BHARATHIDASAN UNIVERSITY
Tiruchirappalli- 620024
Tamil Nadu India

Programme: M.Sc., Biotechnology (Environment)

Course Title : Genetic Engineering
Course Code: CC 07

Unit-IV
DNA Sequencing and In Vitro Translation

Dr T Sivasudha

Professor

Department of Environmental Biotechnology

„Next-Generation“ Sequencing?

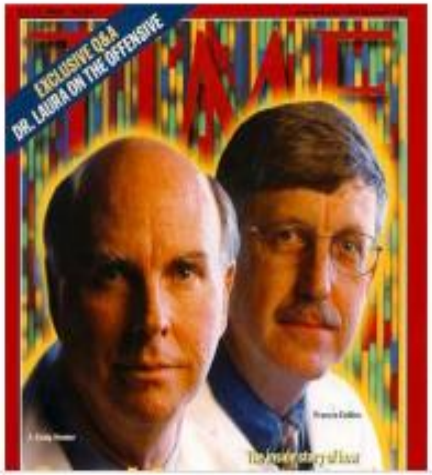
- Sanger technique published in 1977, based on the polymerase chain reaction (PCR)
  - → high quality, long reads
  - → slow, expensive
  - ➔ „First Generation“

http://upload.wikimedia.org/wikipedia/commons/thumb/f/f8/Frederick_Sanger2.jpg/195px-Frederick_Sanger2.jpg

http://www.lifesciencesfoundation.org/content/im

- Human Genome Project: 1990 – 2004
- Investments >3,000,000,000 $ (~1$/bp)

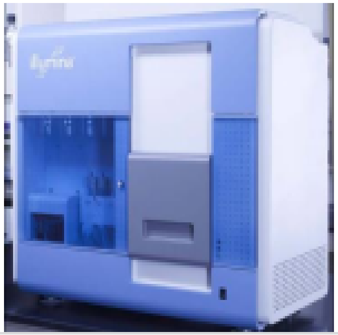
- Development of cheaper „NGS“ technologies
- Compete with microarray technology*

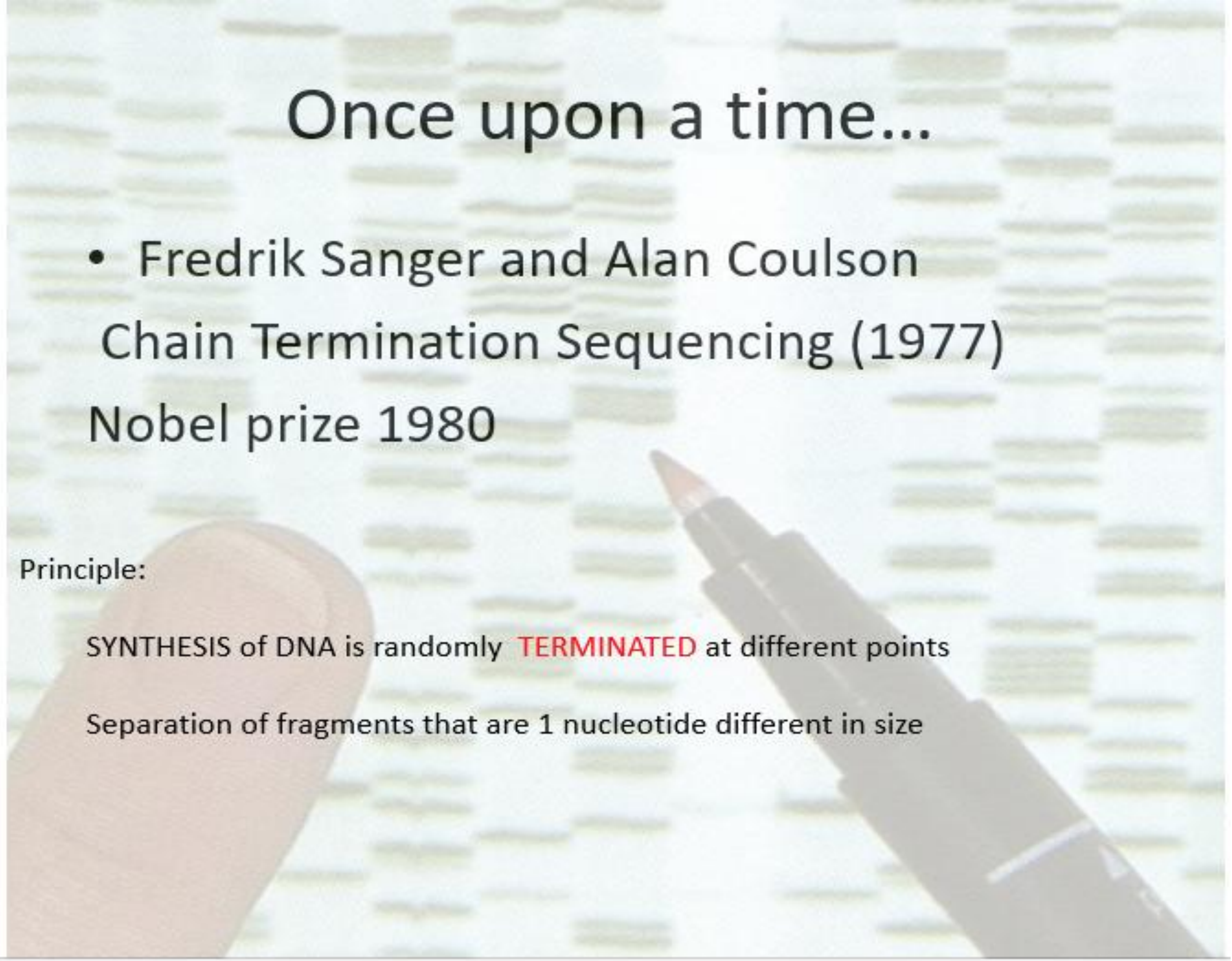


yhub.com/wp-
/2010/05/Venter_Collins_Genome.jpg

NGS depends on massive, molecular parallelization

- Four major platform types (and further ones):
  - Solexa/Illumina (Genome Analyzer, MiSeq, HiSeq)
  - Applied Biosystems (ABi; SOLiD)
  - Roche/454 (GS FLEX, GS Junior)
  - Ion (Torrent PGM, Proton)
- Advantages: fast (massively parallel) & cheap (low cost per bp)
- Disadvantages: error-prone, extensive needs in computation time and storage





Once upon a time…

- Fredrik Sanger and Alan Coulson

Chain Termination Sequencing (1977)

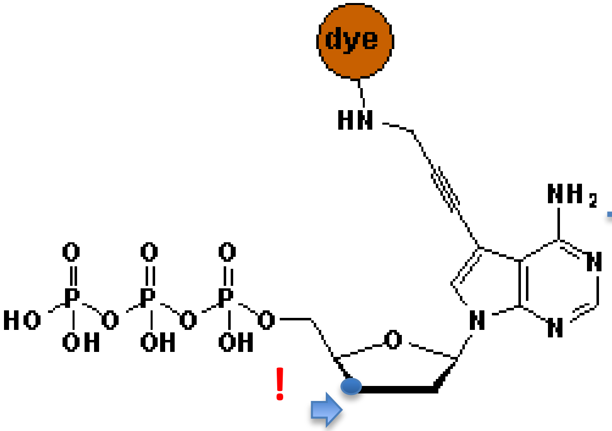
Nobel prize 1980

Principle:

SYNTHESIS of DNA is randomly TERMINATED at different points

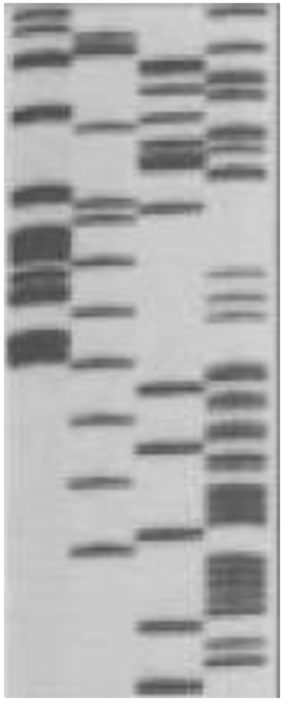
Separation of fragments that are 1 nucleotide different in size

Sanger’s sequencing

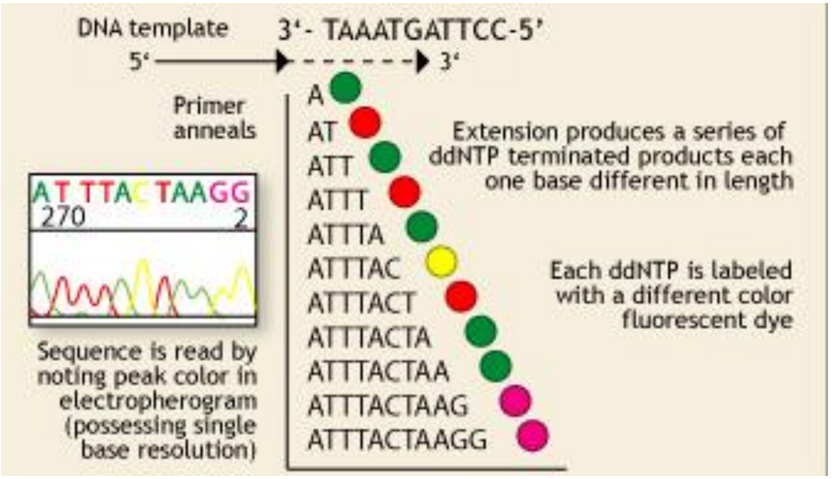


Lack of OH-group at 3’ position of deoxyribose

P^{32} labelled ddNTPs



Fluorescent dye terminators



Extension produces a series of ddNTP terminated products each one base different in length

Each ddNTP is labeled with a different color fluorescent dye

Sequence is read by noting peak color in electropherogram (possessing single base resolution)

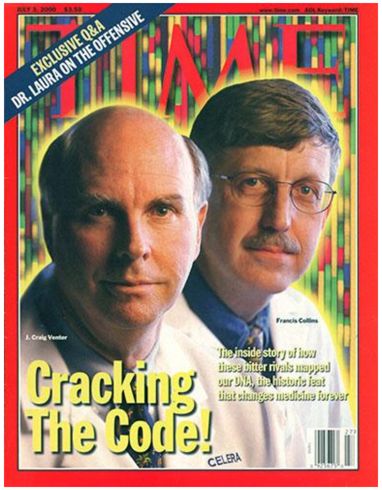
Max fragment length – 750 bp

Sequencing genomes using **Sanger**’s method

- Extract & purify genomic DNA
- Fragmentation
- Make a clone library
- Sequence clones
- Align sequencies (-> contigs -> scaffolds)
- Close the gaps

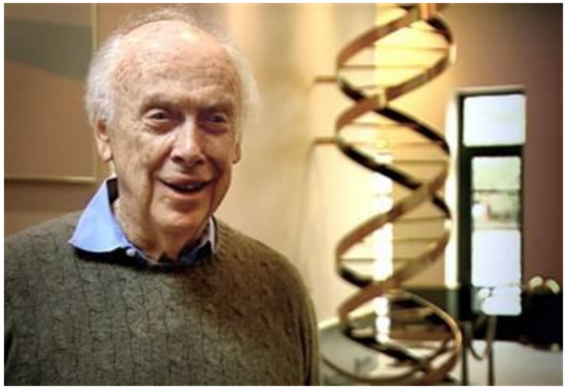
- Cost/Mb=1000 $, and it takes TIME

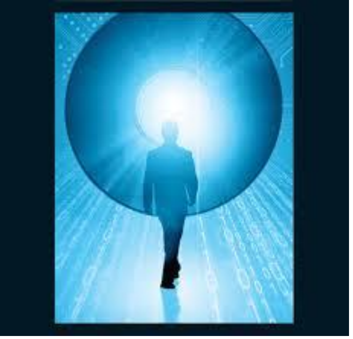
Just an interesting comparison:


TIME
EXCLUSIVE Q&A
DR. LAURA ON THE OFFENSIVE
J. Craig Venter
Francis Collins
Cracking The Code!
The inside story of how these bitter rivals mapped our DNA, the historic feat that changes medicine forever
CELERA

- Human genome project, 2007
  - Genome of Craig Wenter costs 70 mln $
    - Sanger's sequencing
  - Genome of James Watson costs 2 mln $
    - 454 pyrosequencing
  - Ultimate goal: 1000 $ / individual

    Almost there!

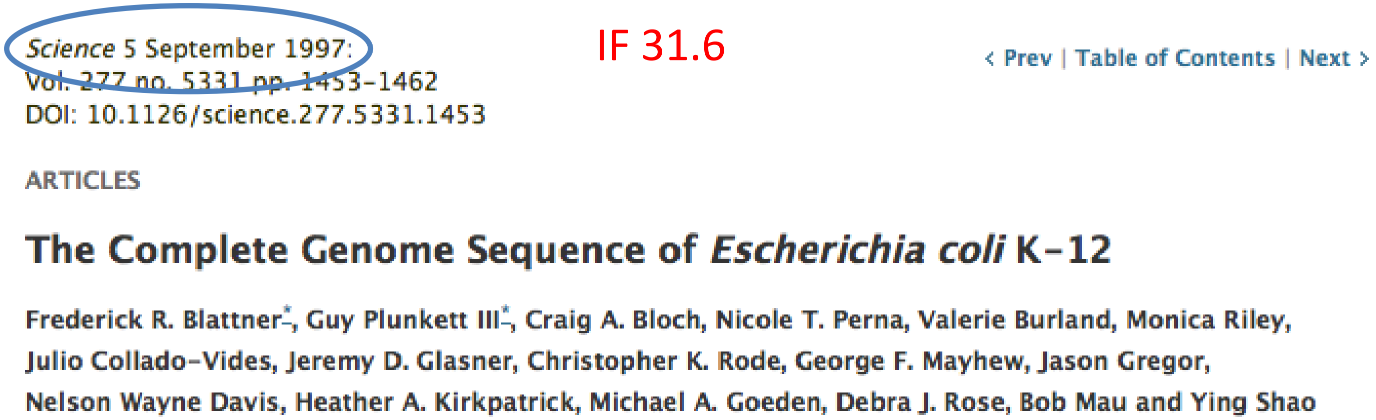




Paradigm change

- From single genes to complete genomes
- From single transcripts to whole transcriptomes
- From single organisms to complex metagenomic pools
- From model organisms to the species you are studying

BEHOLD...THE ANSWER
TO LIFE, THE UNIVERSE AND EVERYTHING
42
SORRY- WHAT WAS THE QUESTION AGAIN?
THE HITCHHIKER'S GUIDE TO THE GALAXY



IF 31.6

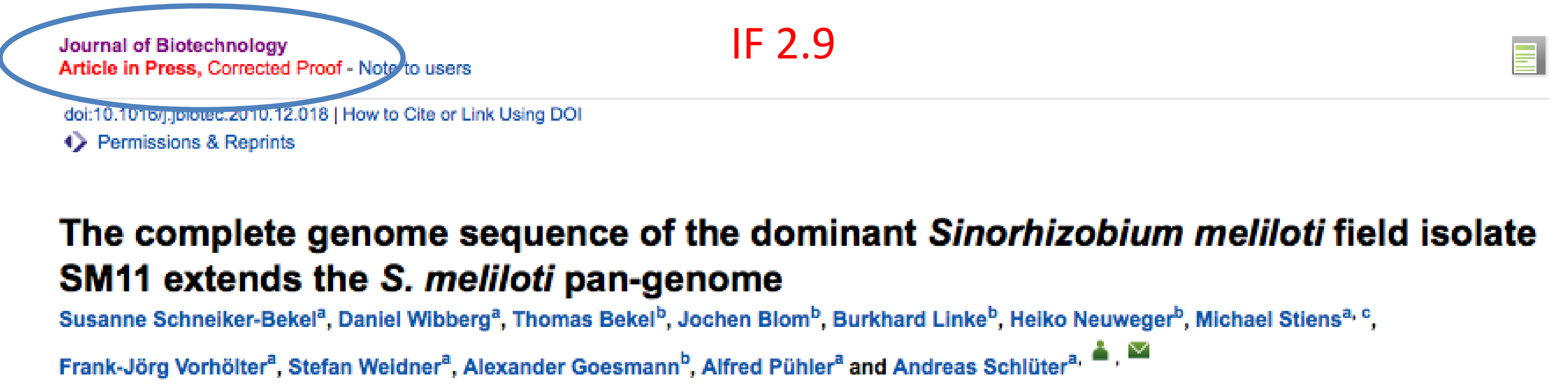
Science 5 September 1997:
Vol. 277 no. 5331 pp. 1453–1462
DOI: 10.1126/science.277.5331.1453

‹ Prev | Table of Contents | Next ›

ARTICLES

The Complete Genome Sequence of *Escherichia coli* K-12

Frederick R. Blattner[*], Guy Plunkett III[*], Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau and Ying Shao



IF 2.9

Journal of Biotechnology
Article in Press, Corrected Proof - Note to users

doi:10.1016/j.jbiotec.2010.12.018 | How to Cite or Link Using DOI
Permissions & Reprints

The complete genome sequence of the dominant *Sinorhizobium meliloti* field isolate SM11 extends the *S. meliloti* pan-genome

Susanne Schneiker-Bekel[a], Daniel Wibberg[a], Thomas Bekel[b], Jochen Blom[b], Burkhard Linke[b], Heiko Neuweger[b], Michael Stiens[a, c], Frank-Jörg Vorhölter[a], Stefan Weidner[a], Alexander Goesmann[b], Alfred Pühler[a] and Andreas Schlüter[a], ,

a bit of… Theory

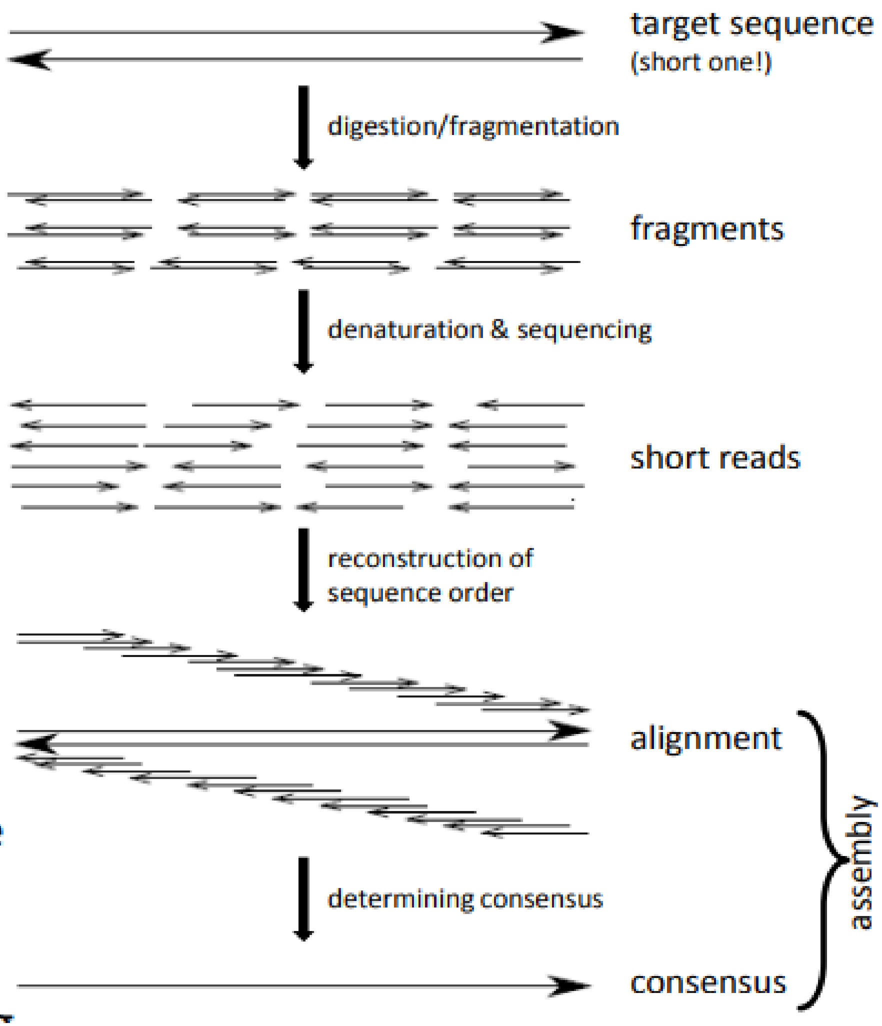
- per-base quality (sequencing errors)
- short length of reads:
  - more gaps
  - more comparisons
  - more uncertain positions
  - may not bridge repetetive regions
  - …
- coverage problem in genomics (Lander-Waterman 1988):

  How much read data must be generated in order to ensure a certain statistical correctness of the consensus sequence?
  - Poisson distribution of residing gaps
  - „Coverage“ (e.g. 10x) applies to read length

→ NGS technology requires a multiple of data to be generated compared to the original sequence

→ but: massive cost reduction (per base)

➔ efforts shift from lab to computing



target sequence
(short one!)

digestion/fragmentation

fragments

denaturation & sequencing

short reads

reconstruction of
sequence order

alignment

determining consensus

consensus

assembly

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

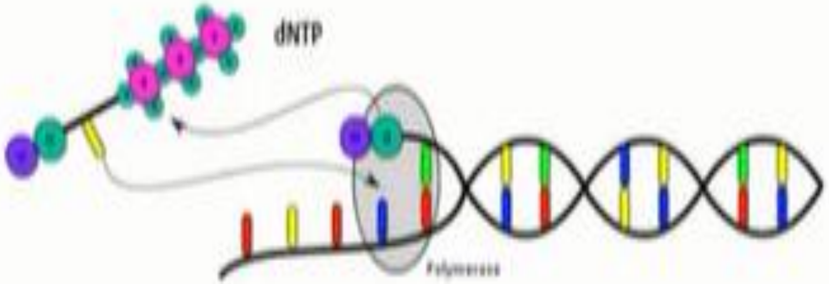
RIP technologies: Helicos, Polonator, etc.

In development: Tunneling currents, nanopores, etc.

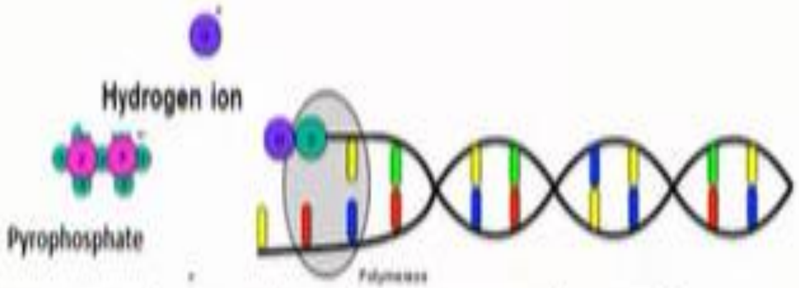
Differences between platforms

- Technology: chemistry + signal detection
- Run times vary from hours to days
- Production range from Mb to Gb
- Read length from <100 bp to > 20 Kbp
- Accuracy per base from 0.1% to 15%
- Cost per base varies

Products of base addition reaction



Polymerase integrates a nucleotide.



Hydrogen and pyrophosphate are released.

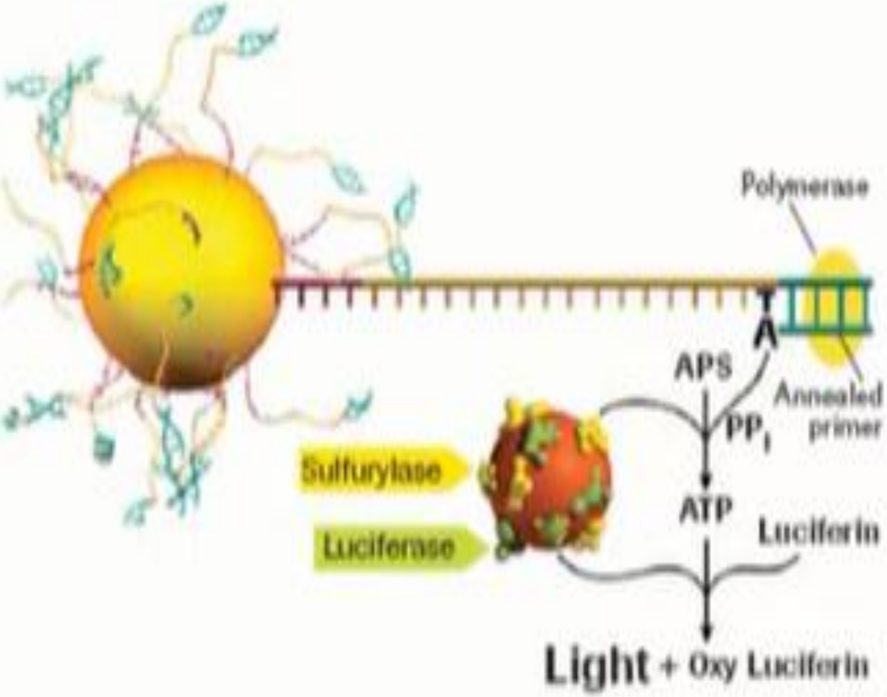
Second (next) generation sequencing technologies

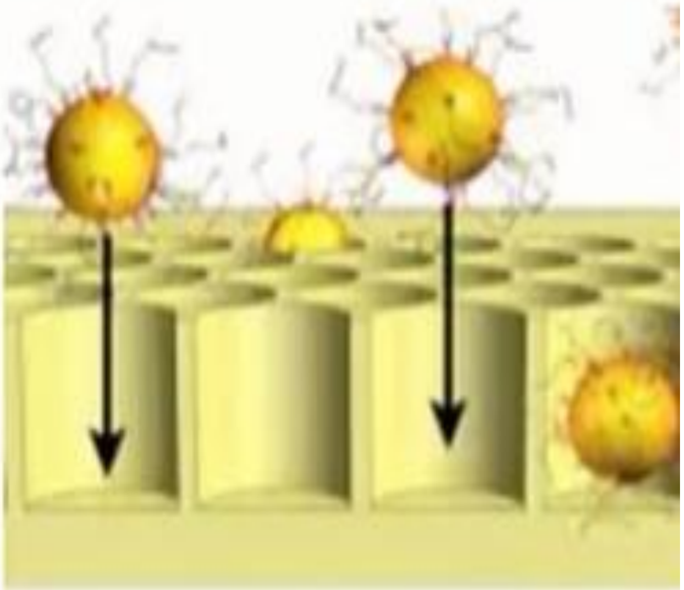
Company	Platform	Method	Detection	Length	Advantages	Disadvantages
Roche/454	FLX genome sequencer	Pyrosequencing Detecion of pyrophosphate release	Optical	0.4-1 Kb	Long read length	High cost; challenging sample prep.
Life Technologies	IonPGM IonProton	Sequencing by synthesis	Released H+ ions	200 bp	Rapid runs, low cost	Lower throughput compared to Ilumina; Maturing technology
Illumina	HiSeq 2500 MiSeq	Rev. terminator sequencing by synthesis	Fluorescence/ optical	2x150 or 2x250 bp	Very high throughput	Long run time for standard runs
Life technologies	5500 SOLiD W system	Sequencing by ligation	Fluorescence/ optical	1x75 or 2x60 bp	Very high throughput	Short read lengths; non-standard data analysis

Illumina platform is market leader – one 30x coverage human genome for $5-10k

Sequencing by synthesis:

Detection and estimation of pyrophosphate release

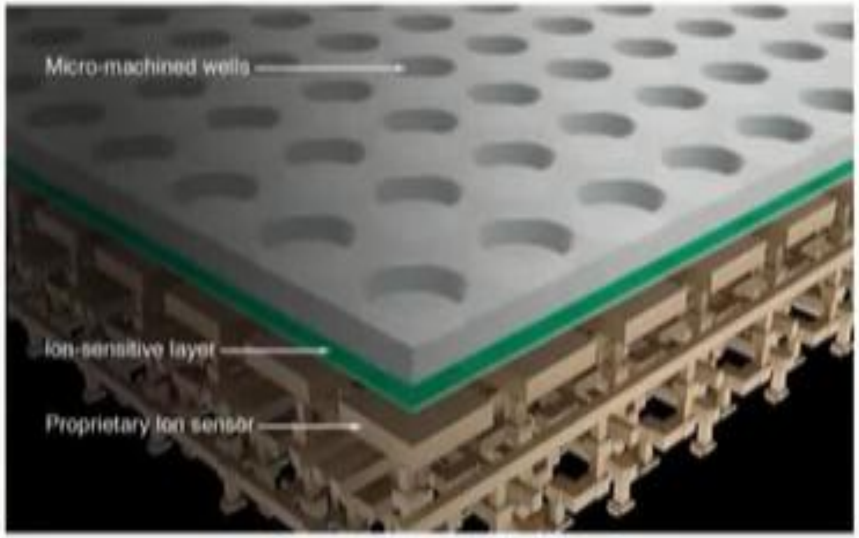


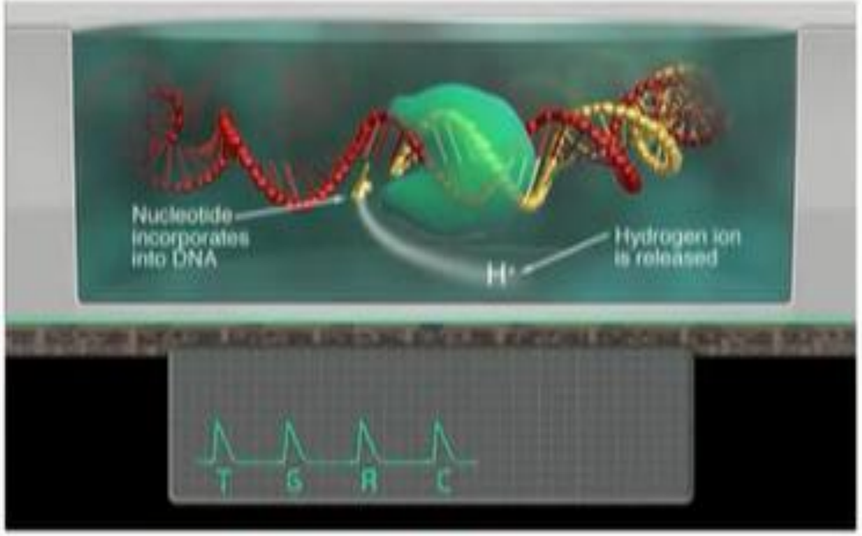


454/Roche: First NGS sequencer in market

Sequencing by detecting H^+ release after addition of the base

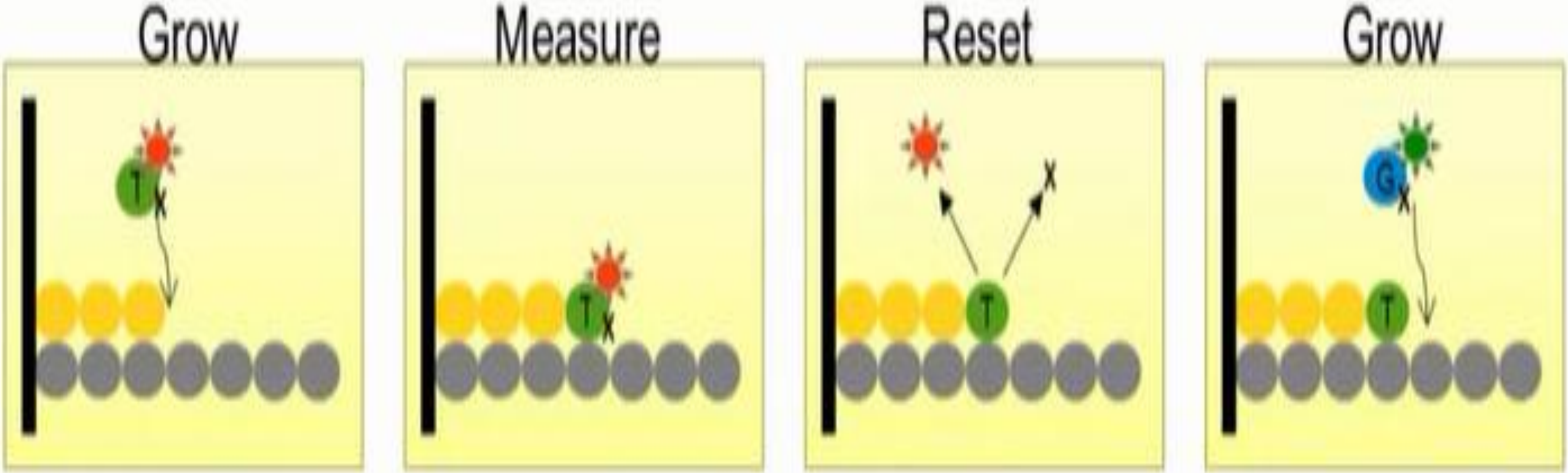
- Polymerase releases H+ during base incorporation
- Measured by semi-conductor wafer
- Essentially a massively parallel pH meter



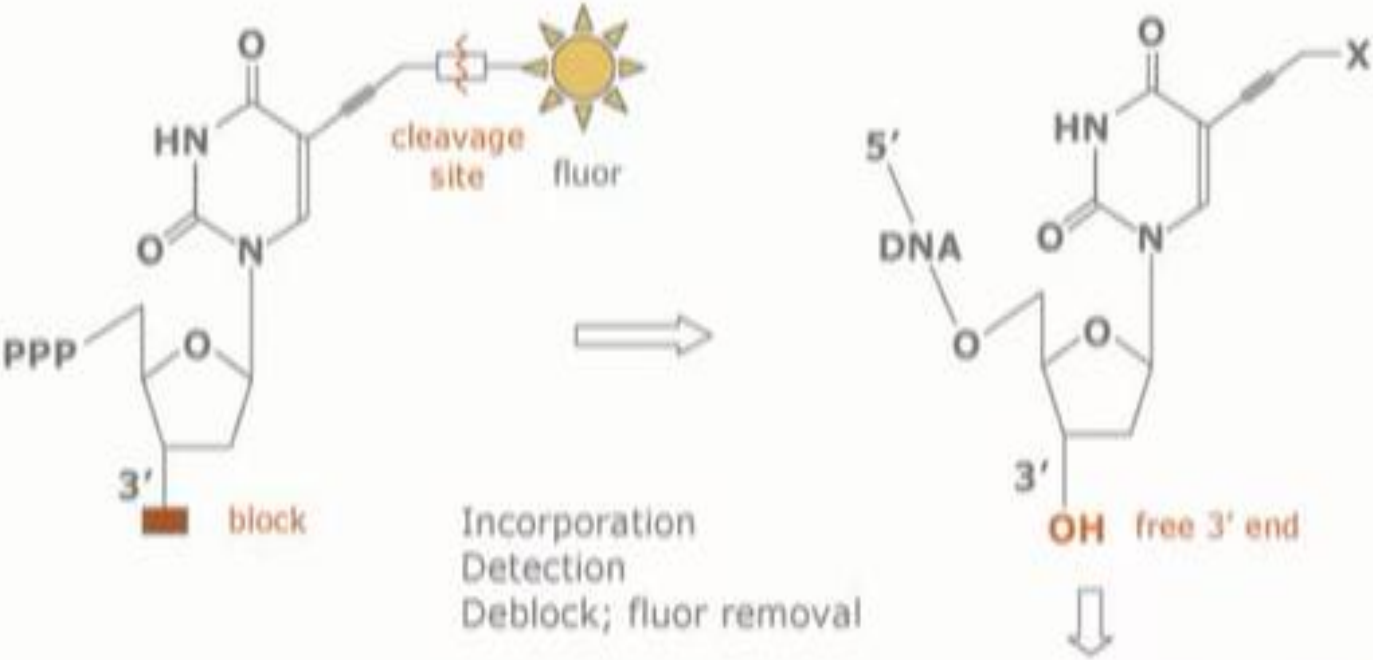


Life/IonTorrent 'Electrical' Sequencing

Dye sequencing by synthesis using reversible terminator



Reversible dye chemistry



Incorporation
Detection
Deblock; fluor removal

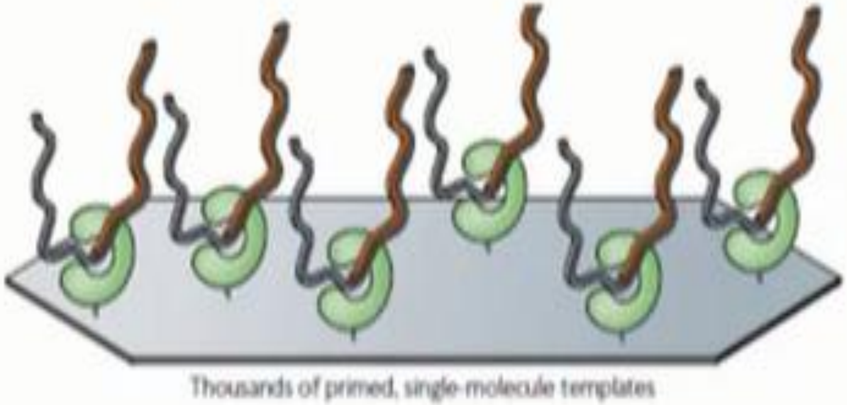
Next cycle

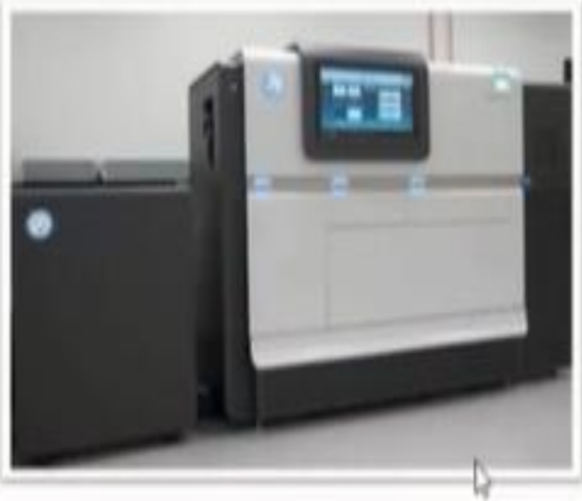
illumina

Limitations of second-generation sequencing

- Second generation sequencing requires amplification to get sufficient number of sequences to meet detection thresholds.
  - Coverage of GC rich sequences
  - Amplification bias
    - Inherently a problem for quantitation
    - Unique molecular identifiers however may solve this problem (Islam et al., Nat Methods, 2014; Kivioja et al., Nat Methods, 2012)—see Week 1
- Second Gen Seq technologies have practical limits in read length
  - Mapping of long repeat regions in the genome
  - Identification and mapping of duplicate genes and pseudogenes

PacBio real-time sequencing



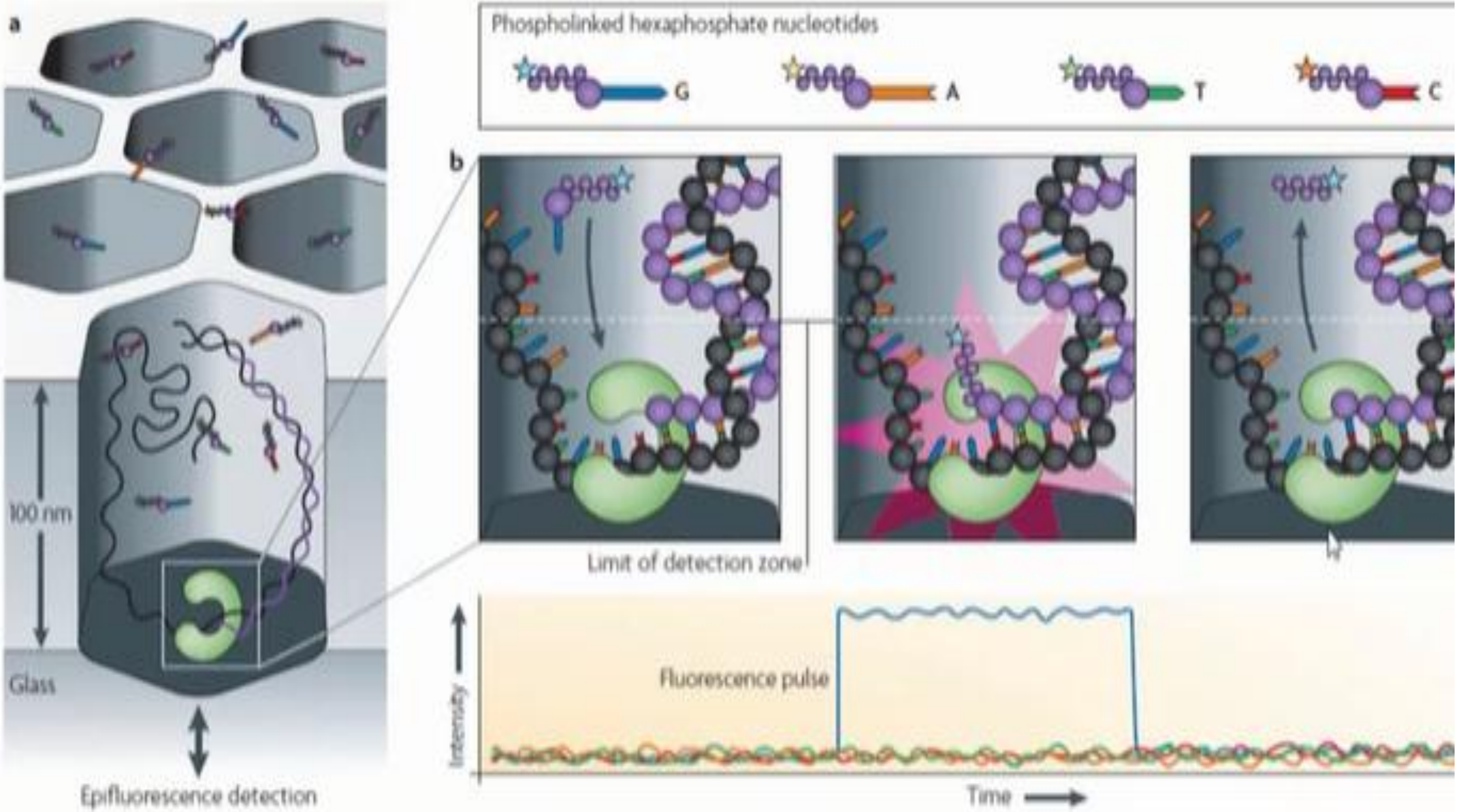


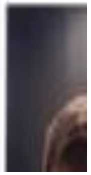
Pacific Biosystems RS
3rd Generation Single Molecule Sequencer

Pacific Biosciences RS
Immobilized Polymerases + fluorescent dNTP +
really, really good optics

The particular polymerase used ("displacing") is on the order of bp/sec

Detection of base incorporation





Extension results in different fluorescent signal for each base

NGS technologies - SUMMARY

Platform	Read length	Accuracy	Projects / applications
454	Medium	Homopolymer 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

Thank You