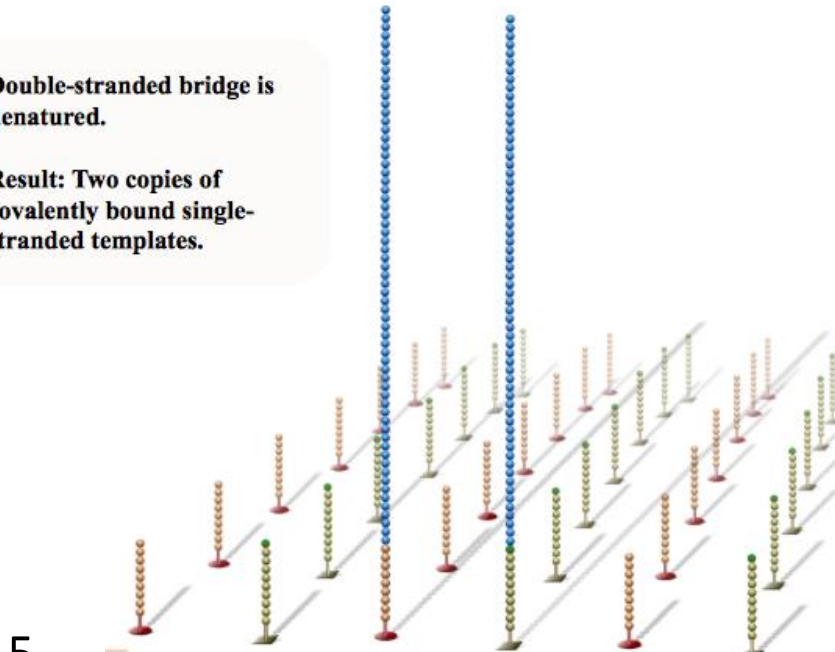
Bridge amplification

→ double-stranded bridge is formed.



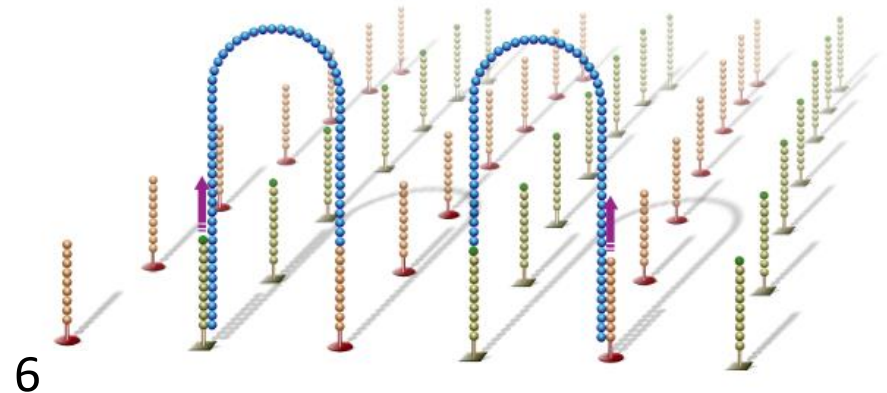
Double-stranded bridge is denatured.

Result: Two copies of covalently bound single-stranded templates.



Single-strands flip over to hybridize to adjacent primers to form bridges.

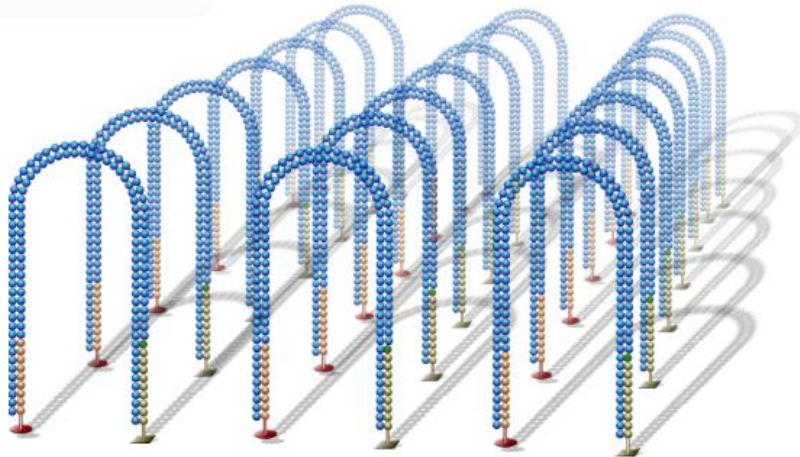
Hybridized primer is extended by polymerase.



Illumina

Bridge amplification
cycle repeated till
multiple bridges
are formed

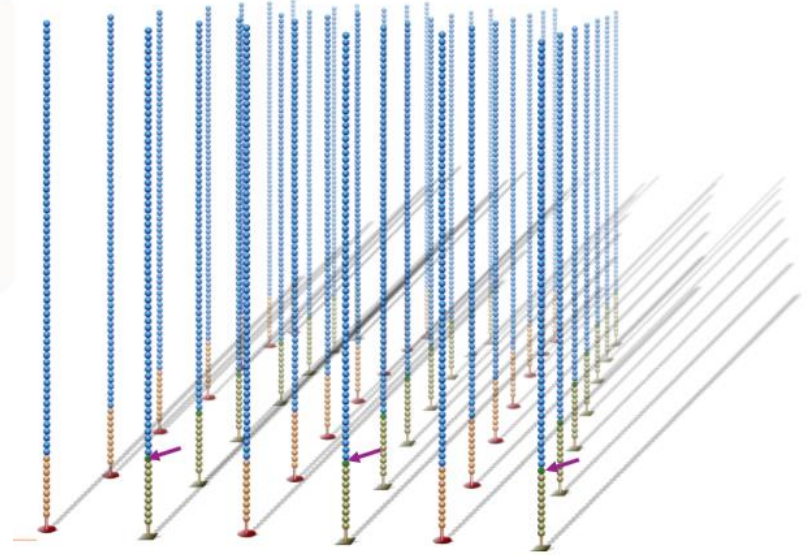
7



dsDNA
bridges
denatured.

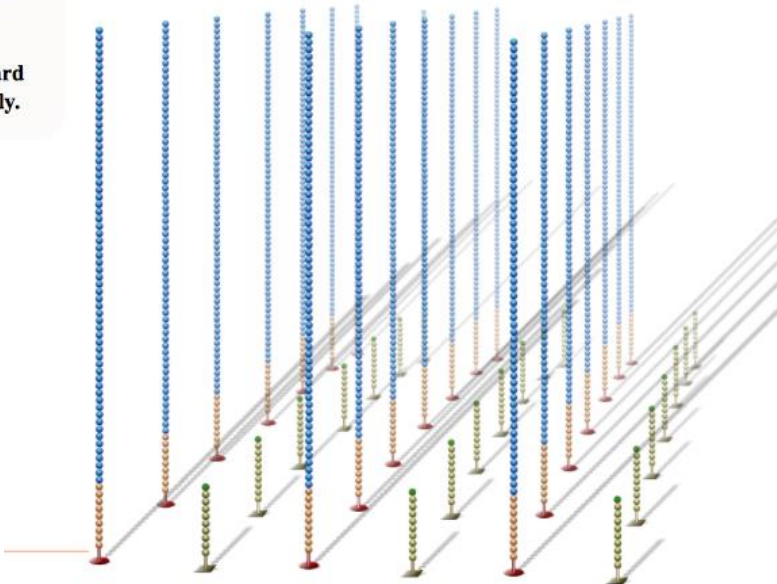
Reverse
strands
cleaved and
washed
away.

8



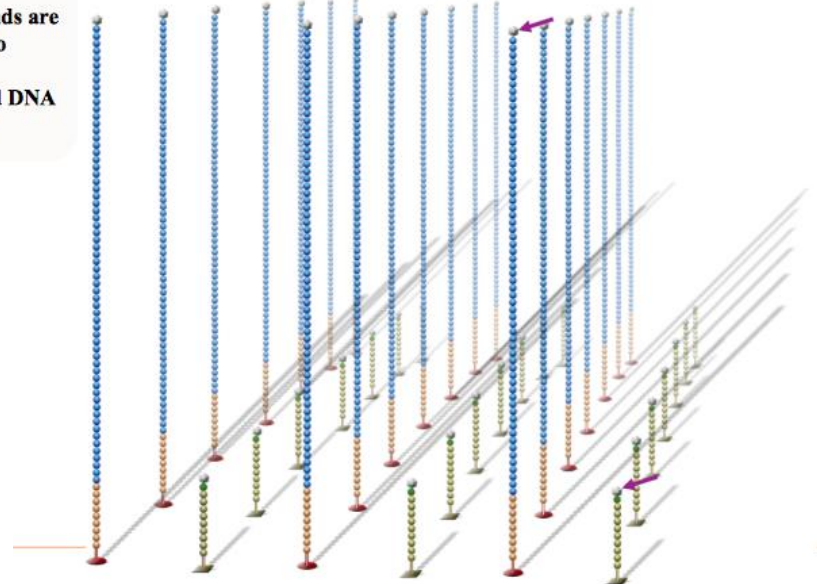
... leaving
a cluster
with forward
strands only.

9

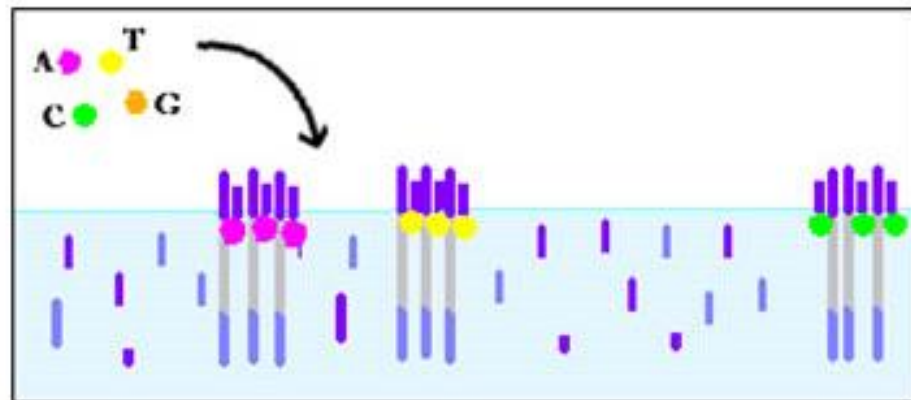
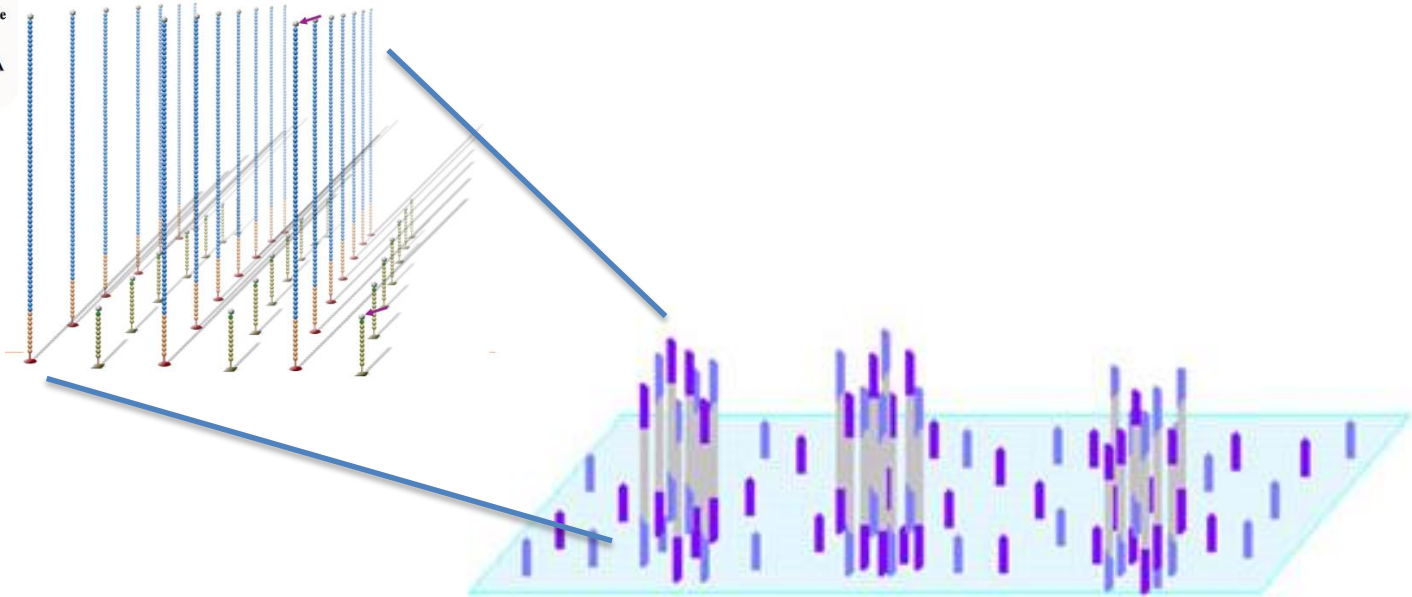


Free 3' ends
are blocked to
prevent
unwanted DNA
priming.

10



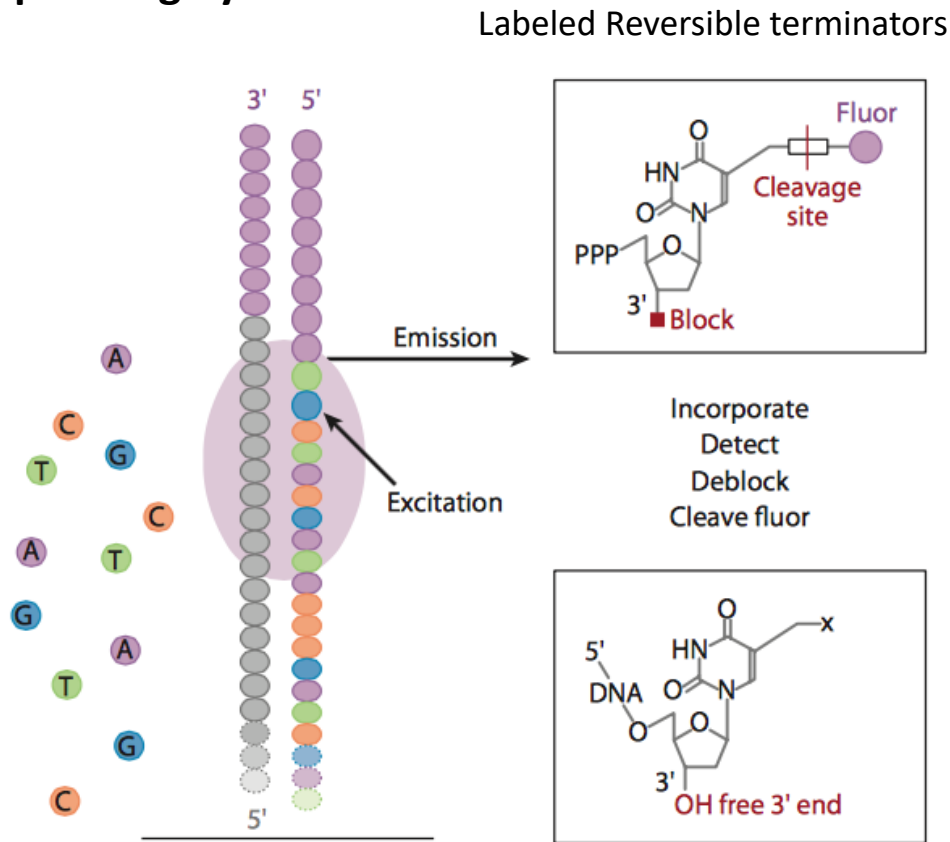
Free 3' ends are blocked to prevent unwanted DNA priming.



LASER

Illumina

C Sequencing Cycle

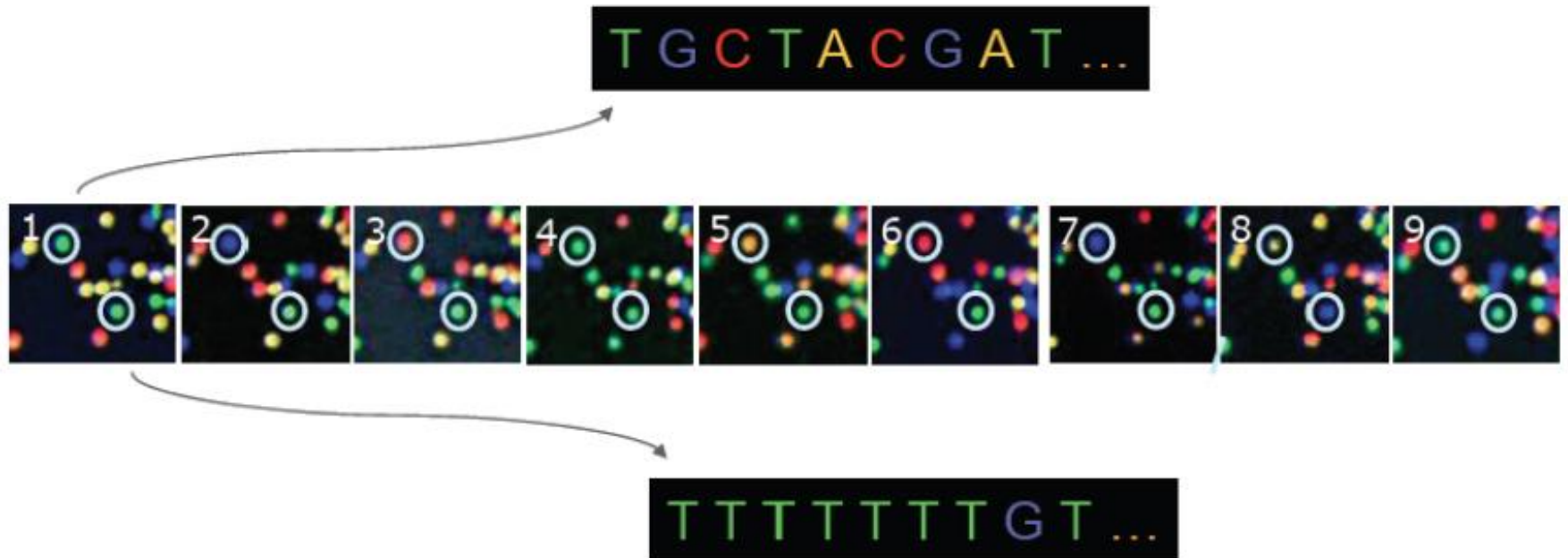


Initiate the first sequencing cycle, all four labeled reversible terminators and DNA polymerase enzyme are first added. Only one base can incorporate at a time

Lasers excite the fluorescent tags and the images are captured via CCD camera. The identify of the first base in each cluster is recorded, and then the fluorescent tag is removed.

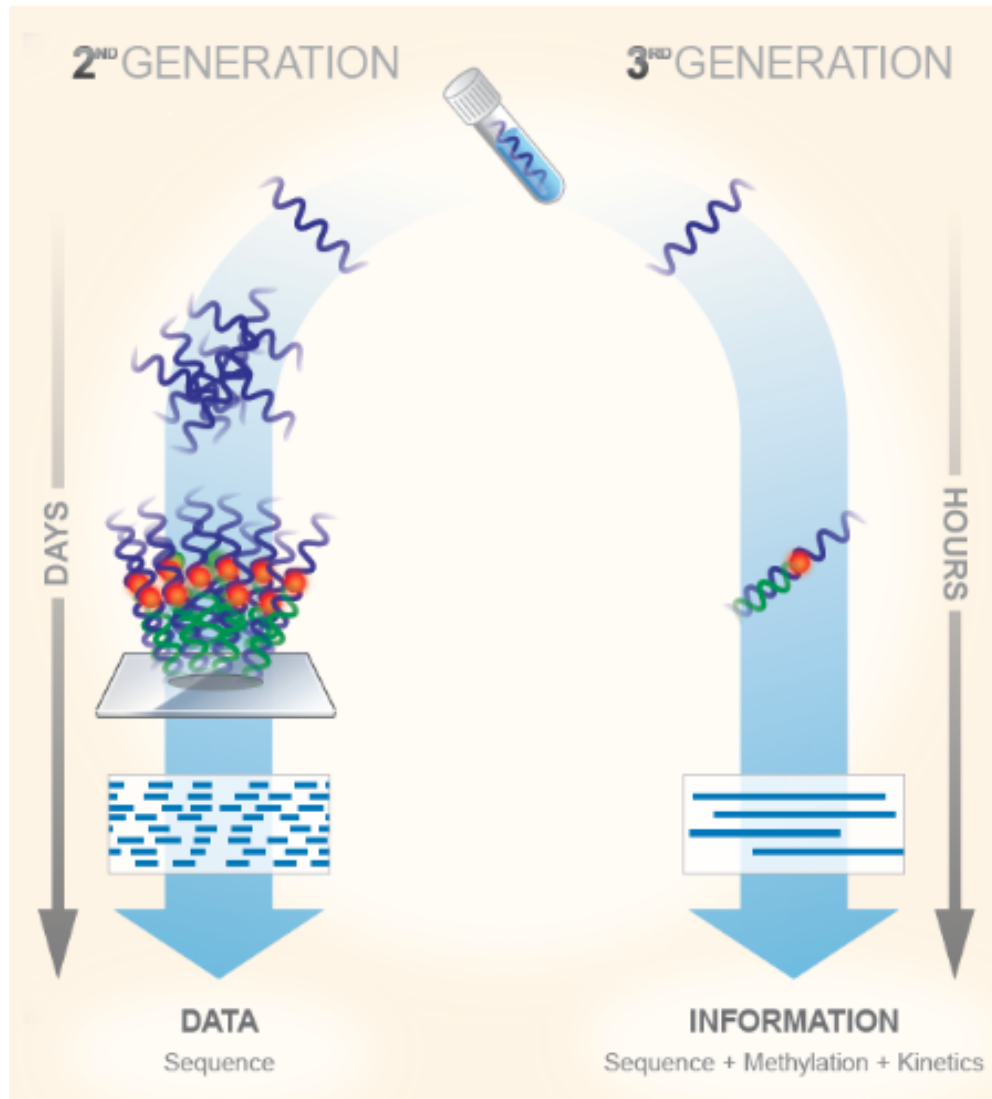
In subsequent cycles, the process of adding sequencing reagents, removing unincorporated bases and capturing the signal of the next base to identify is repeated.

Illumina



Sequencing Platform	Advantages	Disadvantages
Sanger sequencing	<ul style="list-style-type: none"> - Lowest error rate - Long read length (up to 1000 bp) - Gold standard method 	<ul style="list-style-type: none"> - High cost per base - Long time to generate data - Need for cloning - Amount of data per run
454 pyrosequencing	<ul style="list-style-type: none"> - Low error rate - Medium read length (400-800 pb) 	<ul style="list-style-type: none"> - Relatively high cost per base - Must run at large scale - Medium/high start up costs
Ion Torrent	<ul style="list-style-type: none"> - Low start costs - Scalable (10-1000 Mb per run) - Medium/low cost per base - Low error rate - Fast runs (<3 hours) 	<ul style="list-style-type: none"> - Cost not as low as Illumina - Read lengths only (~100-200 pb)
Illumina	<ul style="list-style-type: none"> - Low error rate - Lowest cost per base - Tons of data 	<ul style="list-style-type: none"> - Must run at very large scale - Short read length (50-150 bp) - Run take multiple days - High startup costs - De Novo assembly difficult

Third generation sequencing



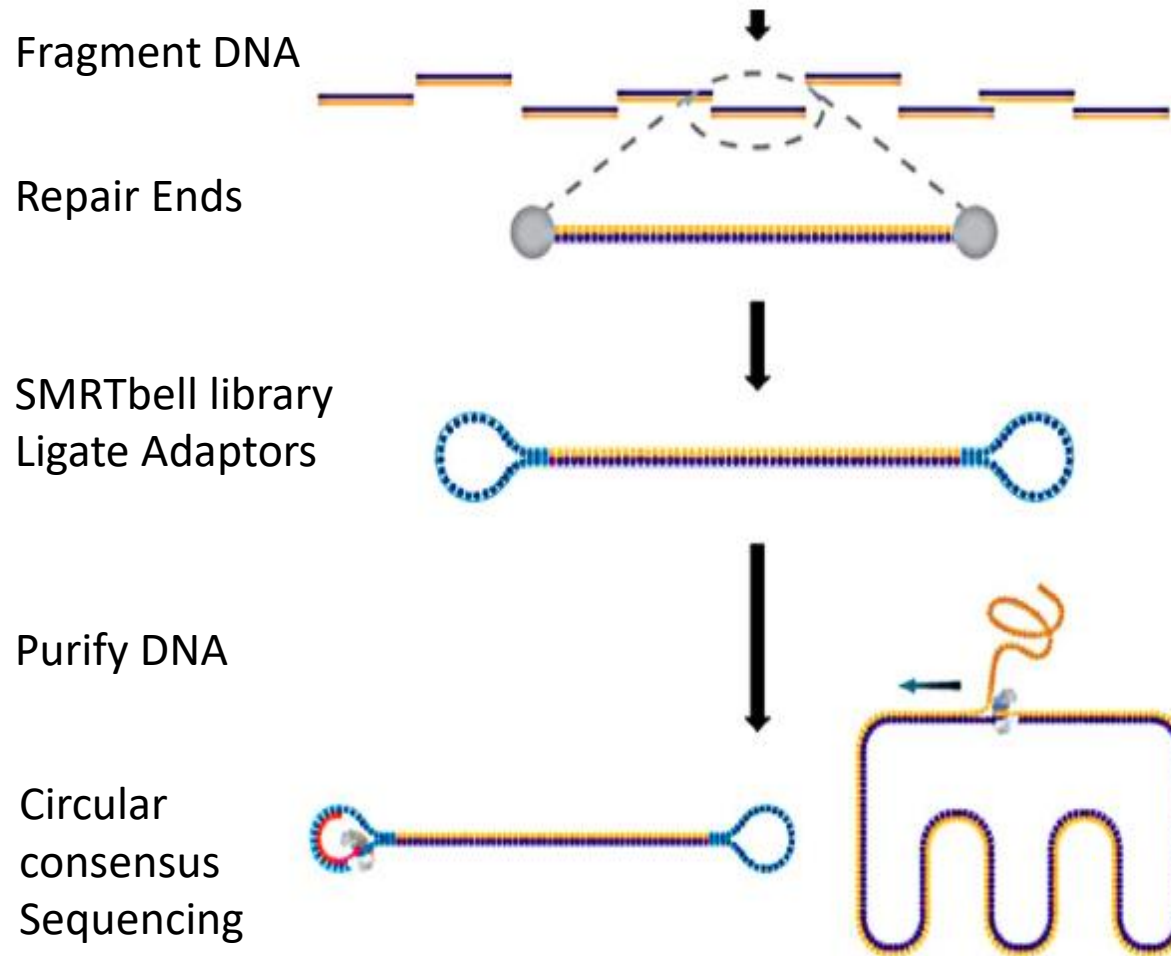
- Single DNA molecules hence much longer fragments are the subject of sequencing
- Real time
- *In situ*

Single molecule real time sequencing (SMRT sequencing)

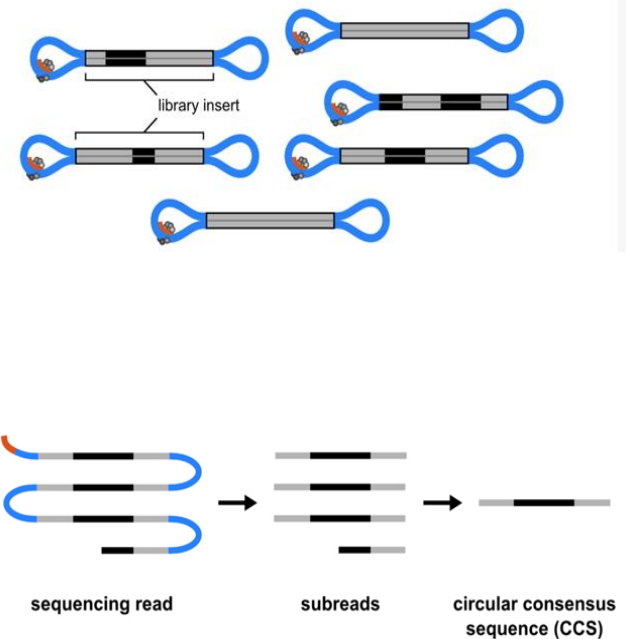


Single molecule real time sequencing (SMRT sequencing)

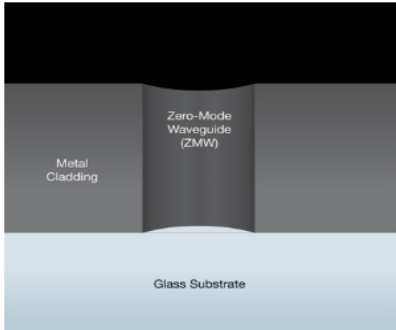
DNA sample



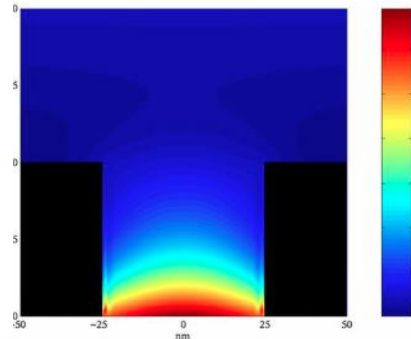
DNA synthesis after attachment of hairpin adapters.



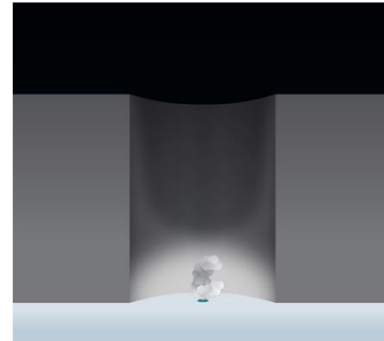
Single molecule real time sequencing (SMRT sequencing)



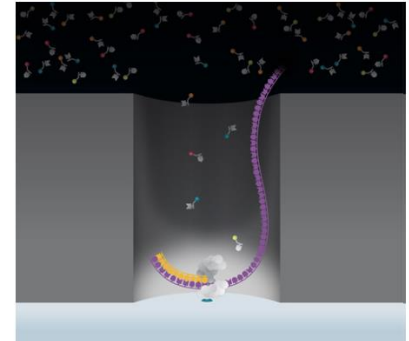
Individual ZMW



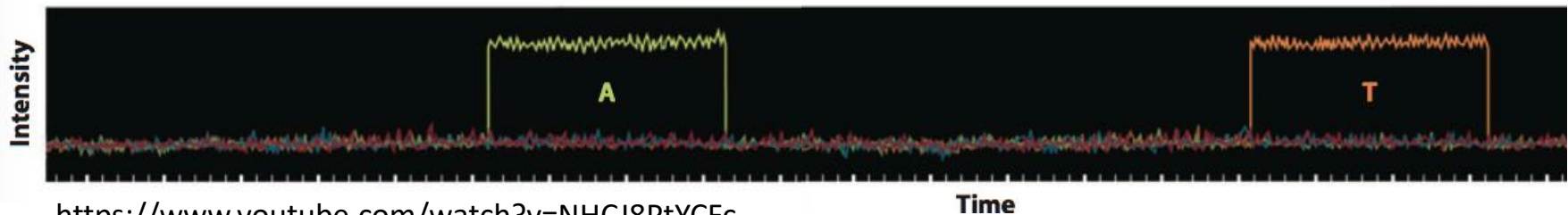
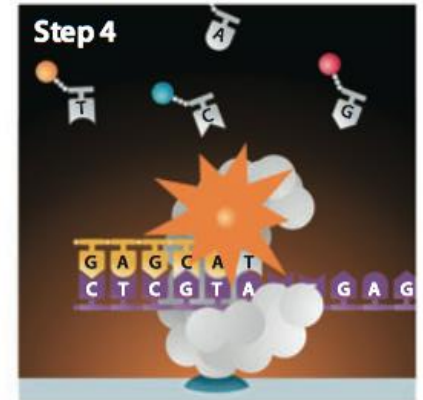
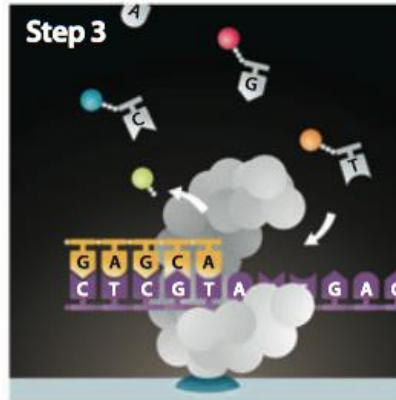
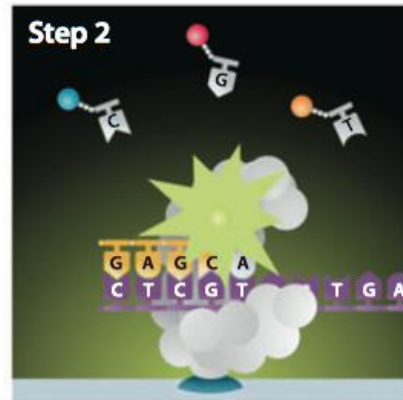
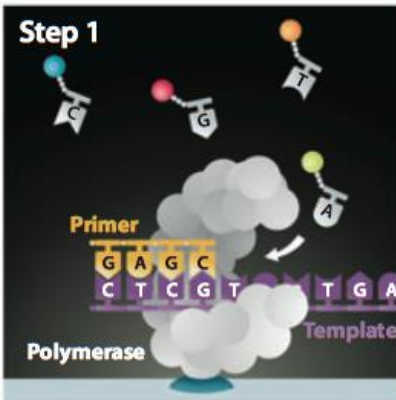
Laser light illuminates the ZMW



ZMW with DNA polymerase



ZMW with DNA polymerase and phospholinked nucleotides



<https://www.youtube.com/watch?v=NHCJ8PtYCFc>

Oxford Nanopore sequencing



MinION
512 nanopore channels

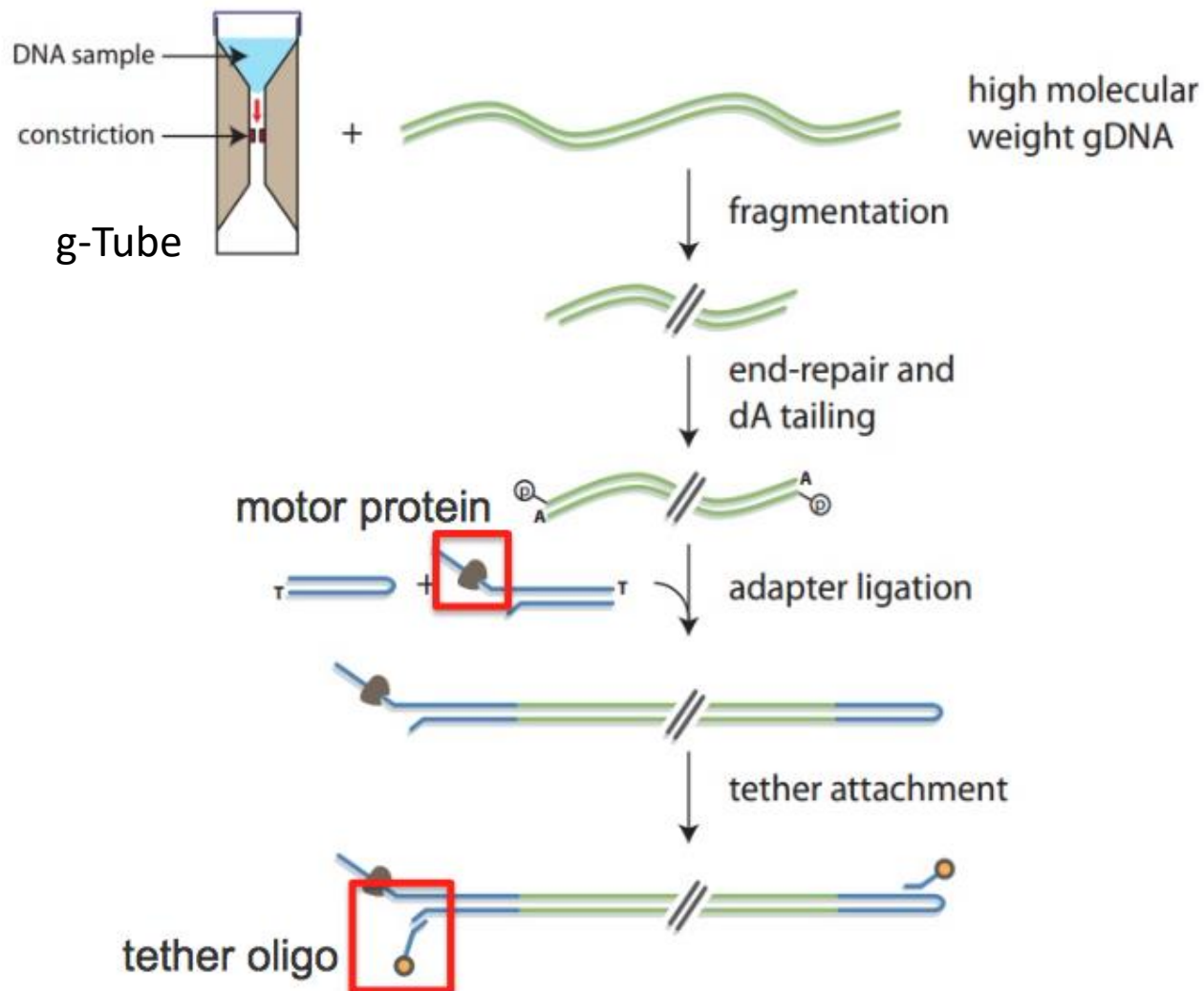


GridION = MinIONx5
2560 nanopore channels

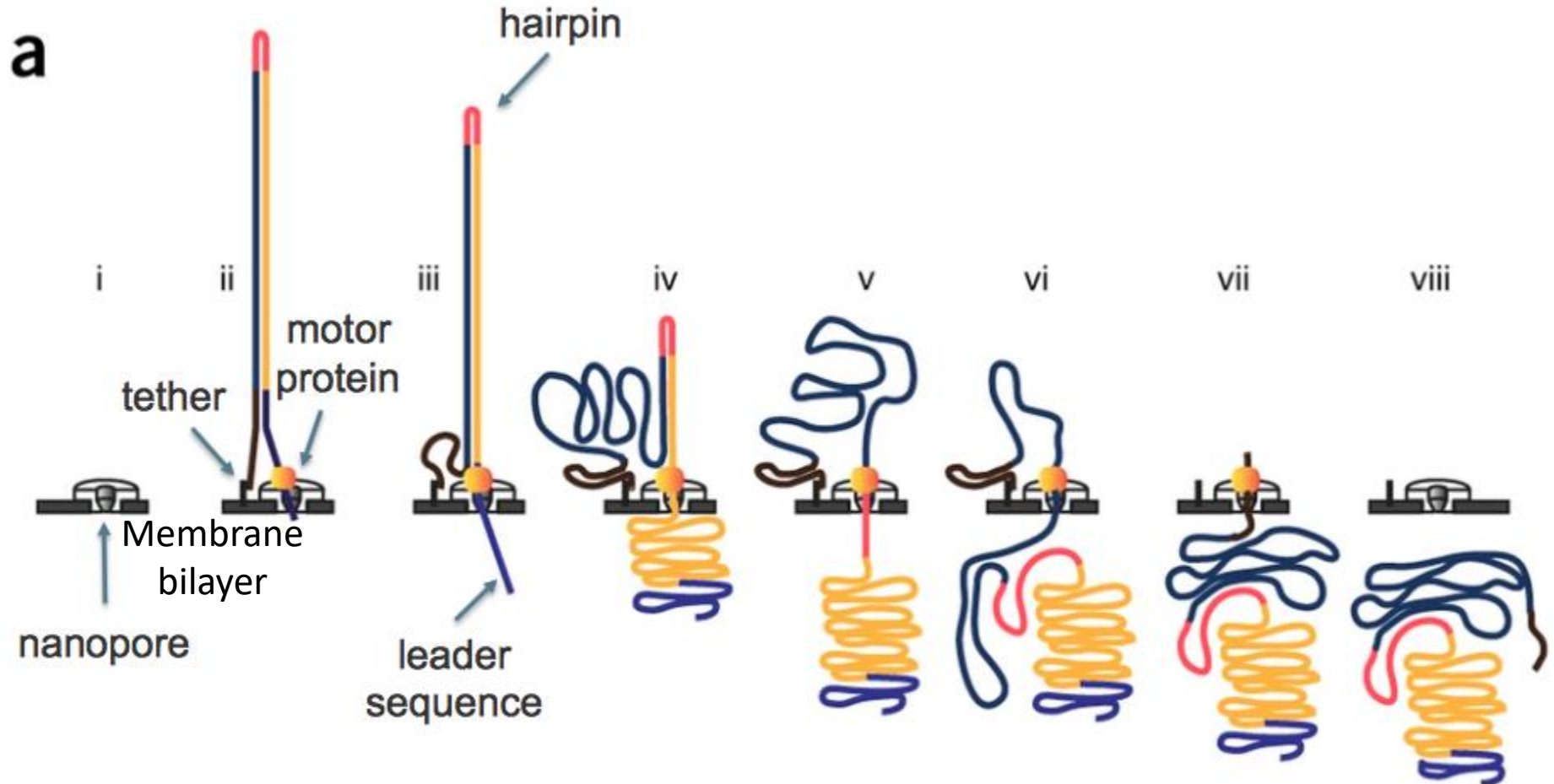


PromethION
48 flow cell, each with up to
3000 nanopore channels
Total 144,000 nanopore

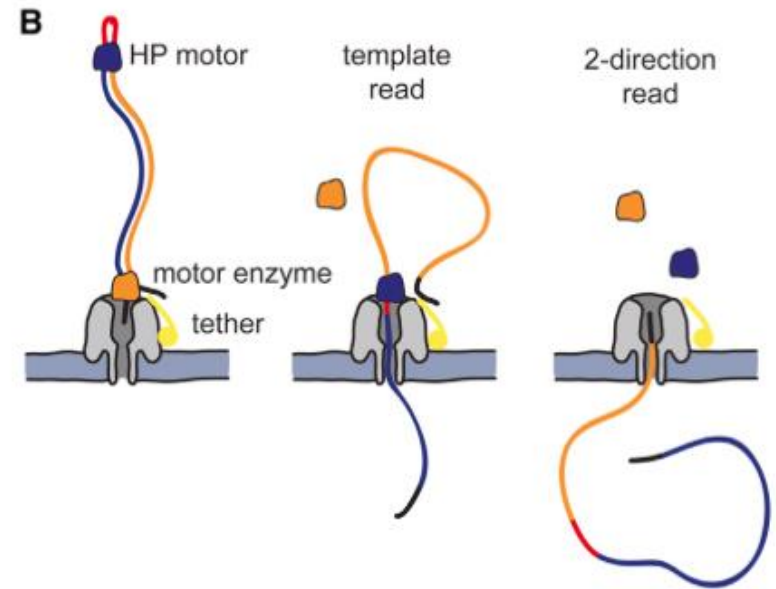
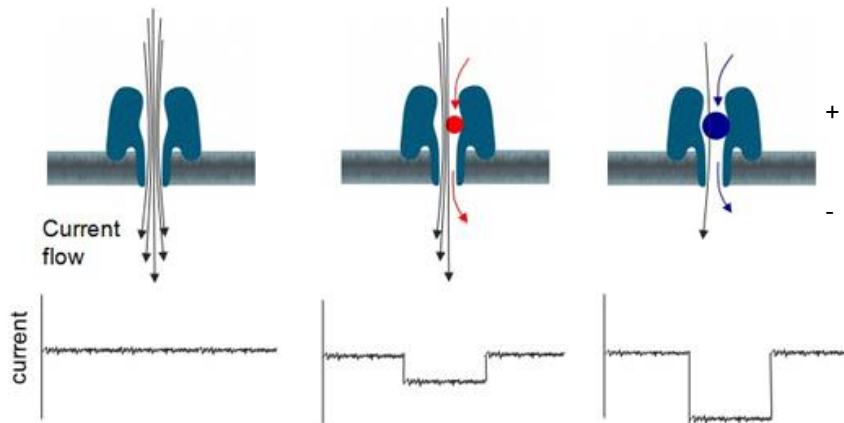
Oxford Nanopore sequencing



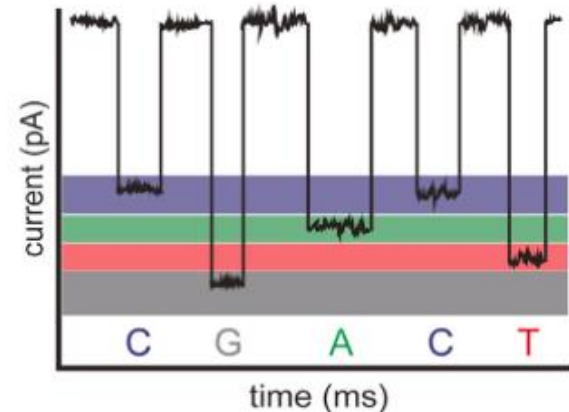
Oxford Nanopore sequencing










Oxford Nanopore sequencing



Albacore base callers
neural network



Select MinION starter pack

		Basic	Enhanced	Development
		Select	Select	Select
MinION		1	1	Up to 2*
Flow cells		2	4	16
Sequencing kits		1	2	4
Wash kits		1	1	1
Community Support		Included	Included	Included
Enhanced Support		Optional	8 weeks included	8 weeks included
Rapid Start Day		Optional	Optional	Optional
		\$1,000.00	\$4,999.00	\$15,677.00

Sequencing Platform	Advantages	Disadvantages
Sanger sequencing	<ul style="list-style-type: none"> - Lowest error rate - Long read length (up to 1000 bp) - Gold standard method 	<ul style="list-style-type: none"> - High cost per base - Long time to generate data - Need for cloning - Amount of data per run
454 pyrosequencing	<ul style="list-style-type: none"> - Low error rate - Medium read length (400-800 pb) 	<ul style="list-style-type: none"> - Relatively high cost per base - Must run at large scale - Medium/high start up costs
Ion Torrent	<ul style="list-style-type: none"> - Low start costs - Scalable (10-1000 Mb per run) - Medium/low cost per base - Low error rate - Fast runs (<3 hours) 	<ul style="list-style-type: none"> - Cost not as low as Illumina - Read lengths only (~100-200 pb)
Illumina	<ul style="list-style-type: none"> - Low error rate - Lowest cost per base - Tons of data 	<ul style="list-style-type: none"> - Must run at very large scale - Short read length (50-150 bp) - Run take multiple days - High startup costs - De Novo assembly difficult
PacBio	<p>Can use single molecule as template</p> <p>Potential for very long reads (several 10 kb+)</p>	<p>High error rate (~10-15%)</p> <p>Medium/high cost per base</p> <p>High startup costs</p>
Nanopore	<p>Can use single molecule as template</p> <p>Potential for very long reads (up to 200kb+) 4000 bp for mean</p>	<p>High error rate (~5-15%)</p> <p>Medium/high cost per base</p> <p>Low startup costs</p>

Error and bias

- Some of sequence errors may mimic true biological signals (mutation).
- How to solve the problem?

NGS limitation

- Big computing infrastructure
- Bioinformatic tools can be used for sequencing analysis and the whole organism is sequenced

Wide range of applications

de novo whole genome sequencing,

whole genome re-sequencing

RNA (RNA-seq)

ChIP-seq

Exome sequencing

Metagenomics

MicroRNA profiling

Methylation analysis

Which technology should we use?

NGS Technologies Platforms

Table: the platforms and the detailed information for the NGS technologies.

	Pyrosequencing	Ion Torrent		Illumina	PacBio		Oxford Nanopore
Instrument	GS-FLX Titanium	PGM 318	Proton II	HiSeq 3000	RS II	Sequel	MinION
Sequencing by synthesis	Pyro-sequencing	Semiconductor-based pH sequencing		Bridge amplification	Single molecule real time DNA sequencing		Nanopore exonuclease sequencing
Average read length	400 – 600 bp	Up to 400 bp	200 bp	2 x 150 bp	10-15 kb	10-15 kb	Variable (up to 900 kb)
Error rate	1%	1%	<1%	<1%	10-15%	10-15%	5-15%
Output (per run)	500-700 Mbp	1 Gbp	100Gb	650-750 Gbp	500 Mb-1Gb	5 Gb-10Gb	~ 5Gb
# of reads	1M	6M	80M	340M	~50k	500k	Variable (up to 1 M)
Run price	~\$6000	\$800	\$1000	\$1000	~\$400	~\$850	\$500-\$90