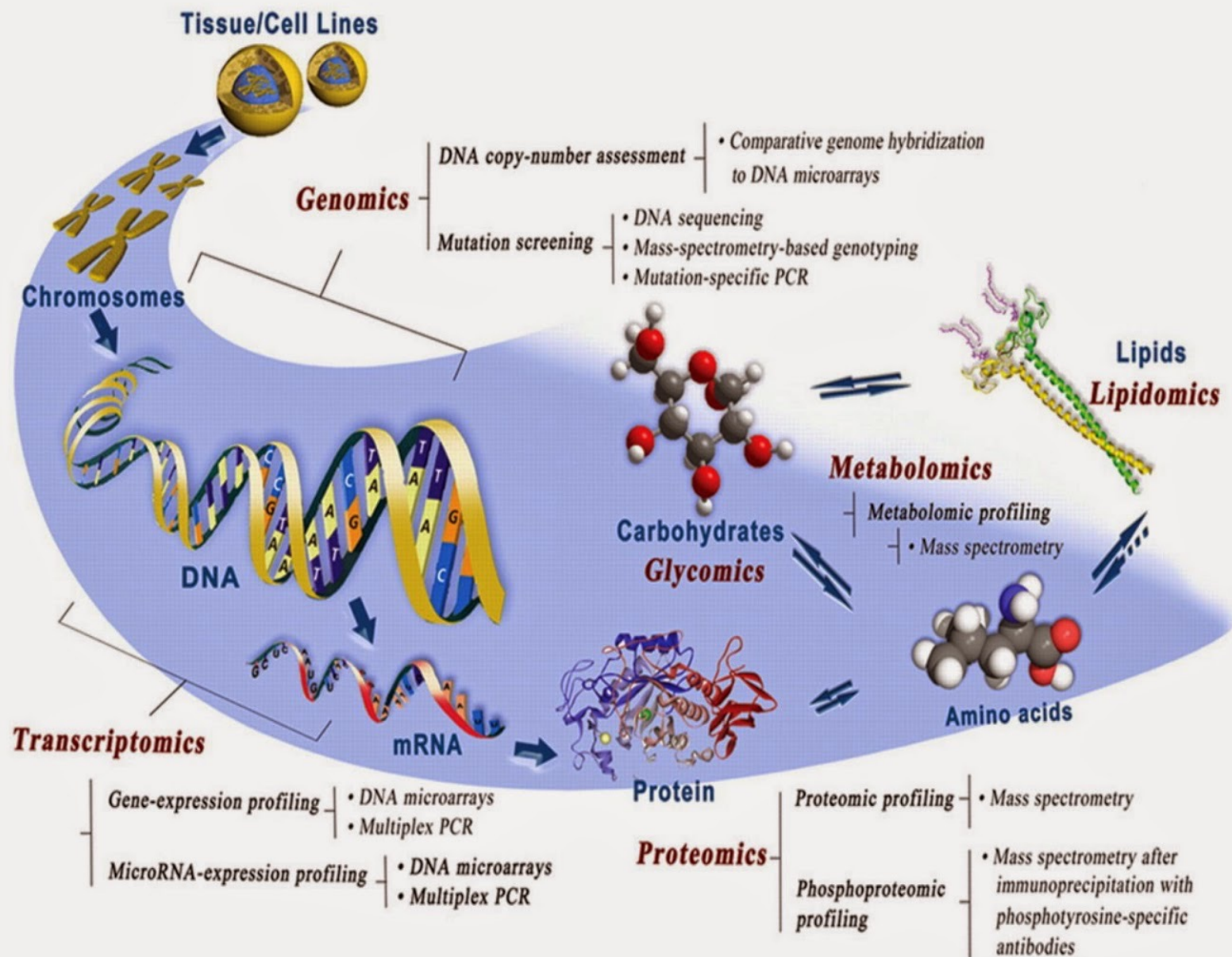




Introduction to NGS and sequencing files

Paúl Cárdenas, MD, PhD



¿Por qué han llegado a ser importantes?

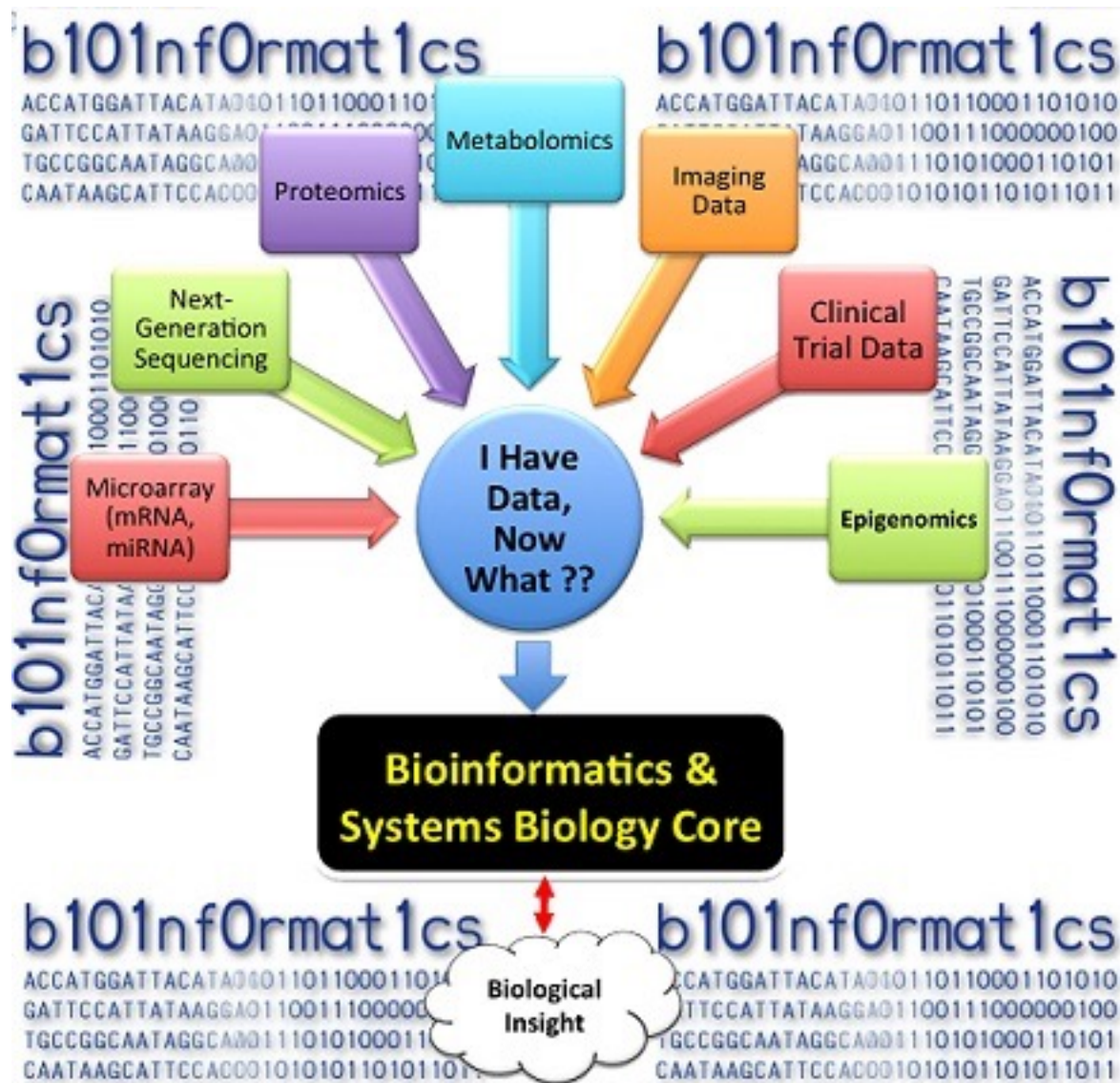


Platform Features



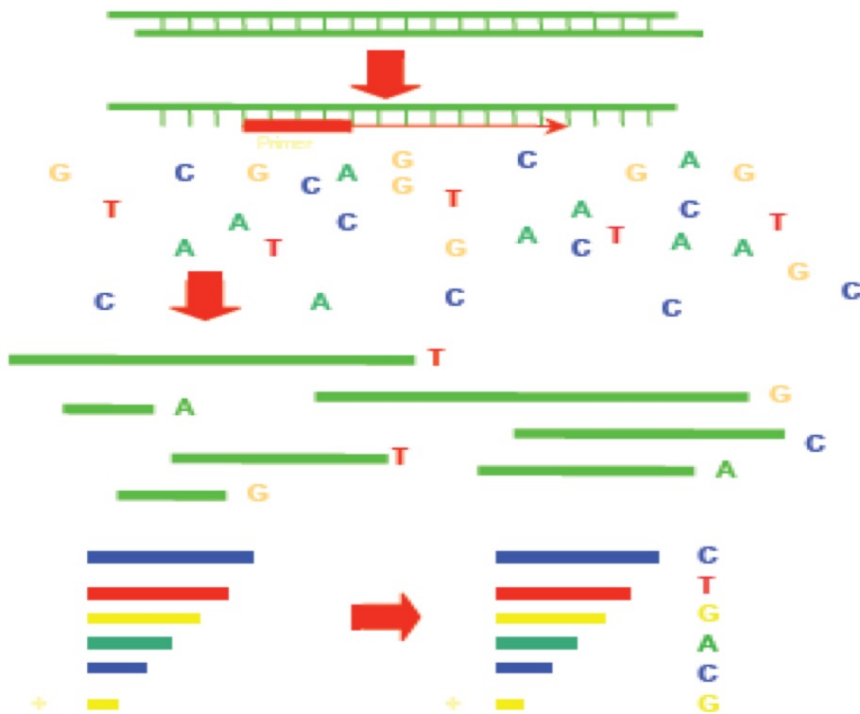
Feature	HiSeq2500 - Highoutput	HiSeq2500 – Rapid mode	MiSeq	PacBio RSII
Number of reads	150-180M/lane	100-150M/lane	12-15M (v2) 20-25M (v3)	50-80K/SMRT cell
Read length	2 x 100 bp	2 x 150 bp	2 x 300 bp (v3)	~ 10-20 kb
Yield per lane (PF data)	up to 35 Gb	up to 45Gb	up to 15 Gb	up to 0.4 Gb
Instrument Time	~12-14 days	~2 days	~2 days	~2 hours
Pricing per Gb	\$59 (PE100)	\$53 (PE150)	\$108 (PE300)	\$697



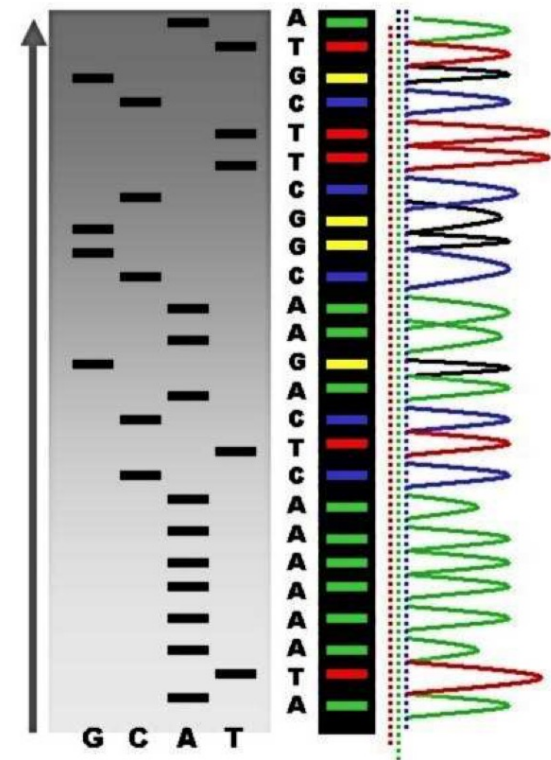


Sanger sequencing: dye-terminator sequencing

1986: 4 Reactions to 1 Lane
fluorescently labelled ddNTPs



Sequencing Reaction Products

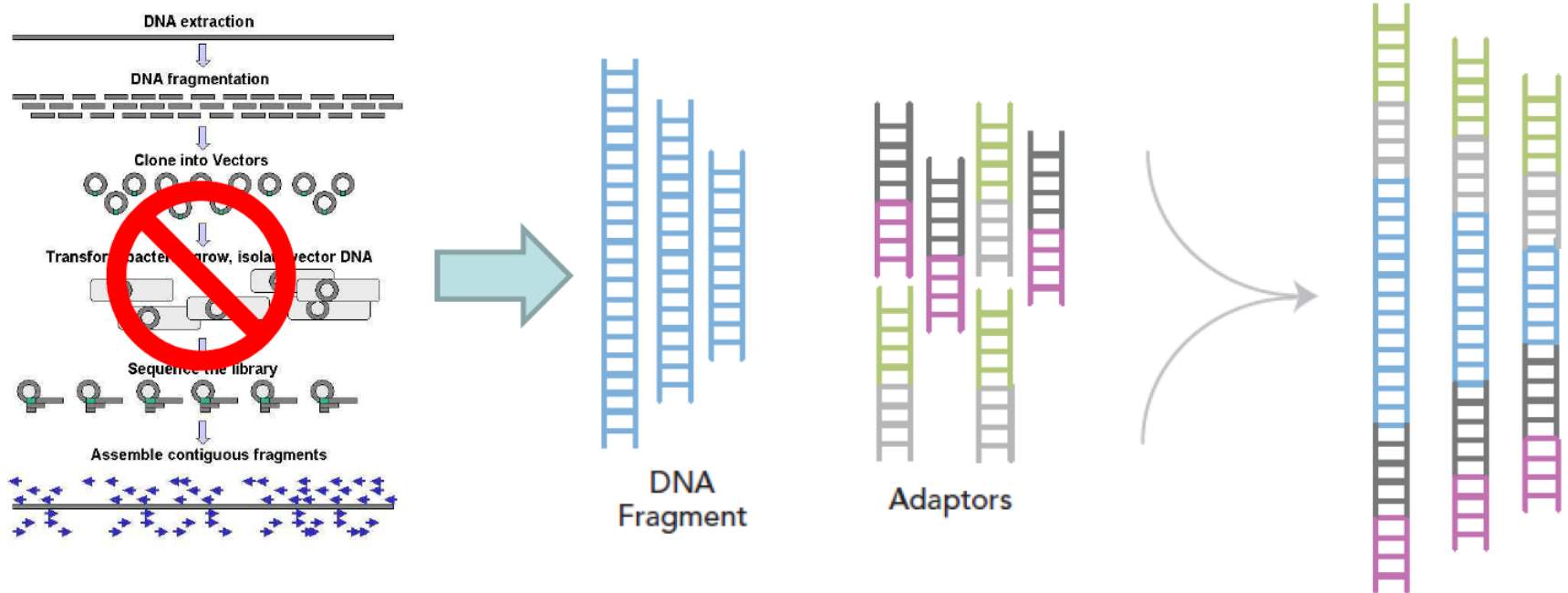


Progression of Sequencing Reaction

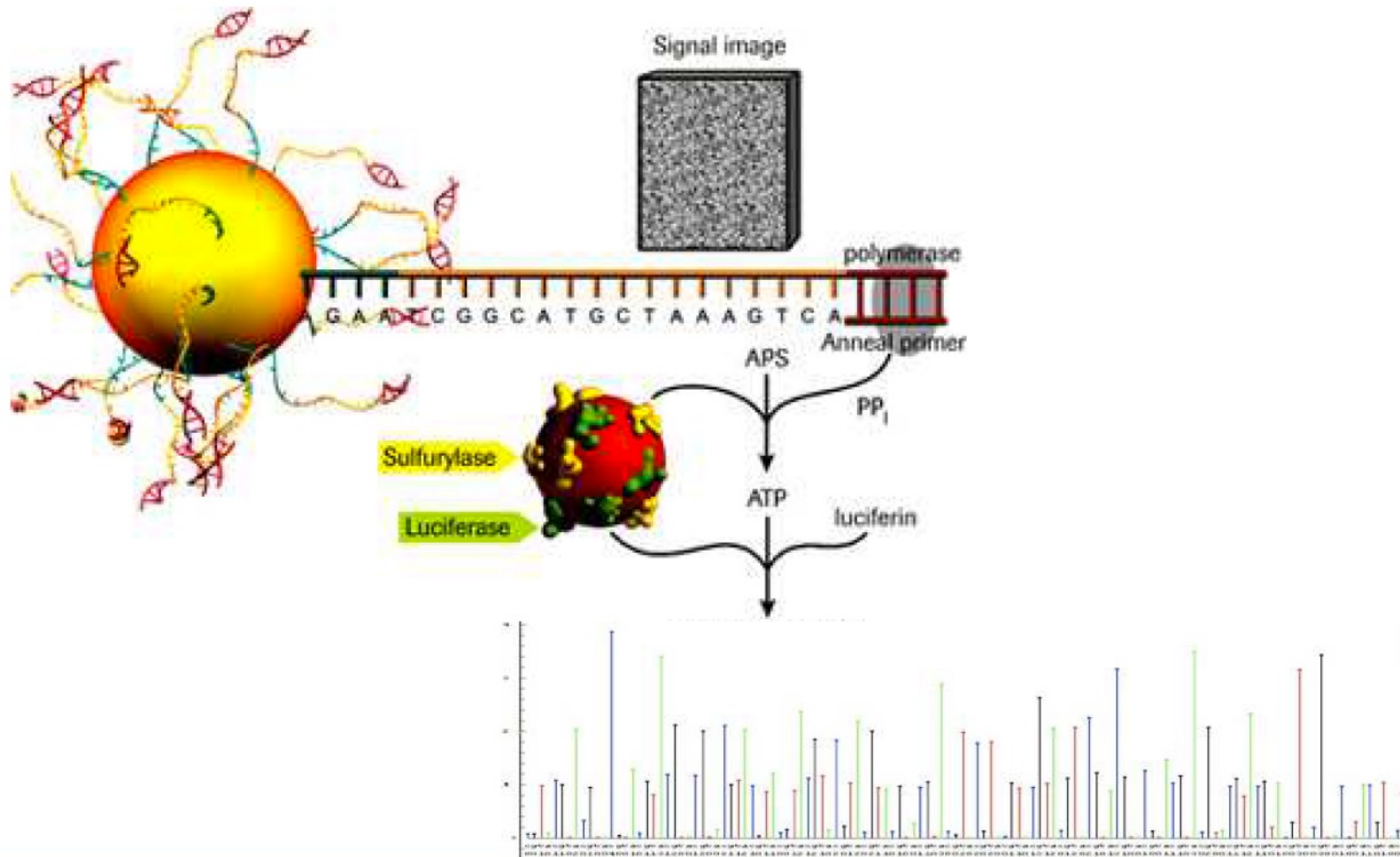
NGS

- Secuenciamiento másivo paralelo
 - Secuenciamiento por síntesis (de novo) cadena complementaria
 - Secuenciamiento por ligación (de novo) cadena complementaria
- Secuenciamiento directo

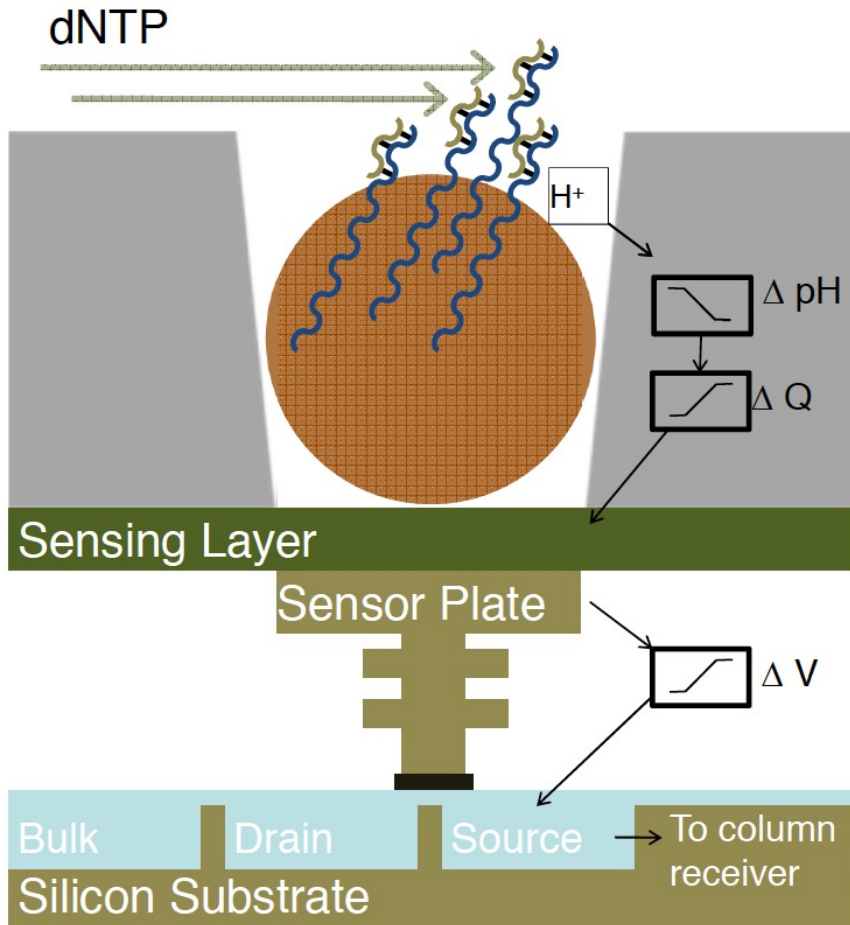
Next-gen sequencing: shotgun library preparation



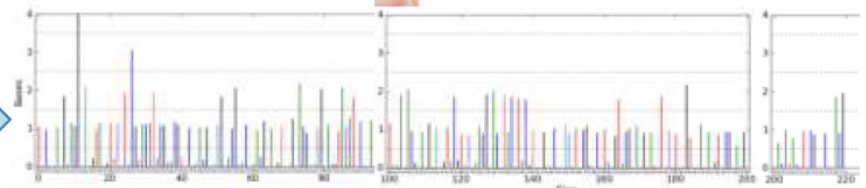
Pyrosequencing



Ion Torrent

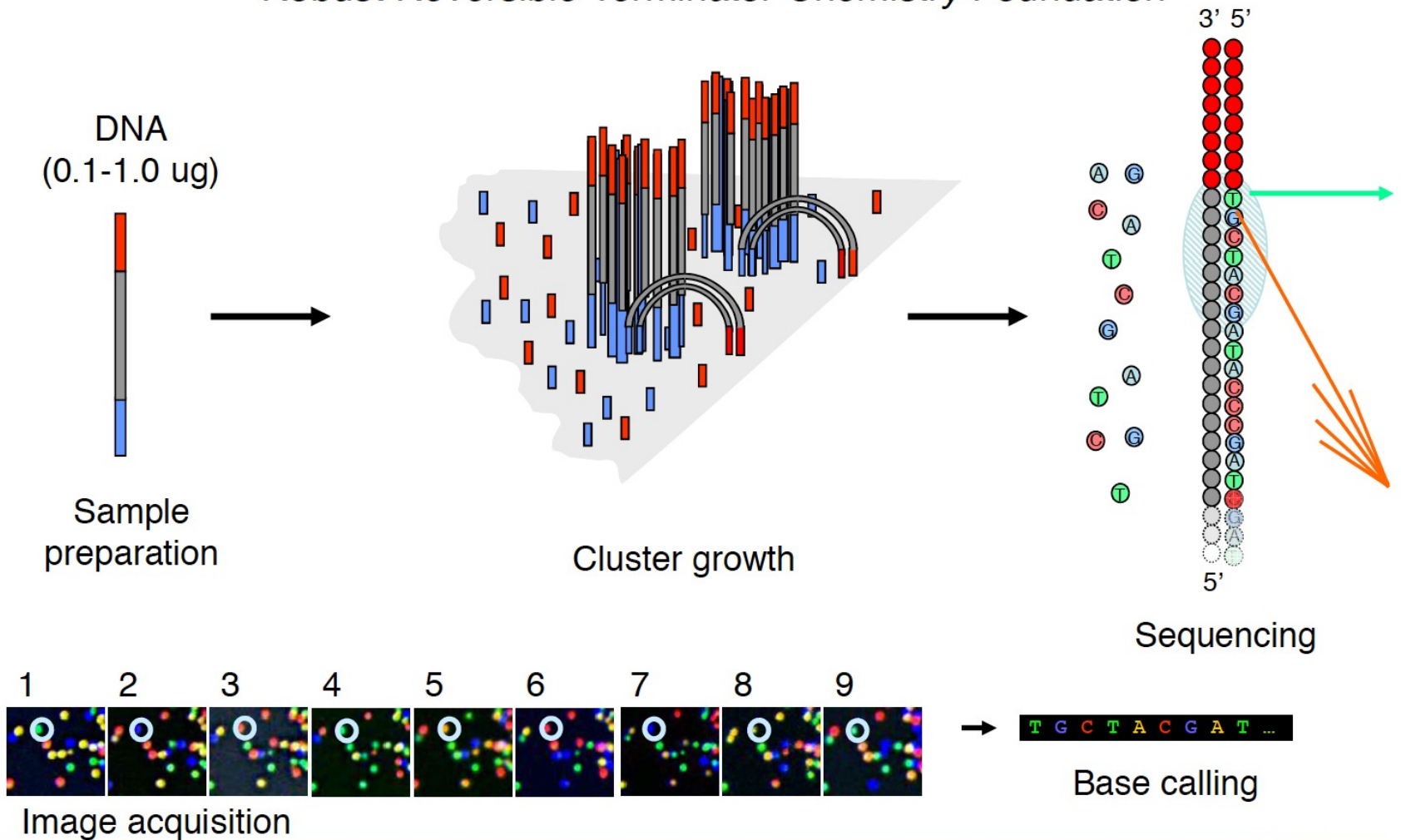


- DNA \rightarrow Ions \rightarrow Sequence
 - Nucleotides flow sequentially over Ion semiconductor chip
 - One sensor per well per sequencing reaction
 - Direct detection of natural DNA extension
 - Millions of sequencing reactions per chip
 - Fast cycle time, real time detection



Illumina Sequencing Technology

Robust Reversible Terminator Chemistry Foundation



Illumina



NextSeq Series +






































HiSeq Series +



NovaSeq Series +



HiSeq X Series†

Popular Applications & Methods	Key Application 	Key Application 	Key Application 	Key Application 
Large Whole-Genome Sequencing (human, plant, animal)				
Small Whole-Genome Sequencing (microbe, virus)				
Exome Sequencing				
Targeted Gene Sequencing (amplicon, gene panel)				
Whole-Transcriptome Sequencing				
Gene Expression Profiling with mRNA-Seq				
miRNA & Small RNA Analysis				
DNA-Protein Interaction Analysis				
Methylation Sequencing				
Shotgun Metagenomics				

Zero Mode Waveguide (Single molecule real time seq) PacBio

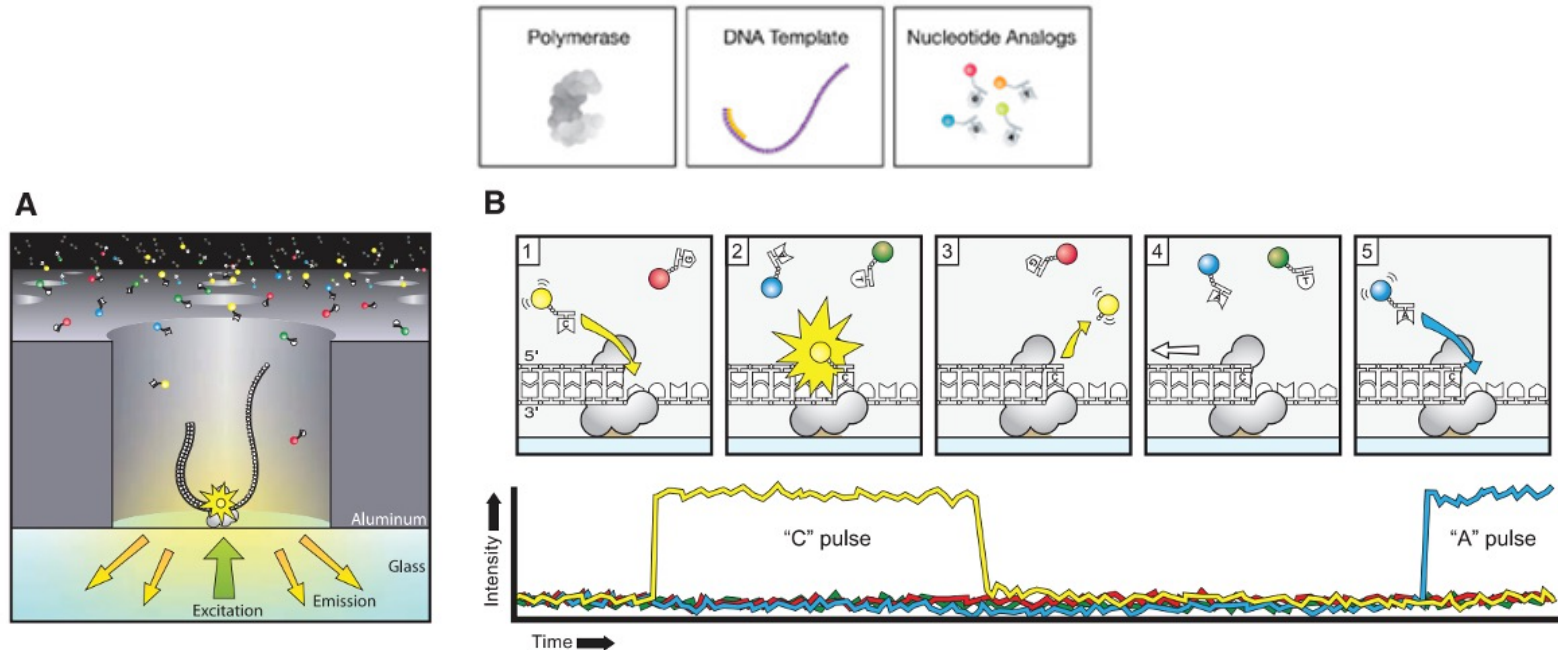
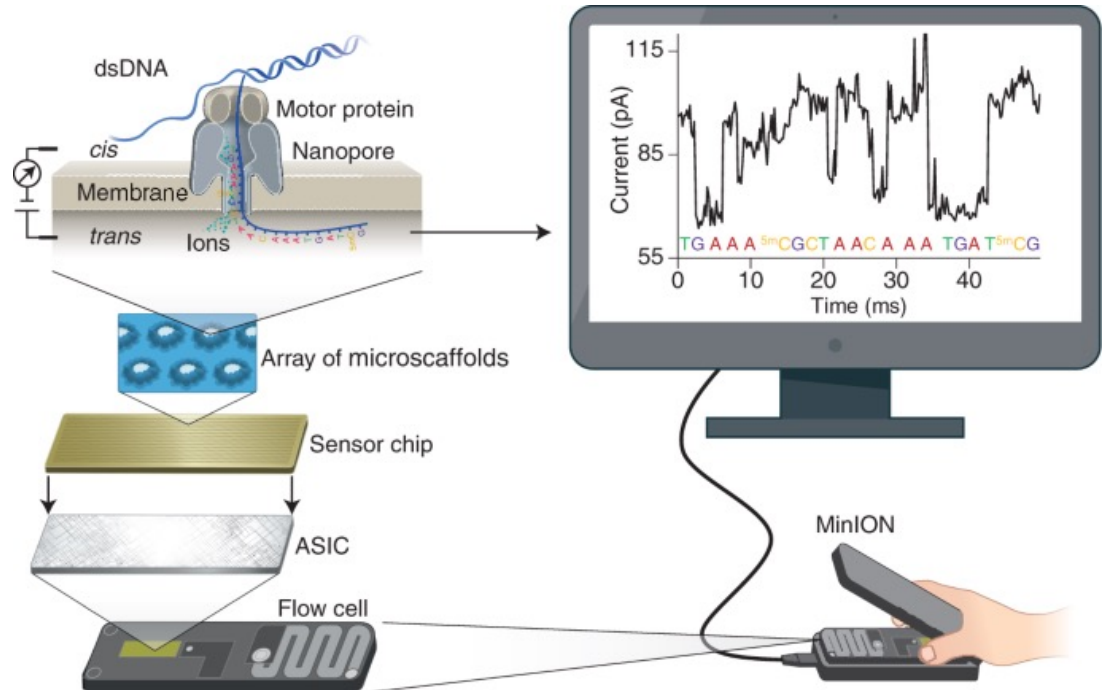


Fig. 1. Principle of single-molecule, real-time DNA sequencing. **(A)** Experimental geometry. A single molecule of DNA template-bound $\Phi 29$ DNA polymerase is immobilized at the bottom of a ZMW, which is illuminated from below by laser light. The ZMW nanostructure provides excitation confinement in the zeptoliter (10^{-21} liter) regime, enabling detection of individual phospholinked nucleotide substrates against the bulk solution background as they are incorporated into the DNA strand by the polymerase. **(B)** Schematic event sequence of the phospholinked dNTP incorporation cycle,

with a corresponding expected time trace of detected fluorescence intensity from the ZMW. (1) A phospholinked nucleotide forms a cognate association with the template in the polymerase active site, (2) causing an elevation of the fluorescence output on the corresponding color channel. (3) Phosphodiester bond formation liberates the dye-linker-pyrophosphate product, which diffuses out of the ZMW, thus ending the fluorescence pulse. (4) The polymerase translocates to the next position, and (5) the next cognate nucleotide binds the active site beginning the subsequent pulse.

Nanopore sequencing (direct reading)

- 3rd generation sequencing



Nanopore sequencing (direct reading)

- 3rd generation sequencing

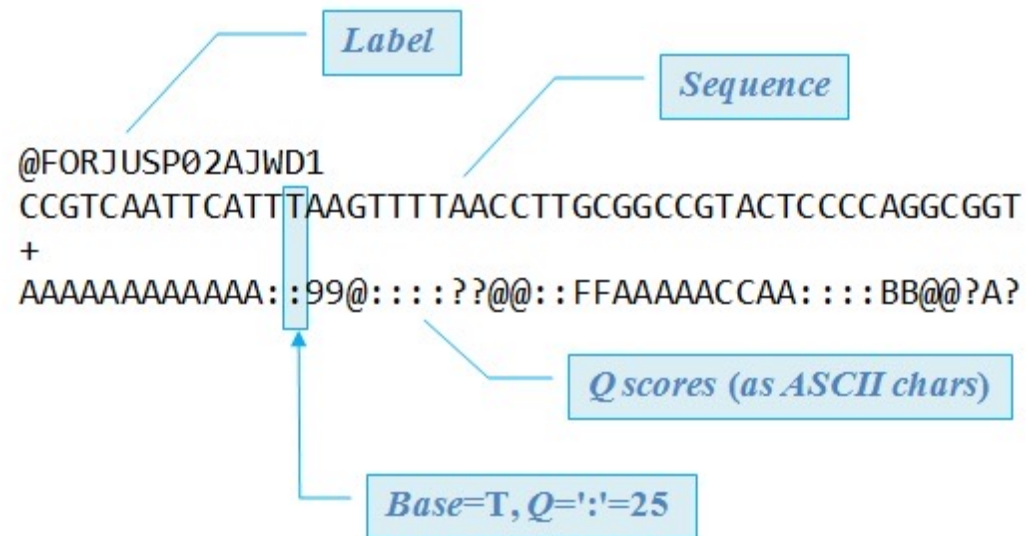


	Flongle	MinION	GridION (5 flow cells)	PromethION (48 flow cells)
				
Maximum run time	16 hours	72 hours	72 hours	64 hours
Theoretical 1D maximum yield	Up to 3.3 Gb	Up to 40 Gb	Up to 200 Gb	Up to 15 Tb
Current 1D maximum yield	Up to 2 Gb	Up to 30 Gb	Up to 150 Gb	Up to 8.6 Tb
Available channels	Up to 126	Up to 512	Up to 2,560	Up to 144,000

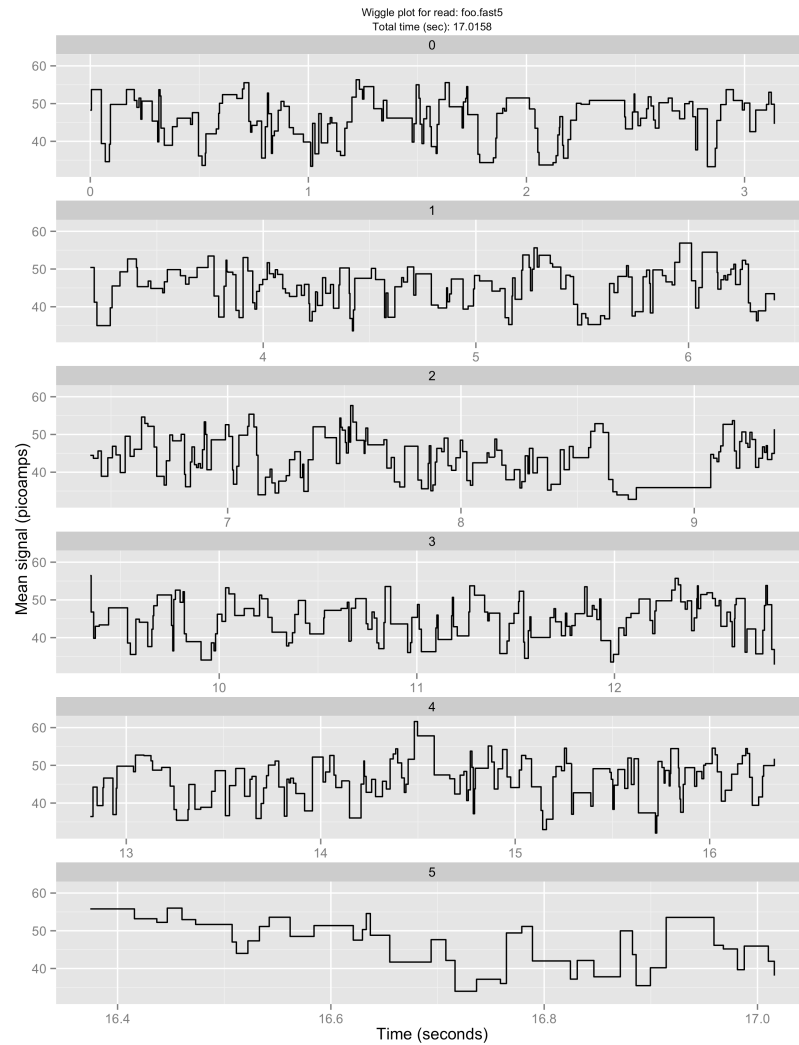
Fasta files

[illegible]

Fastq

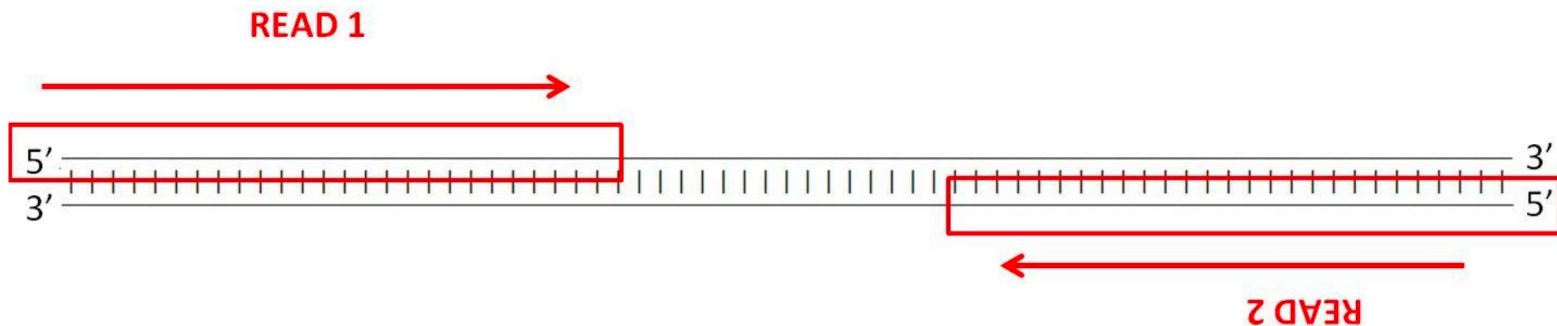


Fast5



Paired sequence reads

(*i.e.* we sequence from both ends of the same molecule)



Calculating Phred Quality Scores

- Q scores are defined as a property that is logarithmically related to the base calling error probabilities (P)².

$$Q = -10 \log_{10} P$$

- For example, if Phred assigns a Q score of 30 (Q30) to a base, this is equivalent to the probability of an incorrect base call 1 in 1000 times (Table 1).
- This means that the base call accuracy (i.e., the probability of a correct base call) is 99.9%.
- A lower base call accuracy of 99% (Q20) will have an incorrect base call probability of 1 in 100, meaning that every 100 bp sequencing read will likely contain an error.

Table 1: Quality Scores and Base Calling Accuracy

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%