



## Secuenciación Masiva y análisis de secuencias

Isabel Cuesta

[isabel.cuesta@isciii.es](mailto:isabel.cuesta@isciii.es)

BU-ISCIII

Unidades Centrales Científico Técnicas - SGSAFI-ISCIII

25 Noviembre 2020  
Master Virología

# Index

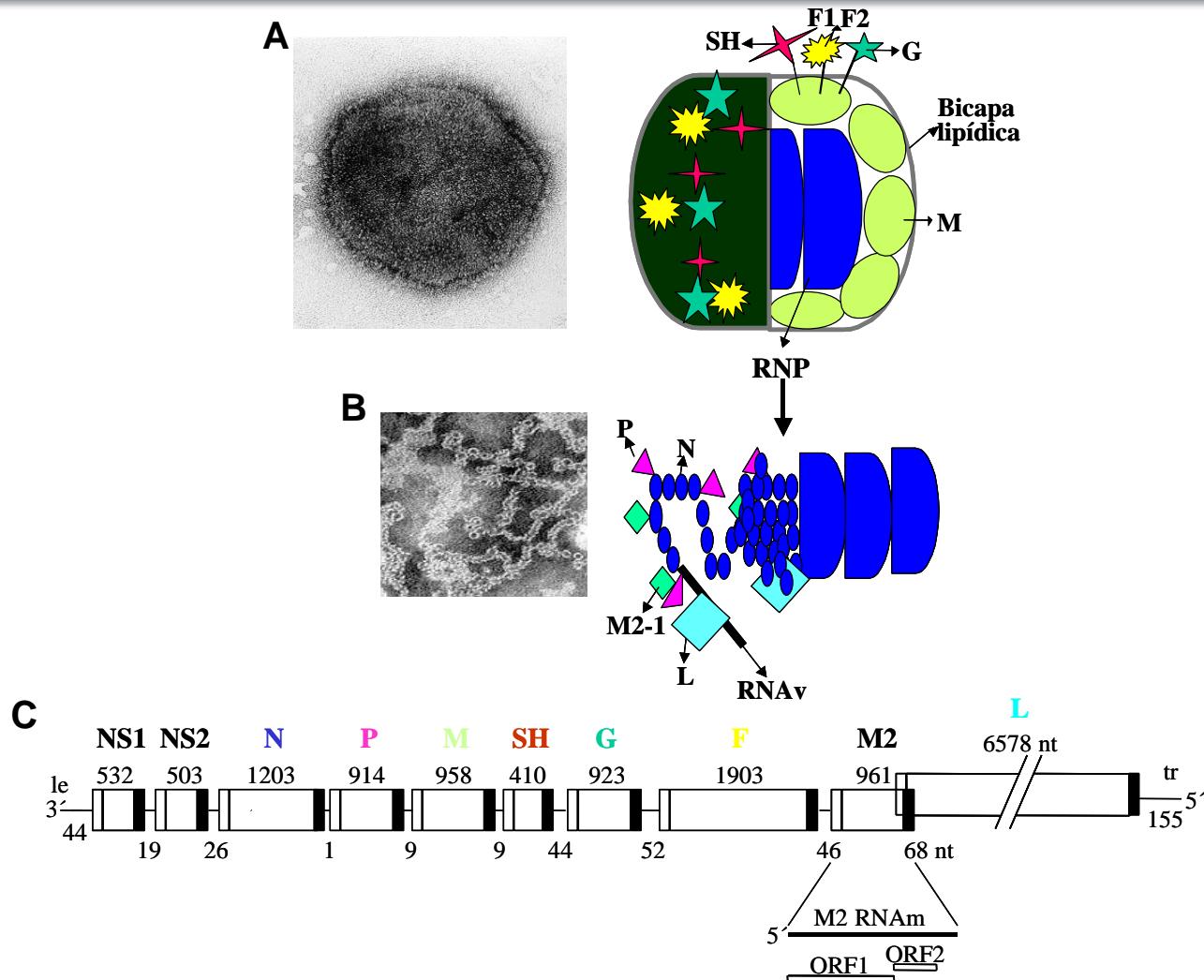
- Secuenciación Masiva (Isabel Cuesta)
- Bioinformática (Isabel Cuesta)
- Conceptos análisis de datos: genoma de SARS-CoV-2 (Sara Monzón)
- Práctica: ViralRecon y Metagenómica (Sarai Varona y Miguel Juliá)





# ¿Para que secuenciar genomas?

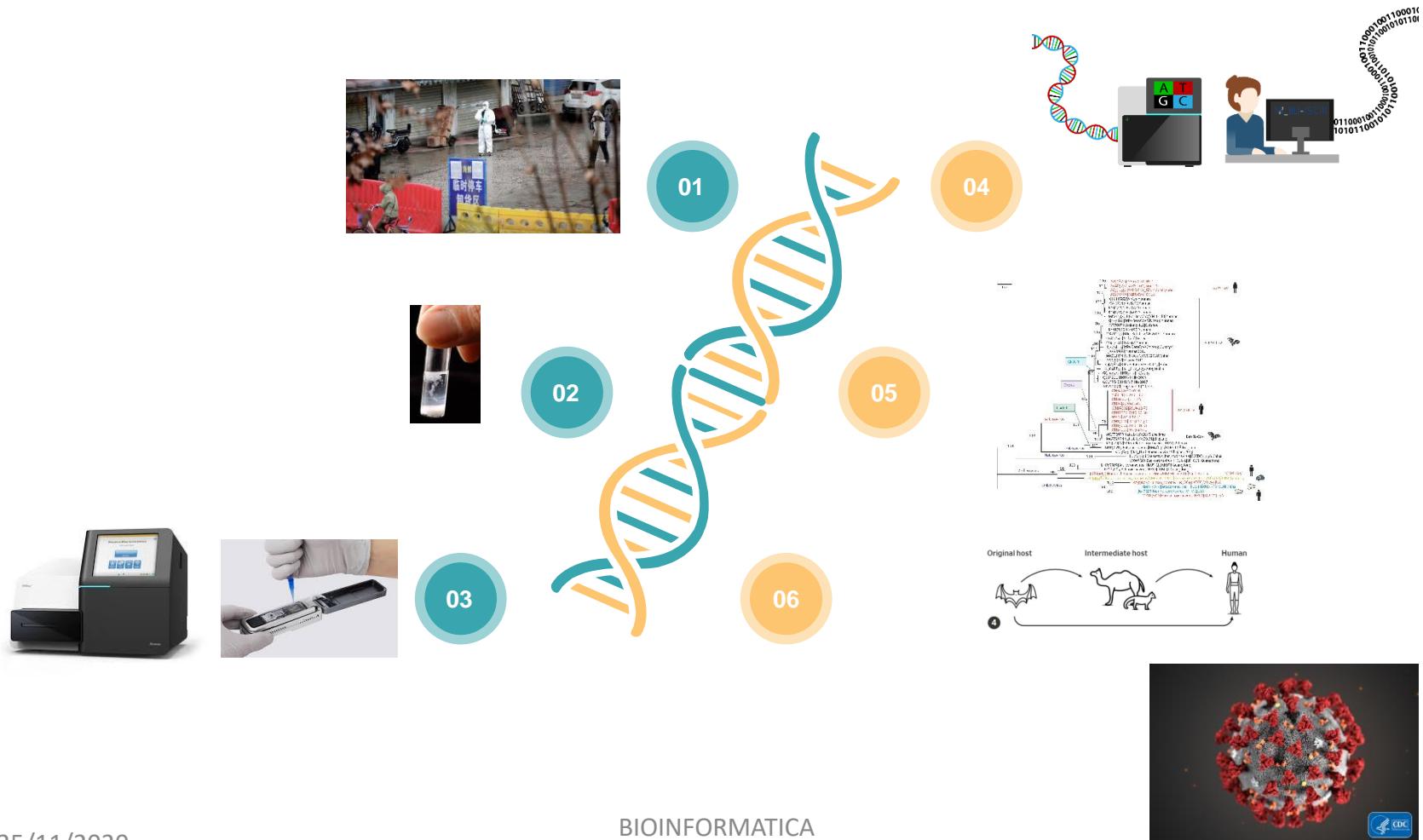
# Virus Respiratorio Sincitial Humano



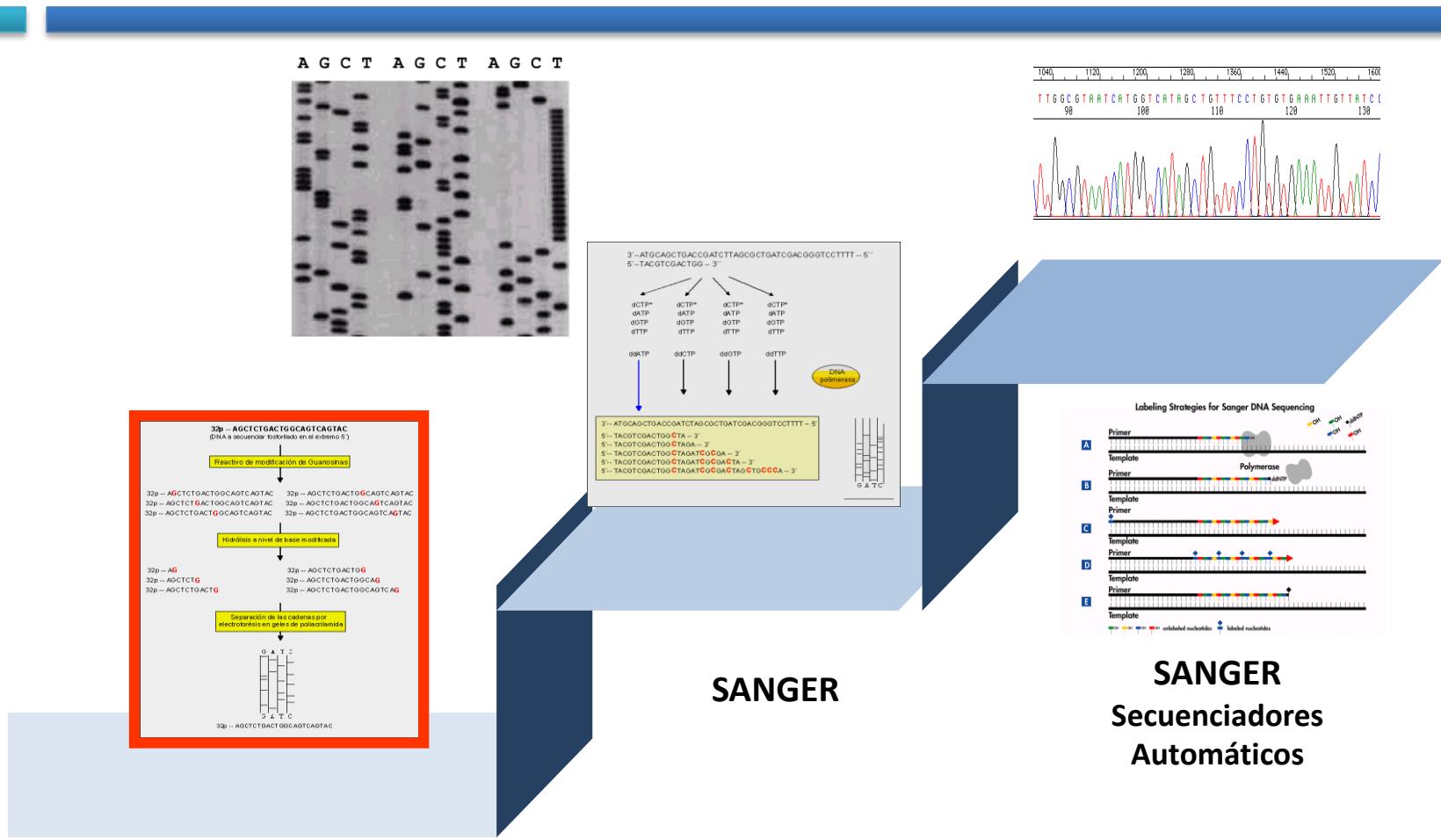


# Descubrimiento de un Nuevo virus

## Secuenciando su genoma



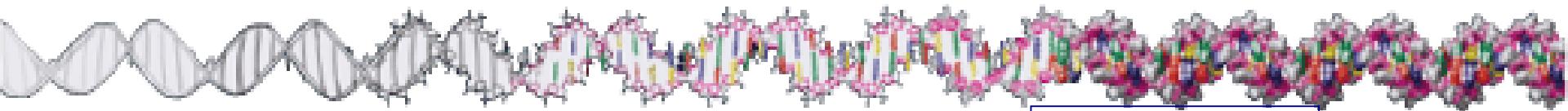
# Métodos de secuenciación de DNA



# Evolution of DNA Revolution

A walk through the biological history: from Sanger to NGS

1953



Watson & Crick: The discovery of the molecular structure of DNA: the double helix (*Nature*, 171, 1953).

1972

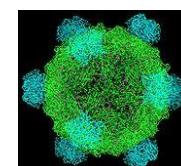
Paul Berg: The first recombinant DNA molecule is build (PNAS 69, 1972).

Development of recombinant genetic engineering

BIOINFORMATICA  
Master Virología

1977

Gilbert & Maxam Sanger Developed new techniques for rapid DNA sequencing.



Bacteriophage ΦX174  
5386nt  
plus and minus method  
  
>X\_BULISCHI

# Evolution of DNA Revolution

A walk through the biological history: from Sanger to NGS

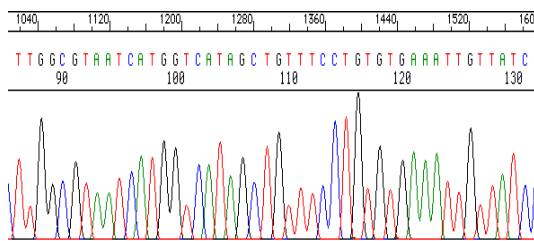
1986

Hood (Applied Biosystems Incorporated, ABI) Developed the automatic sequencer (Nature 321, 1986; Science 254 1991)

1990



BLAST: Myers & Lipman publish the first algorithm to align DNA sequences (JMB, 215 1990).



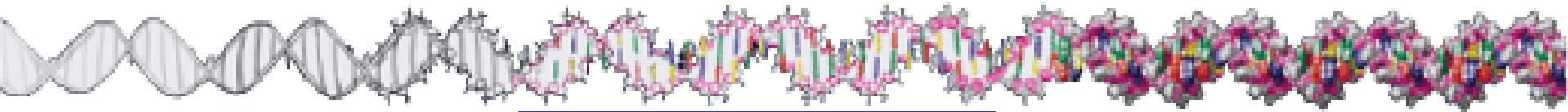
# Evolution of DNA Revolution

A walk through the biological history: from Sanger to NGS

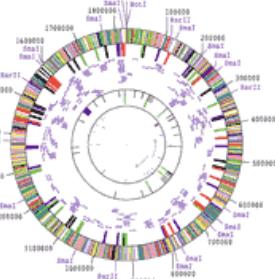
1995

1996

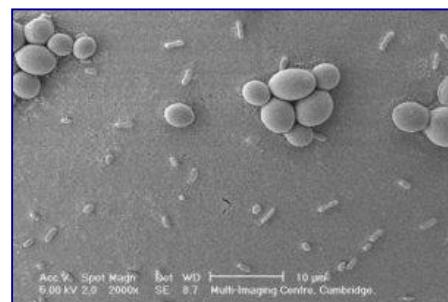
2003



Venter (TIGR)  
*Haemophilus influenzae*  
Genome (1,8Mb),  
*Mycoplasma genitalium*  
(0,5Mb)  
(Science 269, 1995)



Consortium of 600  
Scientist from Europe,  
North America and Japan  
*Saccharomyces cerevisiae*  
genome (Science).



Bacterial Genomes

*Bacillus subtilis*  
*Escherichia coli*

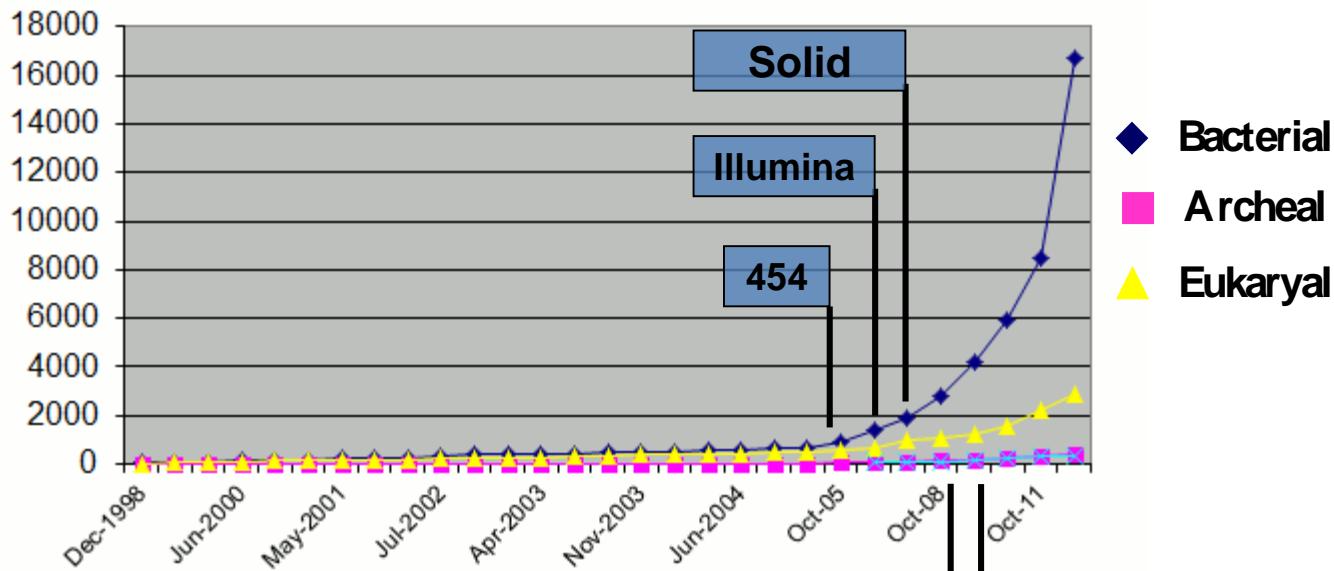
Eukaryotic Genomes

*Caenorhabditis elegans*  
*Arabidopsis thaliana*  
*Drosophila melanogaster*  
Human genome (1986-2003)

# Genomics Revolution Era



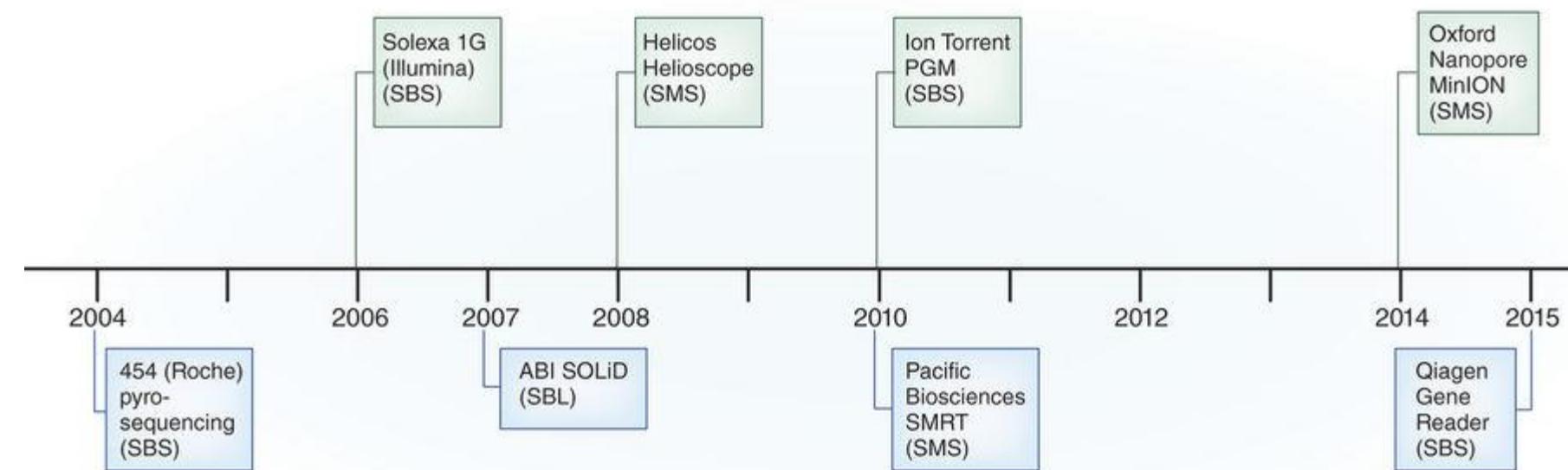
Genome Projects on GOLD according to Phylogenetic Groups ©  
October 2012 - 20327 Projects



Source: <http://www.genomeonline.org>



# DNA sequencing technologies 2006-2016



Mardis, Nature Protocols 2017

# DNA sequencing technologies 2006-2016

Primera  
generación

- Sanger

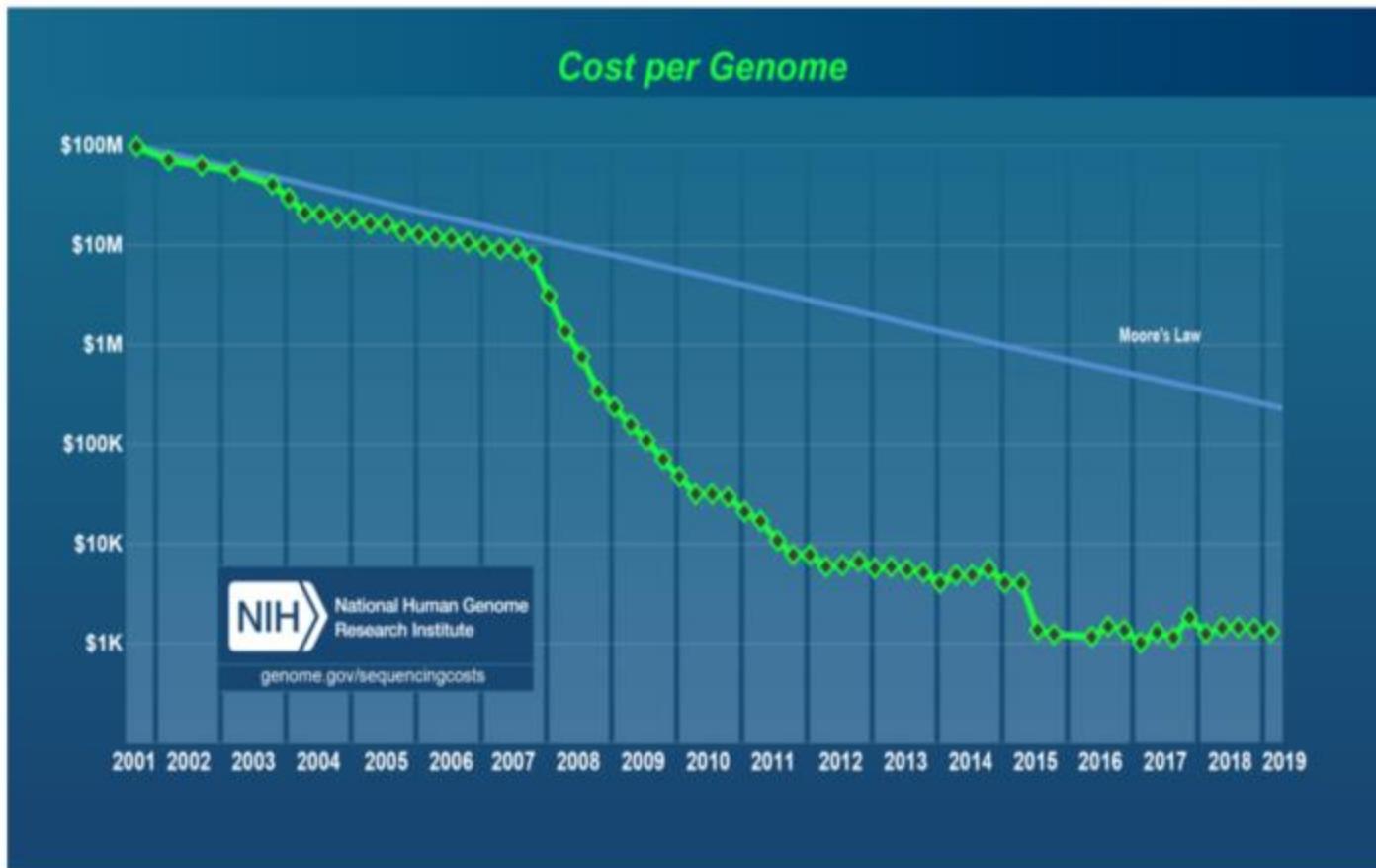
## Segunda Generación

- 454/Roche
- Solexa/Illumina
- Solid
- Ion Torrent

## Tercera Generación

- Pacific Biosciences
- Nanopore

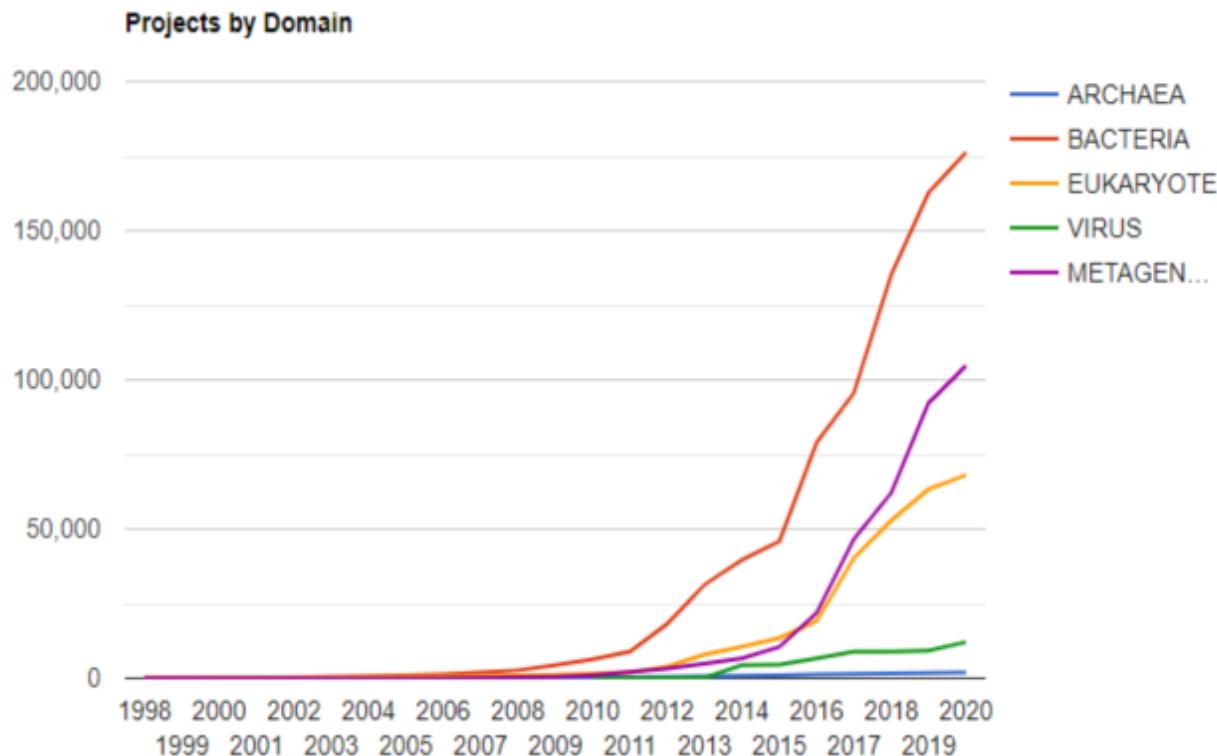
# Sequencing cost



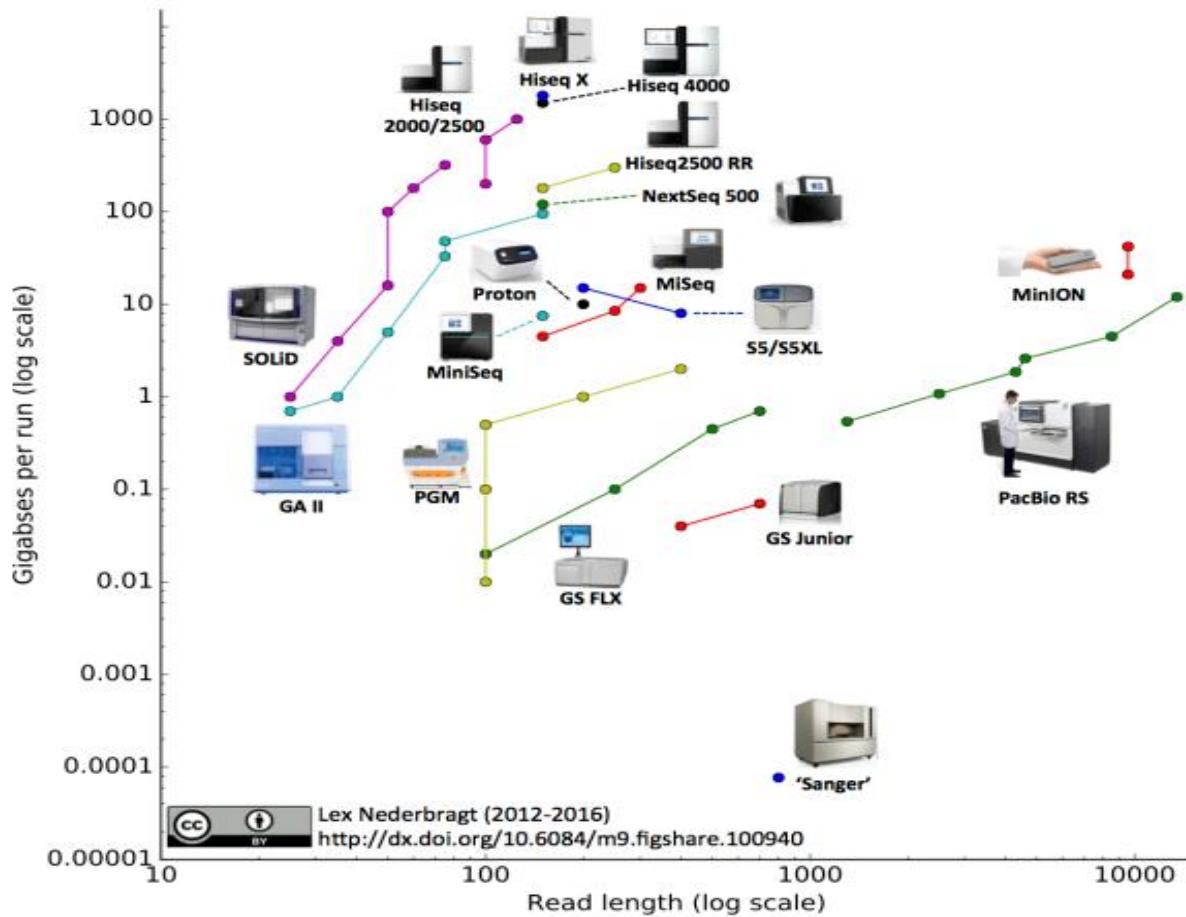
# Sequencing projects

<https://gold.jgi.doe.gov/>

## GOLD, Genome Online DataBase

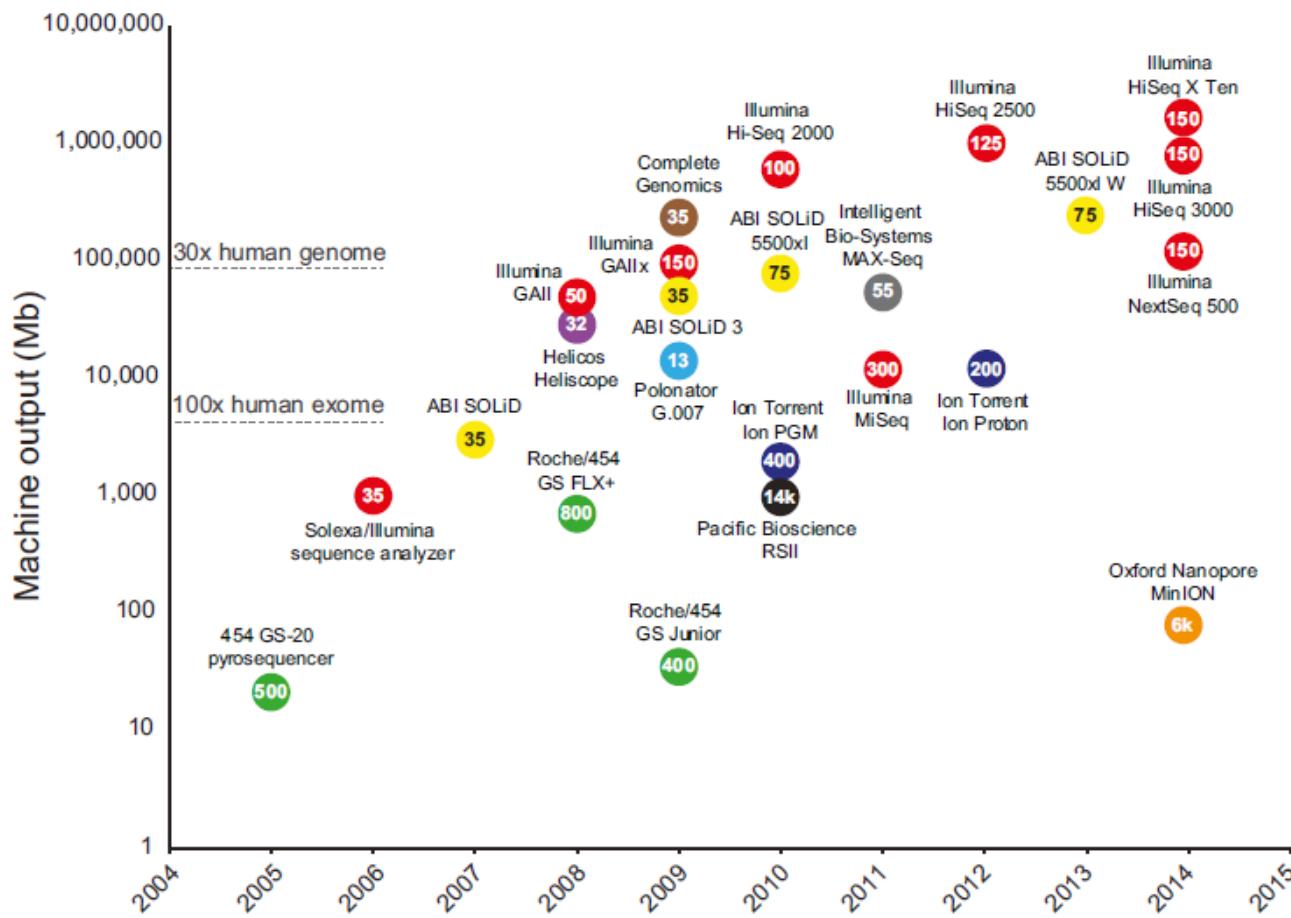


# High-Throughput Sequencing Technologies



<https://flxlexblog.wordpress.com/>

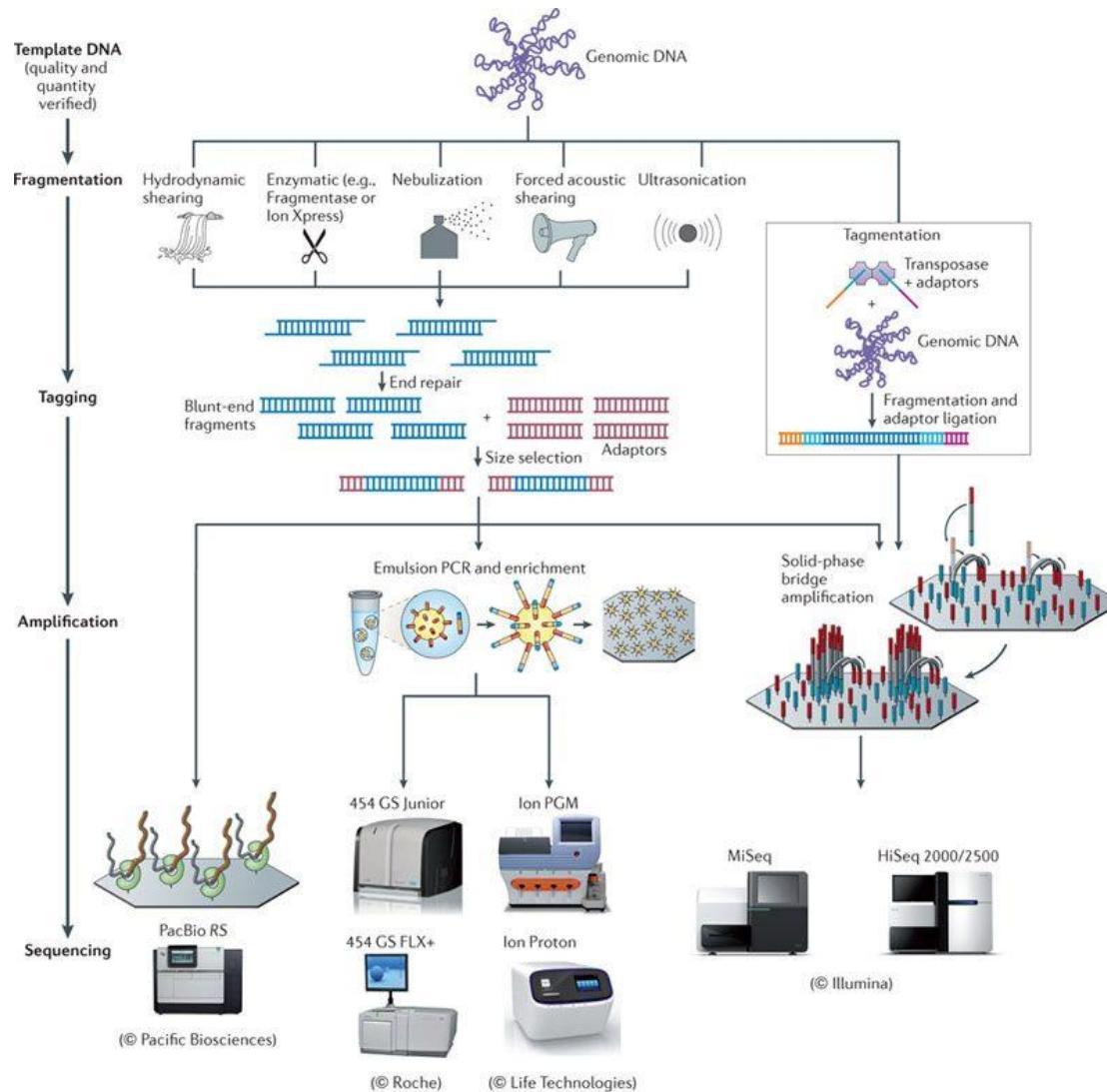
# High-Throughput Sequencing Technologies



Numbers inside data points denote current read lengths.  
Sequencing platforms are color coded.

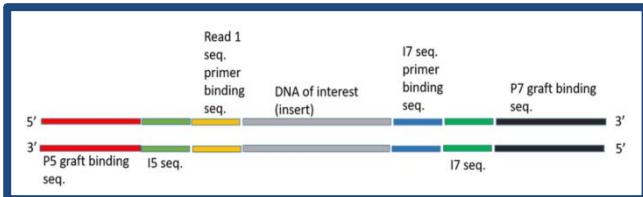
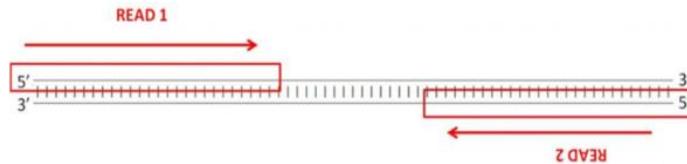
Reuter et al., Mol Cell 2015

# High-throughput sequencing platforms

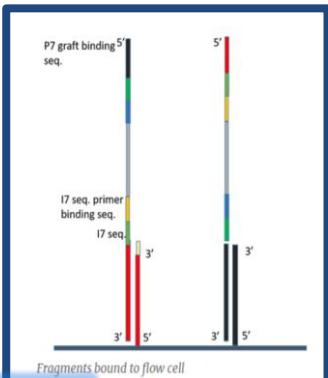


Nature Reviews | Microbiology Loman et al, 2012

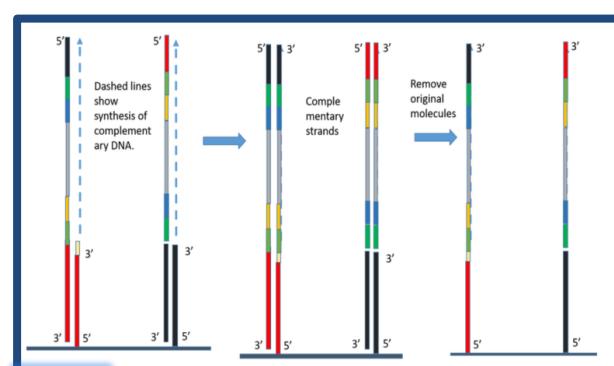
# Illumina sequencing



1



2

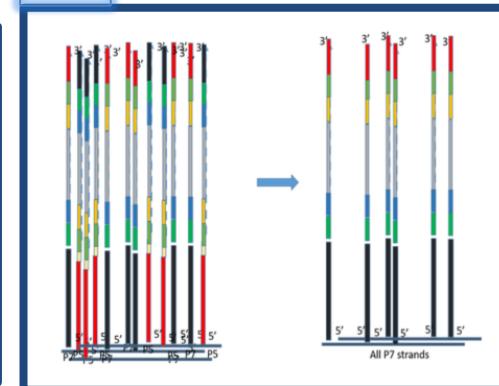


3

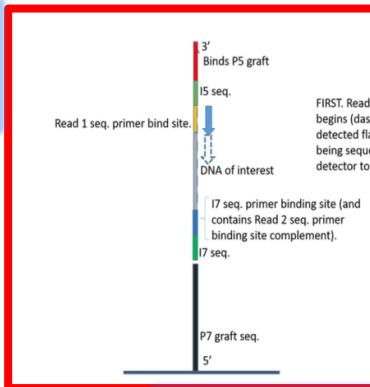
4



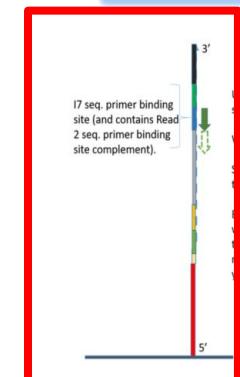
6 R1



7



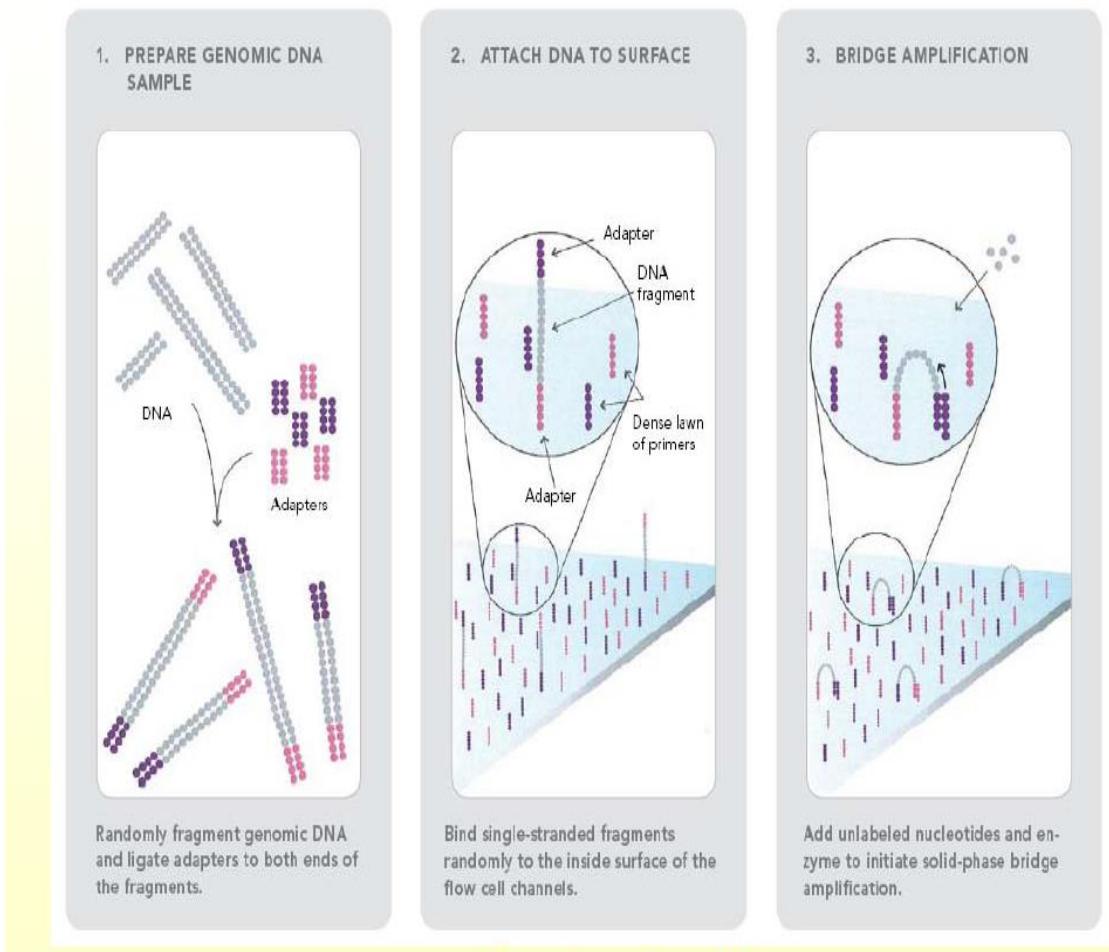
8 R2



<https://kscbioinformatics.wordpress.com/2017/02/13/illumina-sequencing-for-dummies-samples-are-sequenced/>

# Illumina sequencing

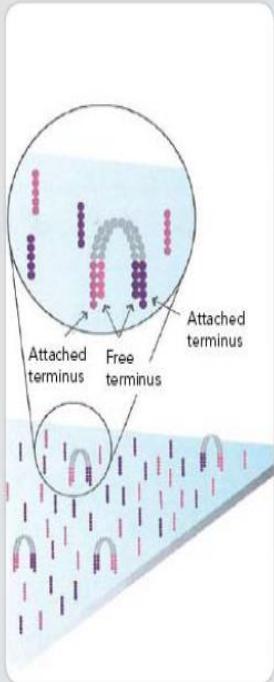
## SEQUENCING TECHNOLOGY OVERVIEW



*amplificación en fase sólida*

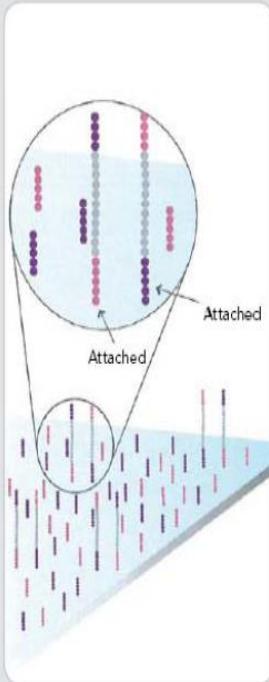
# Illumina sequencing

## 4. FRAGMENTS BECOME DOUBLE-STRANDED



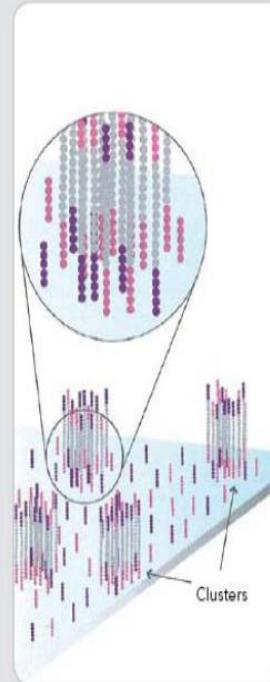
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

## 5. DENATURE THE DOUBLE-STRANDED MOLECULES



Denaturation leaves single-stranded templates anchored to the substrate.

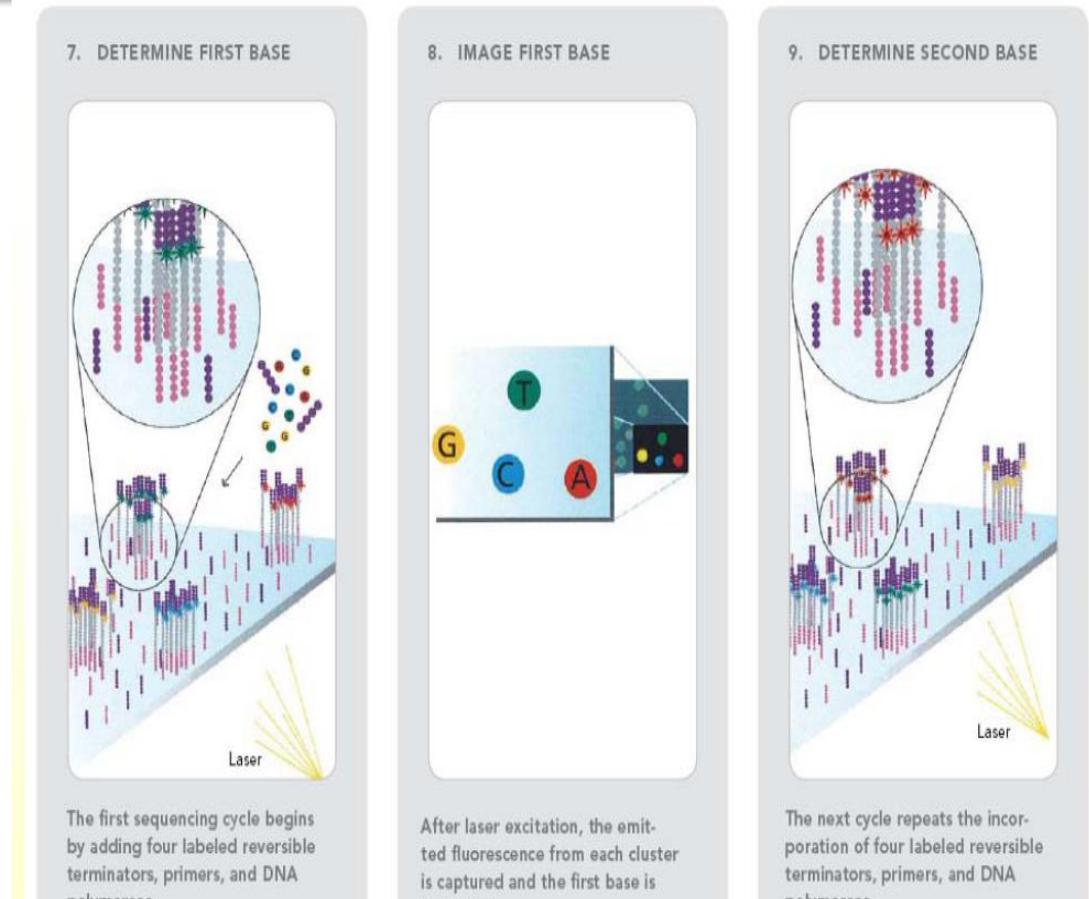
## 6. COMPLETE AMPLIFICATION



Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

$> 1000$  copies in  $\leq 1 \mu\text{m}$ ;  $10^7$  clusters per  $\text{cm}^2$

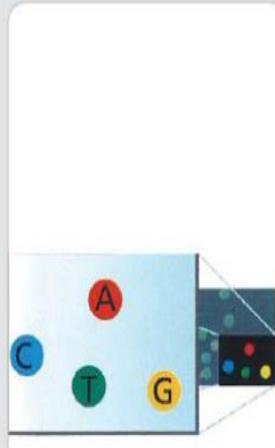
# Illumina sequencing



*Secuenciación con nucleótidos terminadores reversibles*

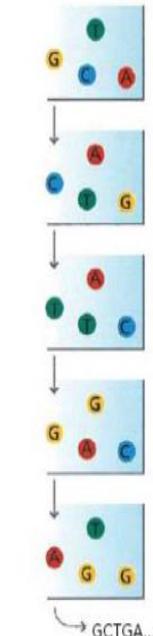
# Illumina sequencing

## 10. IMAGE SECOND CHEMISTRY CYCLE



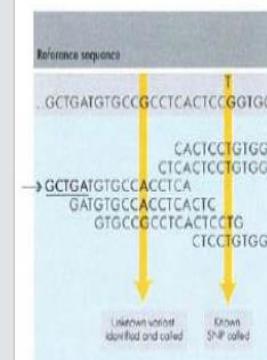
After laser excitation, the image is captured as before, and the identity of the second base is recorded.

## 11. SEQUENCING OVER MULTIPLE CHEMISTRY CYCLES



The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

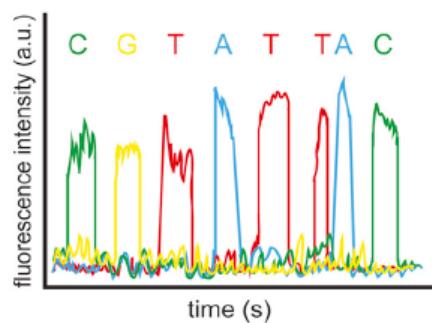
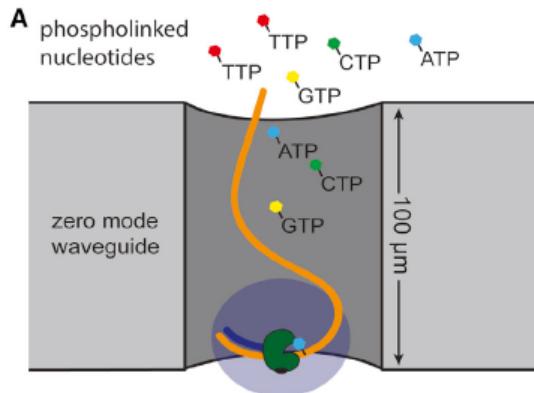
## 12. ALIGN DATA



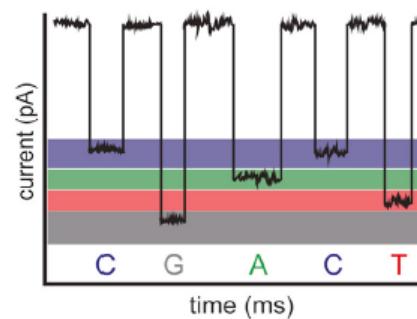
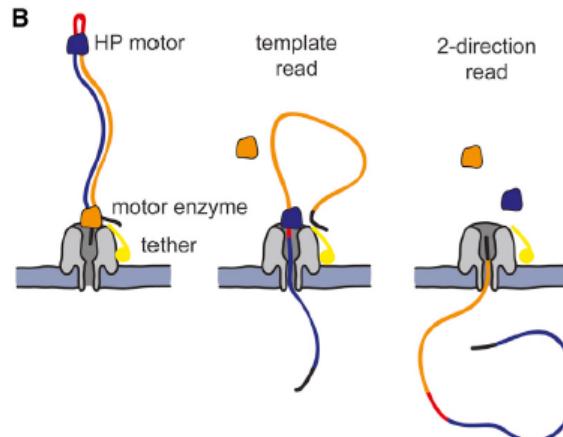
The data are aligned and compared to a reference, and sequencing differences are identified.

# The Third-generation Sequencing Technologies

## Single Molecule Sequencing Platforms



Pacific Bioscience's SMRT sequencing



Oxford Nanopore's sequencing strategy

Reuter et al., Mol Cell 2015

# PacBio sequencing and its applications

Rhoads & Au, Gen Prot Bioinf 2015



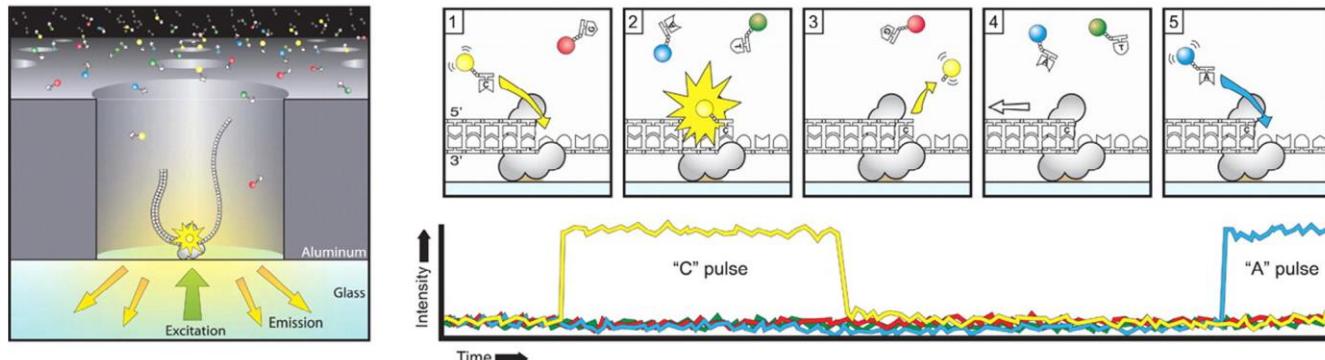
**SMRTbell template:** is a closed, single-stranded circular DNA that is created by ligating hairpin adaptors to both ends of a target dsDNA

**Sequencing by light pulses:** The replication processes in all ZMWs of a SMRTcell are recorder by a movie of light pulses, and the pulses corresponding to each ZMW can be interpreted to be a sequence of bases (**continuous long read, CLR**).

Both strands can be sequenced multiple times (passes) in a single CLR. CLR can be split to multiple reads (subreads) and CCS is the consensus sequence of multiple subreads



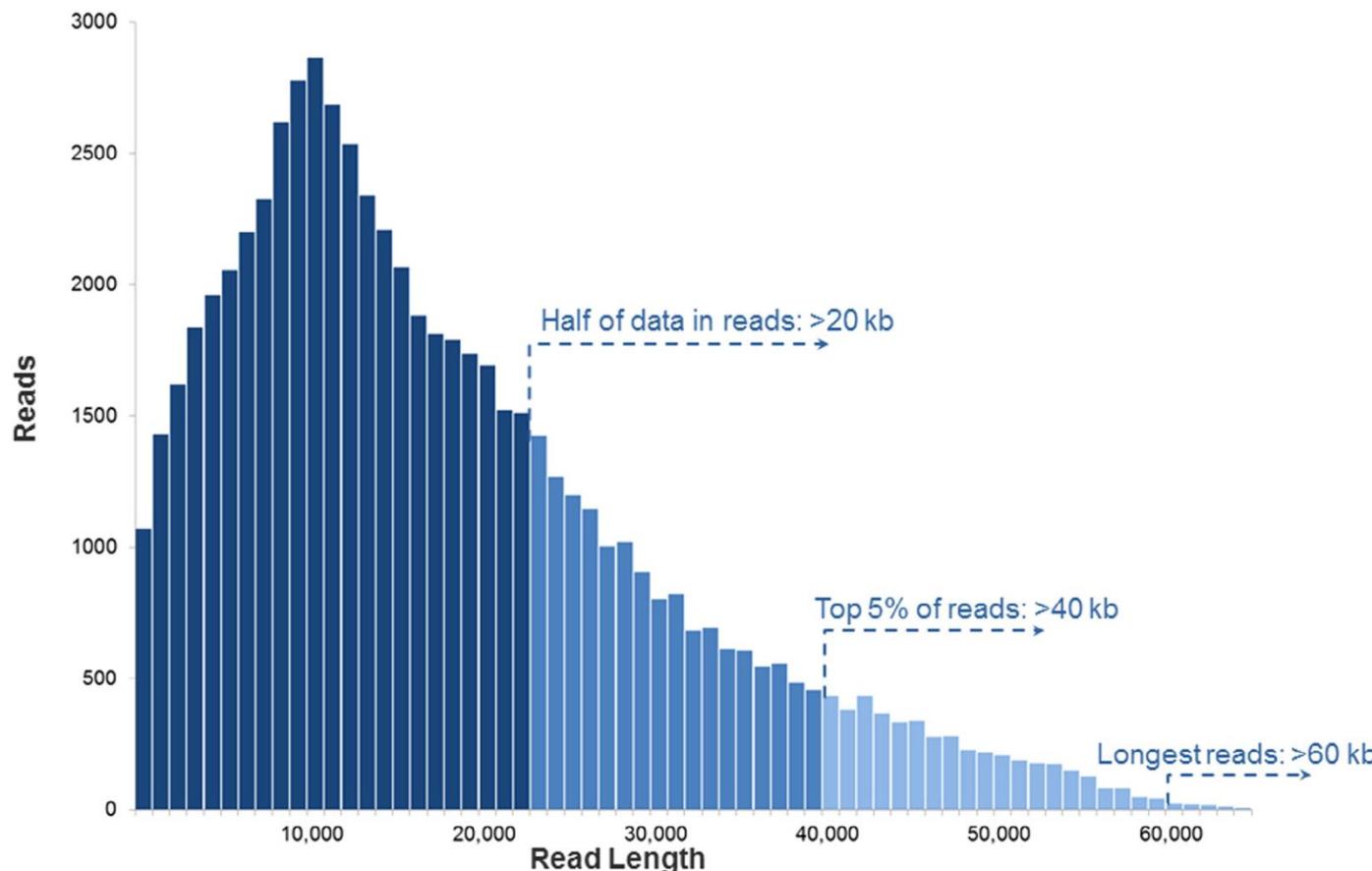
**A single SMRT cell:** this contains 150000 ZMWs (zero-mode waveguide). A SMRTbell diffuses into a ZMW.  
Approx 35000 -75000 ZMWs produce a read in a run lasting 0,5-4h resulting in 0,5-1Gb.



# PacBio sequencing and its applications

Rhoads & Au, Gen Prot Bioinf 2015

**PacBio RS II read length distribution** using P6-C4 chemistry. Data are based on a 20kb size-selected E. coli library using a 4-h movie. A SMRTcell produces 0,5-1 billion bases.



# PacBio sequencing and its applications

Rhoads & Au, Gen Prot Bioinf 2015

## Advantage

Closes gaps and completes genomes due to longer reads

Identifies non-SNP SVs

## Achievements

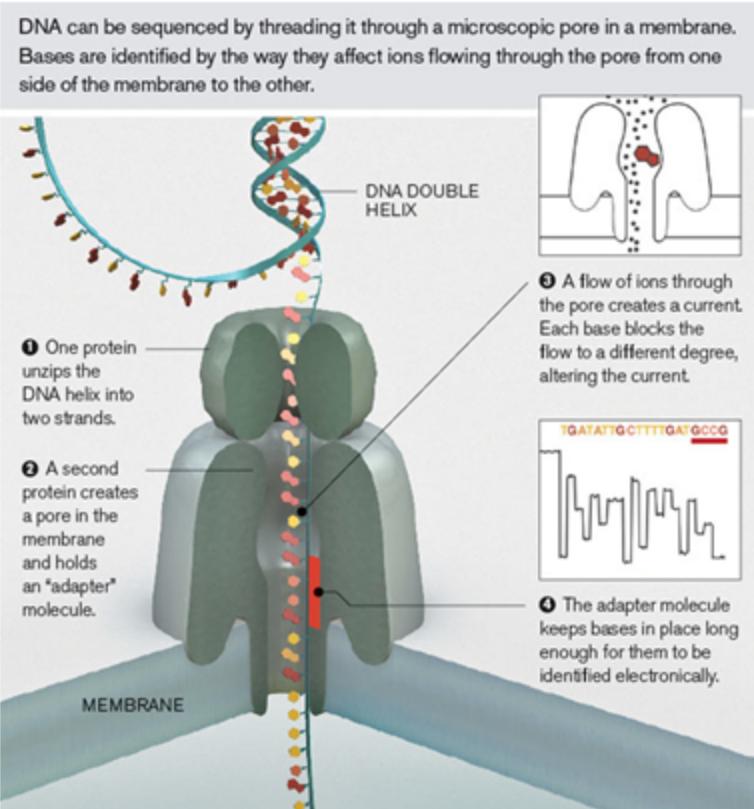
Produced highly-contiguous assemblies of bacterial and eukaryotic genomes

Discovered STRs (short tandem repeats)

## Limitations

Both strands can be sequenced several times if the lifetime of the polymerase is long enough.

# Nanopore-based fourth-generation DNA sequencing technology. ONT, Oxford Nanopore Technologies



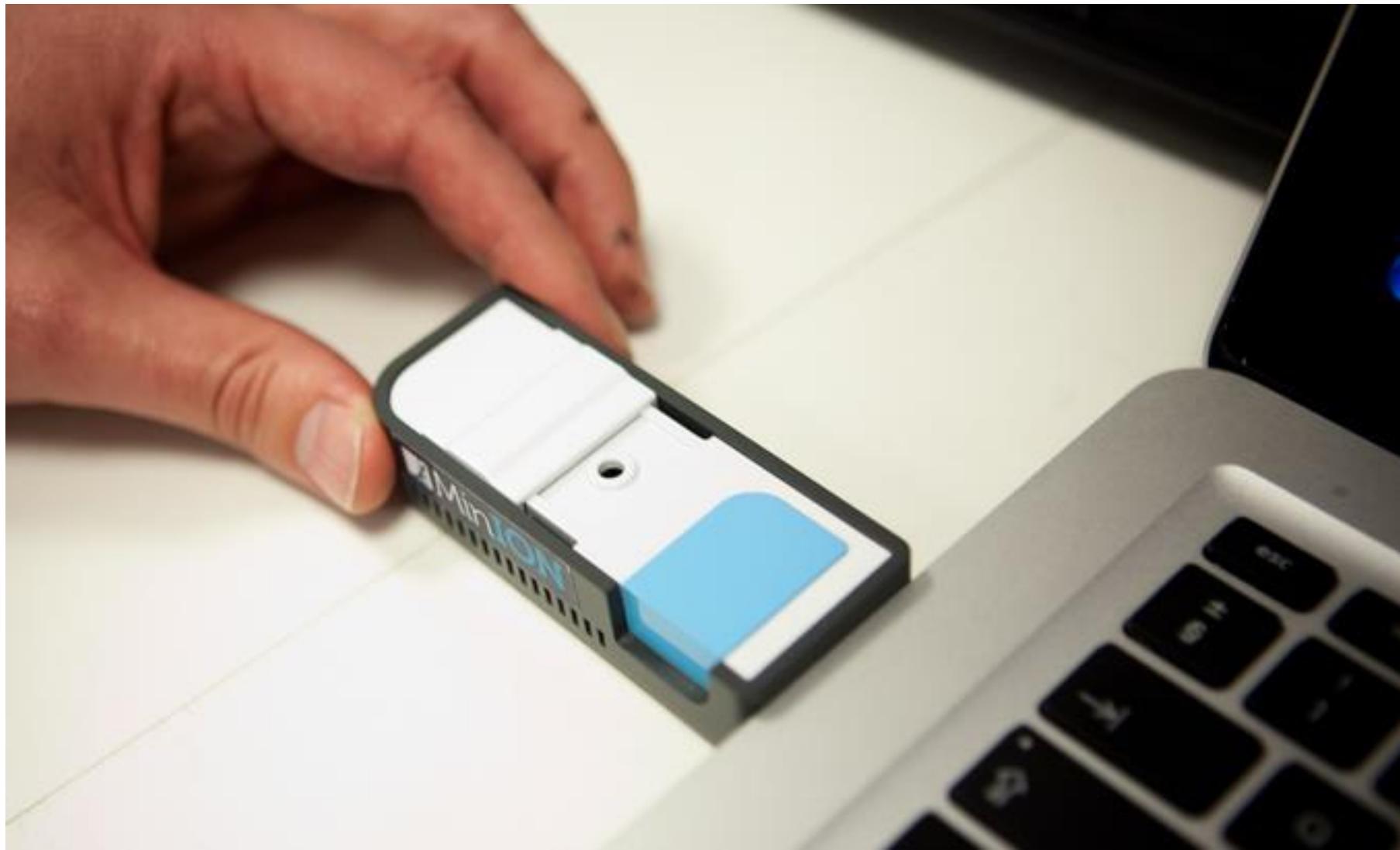
'Strand sequencing' is a technique that passes intact DNA polymers through a protein nanopore, sequencing in real time as the DNA translocates the pore.

Nanopore sequencing also offers, for the first time, direct RNA sequencing, as well as PCR or PCR-free cDNA sequencing.

<https://nanoporetech.com/applications/dna-nanopore-sequencing>

Feng et al , Gen Prot Bioinf 2015

# MinIon, OXFORD NANOPORE



<https://nanoporetech.com/news/movies#movie-24-nanopore-dna-sequencing>

# Oxford Nanopore Technologies, MinION



The MinION is a portable sequencer; flow cells contain up to 512 nanopore sensors.

The Oxford Nanopore system processes the reads that are presented to it rather than generating read lengths. Sample-prep dependent, the longest read reported by a MinION user to date is >1 Mb.

Long reads confer many advantages, including simpler assembly and in the analysis of repetitive regions, phasing or CNVs.

# Oxford Nanopore Technologies



**Flongle**

Long read, direct DNA/RNA/epigenetic sequencing, scalable, real time/rapid, on-demand sequencing that is easy to use and install.

- ✓ Your portable device for smaller, individual, rapid tests.
- ✓ When you don't want to multiplex samples or start a larger run.
- ✓ Amplicons, panels/targeted sequencing, quality testing and more.
- ✓ For use with MinIT or a laptop.

**MinION**

- ✓ Your personal sequencer, putting you in control.
- ✓ Whether in your lab or out in the field.
- ✓ Whole genomes/exomes, metagenomics, targeted sequencing, whole transcriptome (cDNA), smaller transcriptomes (direct RNA), multiplexing for smaller samples and more.
- ✓ For use with MinIT or a laptop.

**GridION**

- ✓ High throughput sequencing, in modular form (up to 5 flow cells) to be on-demand.
- ✓ For your lab or to offer as a service.
- ✓ Larger genomes or projects, whole transcriptomes (direct RNA or cDNA) or where you have larger numbers of samples and more.
- ✓ Compute included for real time data analysis and easy installation.

**PromethION**

- ✓ Very high throughput sequencing, in modular form (up to 48 flow cells) to be on-demand.
- ✓ For your lab or as a service.
- ✓ Larger genomes or projects, whole transcriptomes (direct RNA or cDNA), very large numbers of samples and more.
- ✓ Compute included for real time data analysis and easy installation.

# Characteristics, strengths and weaknesses of commonly used sequencing platforms

Table 2

Characteristics, strengths and weaknesses of commonly used sequencing platforms

Platform \ Instrument	Throughput range (Gb) <sup>a</sup>	Read length (bp)	Strength	Weakness
<i>Sanger sequencing</i>				
ABI 3500/3730	0.0003	Up to 1 kb	Read accuracy and length	Cost and throughput
<i>Illumina</i>				
MiniSeq	1.7–7.5	1×75 to ×150	Low initial investment	Run and read length
MiSeq	0.3–15	1×36 to 2×300	Read length, scalability	Run length
NextSeq	10–120	1×75 to 2×150	Throughput	Run and read length
HiSeq (2500)	10–1000	×50 to ×250	Read accuracy, throughput,	High initial investment, run
NovaSeq 5000/6000	2000–6000	2×50 to ×150	Read accuracy, throughput	High initial investment, run
<i>Ion Torrent</i>				
PGM	0.08–2	Up to 400	Read length, speed	Throughput, homopolymers <sup>c</sup>
S5	0.6–15	Up to 400	Read length, speed,	Homopolymers <sup>c</sup>
Proton	10–15	Up to 200	Speed, throughput	Homopolymers <sup>c</sup>
<i>Pacific BioSciences</i>				
PacBio RSII	0.5–1 <sup>b</sup>	Up to 60 kb	Read length, speed (Average 10 kb, N50 20 kb)	High error rate and initial
Sequel	5–10 <sup>b</sup>	Up to 60 kb	Read length, speed (Average 10 kb, N50 20 kb)	High error rate
<i>Oxford Nanopore</i>				
MinION	0.1–1	Up to 100 kb	Read length, portability	High error rate, run length,

<sup>a</sup> The throughput ranges are determined by available kits and run modes on a per run basis. As an example of a 15-GB throughput, thirty-five 5-MB genomes can be sequenced to a minimum coverage of 40× on the Illumina MiSeq using the v3 600 cycle chemistry.

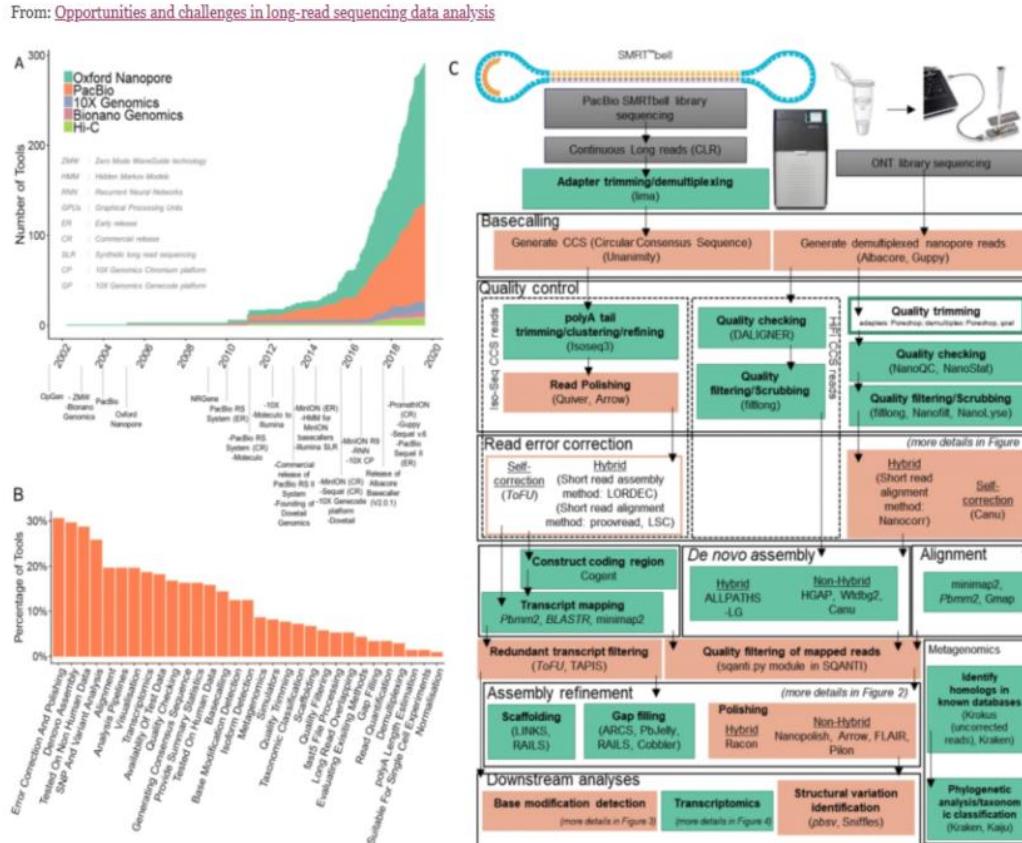
<sup>b</sup> Per one single-molecule real-time cell.

<sup>c</sup> Results in increased error rate (increased proportion of reads containing errors among all reads) which in turn results in false-positive variant calling.

Besser et al., Clin Micr Infect, 2018

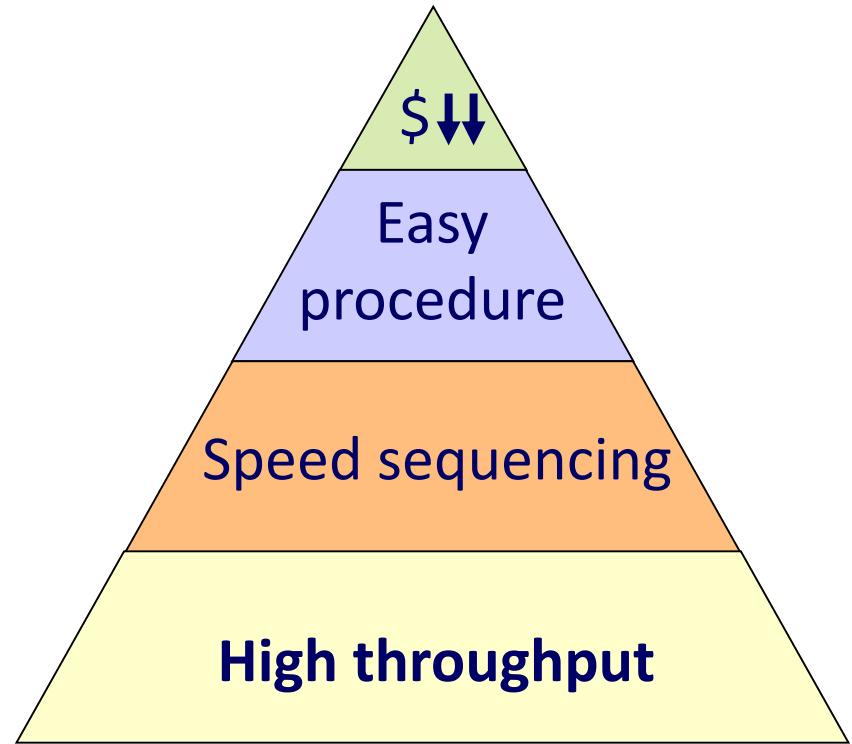
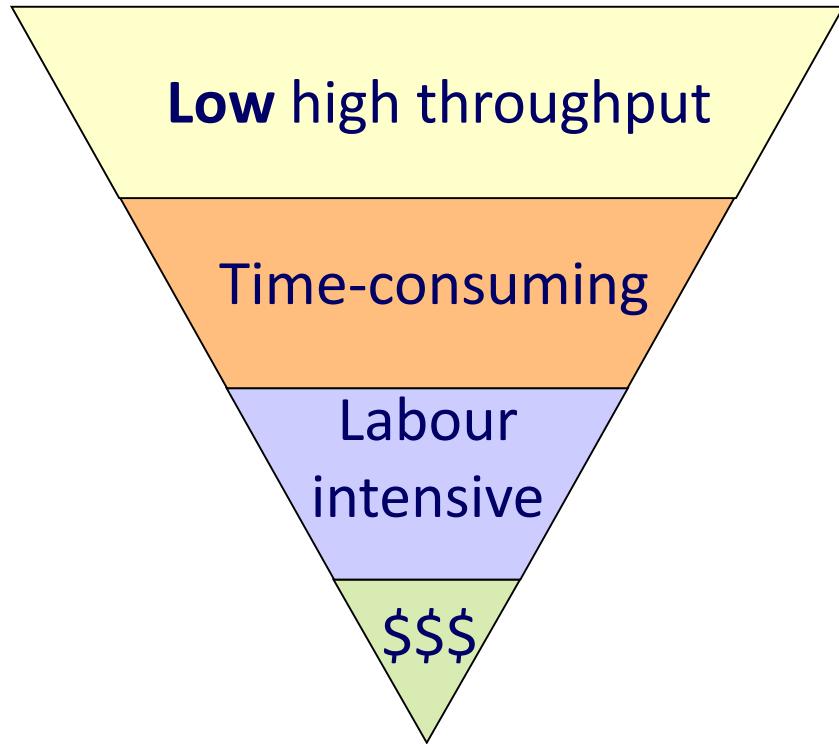
# Long-read sequencing data analysis

Amarasinghe et al., Genome Biology 2020, 21:30



Overview of long-read analysis tools and pipelines. **a** Release of tools identified from various sources and milestones of long-read sequencing. **b** Functional categories. **c** Typical long-read analysis pipelines for SMRT and nanopore data. Six main stages are identified through the presented workflow (i.e. basecalling, quality control, read error correction, assembly/alignment, assembly refinement, and downstream analyses). The green-coloured boxes represent processes common to both short-read and long-read analyses. The orange-coloured boxes represent the processes unique to long-read analyses. Unfilled boxes represent optional steps. Commonly used tools for each step in long-read analysis are within brackets. Italics signify tools developed by either PacBio or ONT companies, and non-italics signify tools developed by external parties. Arrows represent the direction of the workflow.

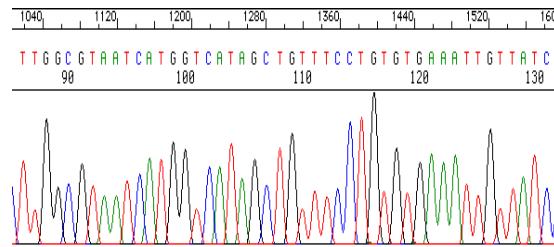
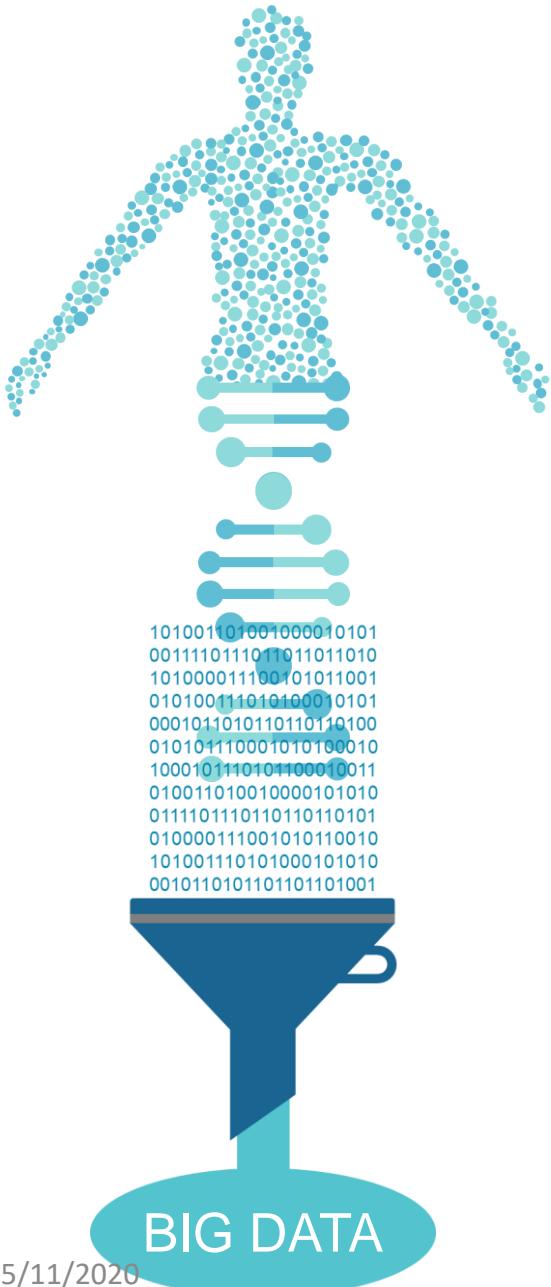
# Sanger vs SM, advantages of new technologies



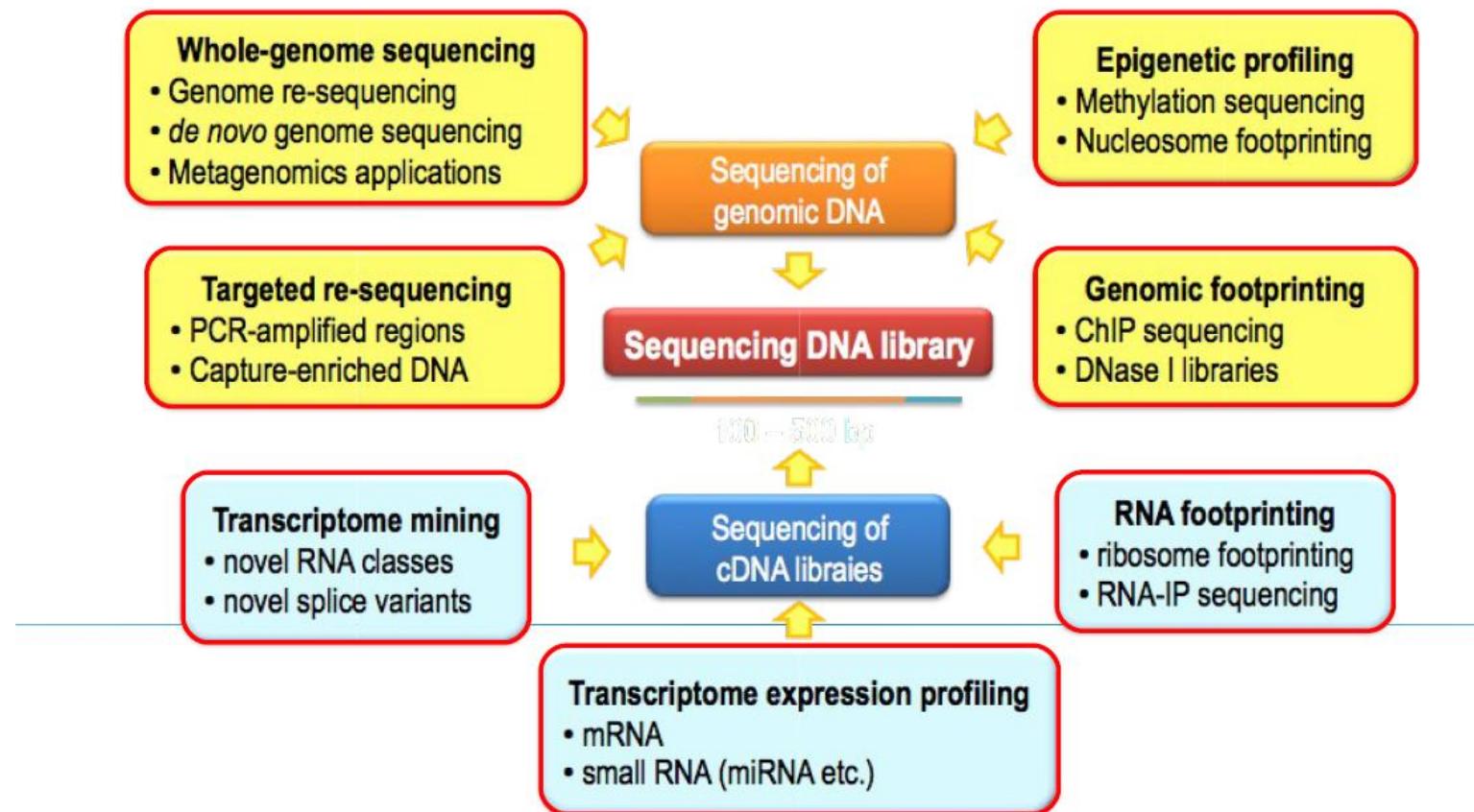
Semiautomatic **Sanger** capillary-based sequencing technology

NGS  
Next Generation Sequencing =  
Now Generation Sequencing

# BIG DATA



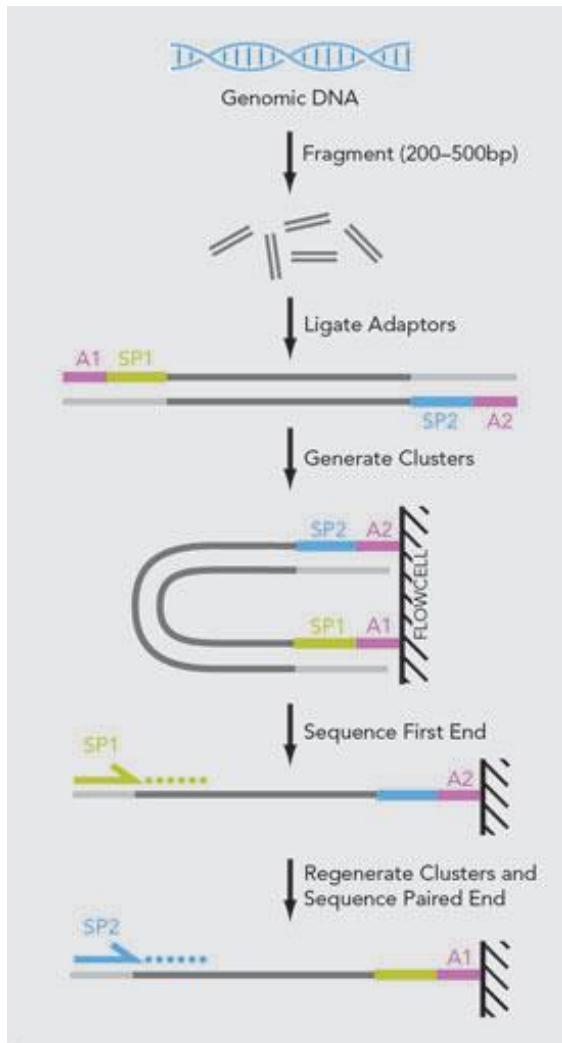
# Massive sequencing applications



# APLICACIONES DE NGS EN EL DIAGNÓSTICO VIROLÓGICO

- ◆ Detección de patógenos virales no identificados en la muestra
- ◆ Identificación de virus nuevos
- ◆ Identificación de virus en muestras tumorales
- ◆ Caracterización del Viroma de un organismo
- ◆ Secuenciación del genoma viral completo
- ◆ Estudio de la Variabilidad genómica (Quasispecies)
- ◆ Monitorización de la resistencia a los Antivirales
- ◆ Epidemiología de las infecciones virales
- ◆ Estudio de la Evolución viral
- ◆ Control de calidad de las vacunas virales vivas atenuadas

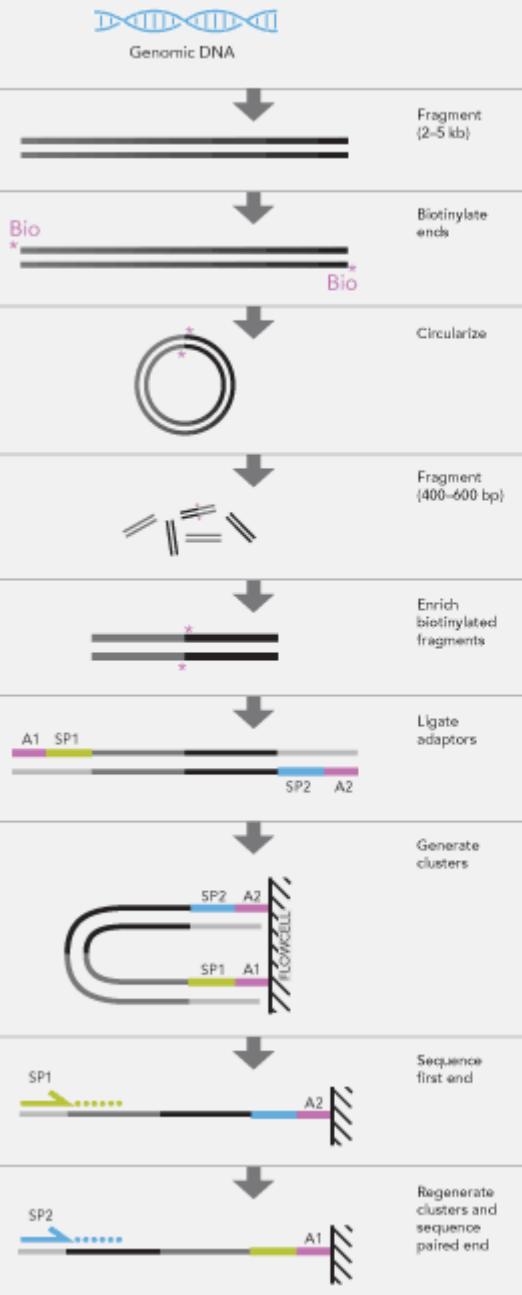
# Que es Pair-end?



**Secuenciación de un fragmento (bp)**

**Modificación de single-read DNA,  
Leyendo por ambos extremos, forward y reverse**

## Mate Pair Library Sequencing for Long Inserts



Mate Pair library preparation is designed to generate short fragments that consist of two segments that originally had a separation of several kilobases in the genome. Fragments of sample genomic DNA are end-biotinylated to tag the eventual mate pair segments. Self-circularization and refragmentation of these large fragments generates a population of small fragments, some of which contain both mate pair segments with no intervening sequence. These Mate Pair fragments are enriched using their biotin tag. Mate Pairs are sequenced using a similar two-adapter strategy as described for paired-end sequencing.

# Que es Mate-pair?

**Secuenciación de dos fragmentos separados kb.**

**Util:**

**Secuenciación de un Genoma de novo  
Finalizar un genoma  
Detección de variantes estructurales**

# Sequencing terms

## Breadth of coverage

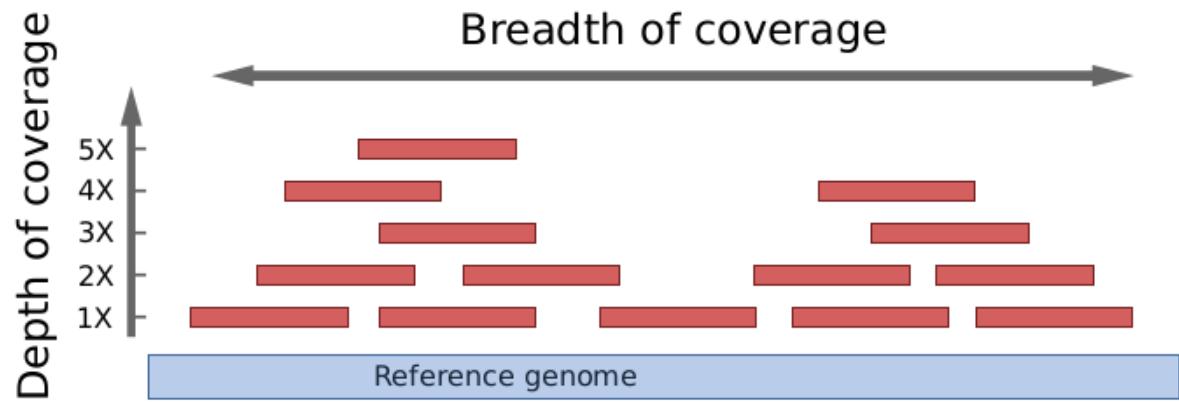
How much of a genome is "covered" by short reads? Are there regions that are not covered, even not by a single read?

Breadth of coverage is the percentage of bases of a reference genome that are covered with a certain depth. For example: 90% of a genome is covered at 1X depth; and still 70% is covered at 5X depth.

## Depth of coverage

How strong is a genome "covered" by sequenced fragments (short reads)?

Per-base coverage is the average number of times a base of a genome is sequenced. The coverage depth of a genome is calculated as the number of bases of all short reads that match a genome divided by the length of this genome. It is often expressed as 1X, 2X, 3X,... (1, 2, or, 3 times coverage).



[www.metagenomics.wiki](http://www.metagenomics.wiki)

# Calculo de cobertura: número de lecturas

Total output required = region size \* coverage / ((1-duplicates/100) \* on target/100)

### Sequencing Coverage Calculator

Support Center:  
Sequencing Coverage Calculator

Application or product: Whole-Genome Sequencing

Coverage: 100 x  
Duplicates: 2 %

Genome or region size (in million bases): 3300 Mb  
Total read length (e.g. 200 for 2x100): 600 cycles

Benchtop Sequencers      Production-Scale Sequencers

iSeq       NextSeq 500/550  
 MiSeq       NovaSeq 6000  
 MiSeq / MiSeq Dx in RUO mode       HiSeq 3000/4000  
 NextSeq 500/550       HiSeq 1500/2500 Rapid Run  
 MiSeq 1500/2500 High Output       NextSeq 1000 Sequencing System  
 NextSeq 2000 Sequencing System

Support Center:  
Sequencing Coverage Calculator

Thank you for using the Illumina coverage estimator.

The results were calculated based on: **coverage needed**. Explain the estimations

Application or product:	Whole-Genome Sequencing
Genome or region size:	3300 Mbases
Read length:	600
Coverage:	100x
Duplicates:	2%
Output Required:	336,734,693,878 bases

Run type	MiSeq	MiSeq	MiSeq	MiSeq
v3 Reagents	v2 Reagents	v2 Nano Reagents	v2 Micro Reagents	
Clusters	25,000,000 per flow cell	15,000,000 per flow cell	1,000,000 per flow cell	4,000,000 per flow cell
Output per unit (flow cell or lane)	15,000,000,000 per flow cell	9,000,000,000 per flow cell	600,000,000 per flow cell	2,400,000,000 per flow cell
Exceeds maximum read length?	Does not exceed maximum (2x300)	Read length exceeds maximum of 2x250	Read length exceeds maximum of 2x250	Read length exceeds maximum of 2x150
Number of units per sample (flow cell or lane)	22,449 flow cells	37,415 flow cells	561,224 flow cells	140,306 flow cells
Samples per unit (flow cell or lane)	-Offlow cell	-Offlow cell	-Offlow cell	-Offlow cell
Comments	Upgraded software: MCS v2.3 or later; MiSeq	Upgraded hardware or from September 2012 and later; MCS v2.0 or later; Reagent Kit v3 (150/600); MiSeq Reagent Kit v2 (50/300/500)	Upgraded hardware or from September 2012 and later; MCS v2.0 or later; MiSeq Reagent Nano Kit v2 (300/500)	Upgraded hardware or from September 2012 and later; MCS v2.0 or later; MiSeq Reagent Micro Kit v2 (300)
Products	MiSeq Reagent Kit v3	MiSeq Reagent Kits v2	MiSeq Reagent Kits v2	MiSeq Reagent Kits v2

Get the results in a comma-separated values (CSV) report:

[https://emea.support.illumina.com/downloads/sequencing\\_coverage\\_calculator.html](https://emea.support.illumina.com/downloads/sequencing_coverage_calculator.html)

# PREPARACIÓN LIBRERÍA, estrategias

## SECUENCIACIÓN GENOMA, EXOMA, TRANSCRIPTOMA

1. Sin amplificación
2. Amplificación con PCR
3. Sondas captura

- Tamaño de fragmento
- Longitud de la lectura
- Single o Pair-end
- Número de bases por muestra
- Profundidad de cobertura x

## SECUENCIACIÓN GENOMAS

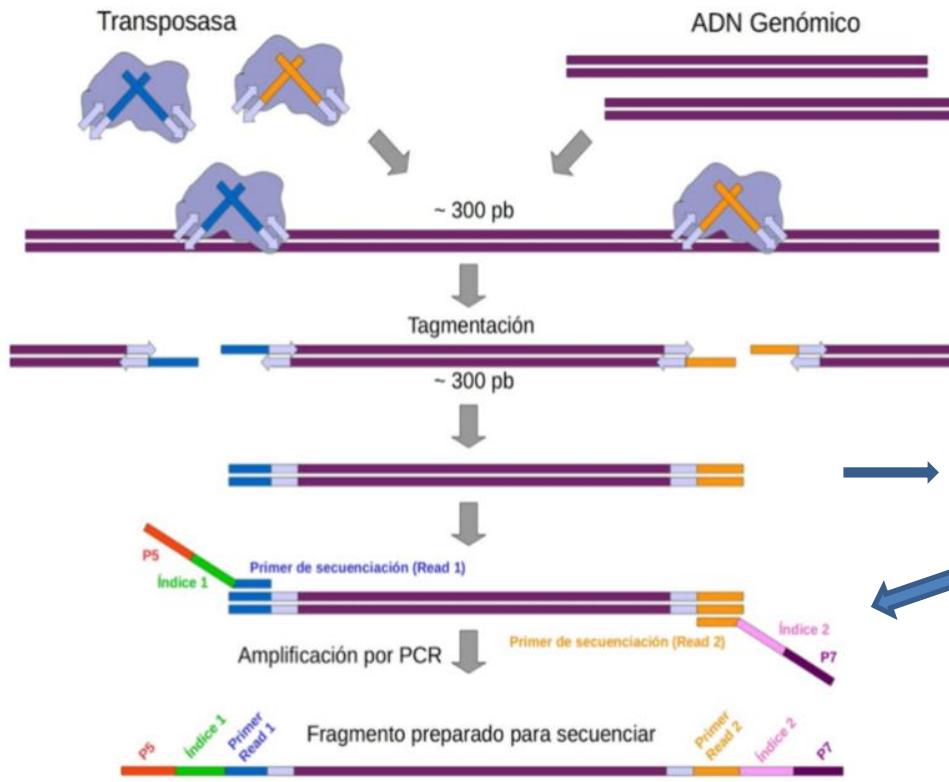
1. Metagenómica

## IDENTIFICACIÓN MICROORGANISMOS

1. Metataxonomía

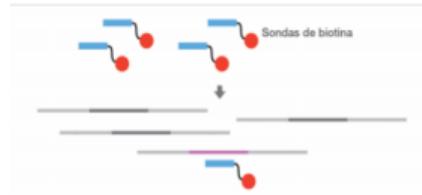
# PREPARACIÓN LIBRERÍA

## ENZIMÁTICA FÍSICA



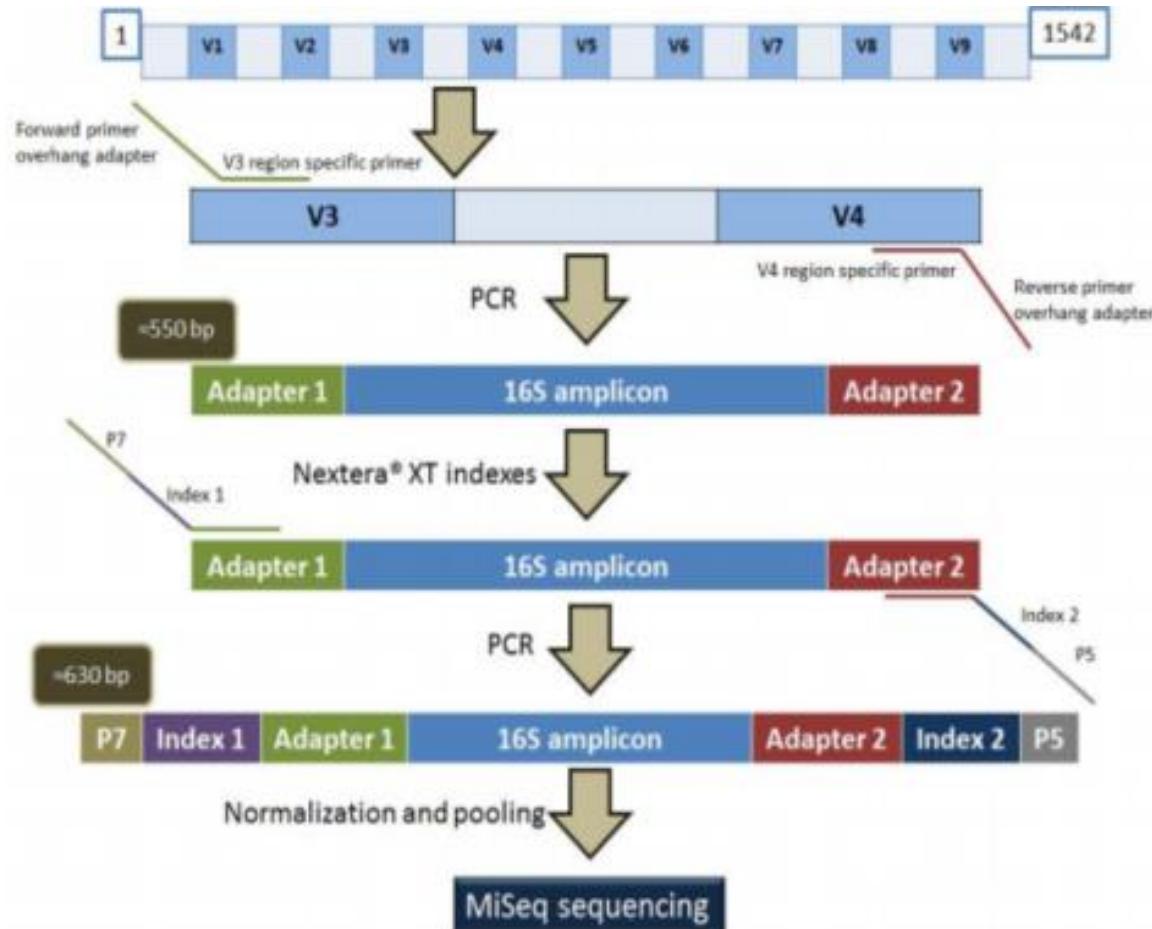
RNA → cDNA

## ENRIQUECIMIENTO: PCR CAPTURA SONDAS



Guia Práctica Genómica [https://www.uv.es/varnau/GM\\_Cap%C3%ADtulo\\_2.pdf](https://www.uv.es/varnau/GM_Cap%C3%ADtulo_2.pdf)

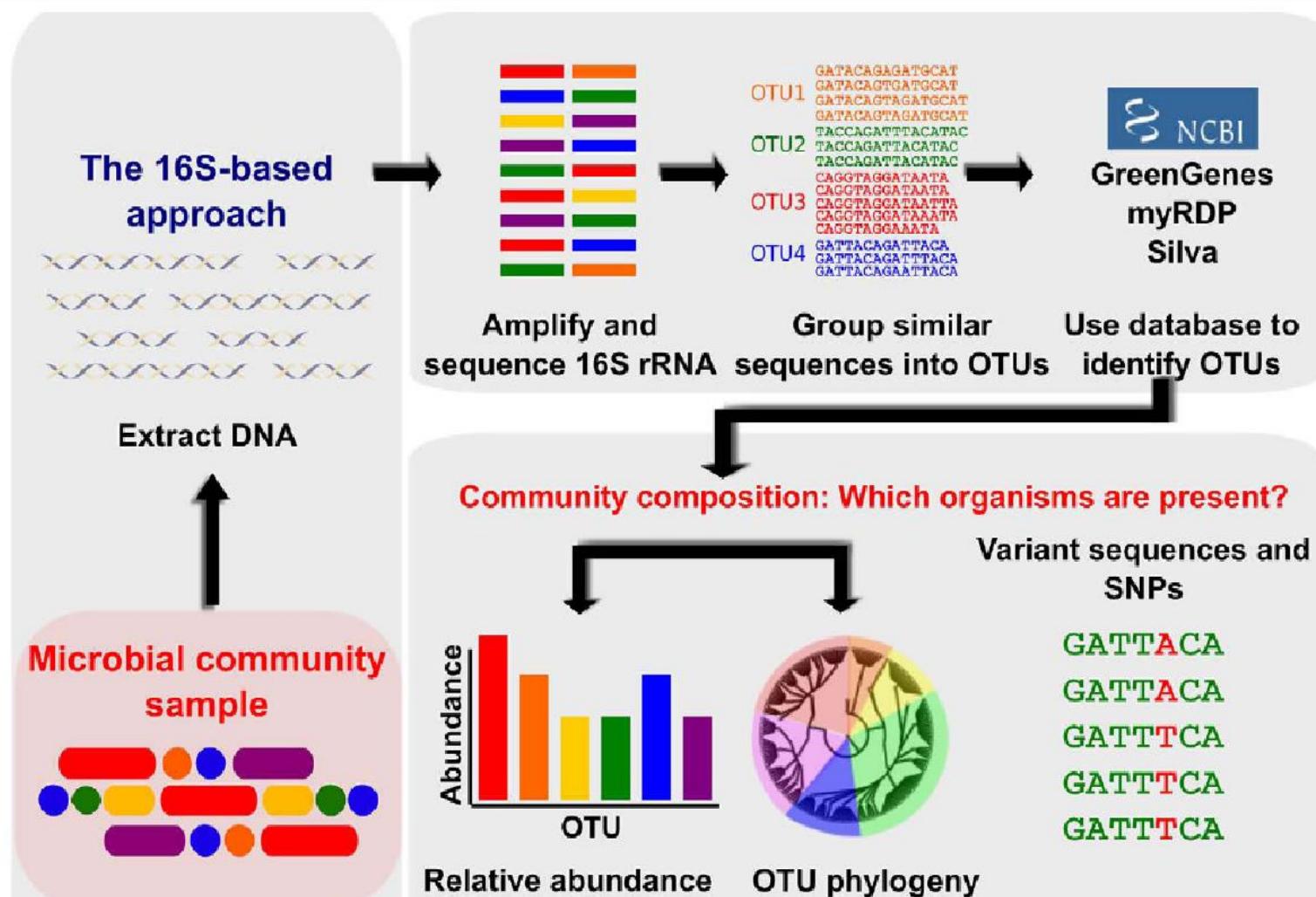
# PREPARACIÓN LIBRERÍA, rRNA 16S, caracterización microbiota



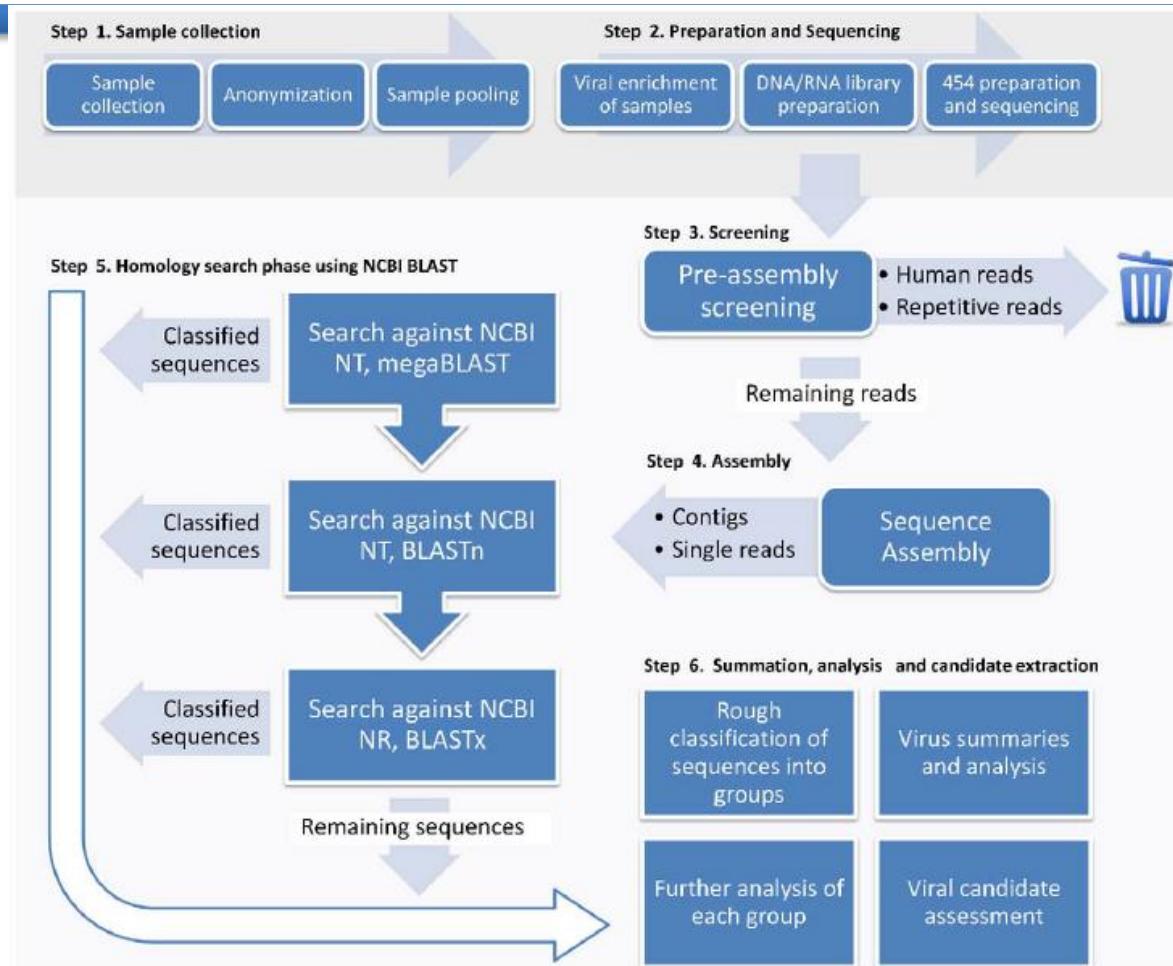
# Metataxonomics vs Metagenomics (16S vs Shotgun)

	Metagenetics	Metagenomics
<b>Amplified sequence</b>	Marker regions	Whole genome
<b>Computing time</b>	Usually short	Usually long
<b>Taxonomic composition</b>	Yes	Yes
<b>New pathogen detection</b>	No	Yes
<b>Genome coverage information</b>	No	Yes

# Metataxonomics



# Metagenómica, pipeline de análisis



Lysholm et al., Plos One 2012:7,2, e30875

# Qué es la Bioinformática?

**PROBLEMAS  
BIOLÓGICOS**



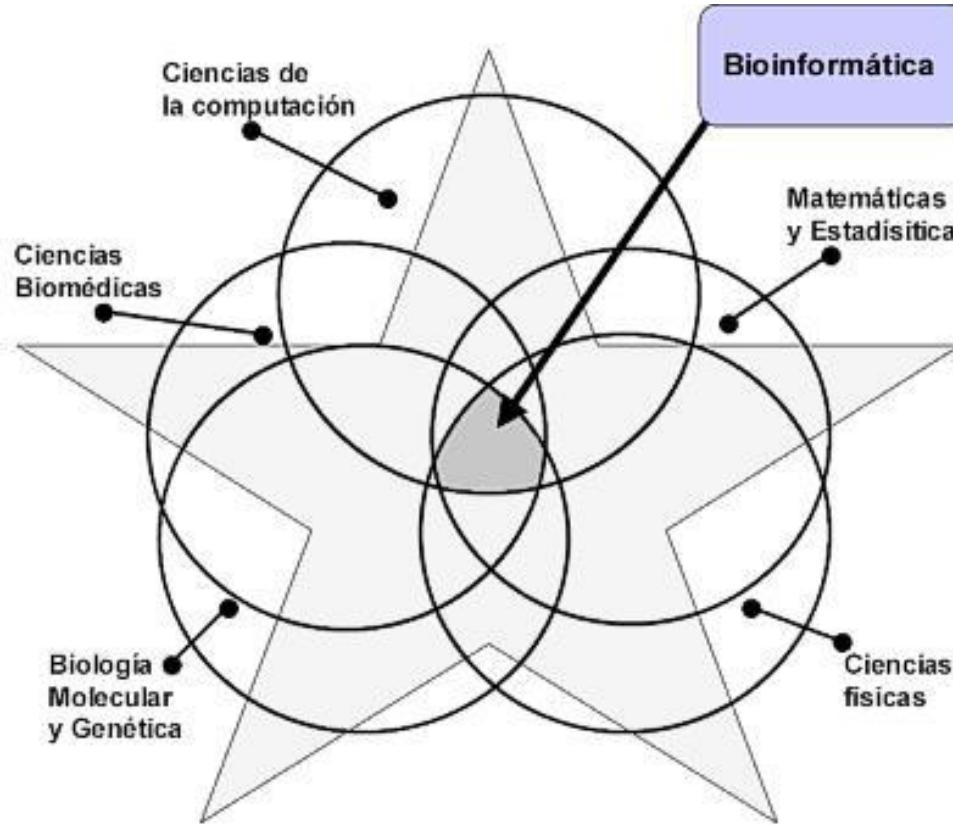
**Procesamiento  
de datos**



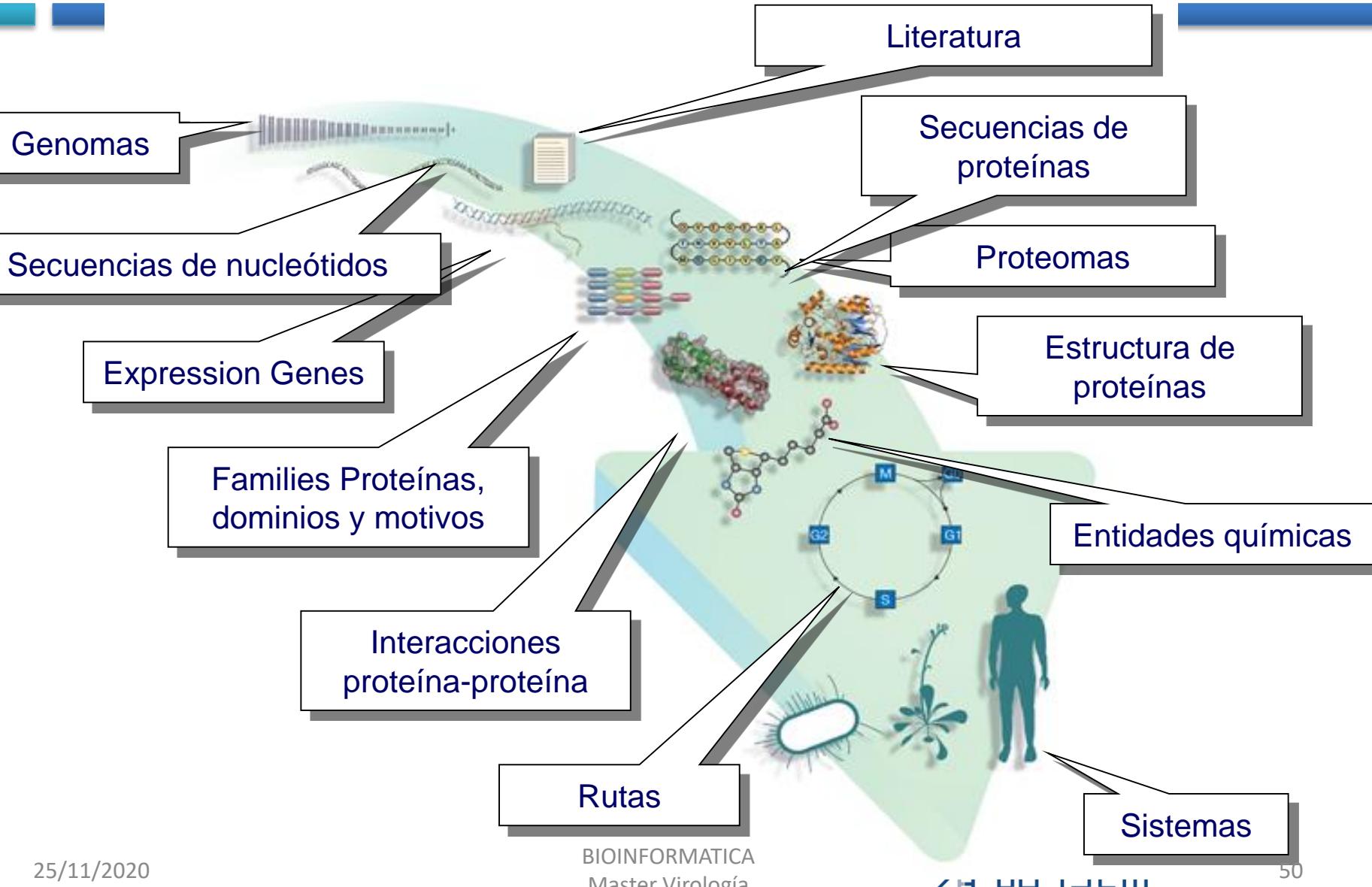
**MÉTODOS  
COMPUTACIONALES**



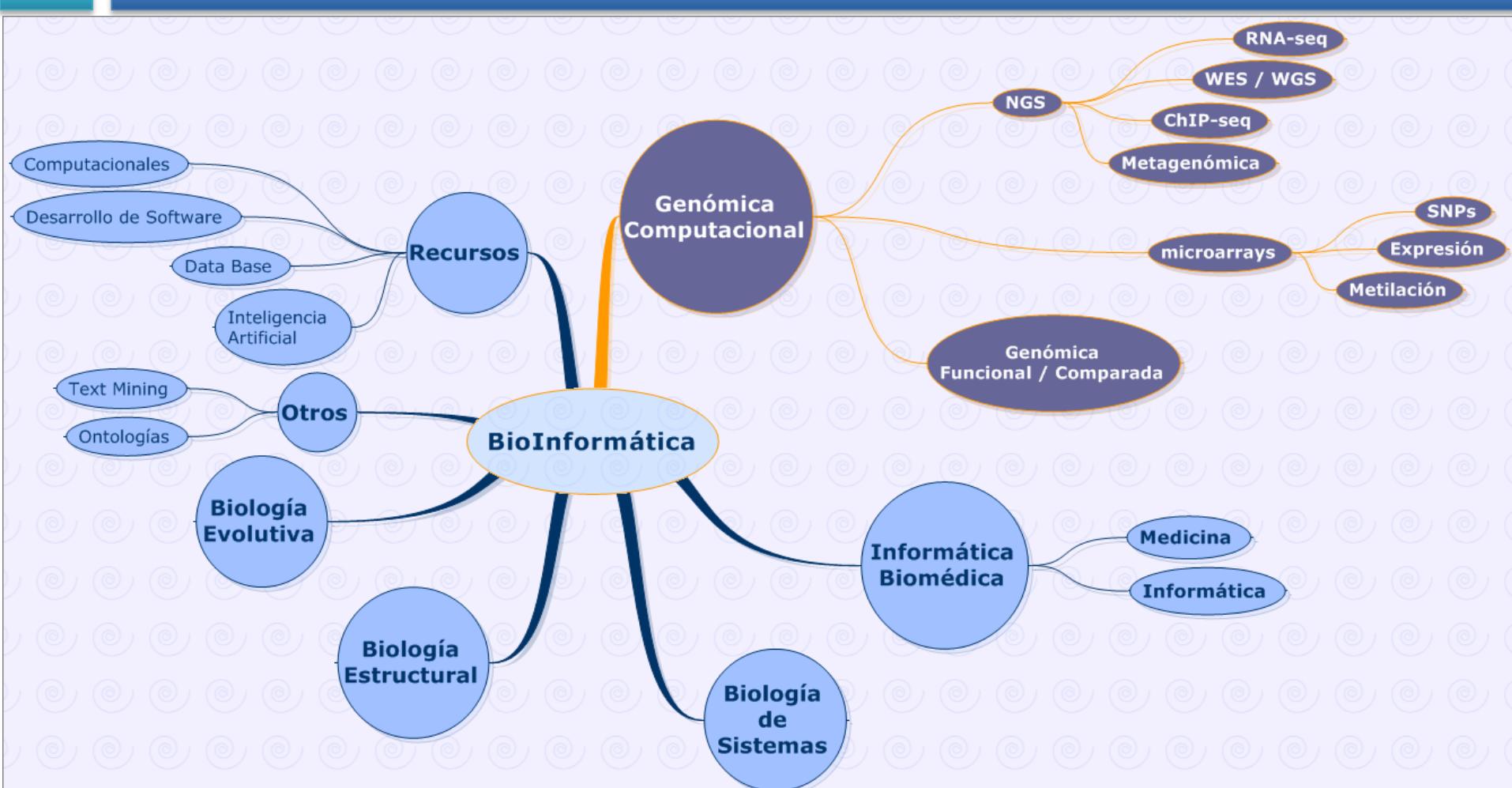
# Bioinformática es multidisciplinaria



# Tipos de datos dan idea de la dimensión de la Bioinformática



# AMBITO DE LA BIOINFORMÁTICA

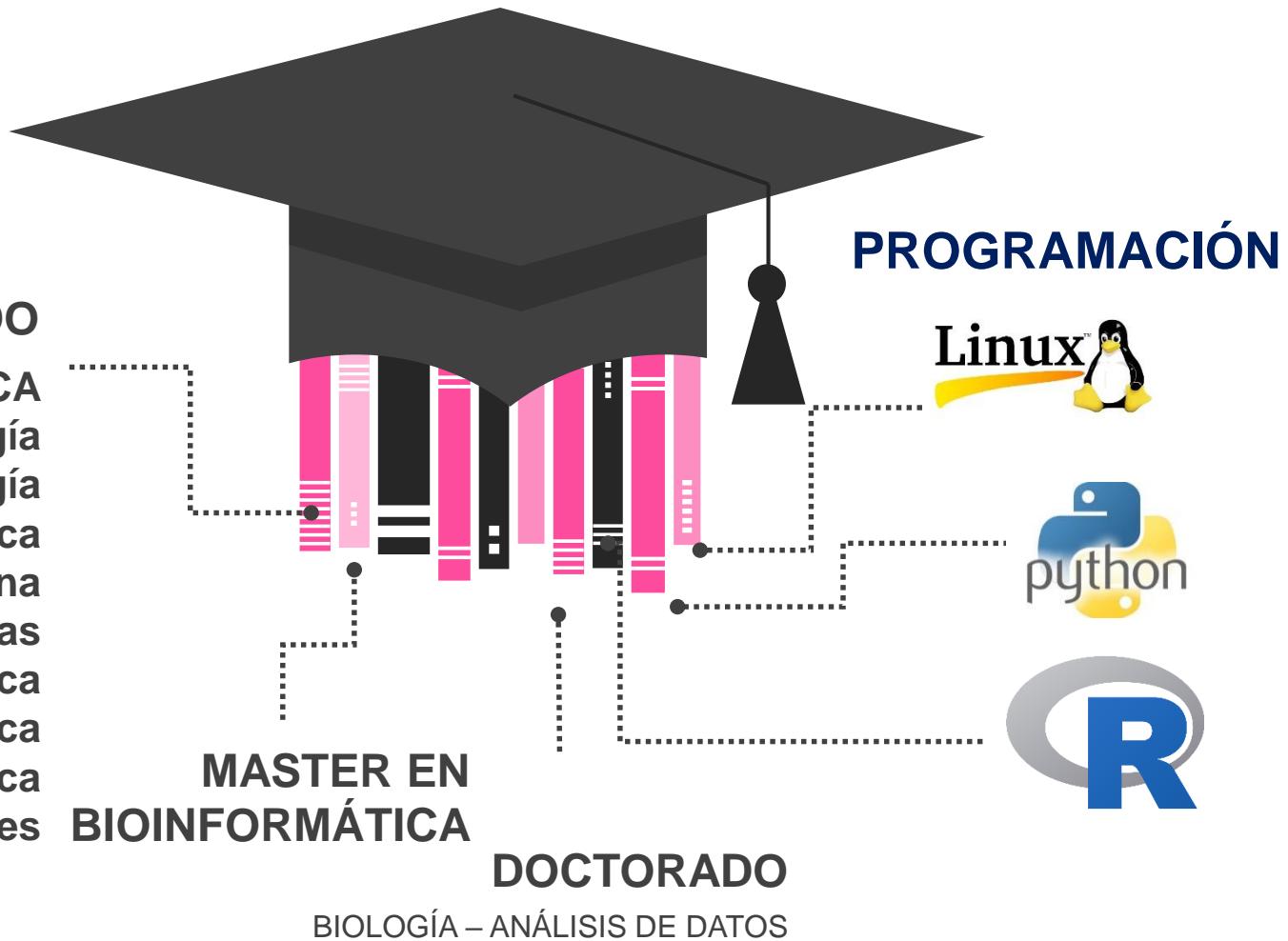


# FORMACIÓN EN BIOINFORMÁTICA

Universidad  
Barcelona.

## GRADO

BIOINFORMÁTICA  
Biología  
Biotecnología  
Bioquímica  
Medicina  
Matemáticas  
Química  
Física  
Informática  
Telecomunicaciones



# ¿Dónde trabaja un Bioinformático?



**UNIVERSIDAD**  
Biociencias  
Informática

**CENTRO DE INVESTIGACIÓN**

Biomedicina  
Agricultura  
Alimentación



**EMPRESA**  
Bioinformática  
Genética  
Genómica

**HOSPITAL BIOINFORMÁTICO CLÍNICO**  
Genética  
Oncología  
Cardiología

Thanks for your attention!

Questions???