Next Generation Sequencing

An Overview

What is sequencing?

DEFINITION

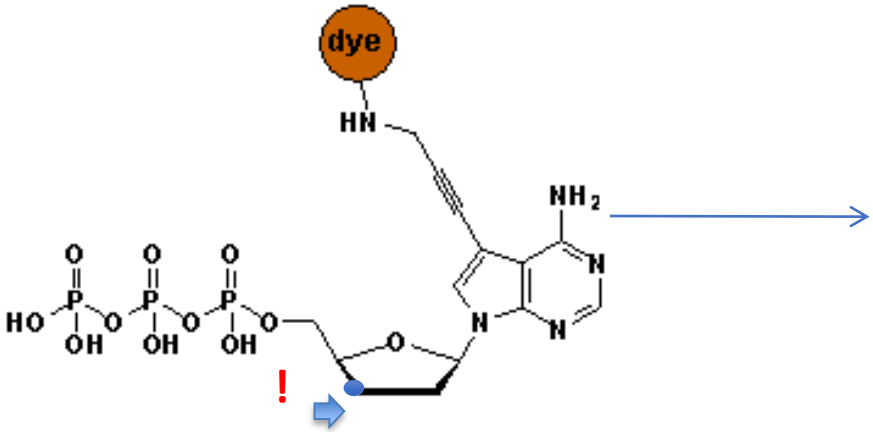
- “In genetics and biochemistry, **sequencing** means to determine the primary structure (or primary sequence) of an unbranched biopolymer.”

Principle

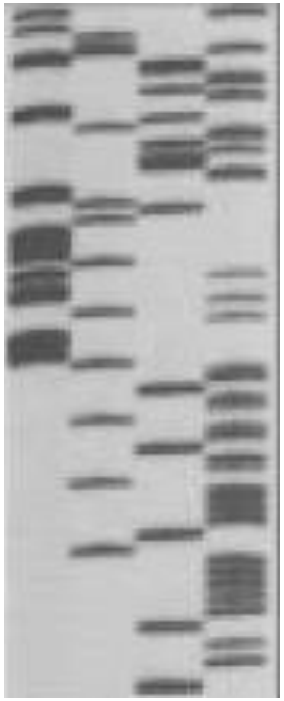
SYNTHESIS of DNA is randomly TERMINATED at different points

Separation of fragments that are 1 nucleotide different in size

Sanger's sequencing

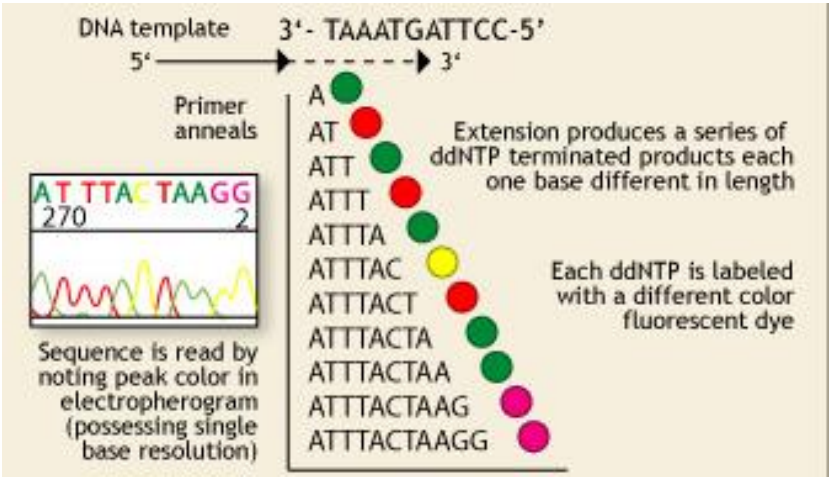


P^{32} labelled ddNTPs



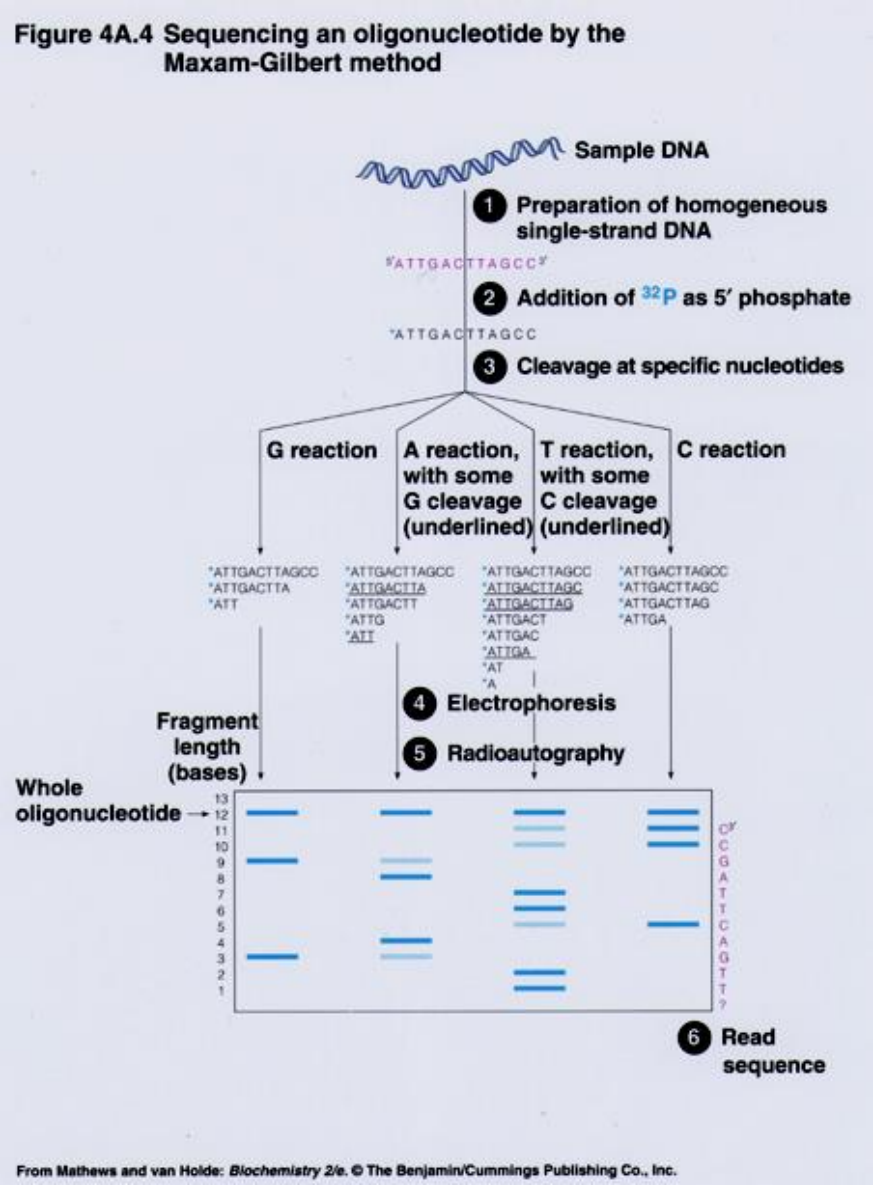
Lack of OH-group at 3' position of deoxyribose

Fluorescent dye terminators



Max fragment length – 750 bp

Maxam & Gilbert Sequencing



Sequencing genomes using **Sanger**'s method

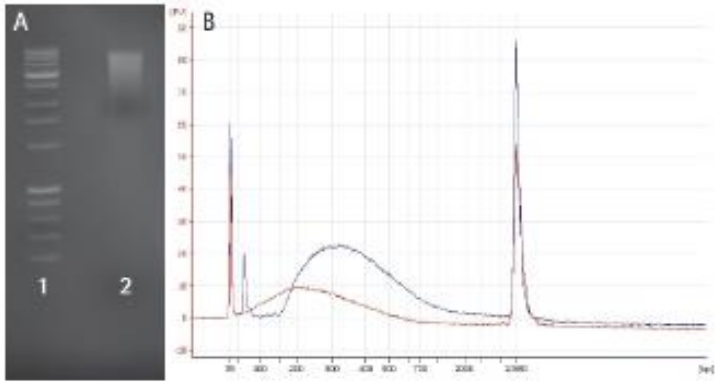
- Extract & purify genomic DNA
- Fragmentation
- Make a clone library
- Sequence clones
- Align sequencies (-> contigs -> scaffolds)
- Close the gaps
- Cost/Mb=1000 $, and it takes TIME

Major NGS technologies

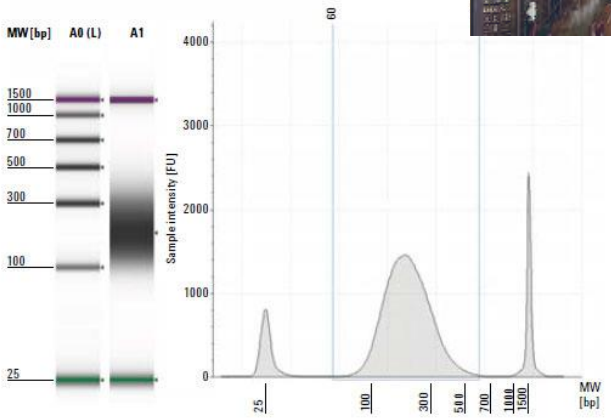
NGS technologies

Company	Platform	Amplification	Sequencing method
Roche	454**	emPCR	Pyrosequencing
Illumina	HiSeq MiSeq	Bridge PCR	Synthesis
LifeTech	SOLiD**	emPCR/ Wildfire	Ligation
LifeTech	Ion Torrent Ion Proton	emPCR	Synthesis (pH)
Pacific Bioscience	RSII	None	Synthesis
Complete genomics	Nanoballs	None	Ligation
Oxford Nanopore*	GridION	None	Flow

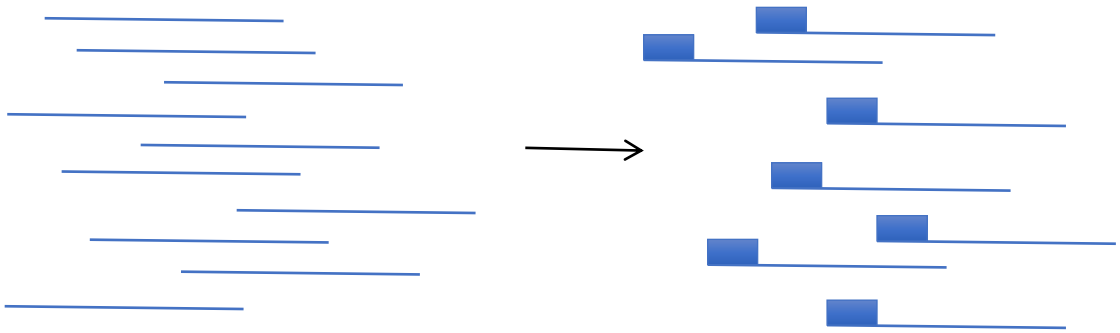
Making a NGS library



DNA QC – **paramount importance**



Sharing & size selection



Amplification

Ligation of sequencing adaptors, technology specific

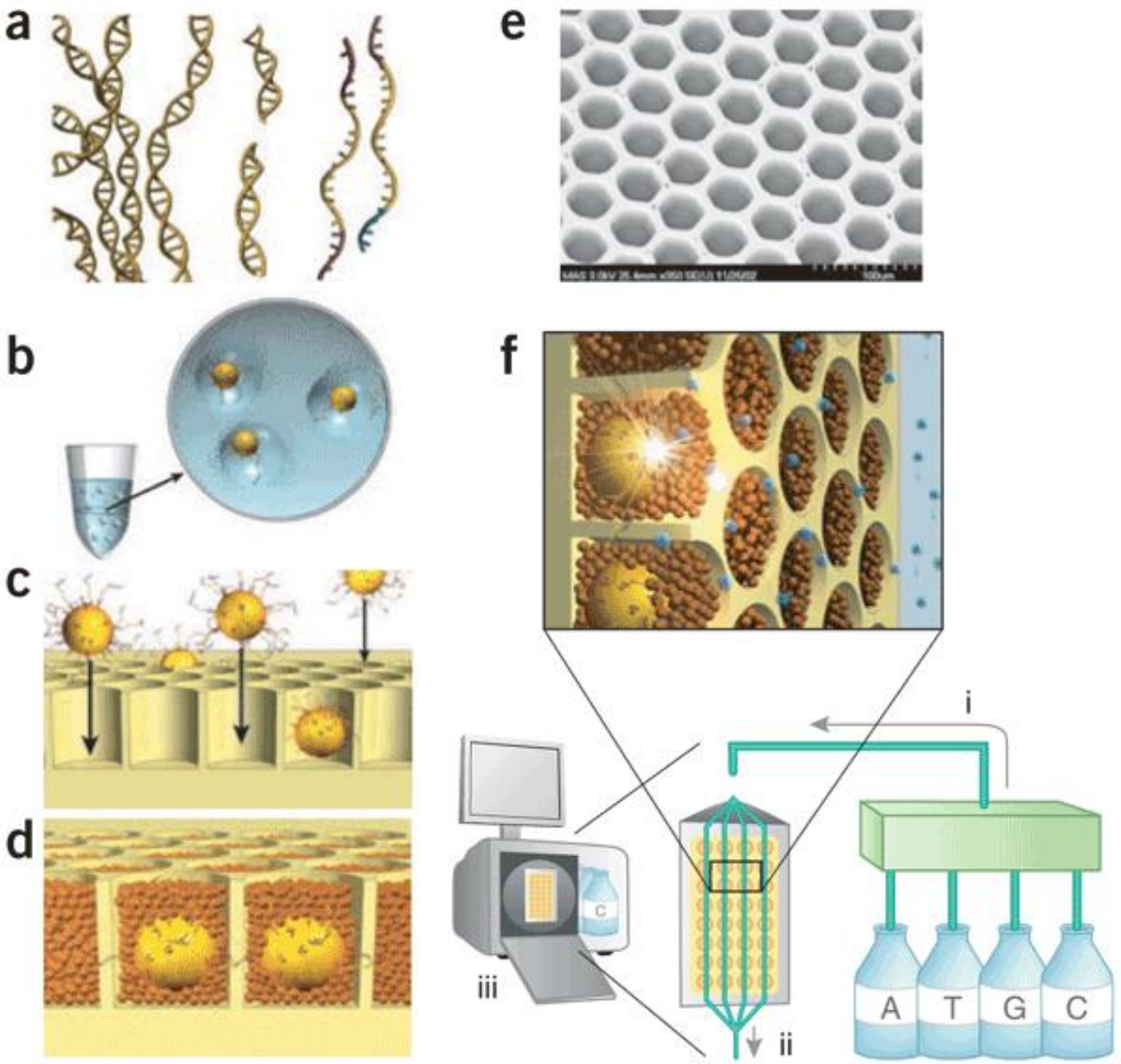
Roche

Instrument	Yield and run time	Read Length	Error rate	Error type
454 FLX+	0.9 GB, 20 hrs	700	1%	Indels
454 FLX Titanium	0.5 GB, 10 hrs	450	1%	Indels
454 FLX Jr	0.050 GB, 10 hrs	400	1%	Indels

Main applications:

- Microbial genomics and metagenomics
- Targeted resequencing

454 Titanium GS FLX



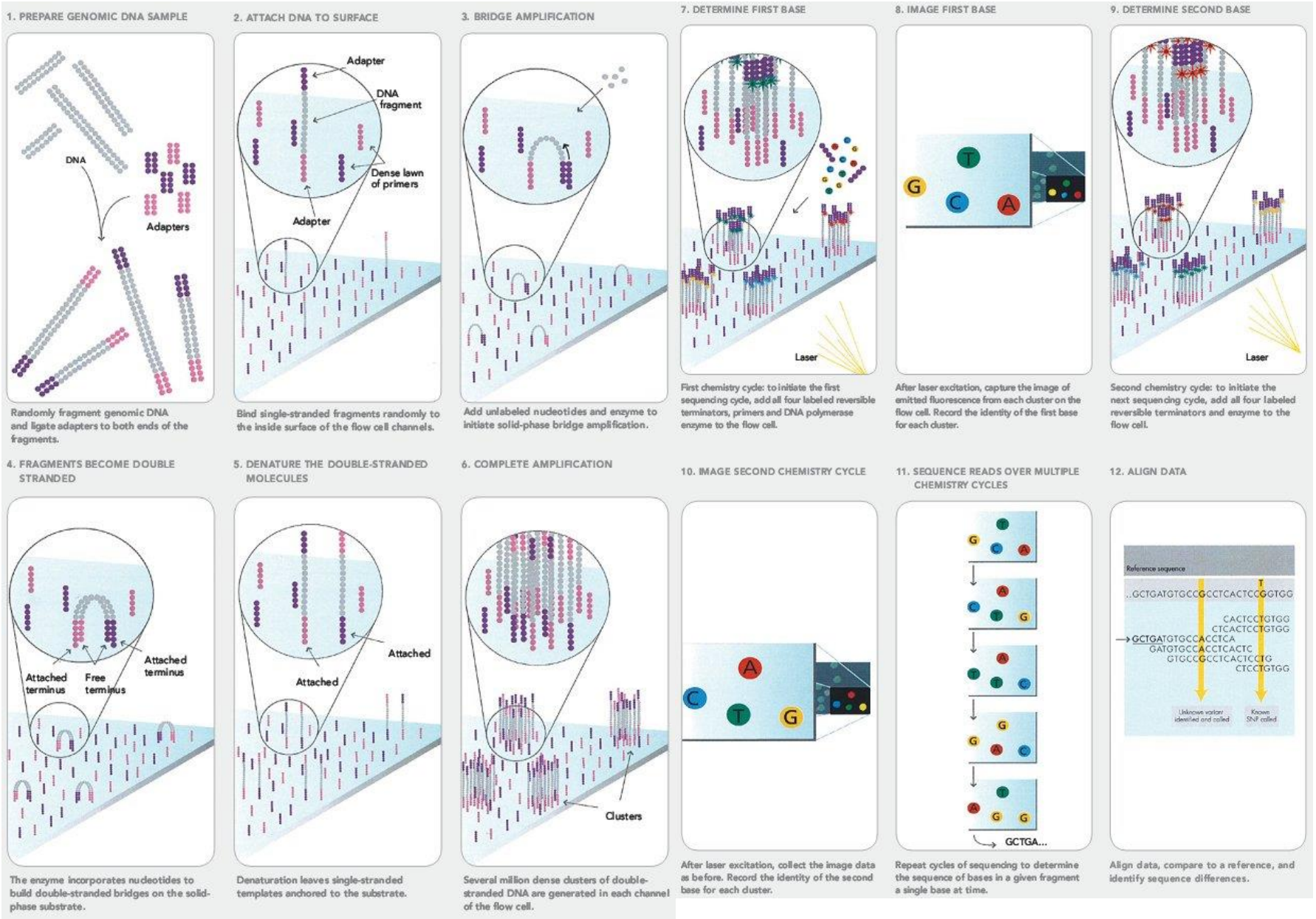
Illumina

Instrument	Yield and run time	Read Length	Error rate	Error type
Upgrade HiSeq2500	120 GB in 27h or standard run	100x100	0.1%	Subst
MiSeq	540 Mb – 15 Gb (4 – 48 hours)	Upp to 350x350	0.1%	Subst

Main applications

- Whole genome, exome and targeted reseq
- Transcriptome analyses
- Methylome and ChiPSeq
- Rapid targeted resequencing (MiSeq)

Illumina


1. PREPARE GENOMIC DNA SAMPLE
DNA
Adapters
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
2. ATTACH DNA TO SURFACE
Adapter
DNA fragment
Dense lawn of primers
Adapter
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
3. BRIDGE AMPLIFICATION
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
7. DETERMINE FIRST BASE
Laser
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.
8. IMAGE FIRST BASE
T
G
C
A
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.
9. DETERMINE SECOND BASE
Laser
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.
4. FRAGMENTS BECOME DOUBLE STRANDED
Attached terminus
Attached terminus
Free terminus
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.
5. DENATURE THE DOUBLE-STRANDED MOLECULES
Attached
Attached
Denaturation leaves single-stranded templates anchored to the substrate.
6. COMPLETE AMPLIFICATION
Clusters
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.
10. IMAGE SECOND CHEMISTRY CYCLE
A
C
T
G
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.
11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES
GCTGA...
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.
12. ALIGN DATA
Reference sequence
..GCTGATGTGCCGCCTCACTCCGGTGG
CACTCCTGTGG
CTCACTCCTGTGG
GCTGATGTGCCACCTCA
GATGTGCCACCTCACTC
GTGCCGCCTCACTCCTG
CTCCTGTGG
Unknown variant identified and called
Known SNP called
Align data, compare to a reference, and identify sequence differences.

Life Technologies SOLiD

Instrument	Yield and run time	Read Length	Error rate	Error type
SOLiD 5500 wildfire	600 GB, 8 days	75x35 PE 60x60 MP	0.01%	A-T Bias

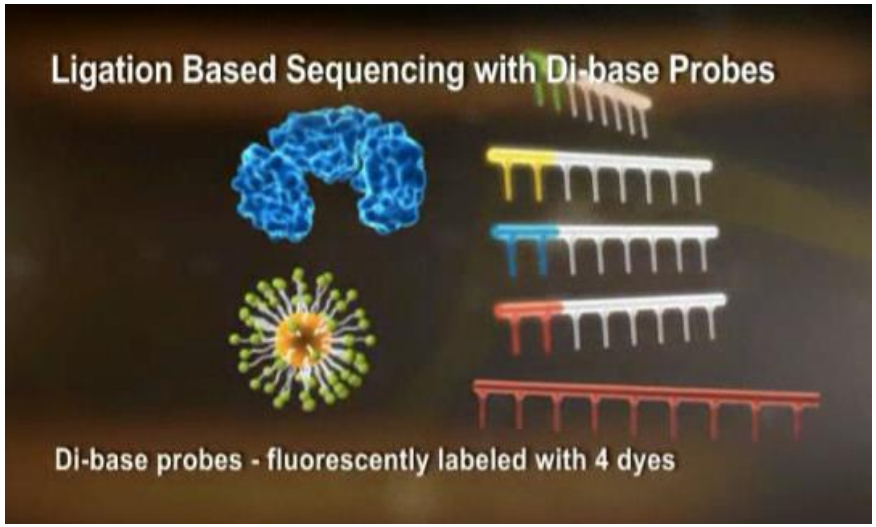
Features

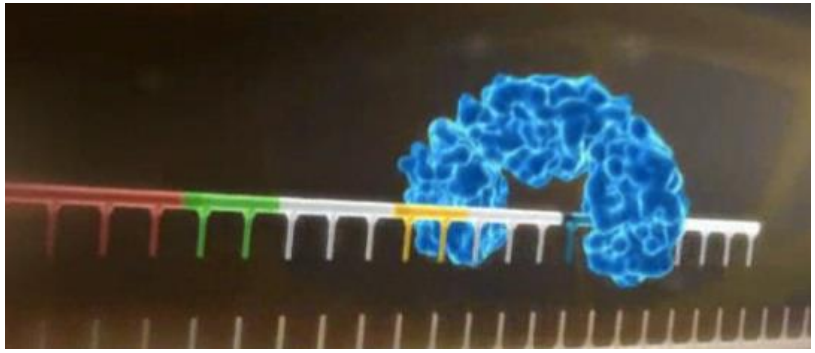
- High accuracy due to two-base encoding
- True paired-end chemistry - ligation from either end
- Mate-pair libraries

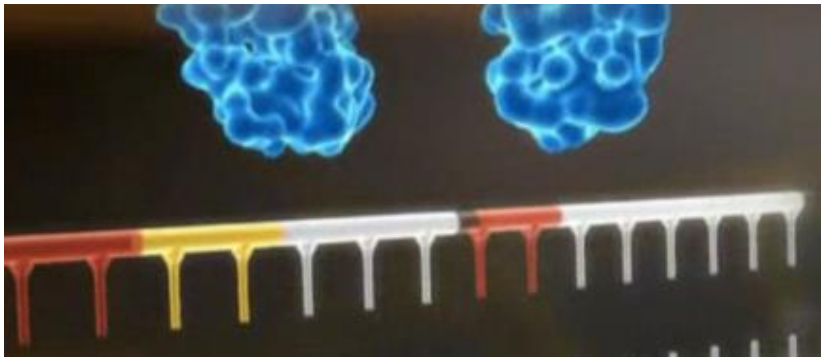
Main applications (currently)

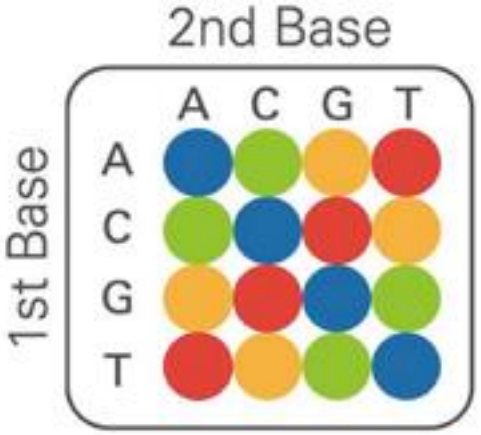
•ChiPSeq

SOLiD - ligation









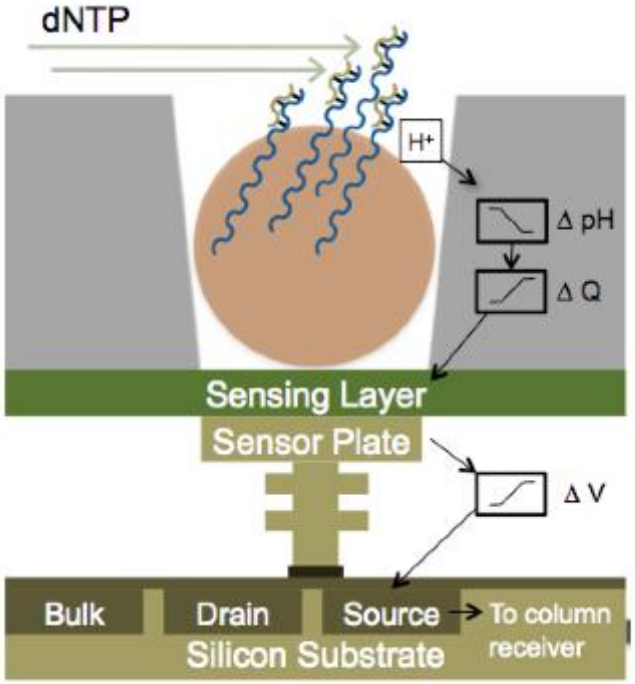
Life Technologies - Ion Torrent & Ion Proton

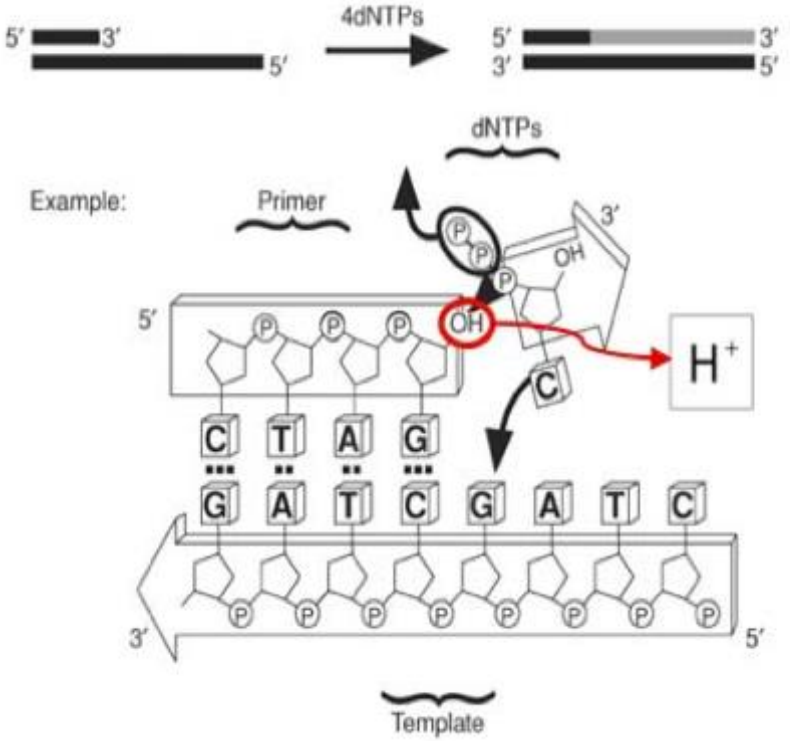
Chip	Yield - run time	Read Length
PGM 314	0.1 GB, 3 hrs	200 – 400
PGM 316	0.5GB, 3 hrs	200 - 400
PGM 318	1 GB, 3 hrs	200 - 400
P-I	10 GB	200

Main applications

- Microbial and metagenomic sequencing
- Targeted resequencing
- Clinical sequencing

Ion Torrent - H^+ ion-sensitive field effect transistors

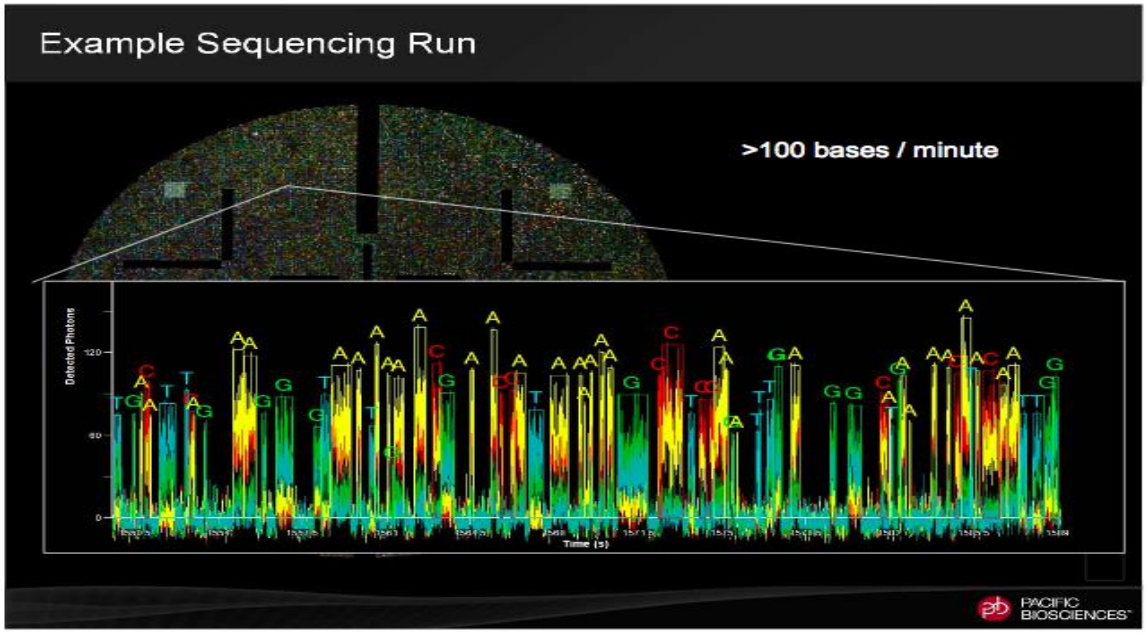




Pacific Bioscience

Instrument	Yield and run time	Read Length	Error rate	Error type
RS II	500 MB/180 min SMRTCell	250 bp – 20 000 bp (35 000 bp)	15% (on a single passage!)	Insertions, **random**

Single-Molecule, Real-Time DNA sequencing



NGS technologies - SUMMARY

Platform	Read length	Accuracy	Projects / applications
454	Medium	Homo- polymer runs	Microbial + targeted reseq
HiSeq MiSeq	Short Medium	High	Whole genome + transcriptome seq, exome
SOLiD	Short	High	Whole genome + transcriptome seq, exome
Ion Torrent	Medium	High	Microbial + targeted reseq
Ion Proton	Short/Medium	High	Exome, transcriptome, genome
PacBio	Long	Low – ultra high*	Microbial + targeted reseq Gap closure & scaffolding

	Illumina HiSeq	Illumina MiSeq	SOLiD Wildfire	Ion Torrent	Ion Proton	PacBio
Read length	100 + 100 bp (150+150 bp)	250 + 250 bp (350+350 bp)	75 bp	200 bp 400 bp (500 bp)	150 bp 200 bp	1 – 20 Kbp
WGS: - human - small	 ++++ +++	 +++	 **(+)** **(+)**	 ++++	 + +++	 **(+)** +++++
De novo	+++	++		+++	++	+++++
RNA-seq miRNA	+++ +++		+++ +++		+++	+++*
ChIP	+++		++++			
Amplicon	++	+++		+++	+++	+++
Metylation	+++					++++*
Target re-seq	++	+++	**(+)**		+++	+++
Exome	+++		**(+)**		++++	**(+)**