Summary & Review

: Human Genome Project

Sung Wook Chi
Division of Life Sciences, Korea University

What we learned in the previous class

quality score

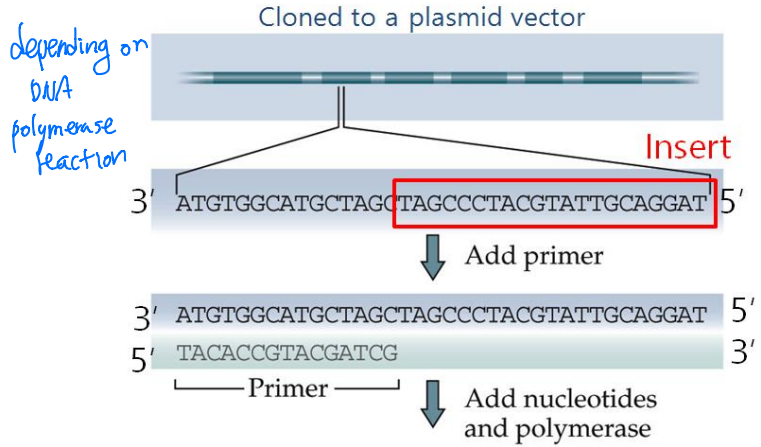
Sanger Sequencing / PHRED score

Human Genome Project

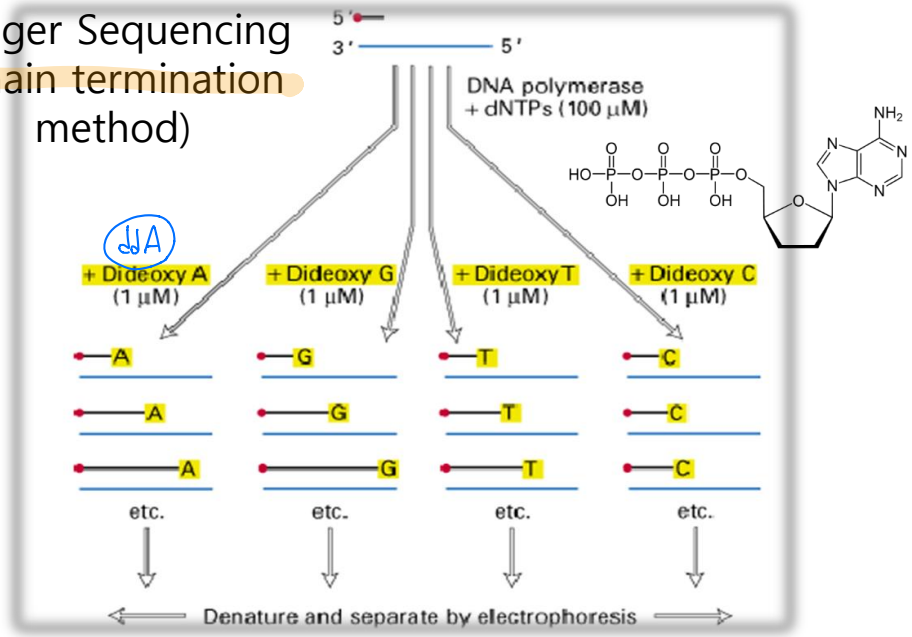
-Hierarchical vs. whole-genome shotgun sequencing

-Human genome

Sanger Sequencing : Chain termination method

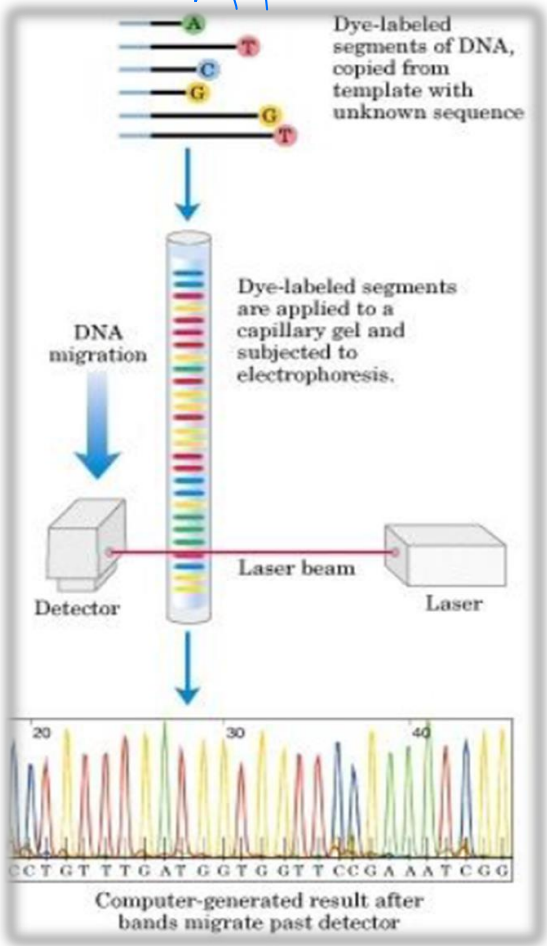


Sanger Sequencing (Chain termination method)



Dideoxy nucleotide

→ Stop point



Base calling

Phred Scores: quality of base calling

Quality score
(PHRED)
: base calling

$$q = -10 \log_{10}(p).$$

decrease error rate
→ increase q score

Error rate
in base calling process
(estimated based on shape/height of peaks)

BOX 3.7 **Phred scores: a measure of quality of sequence determination**

The phred score of a sequence determination is a measure of sequence quality. It specifies the probability that the base reported is correct.

If p = the probability that a base is in error, then the corresponding phred score $q = -10 \log_{10}(p)$.

Here is a short table:

Quality score q	Probability of error	Error rate
10	0.1	1 base in 10 wrong
20	0.01	1 base in 100 wrong
30	0.001	1 base in 1000 wrong
40	0.0001	1 base in 10 000 wrong

Q>20,
Error rate < 0.01

Q=20 (p=0.01)

Q<20,
Error rate > 0.01

Good

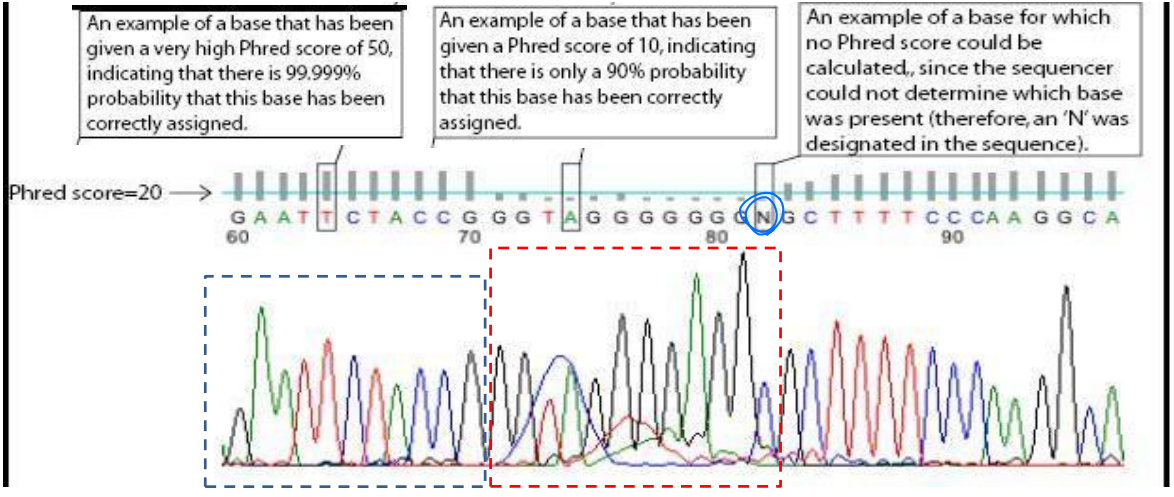
An example of a base that has been given a very high Phred score of 50, indicating that there is 99.999% probability that this base has been correctly assigned.

Bad

An example of a base that has been given a Phred score of 10, indicating that there is only a 90% probability that this base has been correctly assigned.

Bad

An example of a base for which no Phred score could be calculated,, since the sequencer could not determine which base was present (therefore, an 'N' was designated in the sequence).



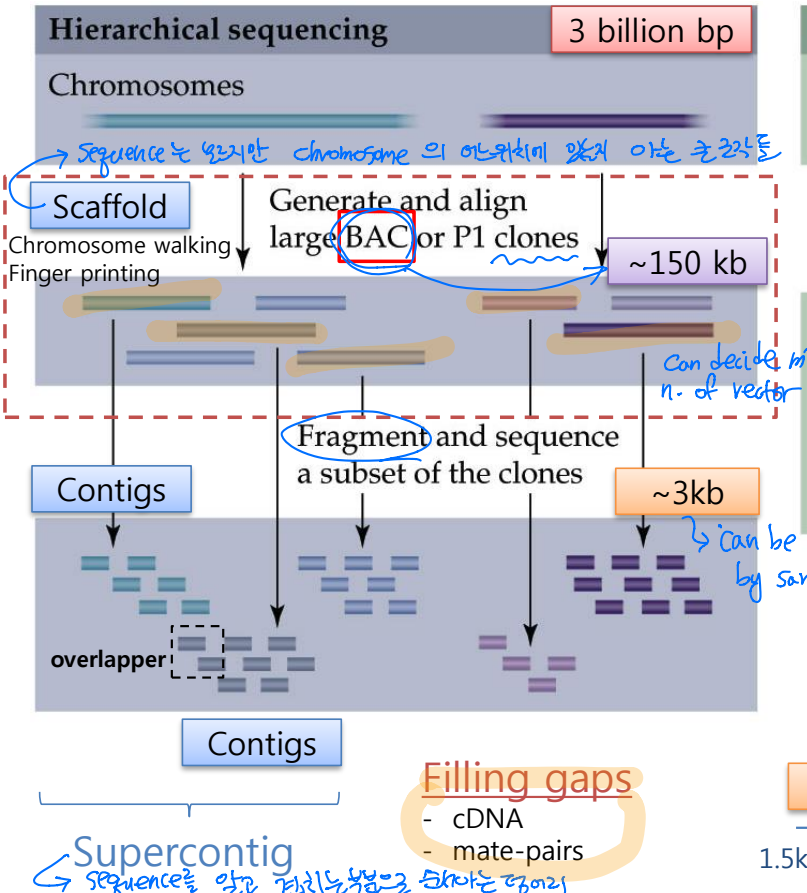
Hierarchical vs. shotgun sequencing

Human genome → Fragmentation → Sequencing → Assembly (De novo)

(read length 길이(한계)있음)

Hierarchical shotgun sequencing

'BAC-to-BAC' method (International consortium)

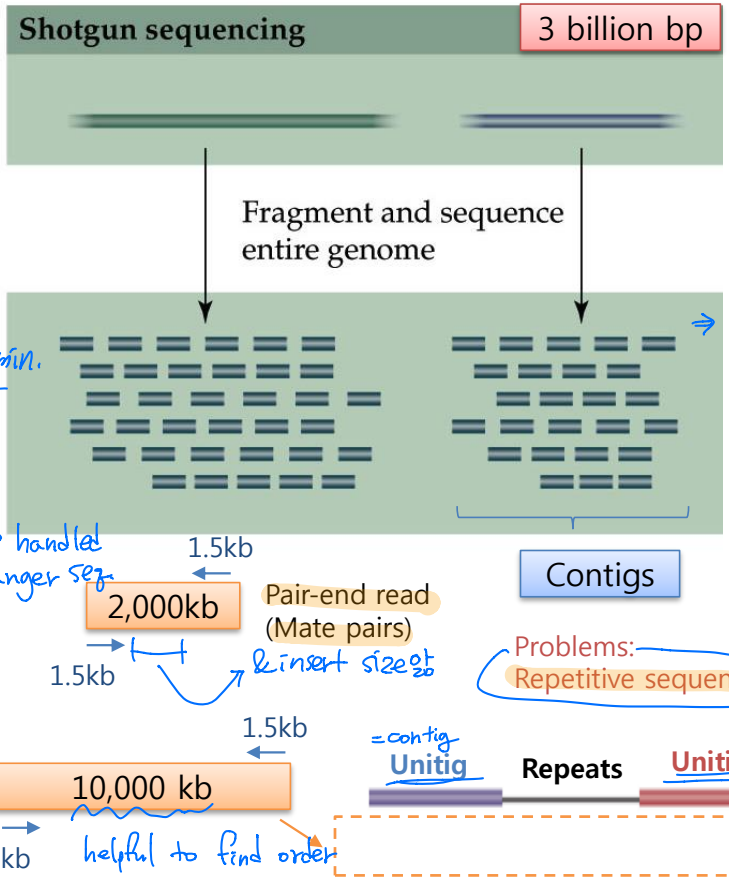


Filling gaps
- cDNA
- mate-pairs

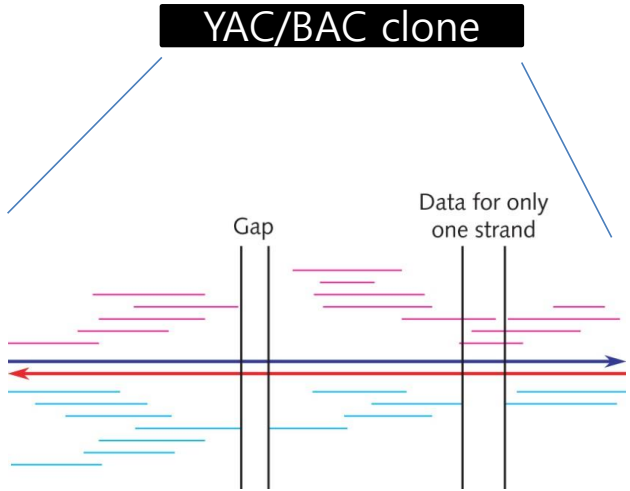
Whole-genome shotgun sequencing

nively fragment everything

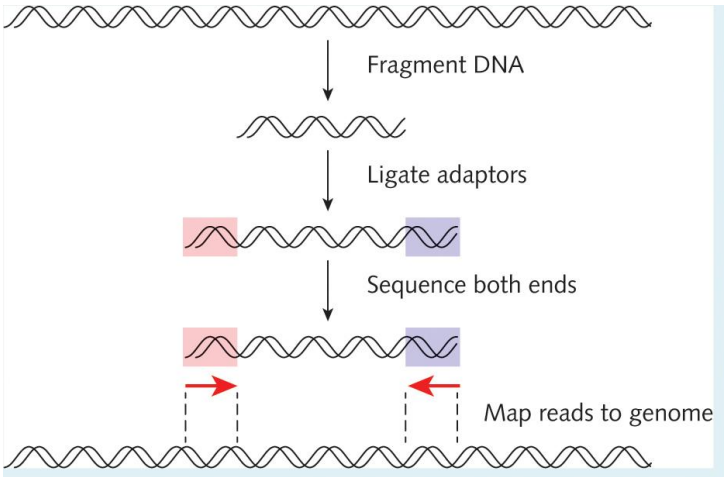
(Cerela)



Assembly of sequence reads: issue of gaps caused by repeat sequences

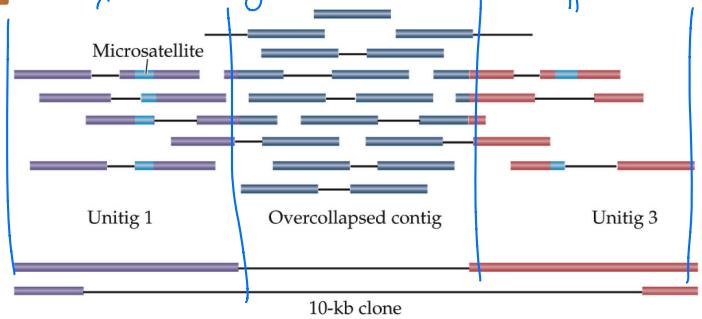


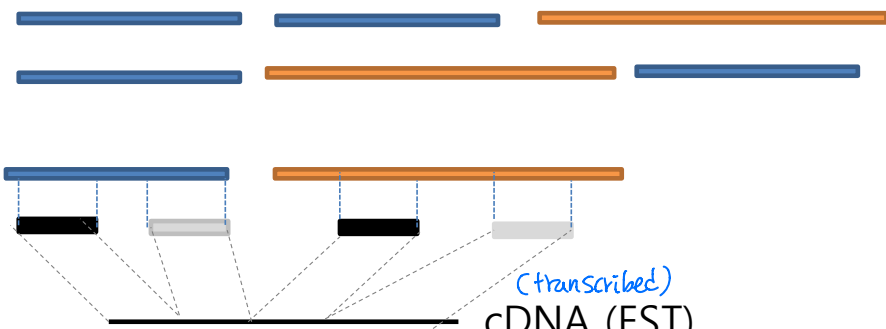
여기 gap은 잘 모르지만 chromosome walking을 통해 알아낼수 있음



Pair-end reads (mate pair)

large clone can cover these two kind of contigs here

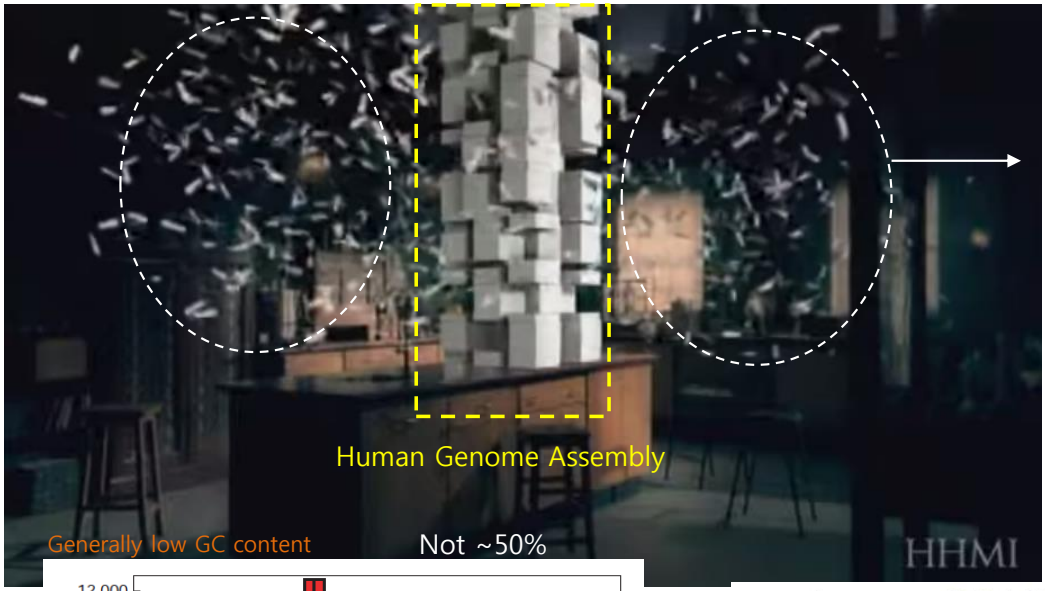




(transcribed) cDNA (EST)

⇒ blue가 left, yellow가 orange 라는걸 알수있음

Human genome

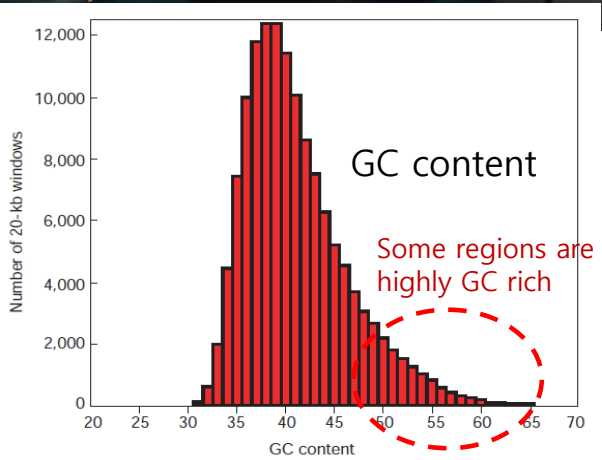


어느 크로모좀인지도 모르겠음
Unlocalized contigs

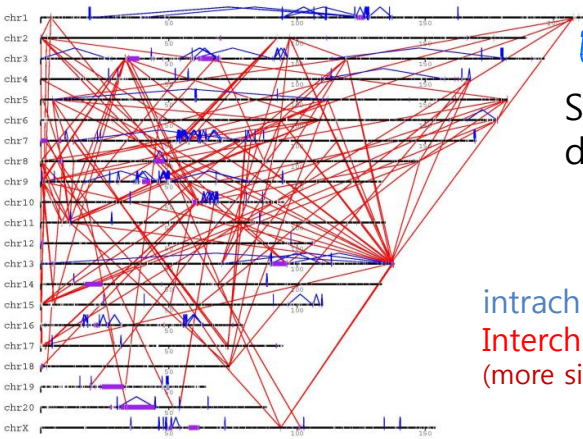
"chrUn"
Consited of repeated seq.

크로모좀은 알지만 위치는 모르겠음
Know the chromosome number, but unplaced

"Chr1_xxxx_random"



CpG island (promoter)



large chunk duplication

Segmental duplication

intrachromosomal
Interchromosomal
(more similar, larger)

Next-Generation Sequencing (NGS)

: High-throughput Sequencing

pallelized

Sung Wook Chi

Division of Life Sciences, Korea University

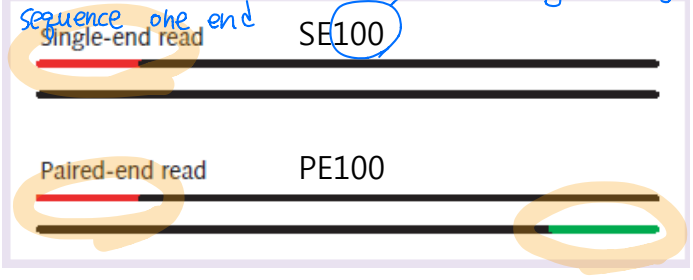
Terminology for DNA sequencing

Fragment: a small piece of genomic DNA – typically several hundred bp in length – subject to an individual partial sequence determination, or *read*.

Single-end read: technique in which sequence is reported from only one end of a fragment (see Figure 1.7).

Paired-end read: technique in which sequence is reported from both ends of a fragment (with a number of undetermined bases between the reads that is known only approximately).

Coverage (or depth) The average number of times a nucleotide is represented by a high-quality base in a collection of random raw sequence. Operationally, a 'high-quality base' is defined as one with an accuracy of at least 99% (corresponding to a PHRED score of at least 20).



Read length
~100bp

reference sequence 없는 새로운 종을 sequence 할 때

Paired-end sequence Raw sequence obtained from both ends of a cloned insert in any vector, such as a plasmid or bacterial artificial chromosome.

Read length: the number of bases reported from a single experiment on a single fragment.

Assembly: the inference of the complete sequence of a region from the data on individual fragments from the region, by piecing together overlaps.

Contig: a partial assembly of data from overlapping fragments into a contiguous region of sequence.

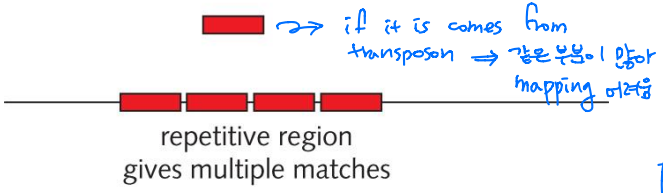
De novo *sequencing*: determination of a full-genome sequence without using a known reference sequence from an individual of the species to avoid the assembly step.

Resequencing: determination of the sequence of an individual of a species for which a reference genome sequence is known. The assembly process is replaced by mapping the fragments onto the reference genome.

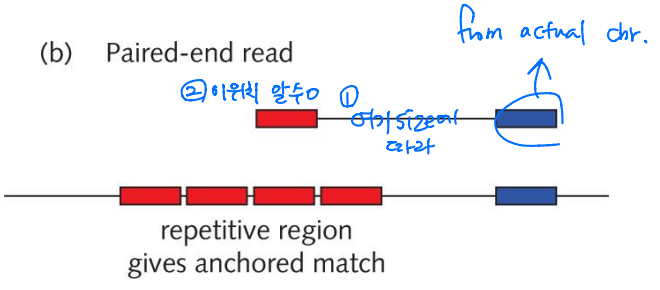
Single-end vs. pair-end read

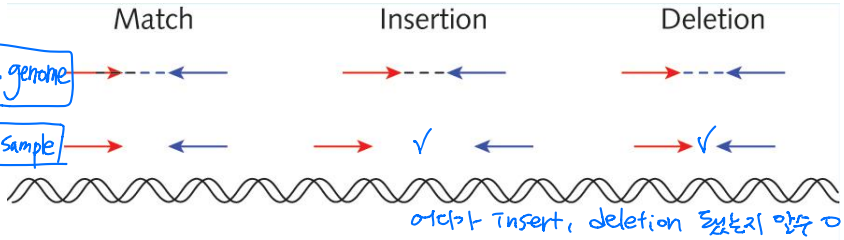
<PE 이용해 알수 있는 정보들>

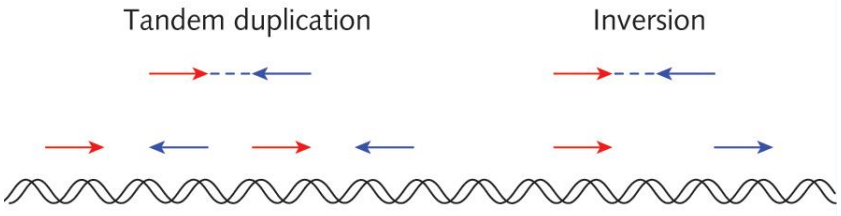
(a) Single-end read

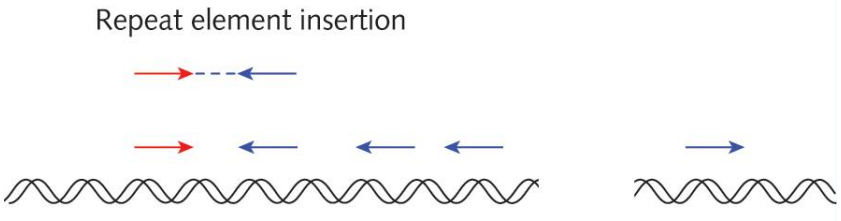


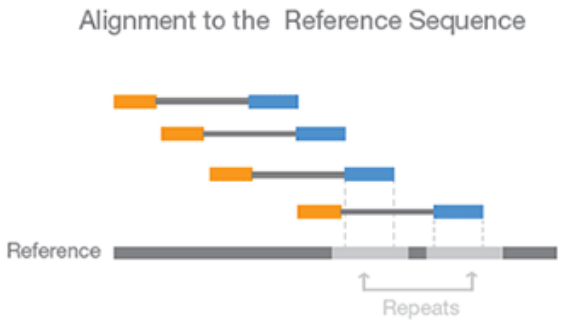
(b) Paired-end read











Next-generation Sequencing (NGS)

Sanger Sequencing

genome : 3GB
⇒ 10x depth need ⇒ 30GB

6 TB in 2 days

~100$!!!!

NovaSeq (Illumina) 2017

Solexa (Illumina) 2006

454 (Roche) 2005
first version of NGS

first version of NGS

Ion torrent (LT) 2011

SOLid (ABI) 2007

PacBio 2011

작동중 ~

Cost of Sequencing One Human Genome

Cost (US$)

10^8
10^7
10^6
10^5
10000
1000

required to seq one human genome

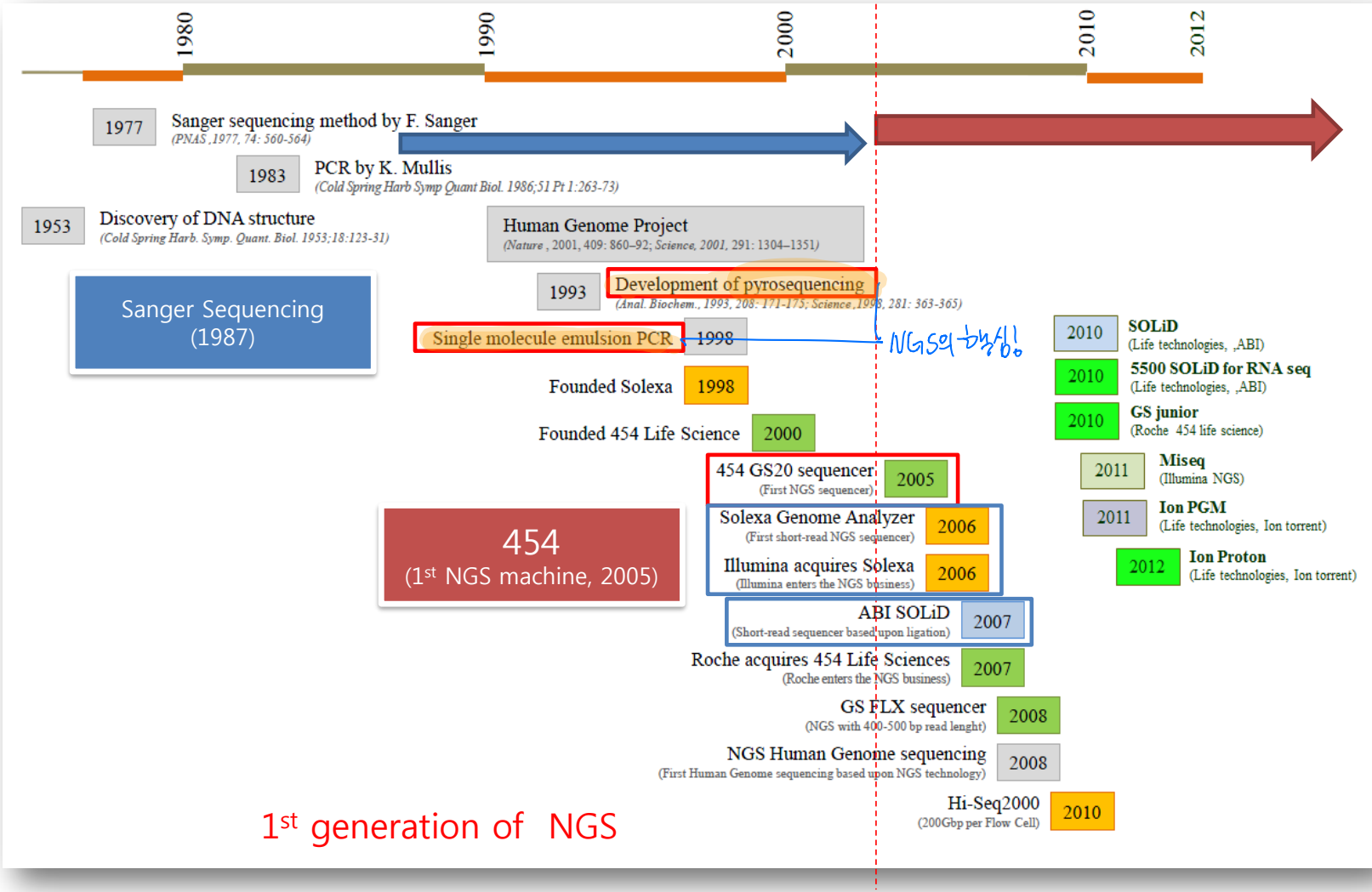
2001 2004 2007 2010 2013 2016

Year

Dramatic decrease !!!

Nanopore (Oxford) 2016

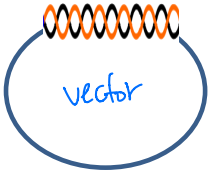
Sequencing Technology : History

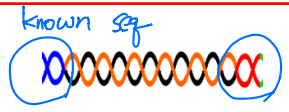


1st generation of NGS

DNA Sequencing

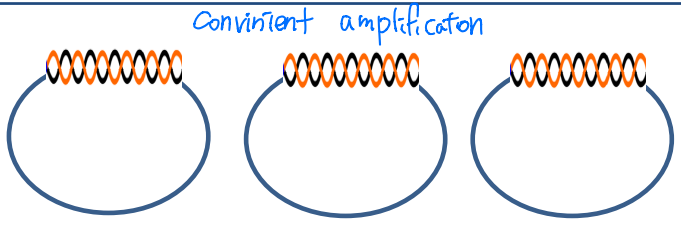
for all fragment
⇒각 end에 adapter 연결(NGS)
vector에 연결 (Sanger seq)

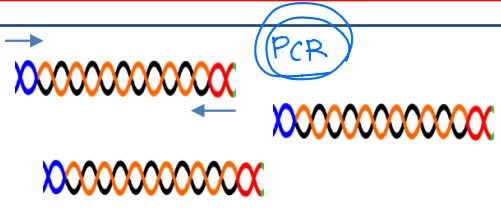




Library Preparation

- No preparation
  ↳ for some kind of method





Amplification

- Emulsion PCR
- Bridge PCR

for some : No amplification

Sequencing Chemistry

DNA polymerase reaction like sanger sequencing

Chain termination method
- -Pyrosequencing
- -SBS (Sequencing by synthesis)
  : reversible chain termination

- Ligase reaction
- translocation

Signal detection

Fluorescence color

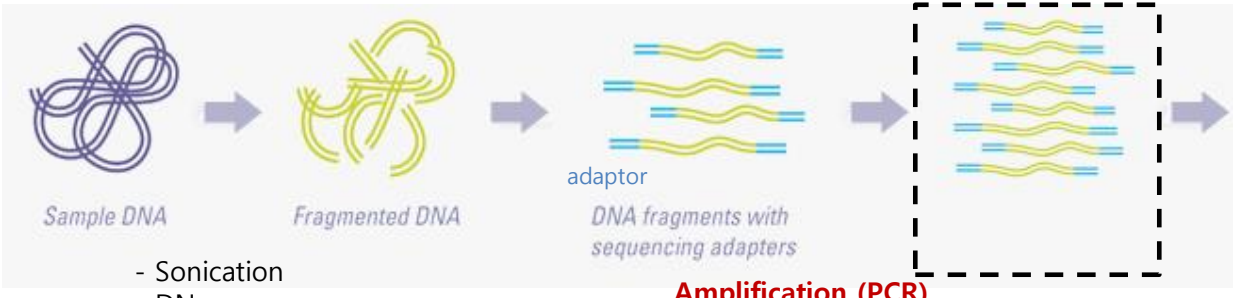
- Light
- Fluorescence color
- Electric signal (without light)
  → directly receive signal

determine A, T, G, C

DNA Sequencing by NGS (flow & principle)

Preparation

Genome → **Fragmentation** → **Adaptor ligation** → **Sequencing Library**



- Sonication
- DNase

Amplification (PCR)

- Depending on NGS platform

NGS sequencing platform
(Sequencer)

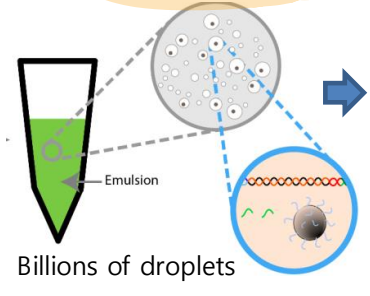
Convert polymerase reaction to light signal

Pyrosequencing

first version of NGS

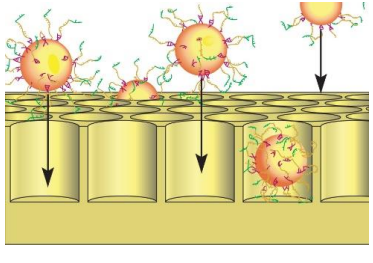
454 (Roche) 2005

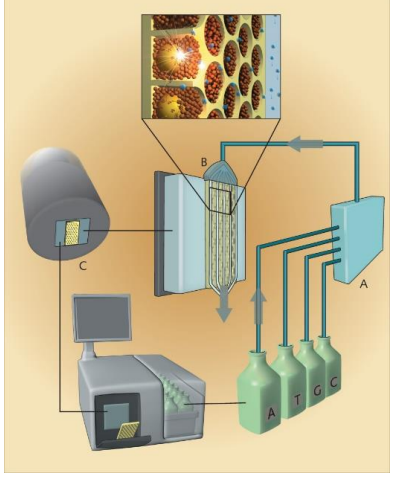
Emulsion PCR



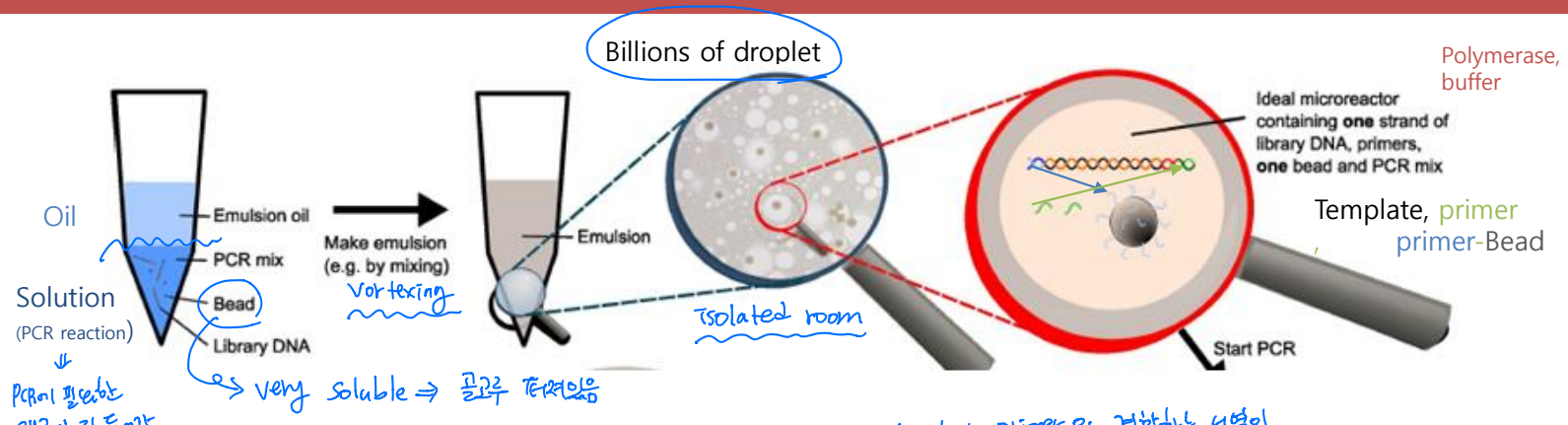
Billions of droplets

Bead deposit

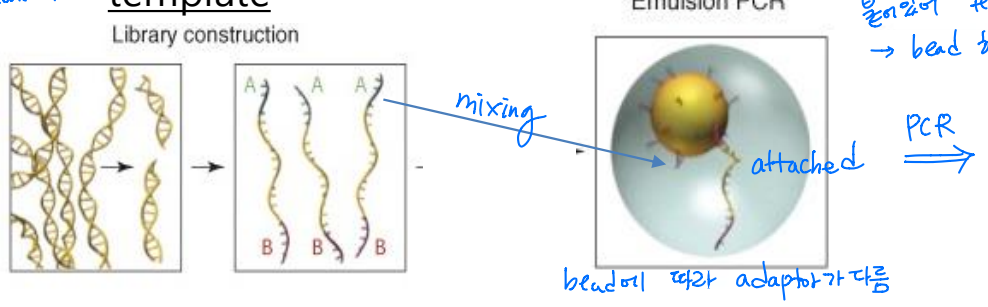


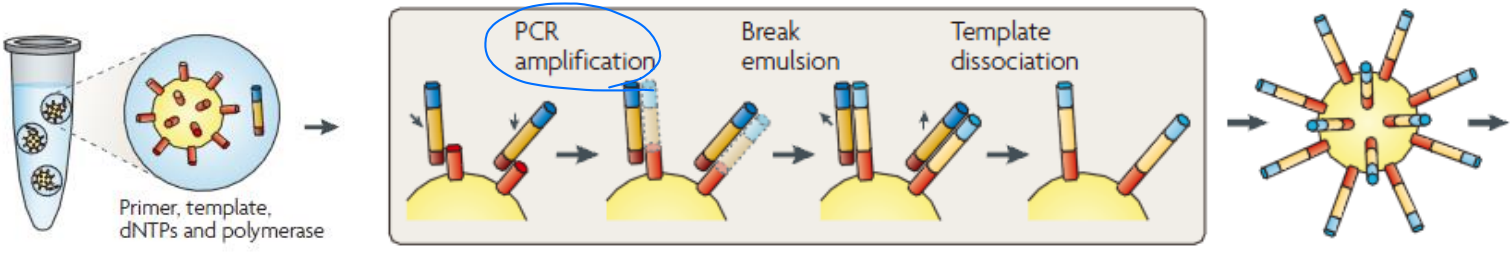


Emulsion PCR (454/Rocher, ABI SOLiD)



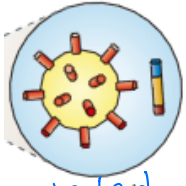
template



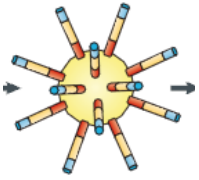


Emulsion PCR & bead loading : critical issues

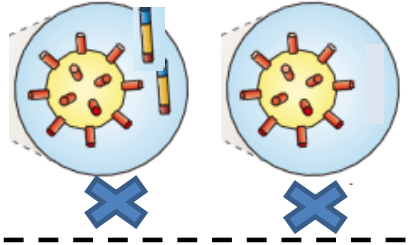
problems



가정 < one bead one templete

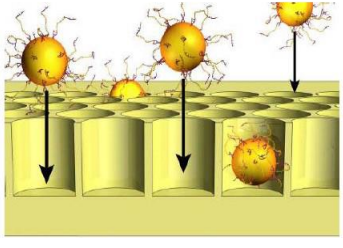


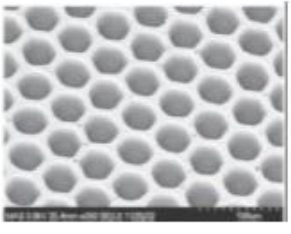
all bead are homogenic



two templete　　no templete

⇒ bead와 templete 비율 중요 !!



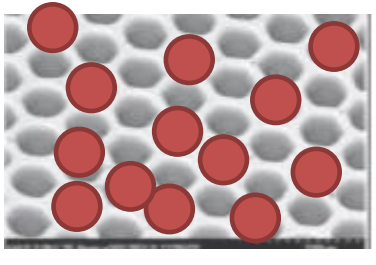


Beads arrayed into PicoTiterPlates

Roche 454

Not 100% loading

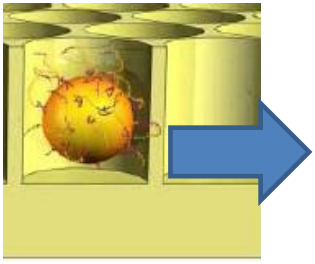
problems



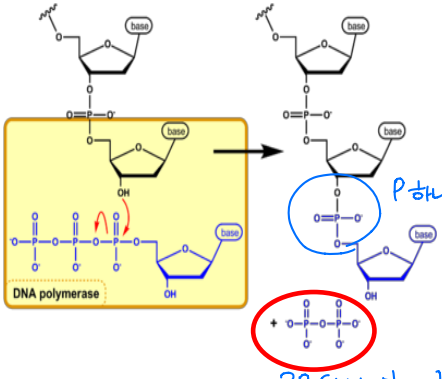
또 구멍에 bead 하나씩 넣기 어려움 (수렴로이 때와 well 회수는게 다른상황)

Pyrosequencing (454/Roche)

Same concept as Sanger seq.



Polymerase reaction

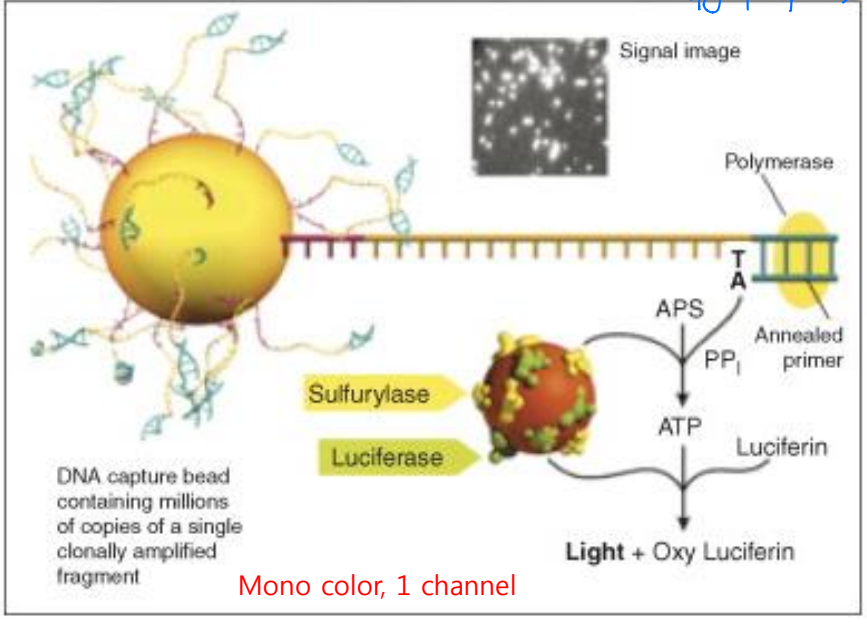


P 하나만 남음

$$AMP\text{-}SO_3H + PP_i \xrightarrow{\text{ATP sulphurylase}} ATP + SO_4^{2-}$$

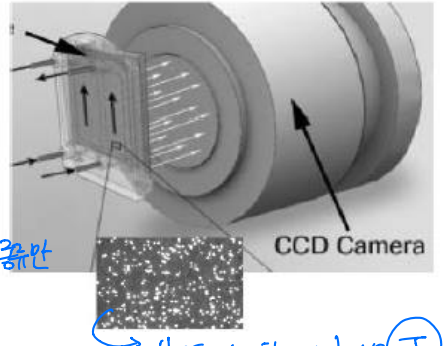
$$ATP \xrightarrow[\text{Luciferin}]{\text{Luciferase}} ADP + \text{light} \quad (\text{Oxyluciferin})$$

2P (pyrophosphate) should get rid of



Mono color, 1 channel

PPi -> light



① 한번에 dNTP 한종만

these positrons have T

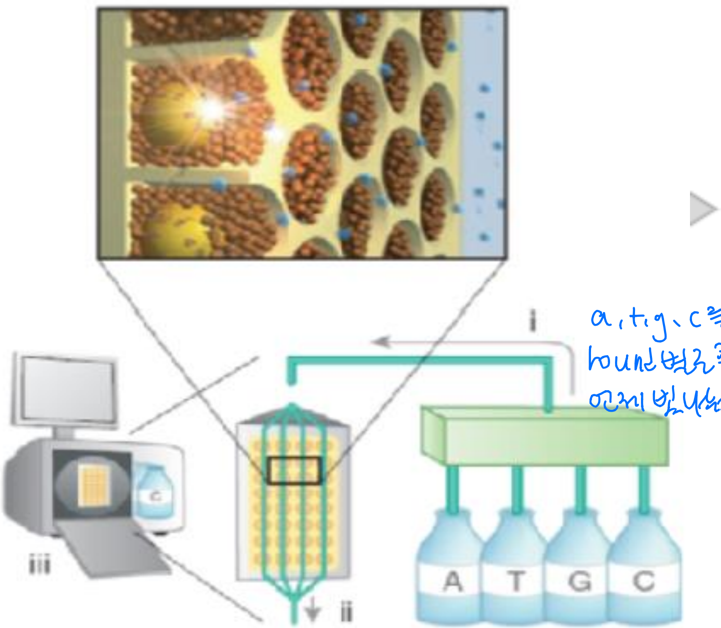
Pyrosequencing (454, Roche) : Problem for homopolymer

③ ② ①

dATP > dTTP > dGTP > dCTP

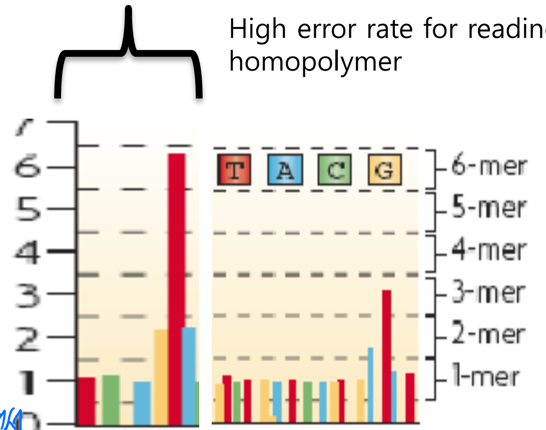
https://www.youtube.com/watch?v=rsJoG-AulNE

TCAGGTTTTTTAA



a, t, g, c를 bound 별로 주입해주면서 언제 빛나는지 detect

High error rate for reading homopolymer



→ intensity에 의존해 몇개가 repeat되는지 판단해야 하는데 not acurate

- Long read (up to 700bp),
  Low throughput (1 million)
  → empty well도 있을거고 하니까 100% 활용할 수 X
- High accuracy (99.997%)
- Low accuracy for homopolymer sequence.
  AAAAA

Sequencing	Amplif.	Chemistry	Read lenght (bp)	Run time (d)	Gbp/ day	DNA required (μg)
Roche 454 GS FLX Titanium	emPCR	Pyrosequencing	250-400	0.35 *	1.3	3-5

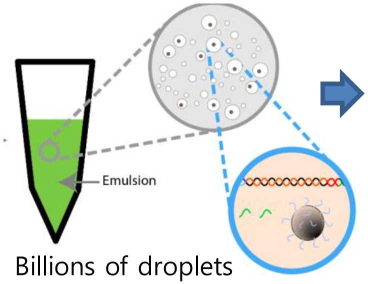
SOLid (ABI): ligation based sequencing method

SOLid (ABI) 2007

It is not available anymore (not important?)

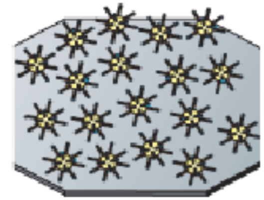
↪ 자세하게 이해할 필요는 X
DNA pol보다 정확한 방법을 만드려는시도

Emulsion PCR



Billions of droplets

Bead deposit



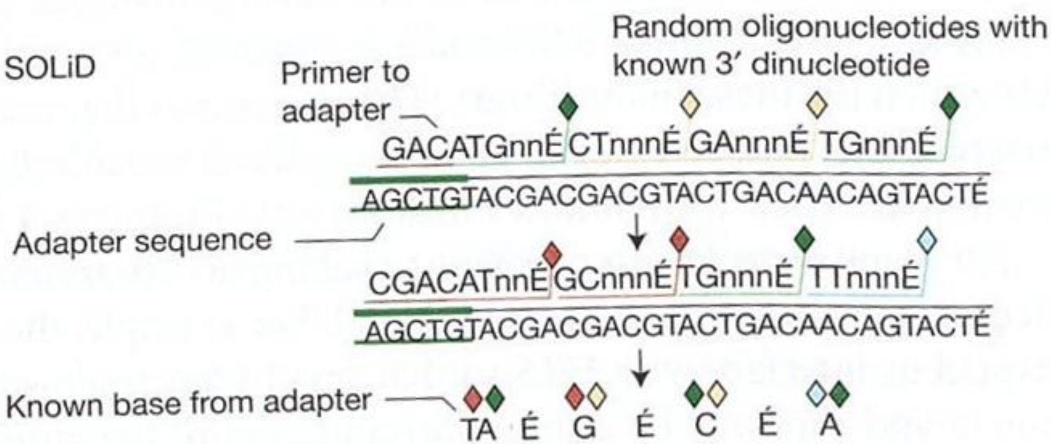
ABi SOLiD

Beads covalently linked to glass surfaces

micro well 사용하지않고
glass plate 이용
→ can attatch any position on surface

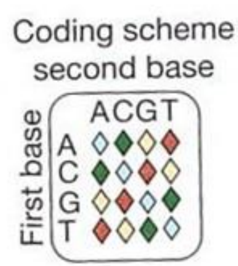
GnnC
AnnT
와 같이 여러 쌍의 oligo들을 만들고
hybridize하면
완벽하게 결합된 경우만
앞 서열과 ligation됨.
(pol 대신 ligase사용)
=> accuracy를 높일 수 있으나
시간이 매우 오래 걸리며
oligo를 만드는 다양한 가능성들을
모두 표지할 색의 다양성이 없음

Ligation based sequencing



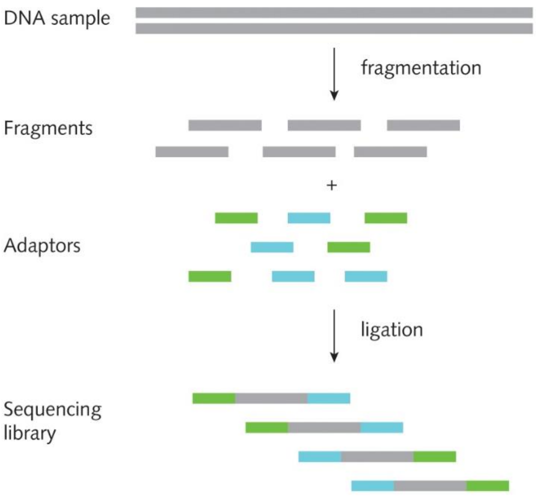
Sequencing by Ligation

- High accuracy
- Large computing time



Illumina NGS platform (Solexa)

↳ 가장 많이 이용되는 platform

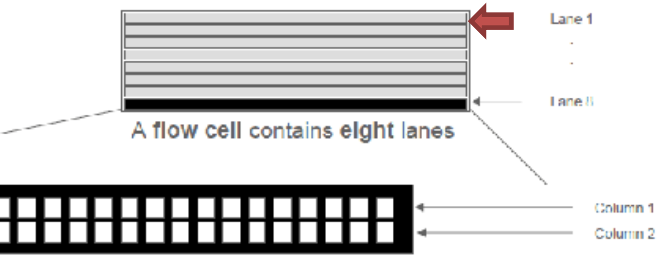


solid plate

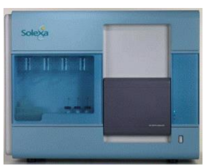
flow cell

flow cell에 넣기전에 dilute.

Loading your library into the lane (6 lane)

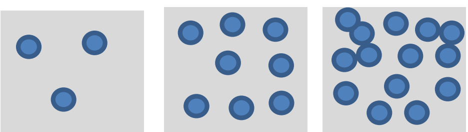


Same as polymerase reaction



⇒ well에 의해 나눠지는게 아니라서 loading concentration 중요!

☆ Library loading

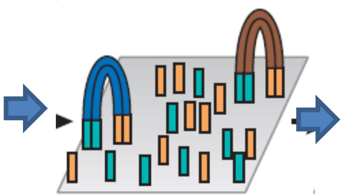


Sparse
(low number of reads)
돈 아깝....

Optimal

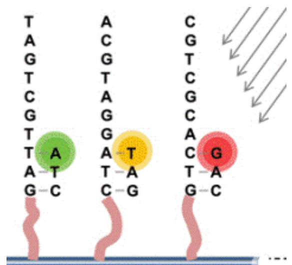
crowded
(Interfering of signal: low quality)

Bridge amplification



Solid phase amplification

SBS
(sequencing by synthesis)

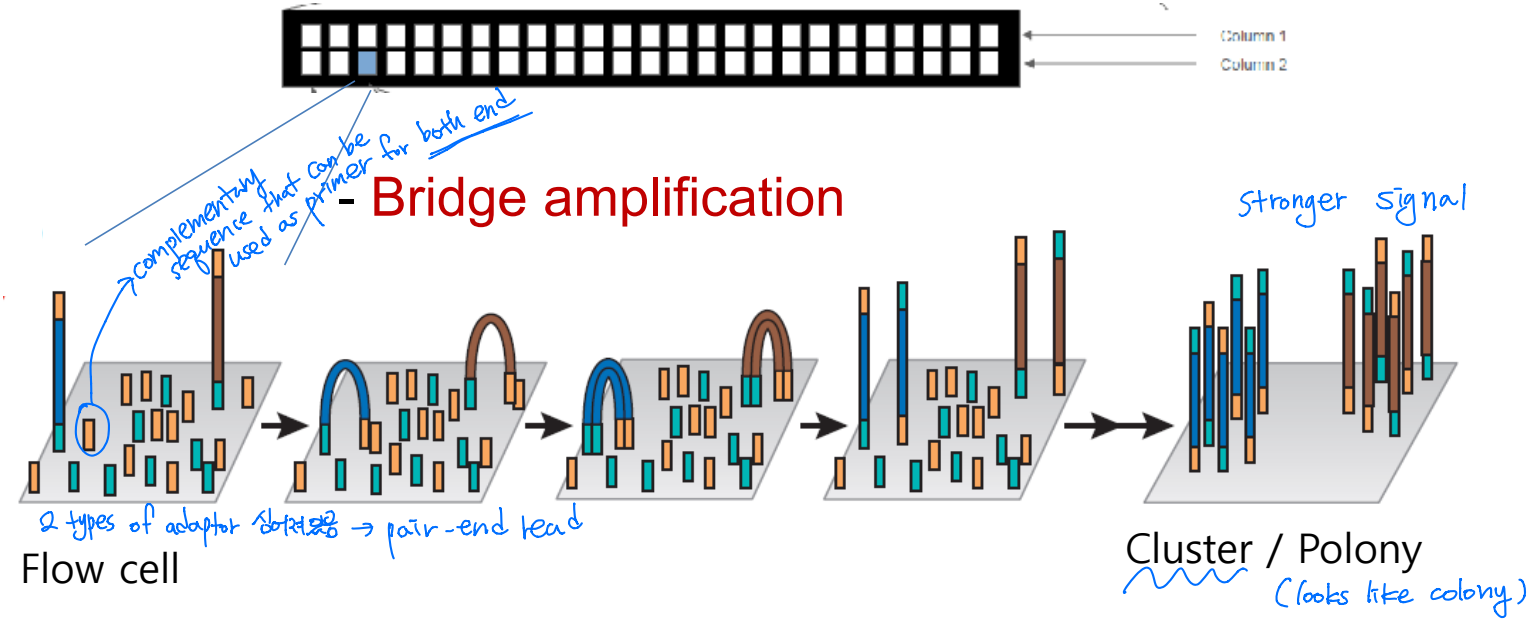


Reversible termination chemistry

Solexa (Illumina)
2006

Loading (spread) your library on glass plate (flow cell)

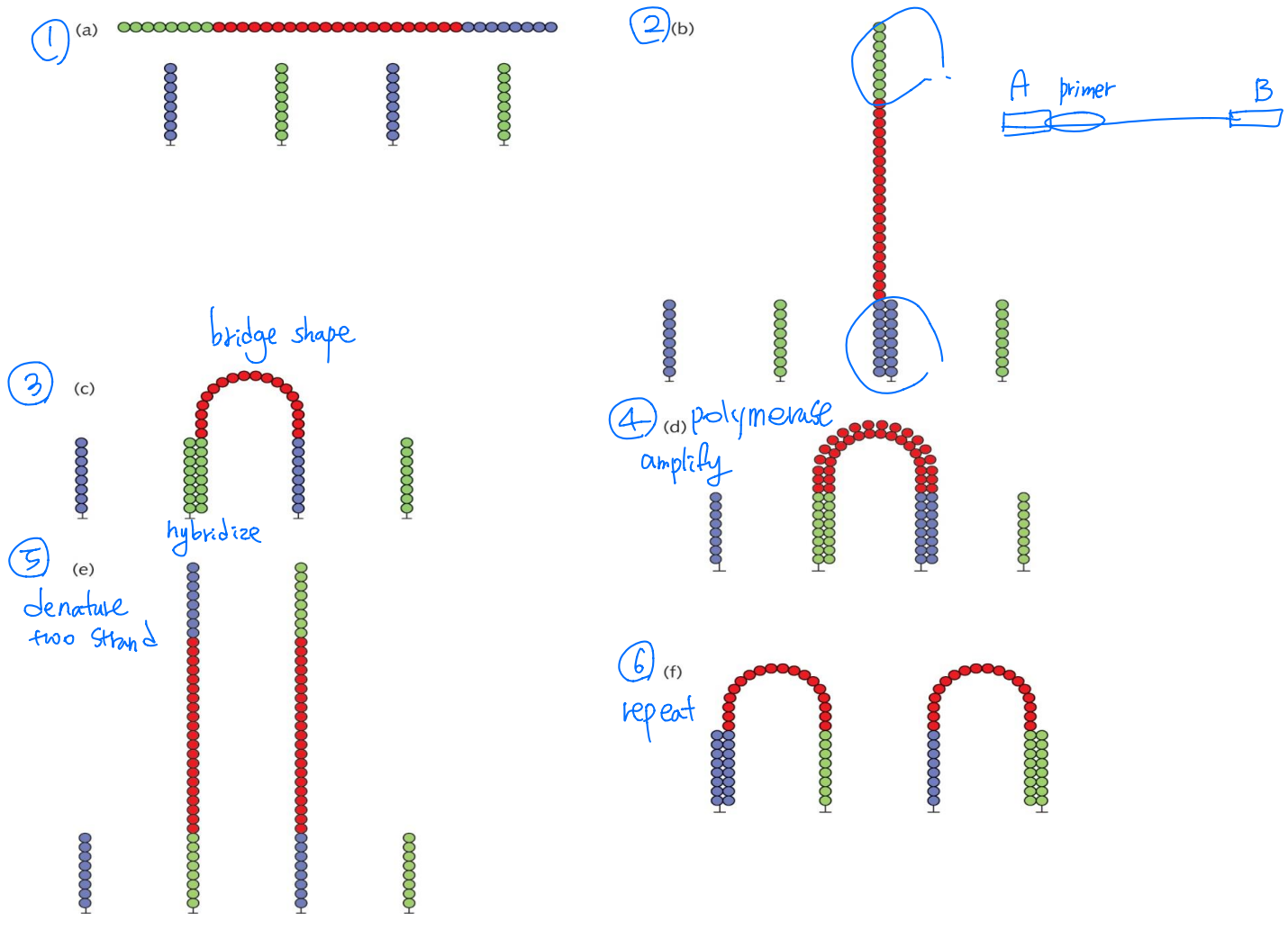
Bridge amplification : Library amplification in Illumina platform



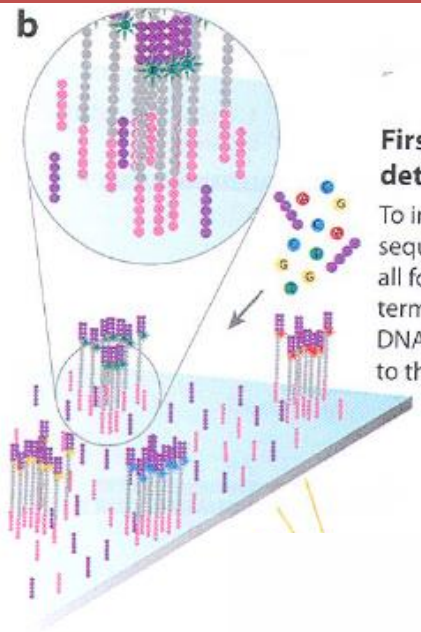
- DNA fragments are flanked with adaptors.
- A flat surface coated with two types of primers, corresponding to the adaptors.
- Amplification proceeds in cycles, with one end of each bridge tethered to the surface.

Cluster (polony) generation on solid-phase

Bridge amplification : Library amplification in Illumina platform



Sequencing chemistry in Illumina platform



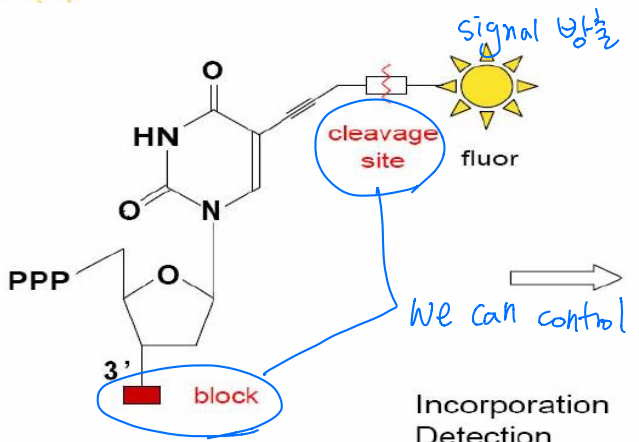
First chemistry cycle: determine first base

To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.

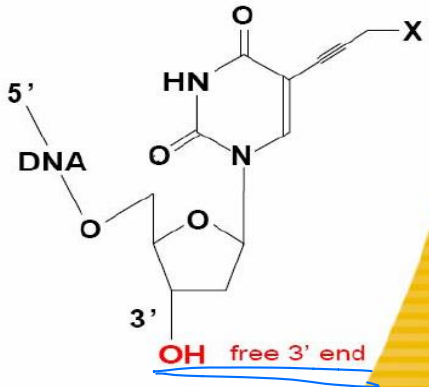
Reversible Terminator Chemistry Solexa

- All 4 labelled nucleotides in 1 reaction
- Higher accuracy
- No problems with homopolymer repeats

SBS (Sequencing by synthesis)

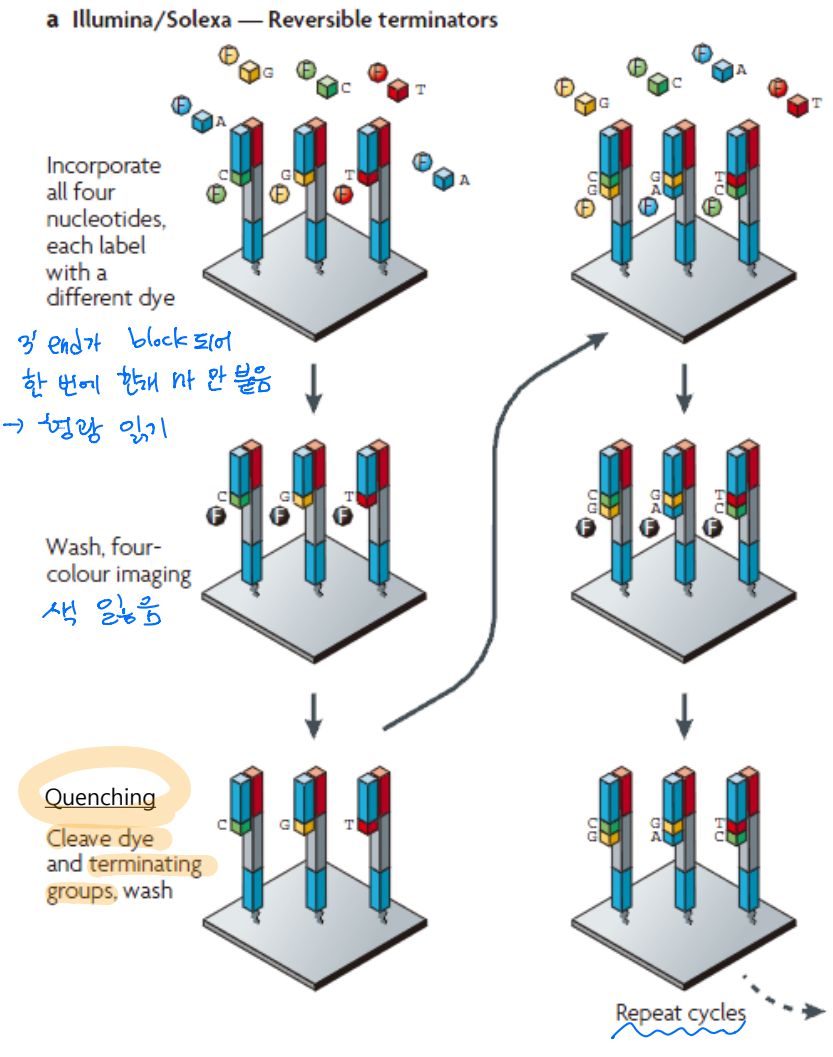


Incorporation
Detection
Deblock; fluor removal



Next cycle

Sequencing chemistry in Illumina platform

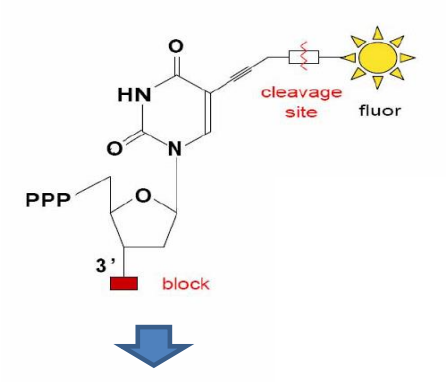


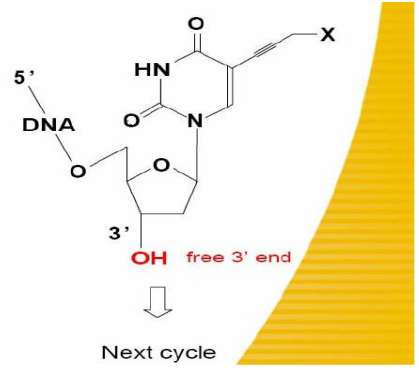
3' end가 block되어 한 번에 한개 씩 반응 → 형광 읽기

색 읽음

one cycle → detect one nt

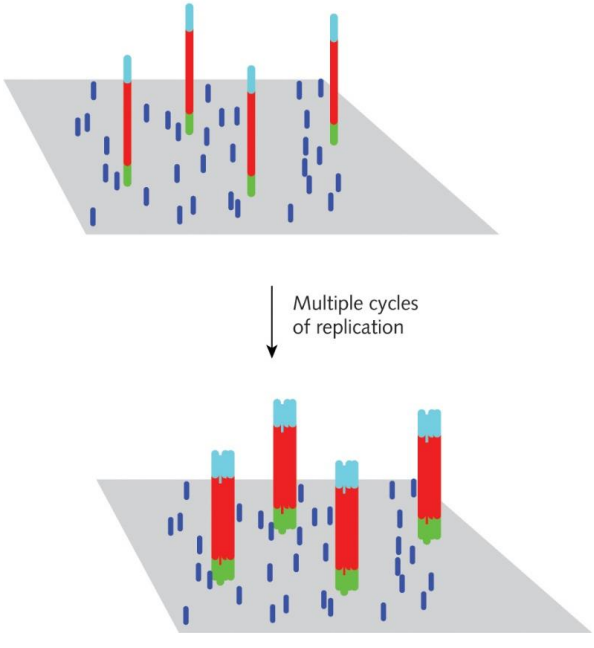
- SBS (Sequencing by synthesis)
  Reversible termination method



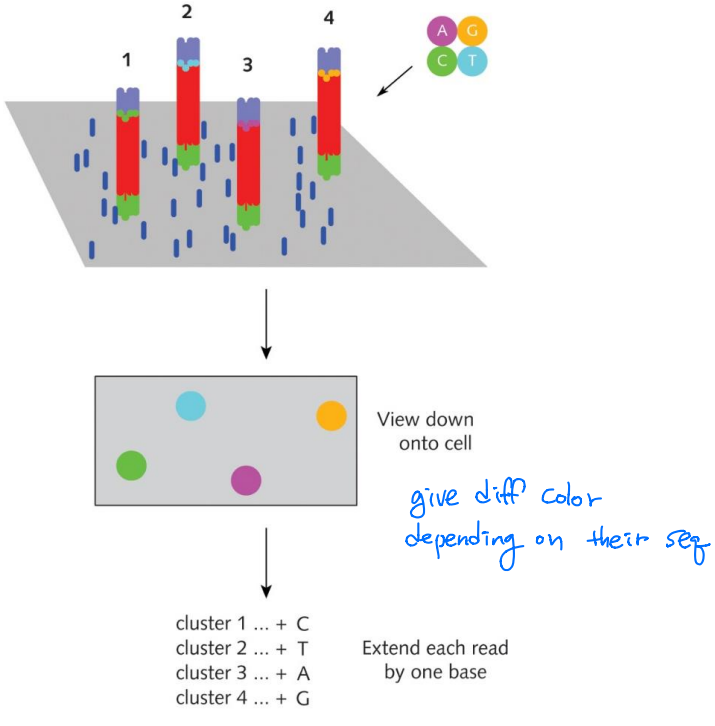


Solexa (Illumina) NGS method

Bridge amplification



SBS
(sequencing by synthesis)



give diff color
depending on their seq

Reversible termination chemistry

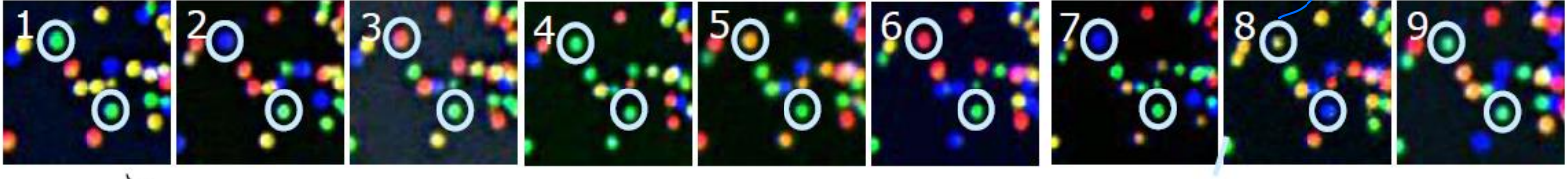
Base calling from image (Illumina)

looks like colors ar mixed

Quenching이 제대로 되지 않아서 interfere

TGCTACGAT...

Intrinsically inevitable for substitution error



homopolymer

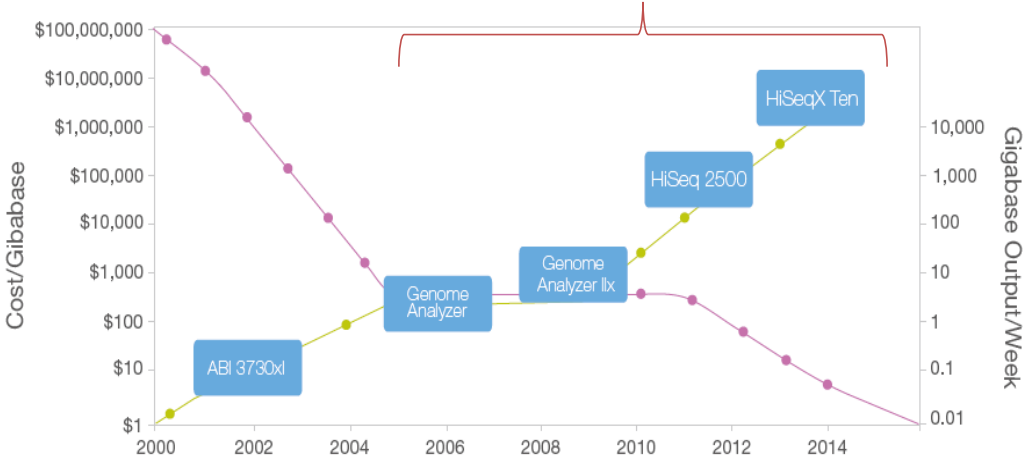
TTTTTTTGT...

Good at reading exact number of homopolymer

⇒ not able to read long length because of reading error

Illumina NGS platform (Solexa)

Winner of NGS market



2017, Jan

- 2 days
~100 $
!!!!

Illumina (NovaSeq)
6 TB in 2 days

https://www.youtube.com/watch?v=qtVM18Twi0I

HiSeq 2500 System (High-Output Mode)	HiSeq 2500 System (Rapid Run Mode)
29 hours–6 days	7–60 hours
1000 Gb	300 Gb
4 billion	600 million
2 × 125 bp	2 × 250 bp

limit of length

HiSeq platform: 300 million reads , 150 PE

1st generation NGS platforms (2005-)

Solexa / Illumina

Illumina GAII

Market winner

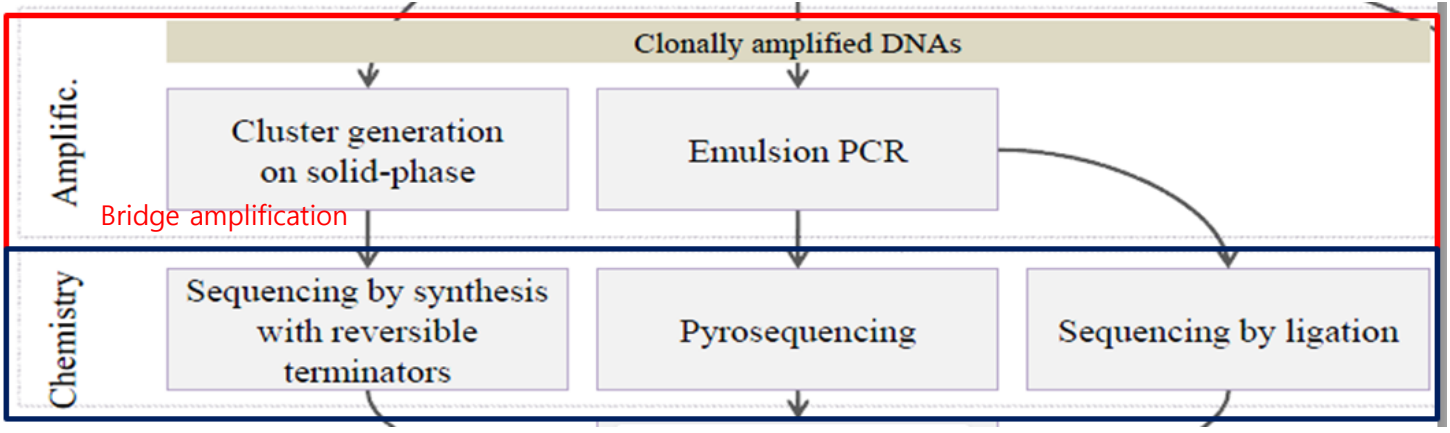
454 /Roche

Roche 454

First NGS sequencer

ABi SOLiD

Terminated



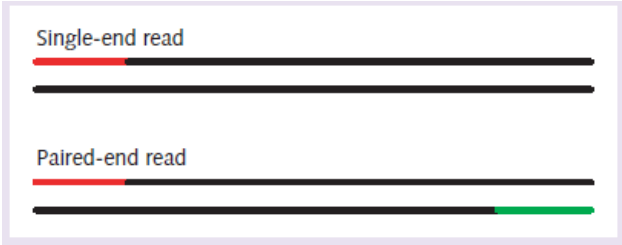
4 channel fluorescence

1 channel light

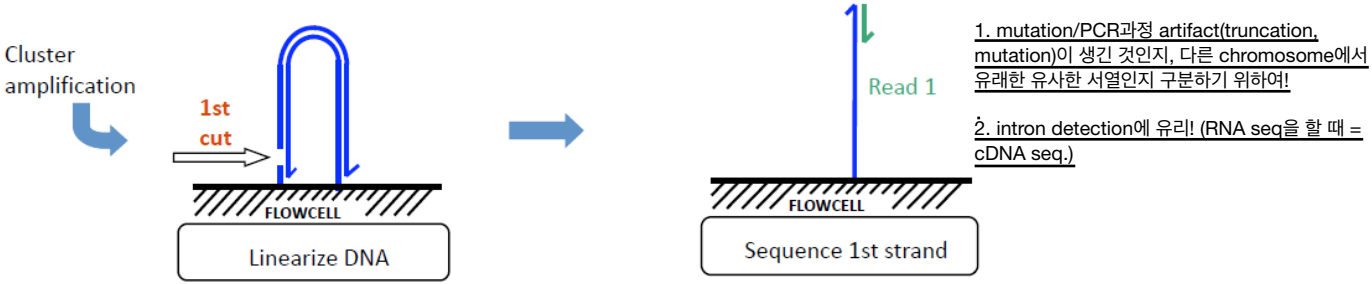
4 channel fluorescence

Paired-End Sequencing (Illumina)

Illumina/Solexa sequencing



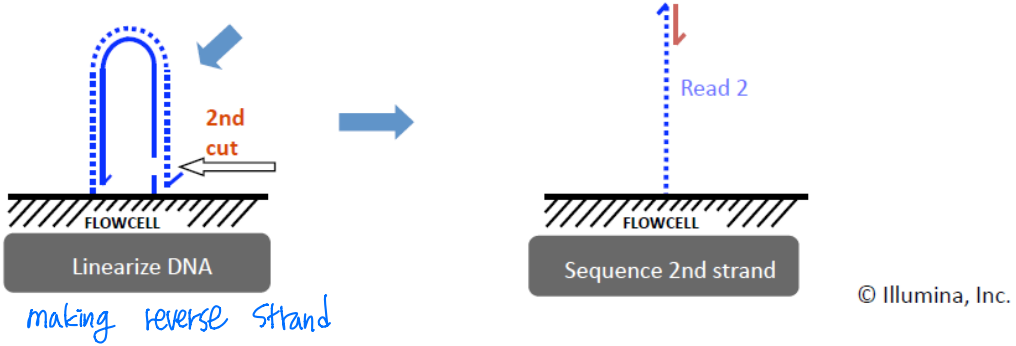
Single-end or Pair-end read



1. mutation/PCR과정 artifact(truncation, mutation)이 생긴 것인지, 다른 chromosome에서 유래한 유사한 서열인지 구분하기 위하여!

2. intron detection에 유리! (RNA seq을 할 때 = cDNA seq.)

reamplify



Multiplexing with Barcode: Sample Indexing

Illumina/Solexa sequencing

put multiple sample in one lane
→ 짧게 여러 개를 하나로 묶음

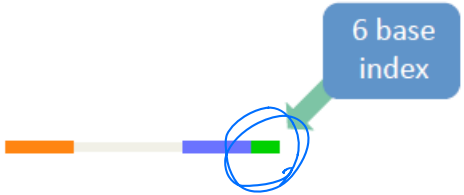
Sequencing multiple samples in a single channel, reducing cost/sample

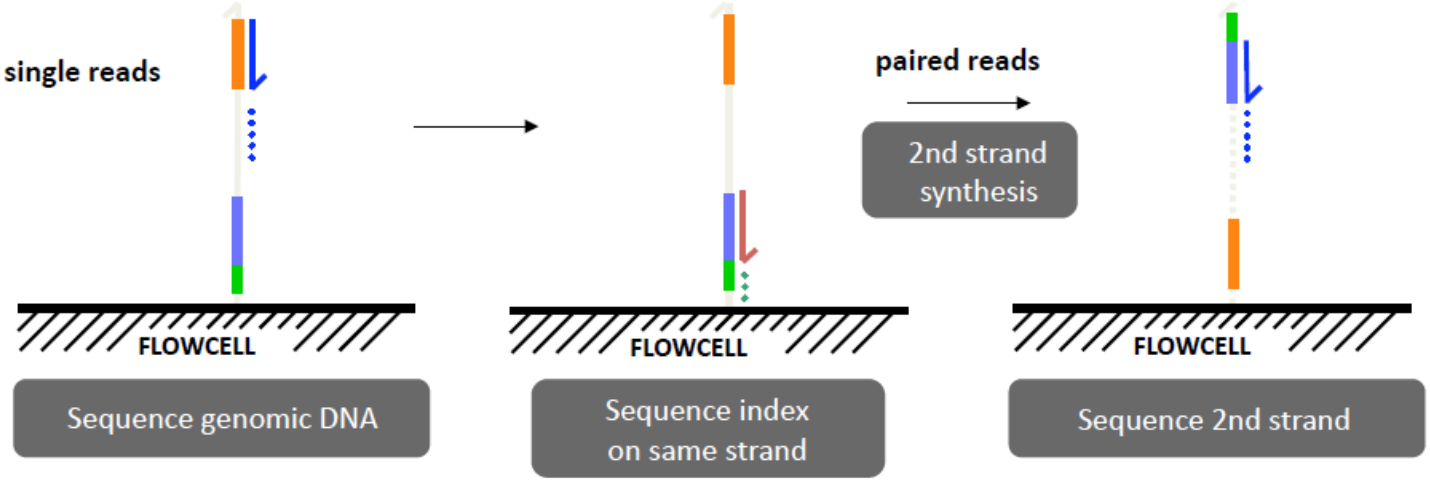
1 lane : 240 million reads

12 samples x 20 million reads

Simple construct design
Based on paired end process

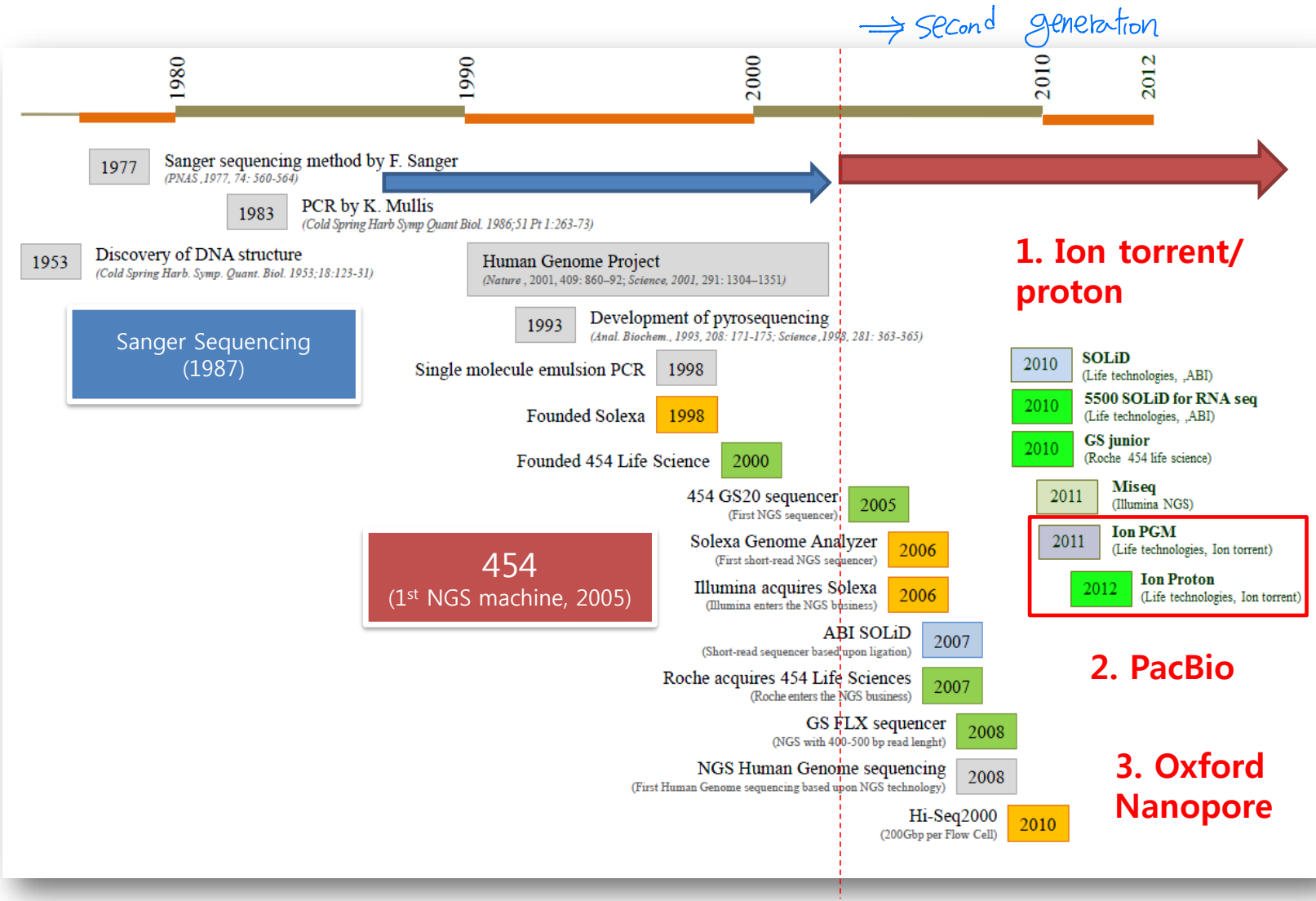
Sequencing protocol
Automated processing of indexing read





https://www.youtube.com/watch?v=HMyCqWhwB8E

2nd Generation of NGS sequencer

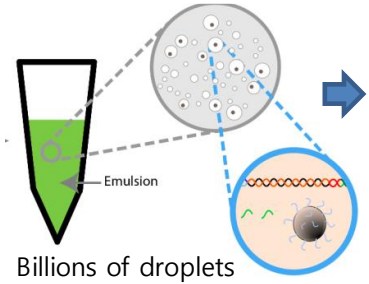


2nd Generation of NGS sequencer

Ion torrent (LT) 2011

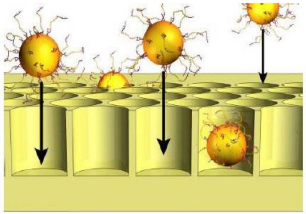
Same as 454

Emulsion PCR



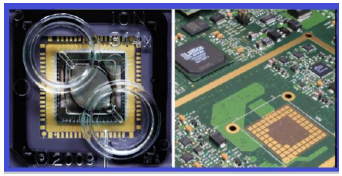
Billions of droplets

Bead deposit



Chemical rea

Semiconductor
: PH detection



Electric signal

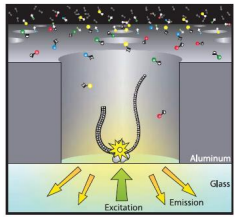
(can get rid of imaging process)
cut down price of machine

PacBio 2011

Single molecule

amp. 필요 X
real time

Real time sequencing



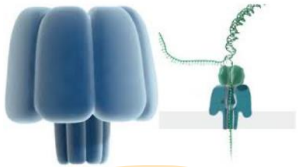
Light
(image process)

DNA polymerase → fast & accurate
→ tracing polymerase reaction in real time

Nanopore
(Oxford) 2016

Single molecule

Real time sequencing



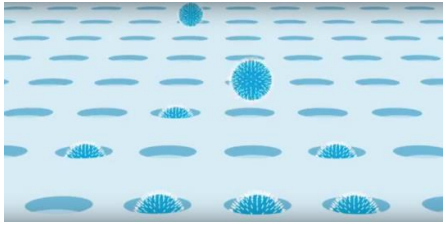
nanopore

Electric signal

DNA pol 이 아닌 nanopore 이용
electric signal 이용 & imaging process 없애기

Ion torrent/proton: semiconductor sequencing

Emulsion PCR

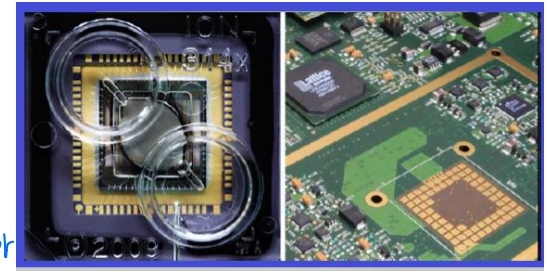


Electric signal
(No camera !!, low price)

1.5 million pH meters semiconductor chip

Ion torrent

use chips with pH detector

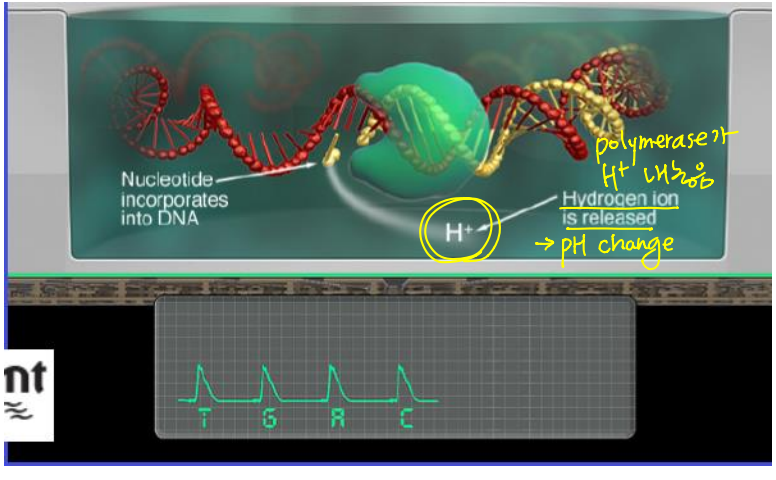


(advanced version)
Ion proton

Camera의 chip사용

80 million pH meters semiconductor chip

각 픽셀을 pH미터로 변환



DNA polymerase reaction

Release of proton (H+) > change in pH > detecting as electric signal

well을 100% 이용 X

homopolymer의 정확한 number X

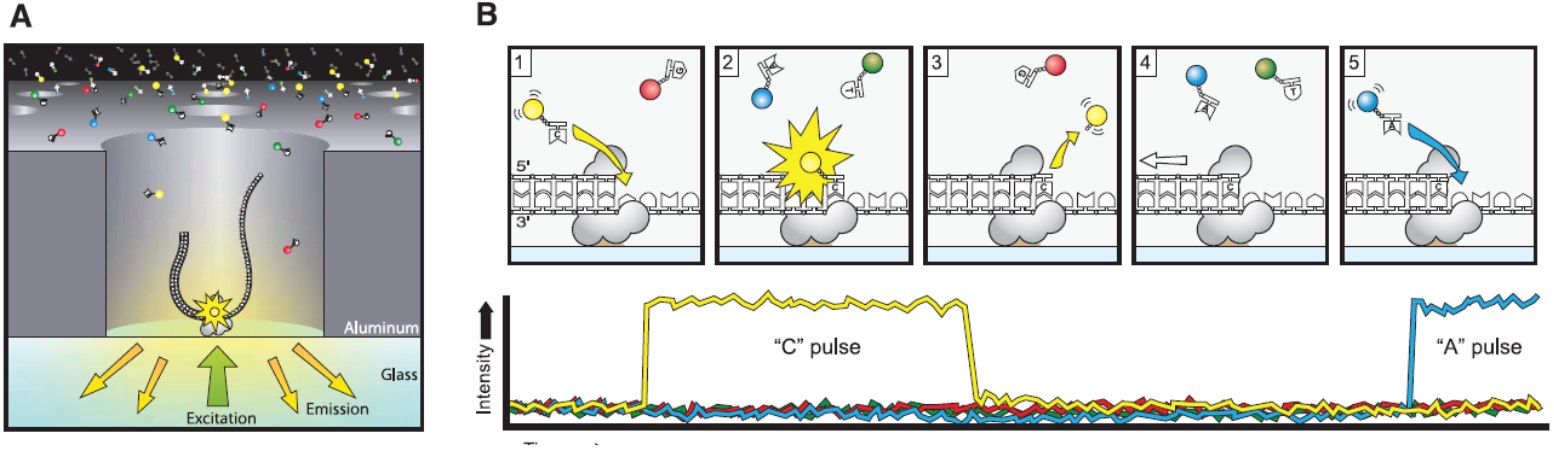
The same problems in
- Emulsion PCR, Pyrosequencing

https://www.youtube.com/watch?v=WYBzbxIfuKs

Pacific Biosciences

No amplification (Single molecule)
Long reads (3kb ~ 15kb) ,
Fast (real time, 30 min)
↳ but too fast to detect with our technology

tiny amount of signal & too fast ⇒ hard to detect
but **low accuracy** (95%)
low throughput (70,000 reads)

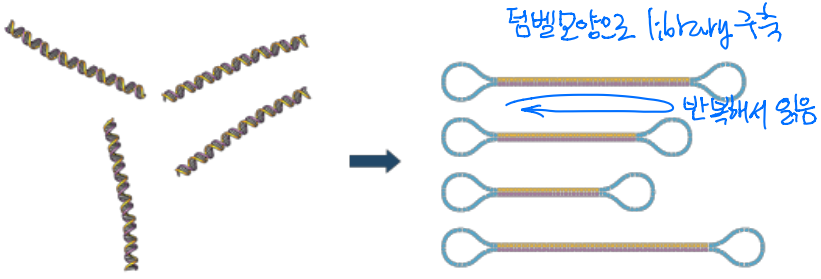


Single Molecule, Real Time Sequencing

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

Useful for long read sequencing (resequencing genomes)

SMRT (Single Molecule RealTime) Sequncing : PacBio



Library Preparation

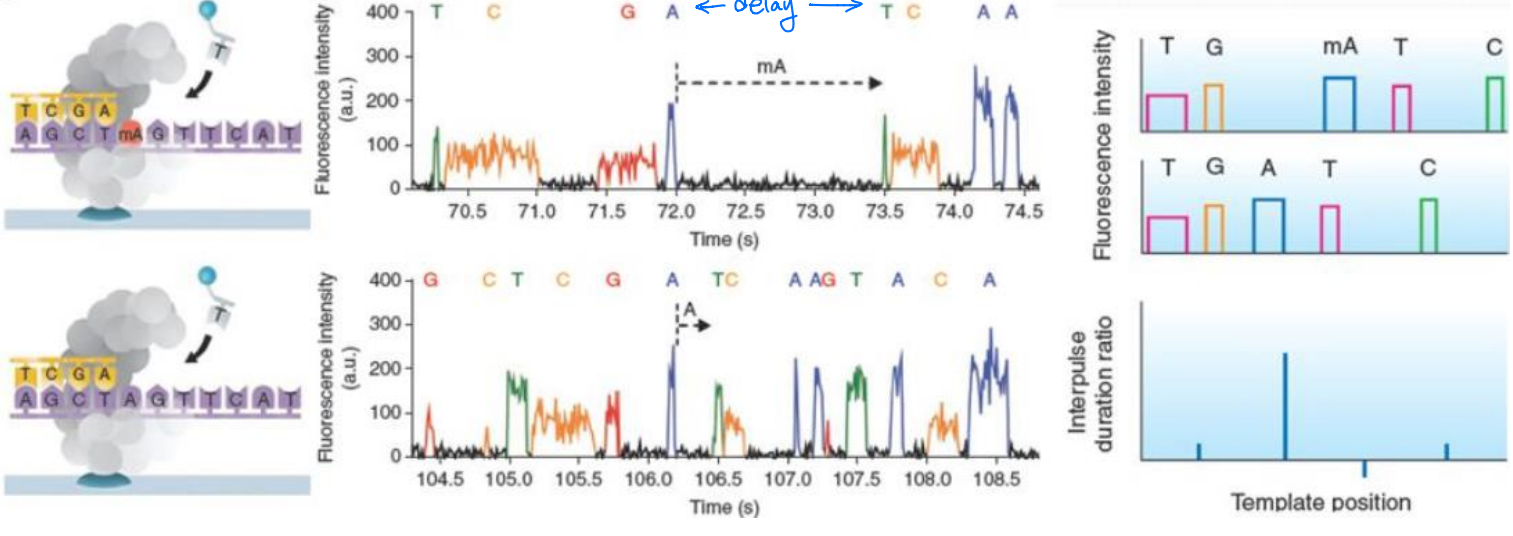
•Universal SMRTbell template accepts insert sizes from 250 bp to 40 kb

Redundant Sequencing Reading

Consensus accuracies > 99.999%

repeat several time → reduce error

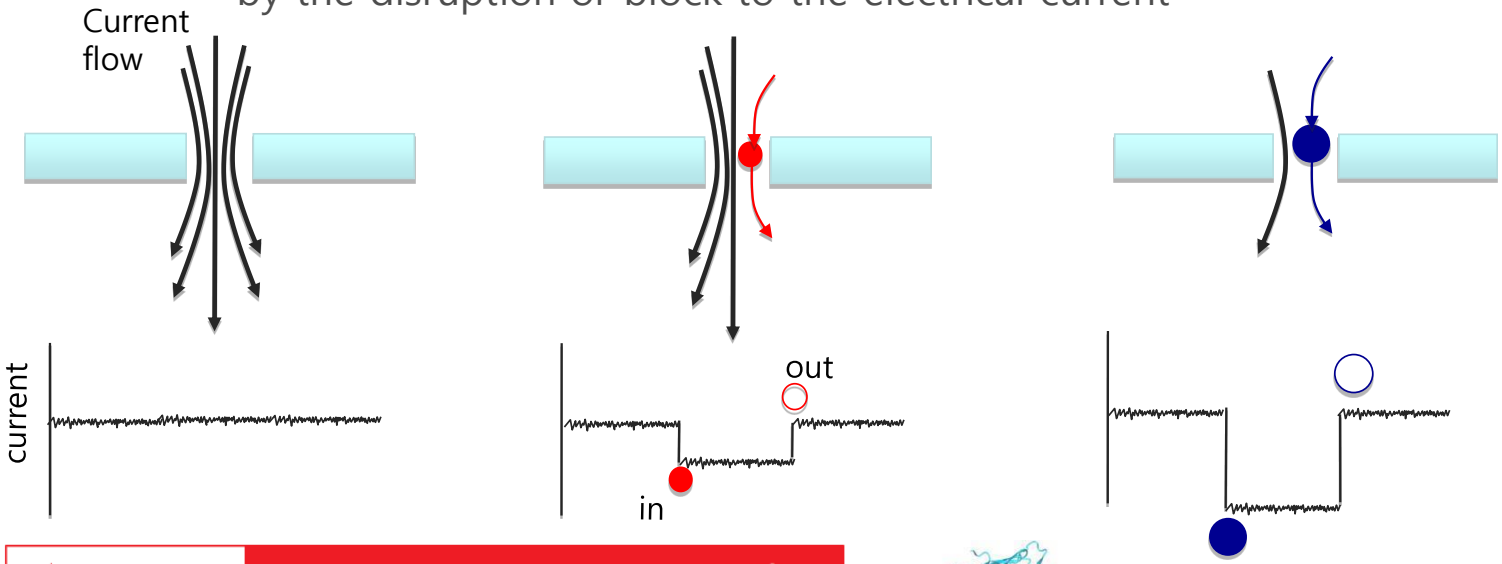
Realtime direct dectection of modified DNA



Oxford Nanopore

Electronic & Single molecule Real time

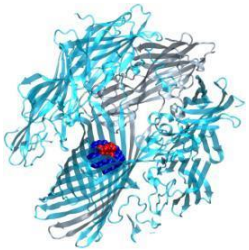
- Nanopore = 'very small hole'
- Electrical current flows through the hole
- Introduce analyte of interest into the hole ➔ identify "analyte" by the disruption or block to the electrical current



nature nanotechnology
ARTICLES
PUBLISHED ONLINE: XX XX 2009 | DOI: 10.1038/NNANO.2009.12

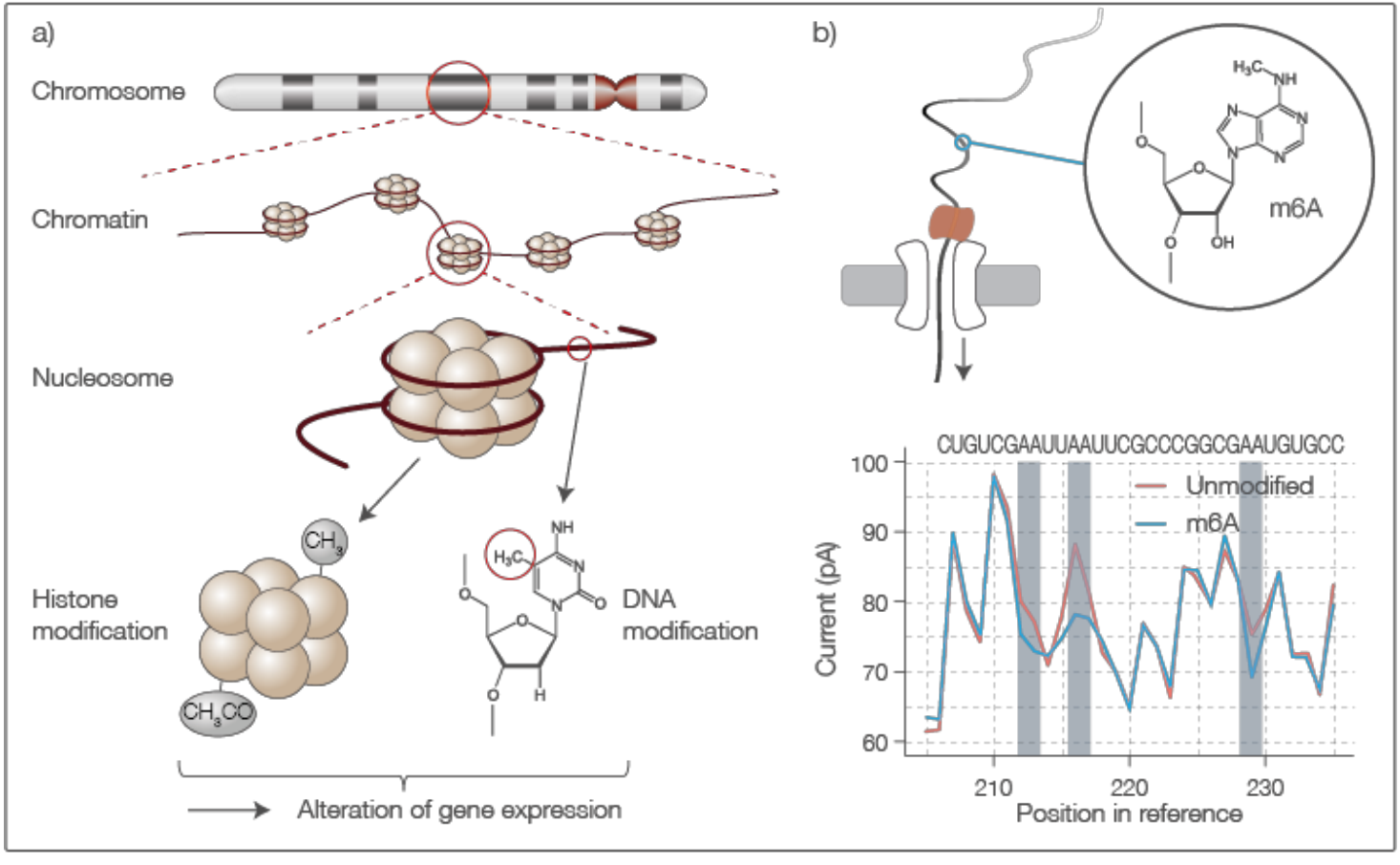
Continuous base identification for single-molecule nanopore DNA sequencing

James Clarke[1], Hai-Chen Wu[2], Lakmal Jayasinghe[1,2], Alpesh Patel[1], Stuart Reid[1] and Hagan Bayley[2]*



https://www.youtube.com/watch?v=BNz880V52rQ

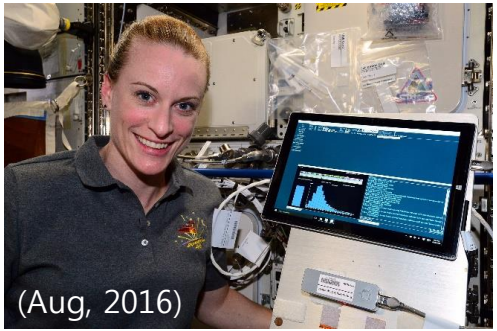
Direct detection of modified DNA/RNA by Oxford Nanopore

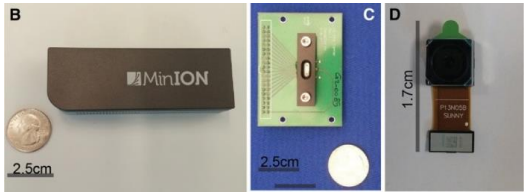


Realtime Portable Sequencer: Oxford Nanopore

Oxford Nanopore (MinION)

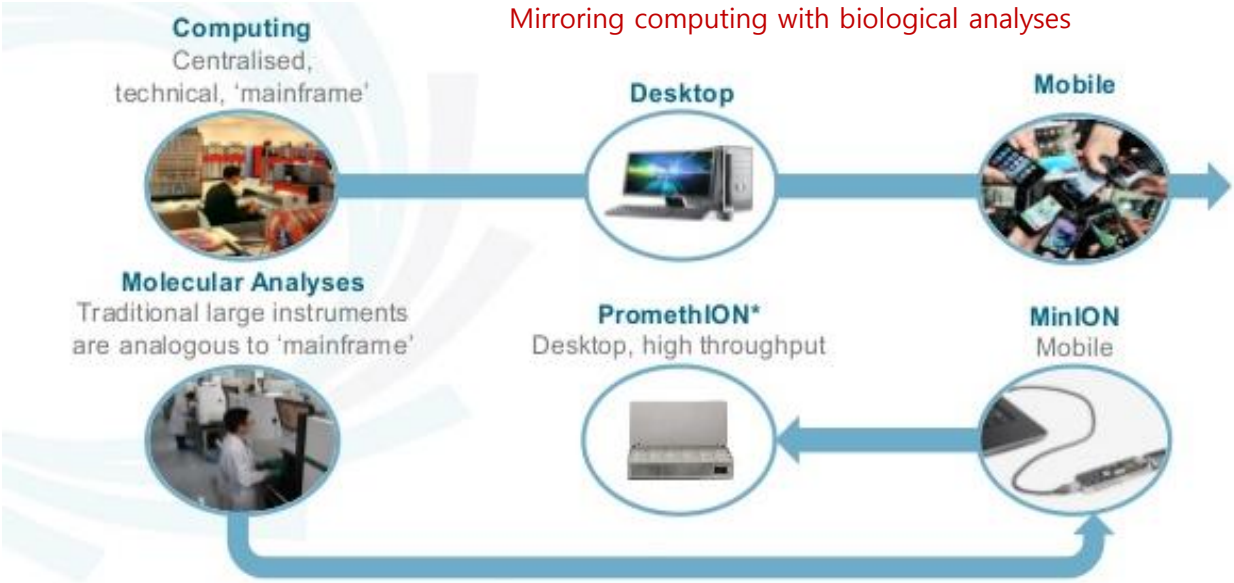
9th NASA/SpaceX commercial cargo





Creation of internet of living things (DNA)

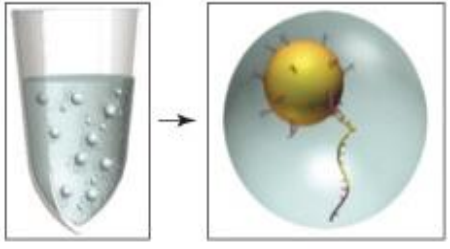
Mirroring computing with biological analyses



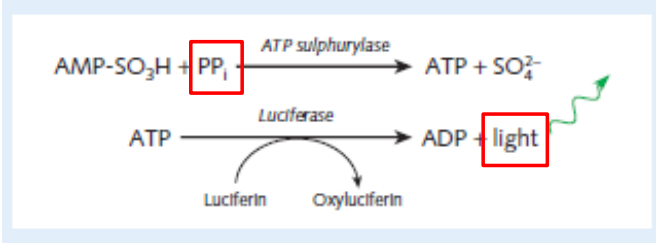
NGS platform comparison

Platform	Amplification	Chemistry	Detection	description
454	emPCR	Pyrosequencing	Image (mono)	Low throuput, long read (~400), ~0.35 day
Abi SOLiD	emPCR	Seq by ligation	Image (color)	High throuput, short read (~50), **~2weeks**
Illumina /Solexa	Solid-phase	Reversible teminator	Image (color)	High throuput, short read (~50), ~1 week
Ion torrent /proton	emPCR	Pyrosequencing	H+ (pH)	low-High throuput, short read (~50), 5 hours
Pacific Bioscience	Single molecule	realtime, polymerase	Image (color)	**longer read (3-15kb), Accuracy (~95%)**, 30min
Oxford Nanopore	Single molecule	realtime, polymerase	H+ (pH)	?

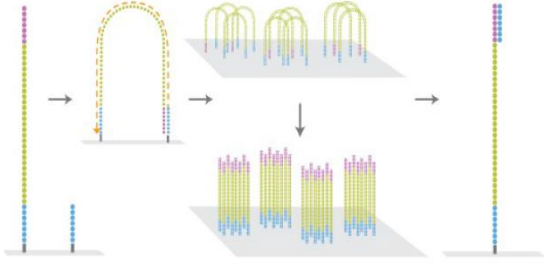
Emulsion PCR



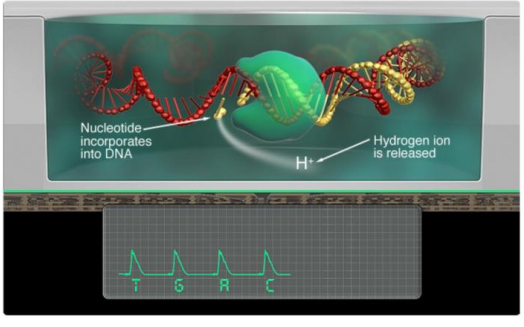
pyrosequencing



Solid-phase (bridge amplification)



Proton (pH meter)



NGS platform comparison

Method	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (US$)
Sanger ABI 3730×1	First	600–1000	0.001	96	0.5–3 h	500
Ion Torrent	Second	200	1	8.2×10^7	2–4 h	0.1
454 (Roche) GS FLX+	Second	700	1	1×10^6	23 h	8.57
Illumina HiSeq 2500 (High Output)	Second	2×125	0.1	8×10^9 (paired)	7–60 h	0.03
Illumina HiSeq 2500 (Rapid Run)	Second	2×250	0.1	1.2×10^9 (paired)	1–6 days	0.04
SOLiD 5500×l	Second	2×60	5	8×10^8	6 days	0.11
PacBio RS II: P6-C4	Third	$1.0–1.5 \times 10^4$ on average	13	$3.5–7.5 \times 10^4$	0.5–4 h	0.40–0.80
Oxford Nanopore MinION	Third	$2–5 \times 10^3$ on average	38	$1.1–4.7 \times 10^4$	50 h	6.44–17.90

From: Rhoads, A. and Au, K.F. (2015). PacBio Sequencing and Its Applications. Genomics, *Proteomics & Bioinformatics,* **13**, 278–289. (This article gives citations of sources of data.)

Application of Next-Generation Sequencing