

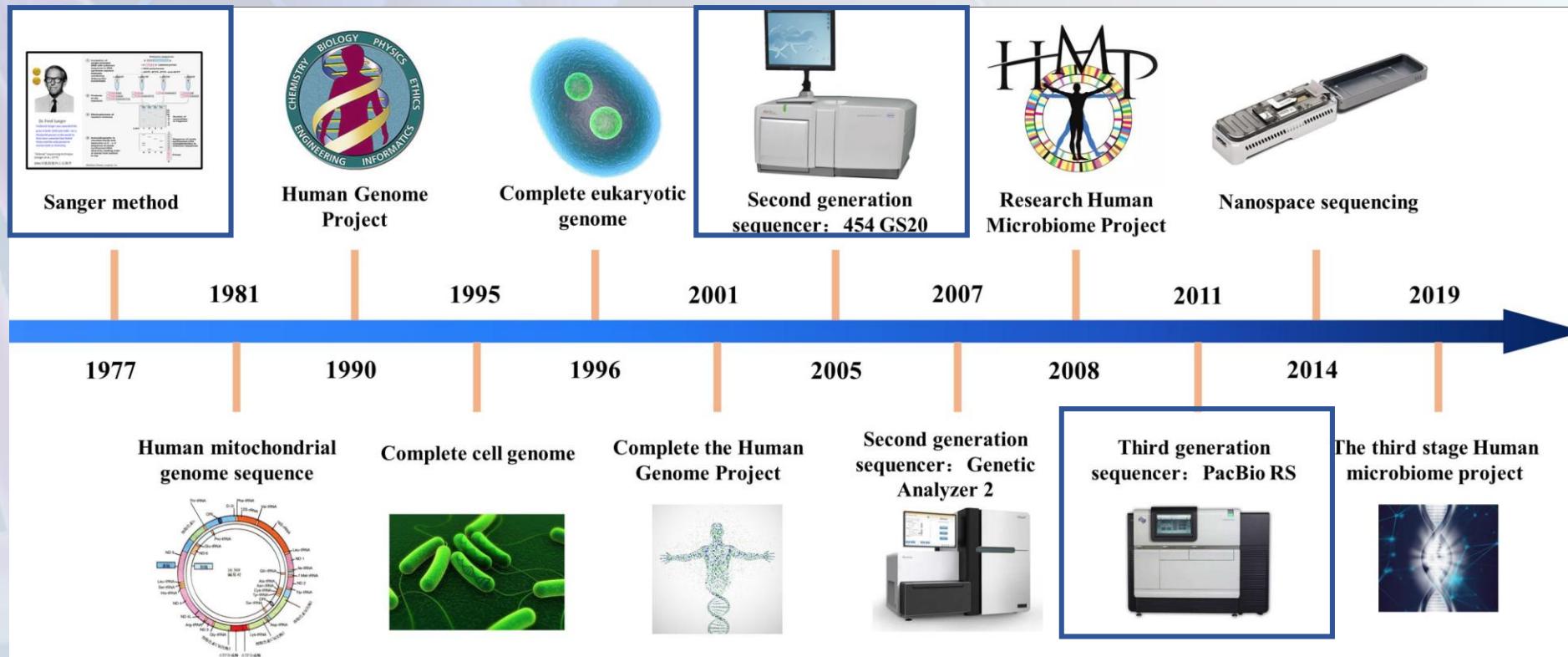
Técnicas de NGS para el estudio de microorganismos patógenos en alimentos y aguas

Dra. María Sol Haim
Unidad Operativa Centro Nacional de Genómica y Bioinformática
ANLIS “Dr. Carlos G. Malbrán”
Buenos Aires, Argentina
mhaim@anlis.gob.ar

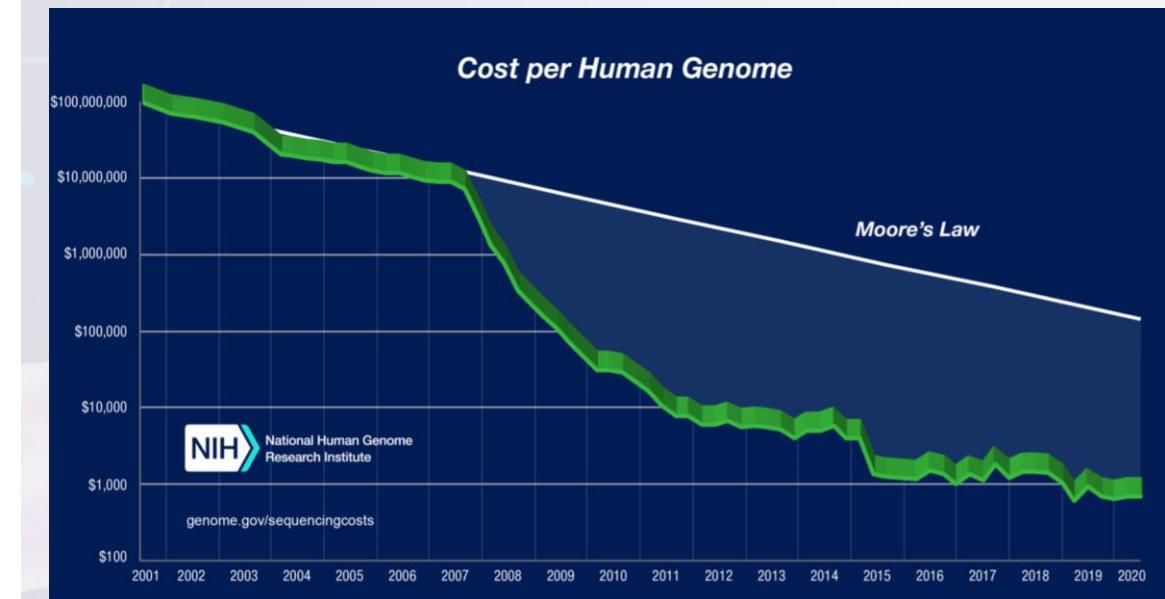
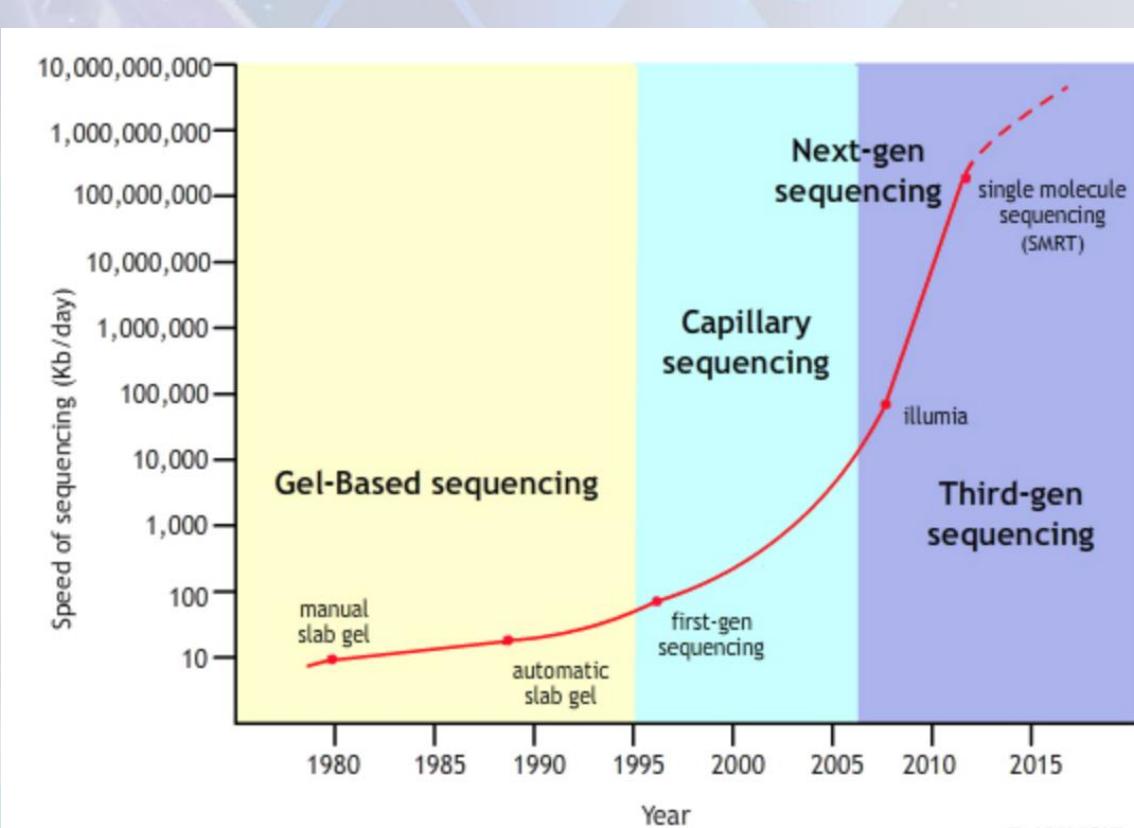
14 de noviembre de 2023



Un breve repaso



Un breve repaso

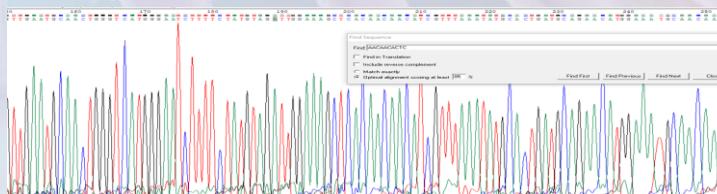


Un breve repaso

SECUENCIACIÓN

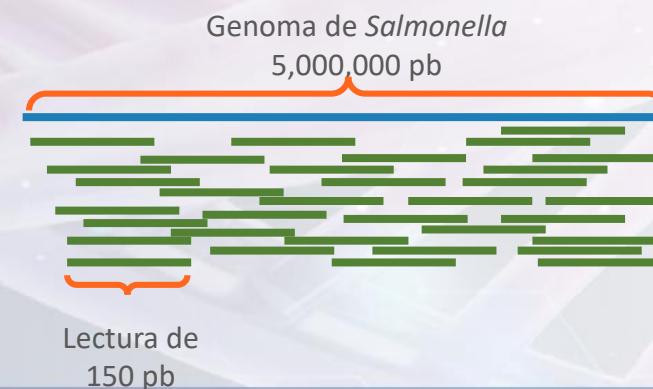
PRIMERA GENERACIÓN

Secuenciación por Sanger
Ej: ABI Applied biosystems



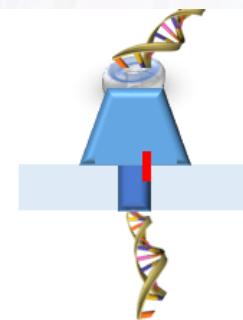
SEGUNDA GENERACIÓN

Secuenciación paralela masiva
Ej: Roche 454, Illumina, Ion Torrent

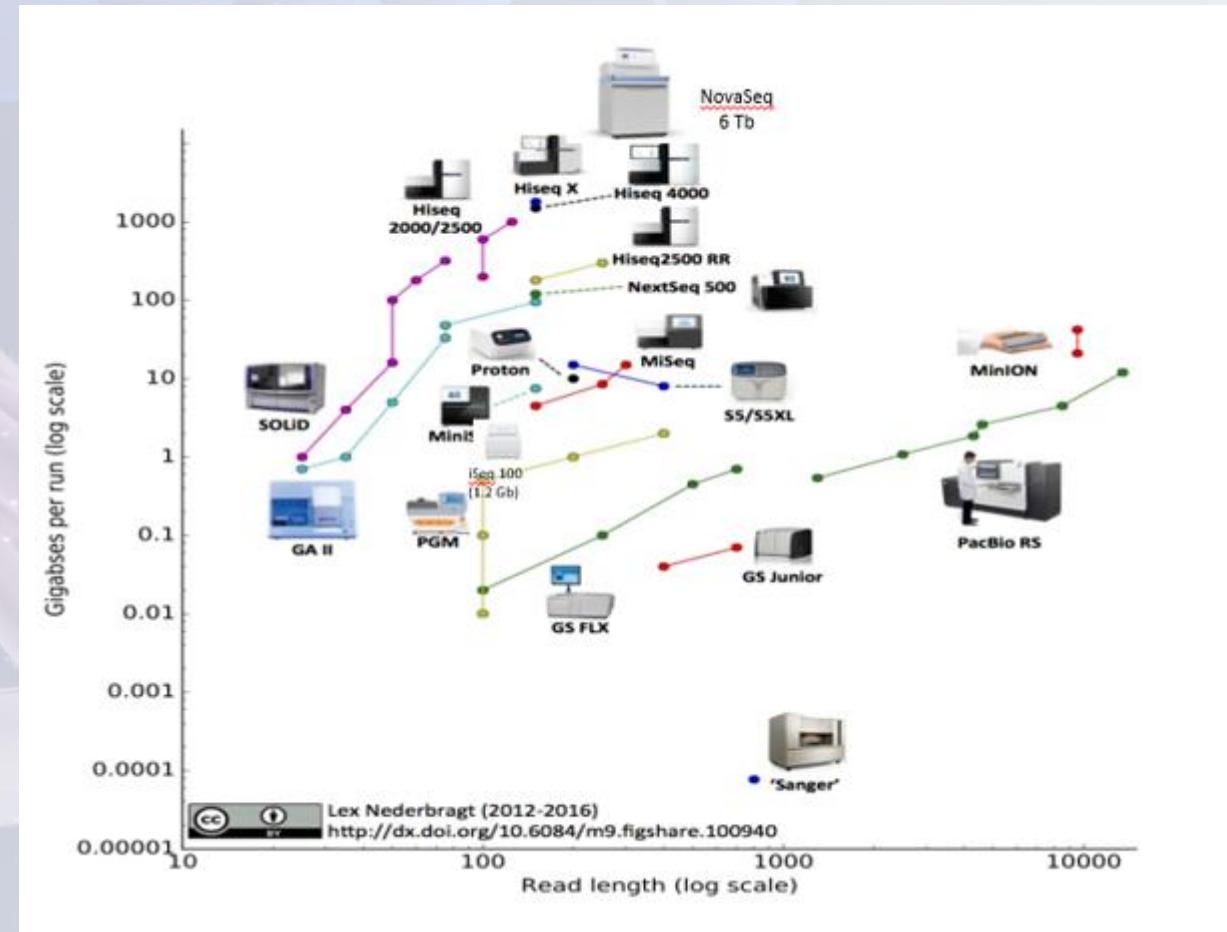


TERCERA GENERACIÓN

Secuenciación de molécula única
Ej: PacBio, Nanopore



Un breve repaso



Cursos Internacional .
Secuenciación y Análisis de Datos Genómicos para la Detección Microbiológica
de Enfermedades Transmitidas por Alimentos y Aguas

Un breve repaso

Short-read Sequencing Platforms and various characteristics.

Company	Illumina							ThermoFisher			
System Platform	iSeq	Miniseq	MiSeq	NextSeq550	NextSeq 1000&2000	NovaSeq 6000	MiSeqDx	NextSeq550 Dx	GeneStudio S5	Genexus	Ion PGM-Dx
Sequencing Principle	Sequence by Synthesis										
Detection	Fluorescent							Ion			
Applications	Small WGS, TS, Small RNA sequencing	Small WGS, TS, ChIP-Seq, Small RNA sequencing	TS, small WGS, exome and transcriptome sequencing	TS, WGS, WES, transcriptome and epigenome sequencing	TS, Small WGS	TS, exome and transcriptome sequencing	TS, epigenetic, exome, and transcriptome sequencing	TS	TS	TS	
Maximum Read length (bases)	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 250 bp	2 × 300 bp	2 × 150 bp	600 bp	400 bp	200 bp	
Flow cells/device	1				2	1					
Output (per flow cell)	1.2 Gb	7.5 Gb	15 Gb	120 Gb	330 Gb	3000 Gb	≥5 Gb	≥90 Gb	15 Gb	24 Gb	1 Gb
Sequencing Run time	9.5-19 hr	5-24 hr	4-56 hr	11-29 hr	11-48 hr	13-44 hr	24 hr	≤35 hr	4.5-21.5 hr	14-31 hr	4.4 hr
Accuracy/Quality	Q30≥ 80% (2 × 150)	Q30≥ 70%	Q30≥ 75% (2 × 150 bp)	Q30≥ 75%	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
Score	bp)		(2 × 300 bp)		(2 × 250 bp)		>99.66%, Q30> 80%	≥99.98%, Q30≥75%	≥99%	≥99%	≥99%
Equipment Cost (USD)	\$19,900	\$49,500	\$99,000	\$275,000	\$335,000	on request					

Hu, 2021

Un breve repaso

System Platform	Sequel	Sequel II	Sequel IIe	Flongle	MinION	GridION	PromethION			
Sequencing Principle	PacBio Single Molecule Sequencing				Nanopore Single molecule Sequencing					
Detection	Fluorescent				Electrical Conductivity					
Applications	Whole genome <i>de novo</i> assembly, variant detection, structural variation detection, full length transcript sequencing, targeted/amplicon sequencing, metagenomics sequencing				DNA, amplicons, cDNA, Direct RNA sequencing					
Maximum Read length (bases)	300 kb				Longest read so far: > 4 Mb					
Flow cells/device	12 SMRT Cells 1M can be used at a time, and 8 SMRT Cell 8M can be used serially				1 (126 channels per flow cell)	1 (512 channels per flow cell)	5 (512 channels per flow cell)	24 or 48 (3000 channels per flow cell)		
Output (per flow cell)	75 Gb	600 Gb	1 - 2 Gb ^a	10 - 30 - 50 Gb ^a		100 - 200 - 300 Gb ^a				
Sequencing Run time	Up to 20 hr	Up to 30 hr	1 min - 16 hr	1 min - 72 hr						
Accuracy/Quality Score	Number of HiFi Reads >99% Accuracy: Up to 5,000,000 reads	Number of HiFi Reads >99% Accuracy: Up to 4,000,000 reads	Single Molecule: R9 modal Accuracy >98.3%, R10 modal Accuracy >97.5%. New chemistry Accuracy >99% (coming soon) Consensus: R9.4.1: Current best Q45 (>99.99%) R10: Current Best Q50 (99.999%)							
Equipment Cost (USD)	approximately \$525,000				\$1,460 (12 flow cells included)	\$9,300	\$69,955	24 flow cells: \$335,455 48 flow cells: \$530,000		

Hu, 2021

¿Qué pregunta queremos responder?
¿Cuál es el objetivo de la secuenciación del genoma completo?

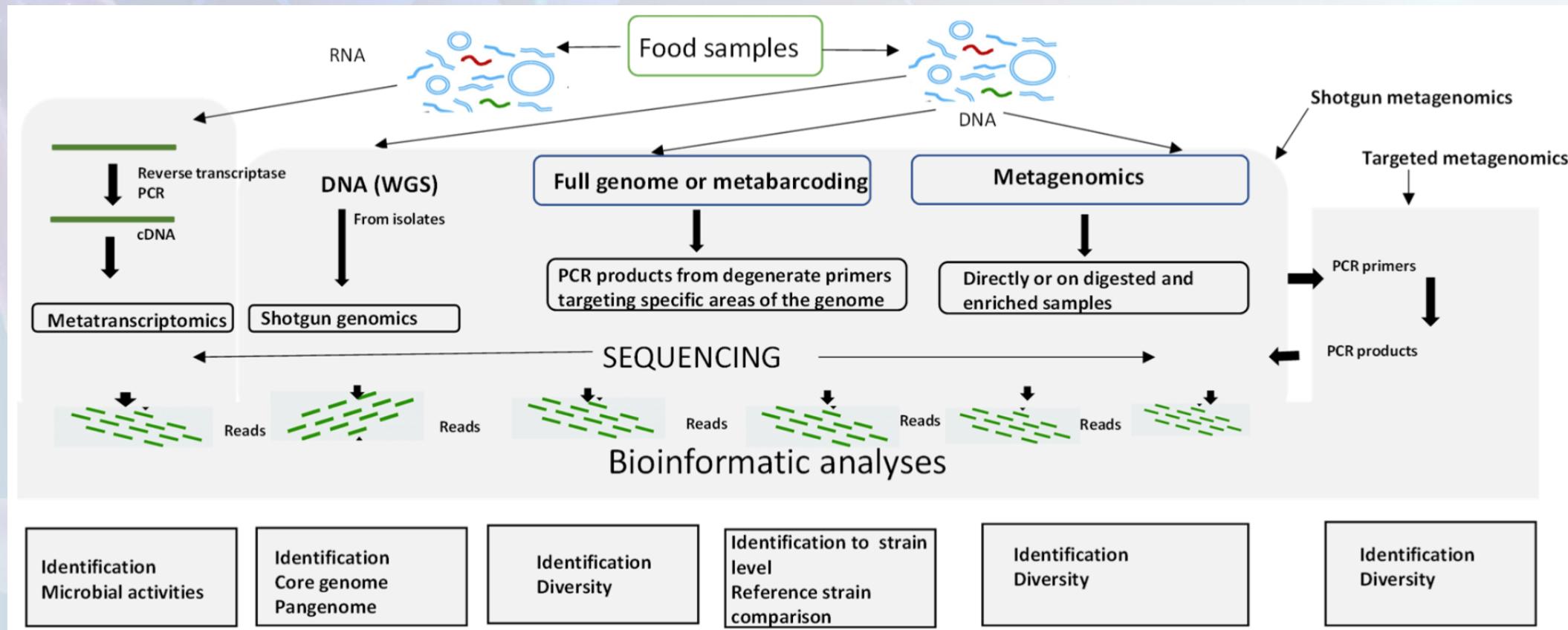


Elección de estrategia de secuenciación
Elección de plataforma de secuenciación
Elección del flujograma de análisis

Algunas preguntas que podemos responder con NGS

- ¿Cuál es la fuente que dio origen a un brote de ETA?
- ¿Cuál es la relación filogenética entre estos aislamientos?
- ¿Cuál es la vía de transmisión de un determinado patógeno?
- ¿Qué especies bacterianas están presentes en un alimento?
- ¿Qué clones/linajes se encuentran presentes en determinados aislamientos/ambientes? ¿Qué diversidad presentan? ¿Presentan características particulares? (RAM, factores de virulencia)

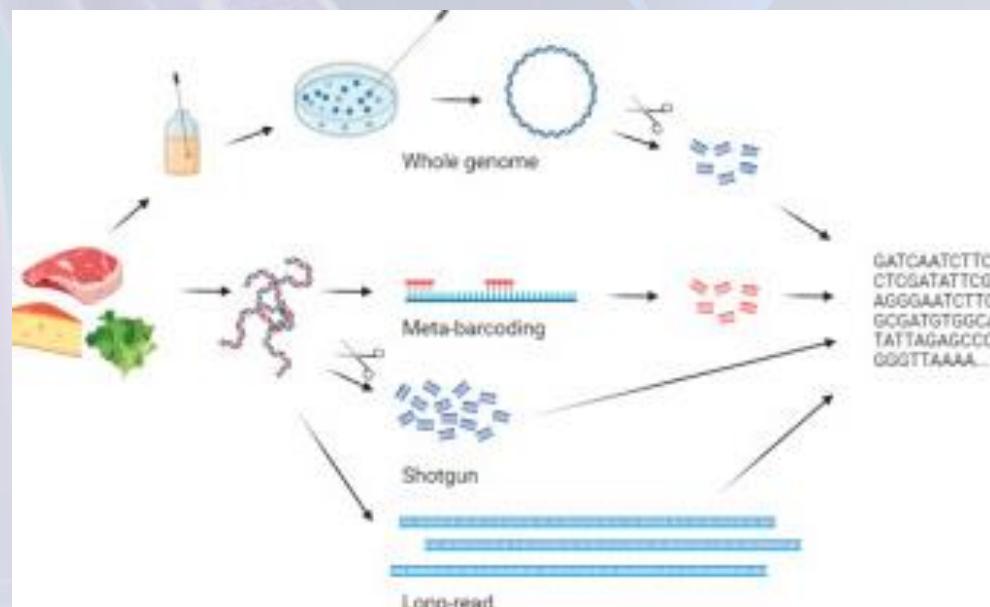
Estrategias de secuenciación



<https://doi.org/10.3390/microorganisms11051111>

Estrategias de secuenciación

Metagenómica



<https://doi.org/10.4315/JFP-21-301>

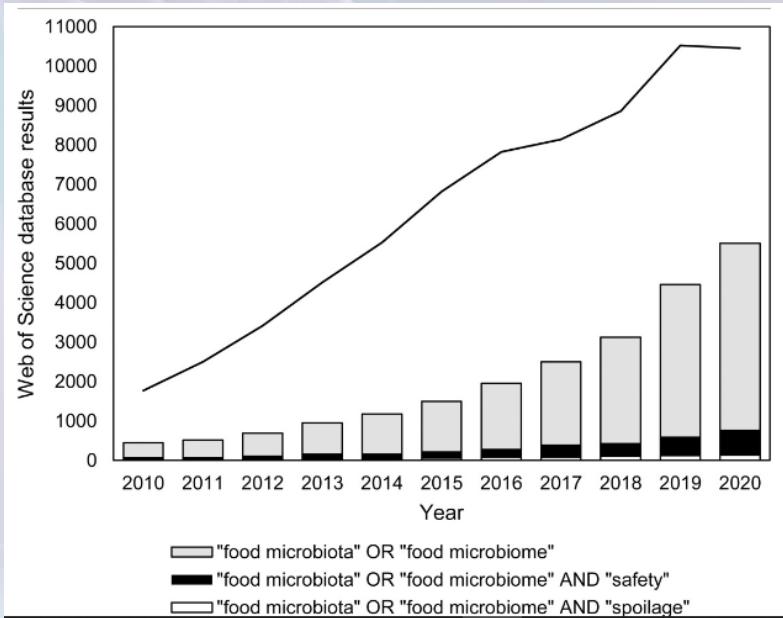
Potenciales aplicaciones para mejorar la seguridad alimentaria:

- Medir los cambios en la estructura de la población microbiana a lo largo del tiempo y en respuesta a estímulos.
- Independiente del cultivo de microorganismos: permitiendo la detección de microorganismos no cultivables, reduciendo sesgos y ahorrando tiempo.
- Identificar riesgos emergentes o previamente desconocidos.

El monitoreo rutinario del microbioma de los ingredientes crudos puede revelar cambios en la cantidad y los tipos de microorganismos presentes que alertan al procesador sobre posibles cambios en la seguridad del producto o en los perfiles de riesgo de calidad, lo que desencadena investigaciones y acciones correctivas.

Estrategias de secuenciación

Metagenómica



Alimento	Aplicación	Target	Estrategia NGS
Lácteos	Control de calidad, búsqueda de genes de RAM, detección de patógenos, rastreo de contaminaciones	Bacterias, hongos, genes de RAM y virulencia, profagos	Metabarcoding (ARNr 16S o 26S, ITS), shotgun y target metagenomics
Mariscos	Detección de patógenos, rastreo de contaminaciones	Bacterias, virus	Metabarcoding (ARNr 16S, VP1, RdRP)
Carnes	Detección de patógenos, rastreo de contaminaciones	Bacterias, virus	Metabarcoding (ARNr 16S), shotgun

<https://doi.org/10.4315/JFP-21-301>

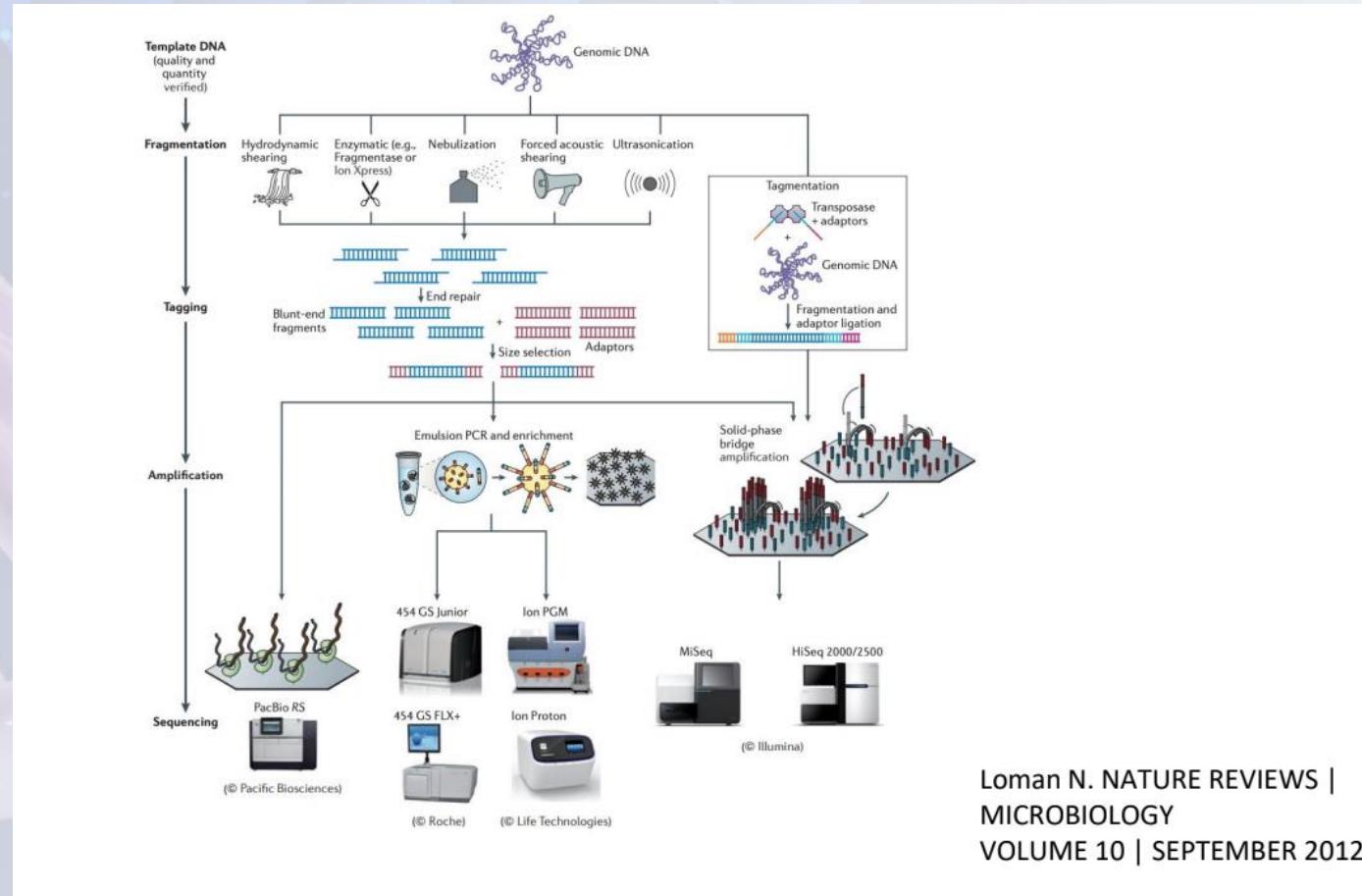
Estrategias de secuenciación

Metagenómica

Estrategia	Ventajas	Desventajas
Metabarcoding	<ul style="list-style-type: none"> • Bajo costo • Se puede realizar en la mayoría de los secuenciadores de escritorio • Apuntar a regiones conocidas independientemente de la prevalencia en una comunidad • Se pueden identificar taxones con baja abundancia (profundidad!) • No es necesaria la reconstrucción del genoma completo 	<ul style="list-style-type: none"> • Conocer previamente del grupo microbiano target • Nos focalizamos exclusivamente en un grupo de microorganismos y no en toda la comunidad microbiiana presente • Posibilidad de errores y sesgos en la amplificación
Shotgun metagenomics	<ul style="list-style-type: none"> • Todo el ADN es secuenciado • No hay sesgos por amplificación • Se puede utilizar para generar targets o marcadores genéticos • Se podría ensamblar genomas de los taxones abundantes • Identificar nuevos microorganismos 	<ul style="list-style-type: none"> • Menos sensible para la detección de patógenos • Costoso • Secuenciadores de mayor rendimiento • Mucha información producida (servidores y recurso humano capacitado)

<https://doi.org/10.4315/JFP-21-301>

Esquema general de secuenciación



Antes de comenzar con la secuenciación

¿A qué nos referimos con “profundidad”?

Profundidad = el número de veces que se ha secuenciado un genoma



La profundidad promedio divide el número total de bases secuenciadas por el tamaño del genoma

Antes de comenzar con la secuenciación

¿A qué nos referimos con “profundidad”?

Profundidad = el número de veces que se ha secuenciado un genoma

La profundidad promedio divide el número total de bases secuenciadas por el tamaño del genoma

⇒ Cuál sería la profundidad promedio de una muestra de *Escherichia coli* para la cual se obtuvieron 2000000 lecturas totales de 150 pb?

Profundidad = cantidad de lecturas x tamaño de lecturas (pb)

tamaño del genoma (pb)

$$\text{Profundidad} = \frac{2000000 \times 150 \text{ pb}}{5000000 \text{ pb}} = 60X$$

Antes de comenzar con la secuenciación

1. Definir organismos a secuenciar y conocer el tamaño del genoma de los mismos
2. Definir la profundidad deseada para cada microorganismo o aquella requerida por los organismos oficiales para la vigilancia genómica

Minimum Depth	<i>Escherichia/Shigella</i>	<i>Salmonella</i>	<i>Listeria</i>	<i>Campylobacter</i>	<i>Vibrio</i>
FDA	20X	20X	20X	20X	20X
PulseNet	40X	30X	20X	20X	40X

1. Definir el kit de reactivos MiSeq y conocer el número de ciclos y su output:
2. Calcular el número de muestras que podrían ser secuenciadas

Antes de comenzar con la secuenciación

⇒ Cuántas muestras de *Escherichia coli* se podrán incluir en una corrida de secuenciación utilizando un MiSeq Reagent kit v2 (2x150)?

1. *Escherichia coli* - 5000000 pb
2. Profundidad deseada= 100X
3. MiSeq Reagent kit v2 (2x150)

	MiSeq Reagent Kit v2				MiSeq Reagent Kit v3	
Read Length	1 × 36 bp	2 × 25 bp	2 × 150 bp	2 × 250 bp	2 × 75 bp	2 × 300 bp
Total Time*	~4 hrs	~5.5 hrs	~24 hrs	~39 hrs	~21 hrs	~56 hrs
Output	540–610 Mb	750–850 Mb	4.5–5.1 Gb	7.5–8.5 Gb	3.3–3.8 Gb	13.2–15 Gb

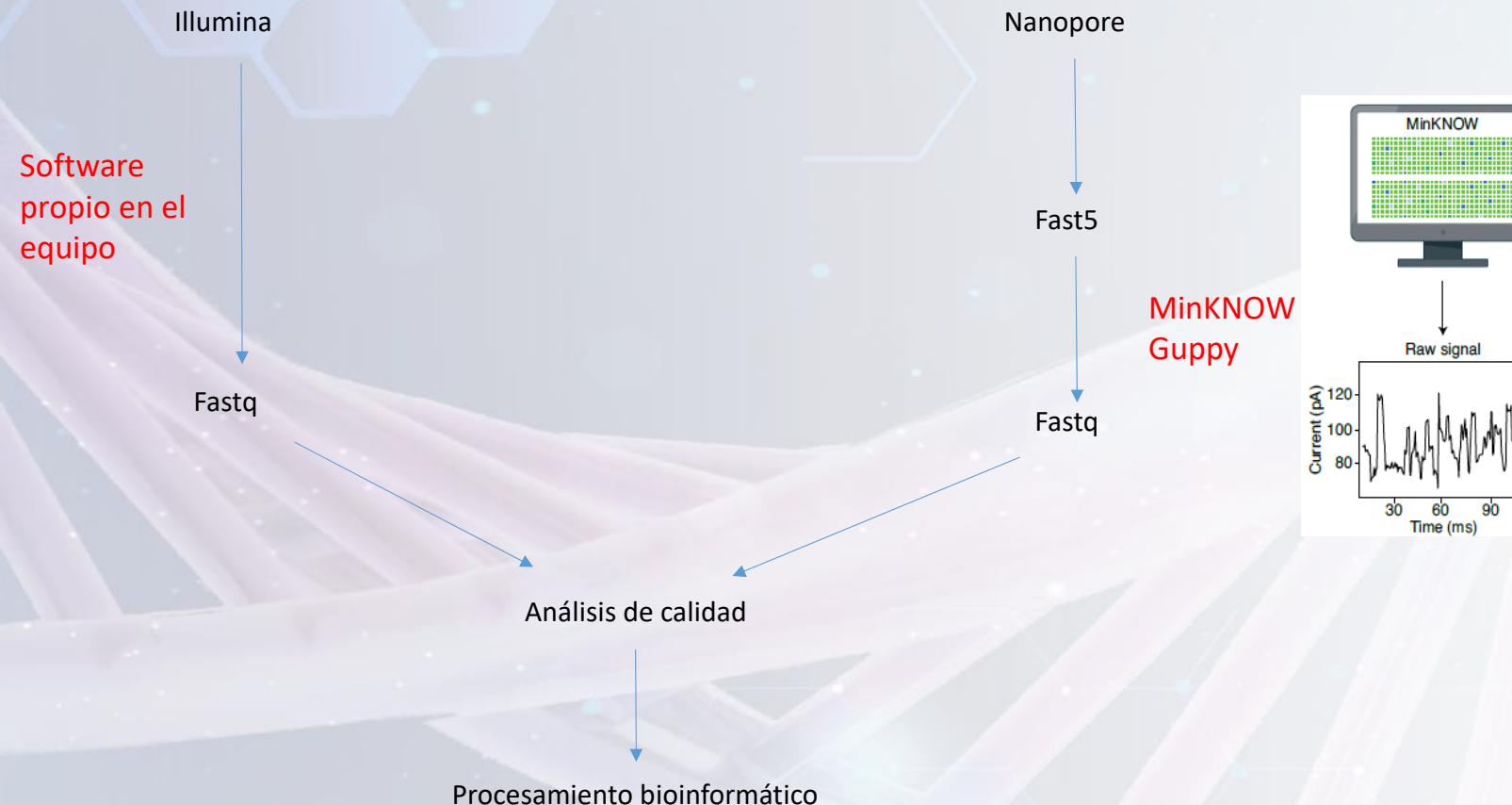
	MiSeq Reagent Kit v2 Micro		MiSeq Reagent Kit v2 Nano	
Read Length	2 × 150 bp	2 × 250 bp	2 × 150 bp	2 × 150 bp
Total Time*	~19 hrs	~28 hrs	~17 hrs	~17 hrs
Output	1.2 Gb	500 Mb	300 Mb	300 Mb

* Total time includes cluster generation, sequencing, and base calling on a MiSeq System enabled with dual-surface scanning.

4. Calcular el número de muestras que podrían ser secuenciadas

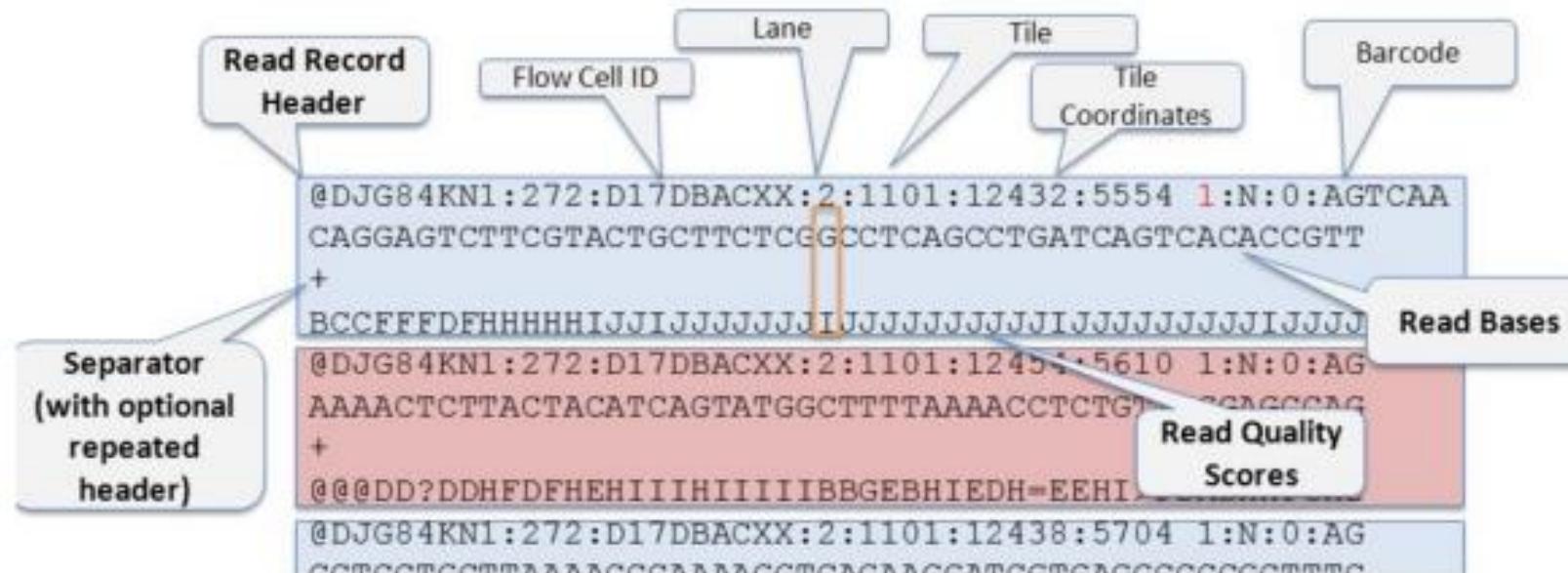
$$5000000\text{pb} * 100 = 500000000 \text{ pb} = 0,5 \text{ Gb} \Rightarrow 5,1 \text{ Gb}/0,5 \text{ Gb} = 10 \text{ muestras}$$

Qué obtenemos luego de la corrida?

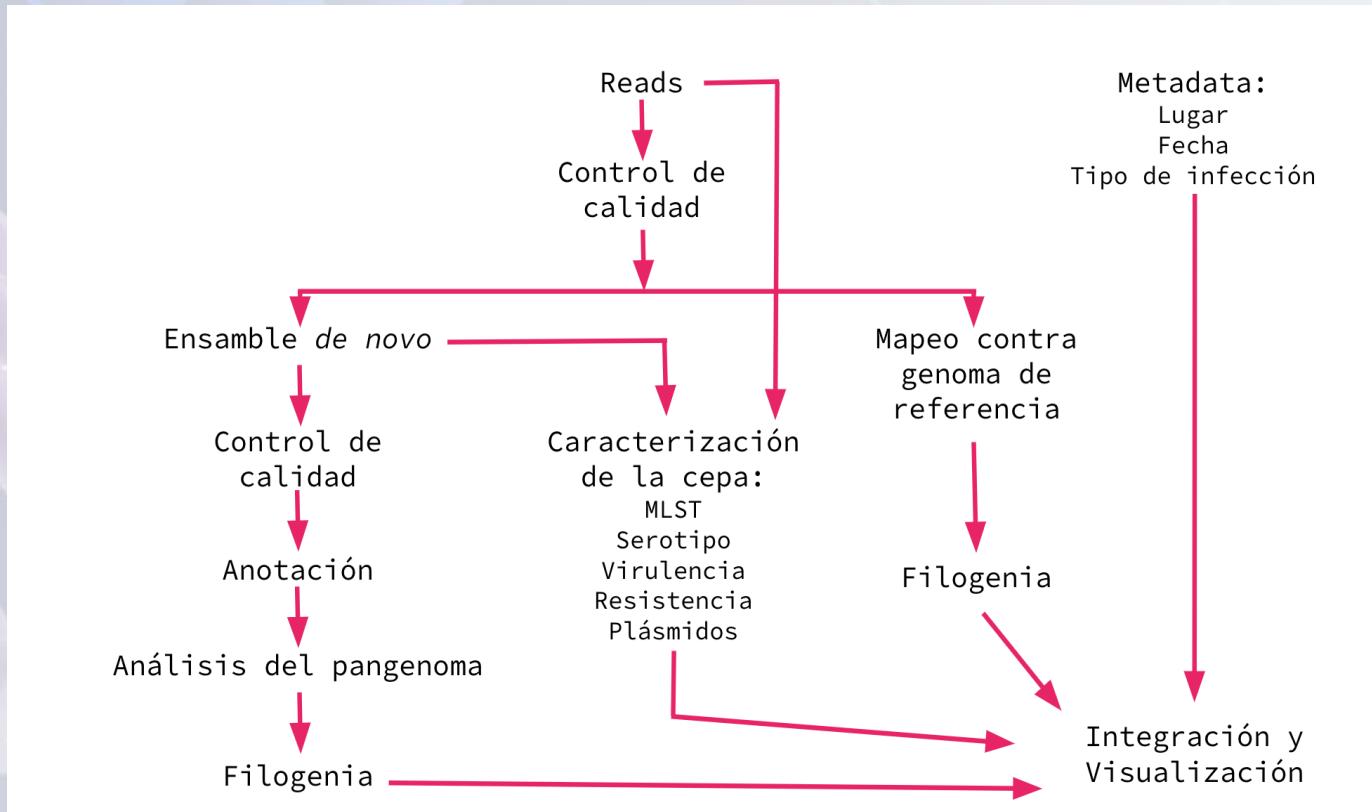


Formato FASTQ

FASTQ Format (Illumina Example)



Flujo general de análisis bioinformático



Control de calidad de las lecturas: FastQC

 **FastQC Report**

Wed 25 Mar 2015
good_sequence_short.txt

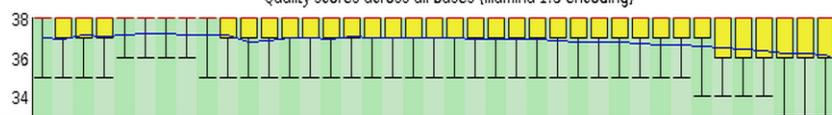
Summary	
 Basic Statistics	
 Per base sequence quality	
 Per tile sequence quality	
 Per sequence quality scores	
 Per base sequence content	
 Per sequence GC content	
 Per base N content	
 Sequence Length Distribution	
 Sequence Duplication Levels	
 Overrepresented sequences	
 Adapter Content	
 Kmer Content	

 **Basic Statistics**

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

 **Per base sequence quality**

Quality scores across all bases (illumina 1.5 encoding)



Control de calidad de las lecturas: FastQC



Control de calidad de las lecturas: Kraken

Kraken

Taxonomic Sequence Classification System

1= % de reads cubiertas por el clado presente en este taxón

2= Número de reads cubiertas por el clado presente en este taxón

3= Número de reads asignadas directamente a este taxón

4= Rango taxonómico: (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, (S)pecies, Otros (-), (U)nclassified,

5= NCBI taxonomy ID

6= Nombre científico

- Permite la clasificación de **reads** de manera rápida a través de un algoritmo que los asocia a los taxones más antiguos (LCA) usando un alineamiento exacto de **kmers**.
- Útil para identificar distintas especies de reads obtenidas por metagenómica
- Kraken2
- Línea de comando/Online (Galaxy)

1	2	3	4	5	6
2.73	4508	4508	U	0	unclassified
97.27	160631	4999	-	1	root
94.24	155631	0	-	131567	cellular organisms
94.24	155630	435	D	2	Bacteria
93.91	155085	20	-	1783272	Terrabacteria group
93.90	155065	70	P	1239	Firmicutes
93.86	154993	95	C	91061	Bacilli
93.79	154889	319	O	1385	Bacillales
93.60	154566	328	F	90964	Staphylococcaceae
93.40	154237	8437	G	1279	Staphylococcus
88.28	145787	106808	S	1280	Staphylococcus aureus
23.60	38971	38723	-	46170	Staphylococcus aureus subsp. aureus
0.13	215	215	-	1074252	Staphylococcus aureus subsp. aureus HO 5096 0412
0.01	10	10	-	1381115	Staphylococcus aureus subsp. aureus Tager 104
0.00	6	6	-	985006	Staphylococcus aureus subsp. aureus LGA251
0.00	4	4	-	685039	Staphylococcus aureus subsp. aureus ED133
0.00	4	4	-	869816	Staphylococcus aureus subsp. aureus JKD6159
0.00	3	3	-	1392476	Staphylococcus aureus subsp. aureus 6850
0.00	2	2	-	1242061	Staphylococcus aureus subsp. aureus ST772 MDCA V

Reporte de Kraken

Wood DE et al.2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. doi: 10.1186/gb-2014-15-3-r46.

Wood DE et al. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. doi: 10.1186/s13059-019-1891-0

Principales estrategias para determinar la relación entre dos cepas

Mapeo contra genoma de referencia



Selección del genoma de referencia: idealmente genoma cerrado y estrechamente relacionado genéticamente con los genomas que se analizan (por ejemplo, el mismo serotipo, ST).

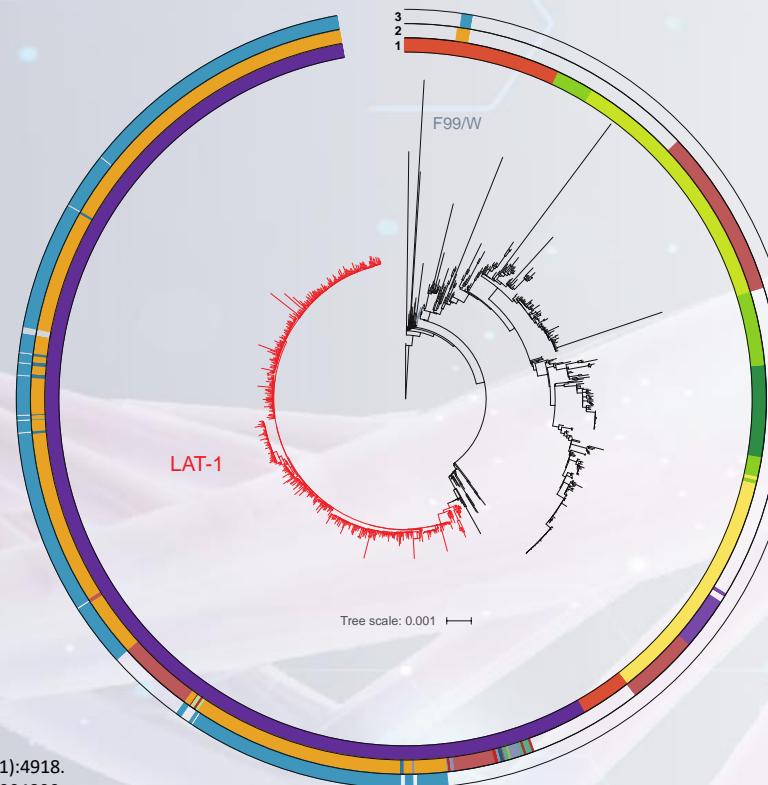
Subestimación de la relación genética, ya que las regiones que se analizan son menos.

El # de SNPs detectados puede variar dependiendo de la referencia utilizada y el método de llamado de variantes

Enmascarar zonas de recombinación/EGM

Principales estrategias para determinar la relación entre dos cepas

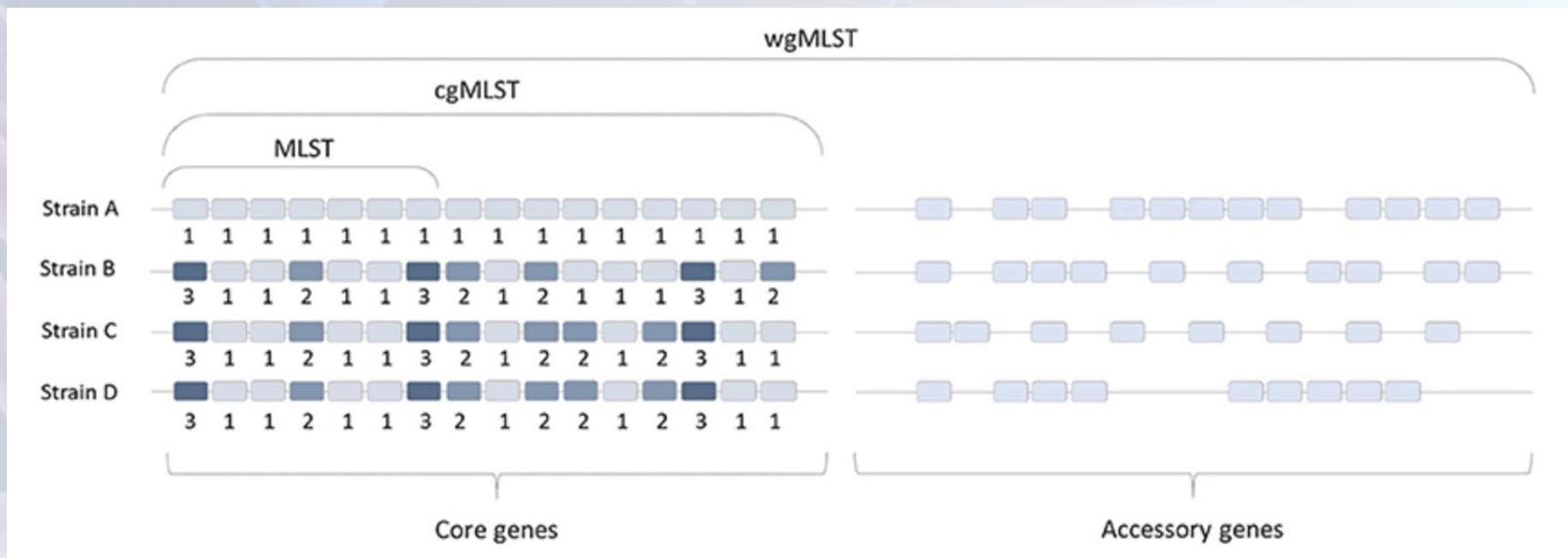
Mapeo contra genoma de referencia



Dorman et al. Nat Commun. 2020 Oct 1;11(1):4918.
doi: 10.1038/s41467-020-18647-7. PMID: 33004800

Principales estrategias para determinar la relación entre dos cepas

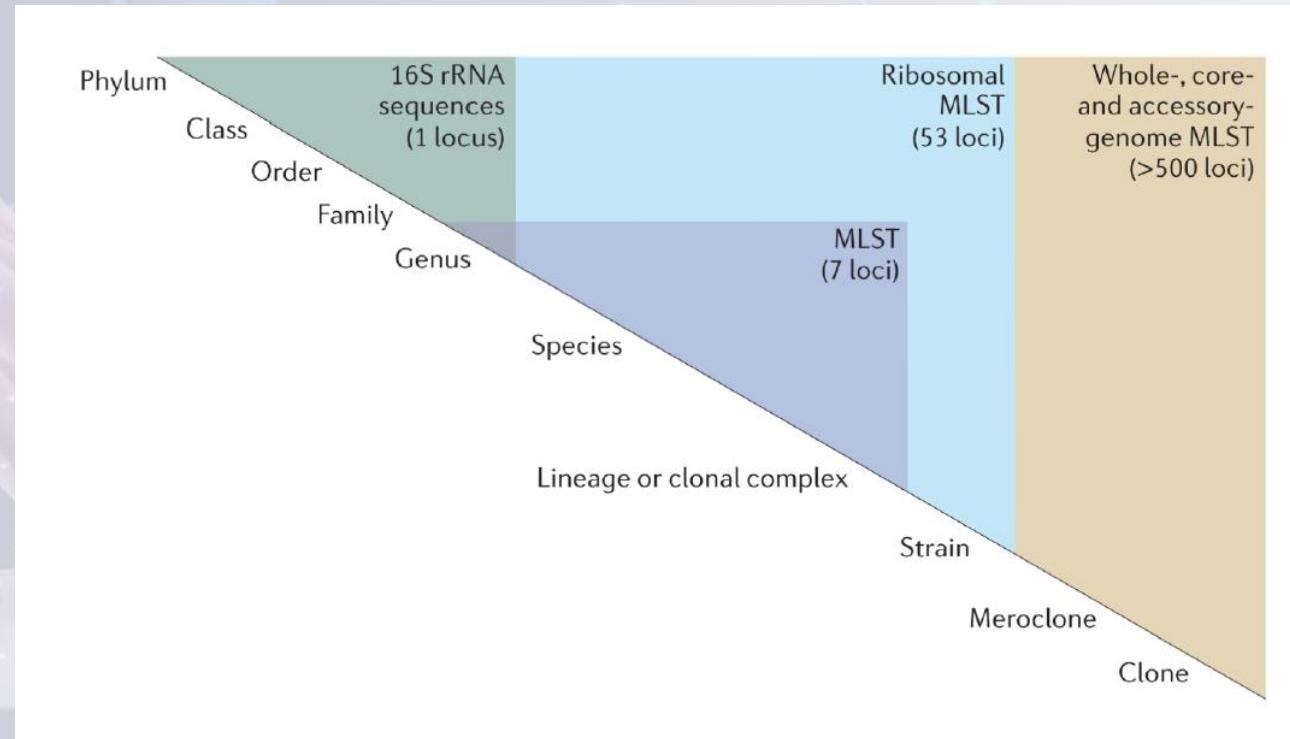
Análisis de genes (MLST, cgMLST y wgMLST)



Janezic S, Rupnik M. 2019. Development and Implementation of Whole Genome Sequencing-Based Typing Schemes for *Clostridioides difficile*. Front Public Health. doi: 10.3389/fpubh.2019.00309.

Principales estrategias para determinar la relación entre dos cepas

Análisis de genes (MLST, cgMLST y wgMLST)



Maiden MC et al. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol.* doi: 10.1038/nrmicro3093.

Principales estrategias para determinar la relación entre dos cepas

Análisis de genes (MLST,
cgMLST y wgMLST)

PubMLST

PubMLST

- <https://pubmlst.org/>
- Secuencias + procedencia + fenotipo para más de 100 especies y géneros microbianos diferentes

Enterobase

Enterobase

- <https://enterobase.warwick.ac.uk/>
- *Salmonella, Escherichia/Shigella, Clostridioides, Vibrio, Yersinia, Helicobacter, Moraxella*

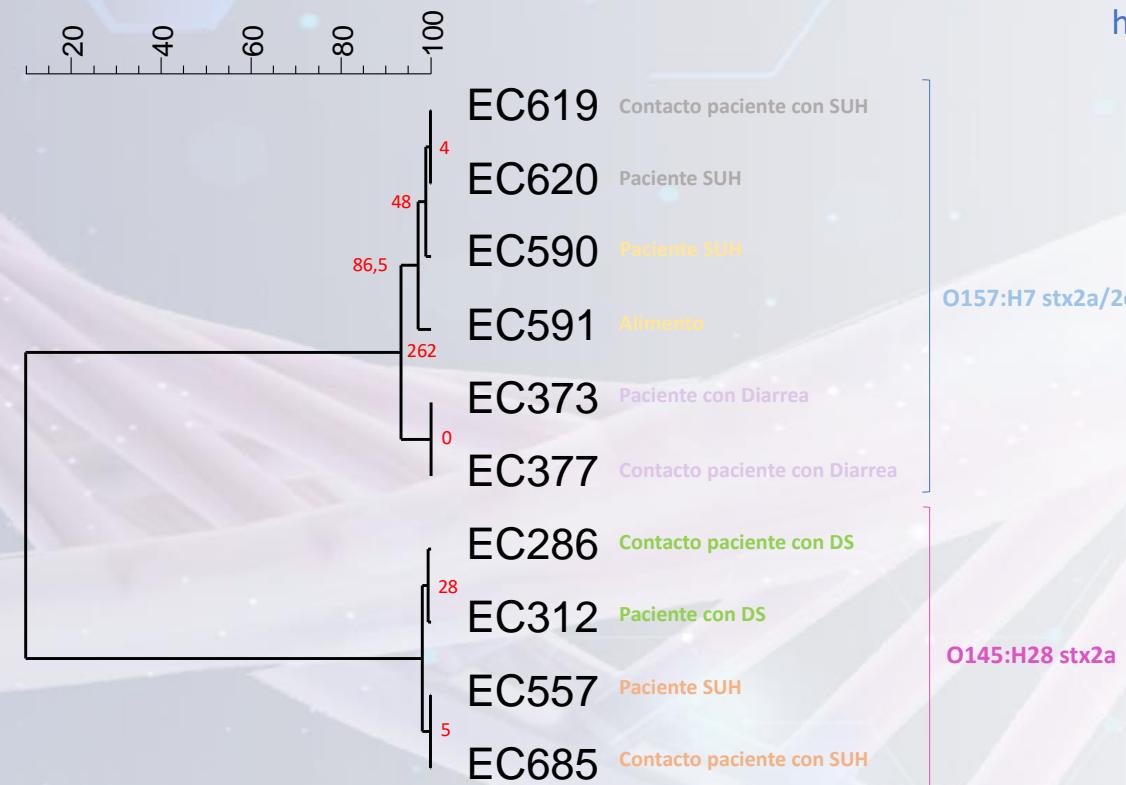
Institut Pasteur MLST



- <https://bigsdb.pasteur.fr/>
- MLST y wgMLST 19 especies/géneros

Principales estrategias para determinar la relación entre dos cepas

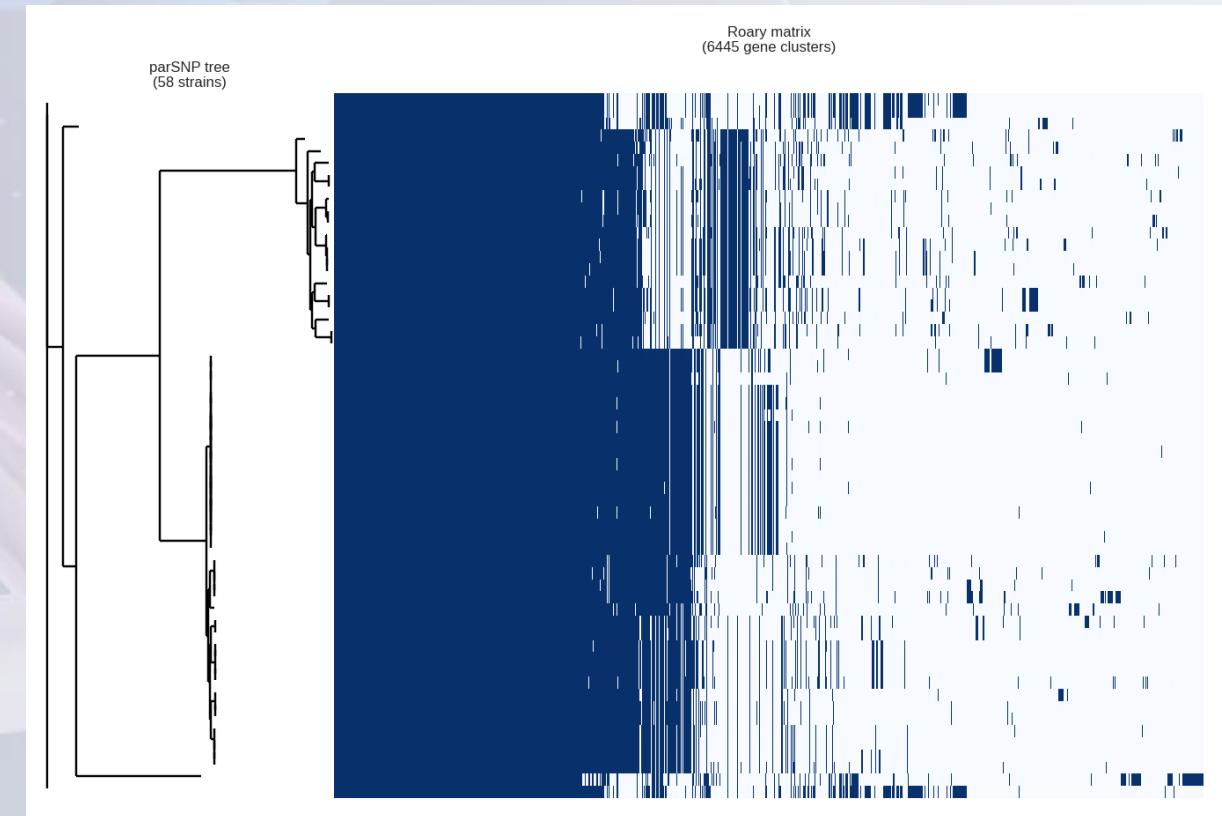
Análisis de genes (MLST, cgMLST y wgMLST)



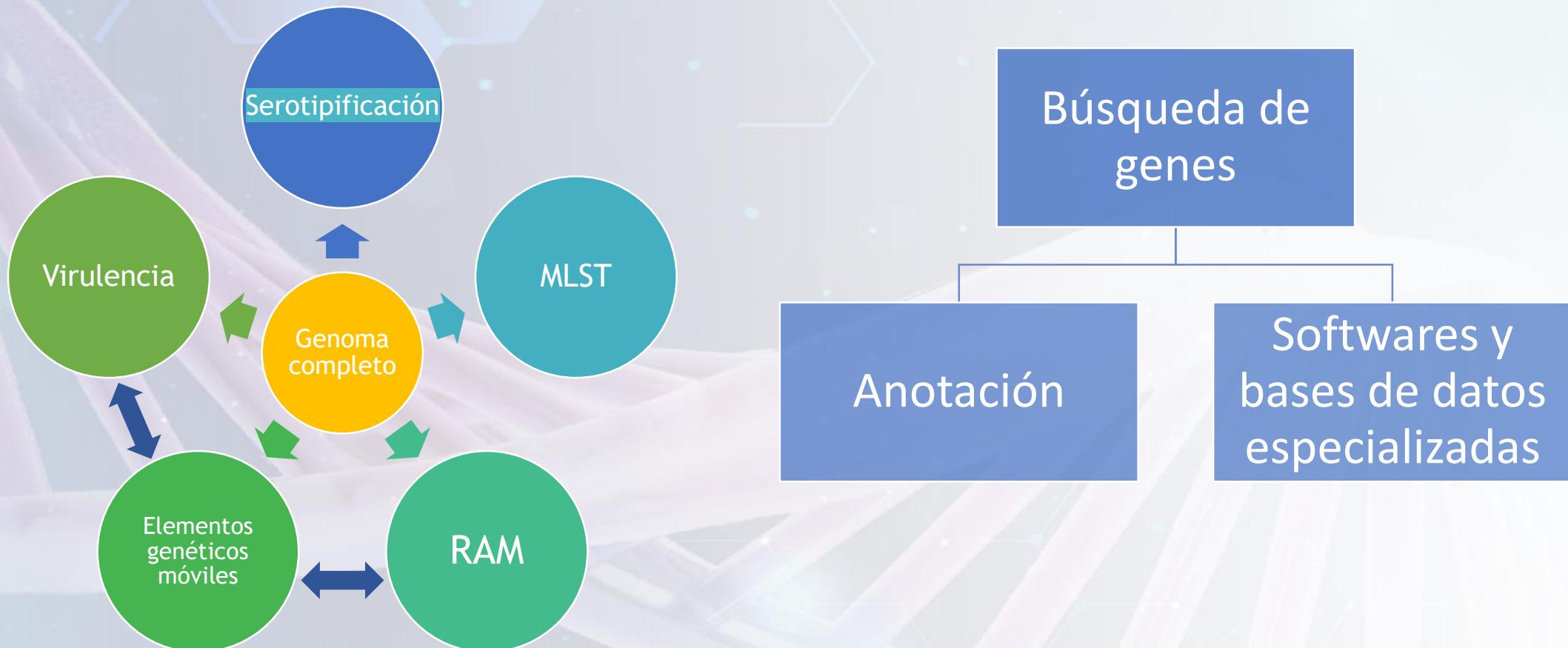
Análisis de brotes de cepas STEC
 humanas y alimentos
 -Argentina 2022-
 Bionumerics 7.6

Principales estrategias para determinar la relación entre dos cepas

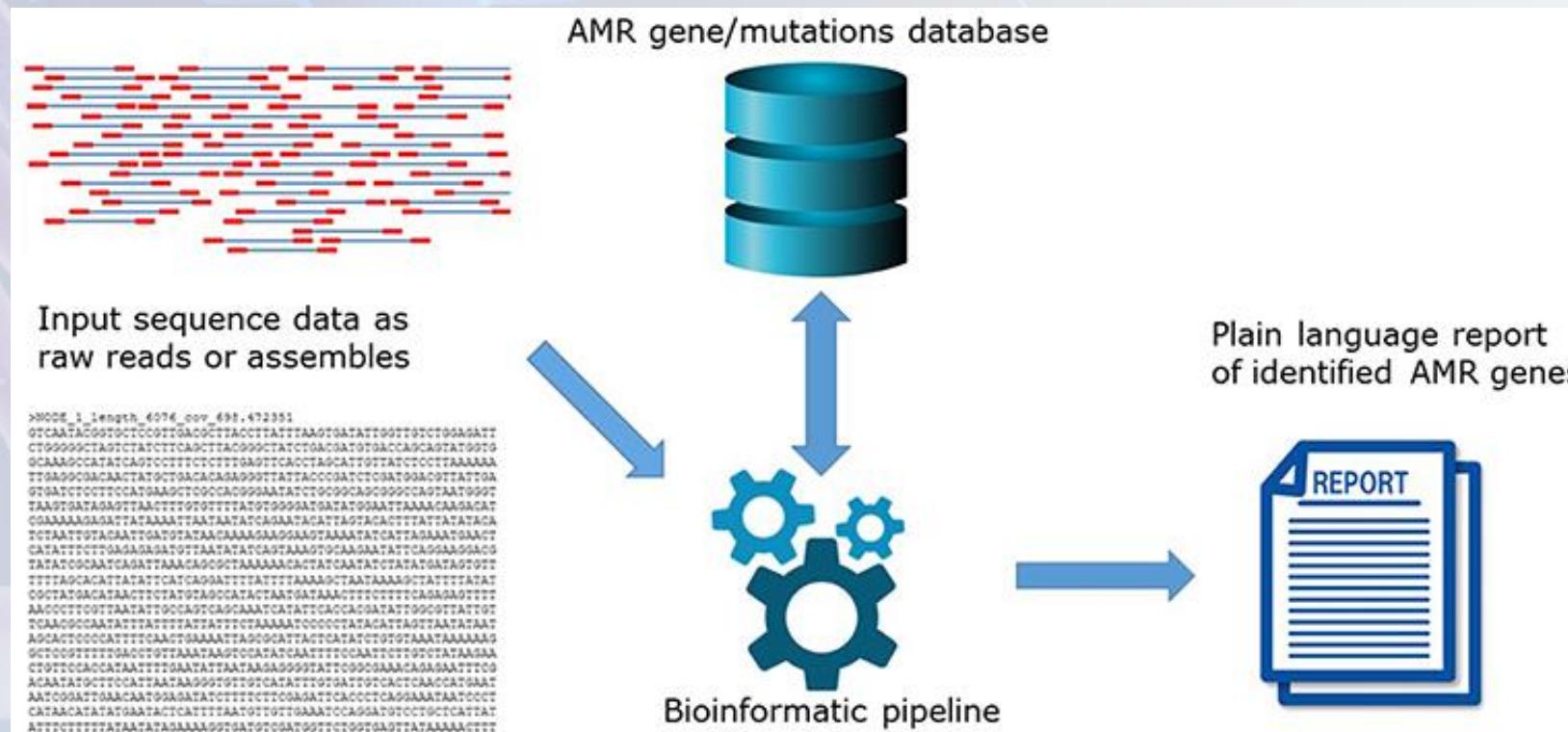
Análisis del pangenoma



Estrategias para caracterizar los genomas

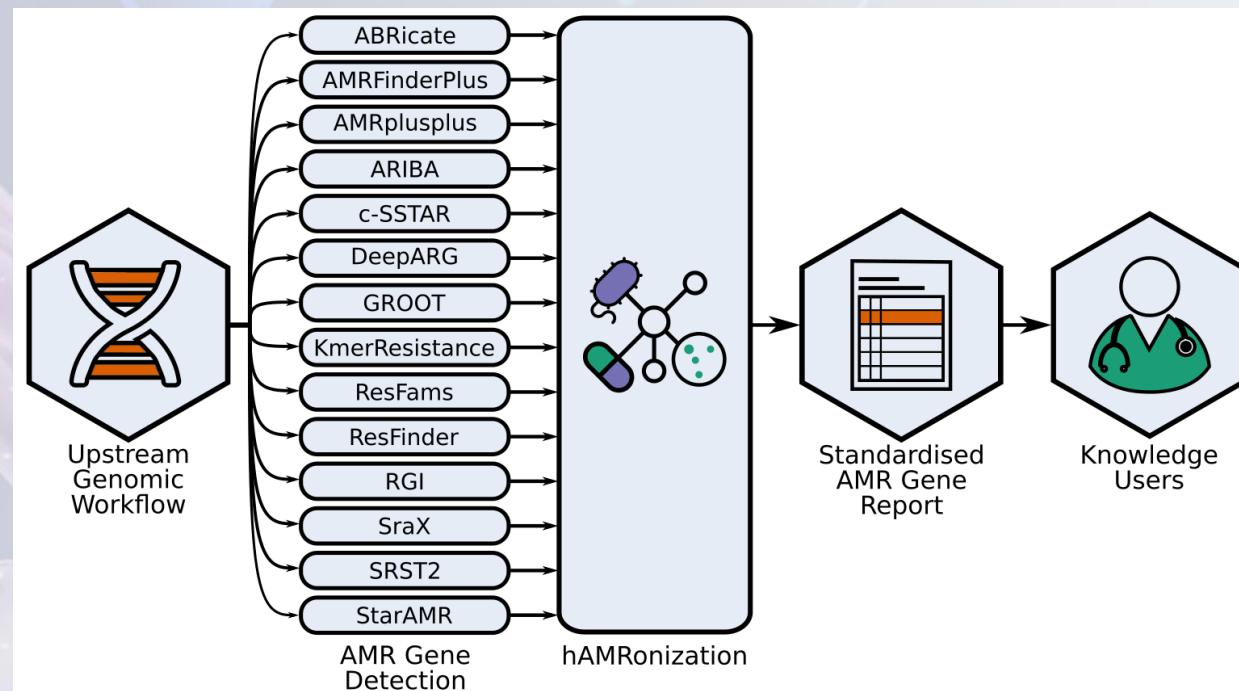


Estrategias para caracterizar los genomas



La necesidad de la estandarización

hAMRonization



<https://github.com/pha4ge/hAMRonization>

La necesidad de la estandarización

Defining genomic epidemiology thresholds for common-source bacterial outbreaks: a modelling study

Audrey Duval, Lulla Opatowski, Sylvain Brisse

- Desarrollo de modelo (SameStrain) para estimar los umbrales de distancia genética para brotes causados por una única cepa de una fuente ambiental o alimentaria contaminada
- Se contemplan características microbiológicas y epidemiológicas específicas del brote en cuestión: la tasa de mutación genética del patógeno y el tiempo desde la contaminación de la fuente inicial
- Tuvo una alta sensibilidad y especificidad para la clasificación de aislamientos cuando se probaron mediante simulaciones y los resultados de 16 conjuntos de datos publicados de brotes transmitidos por alimentos en el mundo real.
- Este enfoque se basa en la biología evolutiva y alivia la necesidad de umbrales predefinidos, que a menudo no están justificados y pueden resultar inapropiados en la mayoría de los casos.

¡Gracias!

