

## Introducción al análisis de datos de secuenciación masiva y sus aplicaciones en microbiología

**Isabel Cuesta**

7 Abril 2022

UAH - ISCIII

BU-ISCIII

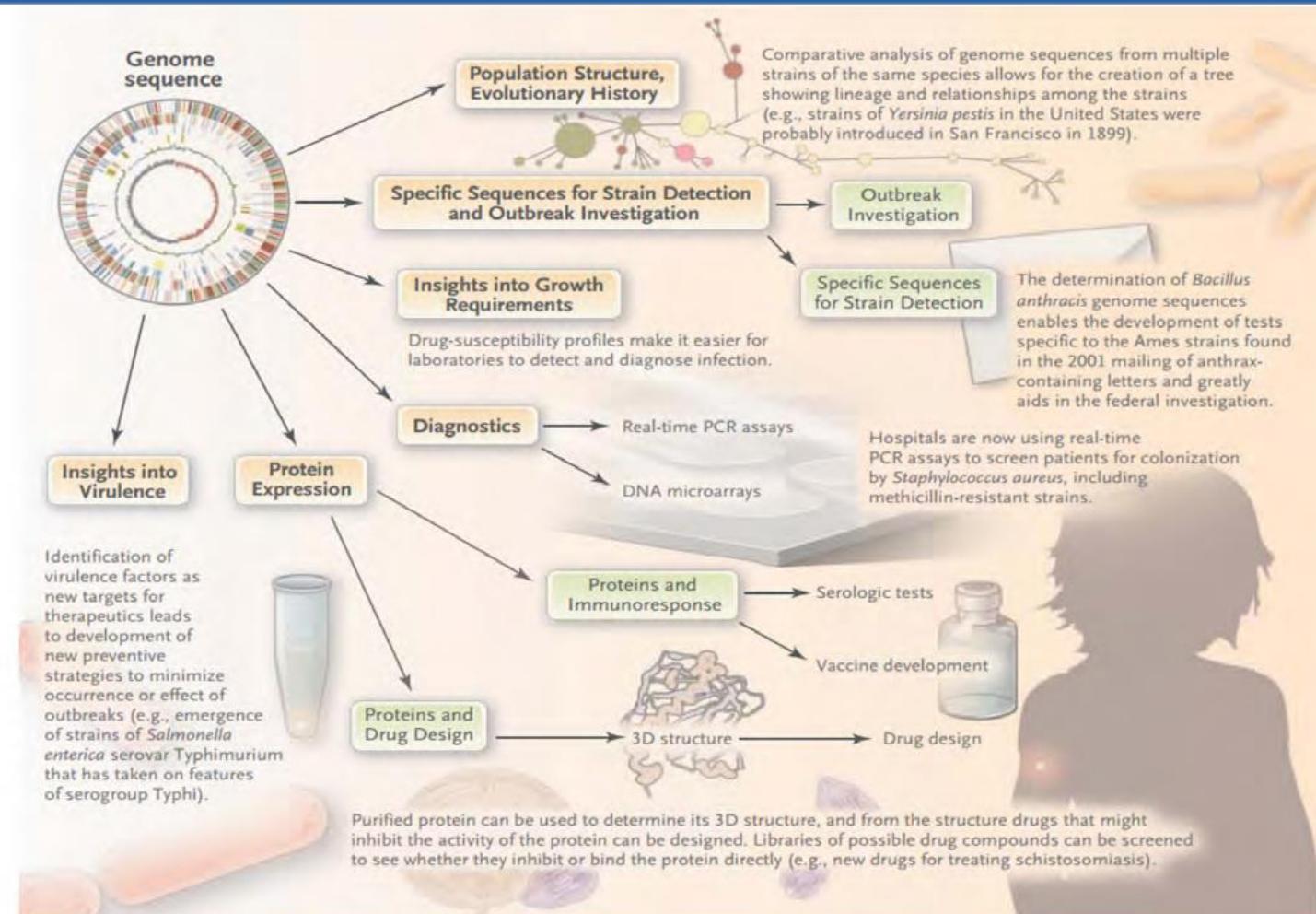
Unidades Centrales Científico Técnicas - SGSAFI-ISCIII

## Index

- Conocer los fundamentos de la secuenciación masiva.
- Aprender a usar de Galaxy, herramienta web que permite el manejo y análisis de datos procedentes de técnicas de secuenciación masiva.
- Conocer los ficheros generados por plataformas como Illumina, y evaluar su calidad.
- Reconstruir la secuencia consenso del genoma de SARS-CoV-2 e identificar las mutaciones y variantes asociadas.
- Ensamblar genomas secuenciados con plataforma Illumina y analizar la calidad del ensamblado.

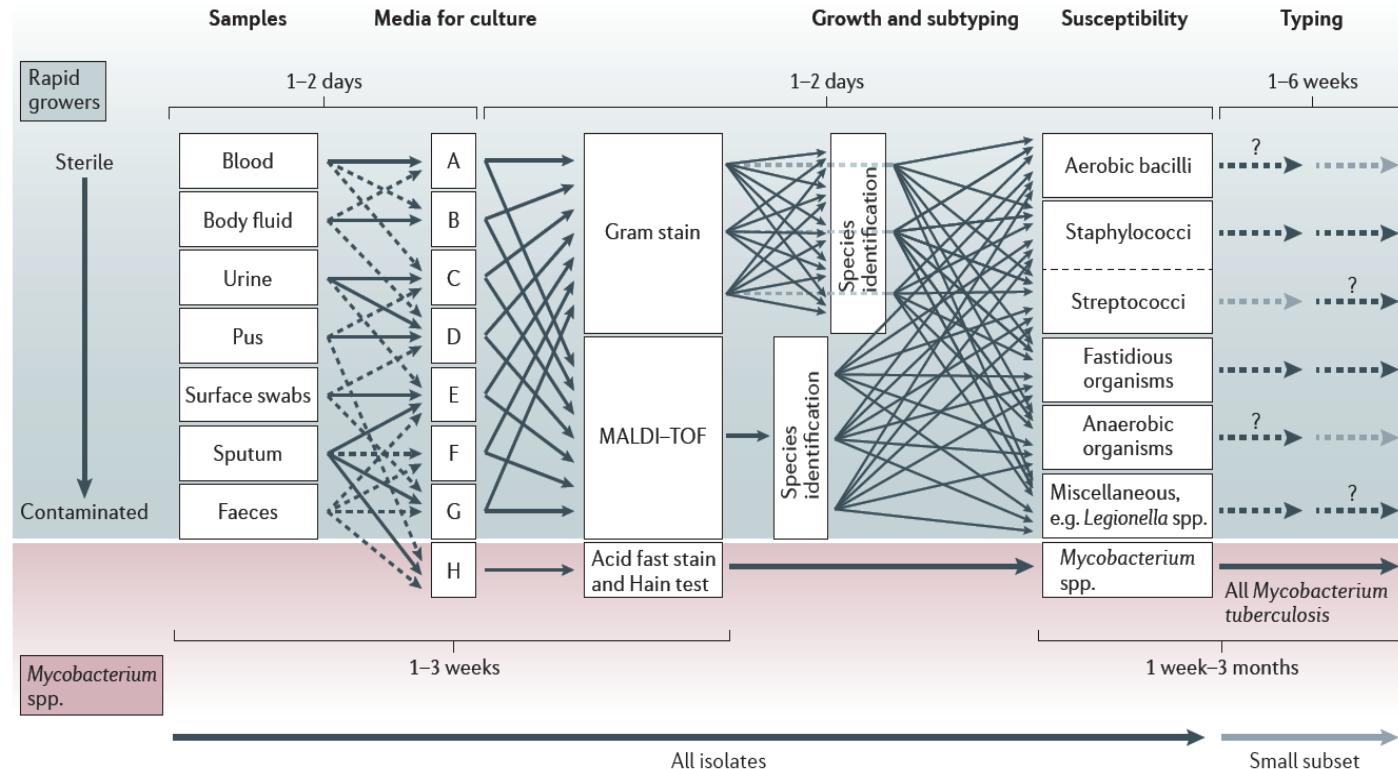
# Use of microbial genomics for tool development

Report from The American Academy of Microbiology, 2015



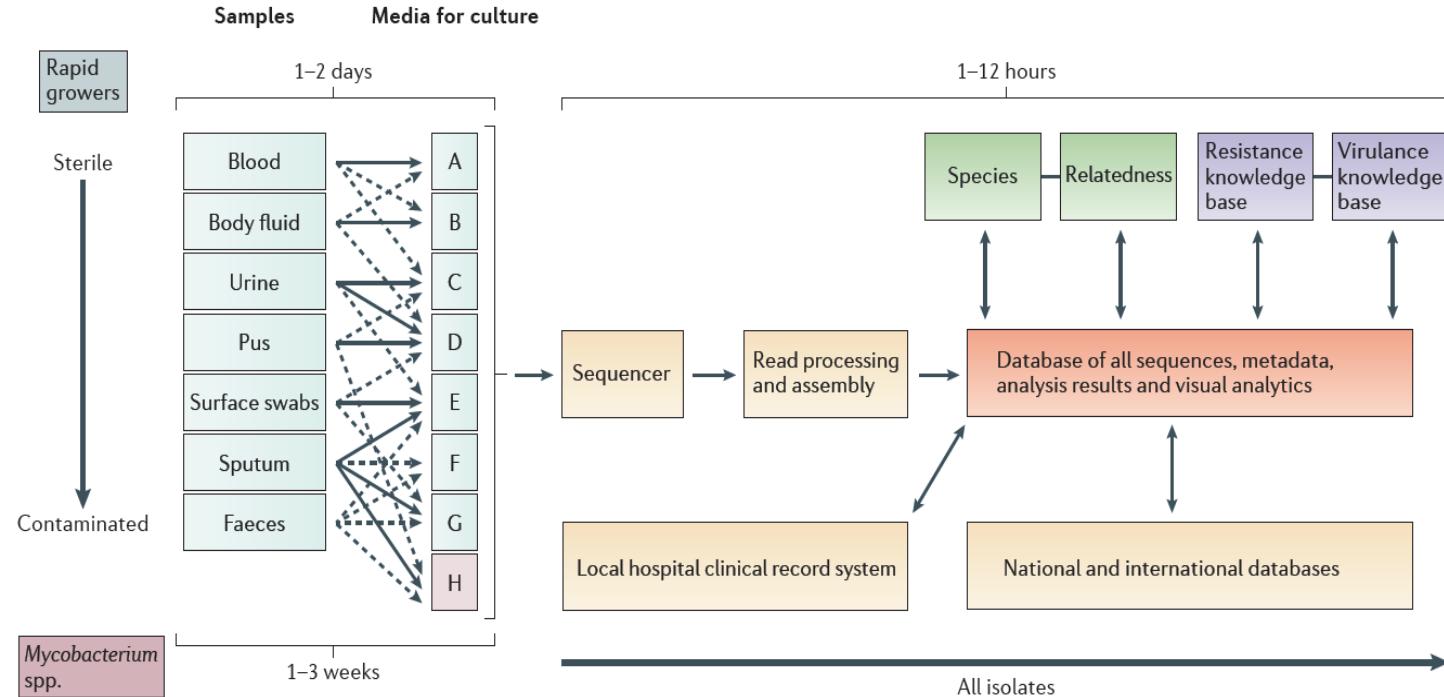
# Workflow for processing samples for bacterial pathogens

Didelot et al., Nature Genet Review 2012, 13:601-612



Ongoing developments in DNA-sequencing technologies are likely to affect the diagnosis and monitoring of all pathogens, including viruses, bacteria, fungi and parasites.

# The diagnostic and clinical applications of bacterial WGS



Didelot et al., Nature Genet Review 2012, 13:601-612

## Foodborne outbreak identification “Crisis del pepino”

2011

- Mayo 24 Primera muerte en Alemania
- 26 Alemania acusa a los pepinos españoles
- 30 Prohibición de importaciones de verduras de España y Alemania
- 31 Laboratorios alemanes desmienten oficialmente que los pepinos españoles sean el foco de infección
- Junio 10 Resolución de la crisis

Causado por la toxi-infección de Escherichia coli enterohemorrágica (EHEC) (*Escherichia coli* O104:H4)

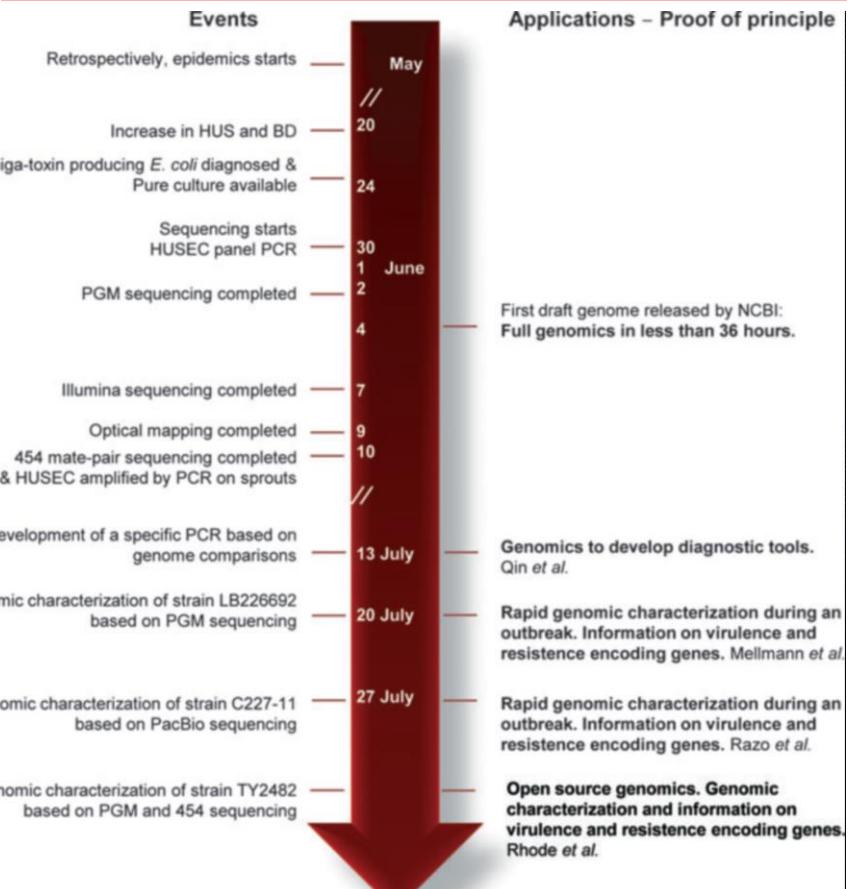
Muerte: 32 personas en Alemania, 1 Suecia y 1 Francia y 2263 infectados en 12 países de Europa.

Crisis Política y Económica Europa:  
Alto impacto en la Economía Europea, mayor afectación en la Española

**Secuenciación Genoma**



## The *Escherichia coli* O104:H4 epidemics: event timeline and major outputs



# Andalusian Listeria Outbreak

**Actualización de información sobre el brote de intoxicación alimentaria causado por *Listeria monocytogenes*.**

Publica: Agencia Española Seguridad alimentaria y Nutrición  
Fecha: 29 agosto 2019  
Sección: Seguridad Alimentaria

Jueves 29 de agosto de 2019, 12.00 horas

## ACTUALIZACIÓN EN RELACIÓN CON LA DISTRIBUCIÓN DE PRODUCTOS RELACIONADOS CON LA ALERTA.

La Agencia Española de Seguridad Alimentaria y Nutrición (AESAN) recomienda a las personas que tengan en su domicilio algún producto de la marca "La Mechá" se abstengan de consumirlo. Si se dispone del producto se debe devolver al punto de compra y, de no ser posible, desecharlo.

## Brote de listeriosis: sube el número de afectados y se apunta a la falta de higiene en la carne como causa

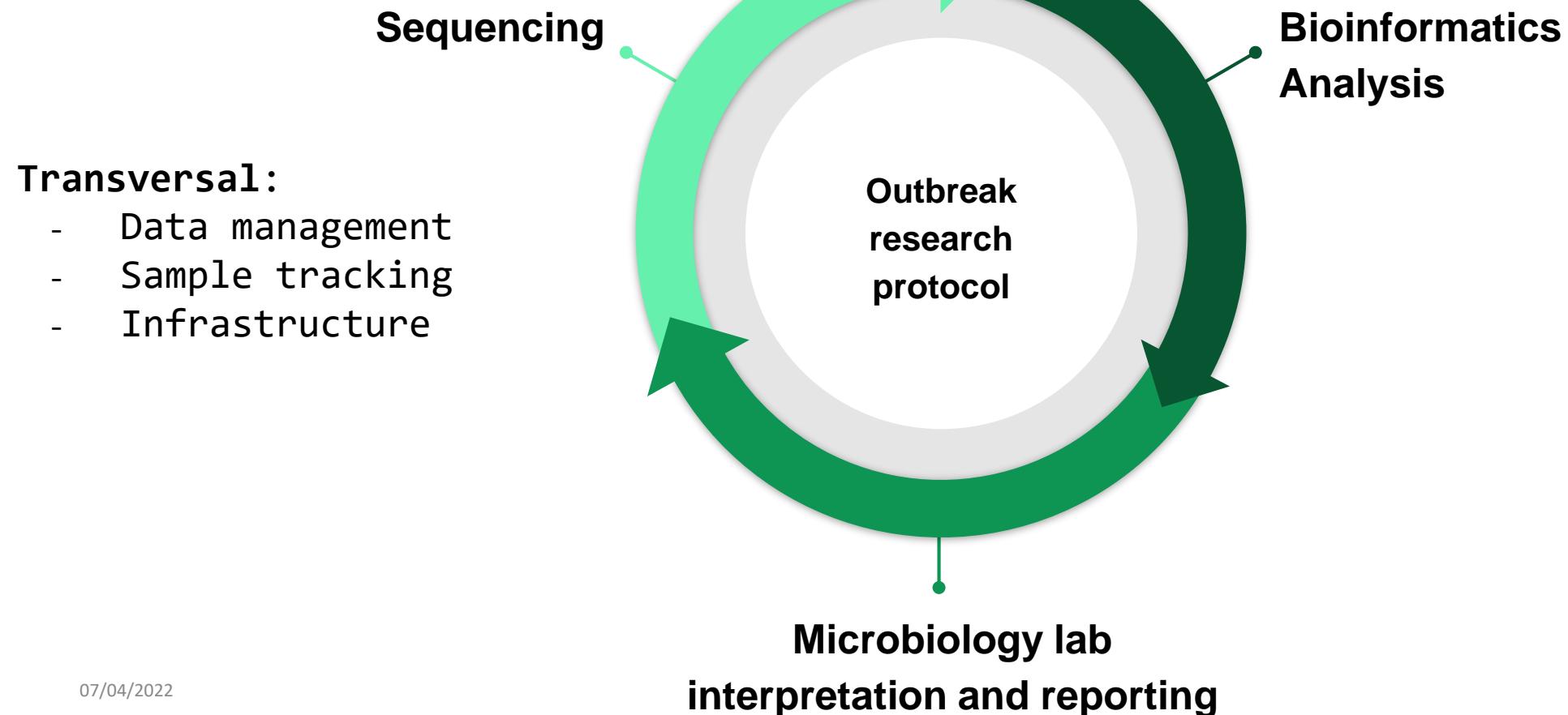
EFE 25.08.2019

- Tres nuevos casos, en Sevilla y Cádiz, dejan el número de personas afectadas en Andalucía en 192.
- [La carne con listeria de la marca blanca se vendió en los municipios de Sevilla.](#)
- La empresa que vendió la marca blanca de Magrudis dice que cumple los protocolos.



- Meat “La Mechá”. Margulis S.L.
- 250 cases related.
- Meat “"La Montanera del Sur". INCARYBE S.L”, suspicion. (Cádiz)
- Meat “Sabores de Paterna” (Málaga)

# Andalusian Listeria Outbreak

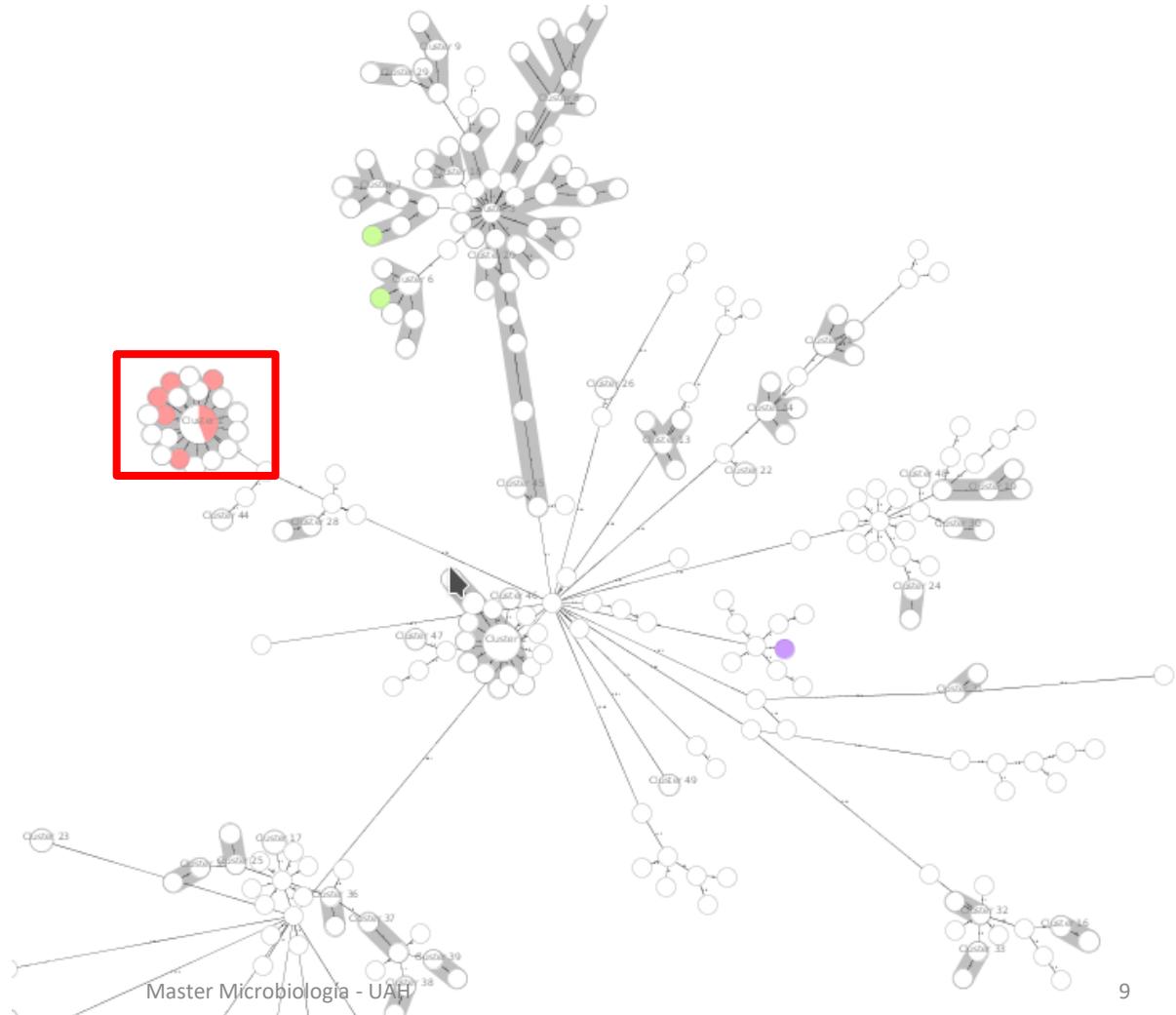


# Andalusian Listeria Outbreak

- 625 listeria samples already sequenced
  - 258 suspected to be related to the outbreak (mid august to mid september)

## Results:

- 233 related to the outbreak, confirmed to be caused by the meat “La Mechá”
  - 25 sporadic cases not related to the outbreak.



# Pathogen discovery: new virus – SARS-CoV-2

## Deep Meta-Transcriptomic Sequencing

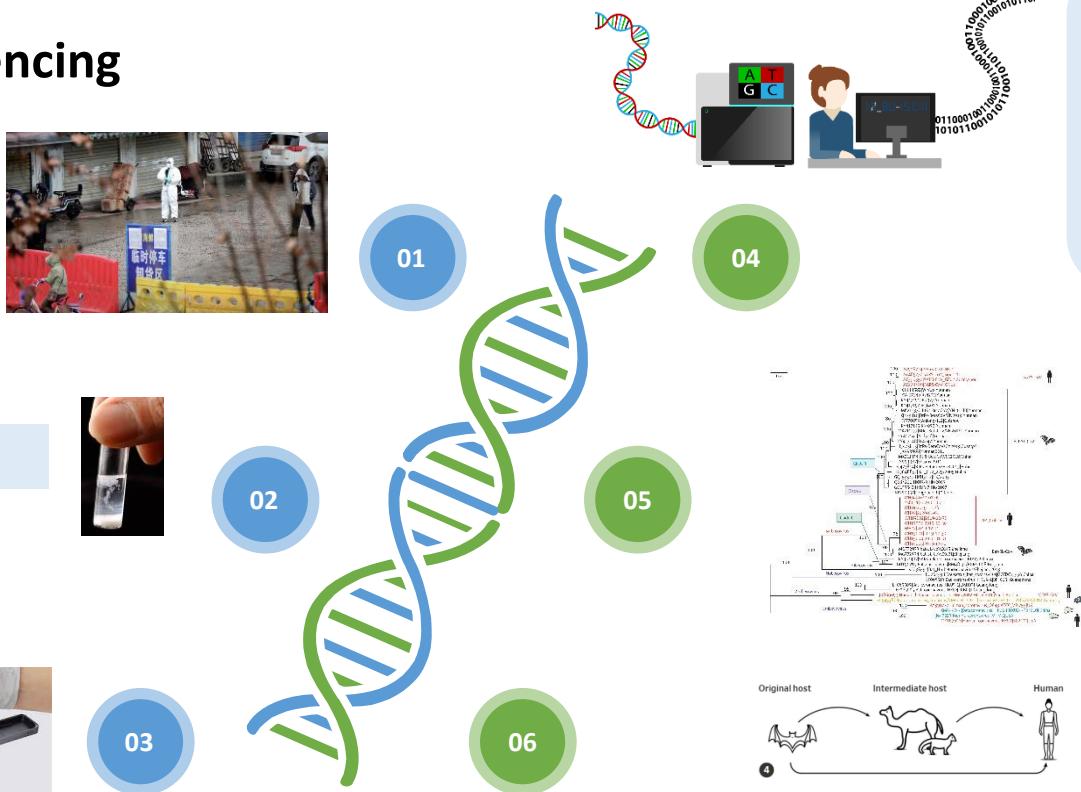
bronchoalveolar lavage fluid (BALF)



Meta-transcriptomic library

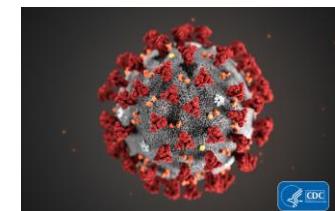
2x150 MiniSeq

56,565,928 sequences reads



De novo-assembled - Megahit  
384,096 Contigs  
Screened for potential aetiological agents  
The longest 30,474 nt

89.1% identity  
Closely related to a bat SARS-like coronavirus



Wu et al., Nature 2020

# One Health approach, infectious diseases could be better controlled and prevented



# Spanish National Microbiology Center (CNM)

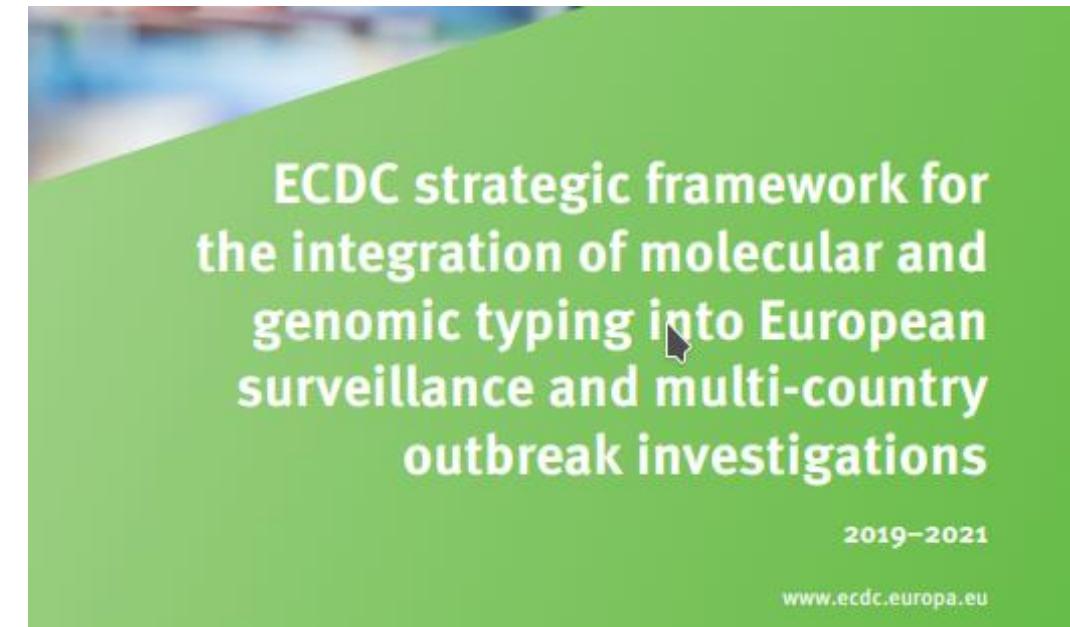


Mission: Provide support to the National Health System and the different Spanish Regions in the diagnosis and control of infectious diseases. In order to fulfill this mission it acts as Reference center offering a series of scientific activities:

- Diagnosis
- **Surveillance** →
- Infectious diseases research
- Training

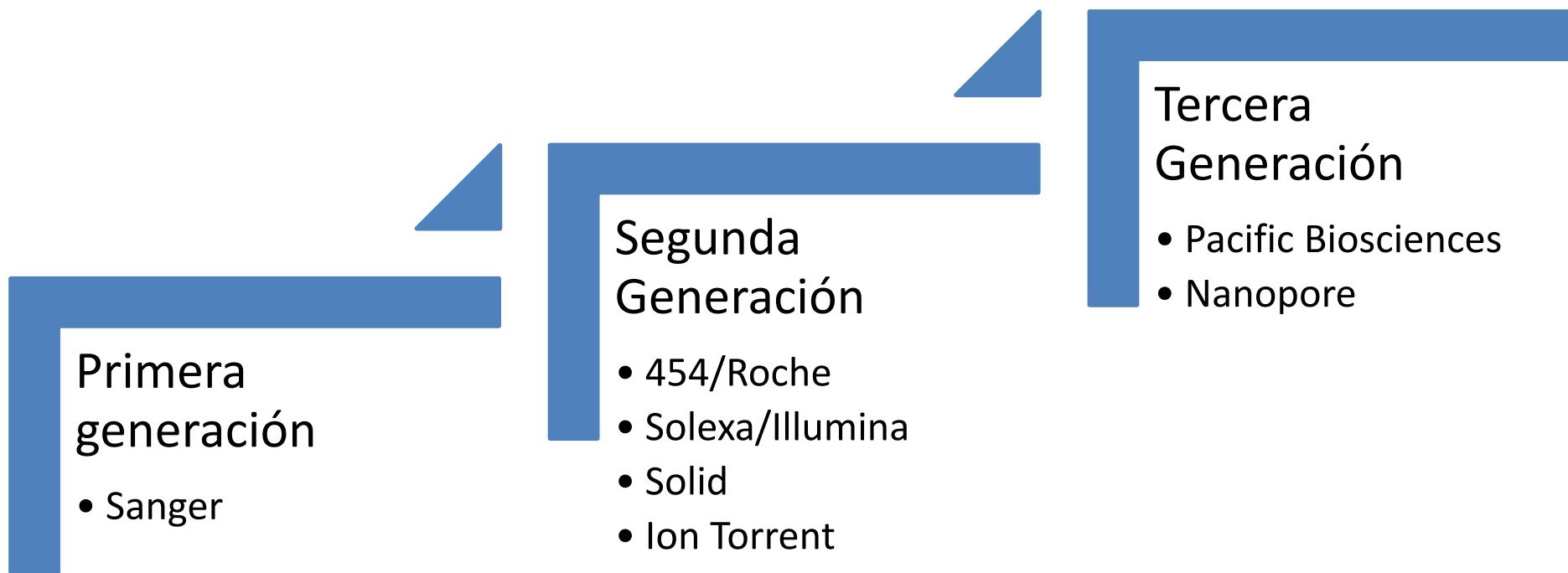
Outbreak research:  
Molecular source  
detection

# ECDC roadmap and international commitment

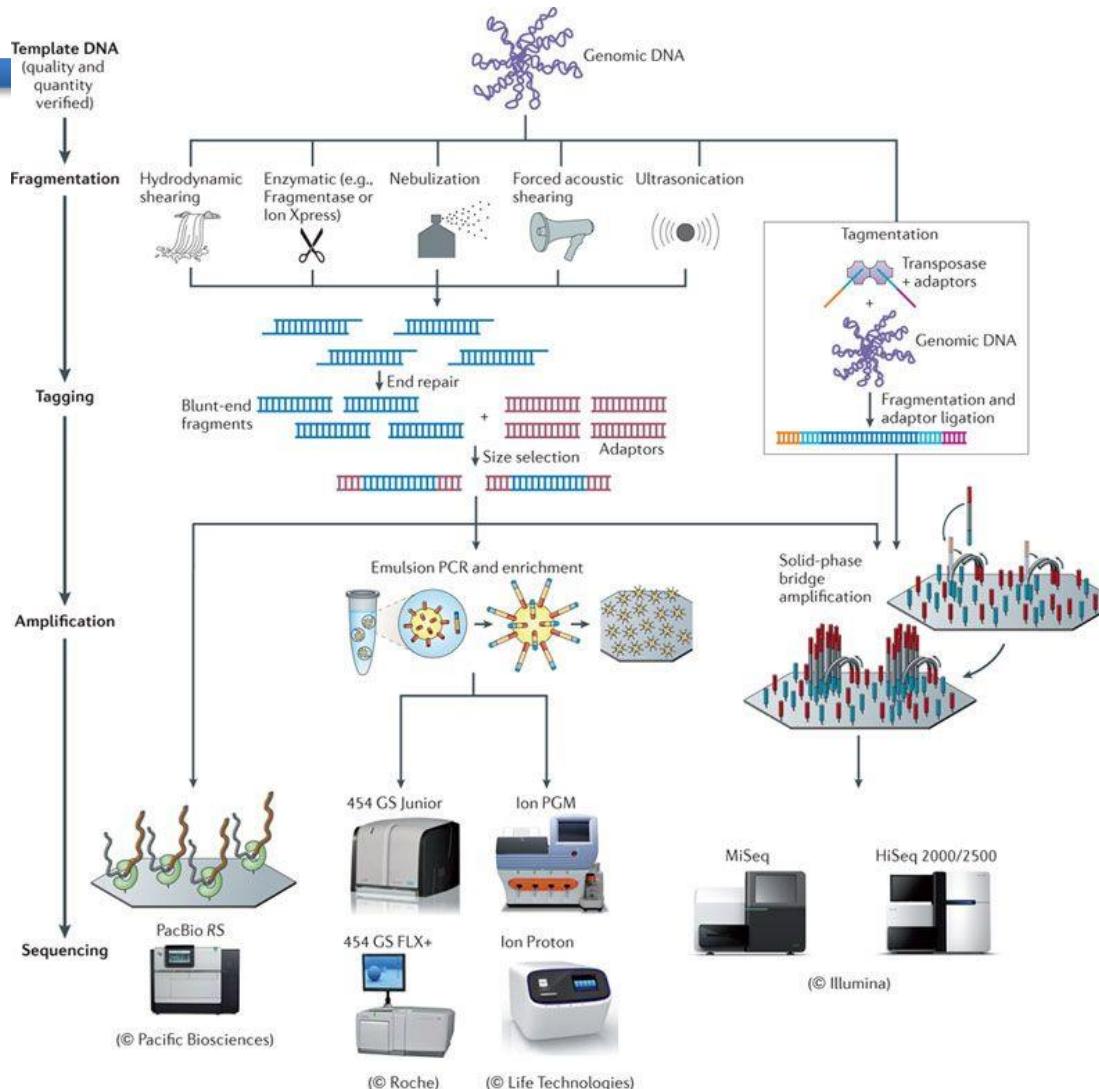


- **Operationalisation of EU-wide WGS-based surveillance systems in the near term:** start implementation of WGS-based surveillance for *Listeria monocytogenes*, *Neisseria meningitidis*, Carbapenemase-producing *Enterobacteriaceae* and antibiotic-resistant *Neisseria gonorrhoeae*; 2018

# DNA sequencing technologies 2006-2016

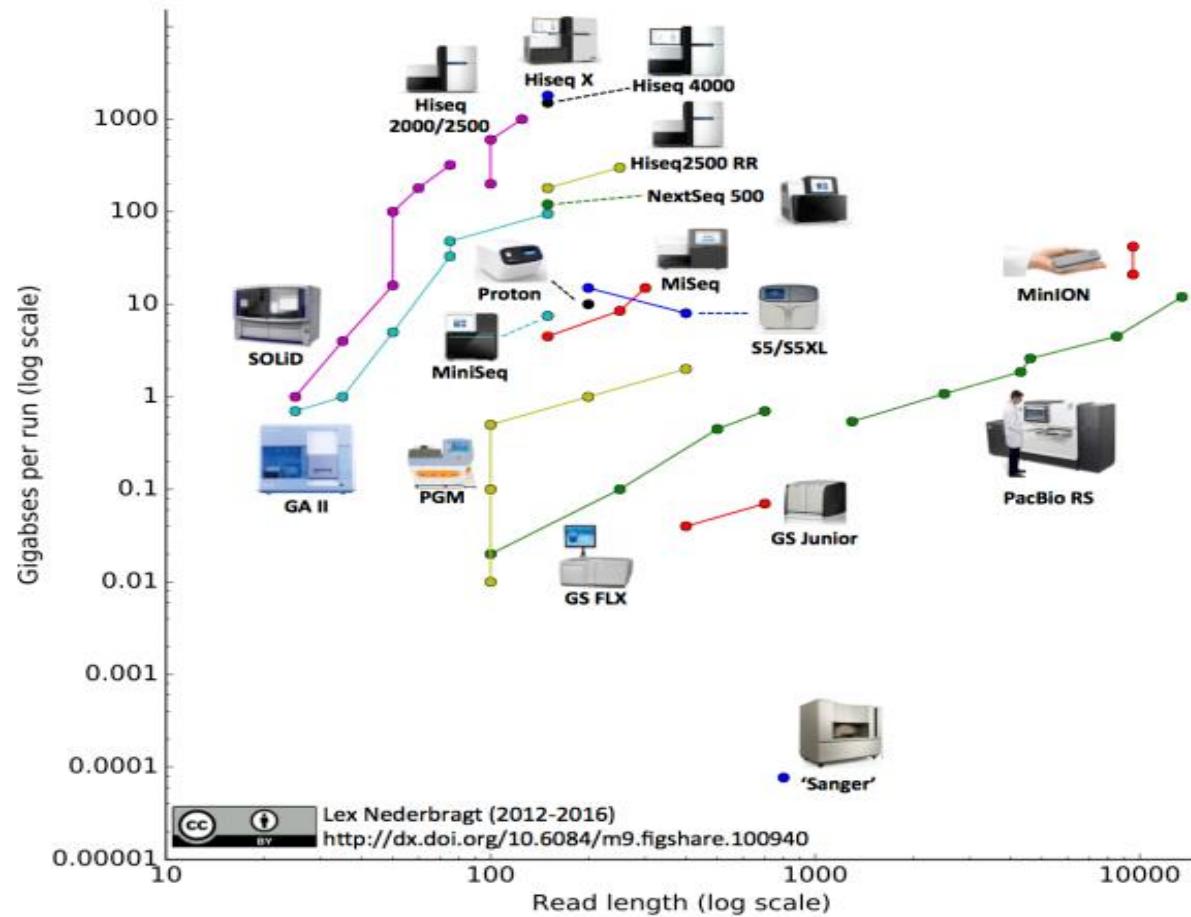


# High-throughput sequencing platforms



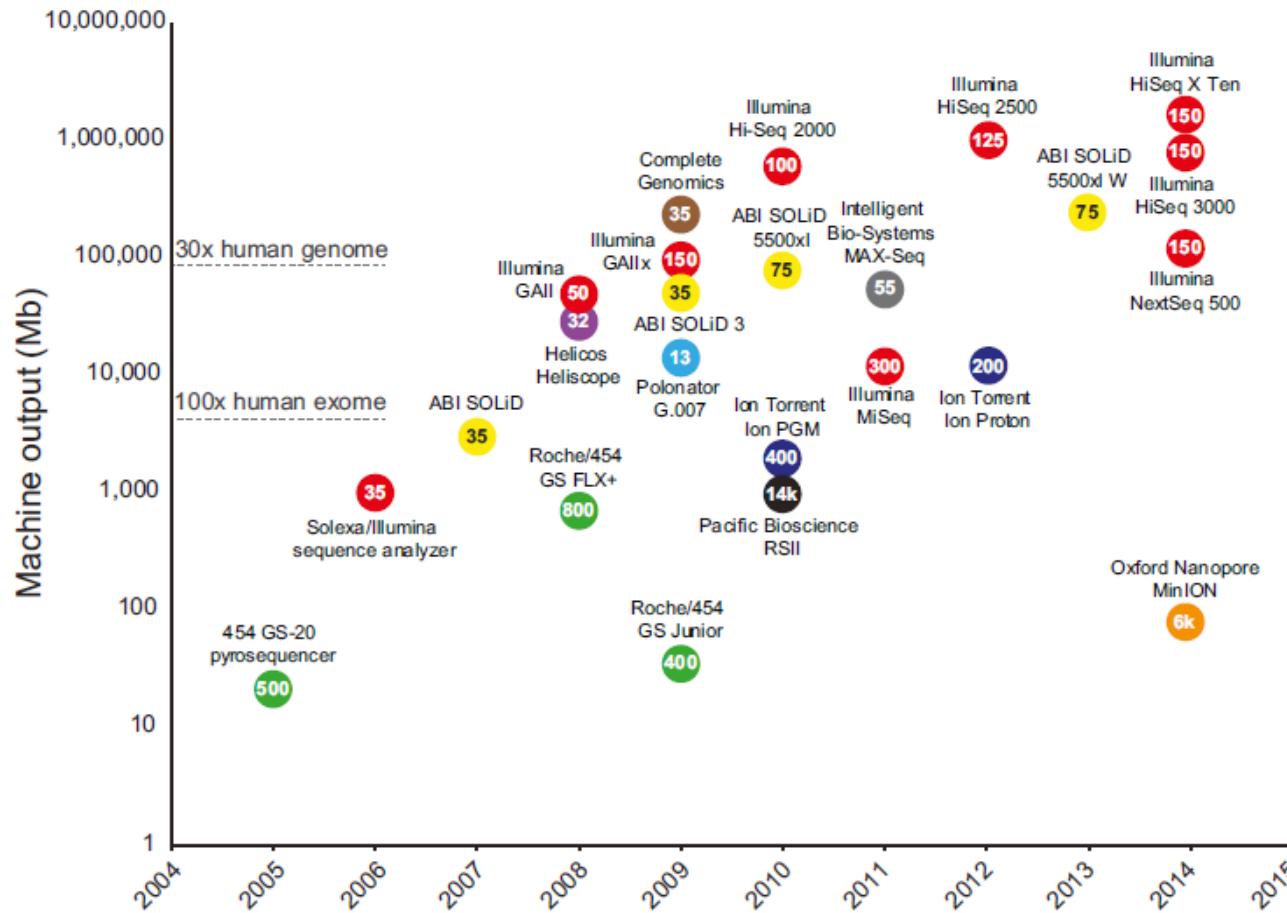
Nature Reviews | Microbiology Loman et al, 2012

# High-Throughput Sequencing Technologies



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

# High-Throughput Sequencing Technologies



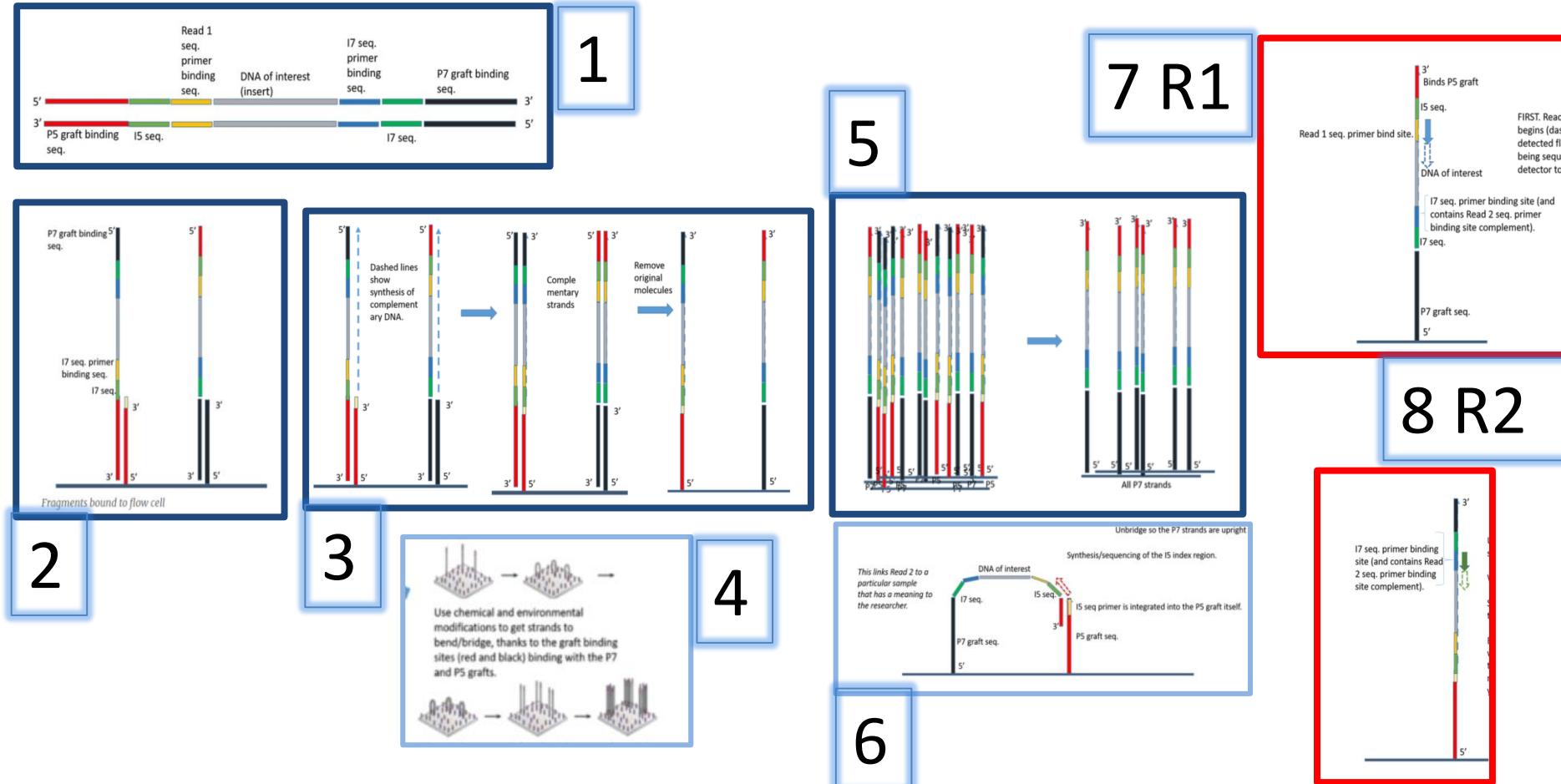
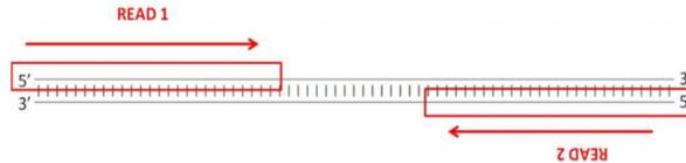
Numbers inside data points denote current read lengths.  
Sequencing platforms are color coded.

Reuter et al., Mol Cell 2015

## NGS PLATFORMS, main characteristics

- Numero de bases que secuencia
- Numero lecturas → aplicaciones
- Longitud de las lecturas -→ importante para las aplicaciones ensamblado genomas, de illumina a PacBio
- Error de la base → Corrección con profundidad de lectura
- Formato fichero salida
- Software dedicado, universal fastq

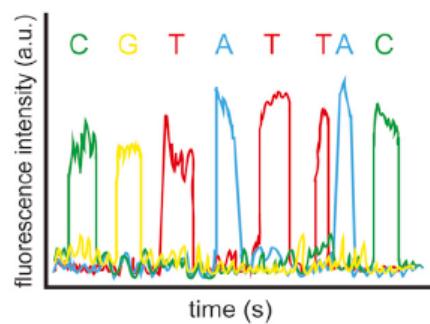
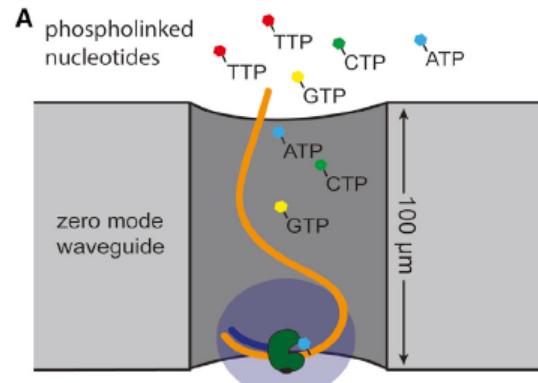
# Illumina sequencing



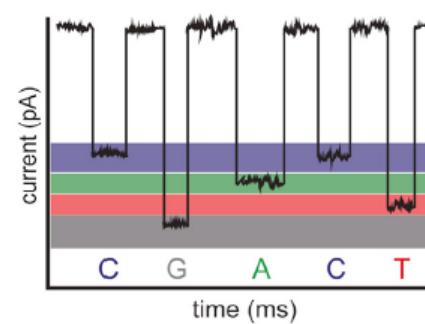
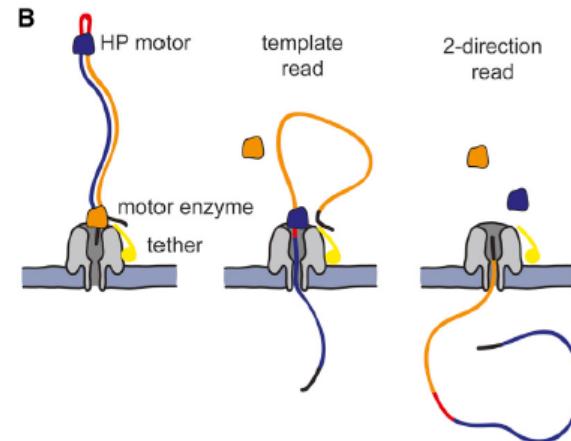
<https://kscbioinformatics.wordpress.com/2017/02/13/illumina-sequencing-for-dummies-samples-are-sequenced/>

# 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



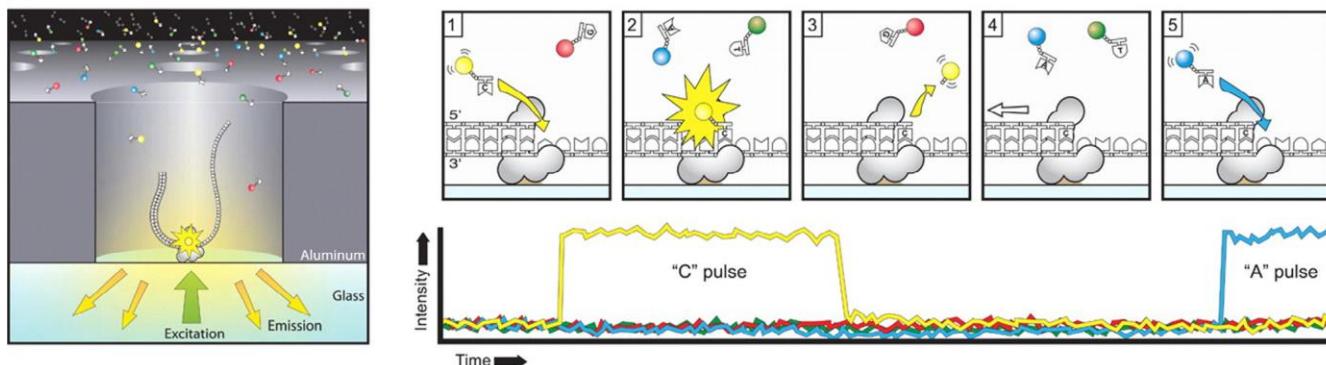
**SMRTbell template:** is a closed, single-stranded circular DNA that is created by ligating hairpin adaptors to both ends of a target dsDNA

**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

Table 2 *De novo* genome assemblies using hybrid sequencing or PacBio sequencing alone

Species	Method	Tools	SMRT cells	Coverage	Contigs	Achievements	Ref.
<i>Clostridium autoethanogenum</i>	PacBio	HGAP	2	179×	1	21 fewer contigs than using SGS; no collapsed repeat regions ( $\geq 4$ using SGS)	[7]
<i>Potentilla micrantha</i> (chloroplast)	PacBio	HGAP, Celera, minimus2, SeqMan	26	320×	1	6 fewer contigs than with Illumina; 100% coverage (Illumina: 90.59%); resolved 187 ambiguous nucleotides in Illumina assembly; unambiguously assigned small differences in two $> 25$ kb inverted repeats	[33]
<i>Escherichia coli</i>	PacBio	PBcR, MHAP, Celera, Quiver	1	85×	1	4.6 CPU hours for genome assembly (10× improvement over BLASR)	[31]
<i>Saccharomyces cerevisiae</i>	PacBio	PBcR, MHAP, Celera	1P	117×	21	27 CPU hours for genome assembly (8× improvement over BLASR); improved current reference of telomeres	[31]
<i>Arabidopsis thaliana</i>	PacBio	PBcR, MHAP, Celera	46	144×	38	1896 CPU hours for genome assembly	[31]
<i>Drosophila melanogaster</i>	PacBio	PBcR, MHAP, Celera, Quiver	42	121×	132	1060 CPU hours for genome assembly (593× improvement over BLASR); improved current reference of telomeres	[31]
<i>Homo sapiens</i> (CHM1cert)	PacBio	PBcR, MHAP, Celera	275	54×	3434	262,240 CPU hours for genome assembly; potentially closed 51 gaps in GRCh38; assembled MHC in 2 contigs (60 contigs with Illumina); reconstructed repetitive heterochromatic sequences in telomeres	[31]
<i>Homo sapiens</i> (CHM1cert)	PacBio	BLASR, Celera, Quiver	243	41×	N/A (local assembly)	Closed 50 gaps and extended into 40 additional gaps in GRCh37; added over 1 Mb of novel sequence to the genome; identified 26,079 indels at least 50 bp in length; cataloged 47,238 SV breakpoints	[32]
<i>Melopsittacus undulatus</i>	Hybrid	PBcR, Celera	3	$5.5 \times$ PacBio + $15.4 \times$ 454 $= 3.83 \times$ corrected	15,328	1st assembly of $> 1$ Gb parrot genome; N50 = 93,069	[34]
<i>Vibrio cholerae</i>	Hybrid	BLASR, Bambus, AHA	195	$200 \times$ PacBio + $28 \times$ Illumina $+ 22 \times$ 454	2	No N's in contigs; 99.99% consensus accuracy; N50 = 3.01 Mb	[30]
<i>Helicobacter pylori</i>	PacBio	HGAP, Quiver, PGAP	8 per strain	446.5× average among strains	1 per strain	1 complete contig for each of 8 strains; methylation analysis associated motifs with genotypes of virulence factors	[35]

Note: N50, the contig length for which half of all bases are in contigs of this length or greater; MHC, major histocompatibility complex; SV, structural variation.

# 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

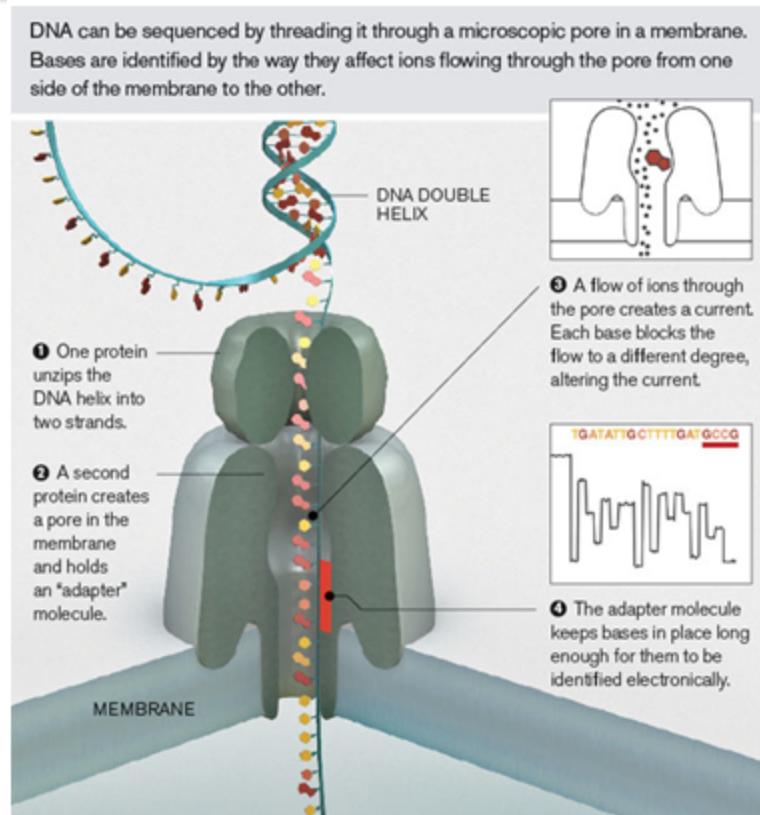
Produced highly-contiguous assemblies of bacterial and eukaryotic genomes

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



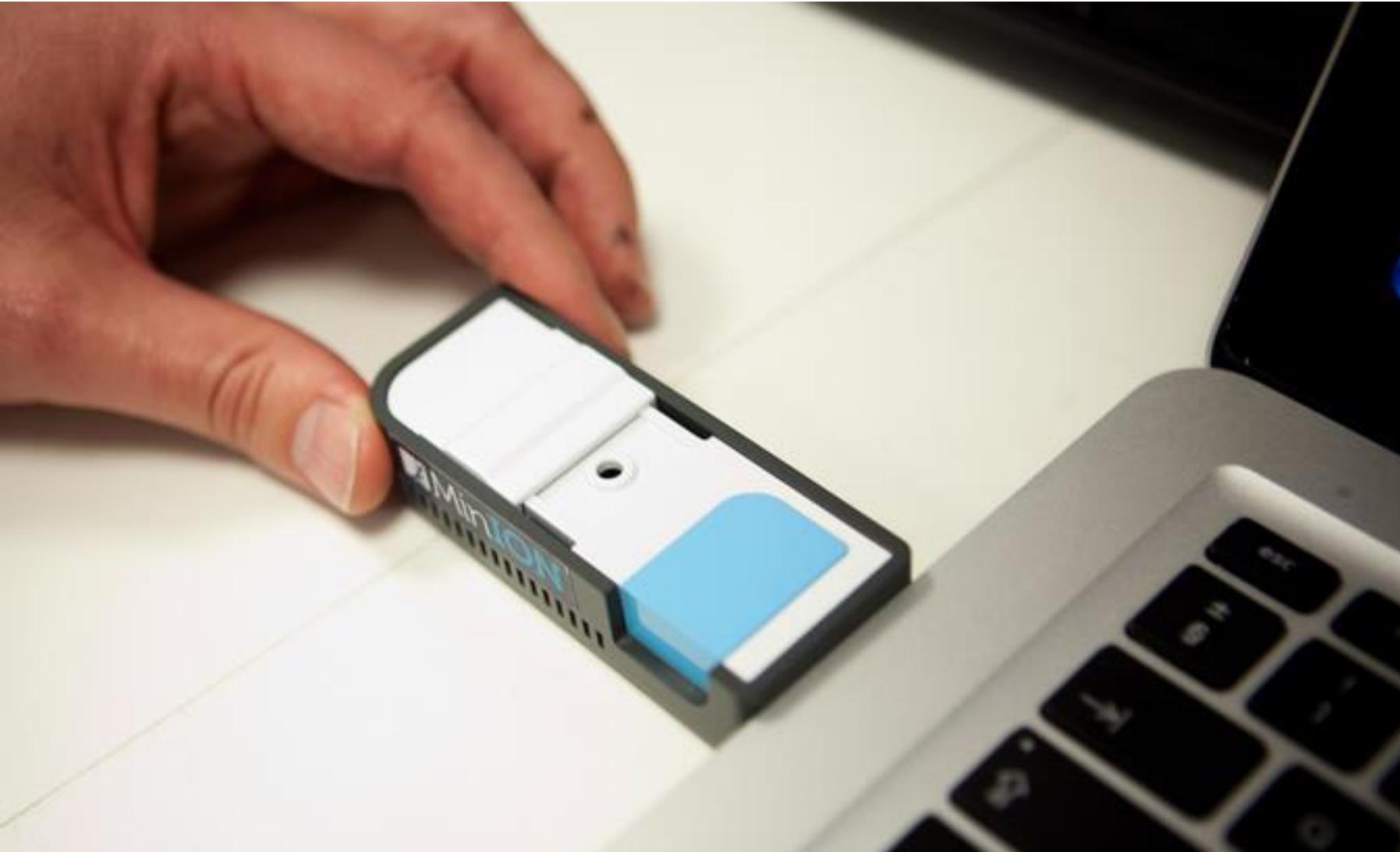
'Strand sequencing' is a technique that passes intact DNA polymers through a protein nanopore, sequencing in real time as the DNA translocates the pore.

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	MinION	GridION	PromethION
Long read, direct DNA/RNA/epigenetic sequencing, scalable, real time/rapid, on-demand sequencing that is easy to use and install.			
<ul style="list-style-type: none"><li>✓ Your portable device for smaller, individual, rapid tests.</li><li>✓ When you don't want to multiplex samples or start a larger run.</li><li>✓ Amplicons, panels/targeted sequencing, quality testing and more.</li><li>✓ For use with MinIT or a laptop.</li></ul>	<ul style="list-style-type: none"><li>✓ Your personal sequencer, putting you in control.</li><li>✓ Whether in your lab or out in the field.</li><li>✓ Whole genomes/exomes, metagenomics, targeted sequencing, whole transcriptome (cDNA), smaller transcriptomes (direct RNA), multiplexing for smaller samples and more.</li><li>✓ For use with MinIT or a laptop.</li></ul>	<ul style="list-style-type: none"><li>✓ High throughput sequencing, in modular form (up to 5 flow cells) to be on-demand.</li><li>✓ For your lab or to offer as a service.</li><li>✓ Larger genomes or projects, whole transcriptomes (direct RNA or cDNA) or where you have larger numbers of samples and more.</li><li>✓ Compute included for real time data analysis and easy installation.</li></ul>	<ul style="list-style-type: none"><li>✓ Very high throughput sequencing, in modular form (up to 48 flow cells) to be on-demand.</li><li>✓ For your lab or as a service.</li><li>✓ Larger genomes or projects, whole transcriptomes (direct RNA or cDNA), very large numbers of samples and more.</li><li>✓ Compute included for real time data analysis and easy installation.</li></ul>

# Library preparation



Oxford Nanopore has developed VolTRAX – a small device designed to perform library preparation automatically, so that a user can get a biological sample ready for analysis, hands-free. VolTRAX is designed as an alternative to a range of lab equipment, to allow consistent and varied, automated library prep options.

## VolTRAX V2 Starter Pack

\$8,000.00

VolTRAX V2 is designed to automate all laboratory processes associated with Nanopore Sequencing from sample extraction to library preparation.

# SmidgION, Mobile analysis



Oxford Nanopore has now started developing an even smaller device, SmidgION.

**potential applications** may include remote monitoring of pathogens in a breakout or infectious disease; the on-site analysis of environmental samples such as water/metagenomics samples, real time species ID for analysis of food, timber, wildlife or even unknown samples; field-based analysis of agricultural environments, and much more.

# 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>IonTorrent</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,

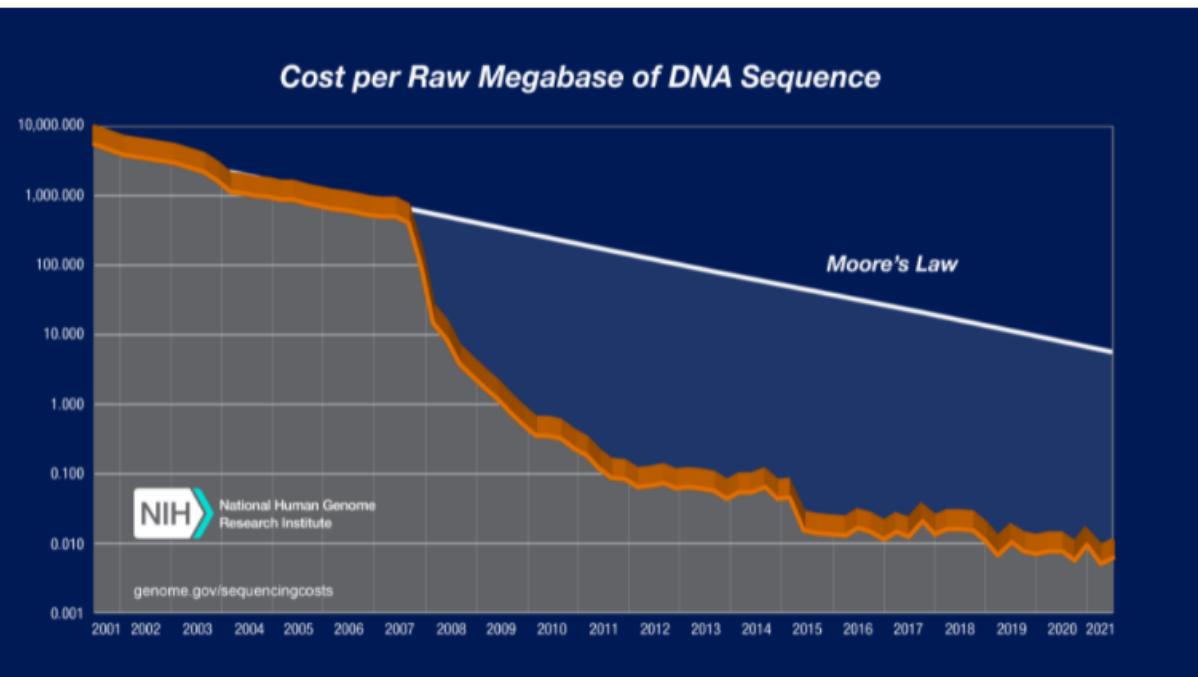
<sup>a</sup> The throughput ranges are determined by available kits and run modes on a per run basis. As an example of a 15-GB throughput, thirty-five 5-MB genomes can be sequenced to a minimum coverage of 40× on the Illumina MiSeq using the v3 600 cycle chemistry.

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

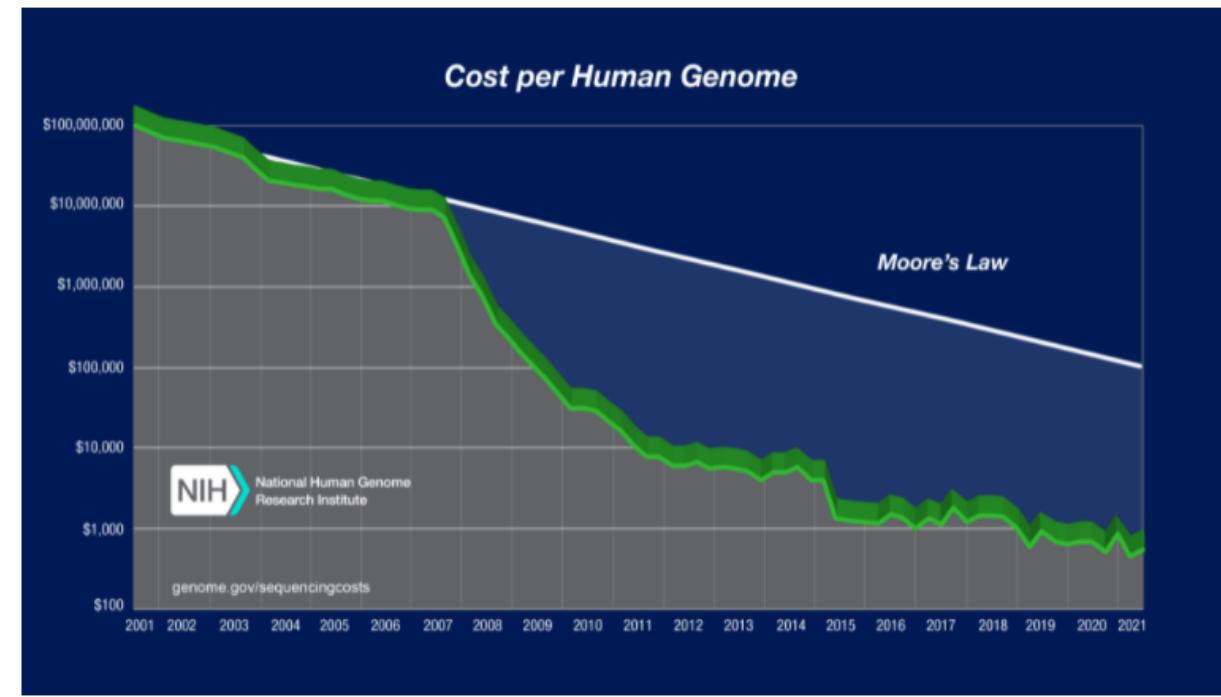
<sup>c</sup> Results in increased error rate (increased proportion of reads containing errors among all reads) which in turn results in false-positive variant calling.

Besser et al., Clin Micr Infect, 2018

# Secuencing cost



Sequencing cost per megabase - 2021



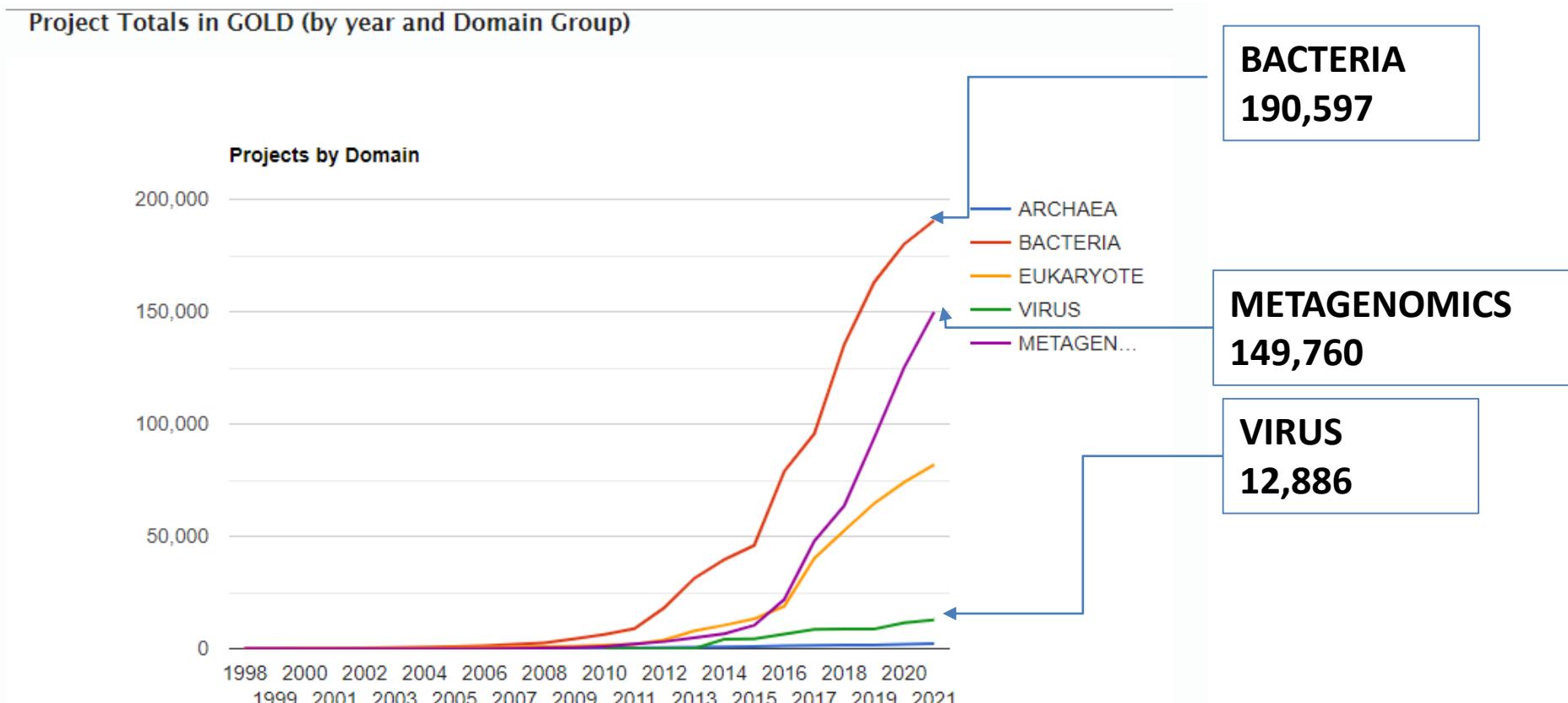
Cost per genome data - 2021

<https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>

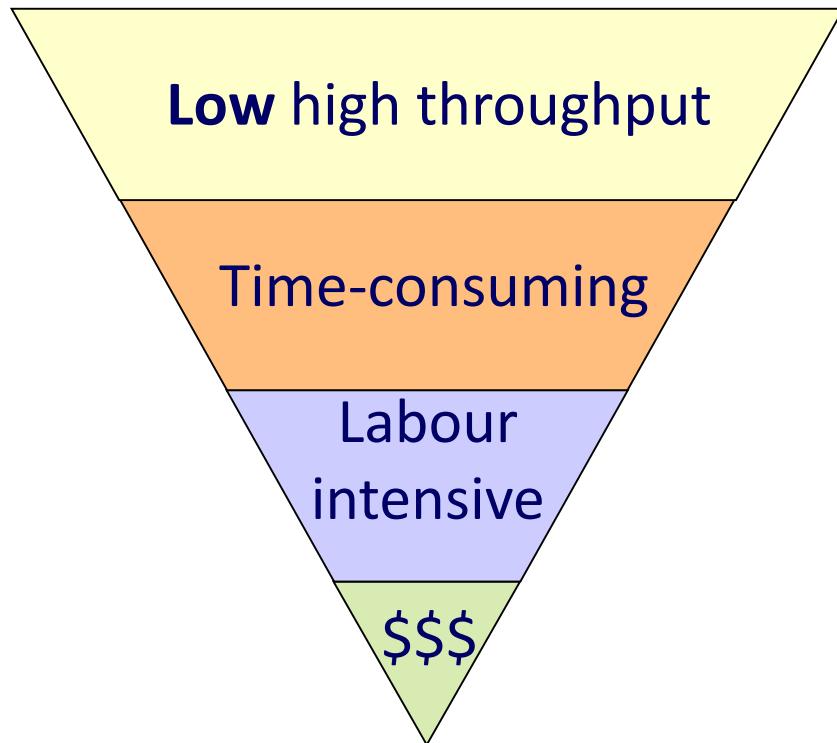
# Sequencing projects

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

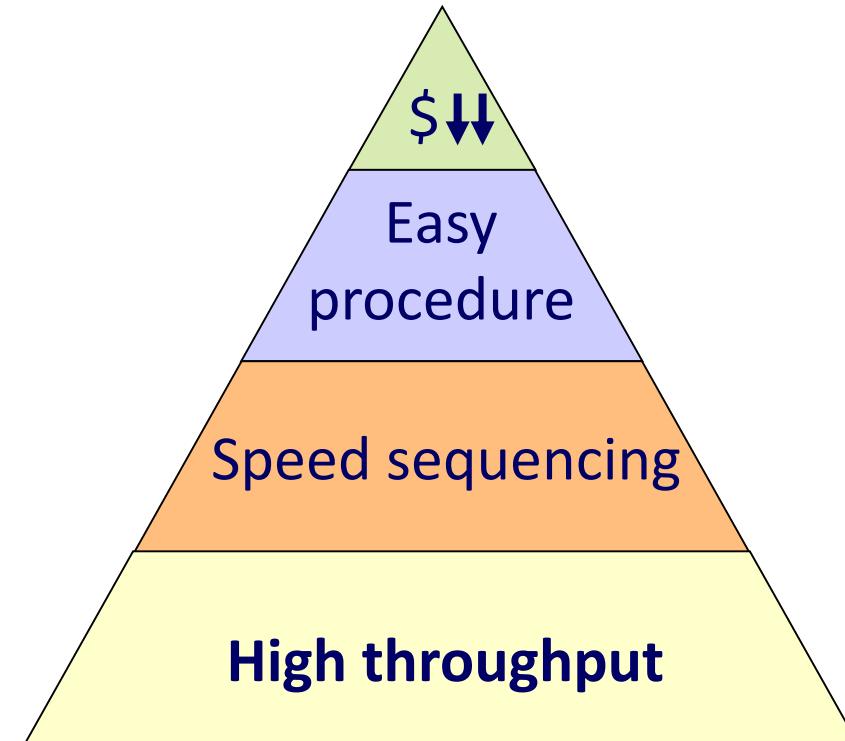
## GOLD, Genome Online DataBase



# Sanger vs SM, advantages of new technologies

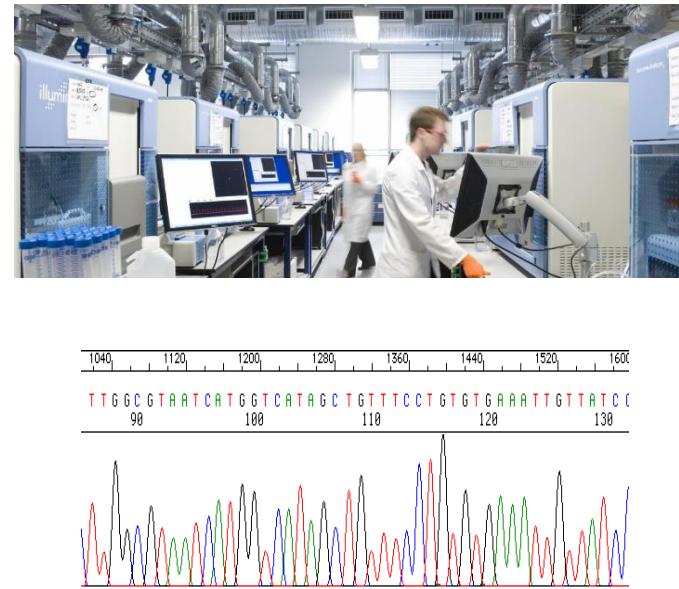
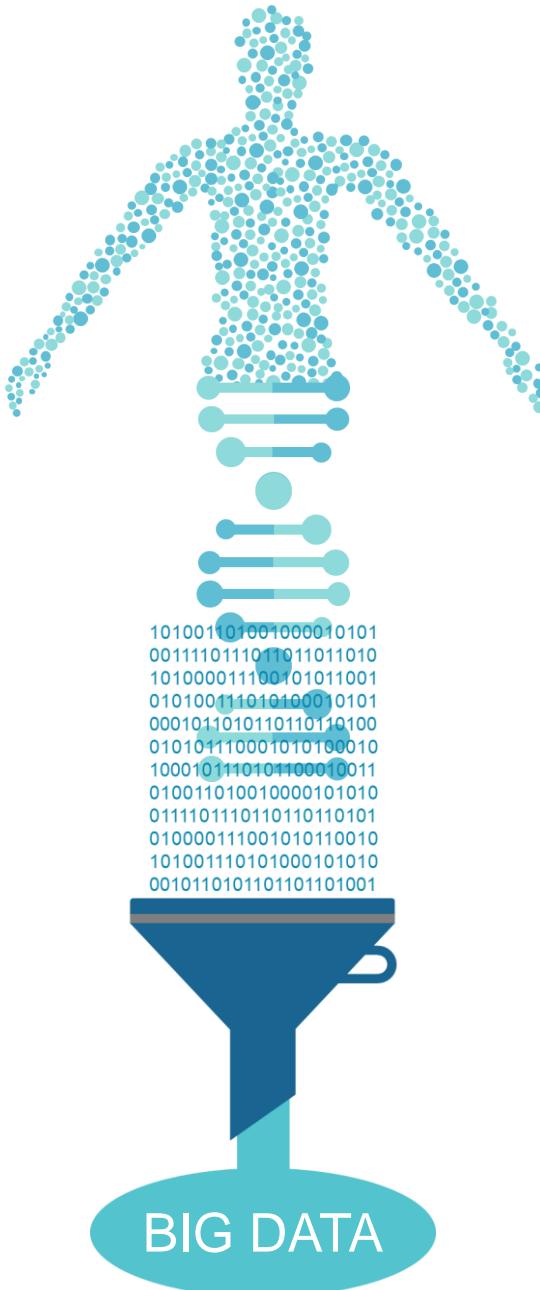


Semiautomatic **Sanger** capillary-based sequencing technology

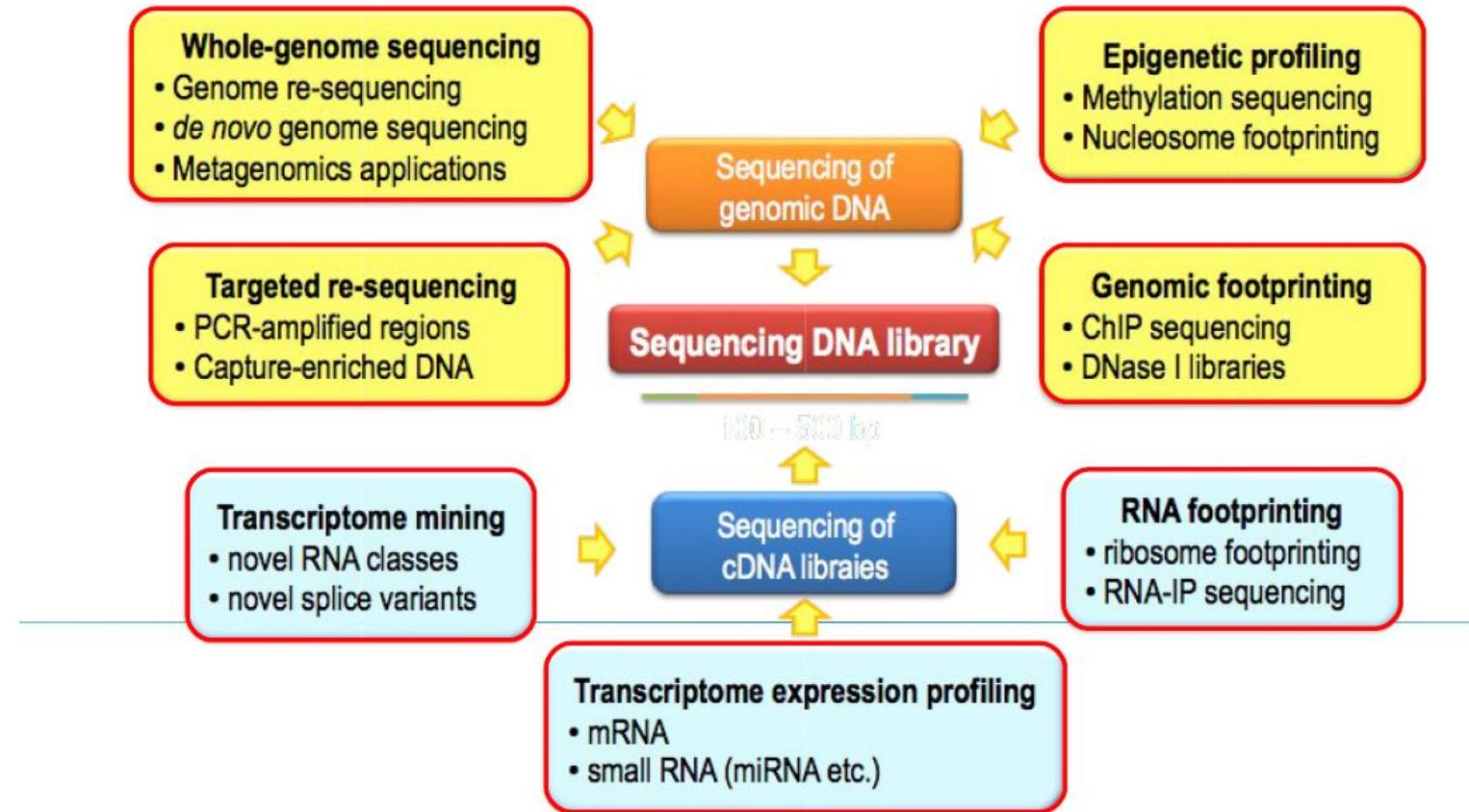


NGS  
Next Generation Sequencing =  
Now Generation Sequencing

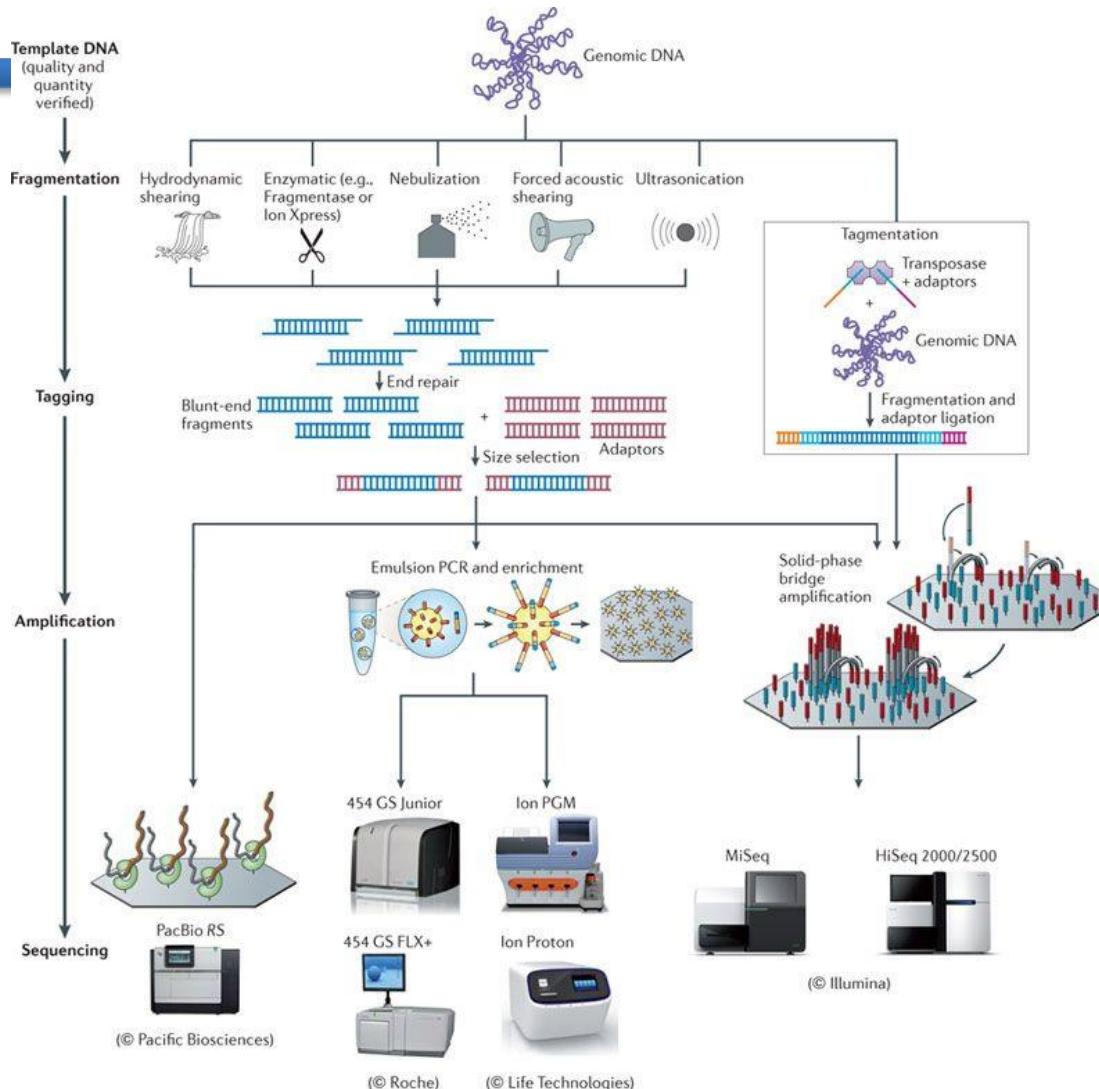
# BIG DATA



# Massive sequencing applications



# High-throughput sequencing platforms



Nature Reviews | Microbiology Loman et al, 2012

## 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 Paired-end
- Número de bases por muestra
- Profundidad de cobertura x

### **SECUENCIACIÓN GENOMAS**

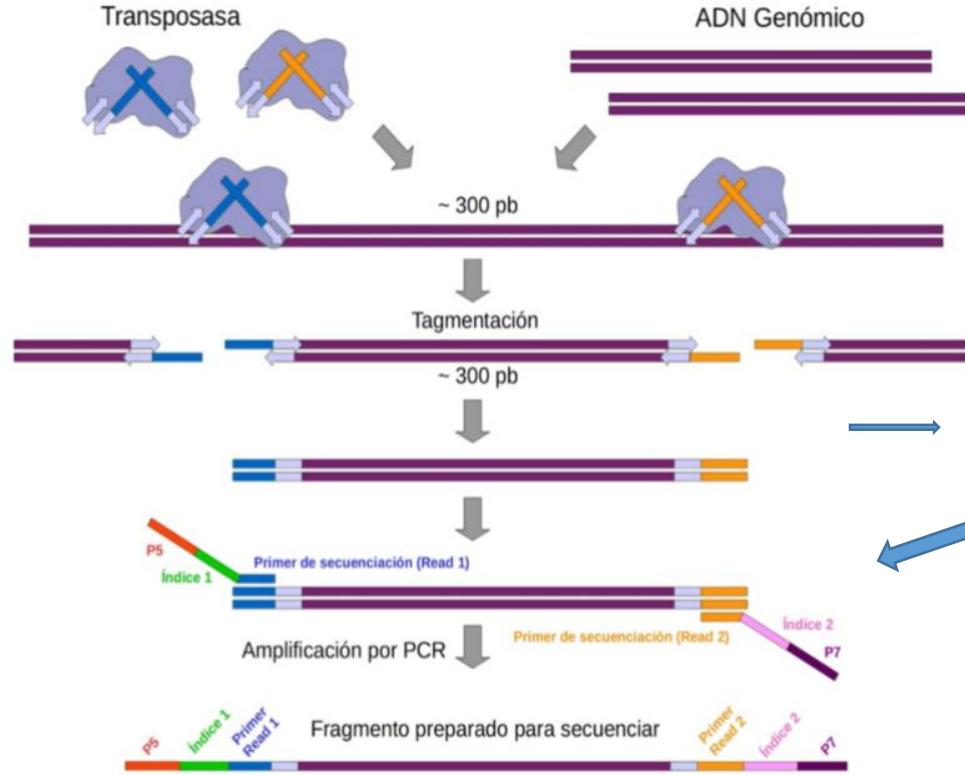
1. Metagenómica

### **IDENTIFICACIÓN MICROORGANISMOS**

1. Metataxonomía

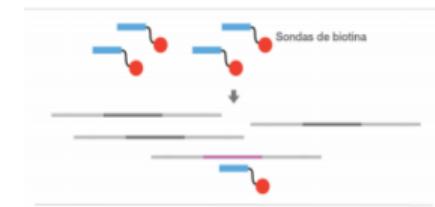
# PREPARACIÓN LIBRERÍA

## ENZIMÁTICA FÍSICA



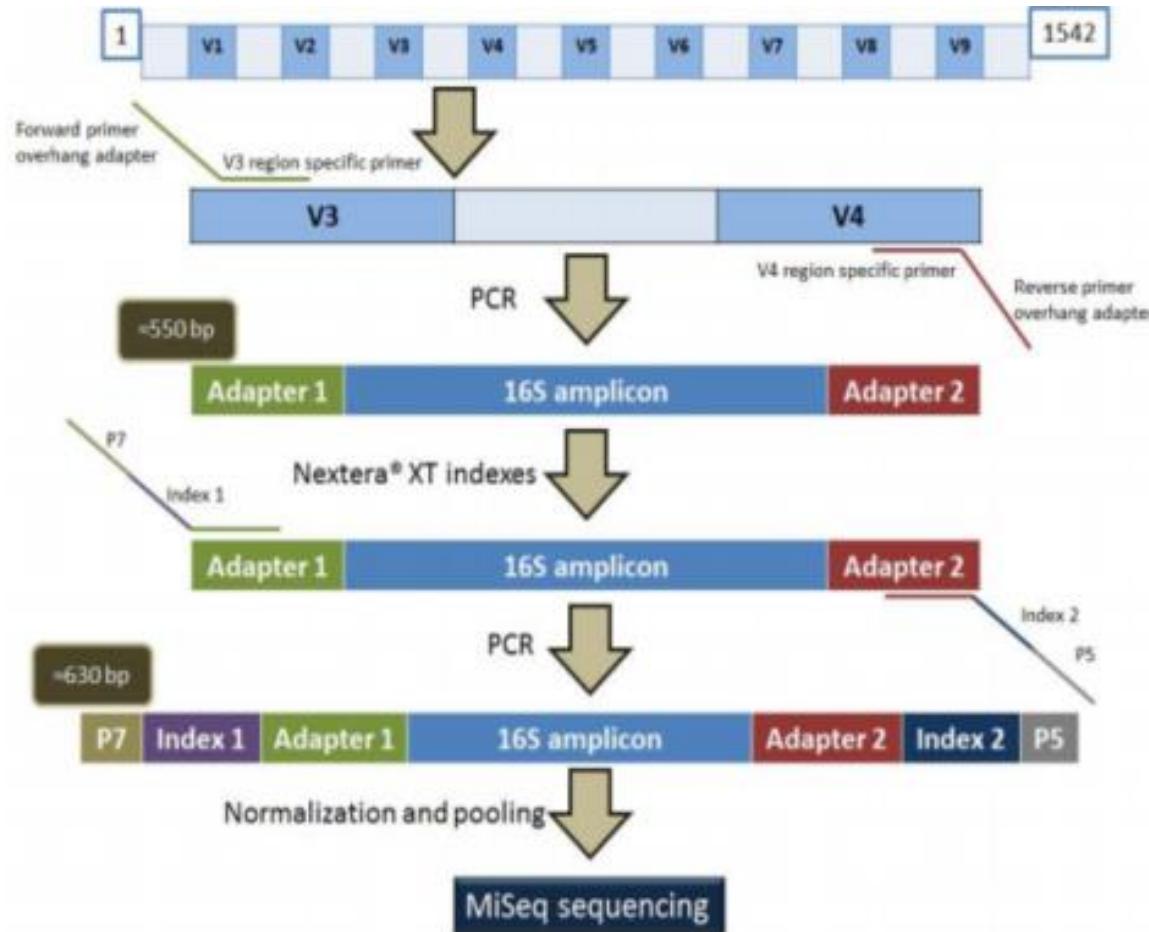
RNA → cDNA

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)

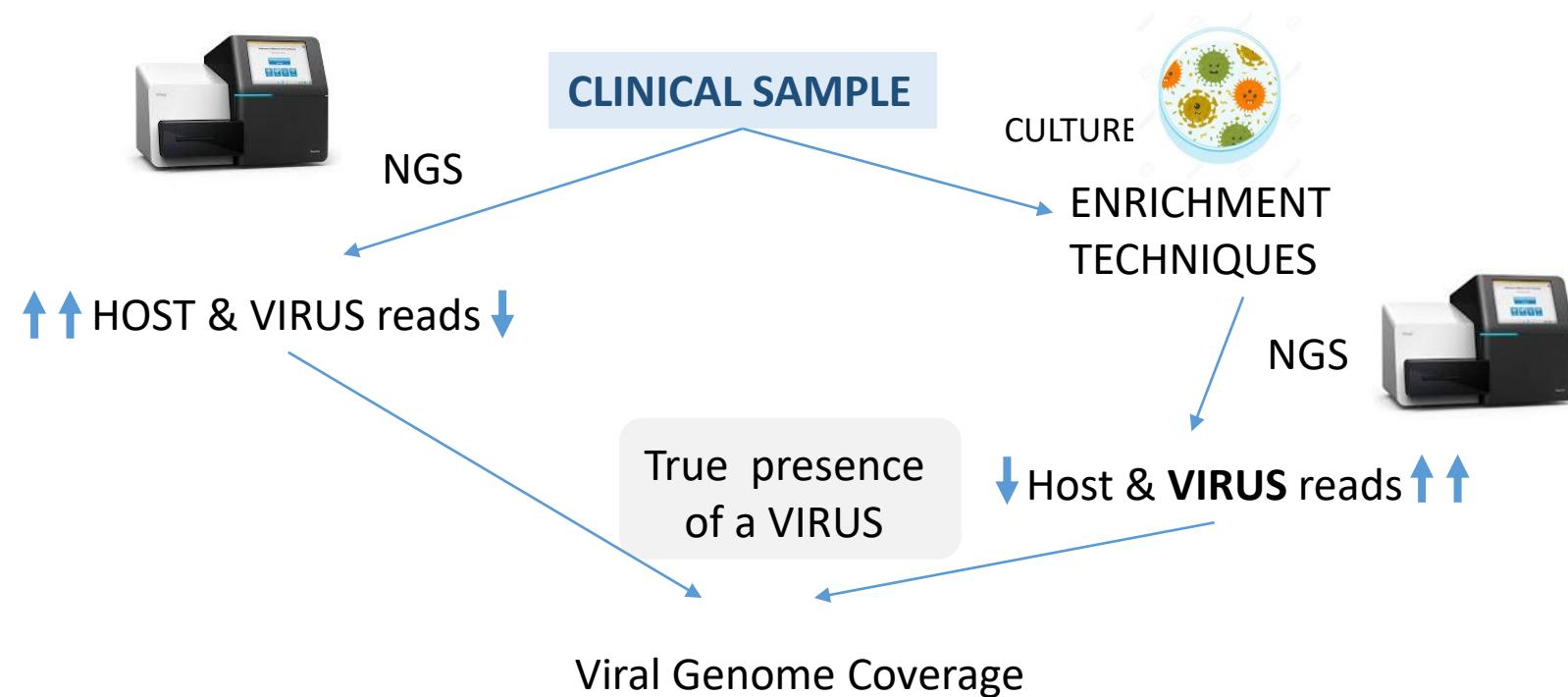
## PREPARACIÓN LIBRERÍA, rRNA 16S, caracterización microbiota



## Main Steps of Viral Genome Sequencing by NGS or HTS

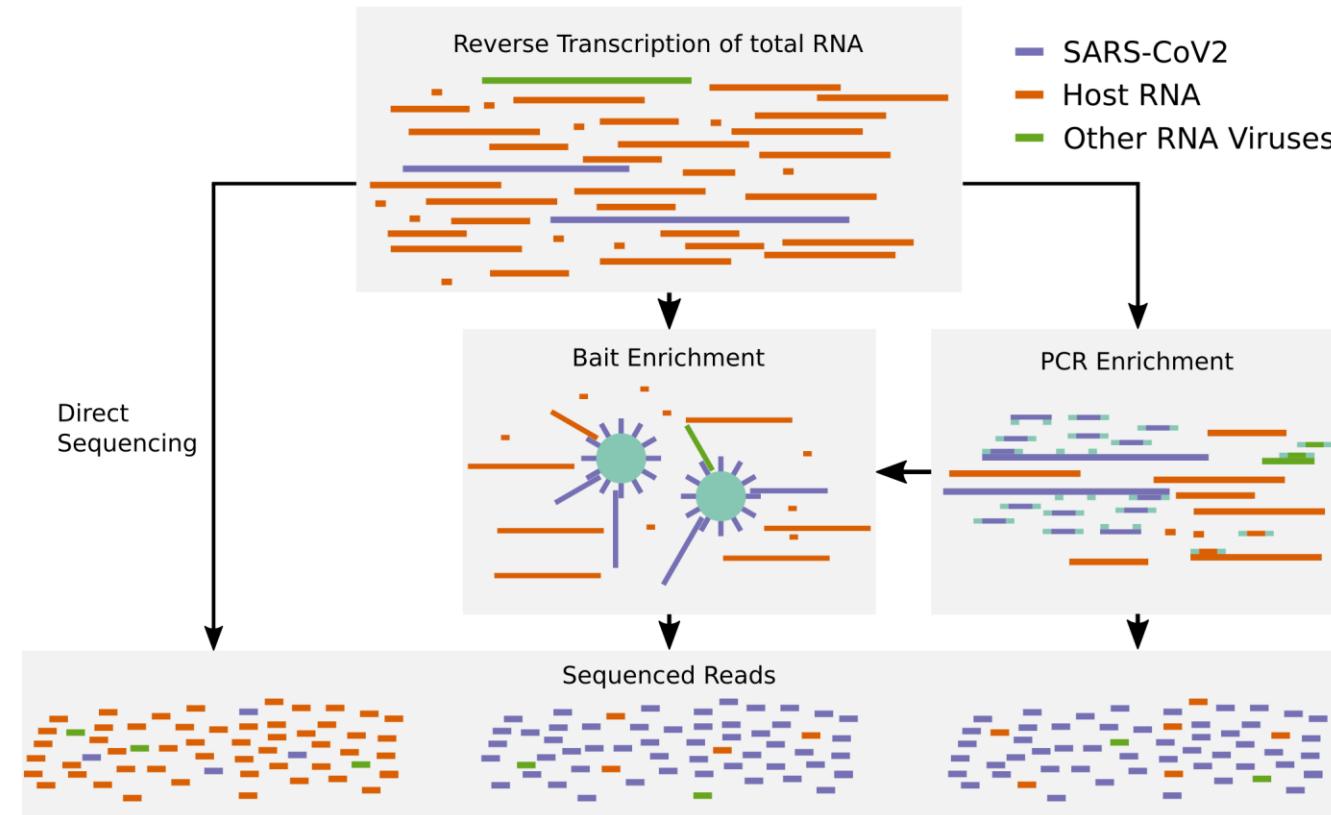
- Nucleic acid amplification
- Library preparation
- High throughput sequencing platforms
- Data analysis

# Viral Genome Sequencing



NGS needs a cutoff to determine the true presence of a pathogen versus carry-over or contamination between specimens or other non-specific reads.

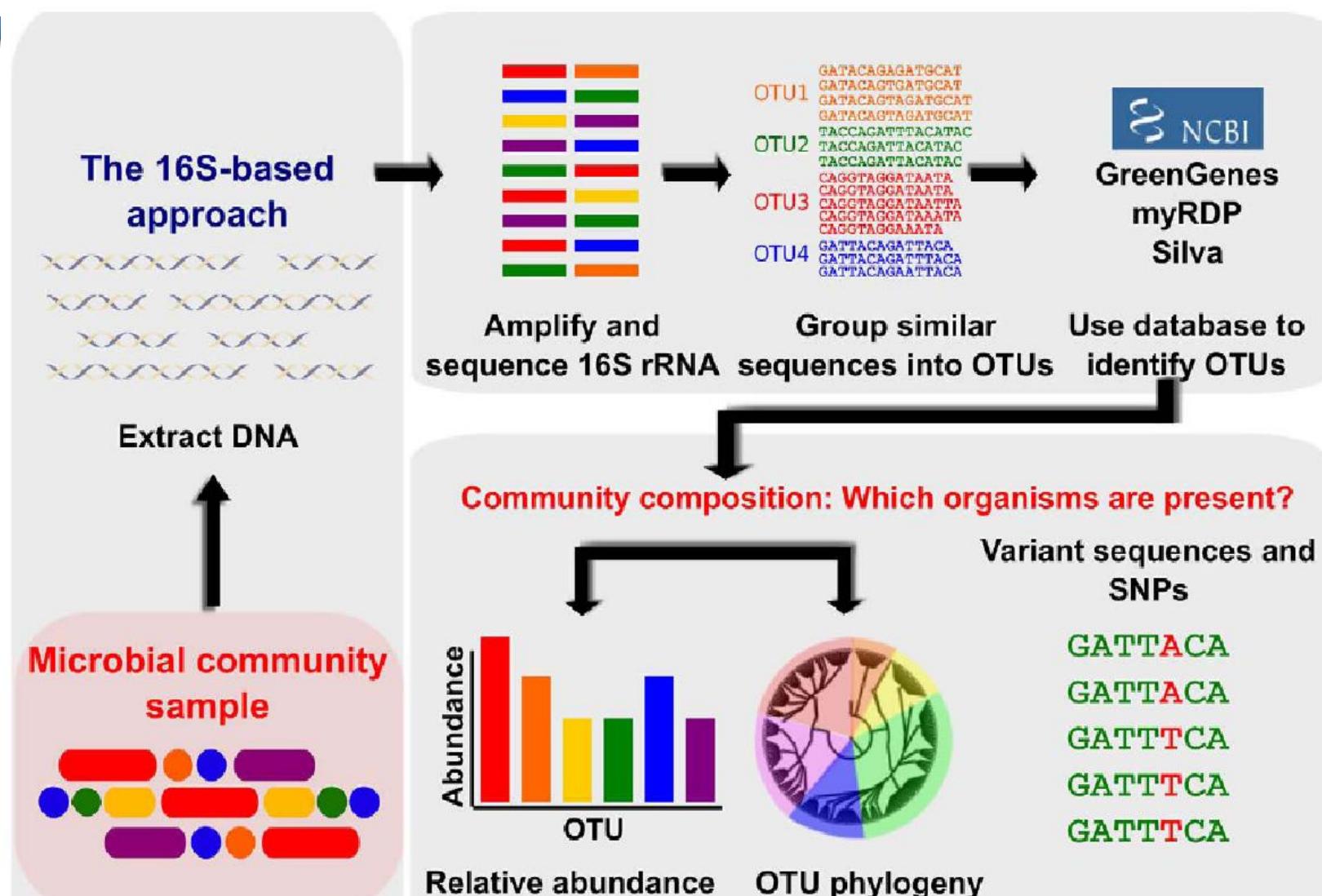
# Enrichment Techniques



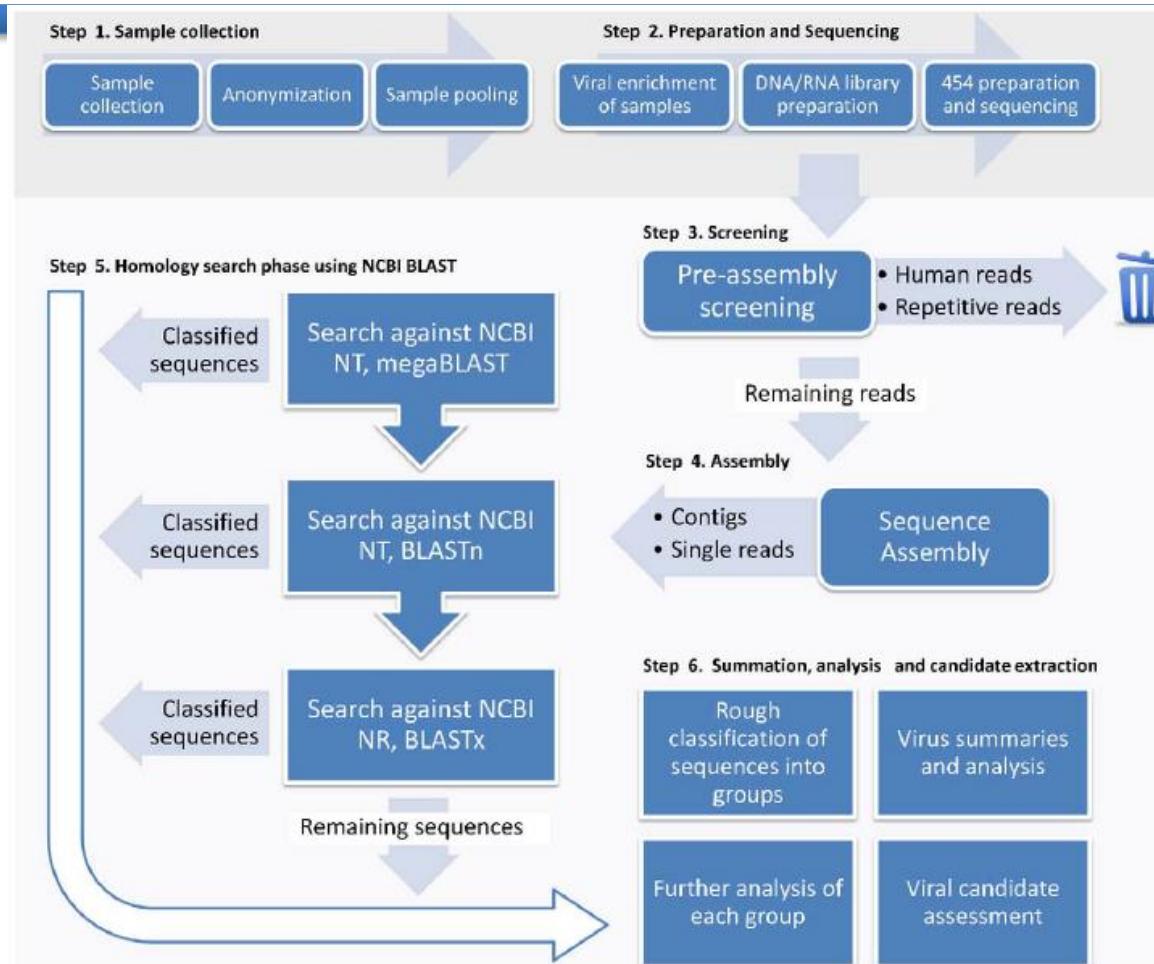
# Metataxonomics vs Metagenomics (16S vs Shotgun)

	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

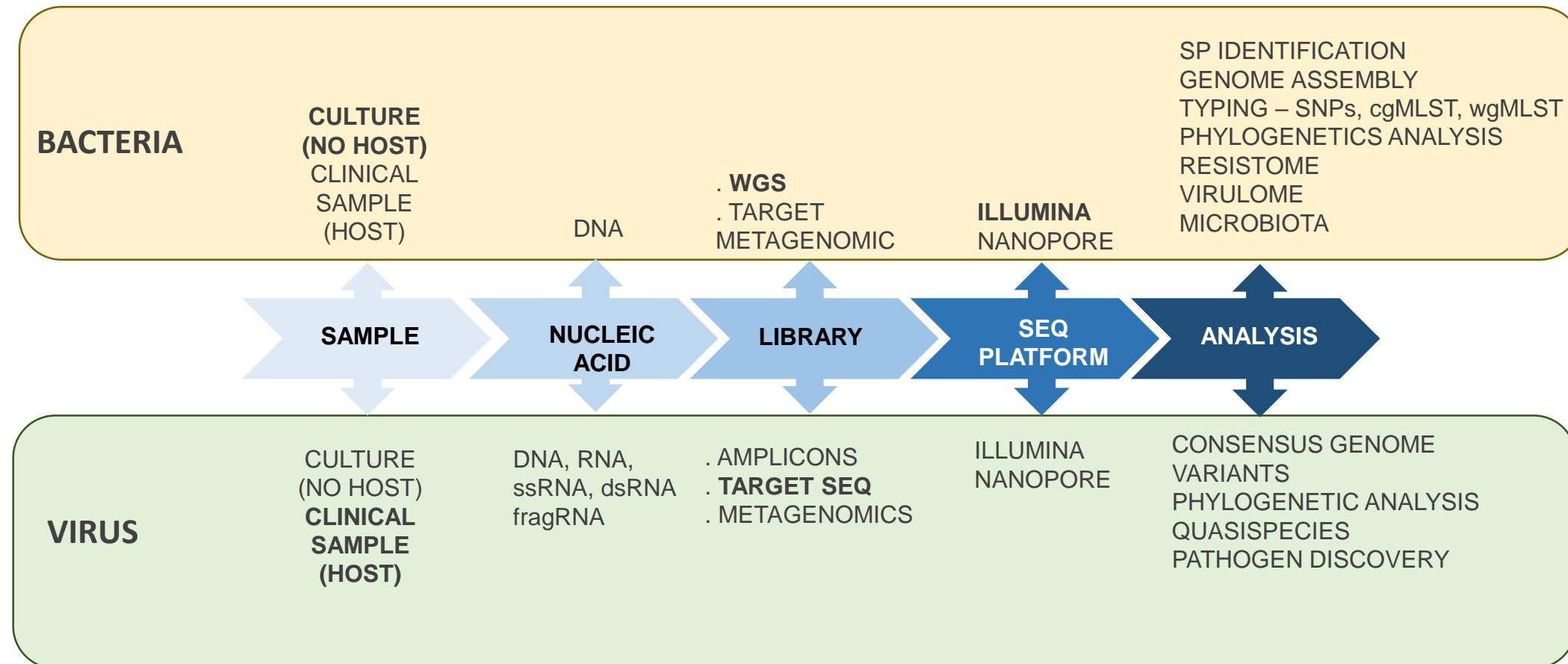


# Metagenómica, pipeline de análisis

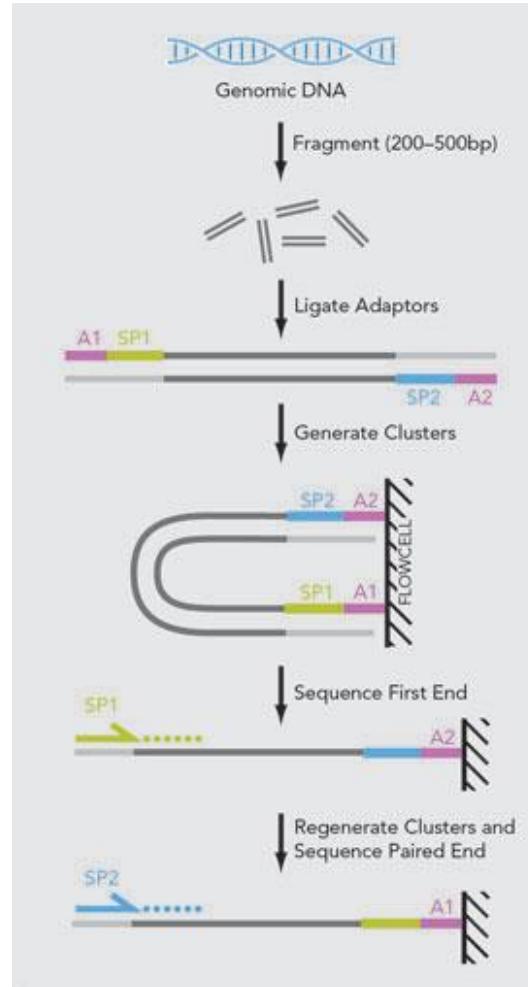


Lysholm et al., Plos One 2012:7,2, e30875

## Bacterial and viral Genome Sequencing

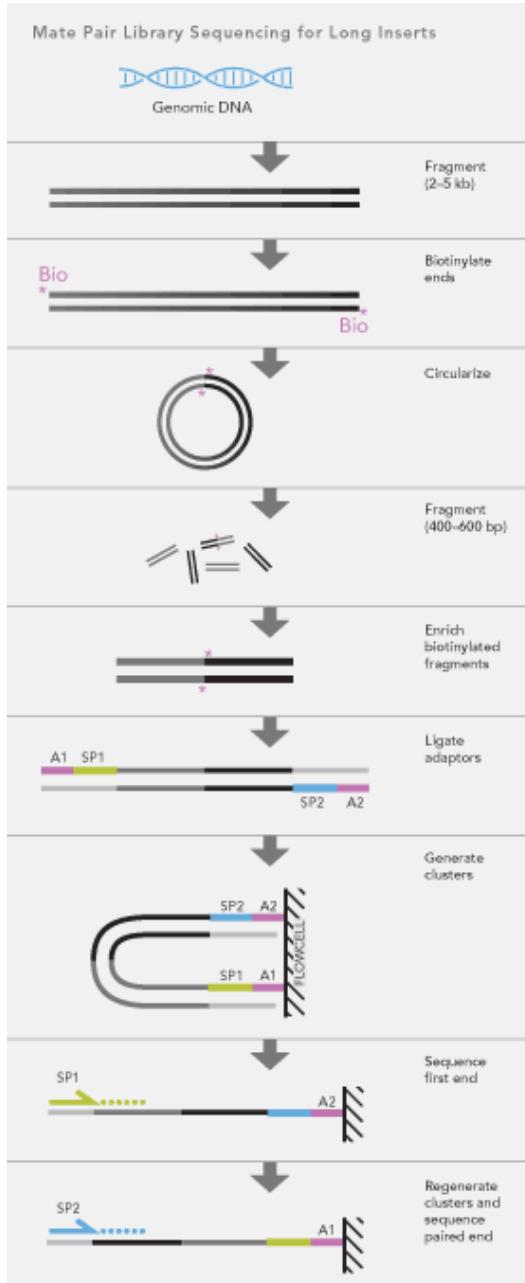


# Que es Pair-end?



**Secuenciación de un fragmento (bp)**

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



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

### Depth of coverage

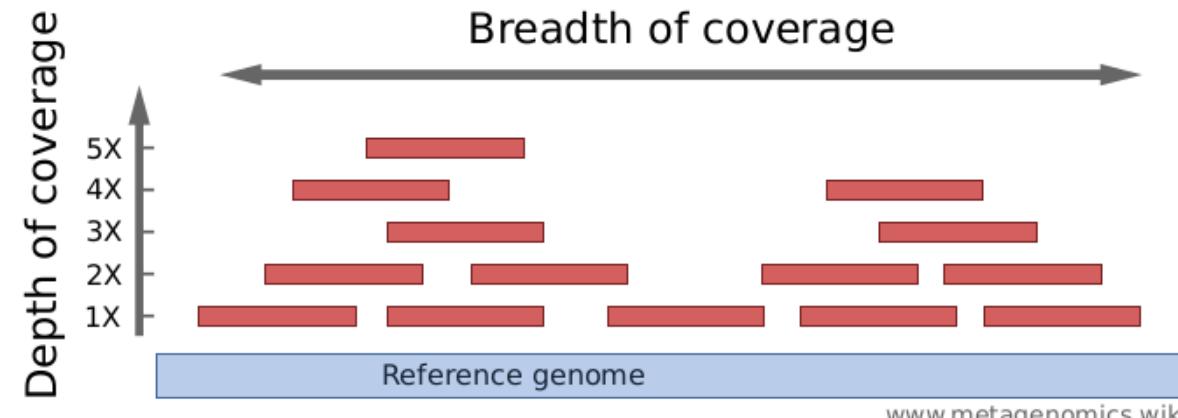
How strong is a genome "covered" by sequenced fragments (short reads)?

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).

### 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.



# 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
- HiSeq 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 v3 Reagents	MiSeq v2 Reagents	MiSeq v2 Nano Reagents	MiSeq 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)	-0/flow cell	-0/flow cell	-0/flow cell	-0/flow cell
Comments	Upgraded software: MCS v2.3 or later; MiSeq 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)	Upgraded hardware or from September 2012 and later: MCS v2.0 or later; MiSeq Reagent Kits v2
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)

# Depth of coverage and genome coverage

## Depth of coverage

the sequencing coverage = 
$$\frac{\text{the number of total reads} \times \text{the read length}}{\text{the length of target sequence or genome}}$$

## Genome coverage

% length sequence genome

### Increase number of raw reads

- For the low-frequency variants
- For assembly (also read lenght)

Thanks for your attention!

Questions???