Bioinformática aplicada a la Microbiología Clínica

Isabel Cuesta y Sara Monzón

BU-ISCIII

Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

9 Marzo 2021

Master Bioinformática aplicada a la Medicina Personalizada y la Salud

Index

- BU-ISCIII
- High-throughput sequencing (HTS) applications in Microbiology
- Concepts: HTS and Outbreak investigation
- Bacterial and Viral Genome Sequencing
- Bioinformatics analysis in microbial genomics
- Viralrecon: SARS-CoV-2 genome reconstruction software

Index

- **BU-ISCIII**
- High-throughput sequencing (HTS) applications in Microbiology
- Concepts: HTS and Outbreak investigation
- Bacterial and Viral Genomics
- Bioinformatics analysis in microbial genomics
- Viralrecon: SARS-CoV-2 genome reconstruction software

Why BU-ISCIII was founded

Genomics Unit

2010

454

2013

NextSeq500

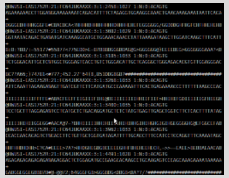
MiSeq

2021

NovaSeq 6000

Bioinformatics Unit

2012



Service & Support to Researchers on HTS Data Analysis

National Microbiology Centre (CNM)

Research Institute for Rare Diseases (IIER)

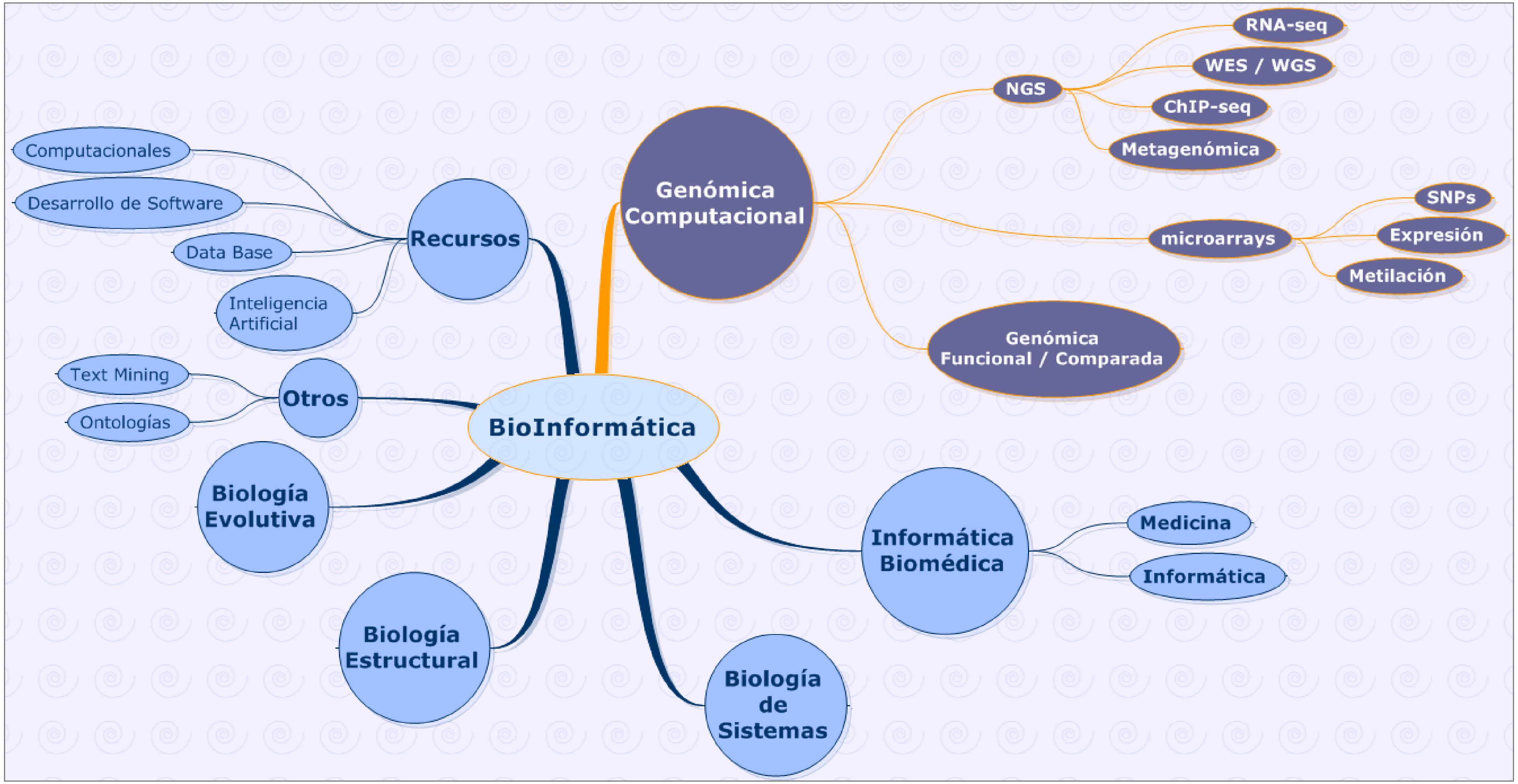
Functional Unit for Research in Chronic Disease

Network of Biological Alerts

National Centre of Tropical Medicine

National Environment Health Centre

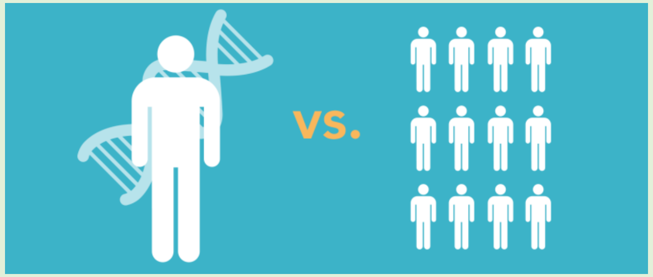
BU-ISCIII Mission - Activities

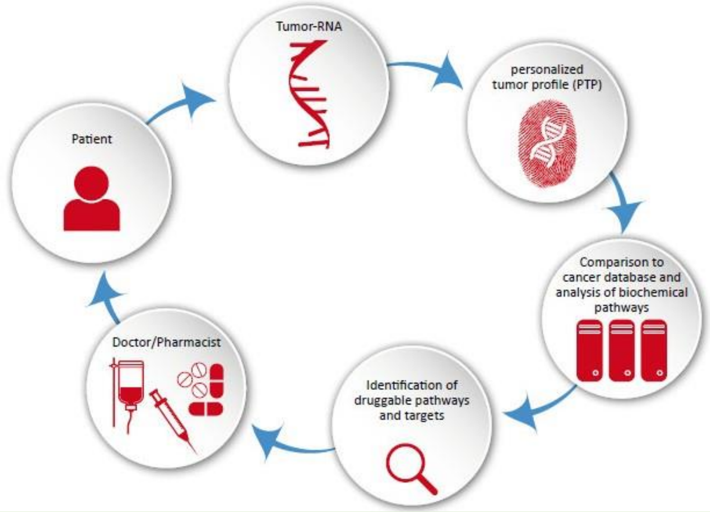


Clinical Bioinformatics - Precision Medicine

IIER

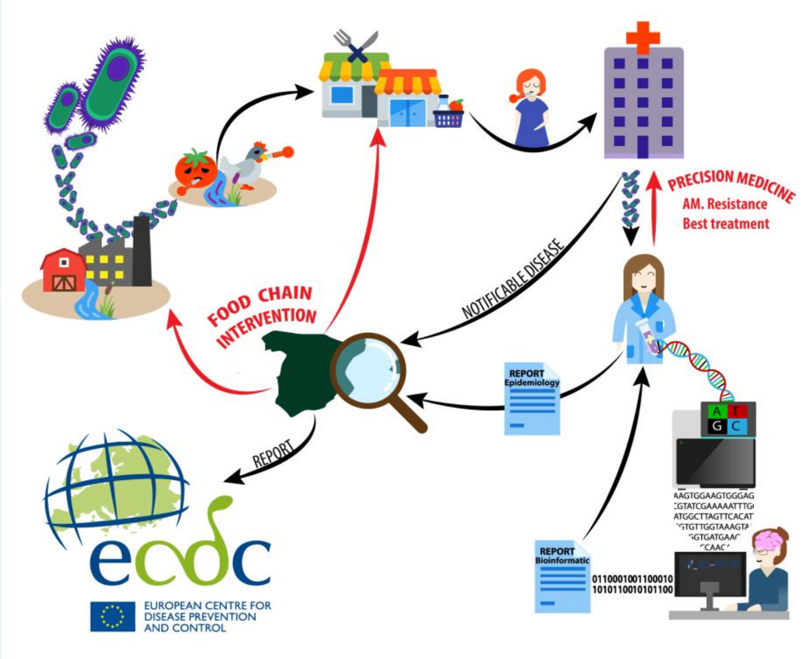
Tumour or Rare Diseases




Tumor-RNA
personalized tumor profile (PTP)
Comparison to cancer database and analysis of biochemical pathways
Identification of druggable pathways and targets
Doctor/Pharmacist
Patient

CNM

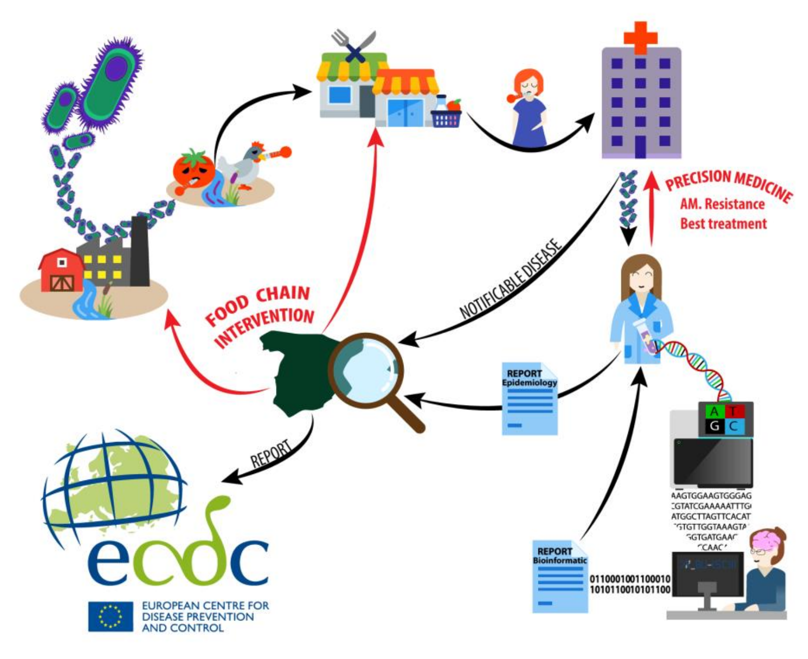
Pathogen associated



Research - Clinical Bioinformatics - Precision Medicine

CNM

Pathogen associated



AESI 2017-2019 BU-ISCIII – Genómica

AESI 2019-2021 BU-ISCIII - Genómica

AESI 2018 – 2021 **PLATAFORMA DE BIOINFORMATICA ISCIII-TransBioNet**

METAGENOMICS EQAE
Special Pathogens Unit,
P. Anda, R. Escudero, I. Jado

EMERGE
Efficient response to highly dangerous and emerging pathogens at EU level

GMI – HTS Standards, Databases Sharing and Guidelines

GMI
Global Microbial Identifier

GMI – UNSGM PT for detection of biological threats by genomic analysis – **AESI 2019**

COMPARE Food Metage

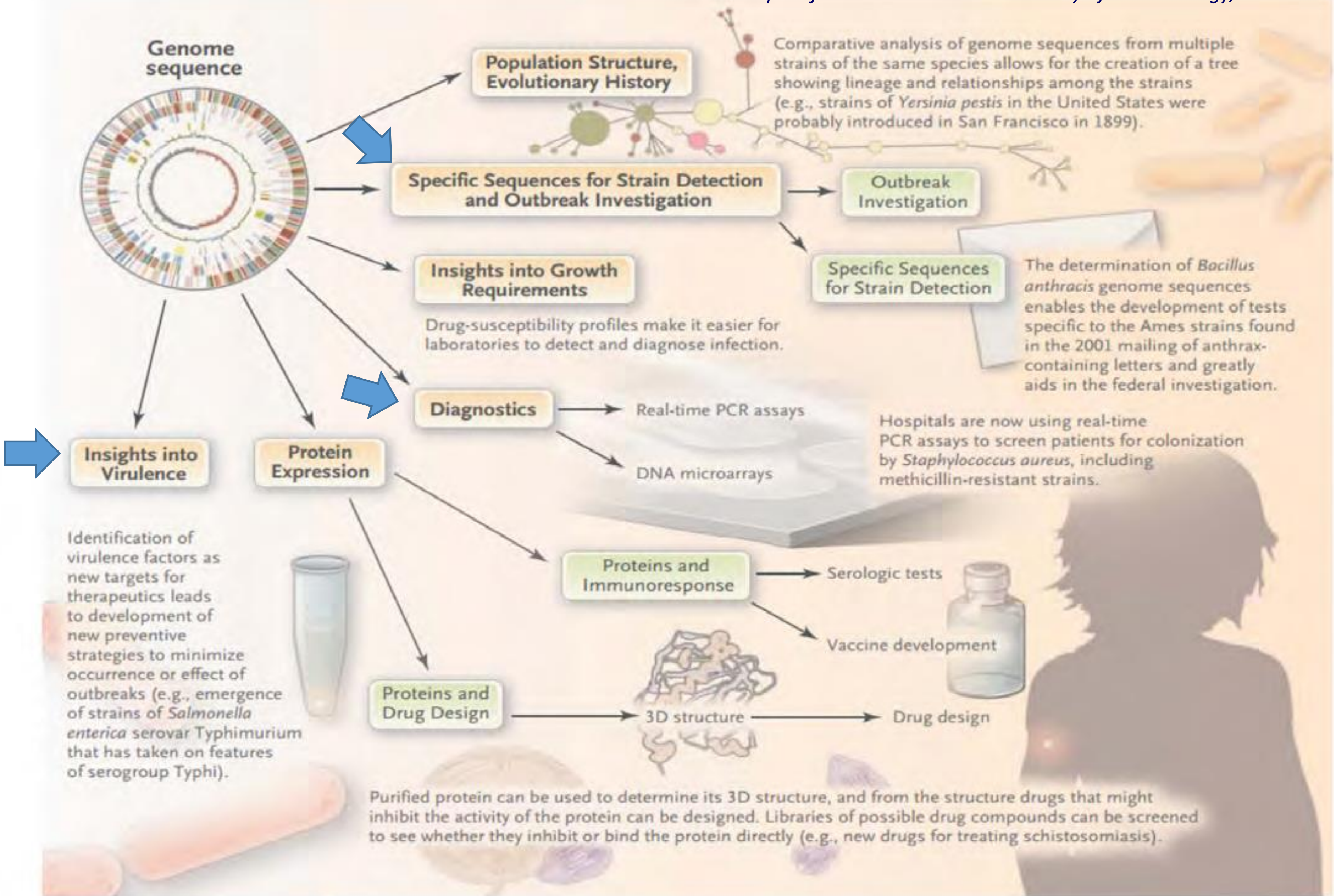
compare

Index

- BU-ISCIII
- **High-throughput sequencing (HTS) applications in Microbiology**
- Concepts: HTS and Outbreak investigation
- Bacterial and Viral Genomics
- Bioinformatics analysis in microbial genomics
- Viralrecon: SARS-CoV-2 genome reconstruction software

Microbial genomics

Report from The American Academy of Microbiology, 2015



Genome sequence

Population Structure, Evolutionary History

Comparative analysis of genome sequences from multiple strains of the same species allows for the creation of a tree showing lineage and relationships among the strains (e.g., strains of *Yersinia pestis* in the United States were probably introduced in San Francisco in 1899).

Specific Sequences for Strain Detection and Outbreak Investigation

Outbreak Investigation

Specific Sequences for Strain Detection

The determination of *Bacillus anthracis* genome sequences enables the development of tests specific to the Ames strains found in the 2001 mailing of anthrax-containing letters and greatly aids in the federal investigation.

Insights into Growth Requirements

Drug-susceptibility profiles make it easier for laboratories to detect and diagnose infection.

Diagnostics

Real-time PCR assays

DNA microarrays

Hospitals are now using real-time PCR assays to screen patients for colonization by *Staphylococcus aureus*, including methicillin-resistant strains.

Insights into Virulence

Protein Expression

Identification of virulence factors as new targets for therapeutics leads to development of new preventive strategies to minimize occurrence or effect of outbreaks (e.g., emergence of strains of *Salmonella enterica* serovar Typhimurium that has taken on features of serogroup Typhi).

Proteins and Immunoresponse

Serologic tests

Vaccine development

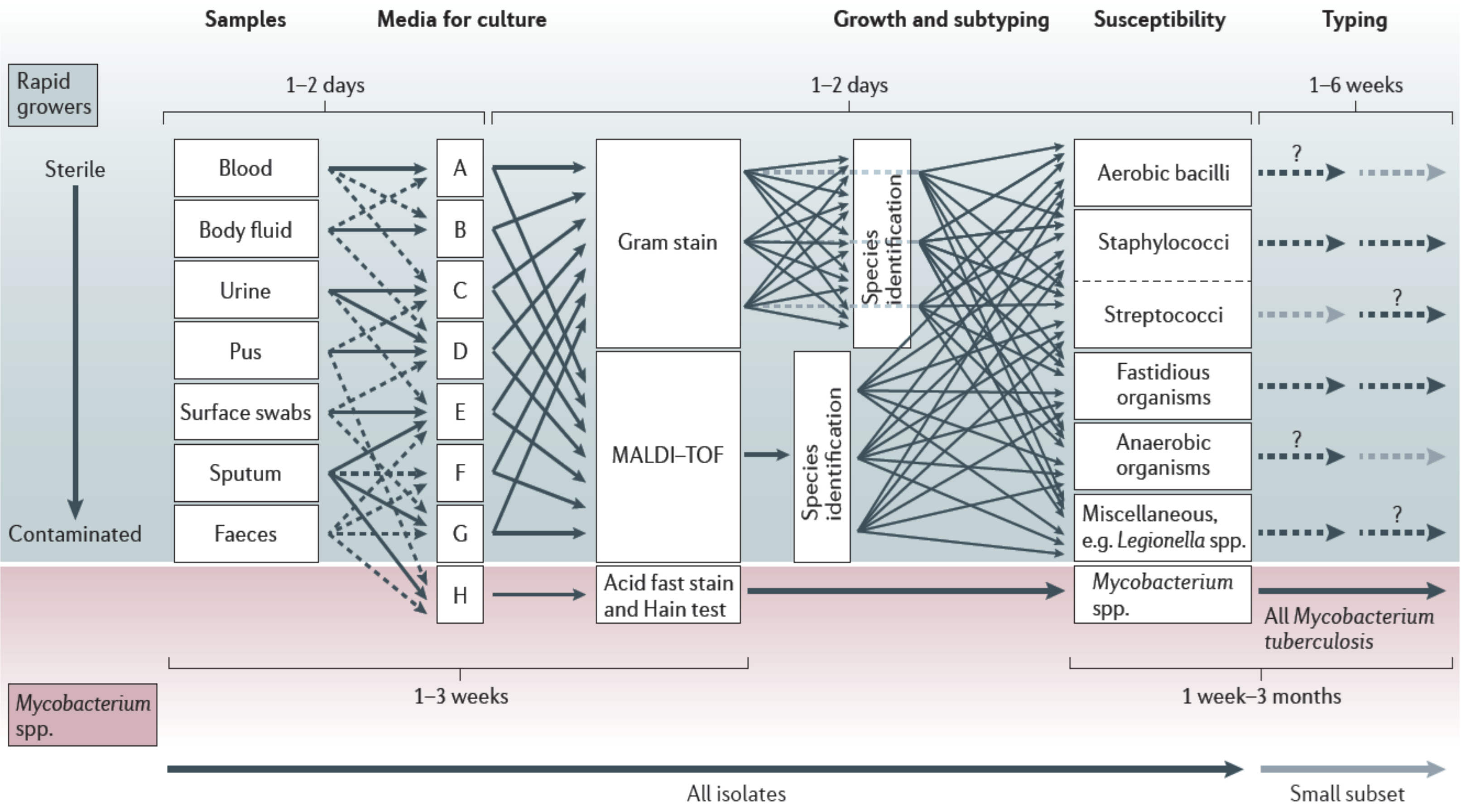
Proteins and Drug Design

3D structure

Drug design

Purified protein can be used to determine its 3D structure, and from the structure drugs that might inhibit the activity of the protein can be designed. Libraries of possible drug compounds can be screened to see whether they inhibit or bind the protein directly (e.g., new drugs for treating schistosomiasis).

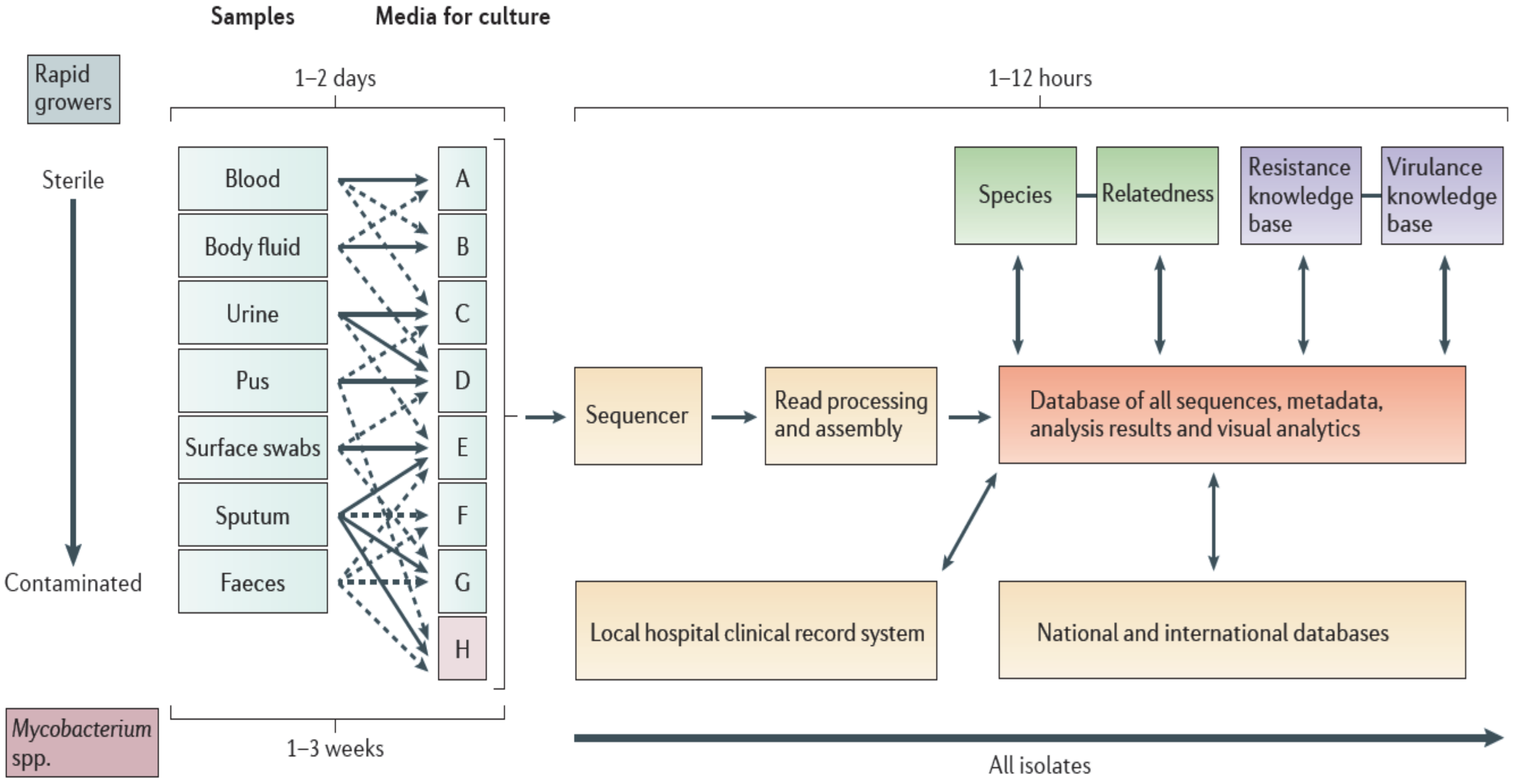
Classic techniques vs Whole Genome sequencing



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

ECDC roadmap and international commitment

Classic techniques vs Whole Genome sequencing



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.

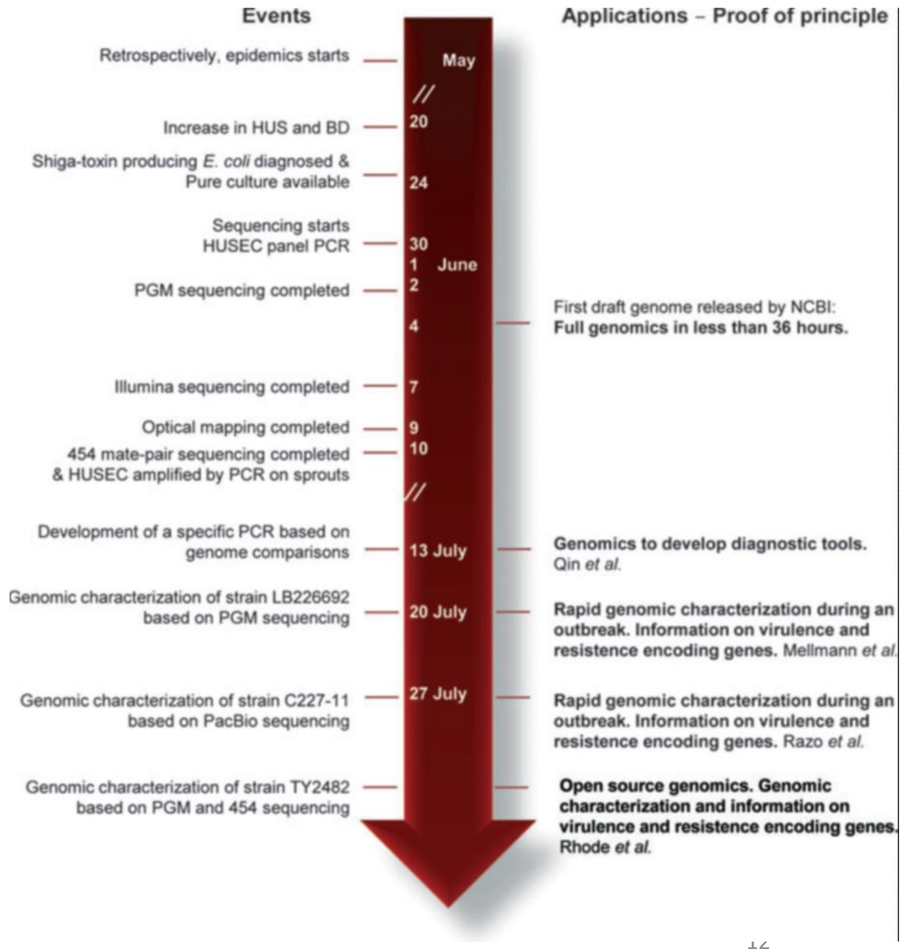
Crisis Política y Económica Europa: Alto impacto en la Economía Europea, mayor afectación en la Española

Secuenciación Genoma

华大基因 BGI

Universitätsklinikum Hamburg-Eppendorf

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


Events
Applications – Proof of principle
Retrospectively, epidemics starts
May
//
20
Increase in HUS and BD
Shiga-toxin producing *E. coli* diagnosed & Pure culture available
24
Sequencing starts HUSEC panel PCR
30
1 June
2
PGM sequencing completed
4
First draft genome released by NCBI: **Full genomics in less than 36 hours.**
Illumina sequencing completed
7
Optical mapping completed
9
10
454 mate-pair sequencing completed & HUSEC amplified by PCR on sprouts
//
Development of a specific PCR based on genome comparisons
13 July
Genomics to develop diagnostic tools. Qin *et al.*
Genomic characterization of strain LB226692 based on PGM sequencing
20 July
Rapid genomic characterization during an outbreak. Information on virulence and resistence encoding genes. Mellmann *et al.*
Genomic characterization of strain C227-11 based on PacBio sequencing
27 July
Rapid genomic characterization during an outbreak. Information on virulence and resistence encoding genes. Razo *et al.*
Genomic characterization of strain TY2482 based on PGM and 454 sequencing
Open source genomics. Genomic characterization and information on virulence and resistence encoding genes. Rhode *et al.*

Bertelli and Greub, Clin Microb and Infect, 2013

Foodborne outbreak identification “Crisis del pepino”

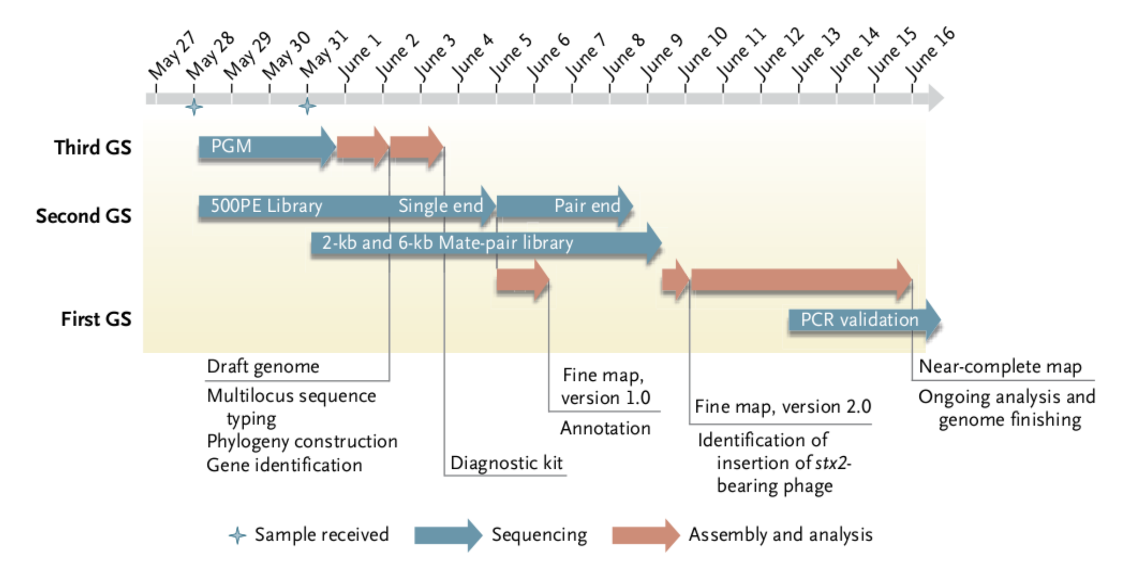


Figure 1. Timeline of the Open-Source Genomics Program.

After receiving the first batch of DNA samples on May 28, 2011, sequencing runs with the use of the Ion Torrent Personal Genome Machine (PGM) and Illumina (small-insert library) were initiated simultaneously. On May 31, the second batch of DNA was received and used for Illumina large-insert sequencing. An assembly of the Ion Torrent reads was released on June 2, which enabled subsequent analyses (multilocus sequence typing, phylogenetic analysis, and genome comparisons). Errors in the Ion Torrent data were corrected with the use of later Illumina data, and a high-quality draft genome sequence was created. GS denotes generation of sequencing technology. The symbols at May 28 and May 31 in the timeline indicate the arrival of DNA samples.

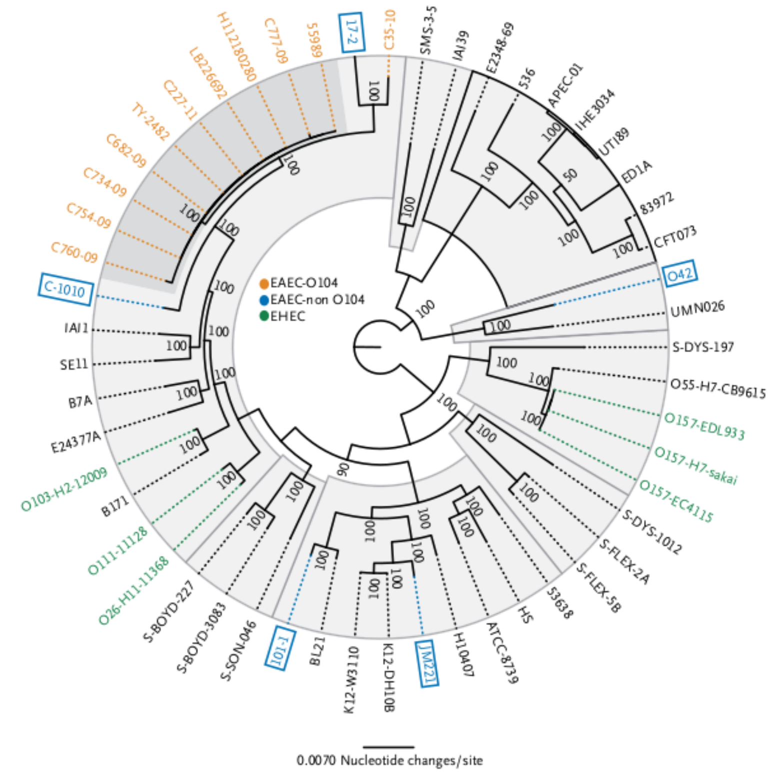
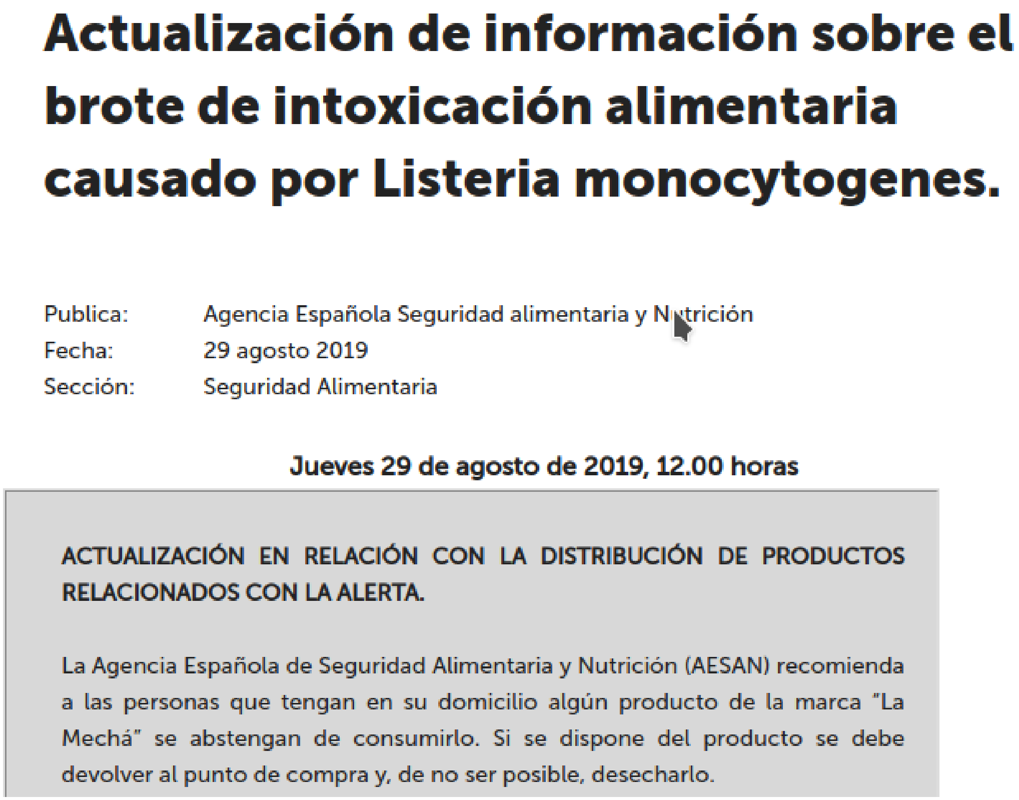


Figure 2. Phylogenetic Comparisons of 53 *Escherichia coli* and Shigella Isolates.

Genomic sequences were compared with the use of 100 bootstrap calculations, as described by Sahl et al.[35] The species-based phylogeny was inferred with the use of 2.56 Mbp of the conserved core genome. The O104:H4 isolates are shown in orange, the reference enteroaggregative *E. coli* (EAEC) isolates in blue, and the enterohemorrhagic *E. coli* isolates in green. (The classification of the other strains is shown in Fig. 4 and Table 4 in the Supplementary Appendix.) The O104:H4 isolates cluster into a single clade (dark gray); in contrast, the reference EAEC isolates are extremely divergent and are represented throughout the phylogeny.

Rohde et al NEJM 2011, 365:718-24

Andalusian Listeria Outbreak



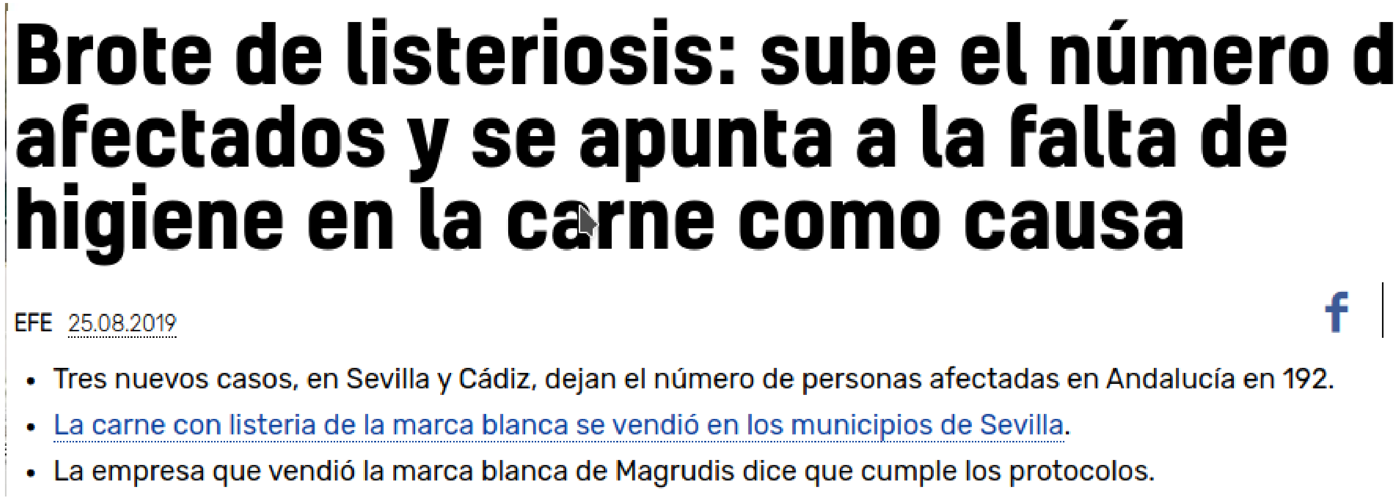
Actualización de información sobre el brote de intoxicación alimentaria causado por Listeria monocytogenes.

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.



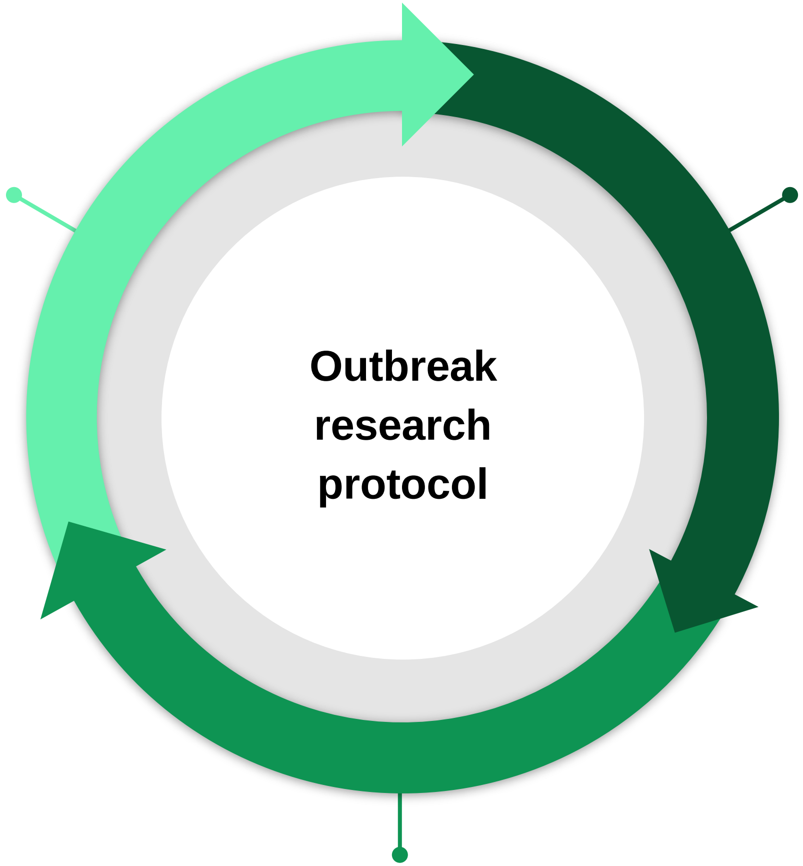
Brote de listeriosis: sube el número d afectados y se apunta a la falta de higiene en la carne como causa

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


Sequencing

Bioinformatics Analysis

Outbreak research protocol

Microbiology lab interpretation and reporting

Transversal:

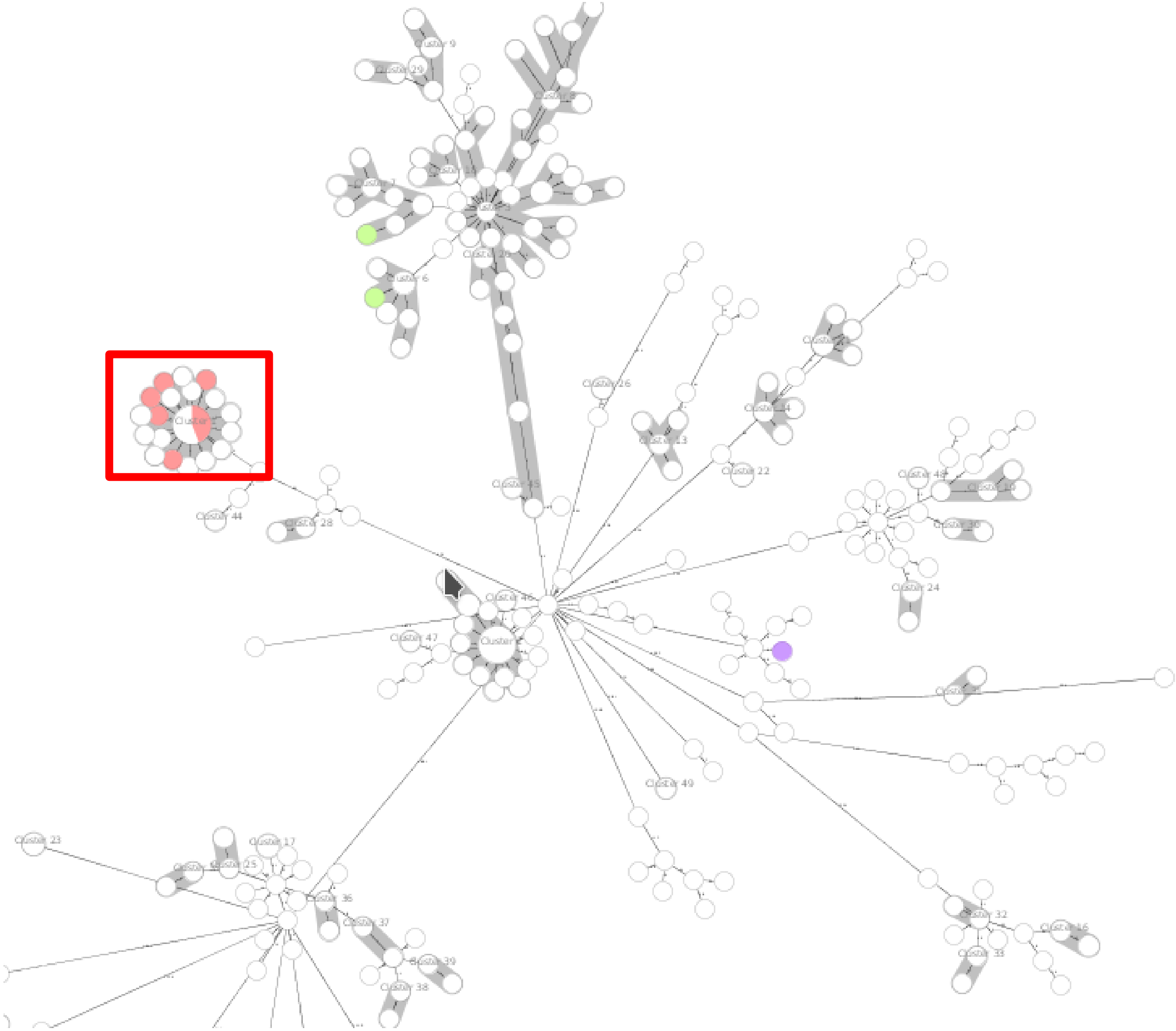
- Data management
- Sample tracking
- Infrastructure

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.



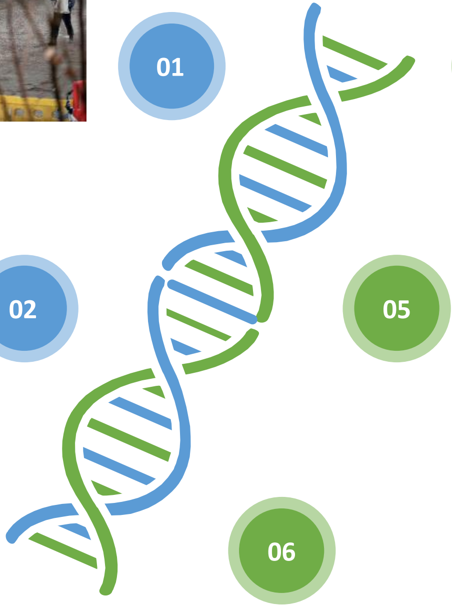
Pathogen discovery: new virus – SARS-CoV-2

Deep Meta-Transcriptomic Sequencing

bronchoalveolar lavage fluid (BALF)

Meta-transcriptomic library

2x150 MiniSeq 56,565,928 sequences reads

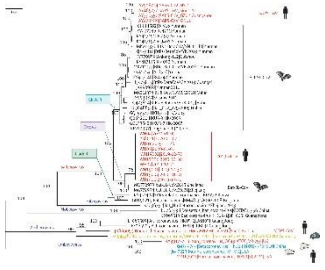


De novo-assembled - Megahit

384,096 Contigs

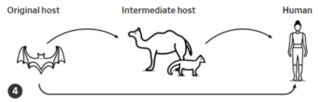
Screened for potential aetiological agents

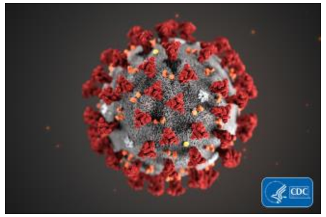
The longest 30,474 nt



89.1% identity

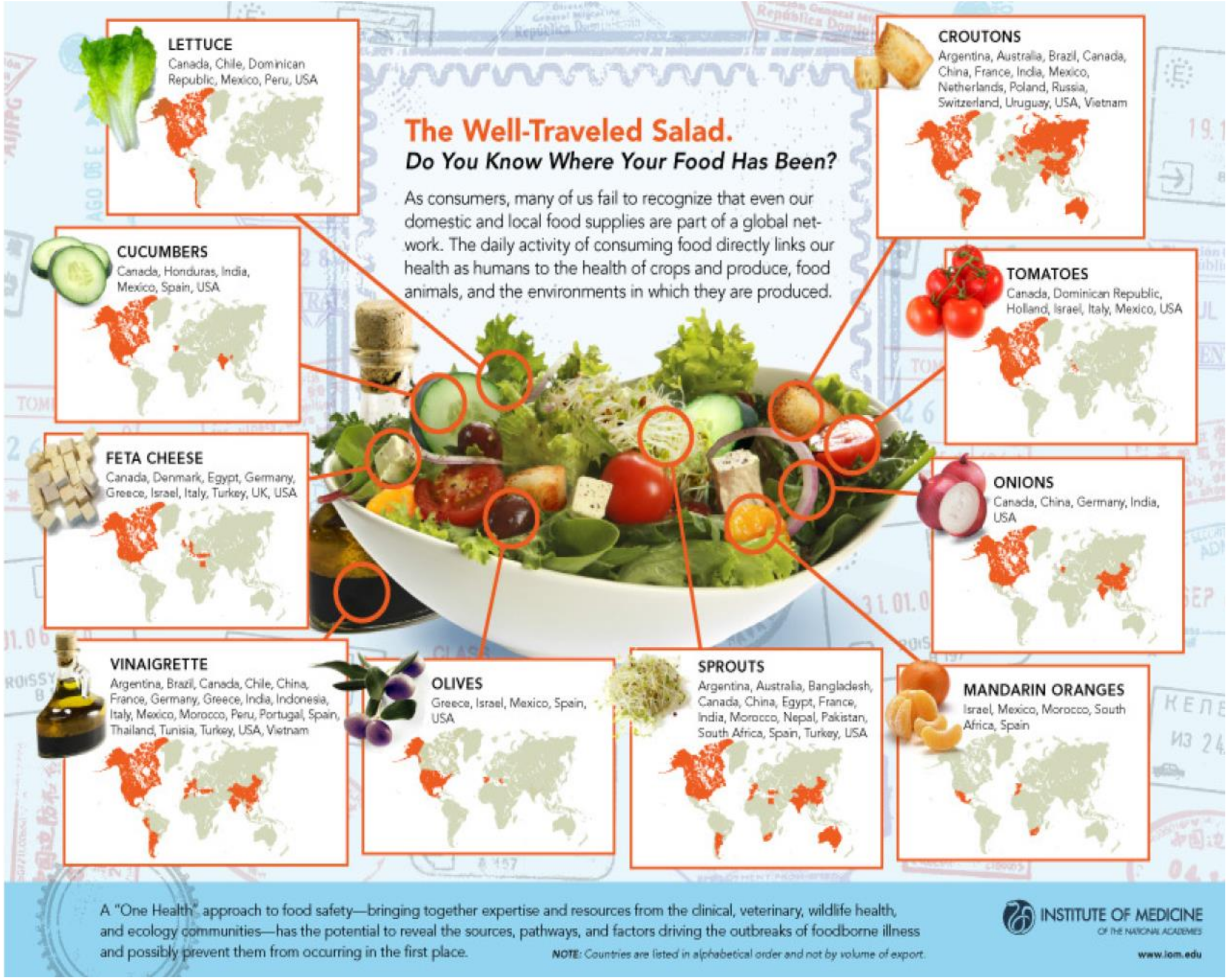
Closely related to a bat SARS-like coronavirus





Wu et al., Nature 2020

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



The Well-Traveled Salad.

Do You Know Where Your Food Has Been?

As consumers, many of us fail to recognize that even our domestic and local food supplies are part of a global network. The daily activity of consuming food directly links our health as humans to the health of crops and produce, food animals, and the environments in which they are produced.

LETTUCE
Canada, Chile, Dominican Republic, Mexico, Peru, USA

CUCUMBERS
Canada, Honduras, India, Mexico, Spain, USA

FETA CHEESE
Canada, Denmark, Egypt, Germany, Greece, Israel, Italy, Turkey, UK, USA

VINAIGRETTE
Argentina, Brazil, Canada, Chile, China, France, Germany, Greece, India, Indonesia, Italy, Mexico, Morocco, Peru, Portugal, Spain, Thailand, Tunisia, Turkey, USA, Vietnam

OLIVES
Greece, Israel, Mexico, Spain, USA

SPROUTS
Argentina, Australia, Bangladesh, Canada, China, Egypt, France, India, Morocco, Nepal, Pakistan, South Africa, Spain, Turkey, USA

CROUTONS
Argentina, Australia, Brazil, Canada, China, France, India, Mexico, Netherlands, Poland, Russia, Switzerland, Uruguay, USA, Vietnam

TOMATOES
Canada, Dominican Republic, Holland, Israel, Italy, Mexico, USA

ONIONS
Canada, China, Germany, India, USA

MANDARIN ORANGES
Israel, Mexico, Morocco, South Africa, Spain

A "One Health" approach to food safety—bringing together expertise and resources from the clinical, veterinary, wildlife health, and ecology communities—has the potential to reveal the sources, pathways, and factors driving the outbreaks of foodborne illness and possibly prevent them from occurring in the first place.

NOTE: *Countries are listed in alphabetical order and not by volume of export.*

INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES
www.iom.edu

GMI
Global Microbial Identifier

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:

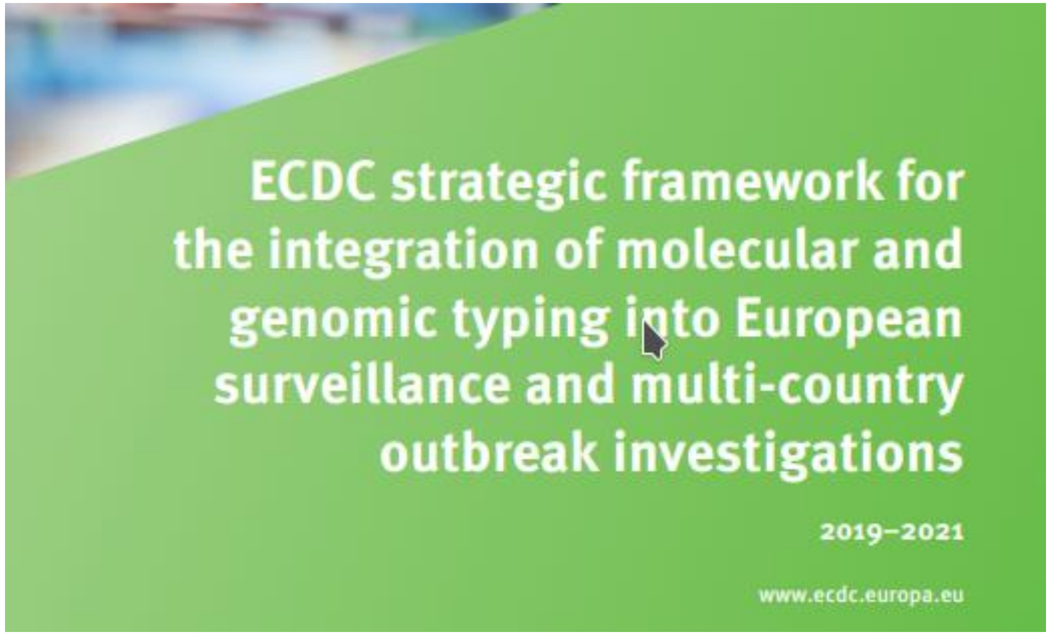
- Diagnosis
- **Surveillance** → Outbreak research: Molecular source detection
- Infectious diseases research
- Training

ECDC roadmap and international commitment

ECDC roadmap for integration of molecular and genomic typing into European-level surveillance and epidemic preparedness

Version 2.1, 2016–2019

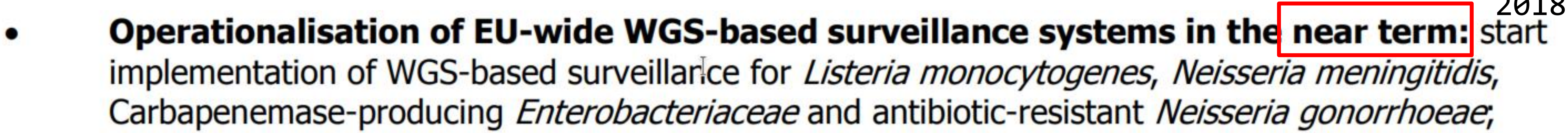
www.ecdc.europa.eu



ECDC strategic framework for the integration of molecular and genomic typing into European surveillance and multi-country outbreak investigations

2019–2021

www.ecdc.europa.eu



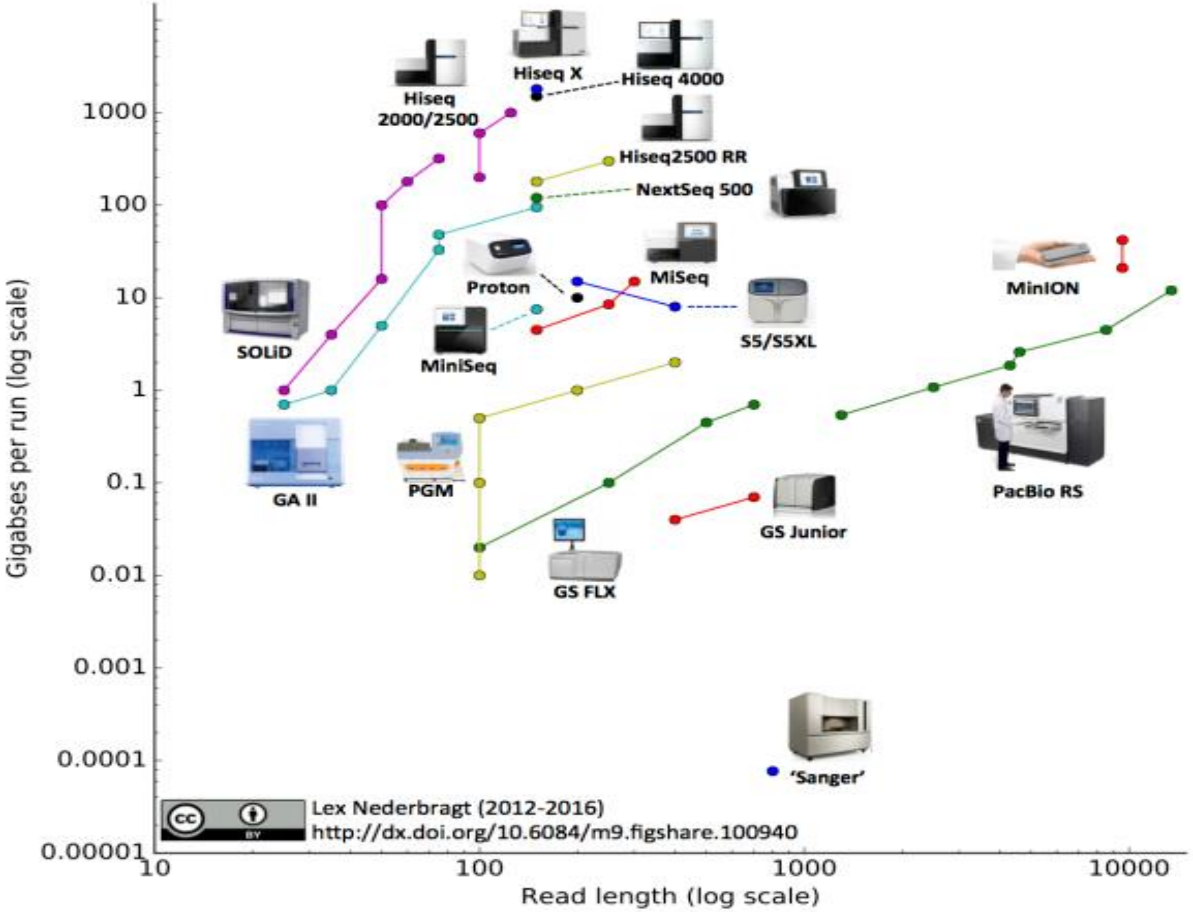
2018

- **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*;

Index

- BU-ISCIII
- High-throughput sequencing (HTS) applications in Microbiology
- **Concepts: HTS and Outbreak investigation**
- Bacterial and Viral Genomics
- Bioinformatics analysis in microbial genomics
- Viralrecon: SARS-CoV-2 genome reconstruction software

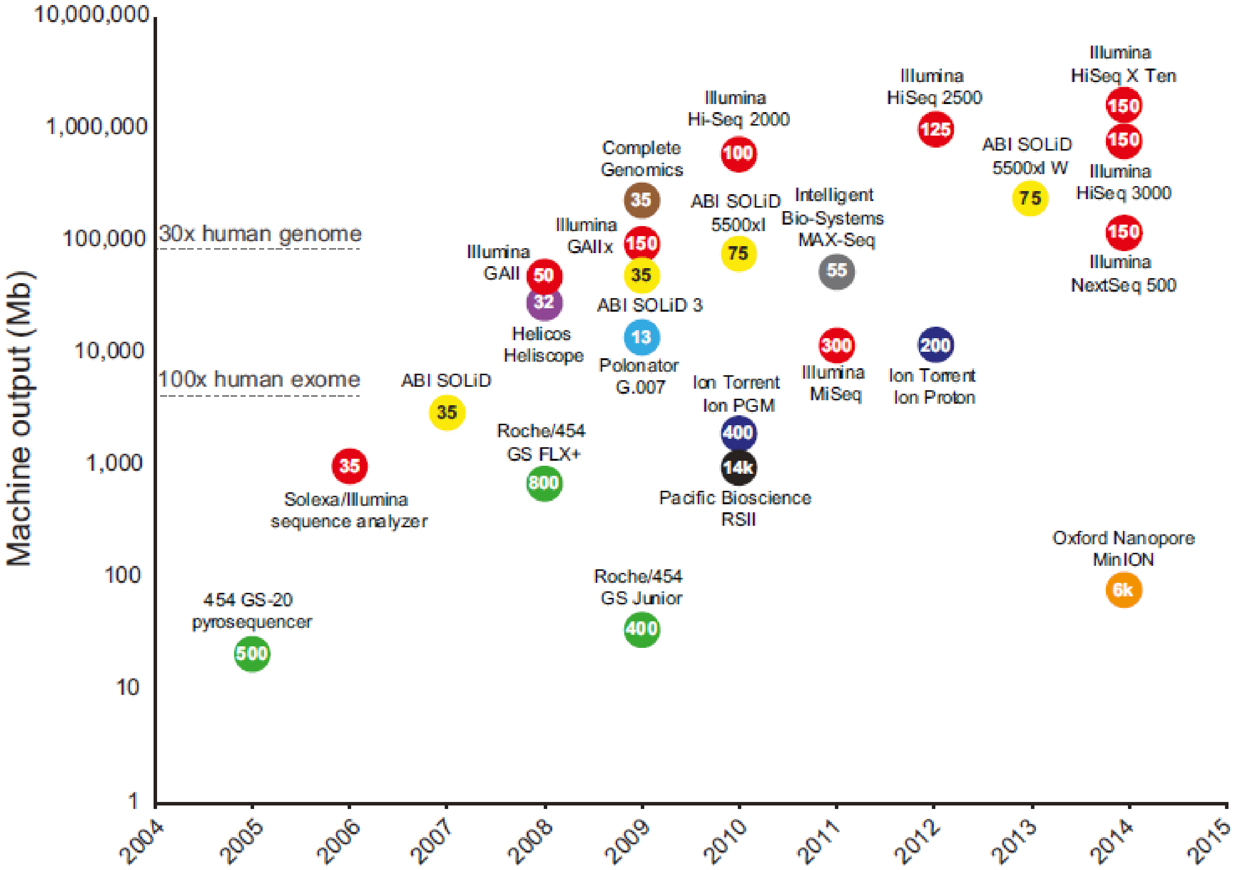
High-Throughput Sequencing Technologies



Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940

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

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

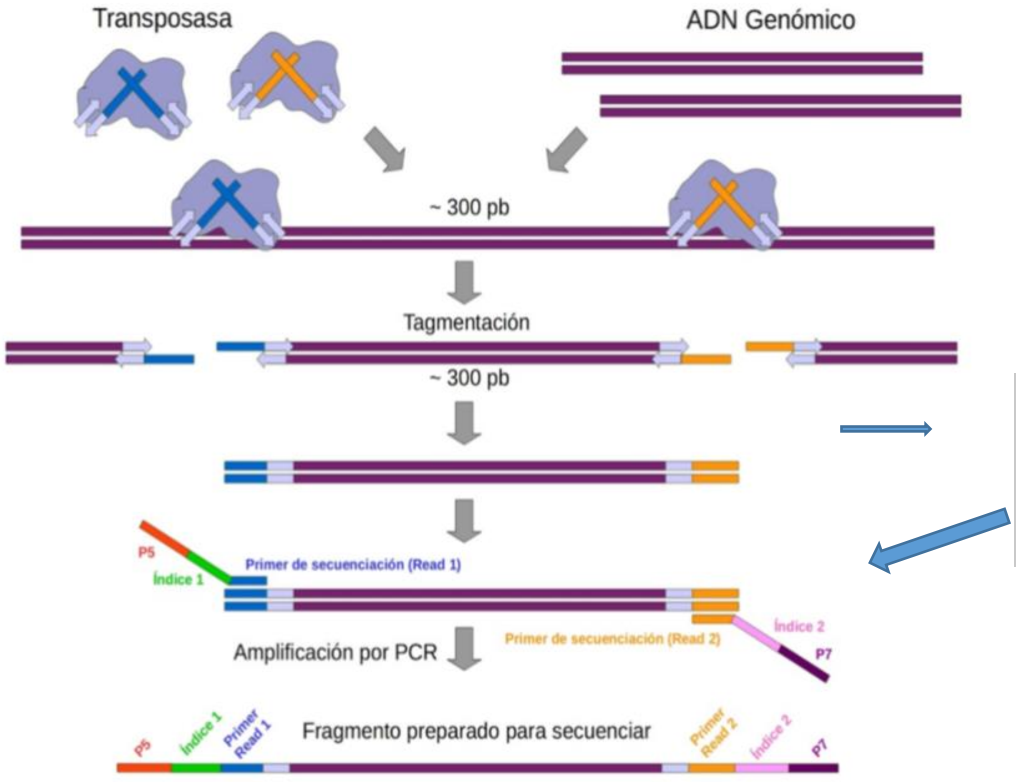
IDENTIFICACIÓN MICROORGANISMOS

1. Metataxonomía

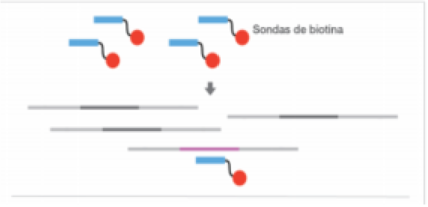
PREPARACIÓN LIBRERÍA

ENZIMÁTICA
FÍSICA

RNA -> cDNA

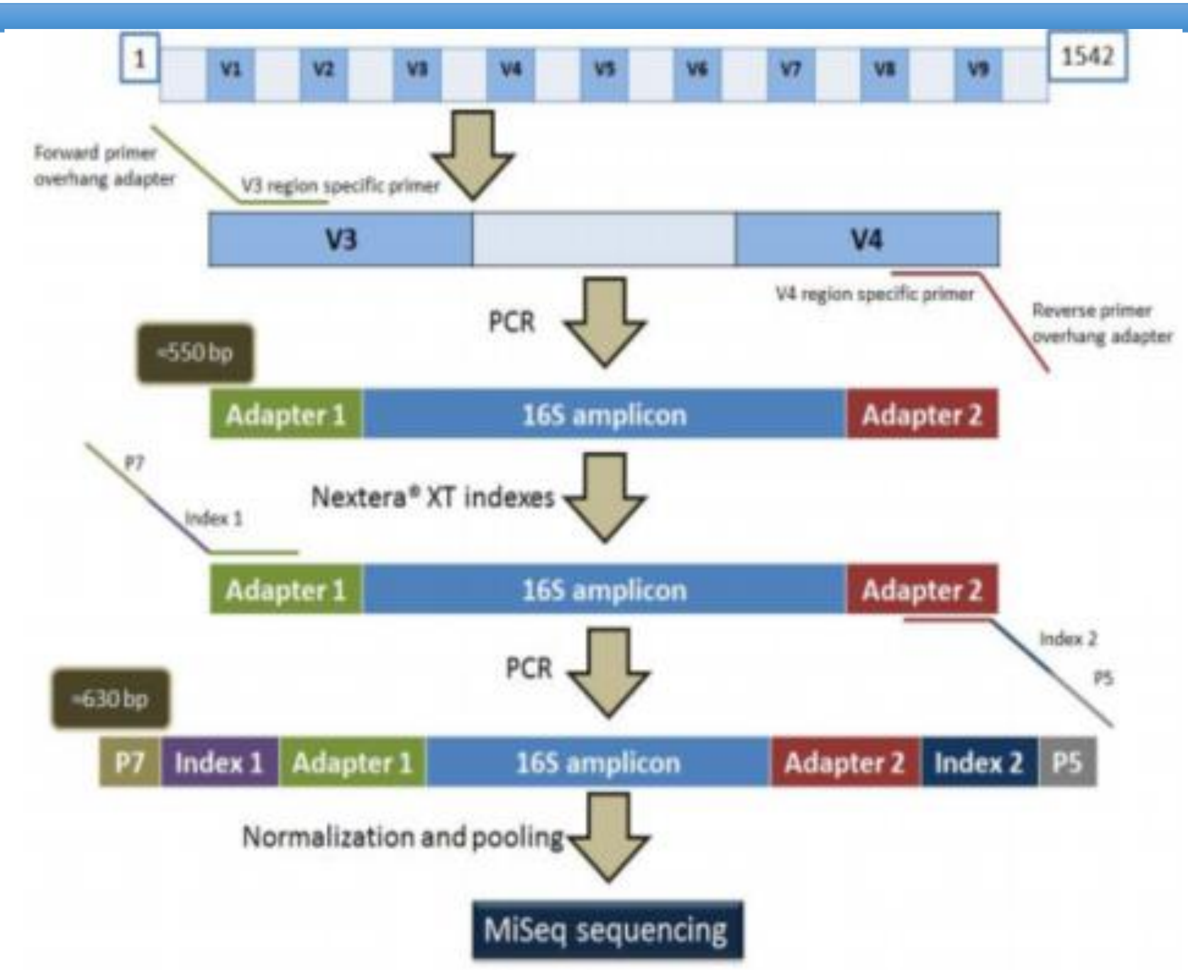


ENRIQUECIMIENTO:
PCR
CAPTURA SONDAS

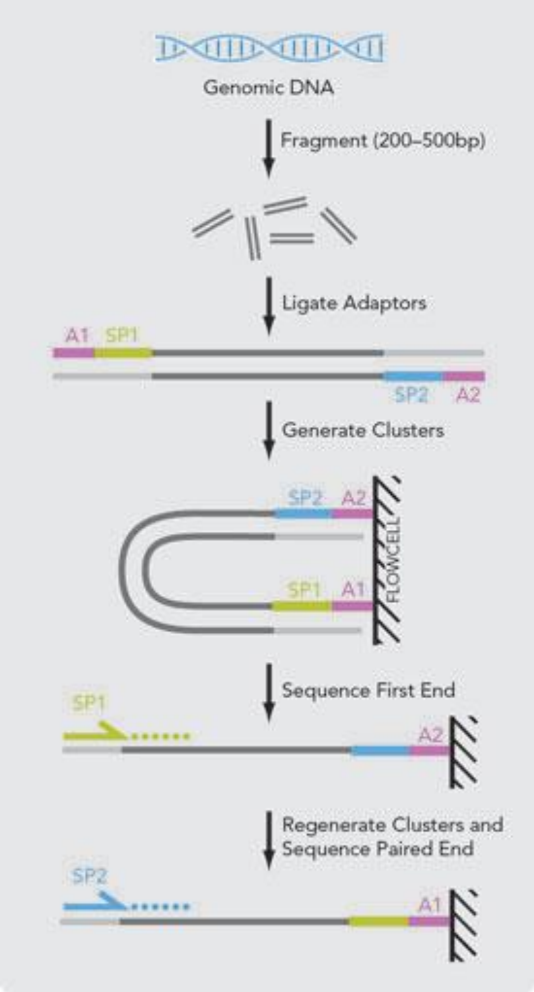


Guia Práctica Genómica https://www.uv.es/varnau/GM_Cap%C3%ADtulo_2.pdf

PREPARACIÓN LIBRERÍA, rRNA 16S, caracterización microbiota



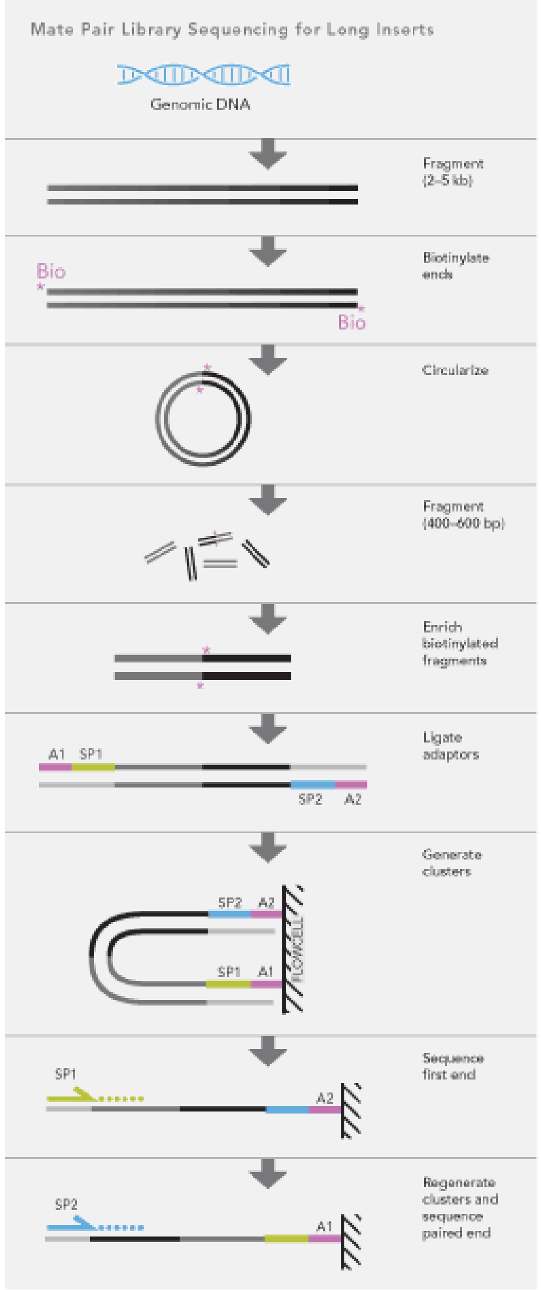
Que es Pair-end?



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

Breadth of coverage

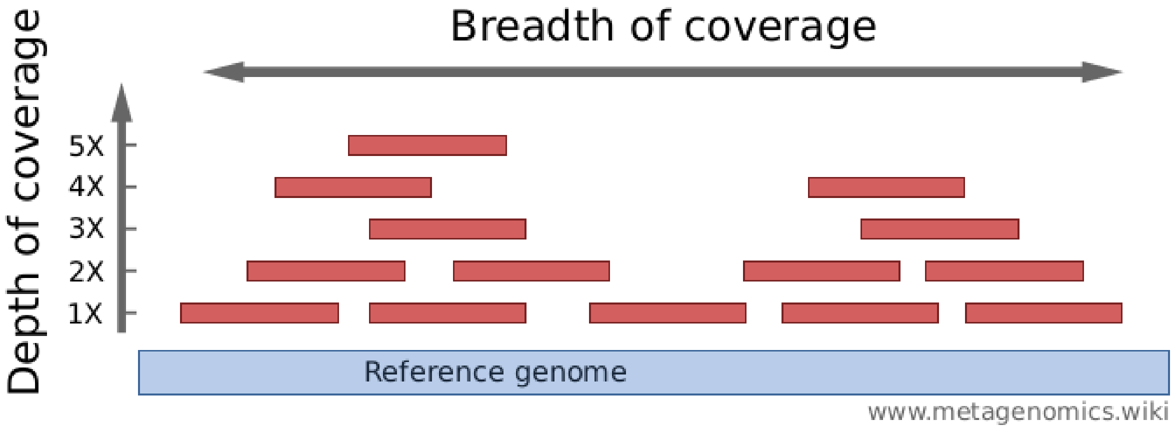
How much of a genome is "covered" by short reads? Are there regions that are not covered, even not by a single read?

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

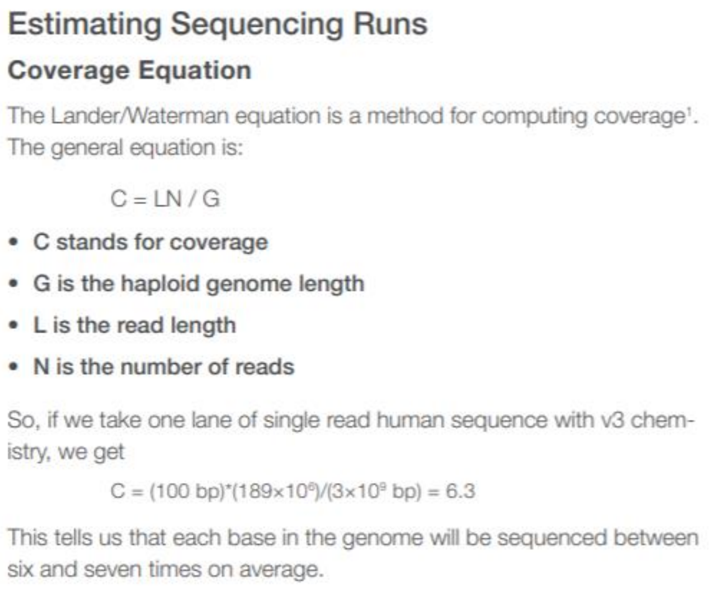
How strong is a genome "covered" by sequenced fragments (short reads)?

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

Calculo de cobertura: número de lecturas



Estimating Sequencing Runs

Coverage Equation

The Lander/Waterman equation is a method for computing coverage[1]. The general equation is:

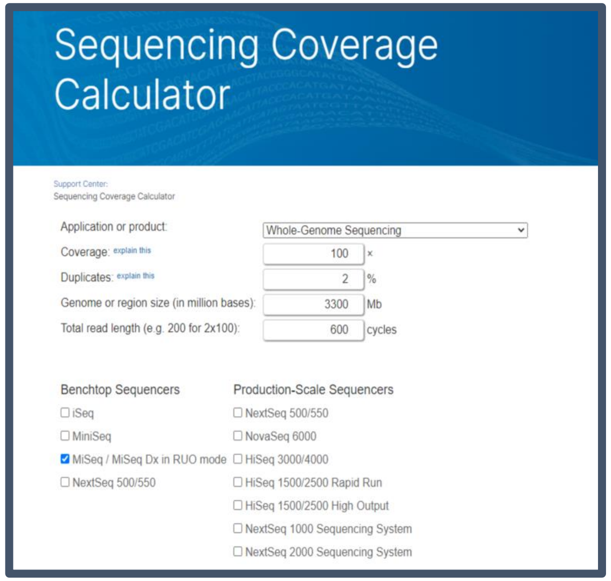
$$C = LN / G$$

- **C stands for coverage**
- **G is the haploid genome length**
- **L is the read length**
- **N is the number of reads**

So, if we take one lane of single read human sequence with v3 chemistry, we get

$$C = (100\ \text{bp})^*(189 \times 10^6)/(3 \times 10^9\ \text{bp}) = 6.3$$

This tells us that each base in the genome will be sequenced between six and seven times on average.



Sequencing Coverage Calculator

Support Center:
Sequencing Coverage Calculator

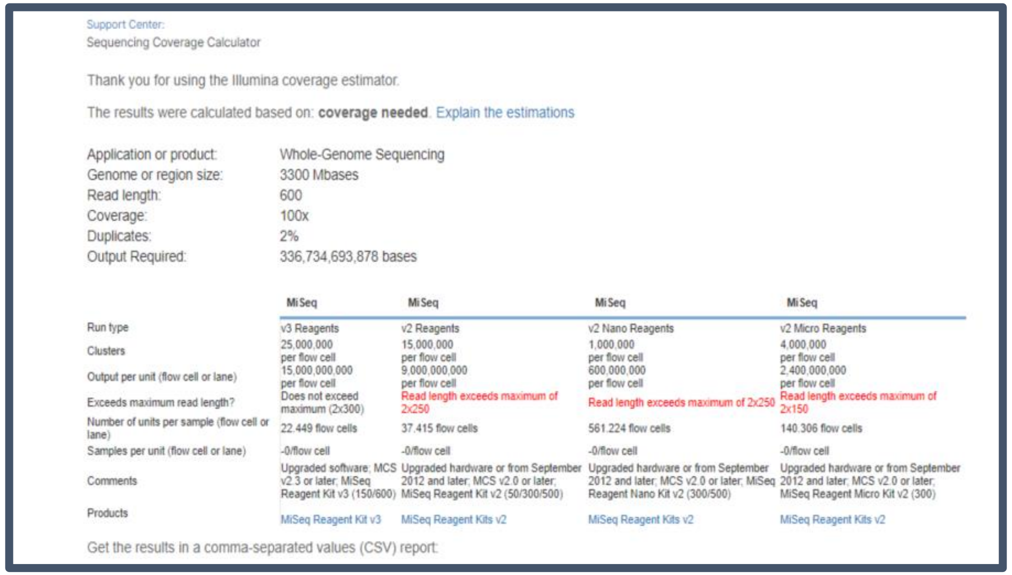
Application or product:	Whole-Genome Sequencing	
Coverage: explain this	100	×
Duplicates: explain this	2	%
Genome or region size (in million bases):	3300	Mb
Total read length (e.g. 200 for 2x100):	600	cycles

Benchtop Sequencers

- ☐ iSeq
- ☐ MiniSeq
- ☑ MiSeq / MiSeq Dx in RUO mode
- ☐ NextSeq 500/550

Production-Scale Sequencers

- ☐ NextSeq 500/550
- ☐ NovaSeq 6000
- ☐ HiSeq 3000/4000
- ☐ 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

	MiSeq	MiSeq	MiSeq	MiSeq
Run type	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)	-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)	Upgraded hardware or from September 2012 and later; MCS v2.0 or later; 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

Index

- BU-ISCIII
- High-throughput sequencing (HTS) applications in Microbiology. Examples
- Concepts: HTS and **Outbreak investigation**
- Bacterial and Viral Genomics
- Bioinformatics analysis in microbial genomics
- Viralrecon: SARS-CoV-2 genome reconstruction software

Outbreak definition and Typing methods: DNA-based methods

A disease **OUTBREAK** is the occurrence of disease cases in excess of normal expectancy.

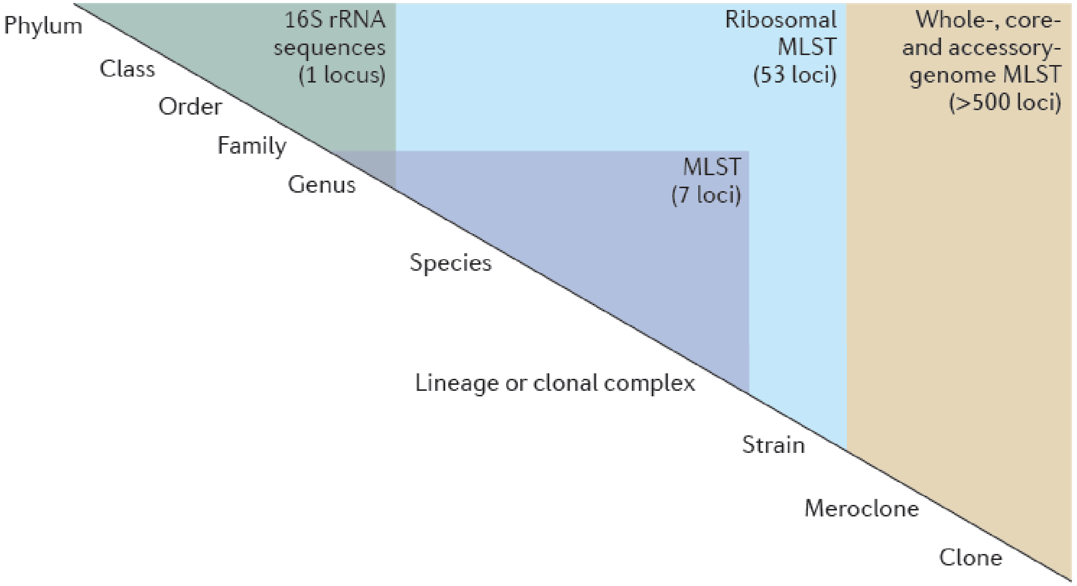
Bacterial identification and characterization at subspecies level is commonly known as **Microbial Typing**. Currently, these methodologies are fundamental tools in Clinical Microbiology and bacterial population genetics studies to track outbreaks and to study the dissemination and evolution of virulence or pathogenicity factors and antimicrobial resistance

Several typing methods have been used in outbreak detection and epidemiological surveillance ranging from **phenotypic methods to fragment based methods and sequence based methods.**

WHAT IS MOLECULAR TYPING?

Molecular typing is a way of identifying specific strains of microorganisms, such as bacteria or viruses, by looking at their genetic material. It is mainly used in outbreak investigation as pinpoint the **source of foodborne outbreaks**. It can also be used to identify which microorganisms are:
Most virulent and cause serious diseases, resistant to antibiotics, or able to survive and multiply.

Sequence data for taxonomy and typing


Phylum
Class
Order
Family
Genus
Species
Lineage or clonal complex
Strain
Meroclone
Clone
16S rRNA sequences (1 locus)
Ribosomal MLST (53 loci)
Whole-, core- and accessory-genome MLST (>500 loci)
MLST (7 loci)

Different levels of sequence information can be associated with different taxonomic levels.

The need for higher-resolution characterization of isolates has led to the development of a wide range of strain-typing methods

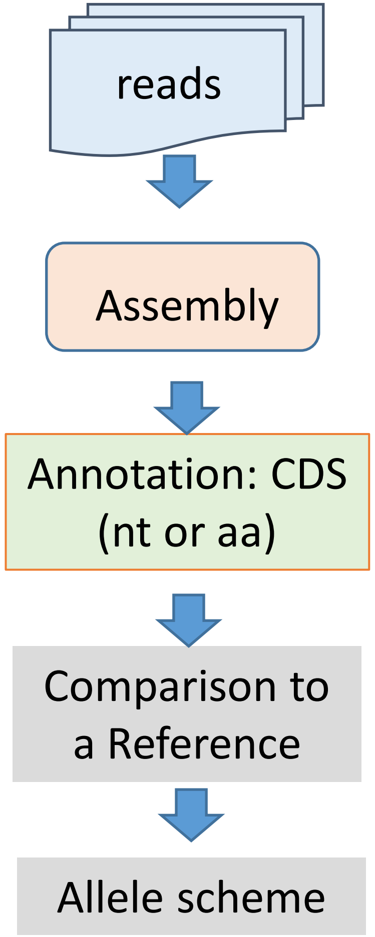
Concepts

Core genome: the number of shared features in a pool of genomes. Shared genes among multiple strains are mostly related to house-keeping genes or central metabolic processes, most of the structural information and main genotypic features. **Orthologues** (sequences have common ancestor and have split due to speciation event) in all genomes of bacteria belonging to the same taxa

Accessory genome or adaptative genome: includes genes conferring adaptive advantages to the strain in order to survive in a specific environment. In most cases, these factors are linked to antibiotic resistance, virulence, capsular serotype, adaptation, and might reflect the organisms predominant lifestyle.

Pangenome: The term "pan-genome" refers to pan (from Greek παν, whole) and genome (genome) referring to the inclusion of the core and the dispensable genome.

General analytical process for cgMLST / wgMLST



reads

Assembly

Annotation: CDS (nt or aa)

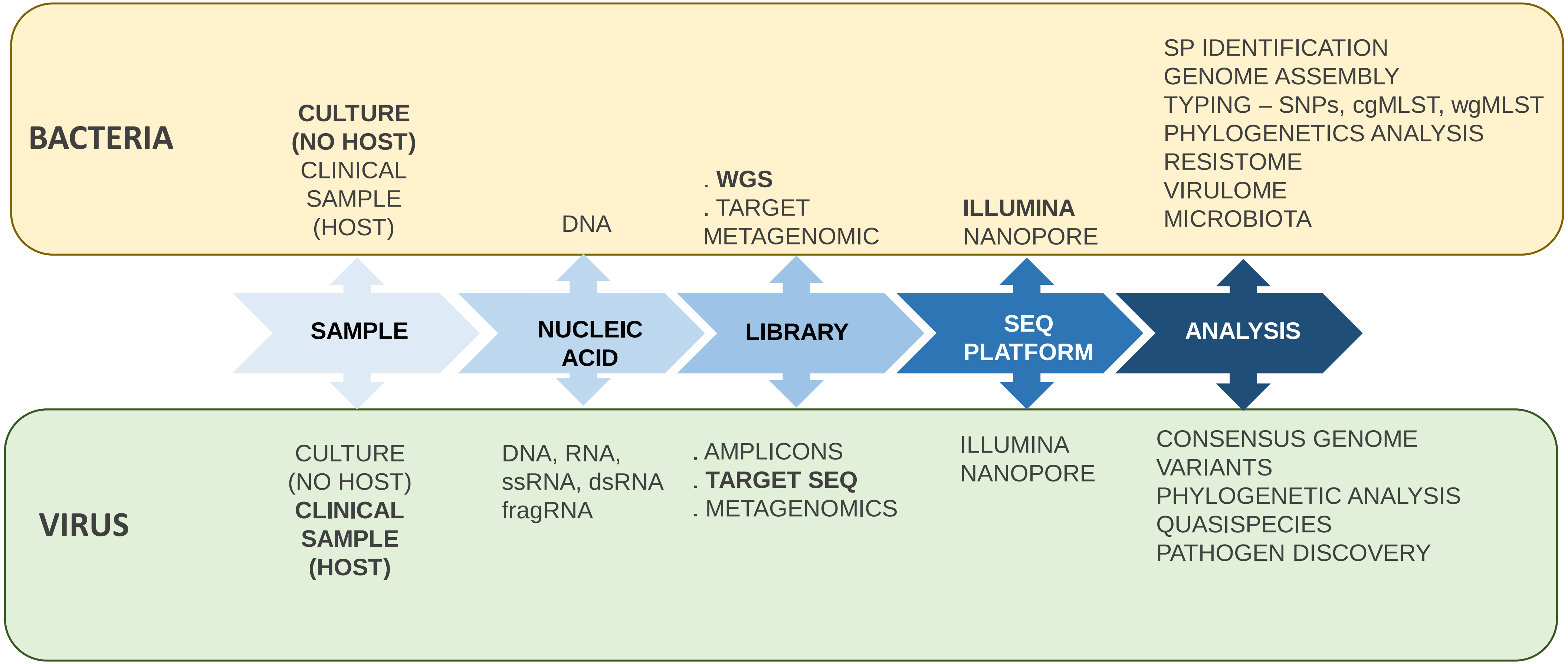
Comparison to a Reference

Allele scheme

Index

- BU-ISCIII
- High-throughput sequencing (HTS) applications in Microbiology. Examples
- Concepts: HTS and Outbreak investigation
- **Bacterial and Viral Genome Sequencing**
- Bioinformatics analysis in microbial genomics
- Viralrecon: SARS-CoV-2 genome reconstruction software

Bacterial and Viral Genome Sequencing



BACTERIA

CULTURE (NO HOST)
CLINICAL SAMPLE (HOST)

DNA

. **WGS**
. TARGET
METAGENOMIC

ILLUMINA
NANOPORE

SP IDENTIFICATION
GENOME ASSEMBLY
TYPING – SNPs, cgMLST, wgMLST
PHYLOGENETICS ANALYSIS
RESISTOME
VIRULOME
MICROBIOTA

SAMPLE → **NUCLEIC ACID** → **LIBRARY** → **SEQ PLATFORM** → **ANALYSIS**

VIRUS

CULTURE (NO HOST)
CLINICAL SAMPLE (HOST)

DNA, RNA,
ssRNA, dsRNA
fragRNA

. AMPLICONS
. **TARGET SEQ**
. METAGENOMICS

ILLUMINA
NANOPORE

CONSENSUS GENOME
VARIANTS
PHYLOGENETIC ANALYSIS
QUASISPECIES
PATHOGEN DISCOVERY

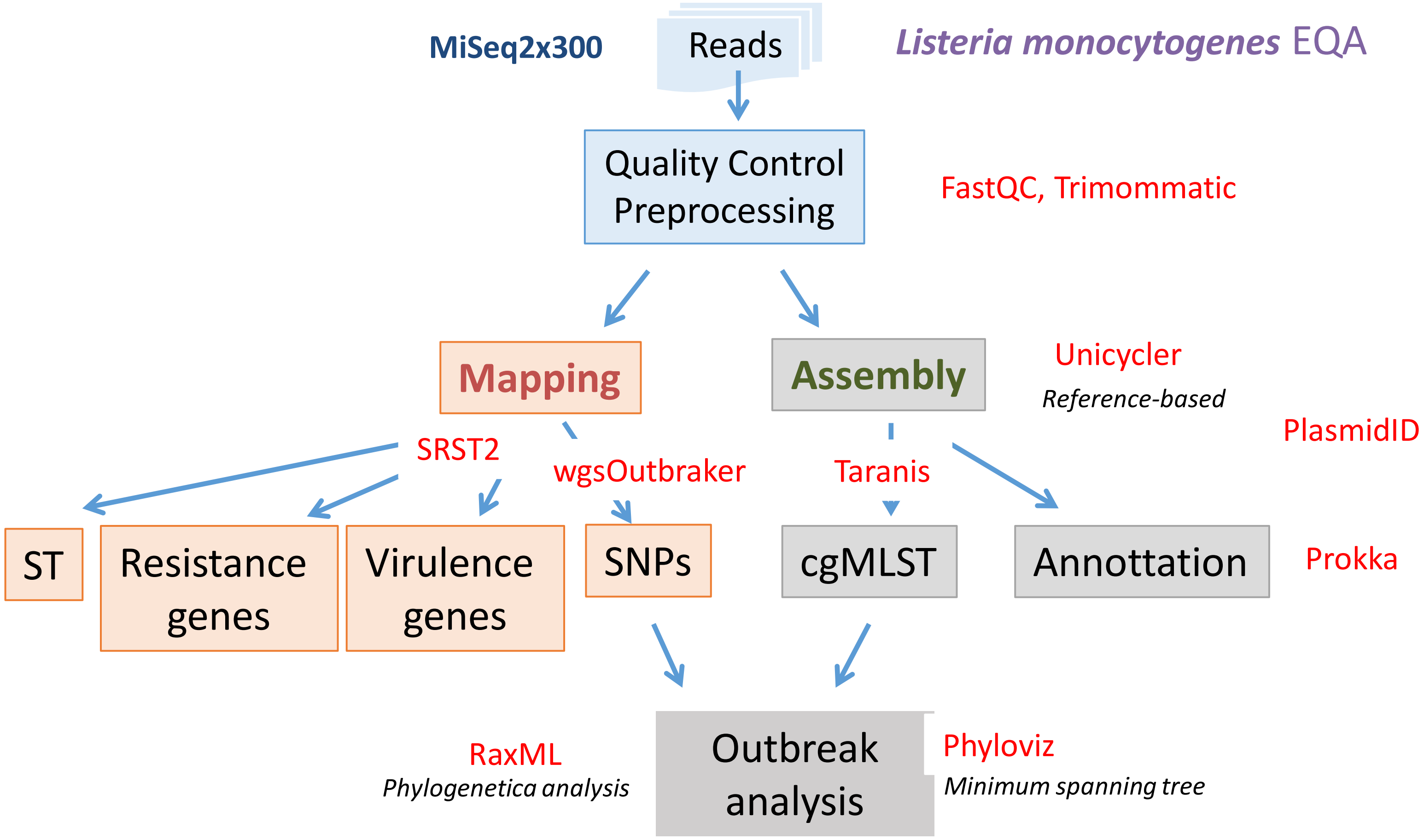
Index

- BU-ISCIII
- High-throughput sequencing (HTS) applications in Microbiology. Examples
- Concepts: HTS and Outbreak investigation
- Bacterial and Viral Genome Sequencing
- **Bioinformatics analysis in microbial genomics**
- Viralrecon: SARS-CoV-2 genome reconstruction software

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
  - HAPLOYD GENOME
  - LOW FREQUENCY VARIANT – QUASISPECIES
  - SNPs MATRIX – PHYLOGENETIC ANALYSIS
- STRUCTURAL AND FUNCTIONAL ANNOTATION
  - RESISTOME, VIRULOME, SEQUENCE-TYPE

Workflow example



MiSeq2x300

Reads

Listeria monocytogenes EQA

Quality Control
Preprocessing

FastQC, Trimommatic

Mapping

Assembly

Unicycler
Reference-based

PlasmidID

SRST2

wgsOutbraker

Taranis

ST

Resistance genes

Virulence genes

SNPs

cgMLST

Annottation

Prokka

RaxML
Phylogenetica analysis

Outbreak analysis

Phyloviz
Minimum spanning tree

Software disponible – VARIANT CALLING

- CFSAN SNP Pipeline

Extracción de SNPs de alta calidad de aislados relacionados

http://snppipeline.readthedocs.io/en/latest/

- GATK, modo haploide
- Samtools
- Varscan
- Snippy

Identificación de variantes haploides y construcción de filogenia usando core genome SNPs

http://github.com/tseemann/snippy

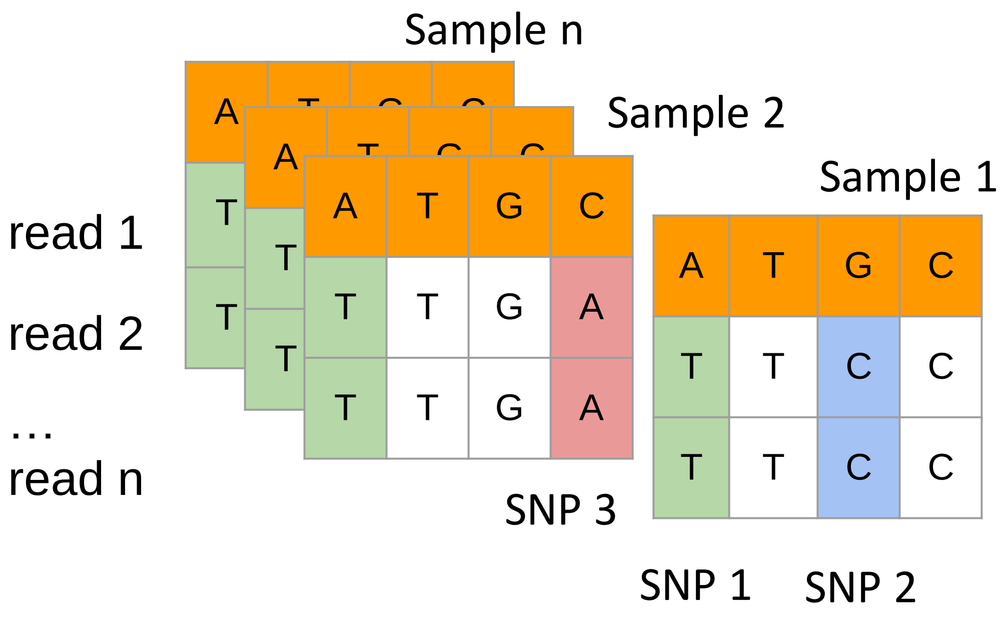
- Live-SET

High-quality SNPs para crear filogenia para investigación de brotes

https://github.com/lskatz/lyve-SET

- WGS-Outbraker

Generación de matriz de SNPs – BACTERIA –OUTBREAK ANALYSIS

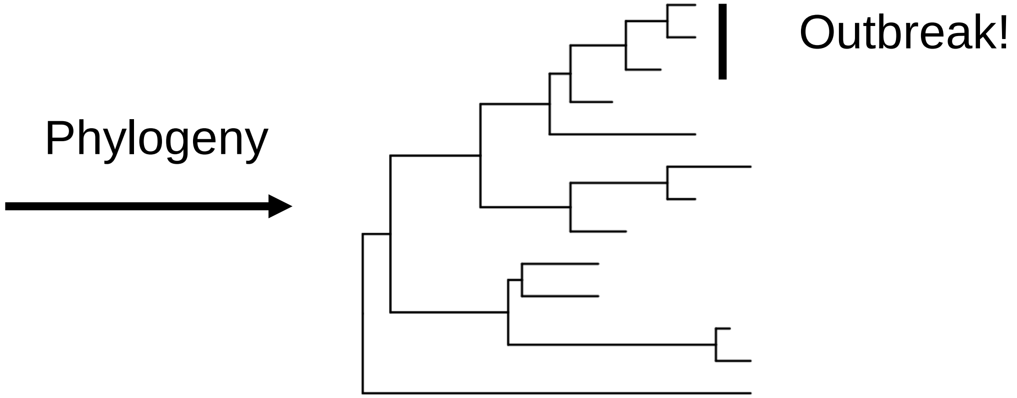


SNP 1	SNP 2	SNP 3	
A	G	C	Reference
T	C	C	Sample 1
T	G	A	Sample 2
			Sample n

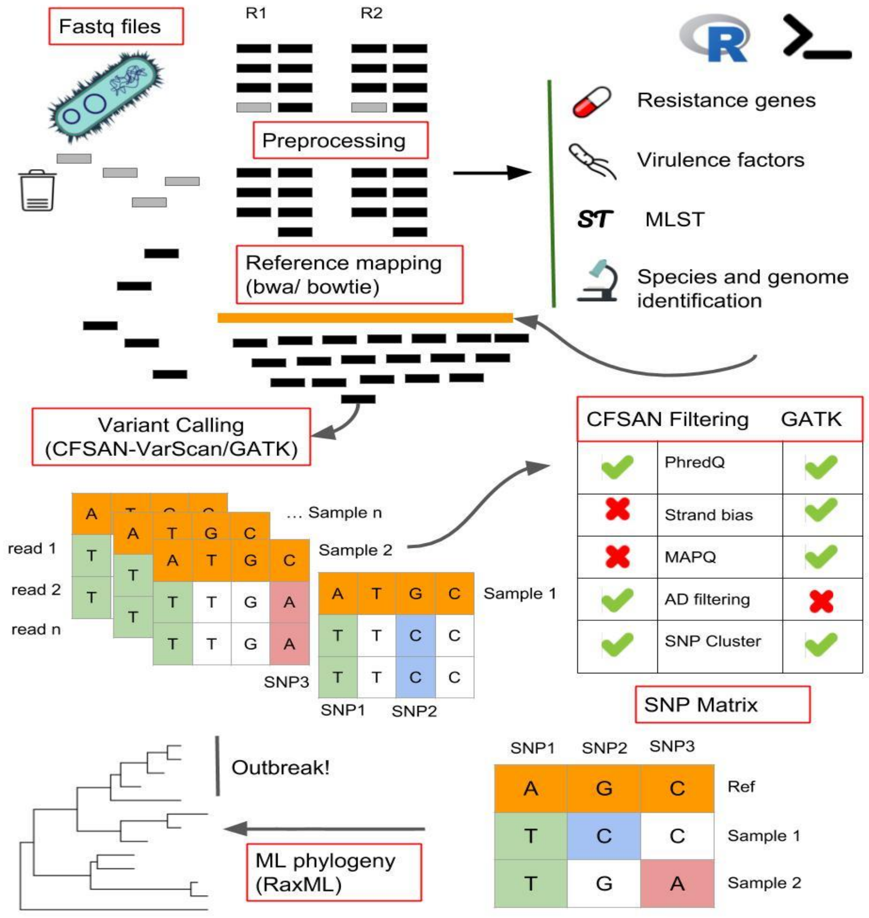
Generación de matriz de SNPs – BACTERIA –OUTBREAK ANALYSIS

SNP matrix

SNP 1	SNP 2	SNP 3	
A	G	C	Reference
T	C	C	Sample 1
T	G	A	Sample 2
			Sample n



WGS-Outbreaker https://github.com/BU-ISCIII/WGS-Outbreaker



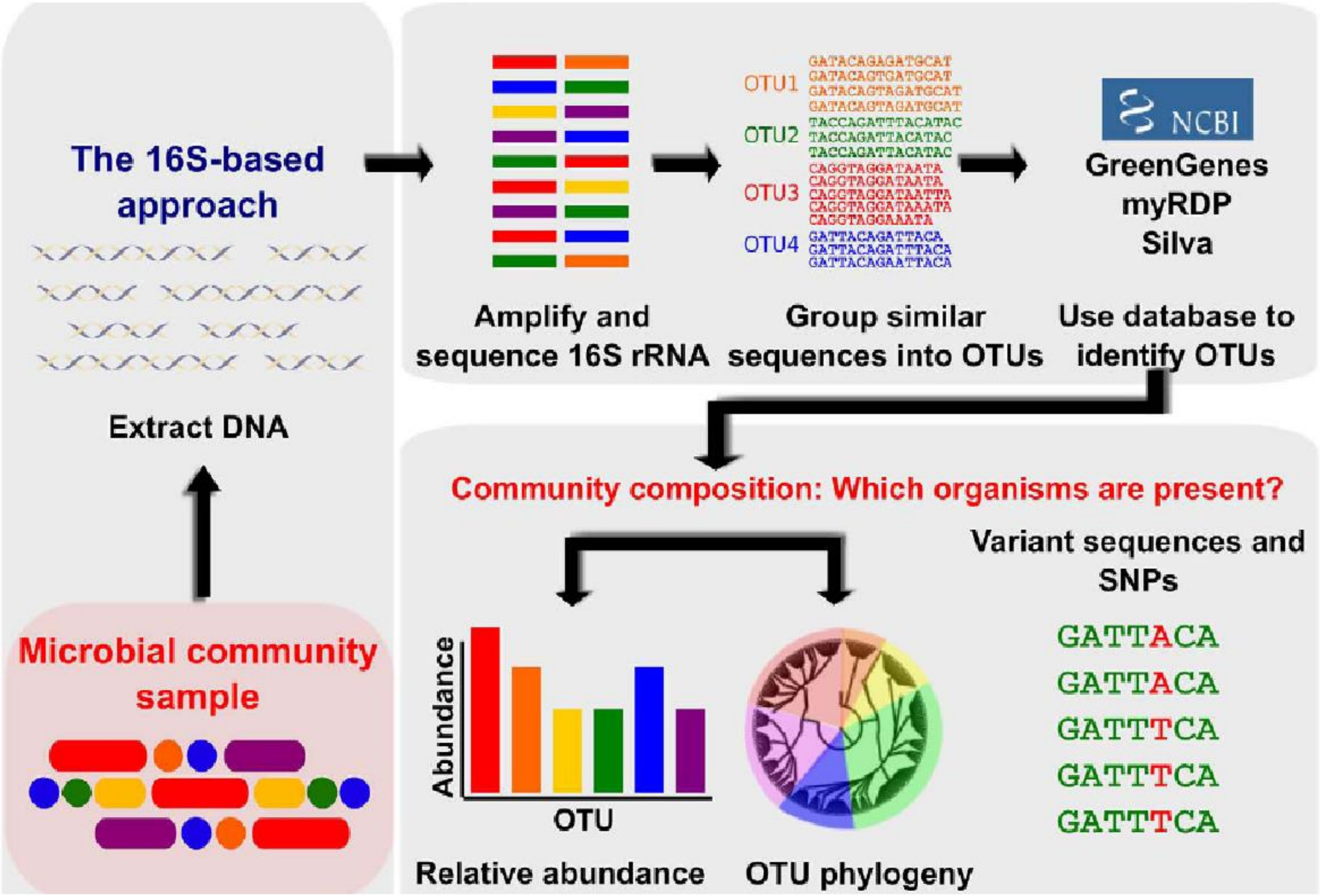
CFSAN Filtering		GATK
✓	PhredQ	✓
✗	Strand bias	✓
✗	MAPQ	✓
✓	AD filtering	✗
✓	SNP Cluster	✓

SNP1	SNP2	SNP3	
A	G	C	Ref
T	C	C	Sample 1
T	G	A	Sample 2

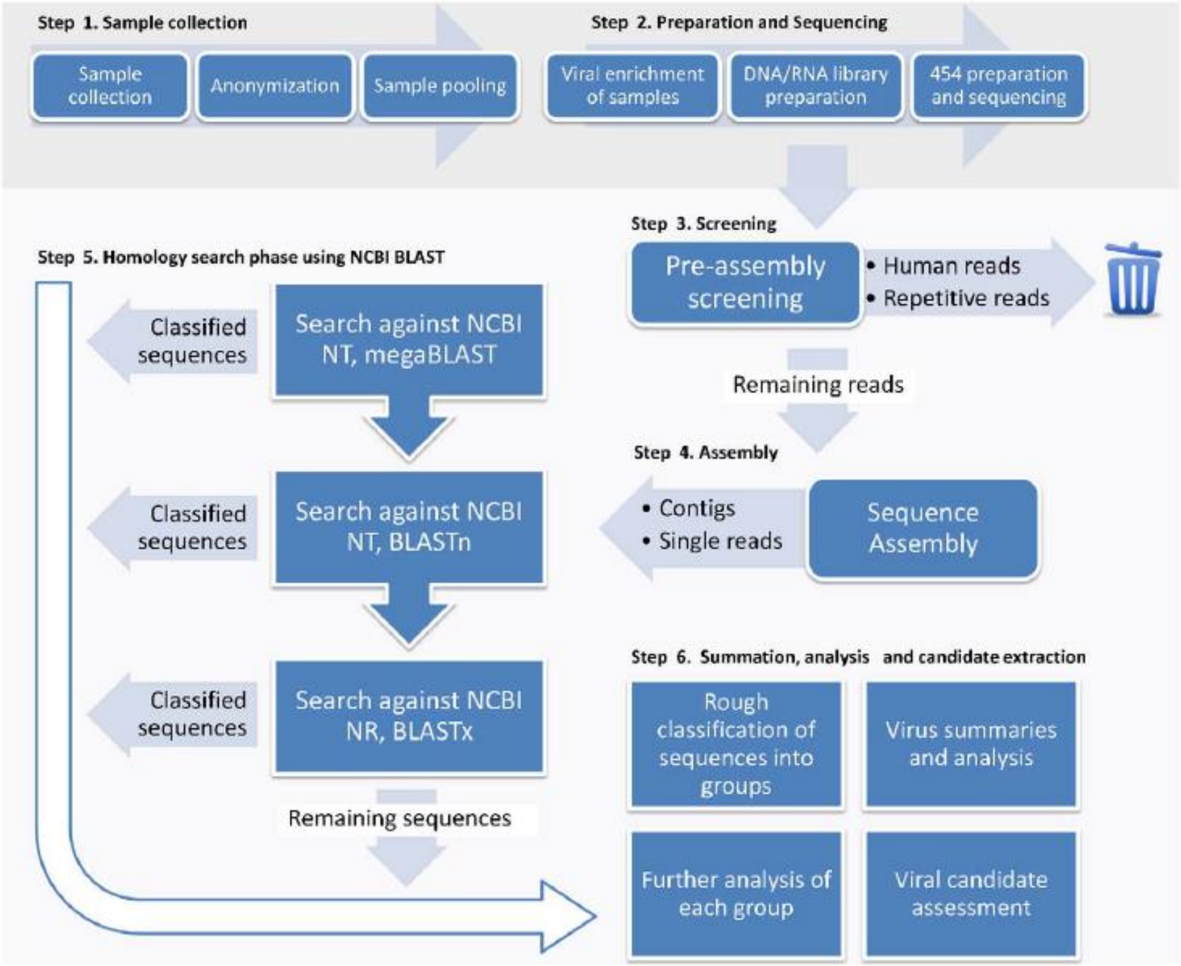
Metataxonomics vs Metagenomics (16S vs Shotgun)

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

Metataxonomics – Target Metagenomics



Metagenomics


Step 1. Sample collection
Sample collection
Anonymization
Sample pooling
Step 2. Preparation and Sequencing
Viral enrichment of samples
DNA/RNA library preparation
454 preparation and sequencing
Step 3. Screening
Pre-assembly screening
• Human reads
• Repetitive reads
Remaining reads
Step 4. Assembly
• Contigs
• Single reads
Sequence Assembly
Step 5. Homology search phase using NCBI BLAST
Classified sequences
Search against NCBI NT, megaBLAST
Classified sequences
Search against NCBI NT, BLASTn
Classified sequences
Search against NCBI NR, BLASTx
Remaining sequences
Step 6. Summation, analysis and candidate extraction
Rough classification of sequences into groups
Virus summaries and analysis
Further analysis of each group
Viral candidate assessment

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

Index

- BU-ISCIII
- High-throughput sequencing (HTS) applications in Microbiology. Examples
- Concepts: HTS and Outbreak investigation
- Bacterial and Viral Genome Sequencing
- Bioinformatics analysis in microbial genomics
- **Viralrecon: SARS-CoV-2 genome reconstruction software**