Next Generation Sequencing Methods

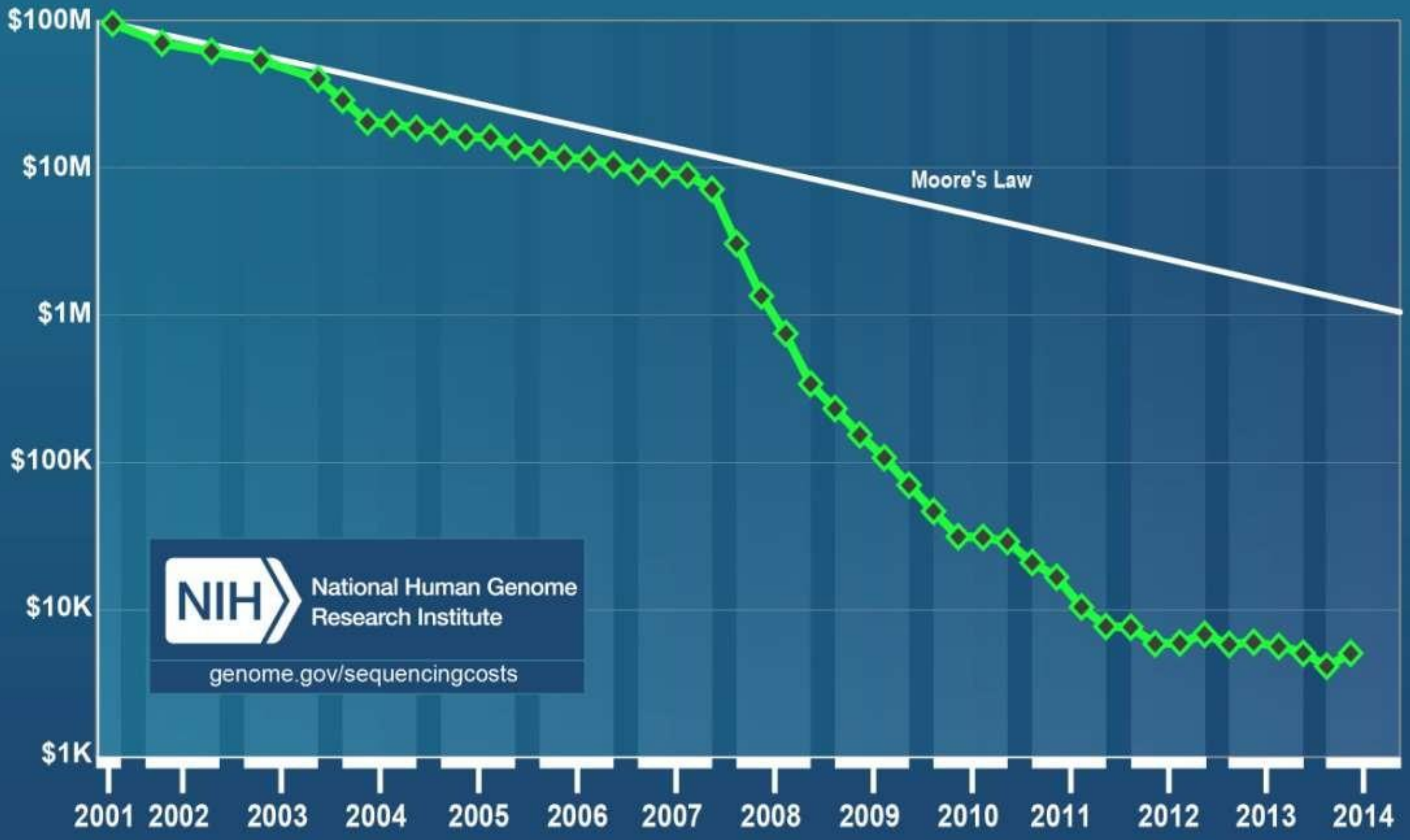
Professor

Dr. Mushtak T.S.Al-Ouqaili

Sequencing

- **Sequencing** is the process of determining the precise order of nucleotides within a **DNA, RNA** molecule.
- In case of **proteins**, amino acids
- It includes any method or technology that is used to determine the order of the four bases—adenine, guanine, cytosine, and thymine—in a strand of **DNA**.
- **Next-generation sequencing (NGS** or **high-throughput sequencing** are collectively technologies developed by:
  - Illumina (Solexa) sequencing
  - Roche 454 sequencing
  - Ion torrent: Proton / PGM sequencing
  - SOLiD sequencing (Thermo Fisher Scientific)

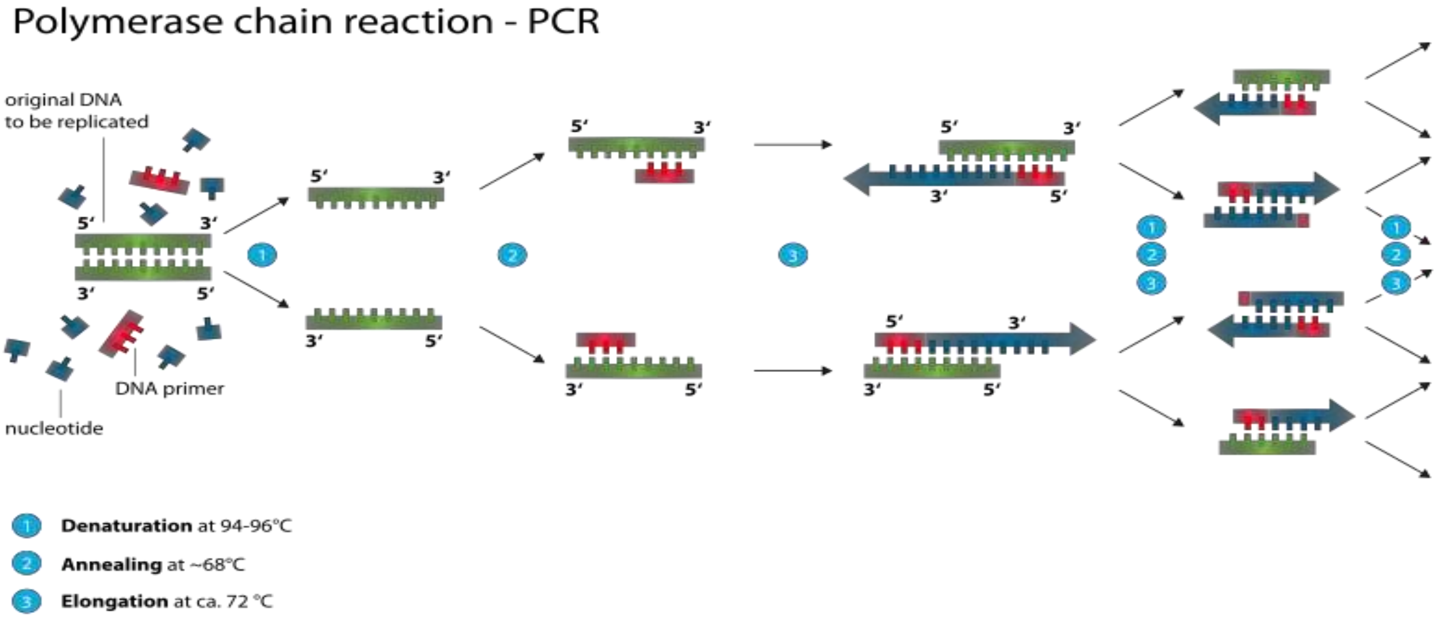
Cost per Genome



NIH National Human Genome Research Institute

genome.gov/sequencingcosts

qPCR



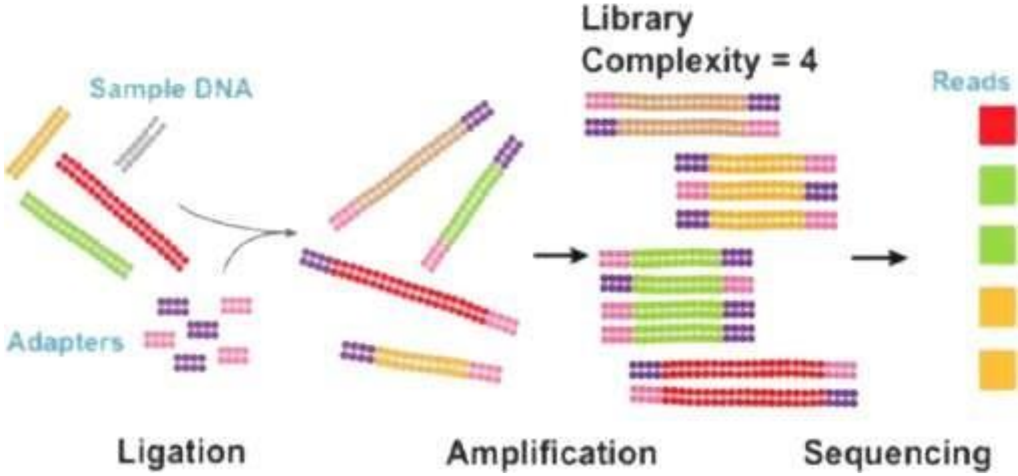
Quantitative PCR measures the level of a particular •
nucleic acid (usually mRNA) defined by a primer
pair
sequence that amplifies it

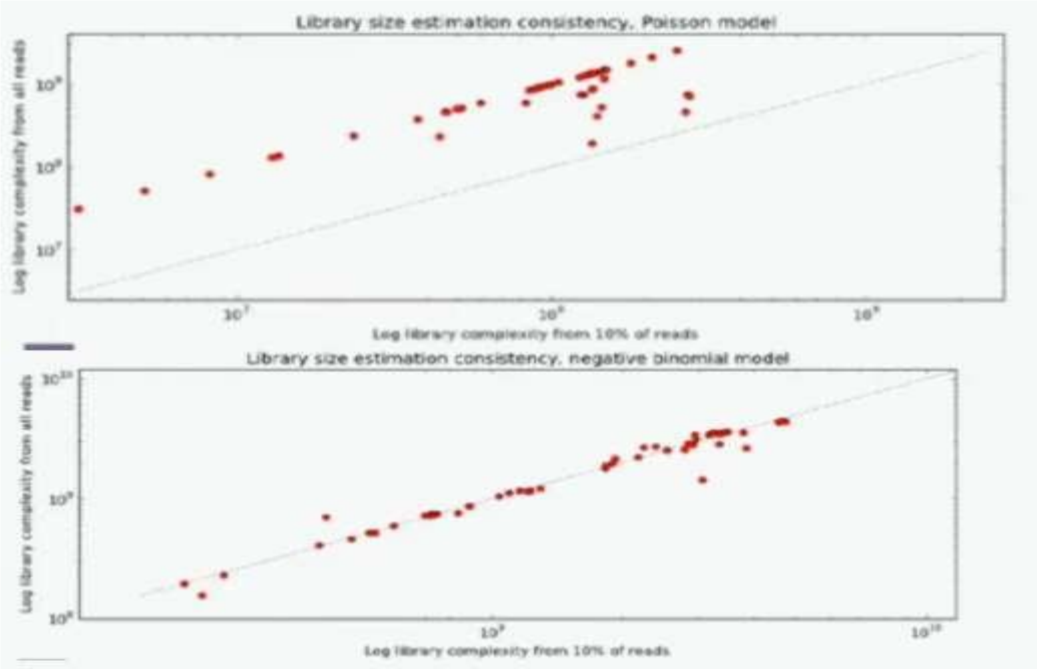
- The mRNA is converted into DNA by reverse transcriptase (a retroviral enzyme)
- Makes use of a high temperature stable polymerase i.e. Taq polymerase
- It is a non-leniar amplification
- **Selective bias of molecules can register less library complexity**

- Quantitaiveness by: measuring the no. of cycles needed to reach a fluorescence intensity threshold
- However, larger no. of cycles = lower original amount of DNA i.e. less DNA available for experimentation purposes
- It is a **medium-throughput** approach and is usually limited to only being of *validative use*

Library Complexity

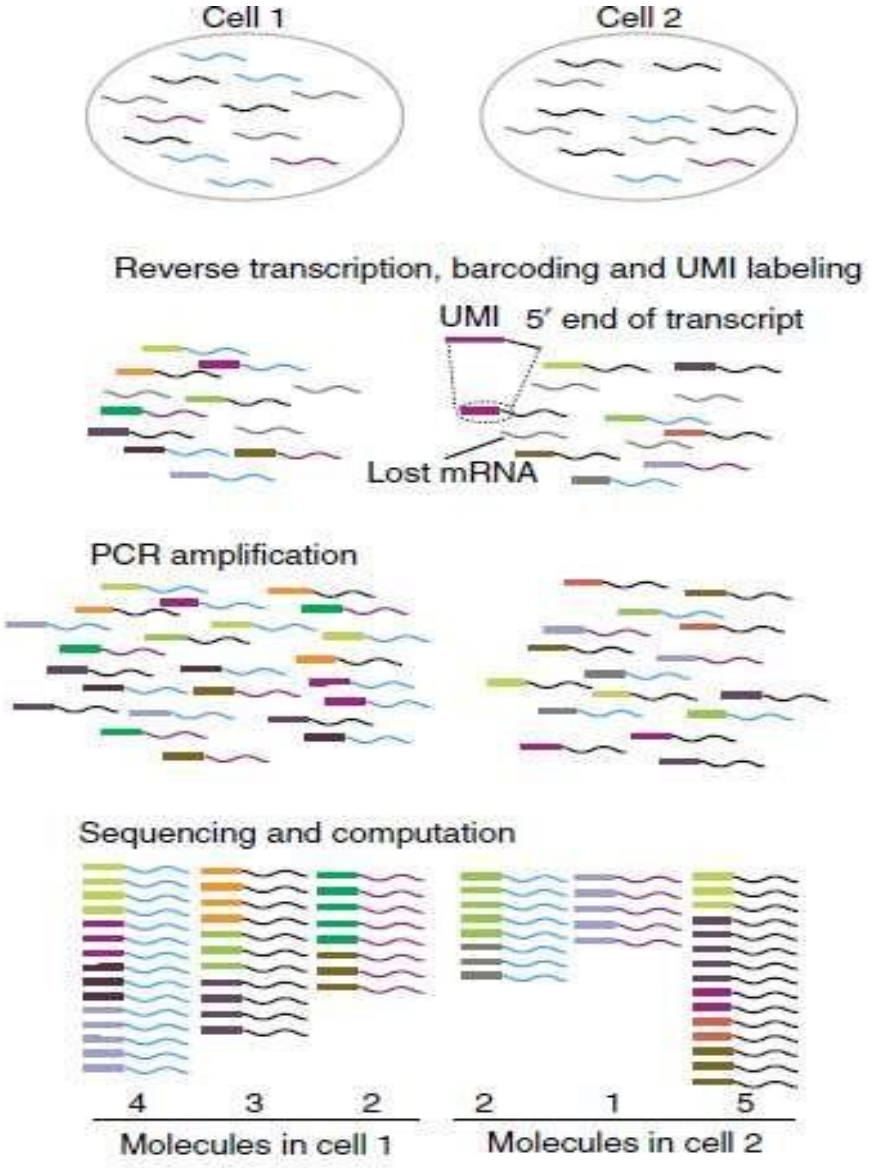
- The no. of unique molecules in the “library” that is sampled by finite sequencing constitutes library complexity.
- Simple representation methods (such as Poisson's distribution) can be wrong in representing the complexity
- Poisson’s dist. does not account for over dispersion, hence Poisson’s sampling is only effective for smaller population sizes
- Negative binomial distribution (Poisson-Gamma dist.) is one method of correctly estimating the library complexity





How to register Lib. Complexity

- Attach a set of 5 nucleotide barcodes (1,024 possible) to the 5’ end of (nearly) every mRNA
- Sequence the 5’ ends and count the number of unique barcodes that appear
- This is enough to ensure unique alignment
- Obtained is the number of mRNA molecules in the original sample



$$Poisson(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$Gamma(x;\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$$

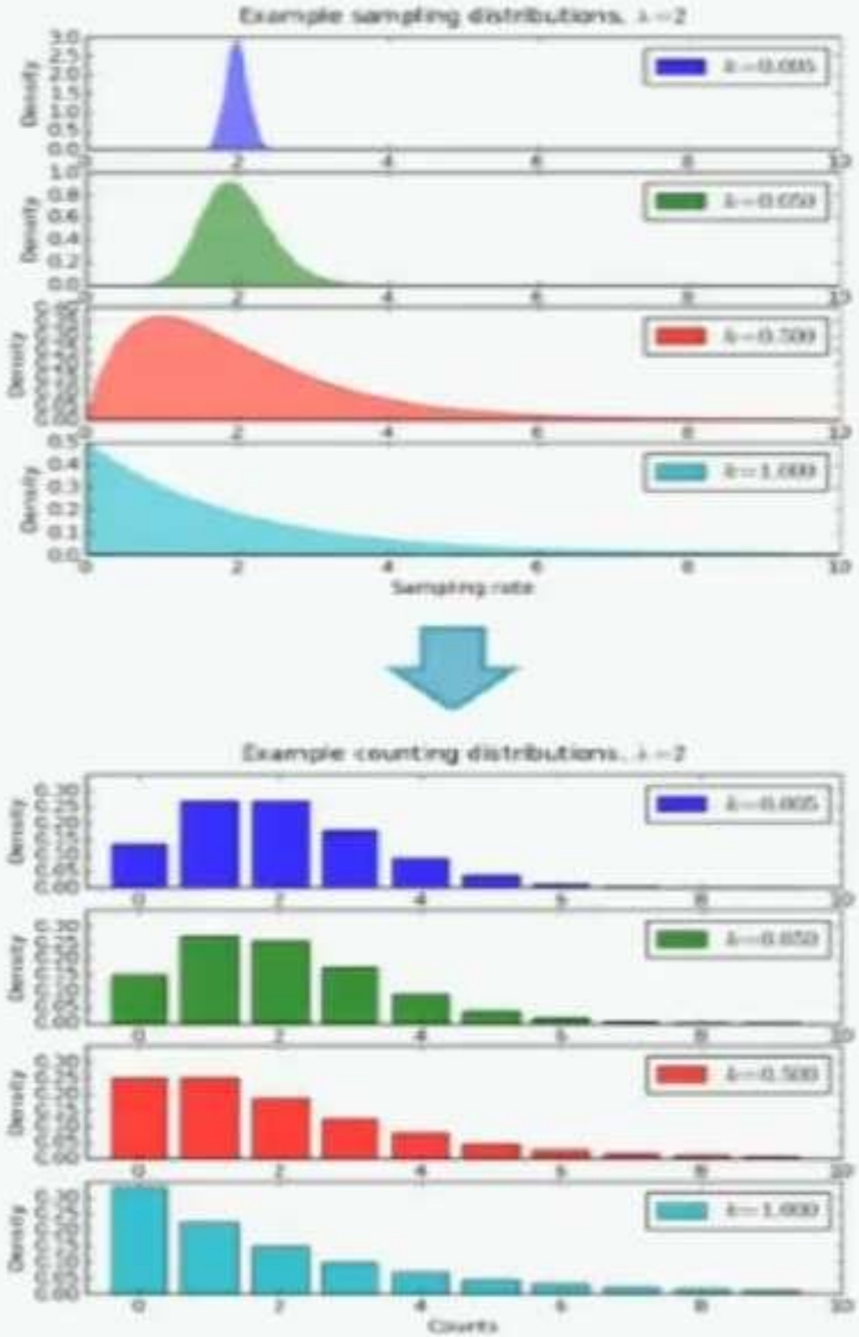
$$NB(y;\alpha,\beta) = \int_0^\infty Poisson(y;x)Gamma(x;\alpha,\beta)dx$$

- Gamma sampling rates describe the entire population library (complexity)
- In high-throughput when reads are as large as hundreds of millions, it becomes useful to ask:

1.) How complex is the original library؟

2.) Is the data really good?

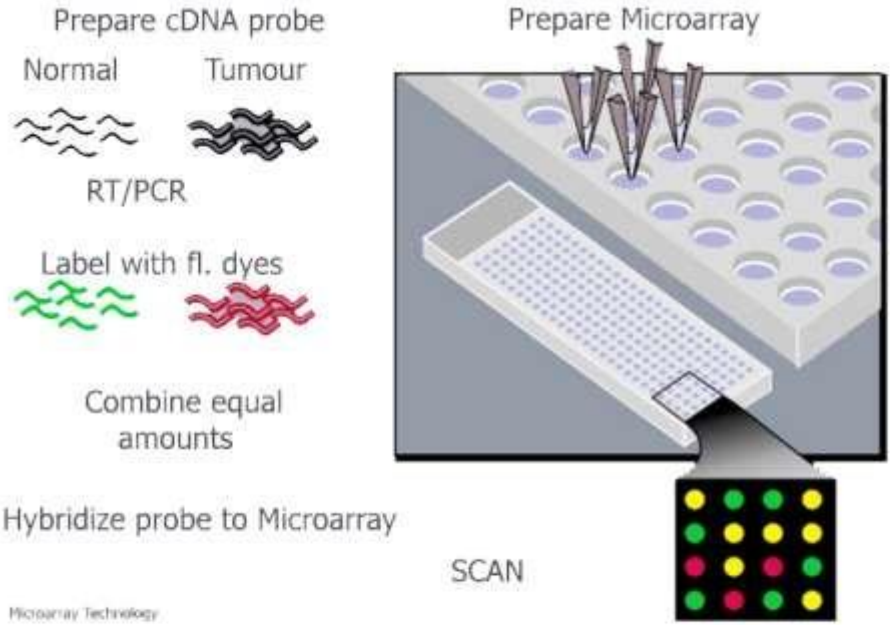
- If the observed complexity matches with the theoretical complexity , then it solves the dilemma whether further sequencing should be done to capture entire complexity



Microarrays

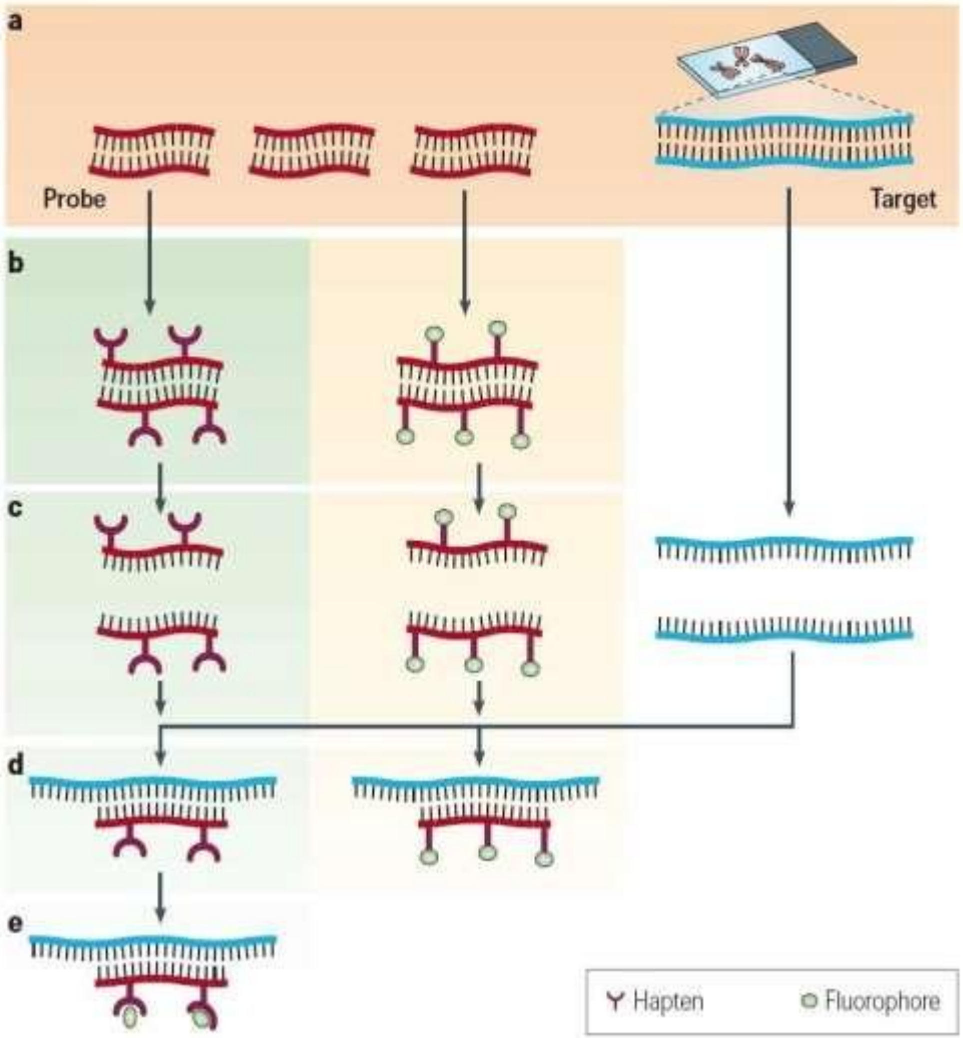
- These were one of the first *omic* methods
- Used for measuring: transcript levels, genotyping, DNA mapping *viz.* DNA copy no., methylome
- Principle: On microarray chip there are spots – each spot has a diff. oligonucleotide – sample binds to complimentary ones and generates a fluorescent signal – we photograph it – analyze the photograpth – get results
- It is primarily used for Transcriptome and registers lib.
- complexity is very well

Instead of chips/slides, beads are also used (as we will see in ILLUMINA seq tech.



FISH

- Fluorescence in situ hybridization uses labeled oligos to hybridize to the target nucleic acids
- It is shown to be done for single molecules (Raj et al, Nat. Met., 5(10), 2008)
- It can also be highly multiplexed with super-resolution and fluropore barcoding (Lubec and cori, Nat. Met. 9(7), 2012)



General rules of thumb

- Few different molecules (RNA): **Northern Blotting** or **FISH**
- Medium throughput: **qPCR**
- High throughput: **Microarrays, Next-Gen Sequencing**

Need of sequencing

- Sequencing such as that of mRNA allows for:

(a.) Quantification of expressed transcriptome (transcript library)

b.) The ratios of splice variant levels i.e. for a gene that differentially expresses two splice isoforms

c.) Identification of novel splice variants

- Seq Techniques for Nucleic acids is broadly divided into two types:

(1.) **Population average**: Large amount of starting material

required (~1 M cells *e.g.* **qPCR, microarrays, Deep Sequencing Technologies** (viz. Whole)

genome, Exome seq, Transcriptome, Bisulphate seq, ChIP seq) *etc.*

(2.) **Single cell:** Detecting analytes from single cell.

-> Main challenge- separating measurement noise

e.g. **qPCR, FISH, RNA seq** *etc.*

Sequencing Based Methods

- Although high-throughput microarrays have certain limitations viz.

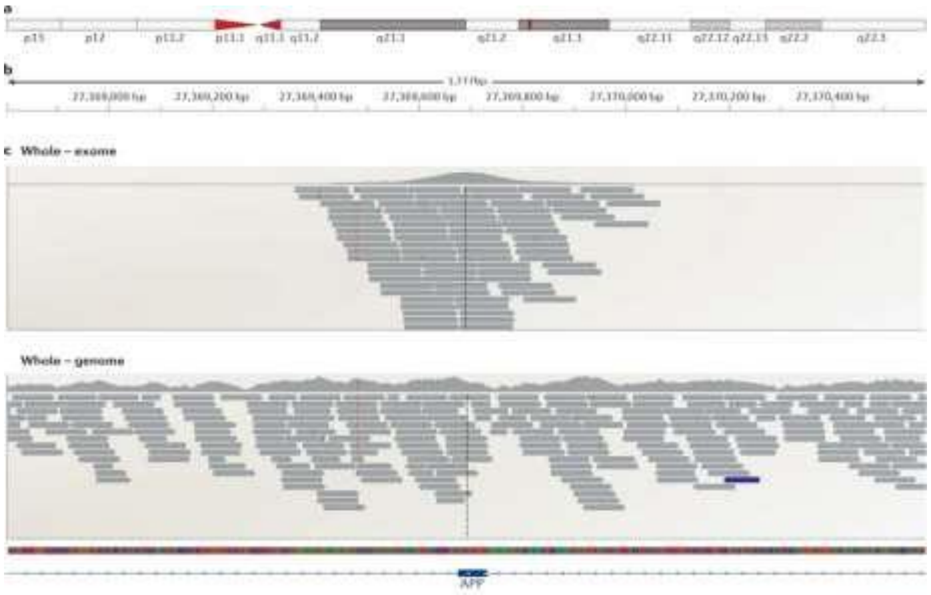
(a.) High background noise

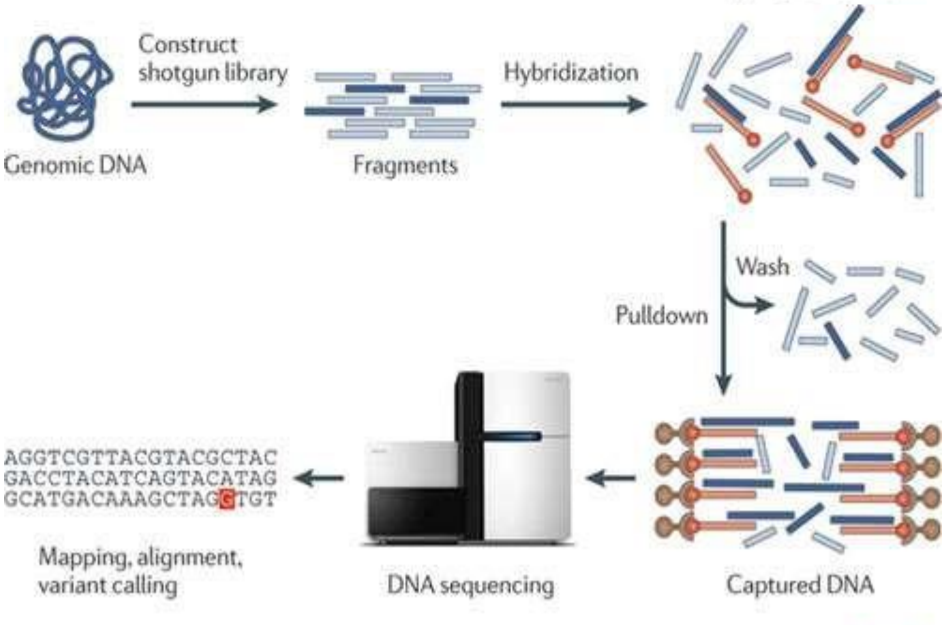
(b.) Need of large starting material

(c.) Microarray is hard to compare across different techs.

(c.) **Limited ability to distinguish isoforms and allelic expression**

- These are overcome by Next Gen RNA sequencers
- Where depth of sequencing or the average no. of short reads per base pair is a key parameter in seq.
- Exome seq give more info. for less short reads
- Apart from the general protocol–

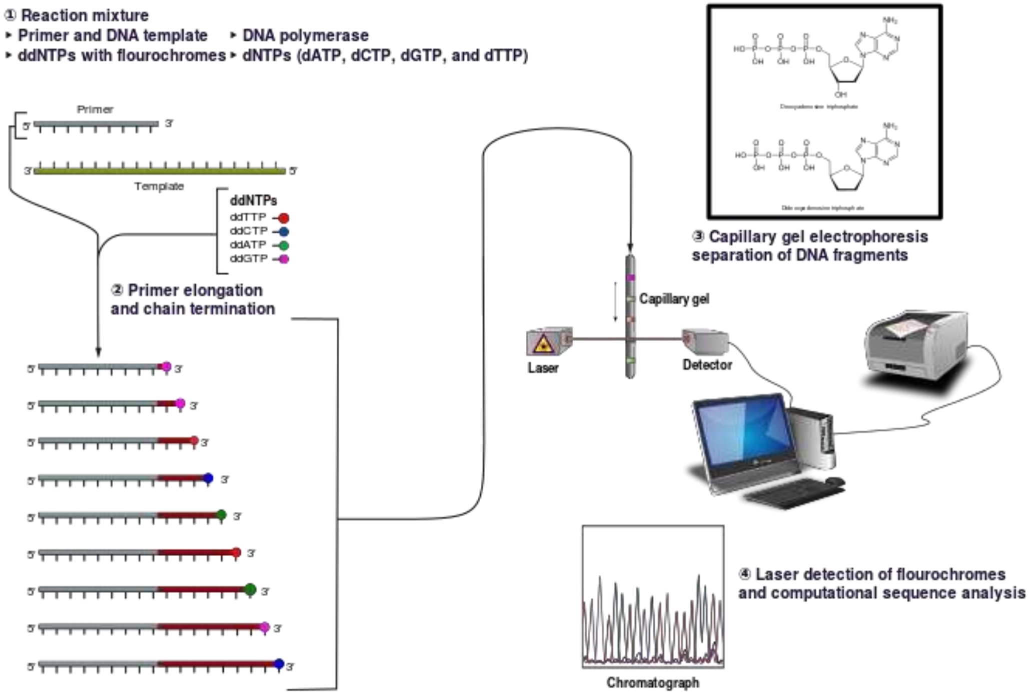




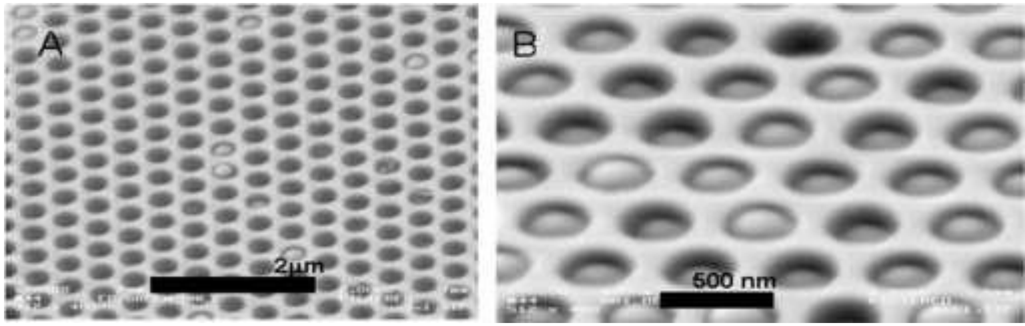
Nature Reviews | Genetics

Generations of RNA sequencers

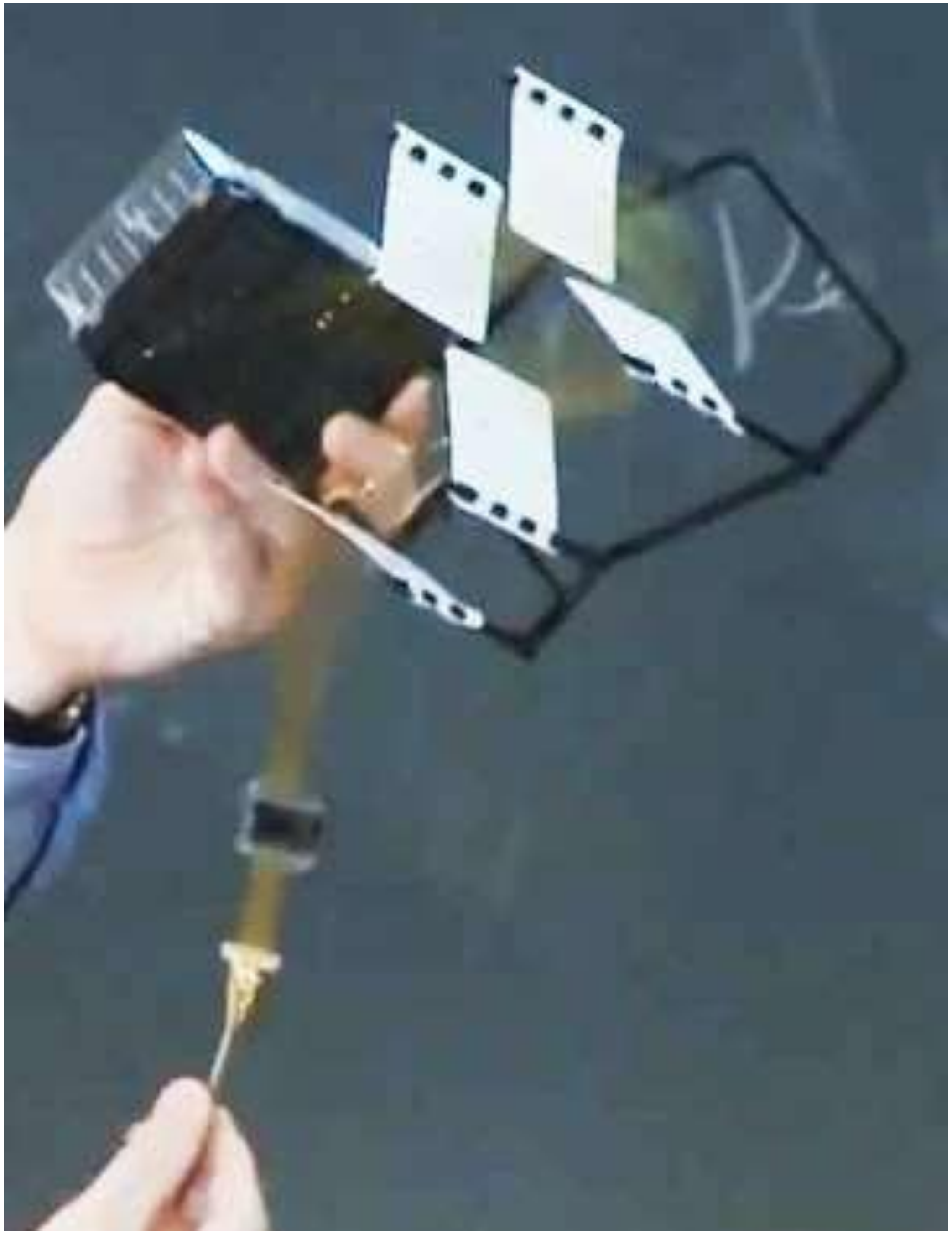
- **1st gen/ Sanger Sequencing**
  - These were primer based methods which used ddNTPs, fluorescent markers and enzymes
  - Low throughput, ~700 bp read length
  - Very slow and expensive but highly accurate
  - Parallel seq was used to increase sequencing speed – 96 and 384 well formats



① Reaction mixture
‣ Primer and DNA template
‣ DNA polymerase
‣ ddNTPs with flourochromes
‣ dNTPs (dATP, dCTP, dGTP, and dTTP)
Primer
Template
5'
3'
ddNTPs
ddTTP
ddCTP
ddATP
ddGTP
② Primer elongation and chain termination
③ Capillary gel electrophoresis separation of DNA fragments
Capillary gel
Laser
Detector
④ Laser detection of flourochromes and computational sequence analysis
Chromatograph



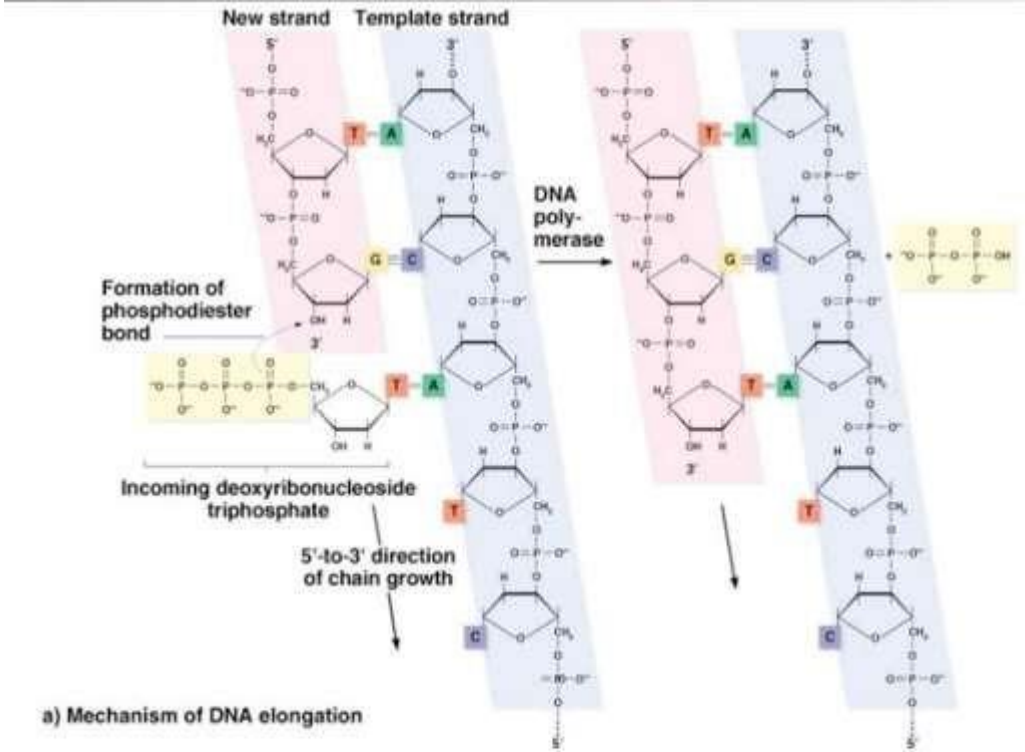
- Fluorescent seq combined with capillary electrophoresis developed automation
- Efficiency? It took 13 years and 2.7 B $ to seq human genome
- Major Drawback: During chain termination it was hard to separate DNA molecules differing in 1 nt BP



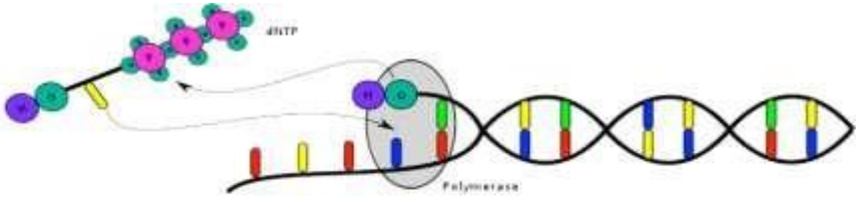
- **<u>2nd gen sequencing</u>**
  - DNA strand sep. problem was eliminated by 2nd gen. seq.
  - Parallel identification of nts. during synthesis
  - A fundamentally different approach

Limitations:

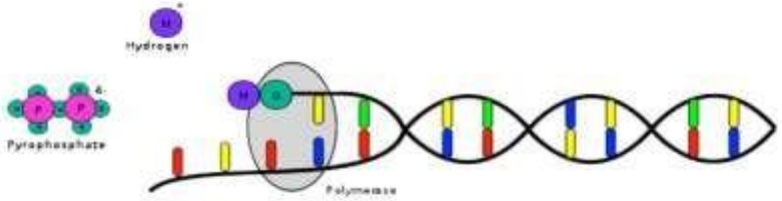
- These require amplification of DNA to meet detection threshold
- **Amplification bias**
- Practical limits in read lengths


New strand
Template strand
5'
3'
DNA poly-merase
Formation of phosphodiester bond
Incoming deoxyribonucleoside triphosphate
5'-to-3' direction of chain growth
T
A
G
C

a) Mechanism of DNA elongation

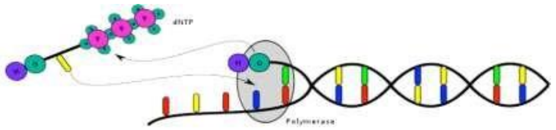


Polymerase integrates a nucleotide.

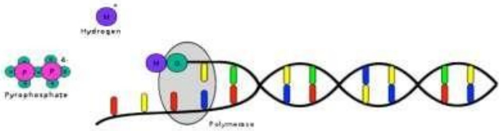


Hydrogen and pyrophosphate are released.

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

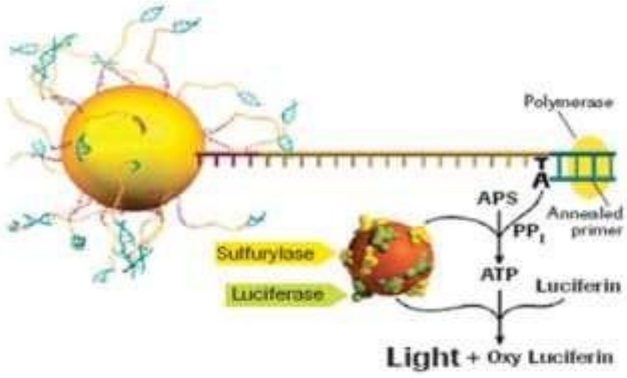


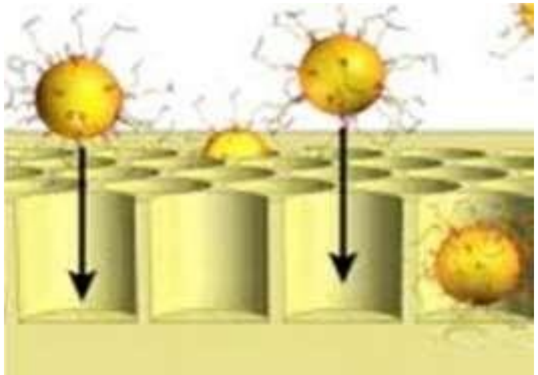
Polymerase integrates a nucleotide.

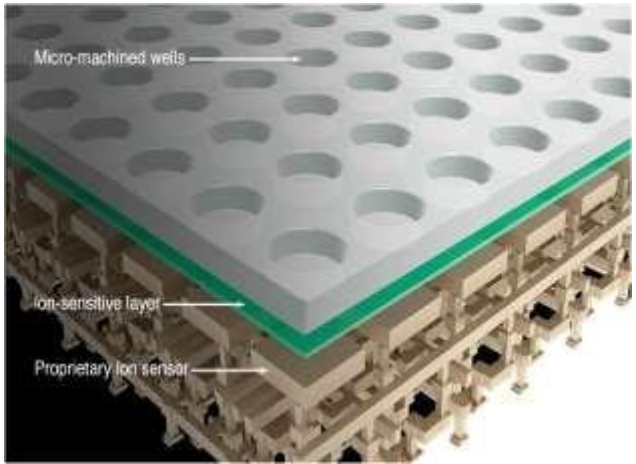


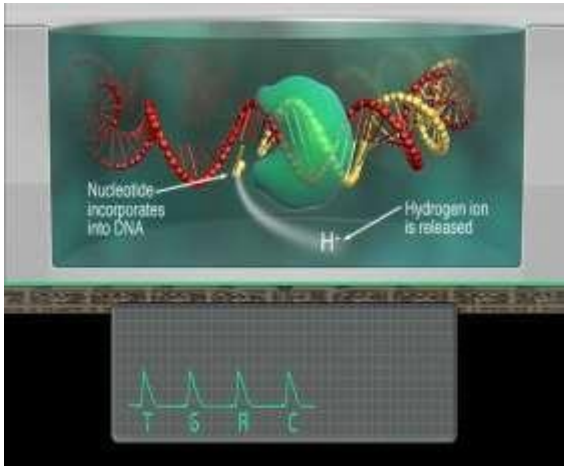
Hydrogen and pyrophosphate are released.

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



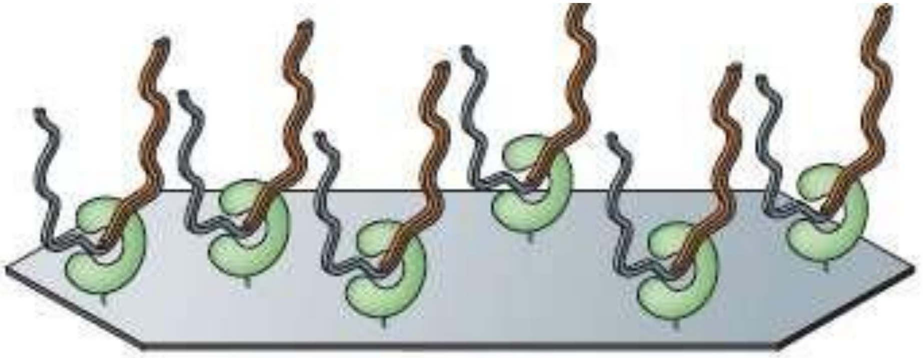




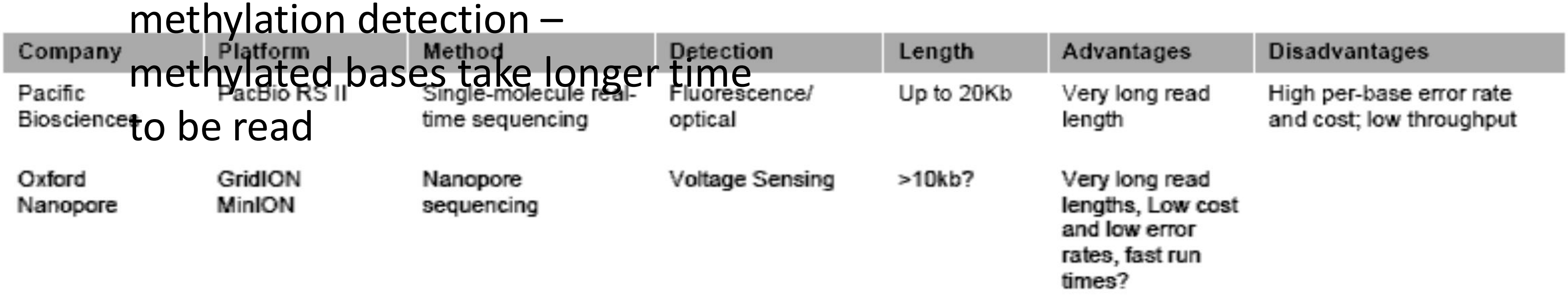


- **3rd gen sequencing**
  - Longer read length without the need of amplification
  - Involves immobilized polymerase + fluorescent DNTPs + highly sensitive optometry
  - DNTPs have 6 Ps instead of 3, thus a longer fluorescence pulse is generated
  - Color of fluorescence is compared to give results
  - It is quite noisy and has a high error rate
  - One powerful aspect is methylation detection – methylated bases take longer time to be read

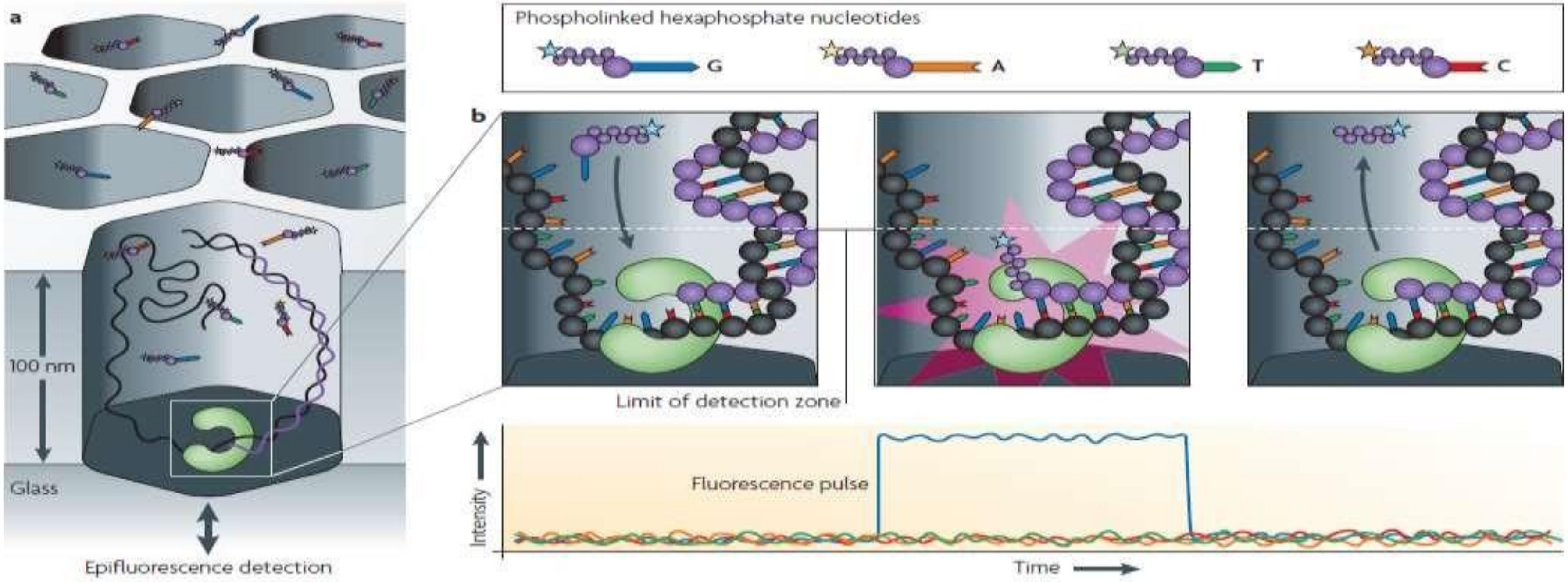
Pacific Biosystems RS
3rd Generation Single Molecule Sequencer

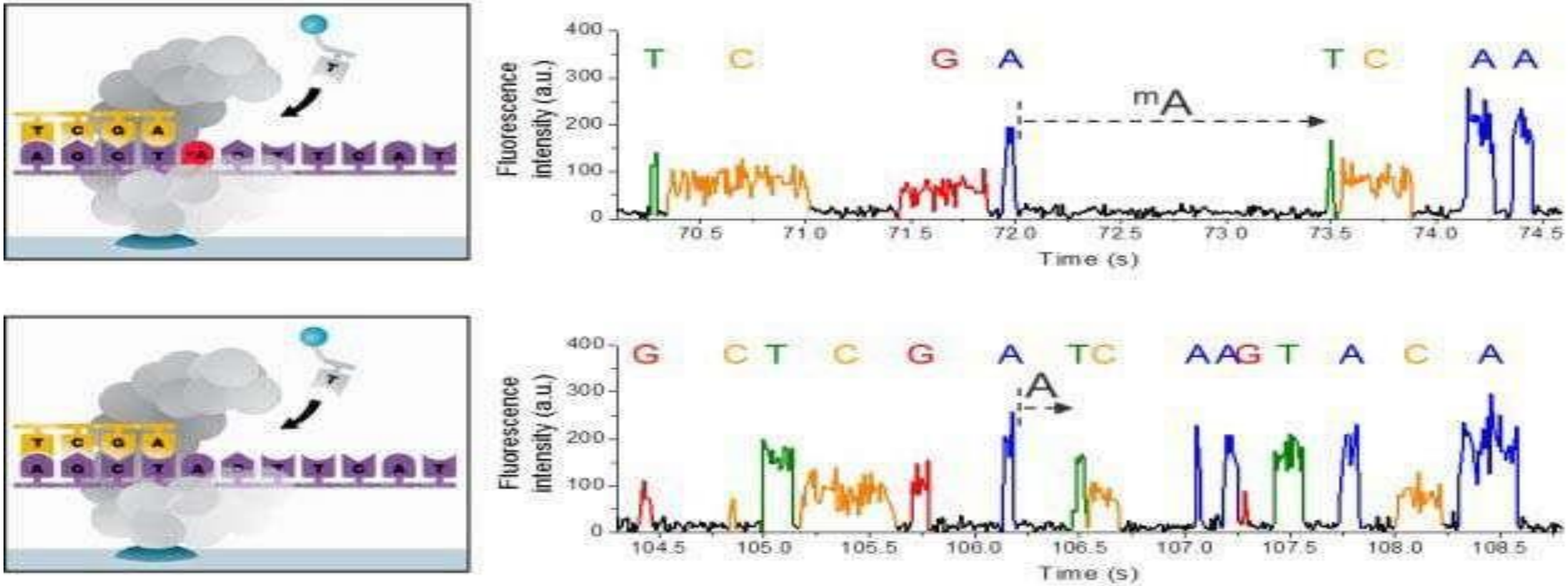


Thousands of primed, single-molecule templates



Company	Platform	Method	Detection	Length	Advantages	Disadvantages
Pacific Biosciences	PacBio RS II	Single-molecule real-time sequencing	Fluorescence/ optical	Up to 20Kb	Very long read length	High per-base error rate and cost; low throughput
Oxford Nanopore	GridION MinION	Nanopore sequencing	Voltage Sensing	>10kb?	Very long read lengths, Low cost and low error rates, fast run times?	


a
Phospholinked hexaphosphate nucleotides
G
A
T
C
b
100 nm
Glass
Epifluorescence detection
Limit of detection zone
Fluorescence pulse
Intensity
Time


T C G A
A G C T A T T C A T
T
T C G A
A G C T A T T C A T
T
T C G A TC AA
mA
Fluorescence intensity (a.u.)
400
300
200
100
0
70.5
71.0
71.5
72.0
72.5
73.0
73.5
74.0
74.5
Time (s)
G C T C G A TC AAG T A C A
A
Fluorescence intensity (a.u.)
400
300
200
100
0
104.5
105.0
105.5
106.0
106.5
107.0
107.5
108.0
108.5
Time (s)

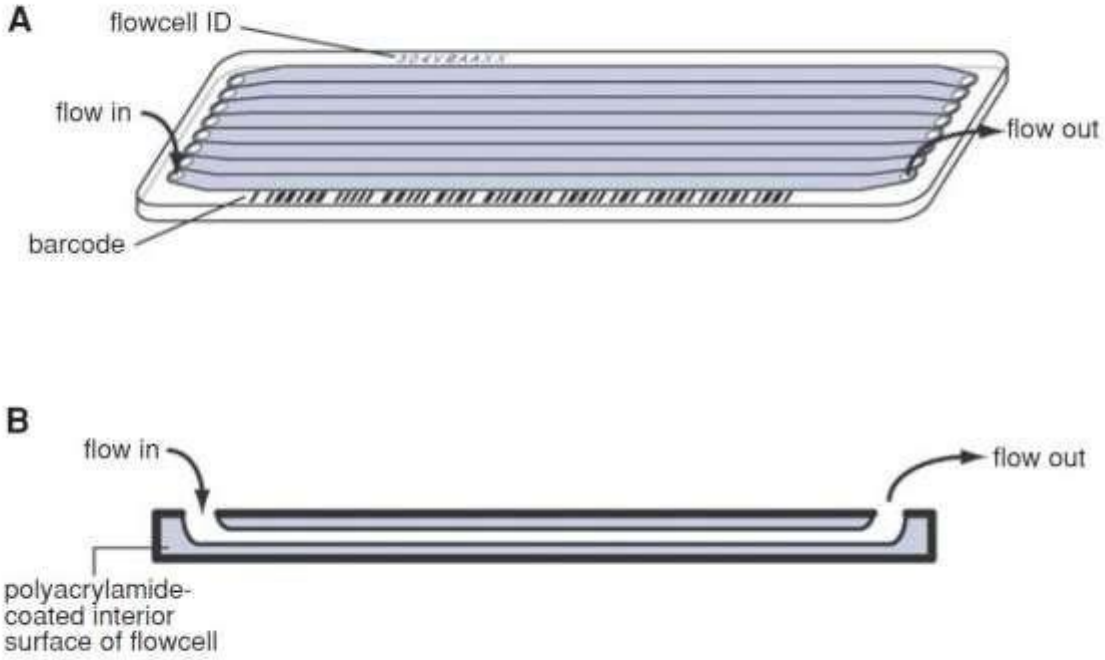
- ILLUMINA Protocol (for 3rd gen se machine)
  - Two components: (a.) RNA seq library preparation (Mol Bio. Component& b.) Actual seq

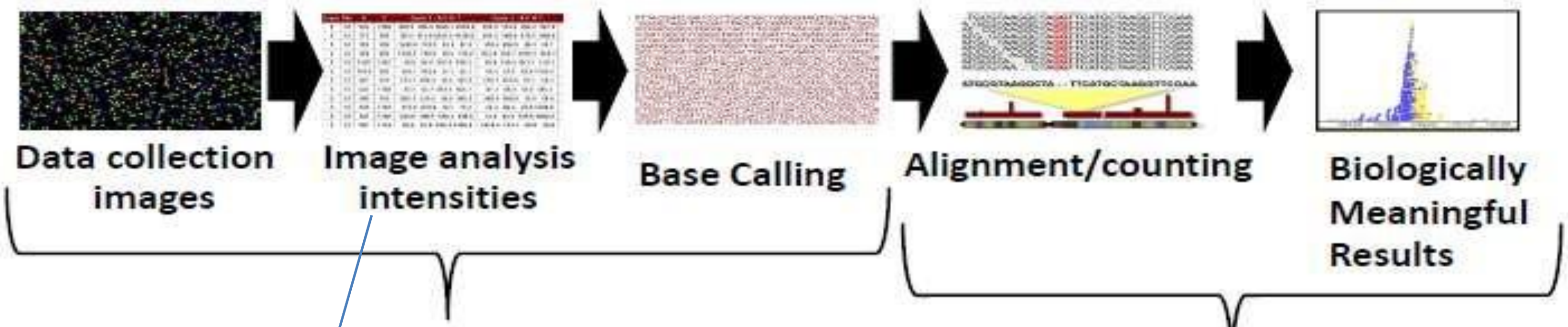
    and Data analysis
  - a.) We use Bioanalyzer chip - a high-throughput electrophoresis on a chip for

    QC of the mRNA sample

    Takes up to 45 min.
  - Next is adapter ligation– automated – takes up to 5 hrs.
  - Preparation of cDNA libraries in PCR – hr.
  - Reanalysis of QC on a DNA chip

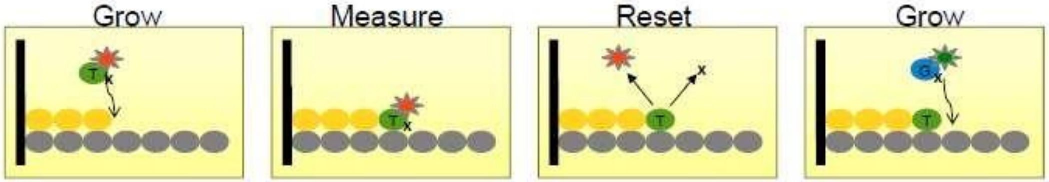
- Libraries are loaded at flow in (tunnel / lane) with a separate automated machine (5 hrs.)
- Each lane – 20 B of the seq data nts.

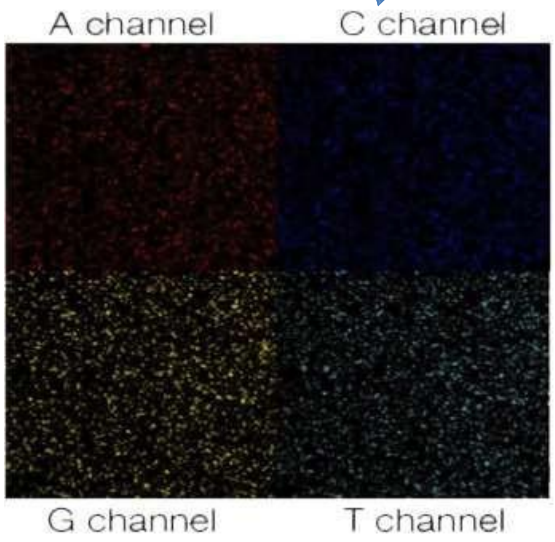


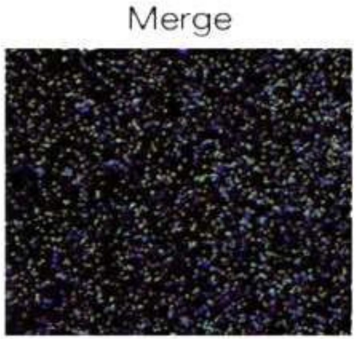


Usually done with platform-specific software

Usually done with open source academic software: e.g. tophat/cufflinks/cuffdiff

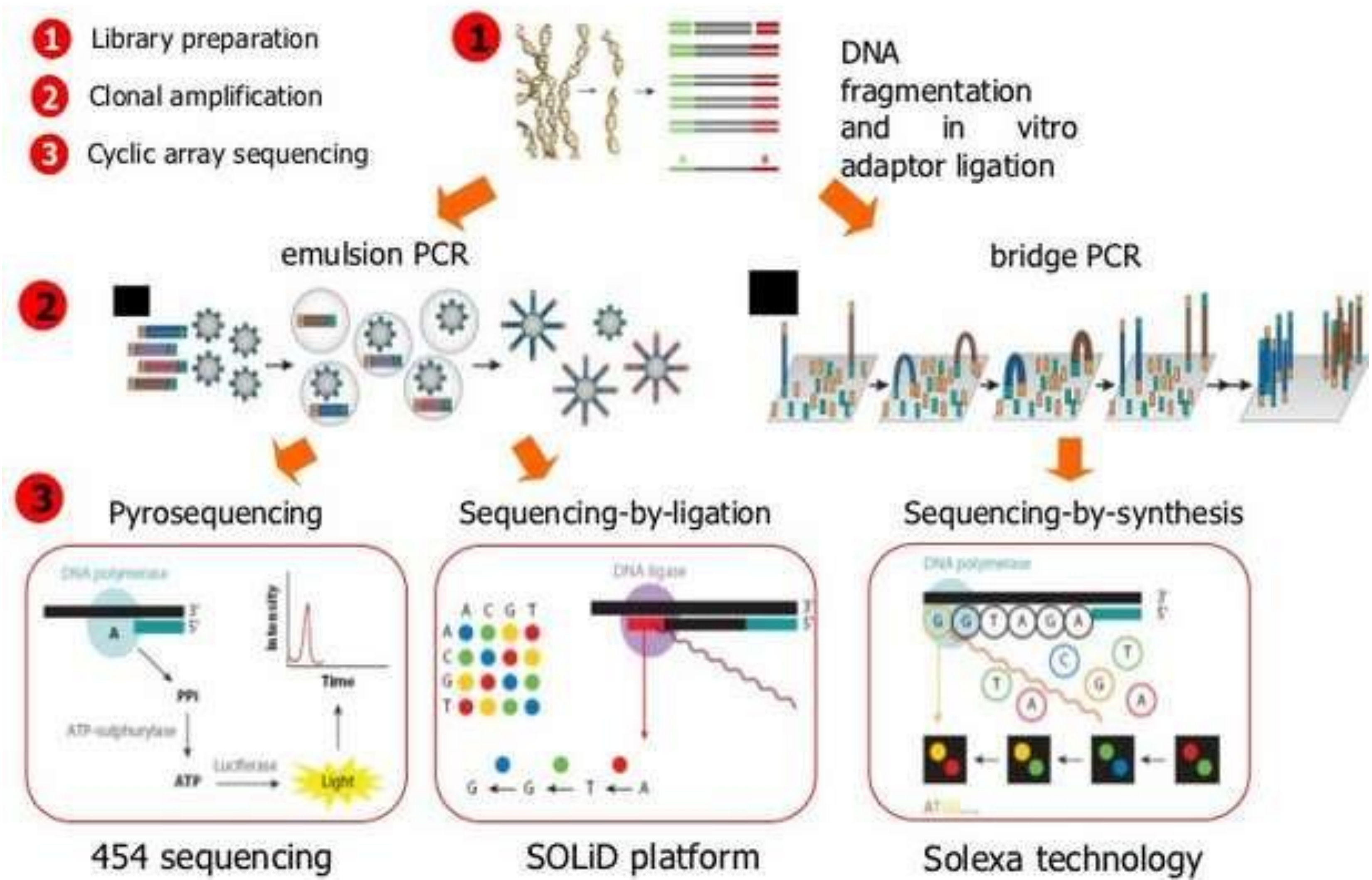




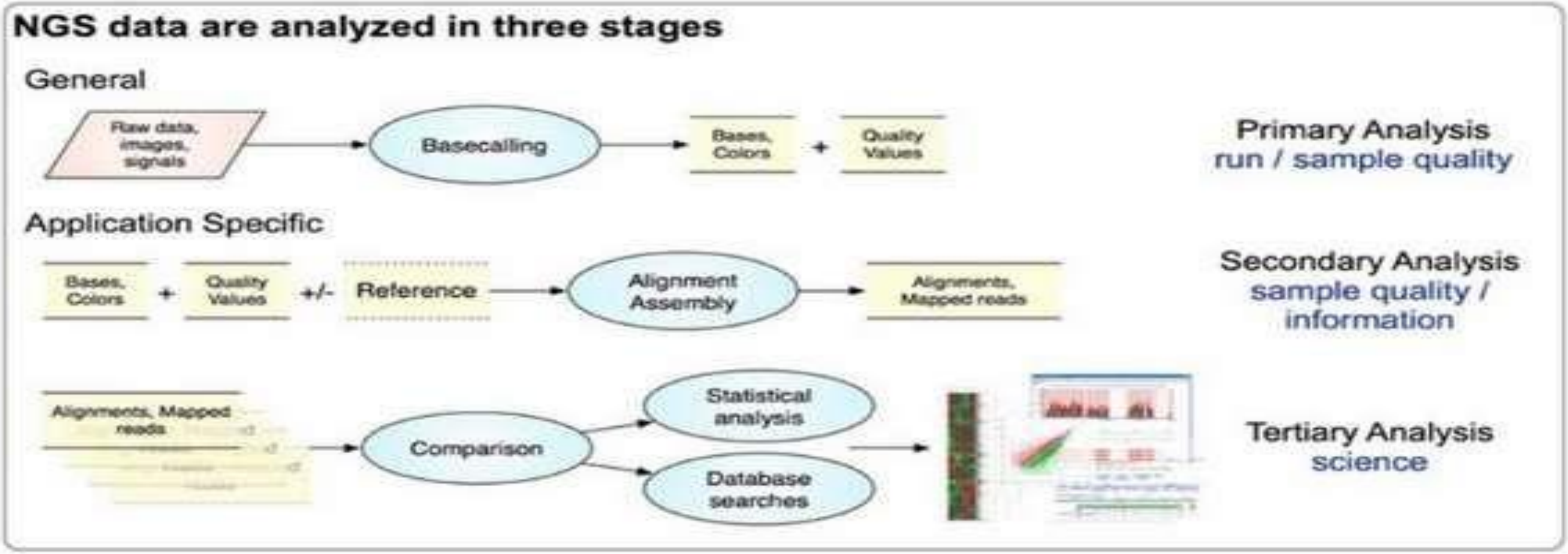


1/4 of one tile (0.03% of a flow cell, GA2)

Comparison of ILLUMINA protocol with other technologies



What next: The fate of seq data



NGS data are analyzed in three stages

General

Raw data, images, signals → Basecalling → Bases, Colors + Quality Values

Primary Analysis
run / sample quality

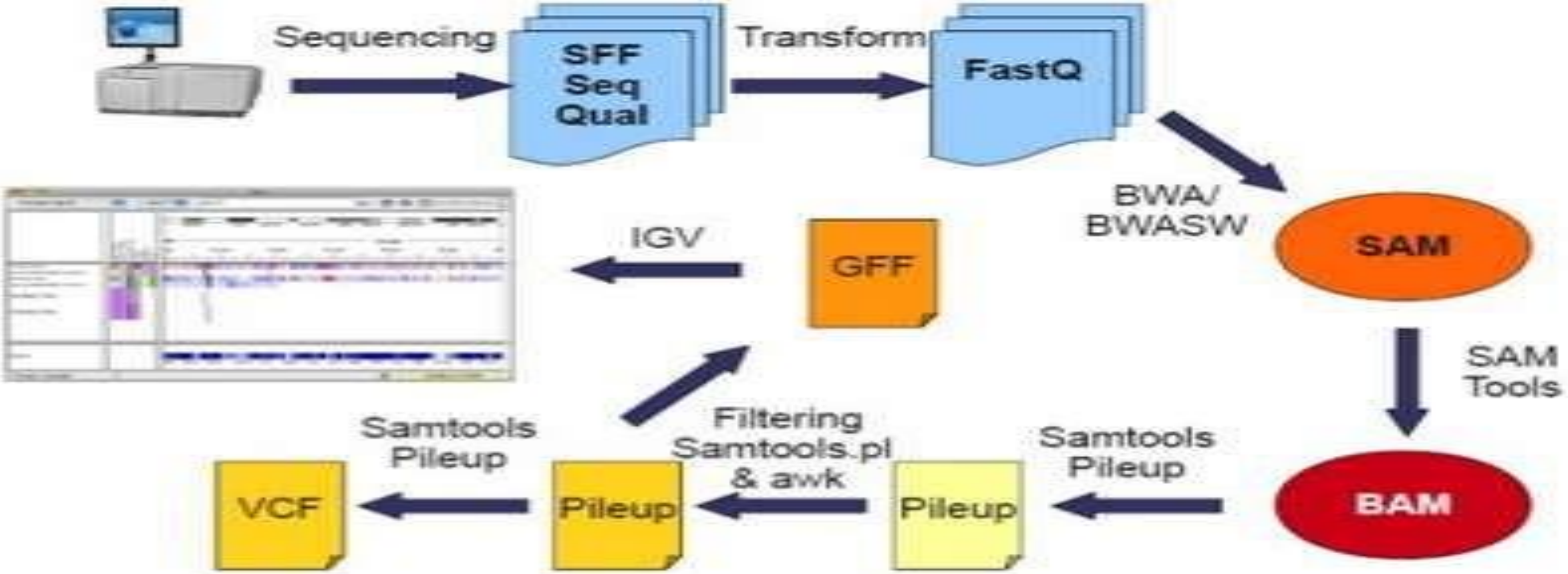
Application Specific

Bases, Colors + Quality Values +/- Reference → Alignment Assembly → Alignments, Mapped reads

Secondary Analysis
sample quality / information

Alignments, Mapped reads → Comparison → Statistical analysis, Database searches →

Tertiary Analysis
science

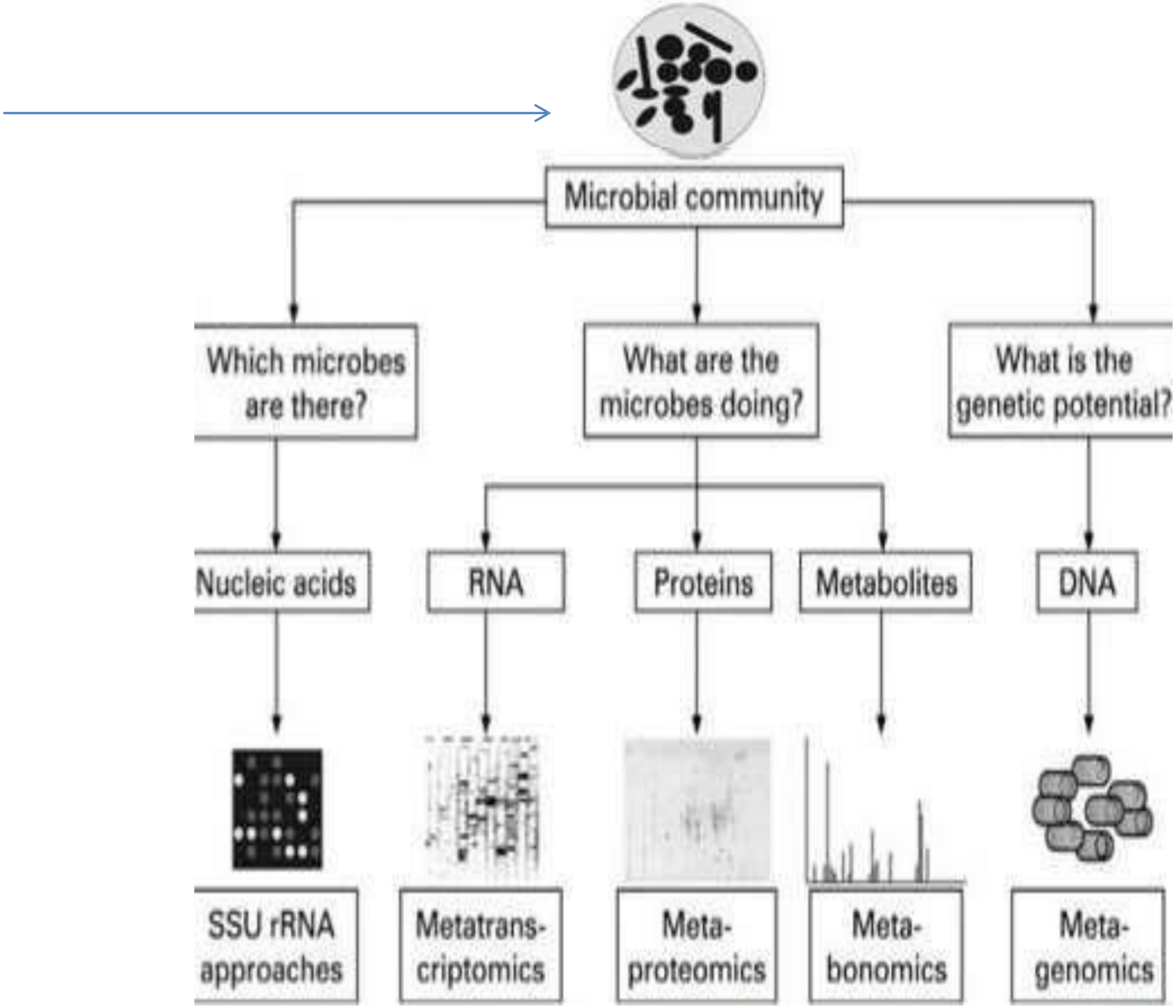


Sequencing → SFF Seq Qual → Transform → FastQ → BWA/ BWASW → SAM → SAM Tools → BAM → Samtools Pileup → Pileup → Filtering Samtools.pl & awk → Pileup → Samtools Pileup → VCF

Pileup → GFF → IGV

Fruits of sequencing

- Answering key questions such as – are certain SNIPs associated with diseases? What critical mutations are there which caused the disease?
- However, data must be quantitative and sampling population should be large to make any such assessment *e.g.* mutation frequencies



Microbial community

Which microbes are there?

What are the microbes doing?

What is the genetic potential?

Nucleic acids

RNA

Proteins

Metabolites

DNA

SSU rRNA approaches

Metatranscriptomics

Metaproteomics

Metabonomics

Metagenomics

Zoetendal E G et al.
Gut 2008;57:1605-1615

Conclusion

- These technologies allows for sequencing of DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing, and as such have revolutionised the study of genomics and molecular biology

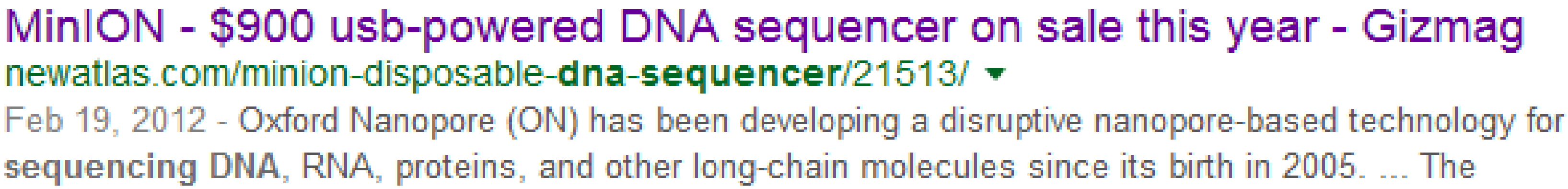
The Latest

- MinION has the potential to revolutionize the field of sequencing completed

- It was field tested recently for seq. EBOLA virus

- A threat to the decade long ILLUMINA dominancy over the market

- Very cheap!



MinION - $900 usb-powered DNA sequencer on sale this year - Gizmag

newatlas.com/minion-disposable-**dna-sequencer**/21513/ ▾

Feb 19, 2012 - Oxford Nanopore (ON) has been developing a disruptive nanopore-based technology for **sequencing** DNA, RNA, proteins, and other long-chain molecules since its birth in 2005. ... The

THANK YOU FOR YOUR PRECIOUS TIME!