

## MODULO 4: Introducción al análisis de datos de secuenciación masiva y sus aplicaciones en Microbiología. Caso de uso: reconstrucción del genoma de SARS-CoV-2.

Isabel Cuesta

20 Diciembre 2023

UAH - ISCIII

BU-ISCIII

Unidades Centrales Científico Técnicas - SGSAFI-ISCIII

## Index

- BU-ISCIII
- Conocer las aplicaciones de la secuenciación masiva en microbiología y los tipos de análisis de datos.
- Conocer los ficheros generados por plataformas como Illumina, y evaluar su calidad.
- Aprender a usar de Galaxy, herramienta web que permite el manejo y análisis de datos procedentes de técnicas de secuenciación masiva.
- 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.

## BIONINFORMATICS UNIT (BU-ISCIII)

- Sara Monzón, Biotecnóloga y Bioinformática (Analista de datos). Titulado Superior Especialista OPIS. Responsable técnico BU-ISCIII
- Sarai Varona, Bioquímica y Bioinformática (Analista de Datos). Contrato Titulado Superior asociado a proyecto (2021-2024)
- Isabel Cuesta, Dra. Biología, Bioinformática (Científico de Datos). Científico Titular de OPIS. Coordinador U. Bioinformática (BU-ISCIII)

# Index

- Que es la Bioinformática. BU-ISCIII

# Qué es la Bioinformática?

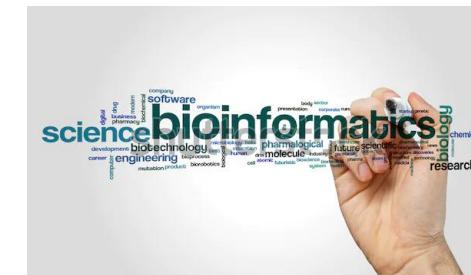
**PROBLEMAS  
BIOLÓGICOS**



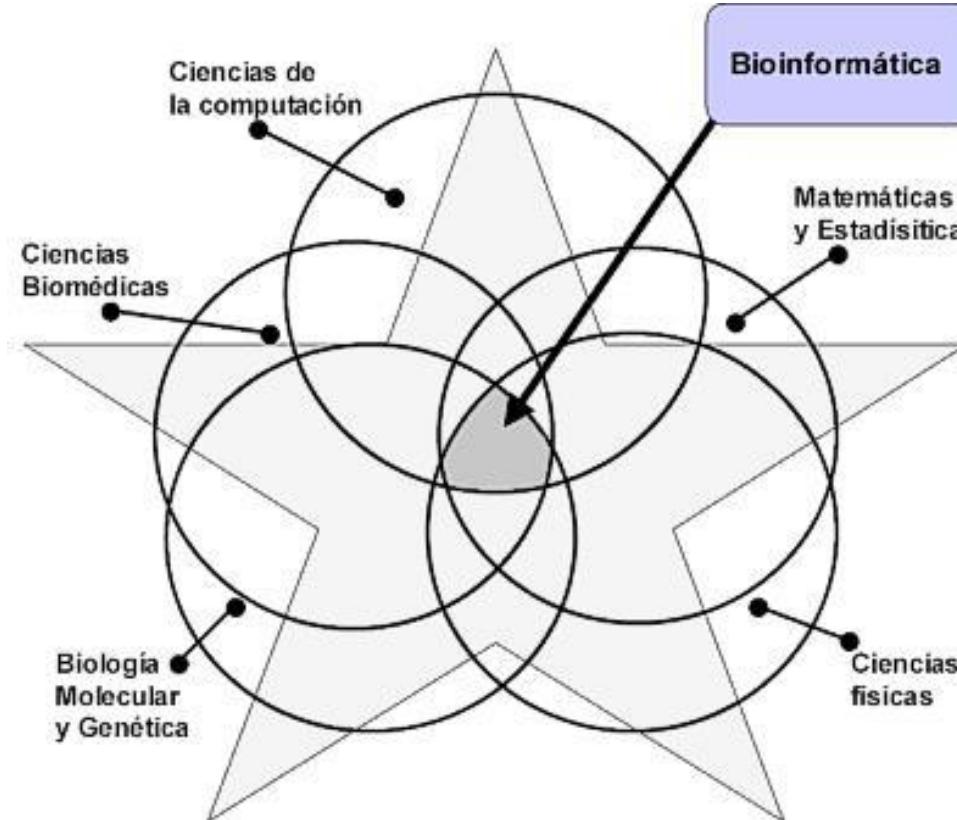
**Procesamiento  
de datos**



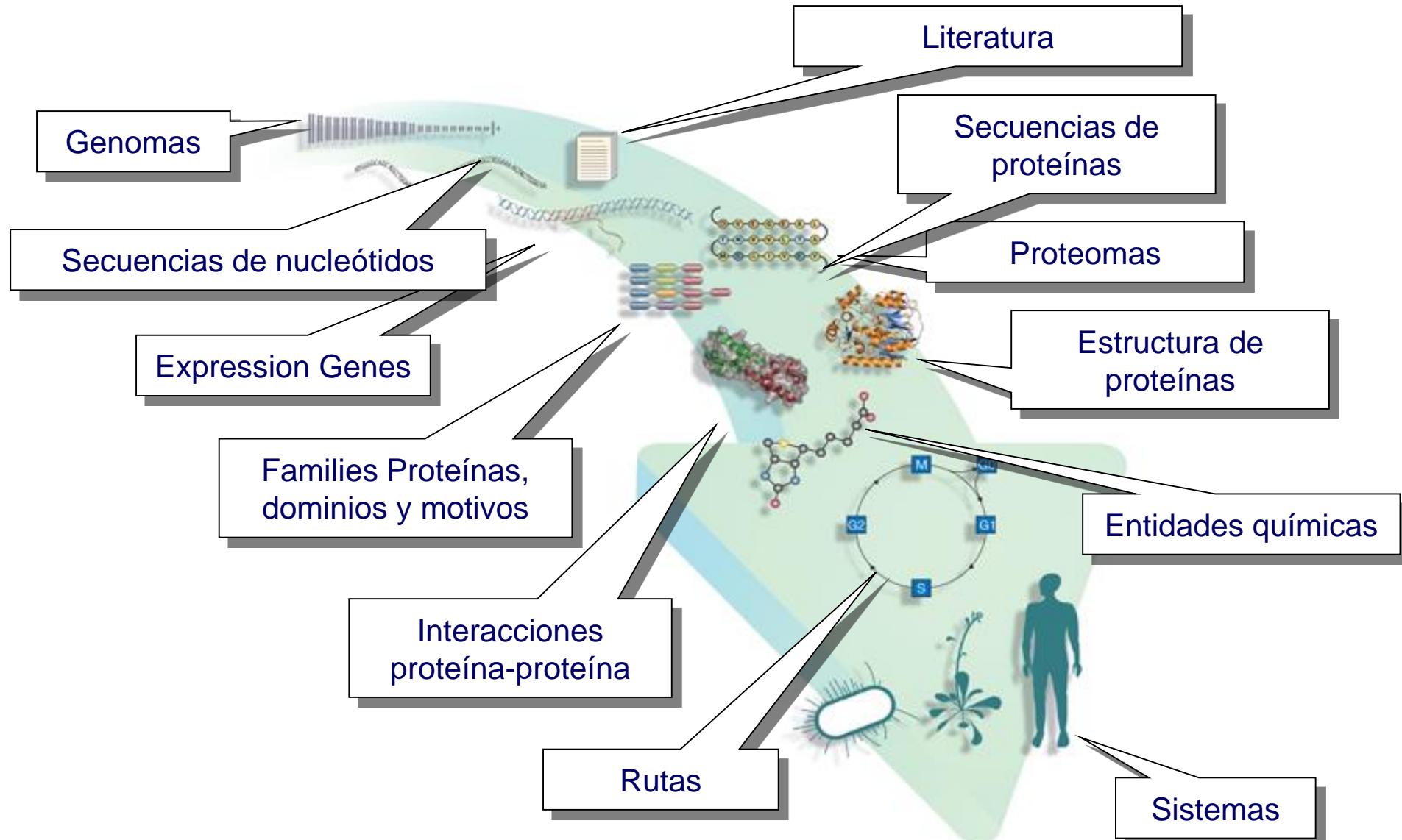
**MÉTODOS  
COMPUTACIONALES**



# Bioinformática es multidisciplinar



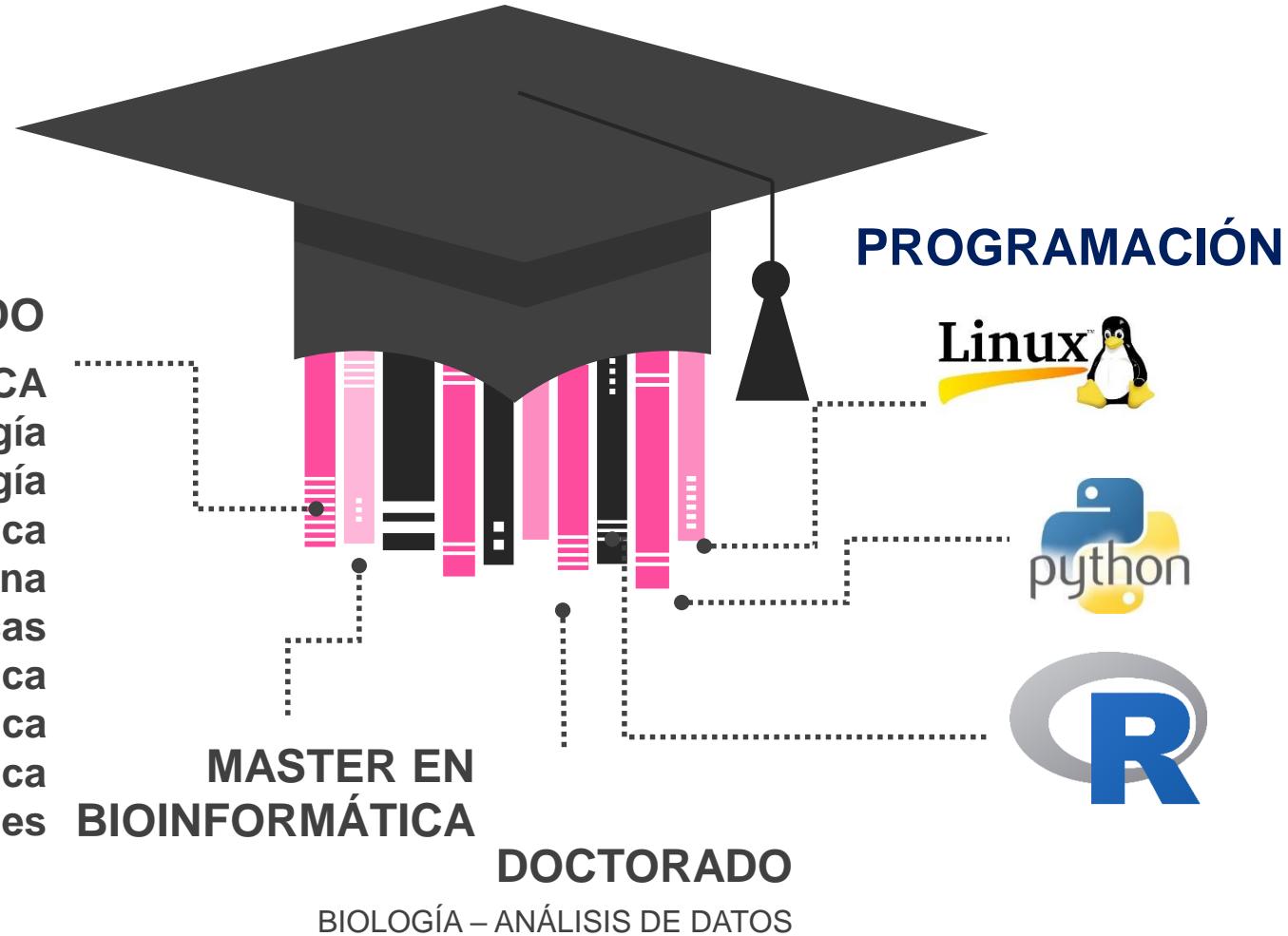
# Tipos de datos dan idea de la dimensión de la Bioinformática



# FORMACIÓN EN BIOINFORMÁTICA

Universidad  
Barcelona.

**GRADO**  
**BIOINFORMÁTICA**  
Biología  
Biotecnología  
Bioquímica  
Medicina  
Matemáticas  
Química  
Física  
Informática  
Telecomunicaciones

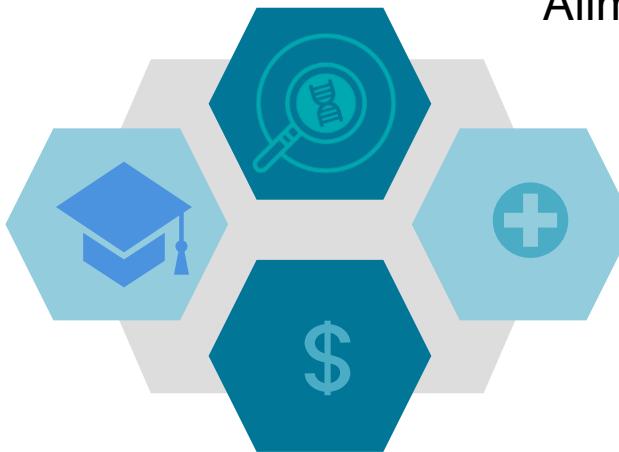


# ¿Dónde trabaja un Bioinformático?



**UNIVERSIDAD**  
Biociencias  
Informática

**CENTRO DE  
INVESTIGACIÓN**



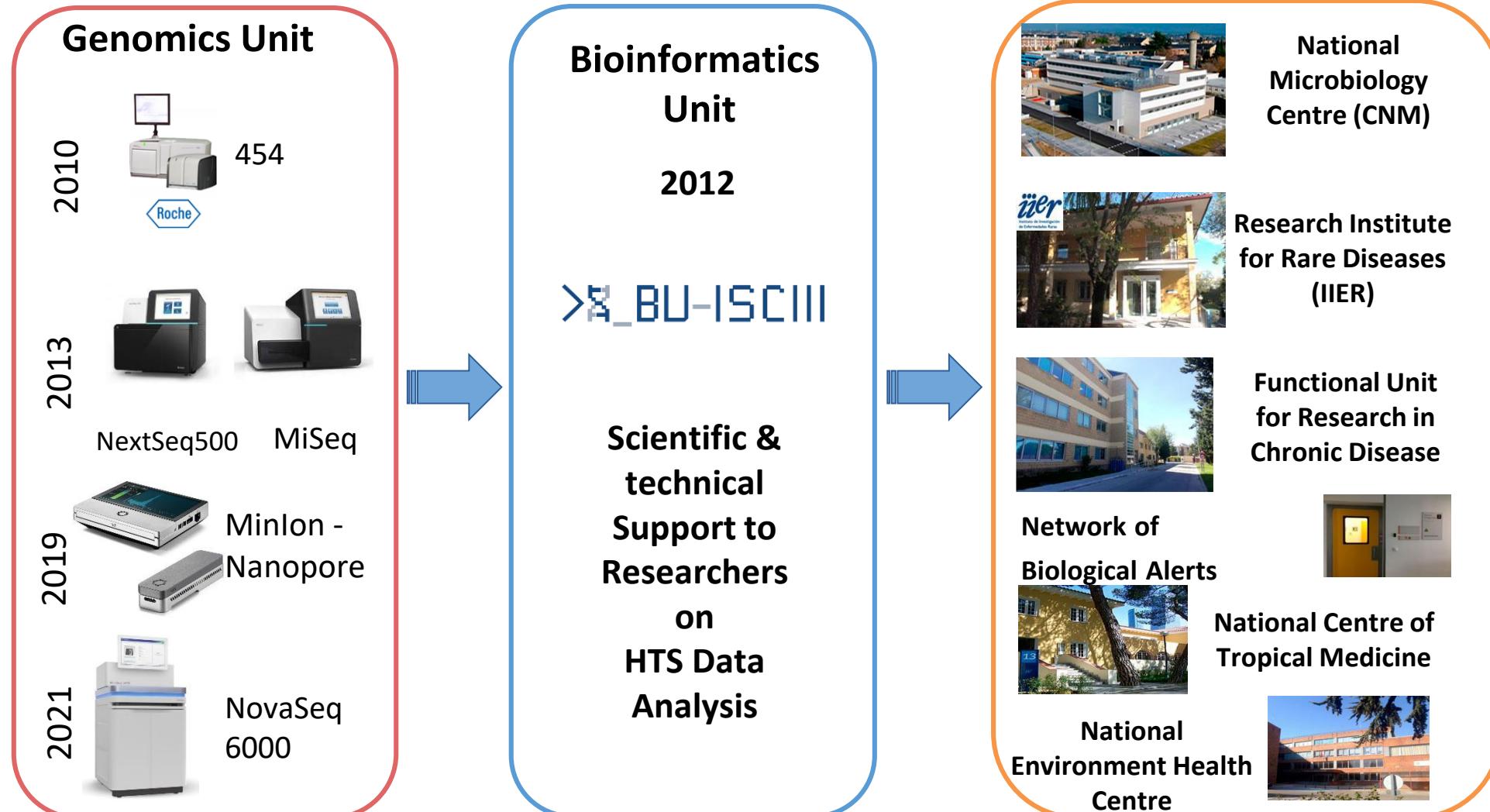
**EMPRESA**

Bioinformática  
Genética  
Genómica

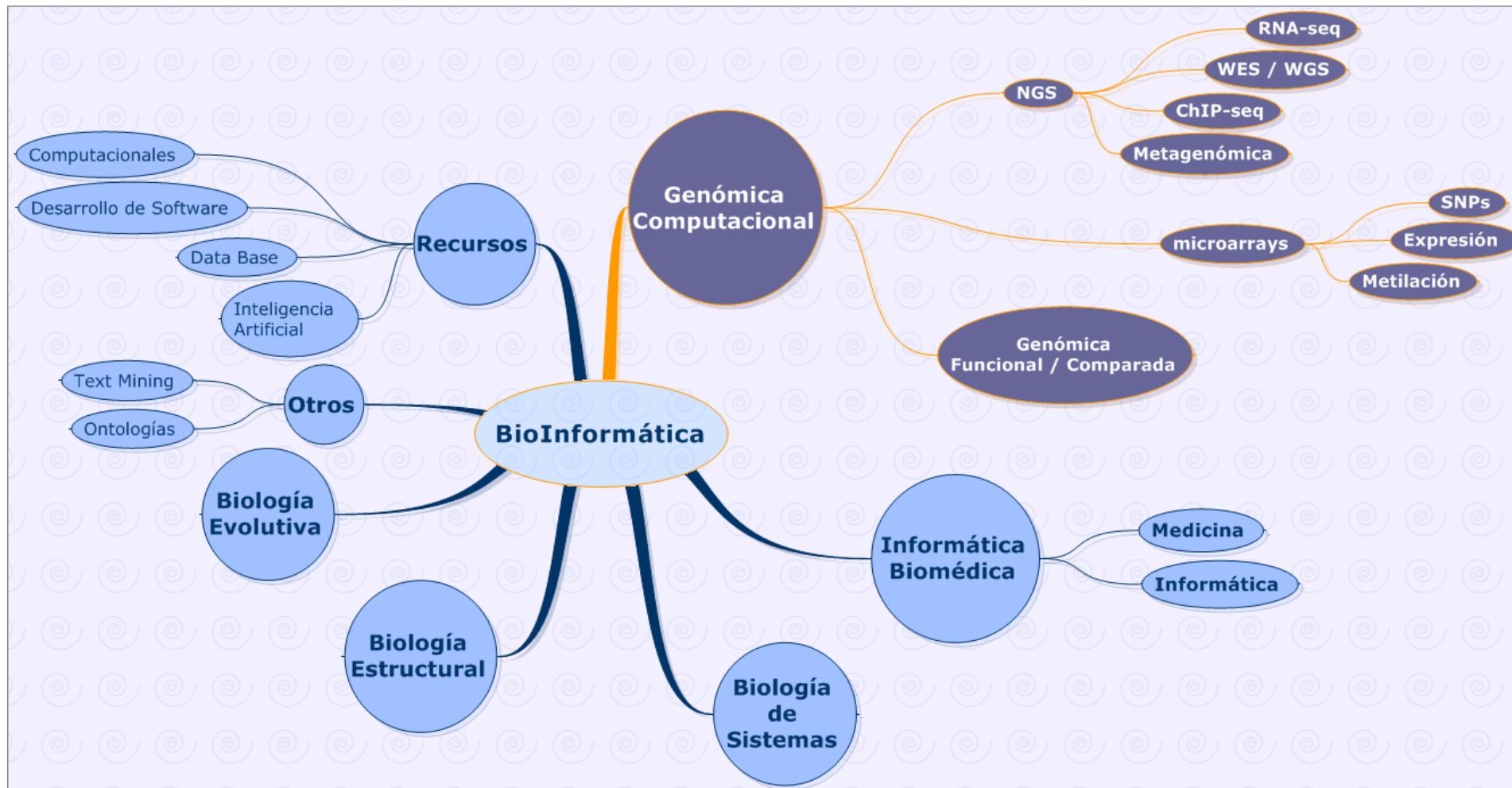
Biomedicina  
Agricultura  
Alimentación

**HOSPITAL  
BIOINFORMÁTICO  
CLÍNICO**  
Genética  
Oncología  
Cardiología

# ¿Por qué nace BU-ISCIII?



# BU-ISCIII Mission - Activities

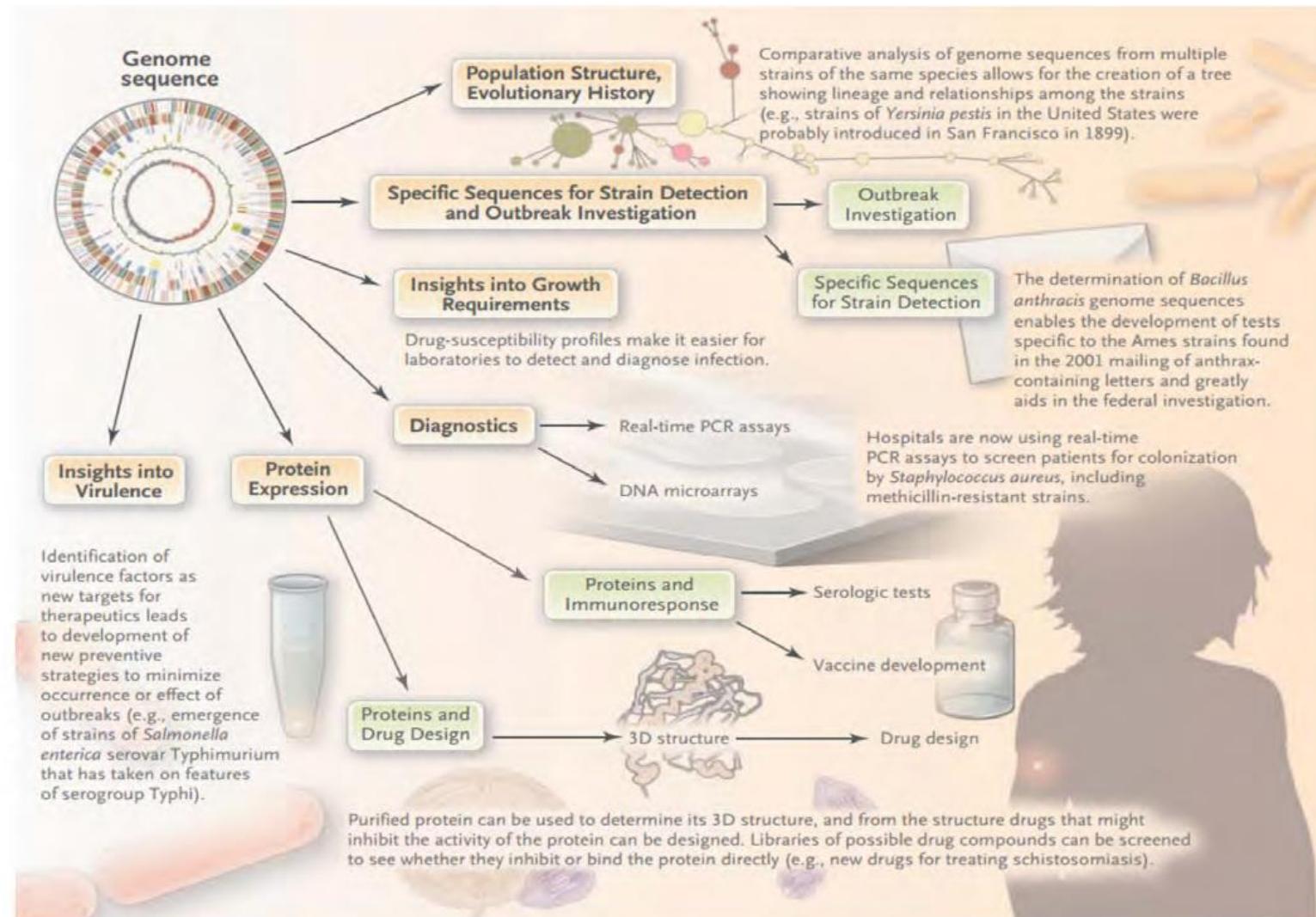


## Index

- BU-ISCIII
- Conocer las aplicaciones de la secuenciación masiva en microbiología

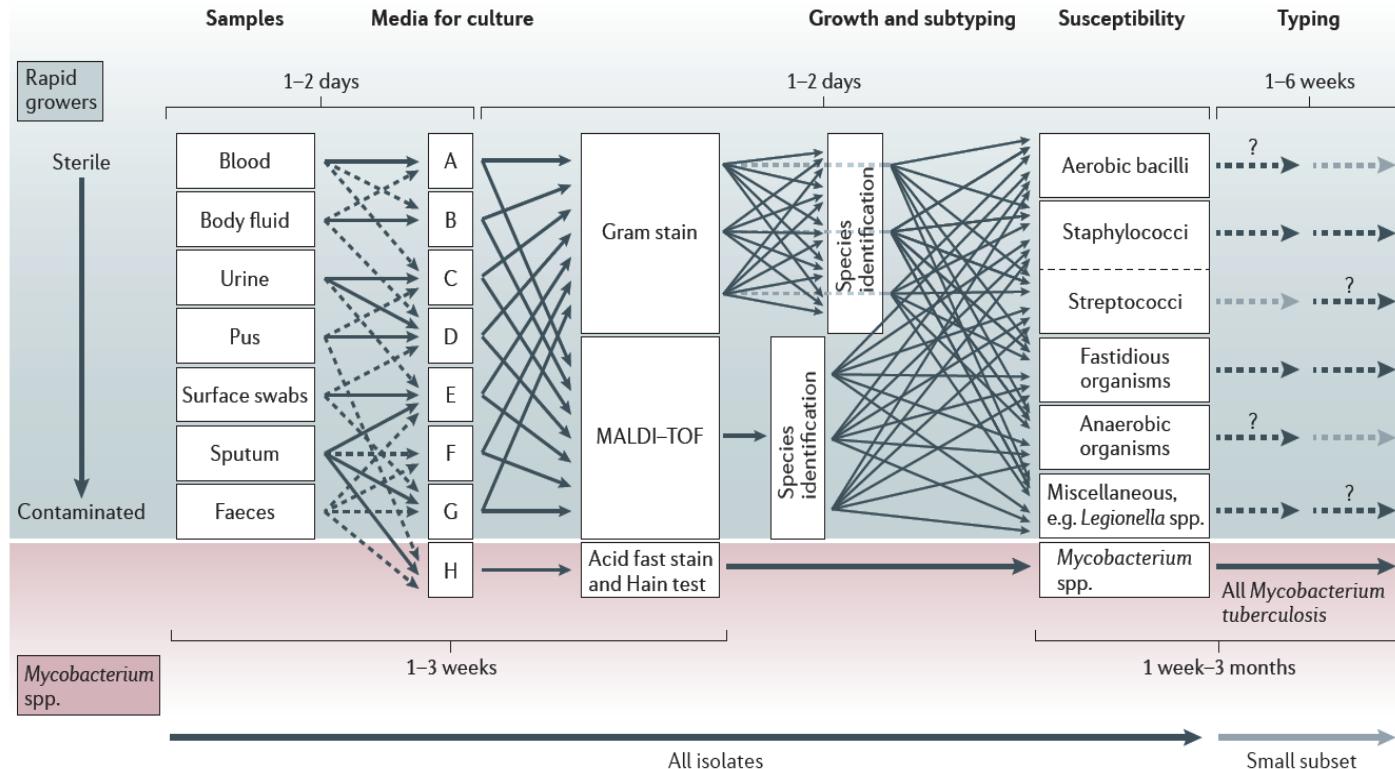
# Use of microbial genomics for tool development

Report from The American Academy of Microbiology, 2015



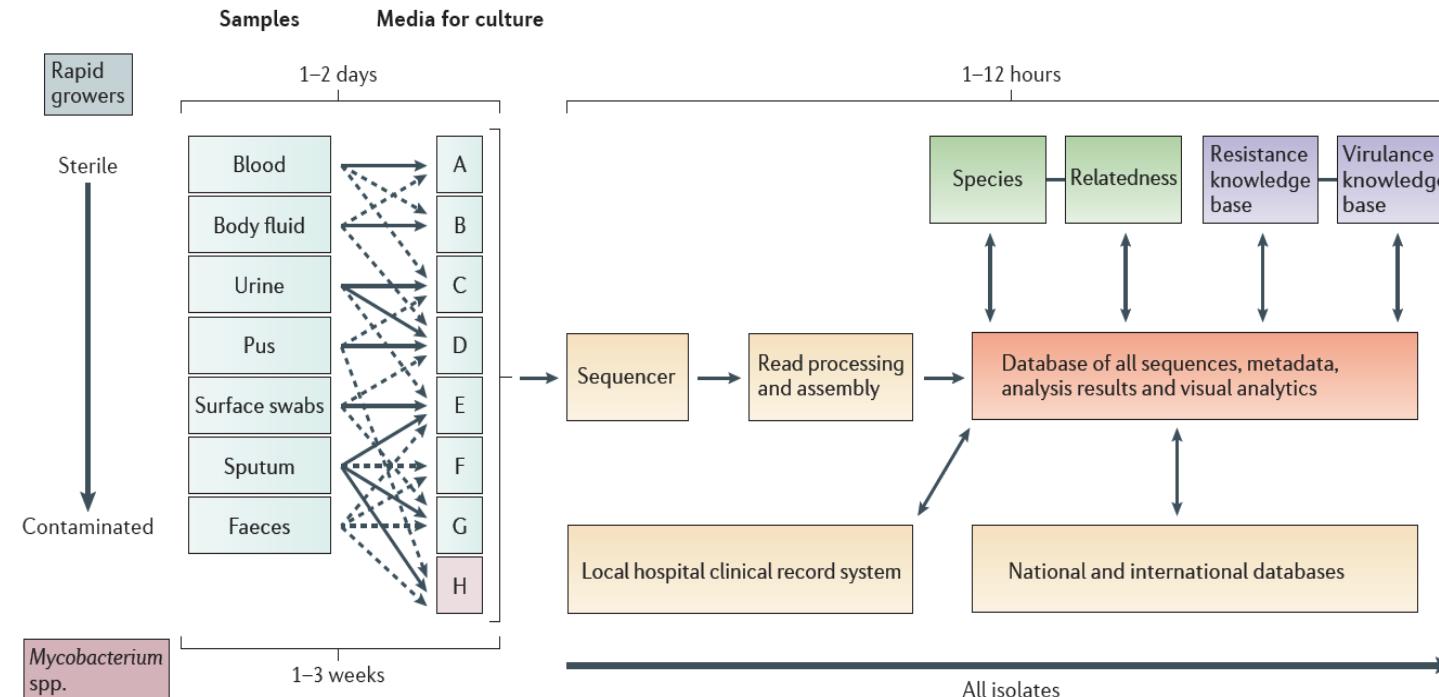
# Workflow for processing samples for bacterial pathogens

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



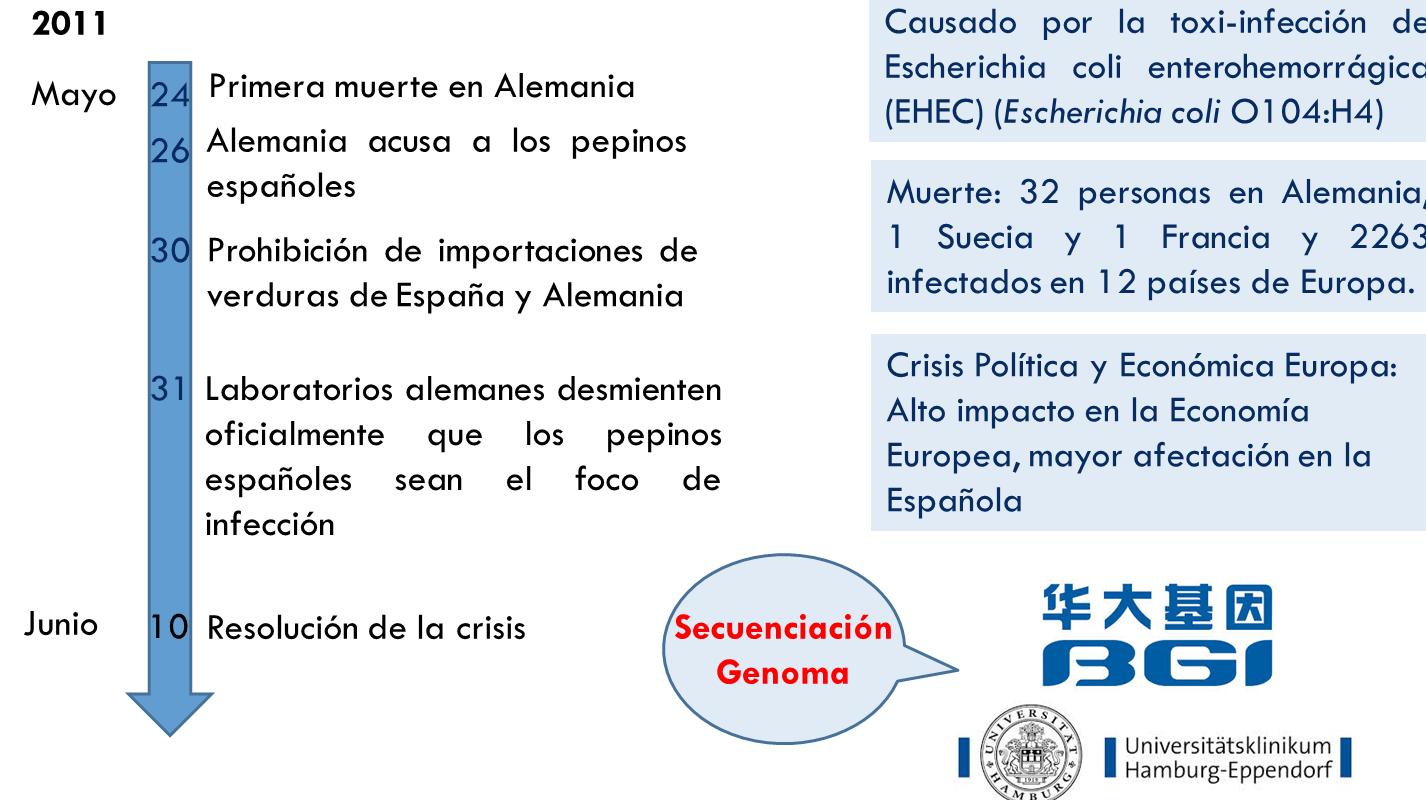
Ongoing developments in DNA-sequencing technologies are likely to affect the diagnosis and monitoring of all pathogens, including viruses, bacteria, fungi and parasites.

# The diagnostic and clinical applications of bacterial WGS



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

# Foodborne outbreak identification “Crisis del pepino”



# 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

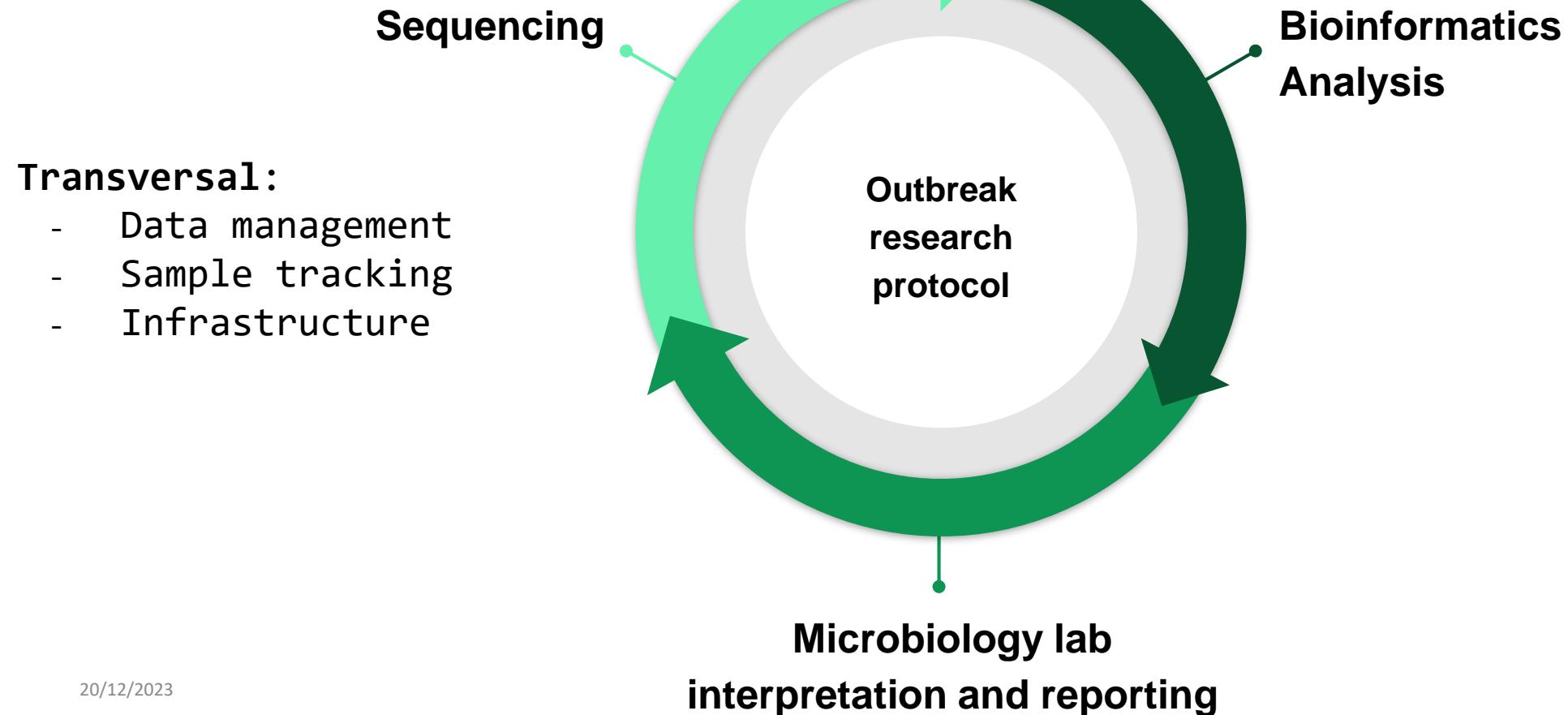
EFE 25.08.2019

- 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

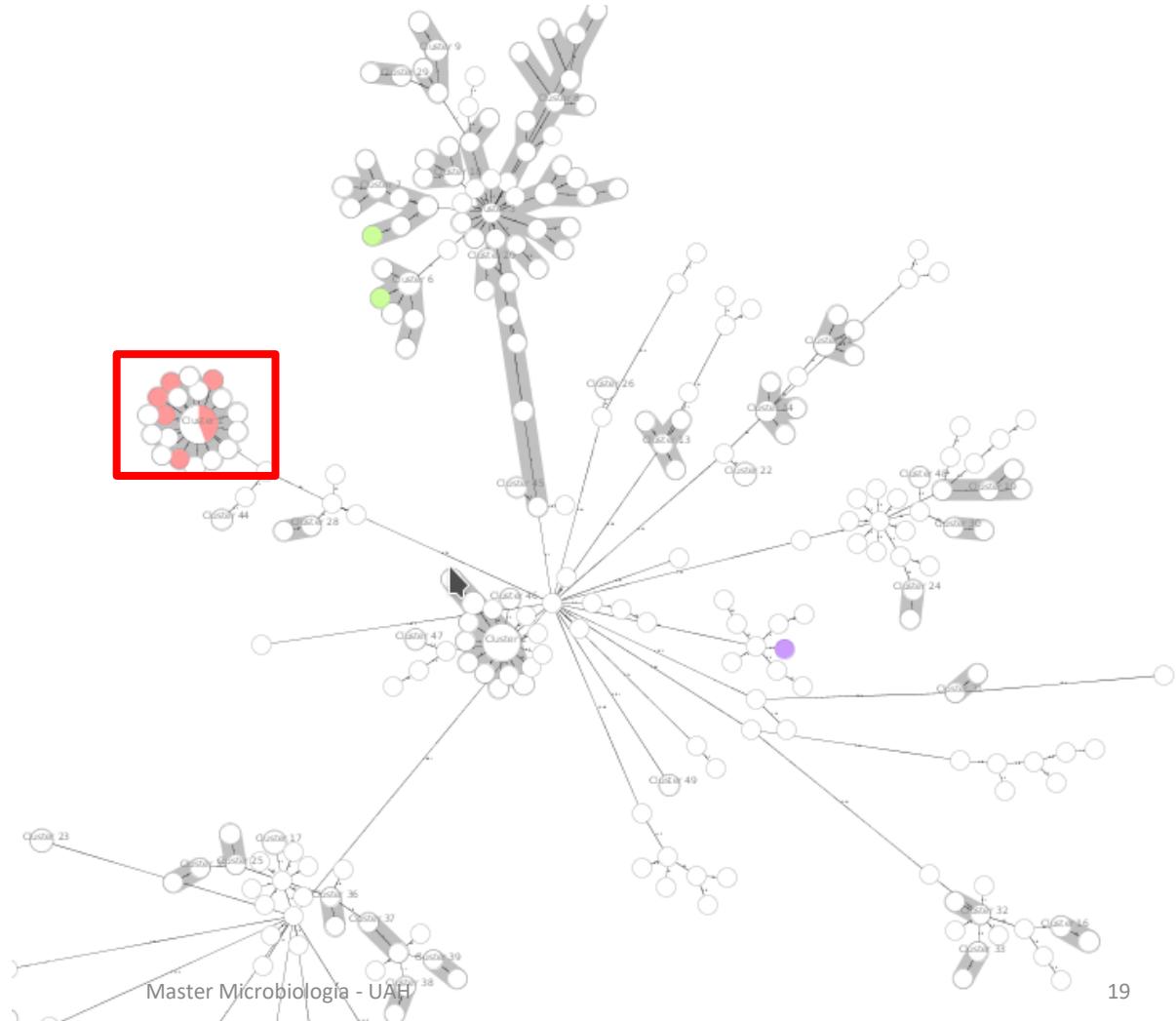


# Andalusian Listeria Outbreak

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

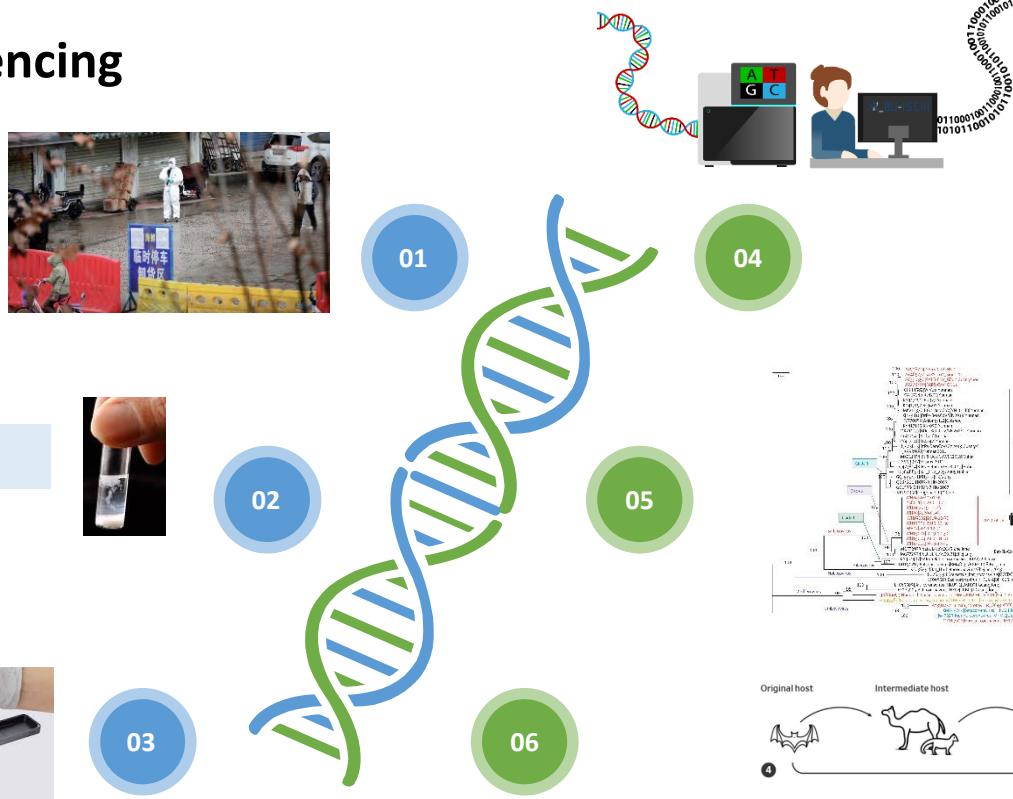
## Results:

- 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)



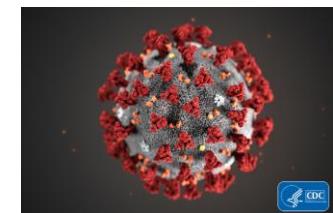
Meta-transcriptomic library

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)

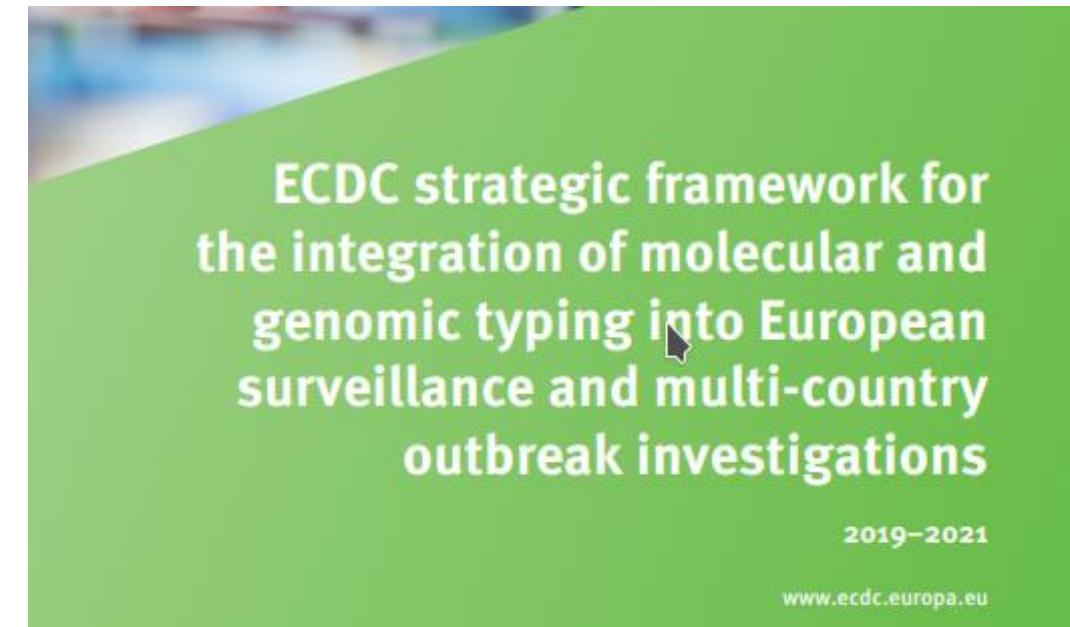


Mission: Provide support to the National Health System and the different Spanish Regions in the diagnosis and control of infectious diseases. In order to fulfill this mission it acts as Reference center offering a series of scientific activities:

- 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

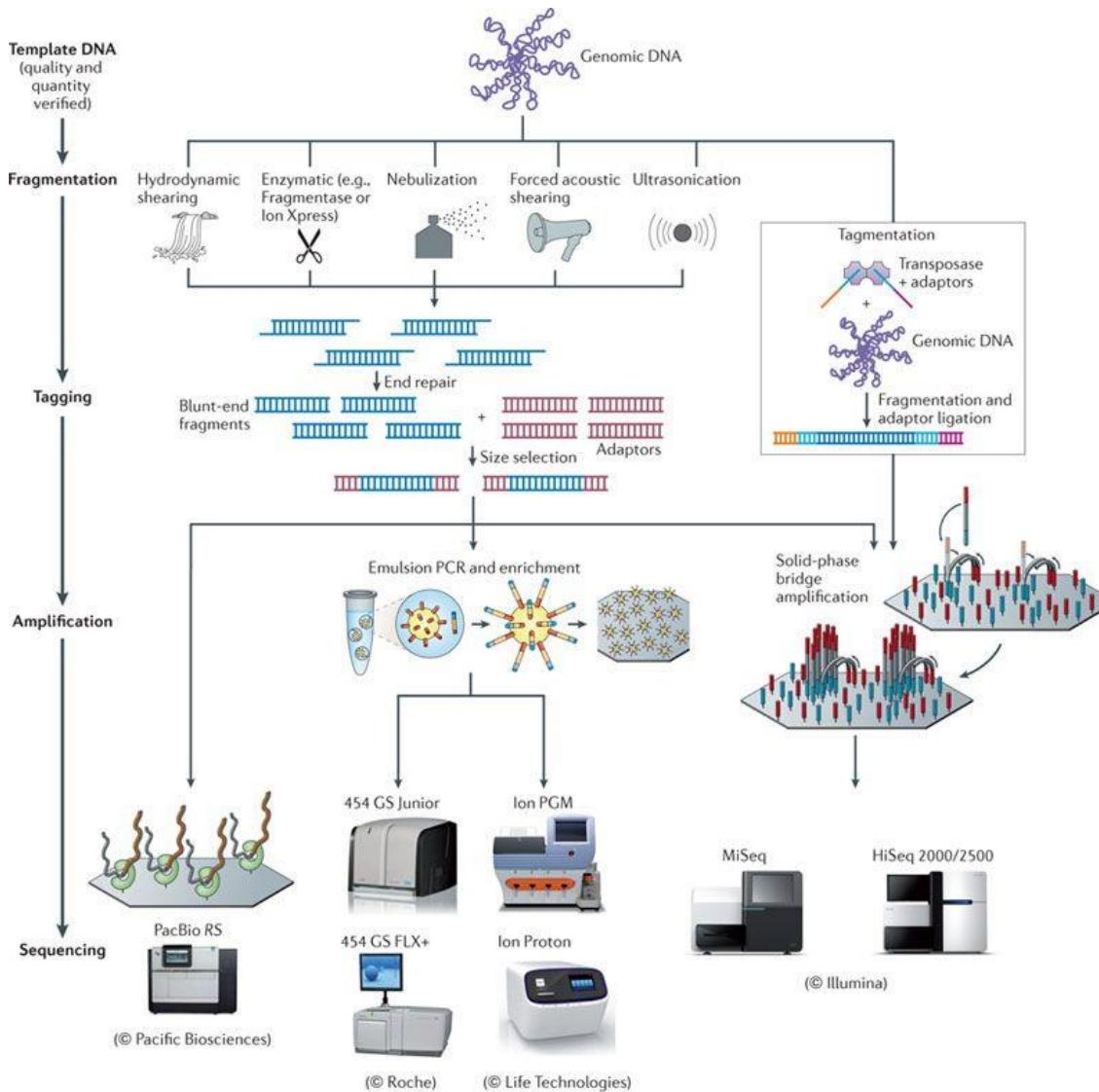
## Index

- BU-ISCIII
- Conocer las aplicaciones de la secuenciación masiva en microbiología
- Repaso conceptos secuenciación masiva

# DNA sequencing technologies 2006-2023

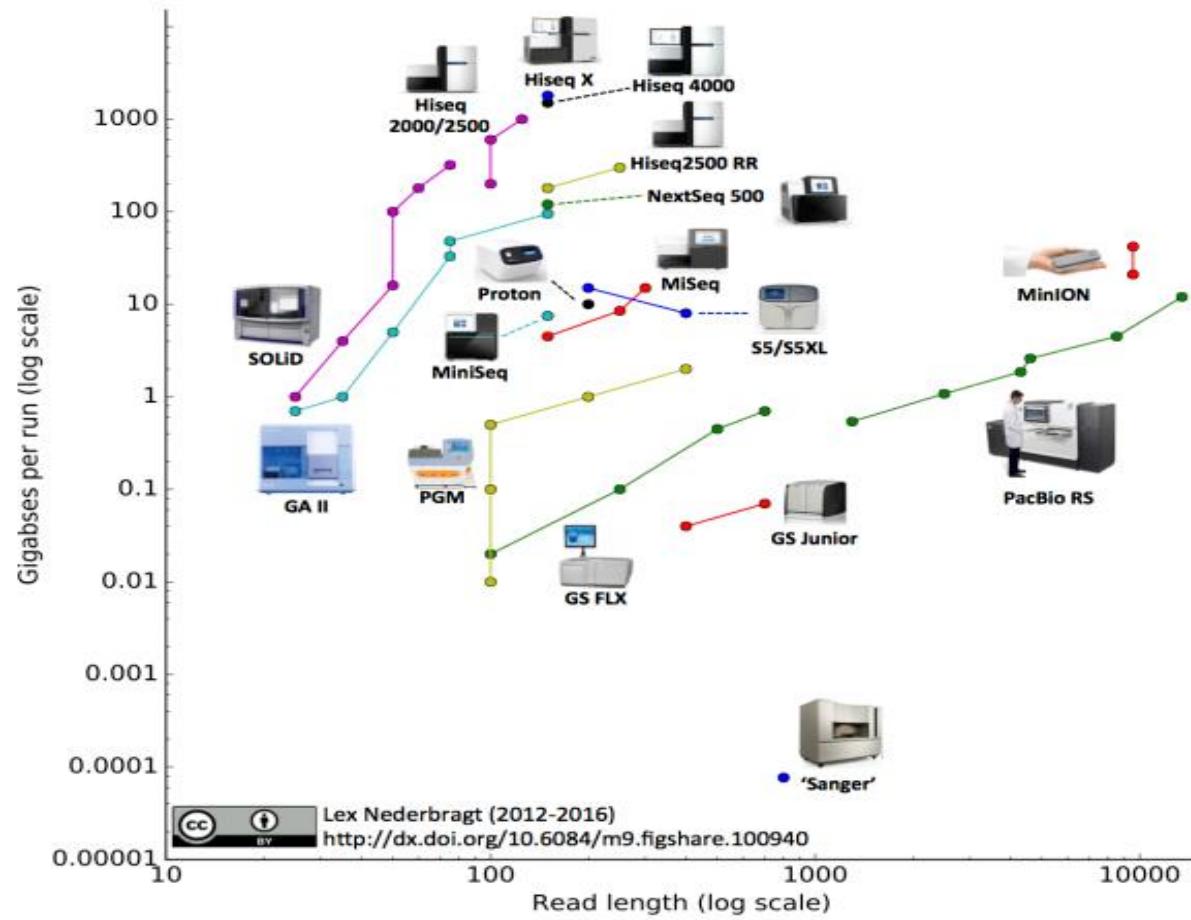


# High-throughput sequencing process



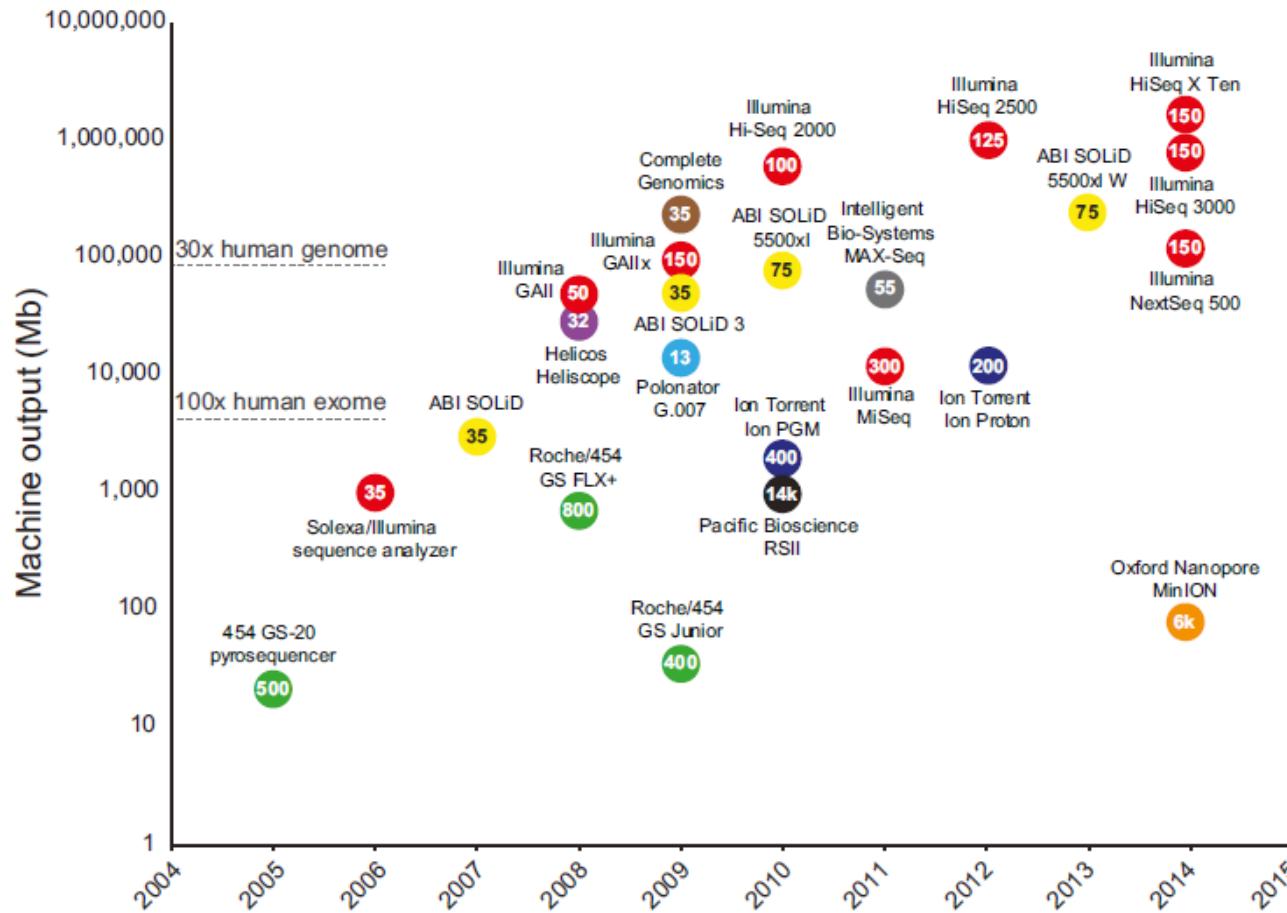
Loman et al, 2012

# High-Throughput Sequencing Technologies



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

# High-Throughput Sequencing Technologies



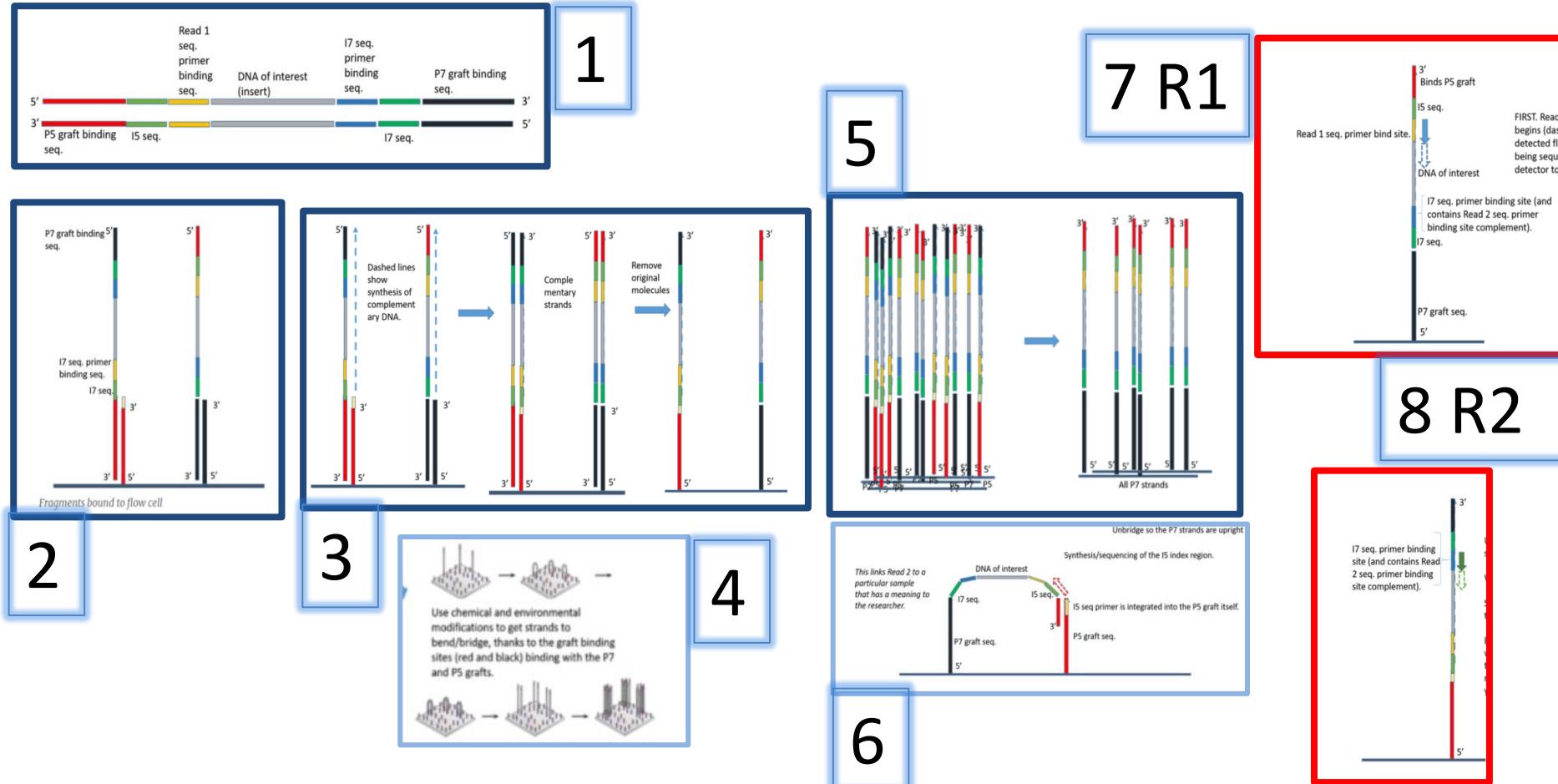
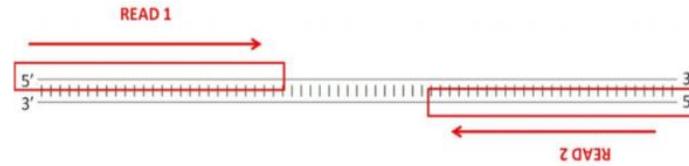
Numbers inside data points denote current read lengths.  
Sequencing platforms are color coded.

Reuter et al., Mol Cell 2015

## NGS PLATFORMS, main characteristics

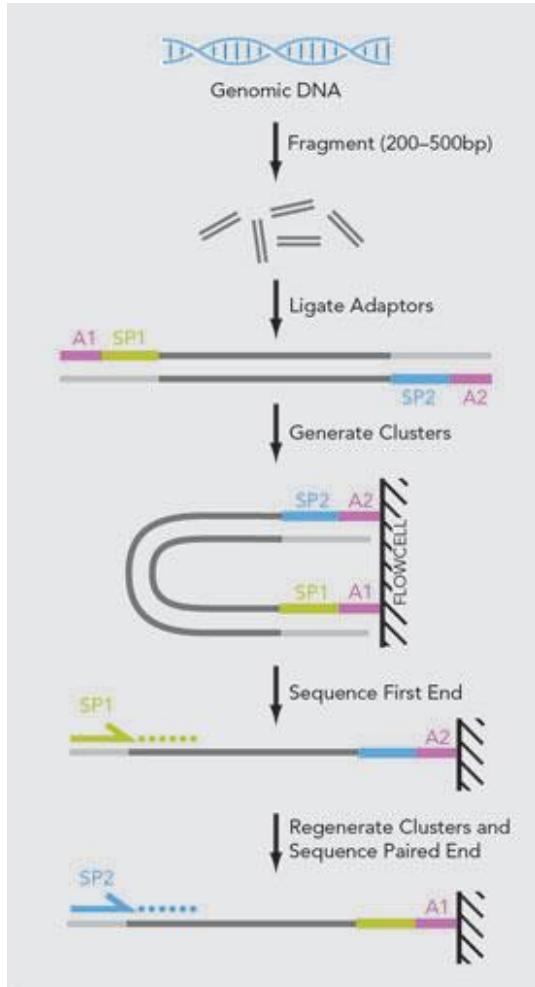
- Numero de bases que secuencia
- Numero lecturas → aplicaciones
- Longitud de las lecturas -→ importante para las aplicaciones ensamblado genomas, de illumina a PacBio
- Error de la base → Corrección con profundidad de lectura
- Formato fichero salida
- Software dedicado, universal fastq

# Illumina sequencing



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

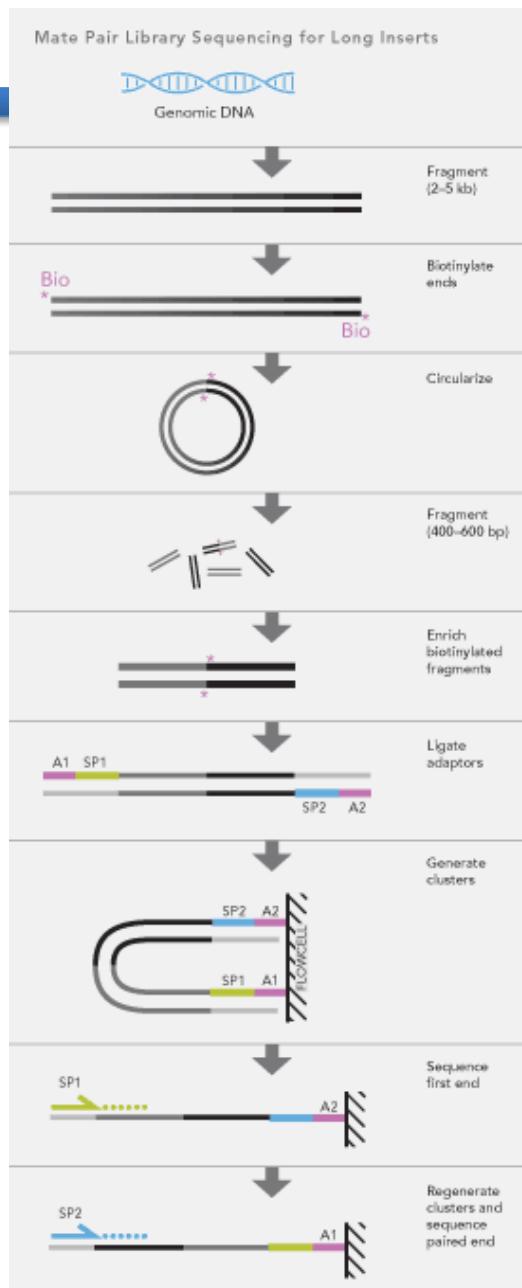
# Que es Pair-end?



**Secuenciación de un fragmento (bp)**

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

# Que es Mate-pair?



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.

**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

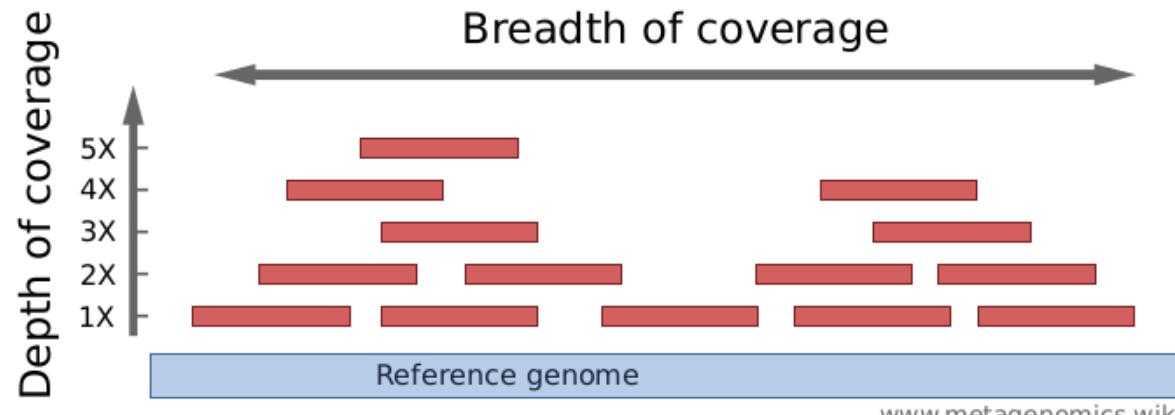
How strong is a genome "covered" by sequenced fragments (short reads)?

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.



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

# 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

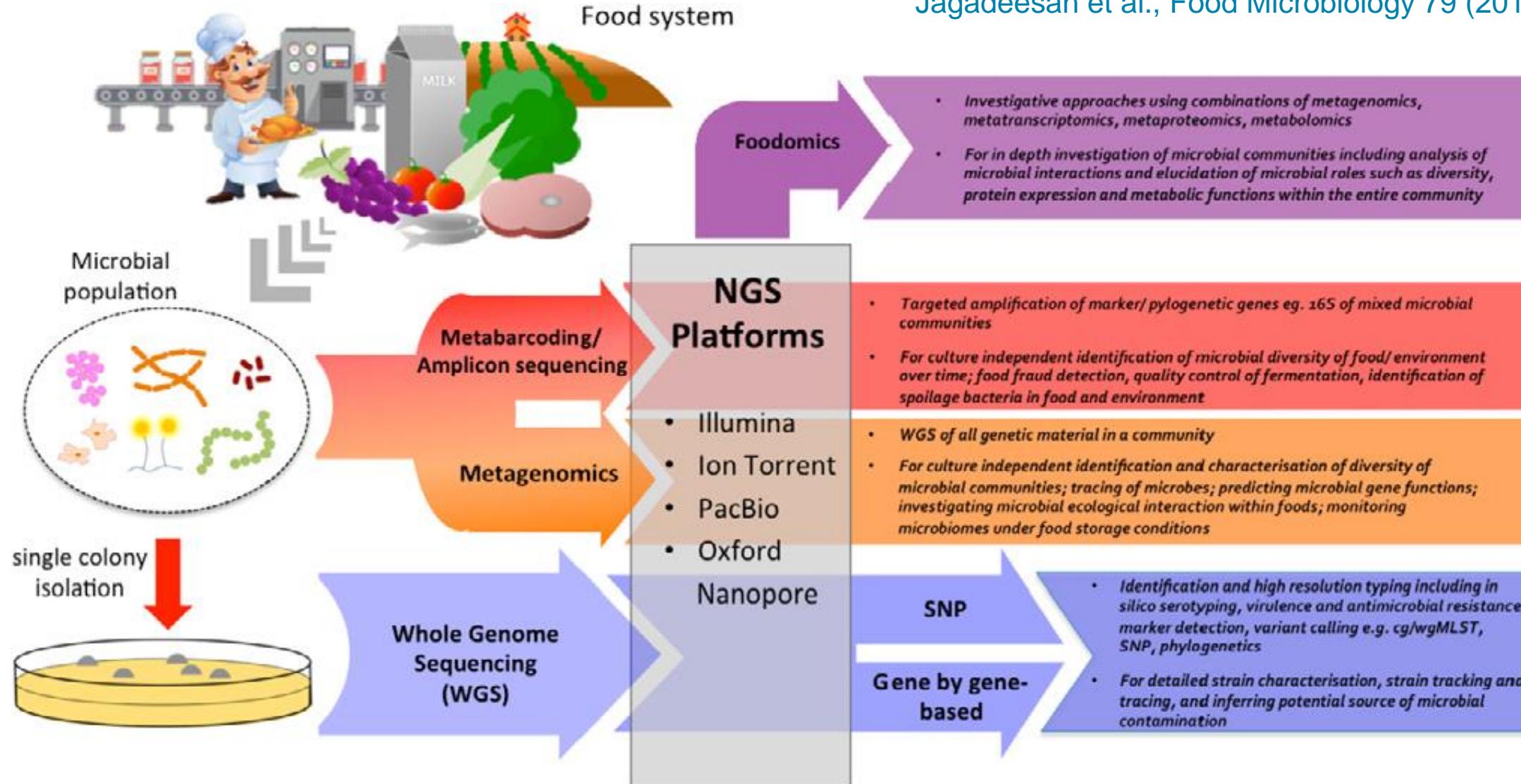
### Increase number of raw reads

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

## Index

- BU-ISCIII
- Conocer las aplicaciones de la secuenciación masiva en microbiología
- Repaso conceptos secuenciación masiva
- Estrategias basadas en preparación de librería

# Estrategias basadas en preparación de librería



# Estrategias basadas en preparación de librería

## 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

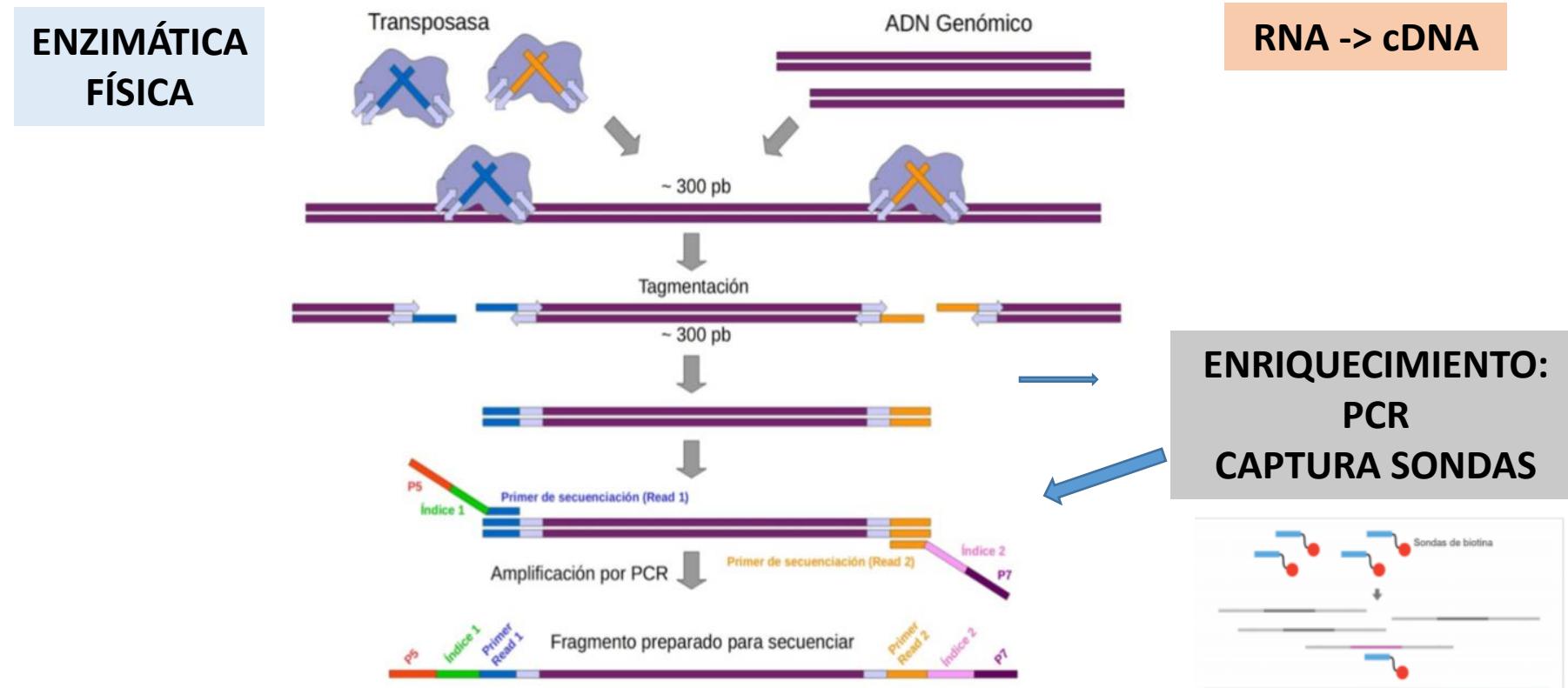
## SECUENCIACIÓN GENOMAS

1. Metagenómica

## IDENTIFICACIÓN MICROORGANISMOS

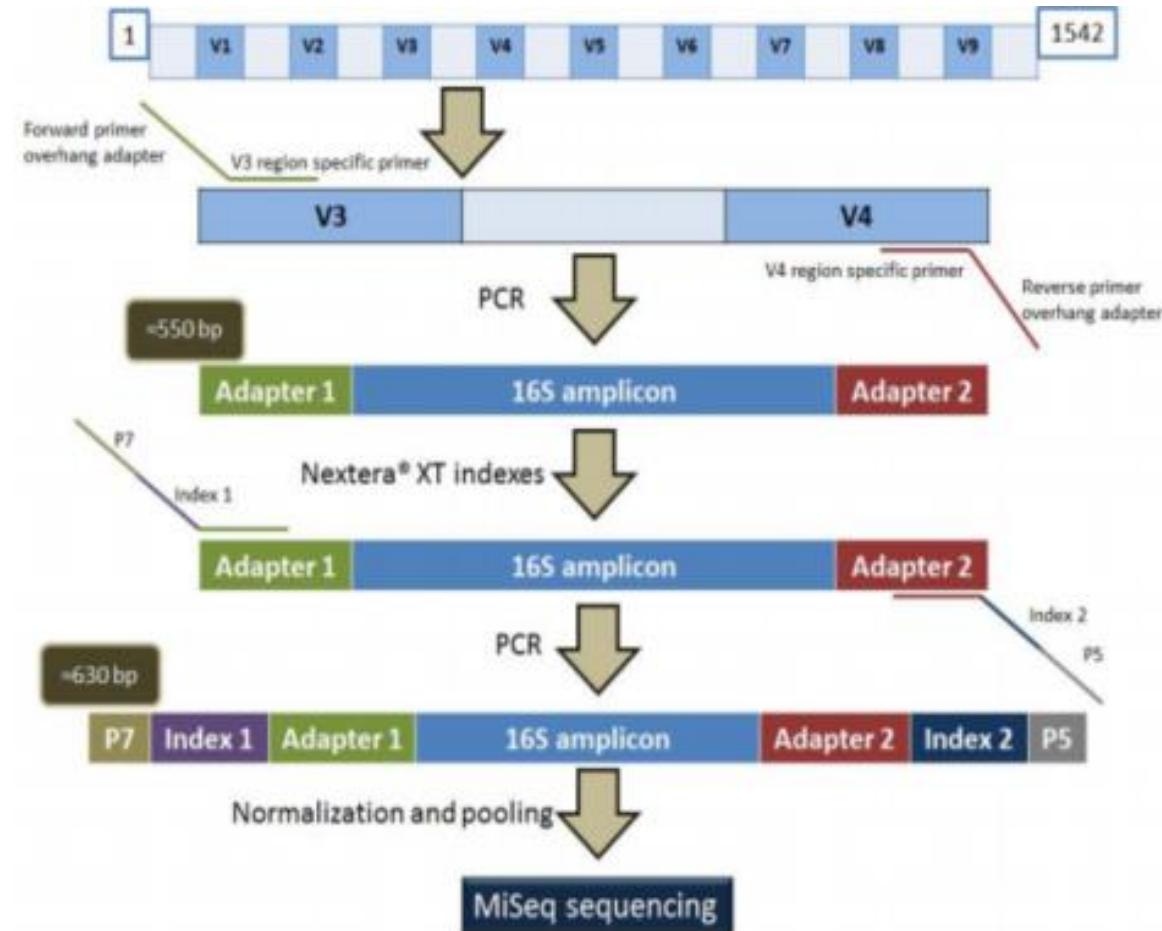
1. Metataxonomía

# Estrategias basadas en preparación de librería



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)

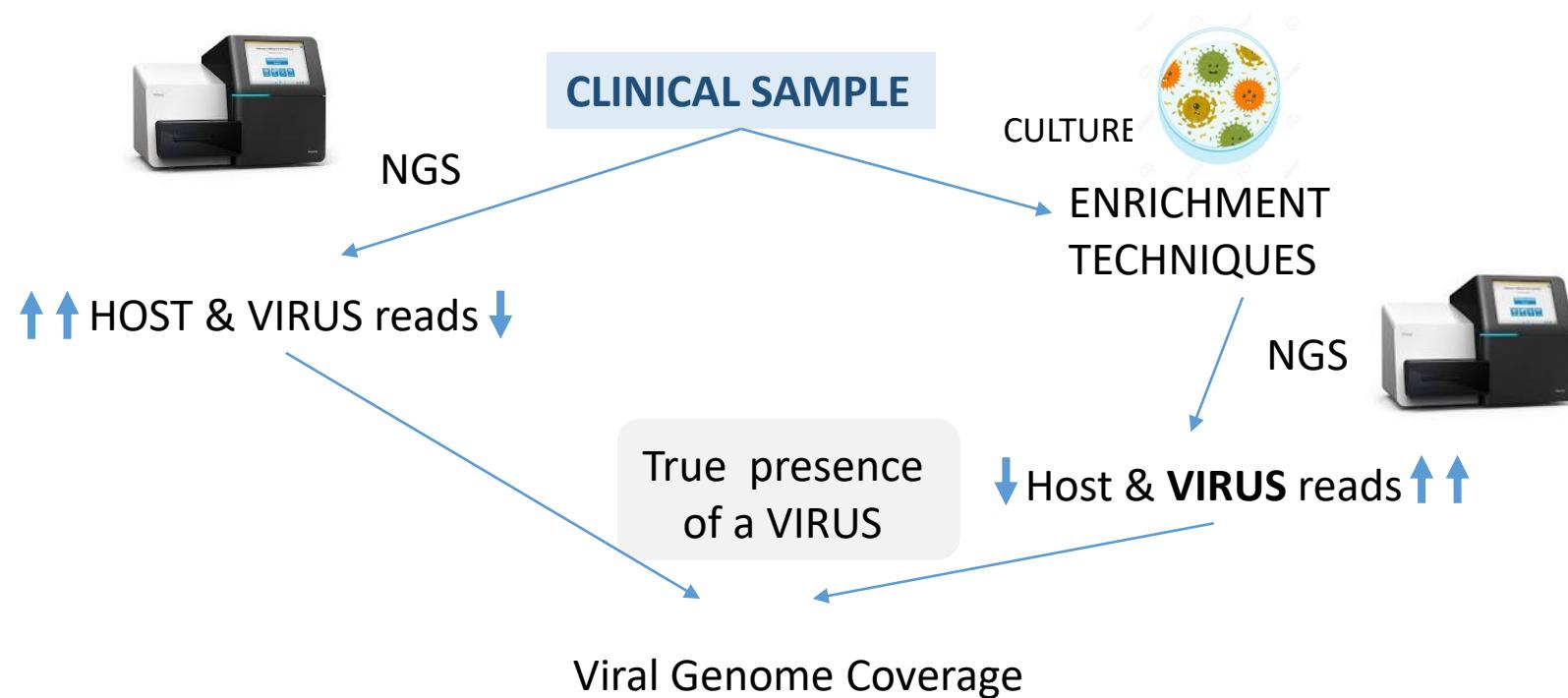
# PREPARACIÓN LIBRERÍA, rRNA 16S, caracterización microbiota



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

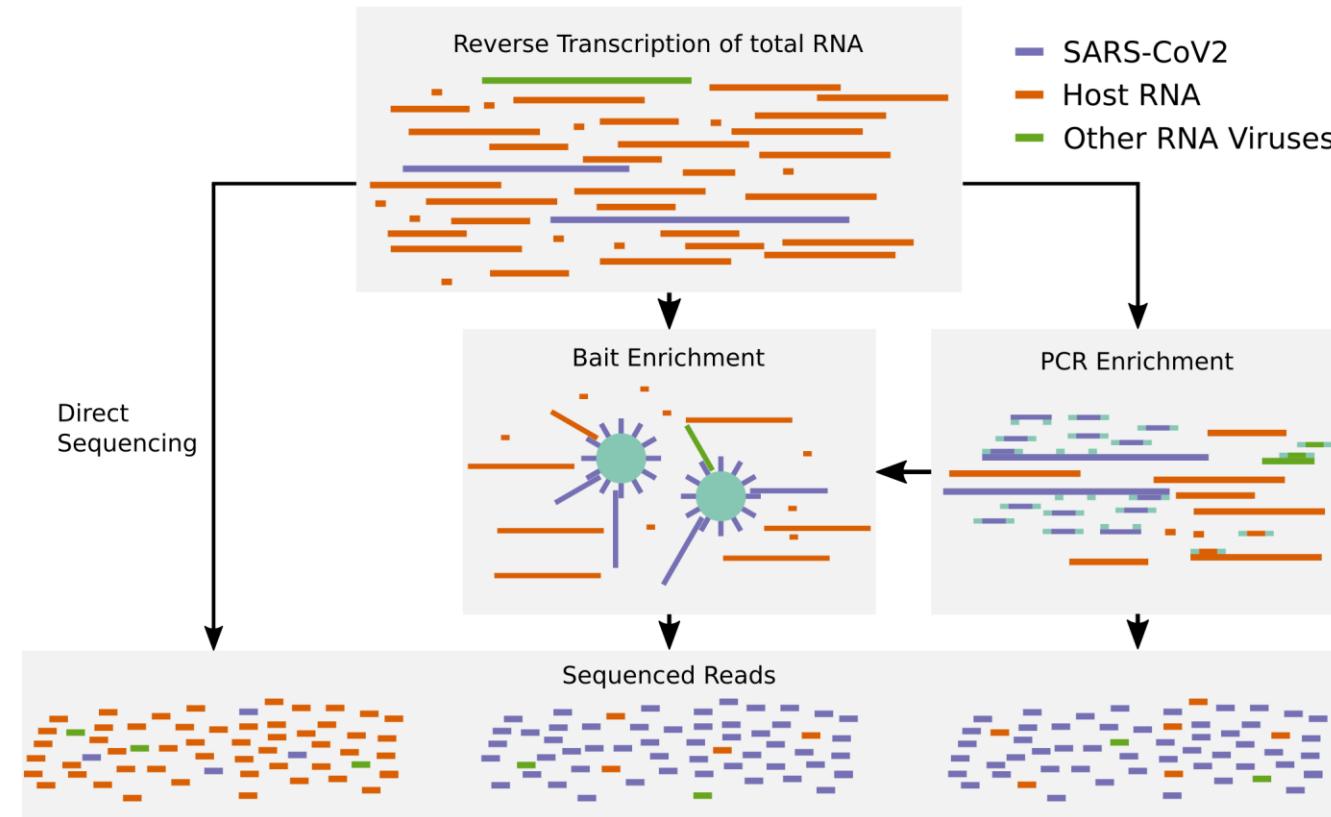
- Nucleic acid amplification
- Library preparation
- High throughput sequencing platforms
- Data analysis

# 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

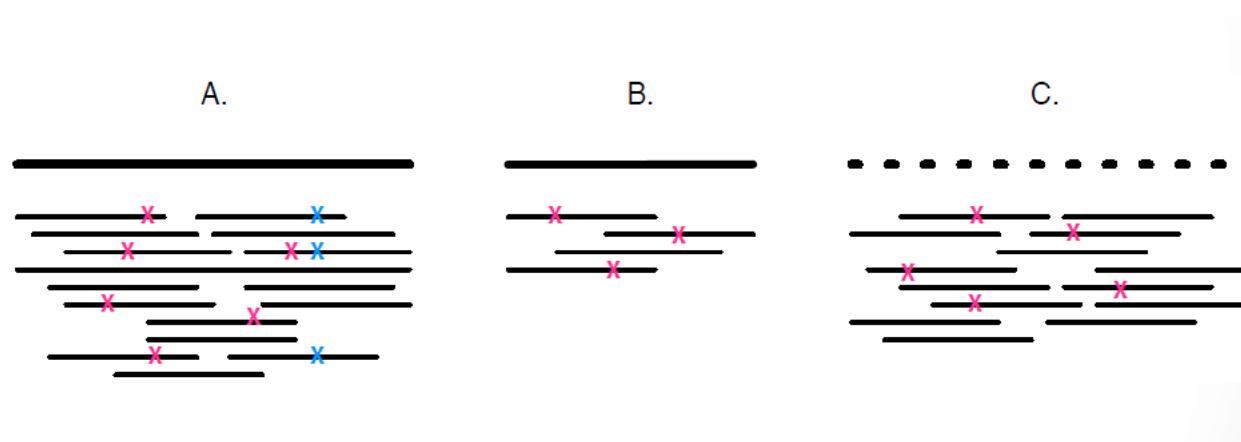


## Index

- BU-ISCIII
- Conocer las aplicaciones de la secuenciación masiva en microbiología
- Repaso conceptos secuenciación masiva
- Estrategias basadas en preparación de librería
- Tipos de análisis de datos.

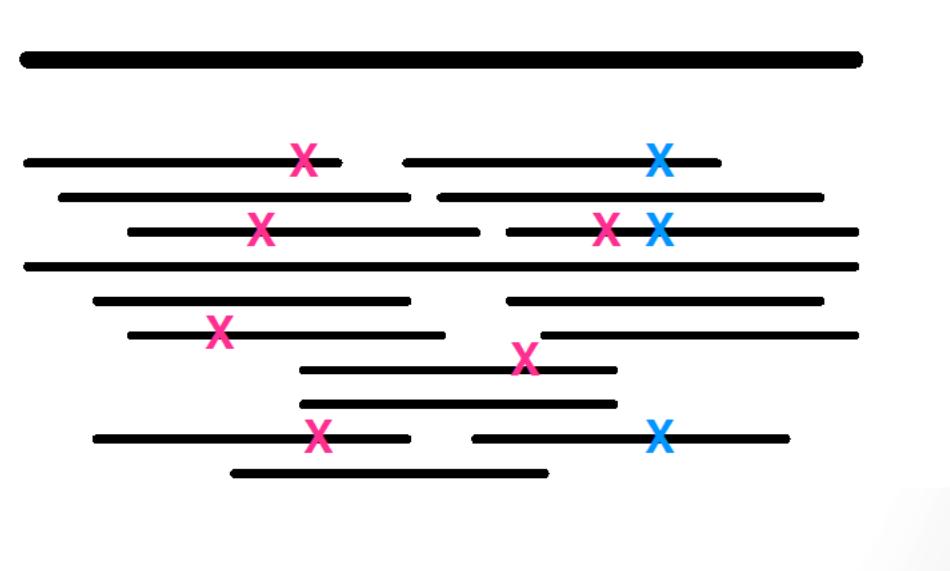
# Básicamente tres problemas

Resecuenciación, Conteo y ensamblado



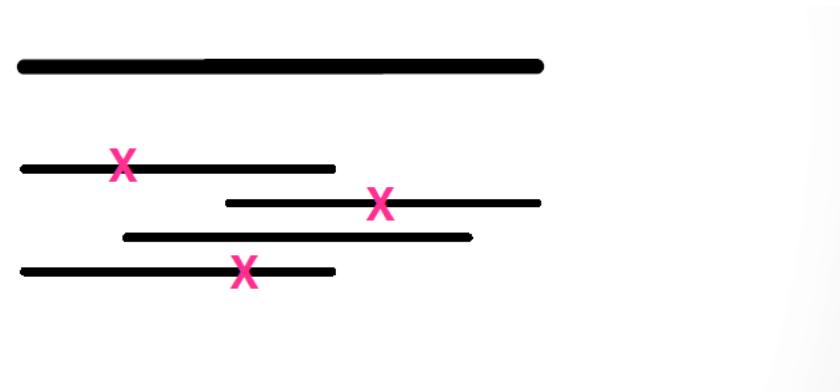
## Resecuenciación

Conocemos el genoma, genoma de referencia, y queremos identificar variaciones SNPs (azul), en un background de errores (rosa). Obtenemos secuencia genoma consenso.



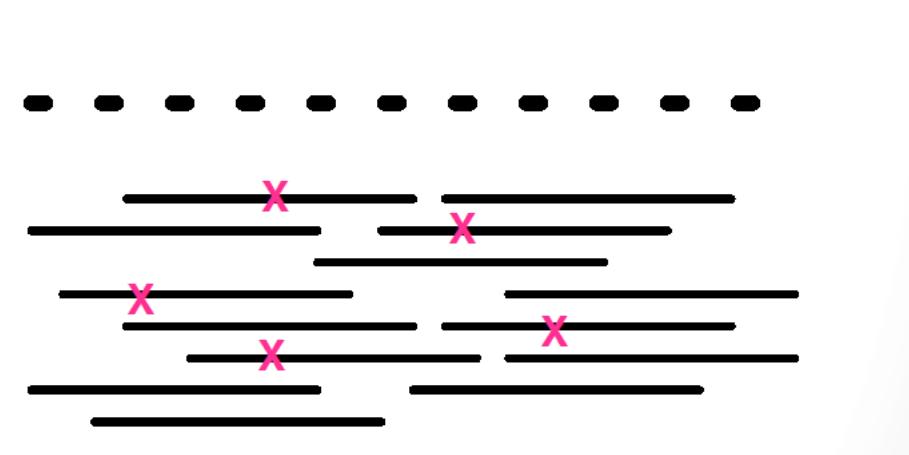
## Conteo

Número de lecturas de un gen (amplicón) o mRNA (RNAseq). Equivalente a expresión en Microarrays.



## Ensamblado

No hay genoma de referencia y lo construimos de novo

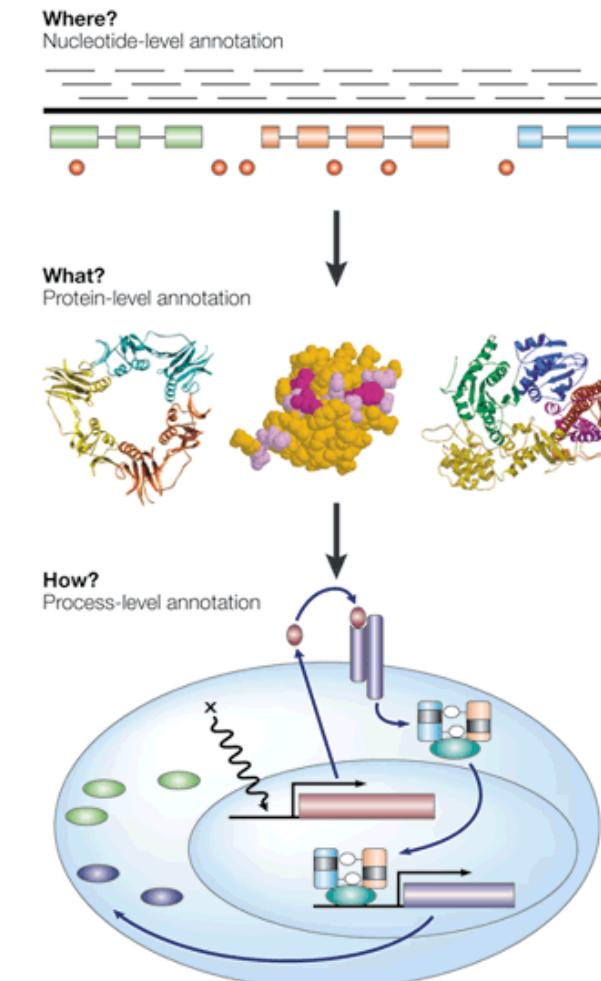


# Anotación

Genome annotation is the process of attaching biological (and positional) information to sequences. It consists of three main steps:

- identifying portions of the genome that do not code for proteins
- Identifying coding elements on the genome, a process called gene prediction
- attaching biological information to these elements

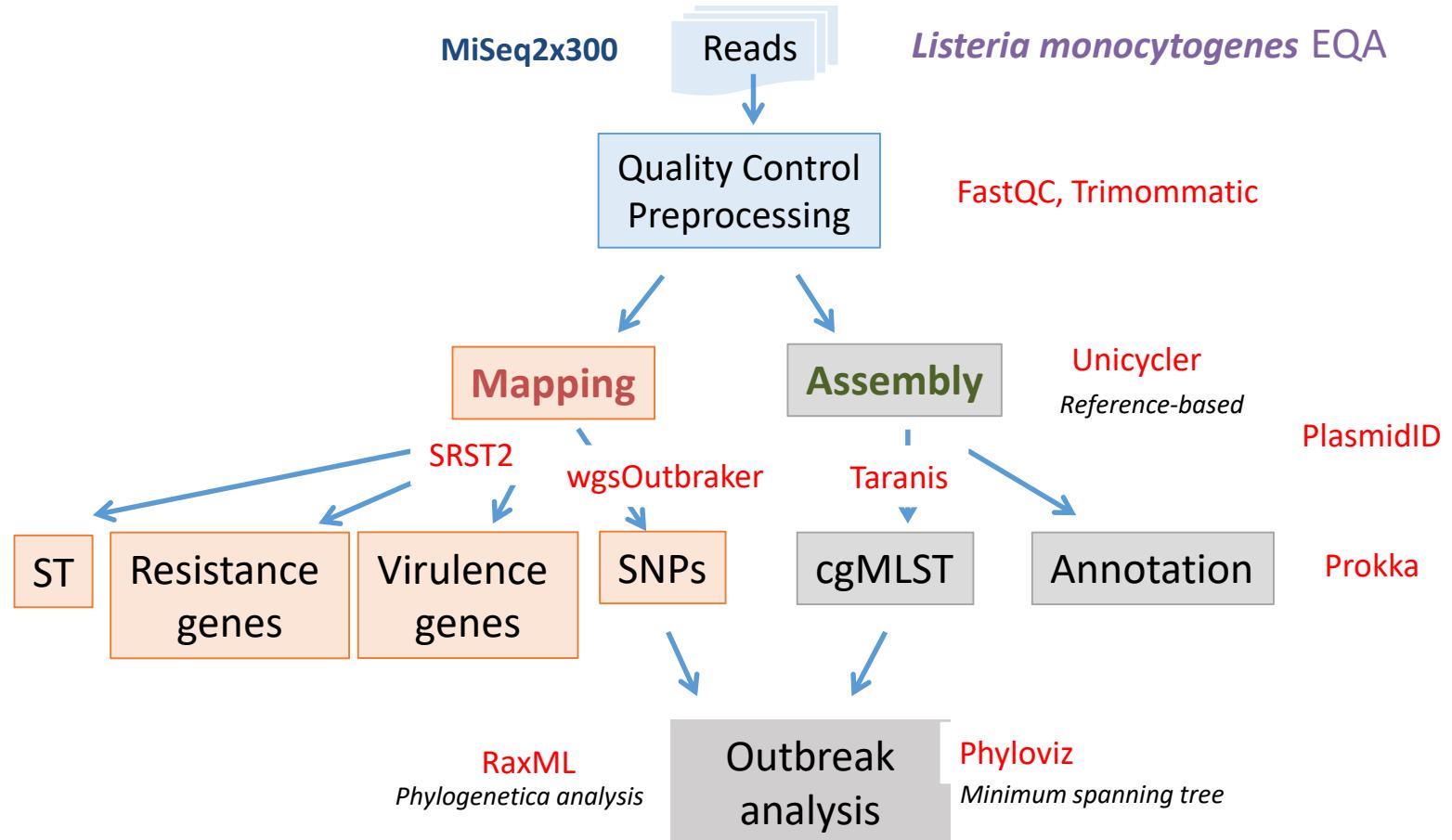
<https://galaxyproject.github.io/training-material/topics/genome-annotation/tutorials/genome-annotation/tutorial.html>



# Bioinformatics analysis in microbial genomics

- SPECIE IDENTIFICATION
  - WGS - Kmers analysis
  - TARGET METAGENOMIC, rRNA - MICROBIOTA
- ASSEMBLY GENOME
  - de NOVO or REFERENCE -BASED
  - cgMLST, wgMLST - MINIMUM SPANING TREE
  - METAGENOMIC - HOMOLOGY -BASED
- VARIANT CALLING
  - REFERENCE GENOME SELECTION
  - HAPLOID GENOME
  - LOW FREQUENCY VARIANT - QUASISPECIES
  - SNPs MATRIX - PHYLOGENETIC ANALYSIS
- STRUCTURAL AND FUNCTIONAL ANNOTATION
  - RESISTOME, VIRULOME, SEQUENCE-TYPE

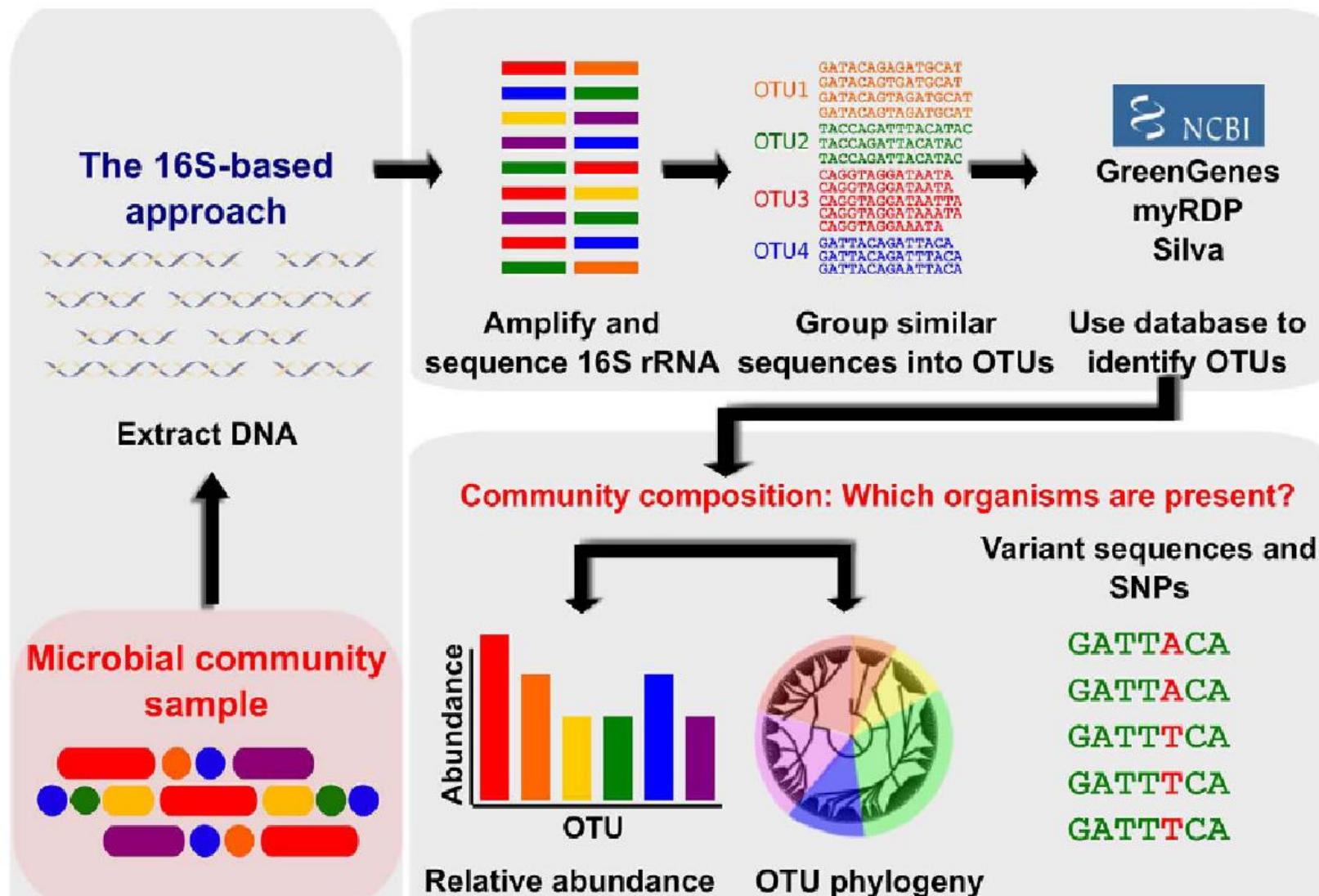
## Workflow example



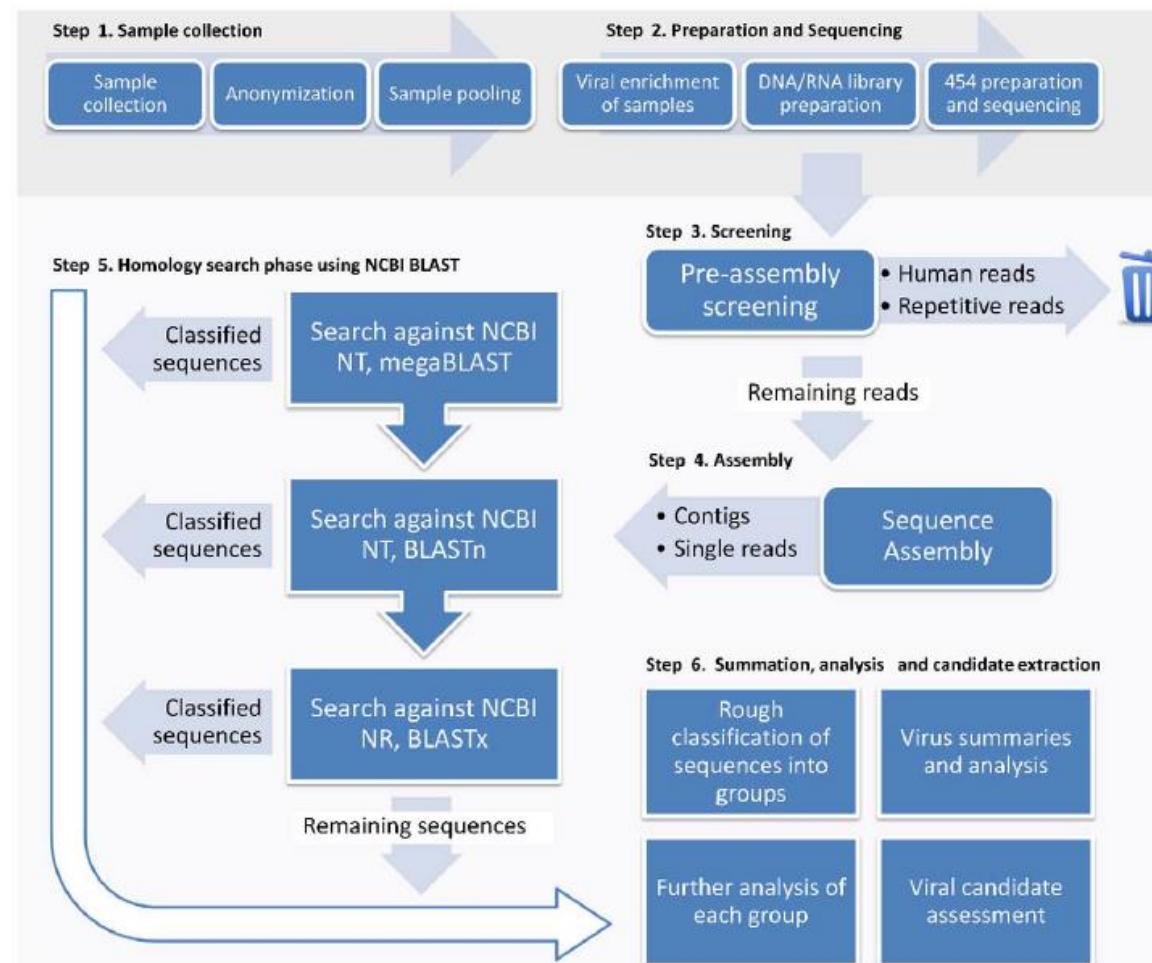
# Metataxonomics vs Metagenomics (16S vs Shotgun)

	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

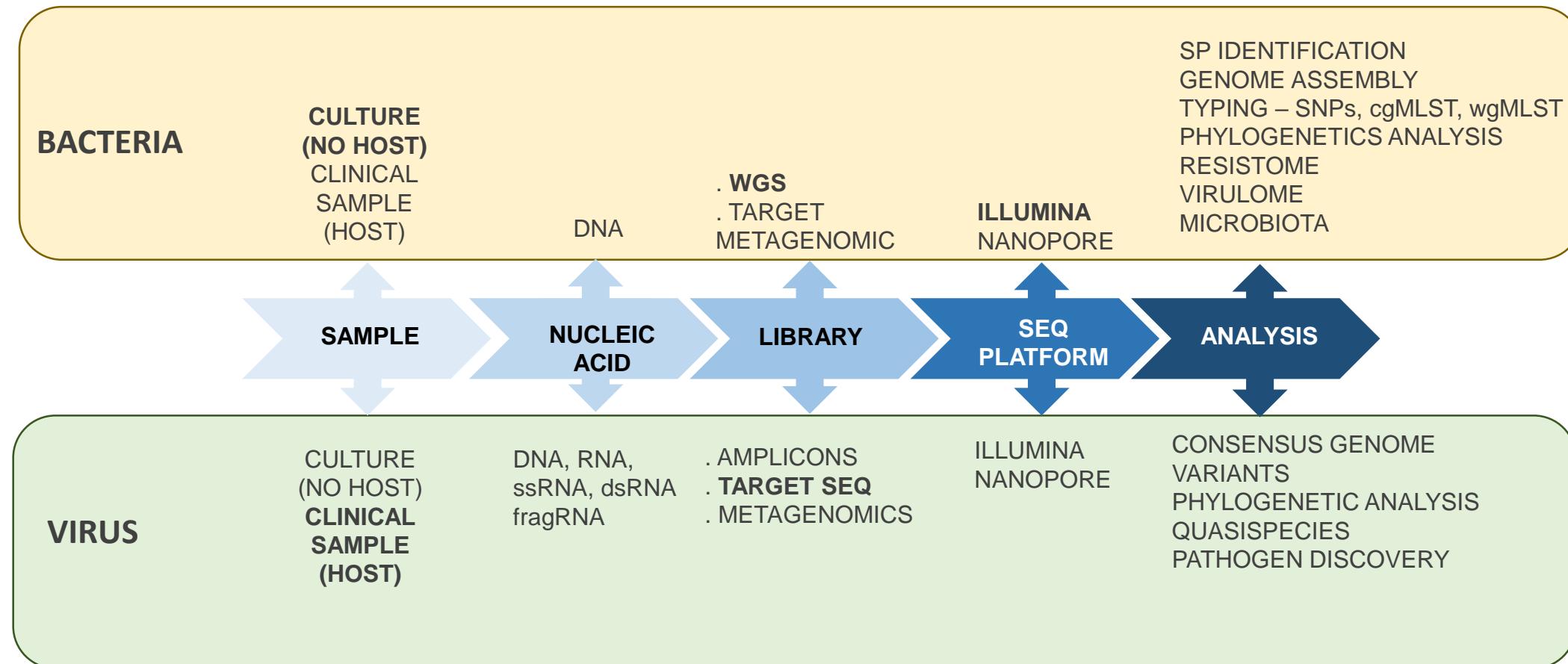


# Metagenómica, pipeline de análisis



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

## Bacterial and viral Genome Sequencing



# Retos de la Bioinformática en High-Throughput Sequencing

- Tecnología que evoluciona muy rápido
  - diferentes plataformas
  - Short vs long reads
  - diferentes errores y outputs
- Coste de la secuenciación disminuye el embudo es el análisis de datos
- Adquisición de secuenciador debe ir ligado a la compra de computo y contratación de bioinformático

# Retos de la Bioinformática en High-Throughput Sequencing

- Necesidades de computo
  - ficheros de gran volumen (10Gb)
  - elevado uso de CPU y/o memoria
  - software no comercial en SO Unix
- Necesidades son dependientes de proyecto
  - No es lo mismo secuenciar un genoma 500Gb que 50 genomas 25Tb
- Si el proyecto es la aplicación en clínica
  - Las necesidades de almacenamiento aumentan por número de pacientes y por tiempo

# Retos de la Bioinformática en High-Throughput Sequencing

- Desarrollo de BD curadas (confianza = reference)
- Algoritmos que resuelvan el problema biológico planteado.
- Necesidades de Bioinformáticos  
Análisis de los datos

# Retos de la Bioinformática en High-Throughput Sequencing

- Tecnología que evoluciona muy rápido
  - nuevos formatos de ficheros
  - nuevas aplicaciones
  - nuevos análisis
  - nuevos algoritmos
- Software en continuo desarrollo (Unix)

Thanks for your attention!

Questions???