Secuenciación de Genomas Bacterianos: Herramientas y Aplicaciones

BU-ISCIII

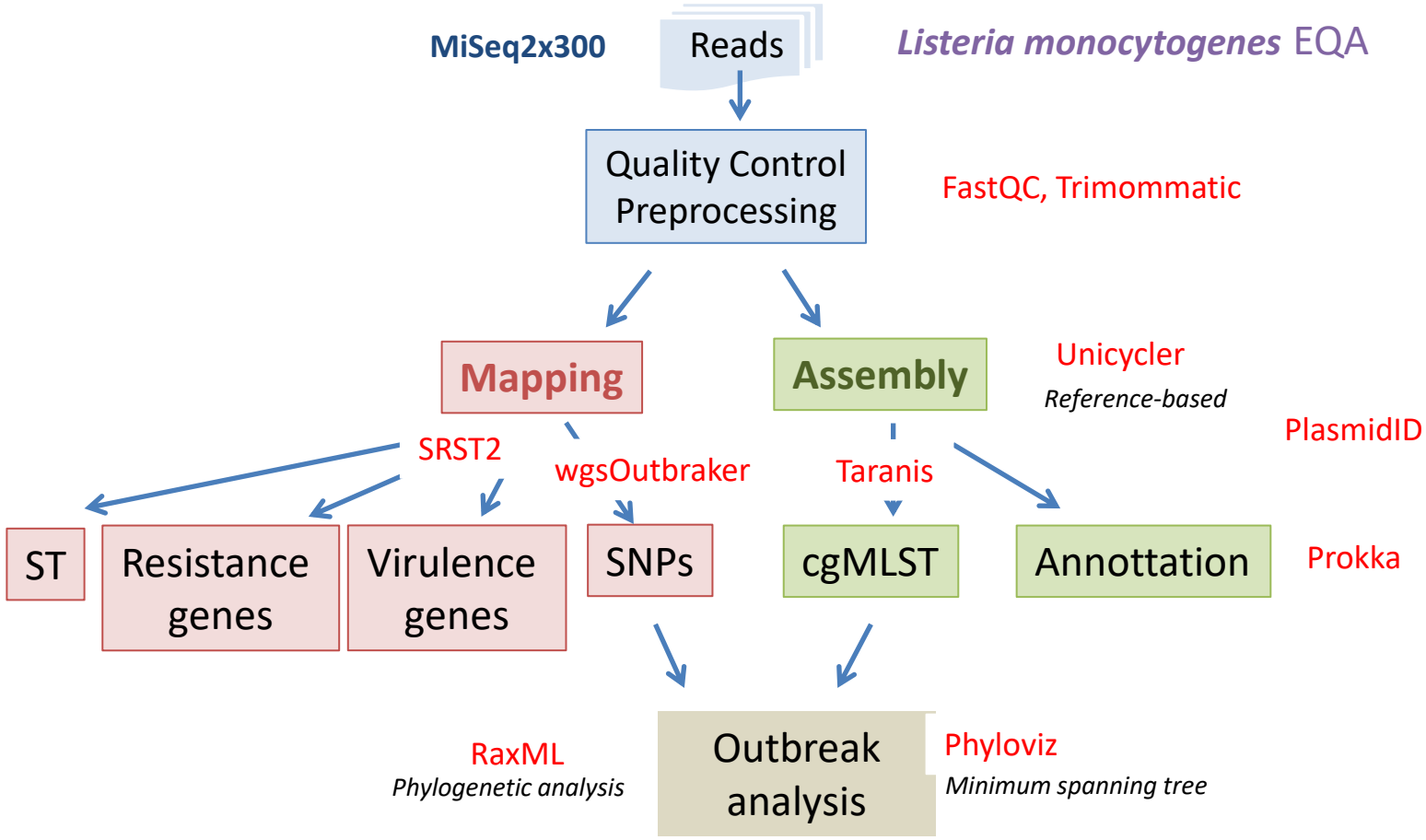
Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

28 al 31 octubre 2023, 6ª Edición
Programa Formación Continua, ISCIII

Learning aims and outcomes

- Understand some principles behind NGS and its applications to whole genome sequencing.
- Know the format files generated in NGS data analysis and the workflow analysis.
- Understand the uses of WGS in: specie, antimicrobial resistance genes and virulence factor genes identification, and for typing.
- Outbreak characterization based on SNPs or gene by gene approaches.

Training workflow



Teachers

- **Sara Monzón Fernández,** Biotecnóloga y Bioinformática (Analista de datos). Titulado Superior Especialista OPIS. Responsable técnico BU-ISCIII
- **Sarai Varona Fernández,** Bioquímica y Bioinformática (Analista de Datos). Titulado Superior Especialista OPIS (2025).
- **Emilia Arjona Bolaños**, Biologa, Bioinformática (Project Manager).
- Contrato Titulado Superior asociado a proyecto (2023-2026).

Session 1.1 – Secuenciación masiva de genomas bacterianos: situación actual

Emilia Arjona Bolaños

BU-ISCIII

Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

16 al 20 octubre 2023, 5ª Edición
Programa Formación Continua, ISCIII

Index

- BU-ISCIII
- High throughput sequencing platforms update
- Bacterial genome sequencing, brief history
- Advantages of WGS
- Use of WGS in Europe
- Library strategies
- Bioinformatics analysis

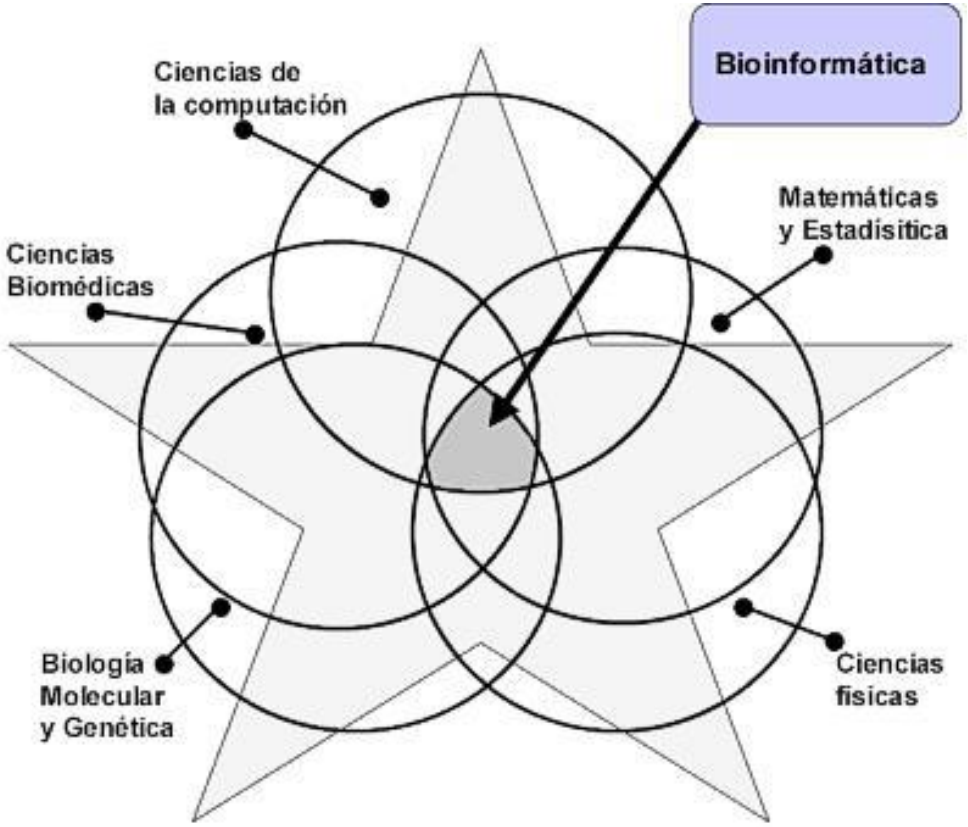
Qué es la Bioinformática?

PROBLEMAS BIOLÓGICOS

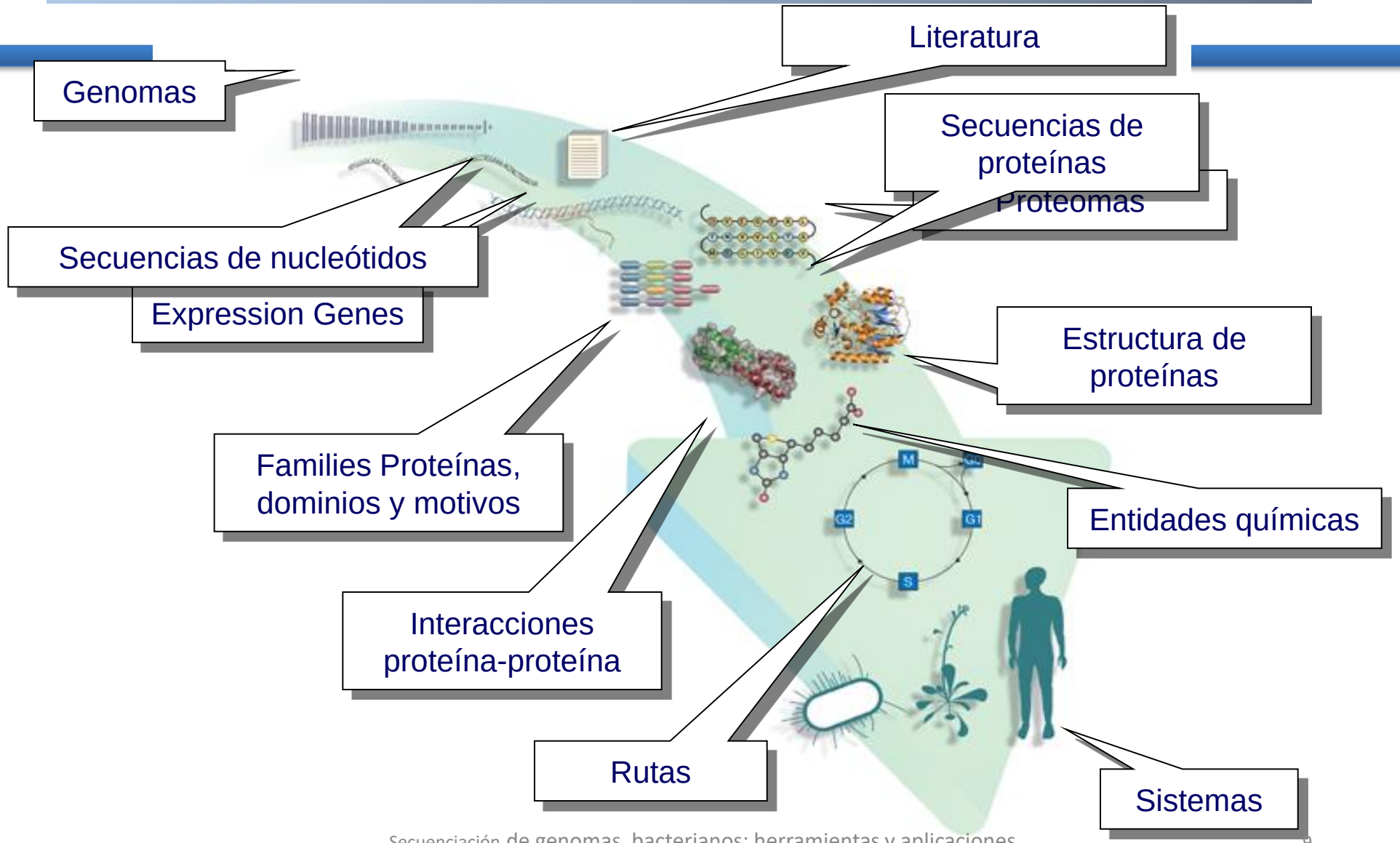
Procesamiento de datos

MÉTODOS COMPUTACIONALES

Bioinformática es multidisciplinar



Tipos de datos dan idea de la dimensión de la Bioinformática



Por qué nace BU-ISCIII?

Genomics Unit

2010 — 454 (Roche)

2013 — NextSeq500, MiSeq

2019 — MinIon - Nanopore

2021 — NovaSeq 6000

Bioinformatics Unit

2012

BU-ISCIII

Scientific & technical Support to Researchers on HTS Data Analysis

National Microbiology Centre (CNM)

Research Institute for Rare Diseases (IIER)

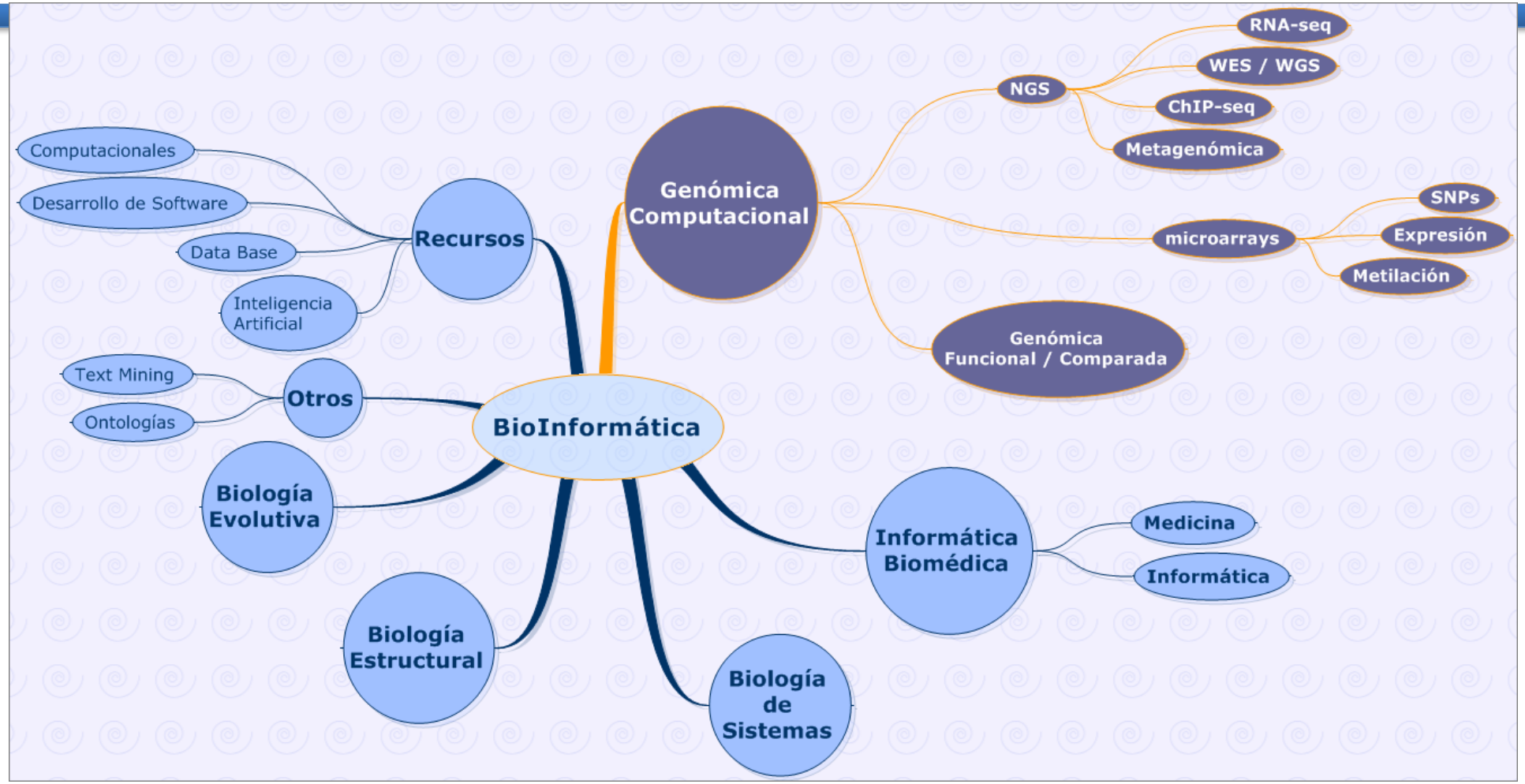
Functional Unit for Research in Chronic Disease

Network of Biological Alerts

National Centre of Tropical Medicine

National Environment Health Centre

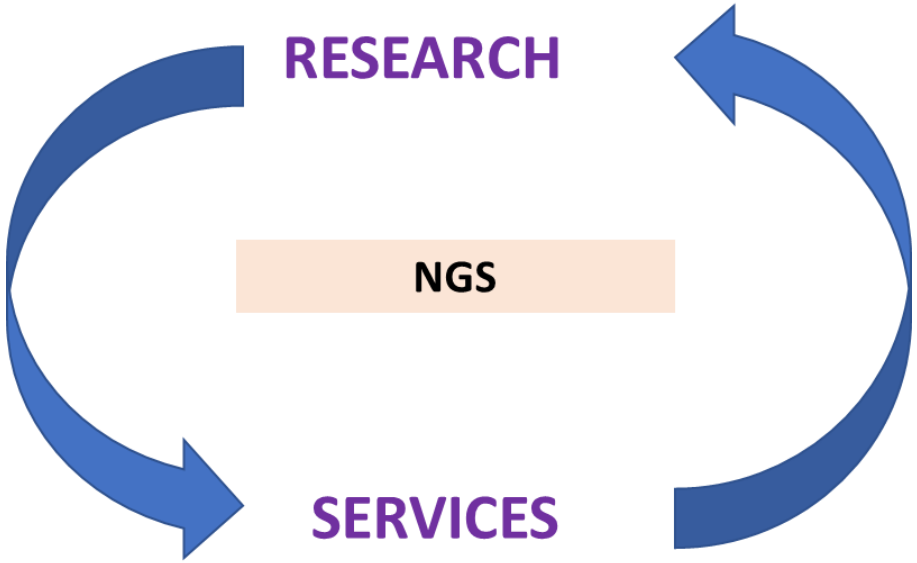
BU-ISCIII Mission - Activities



Bioinformatics Unit Activities

- Identify biological problems (PI / Groups) that could be target of NGS
- Early adopters: establish collaboration with.
- Be strategic providing transversal solutions → reusable tools

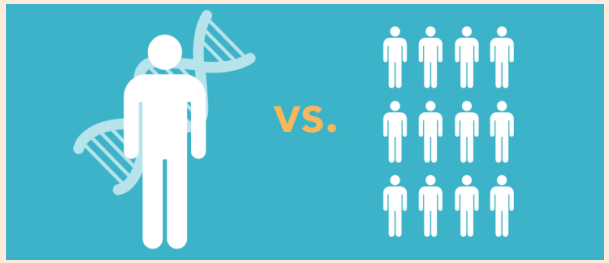
Provide scientific and technical solutions for using NGS in the diagnostic routine or research activity from different ISCIII labs

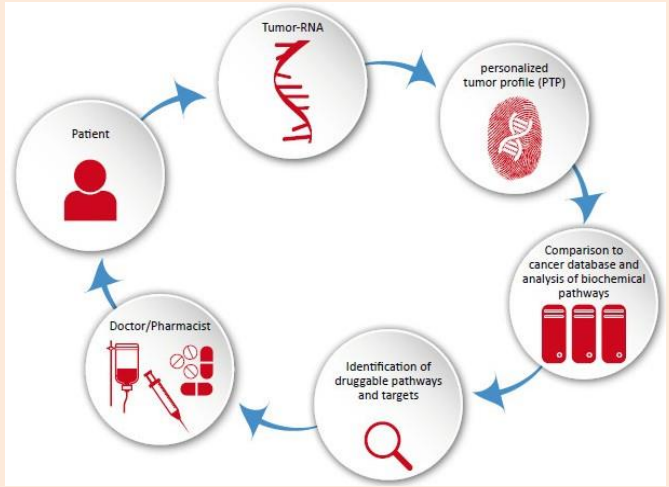


Clinical Bioinformatics - Precision Medicine

IIER

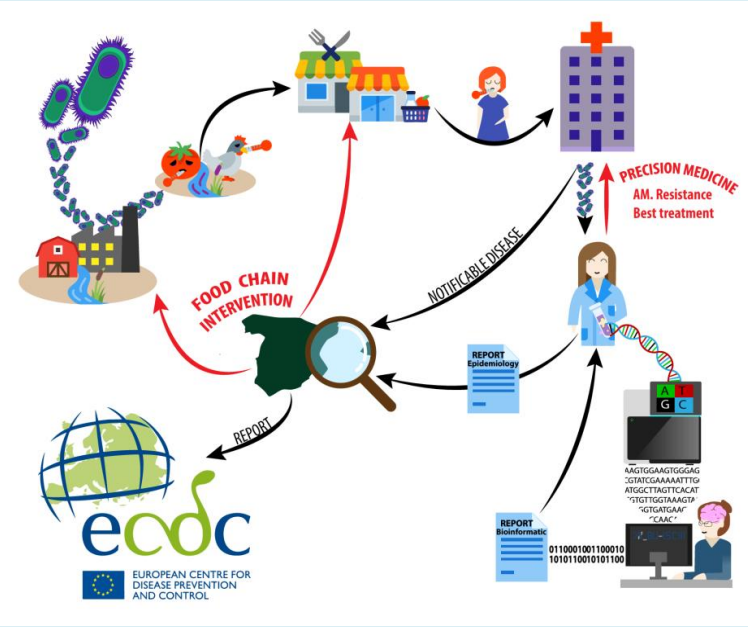
Tumour or Rare Diseases



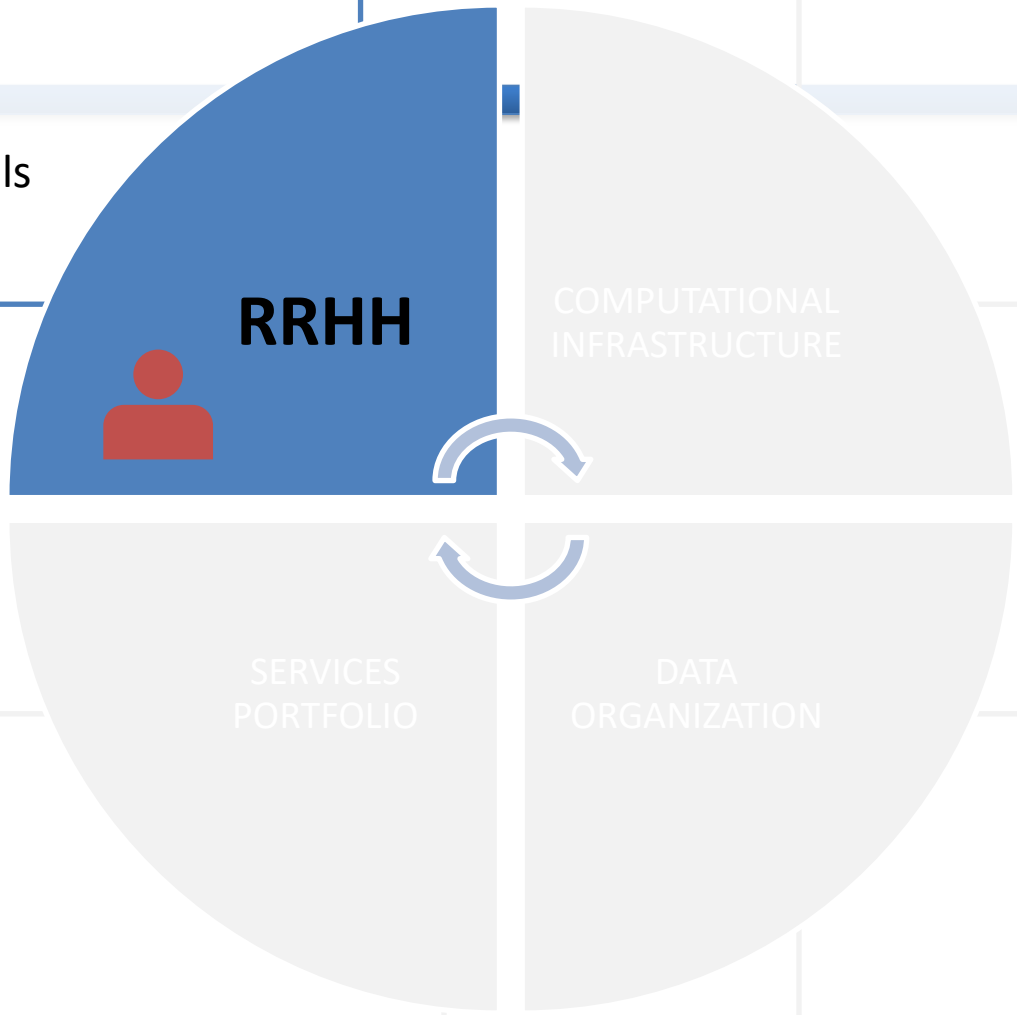


CNM

Pathogen associated



- TEAM: Bioinformatician & IT
- SOURCE OF FUNDING:
  - Research Project
  - National or International Calls
  - Permanent position – Civil Servant



Human resources

Nombre	Disciplina	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Isabel Cuesta	Dr. Biología Molecular	CIENTIFICO TITULAR OPIS												
Sara Monzón	Biotecnología	TFM	CIBERER		PTA MINECO				T.SUP.Esp OPIS					
Bruno Lobo	Administrador Sistemas		PROYECTO		PTA MINECO		SISTEMAS							
Sarai Varona	Bioquímica								PROYECTO (2024)					T.SUP. Esp OPIS
Luis Chapado	Telecomunicaciones					COLABORADOR								
Pablo Mata	Farmacia												PROYECTO	
Jaime Ozaez	Biología												PROYECTO	
Daniel Valle	Biología												PROYECTO	
Alba Talavera	Biología												PROYECTO	
Juan Ledesma	Biología													CT
Victor López	Biología													GJ
Alex Bernabeu	Biotecnología												PROYECTO	
Emi Arjona	Biología												PROYECTO	
Personal (número)		2	4	4	5	5	4	6	6	6	8	8	8	10+4

Master Bioinformática

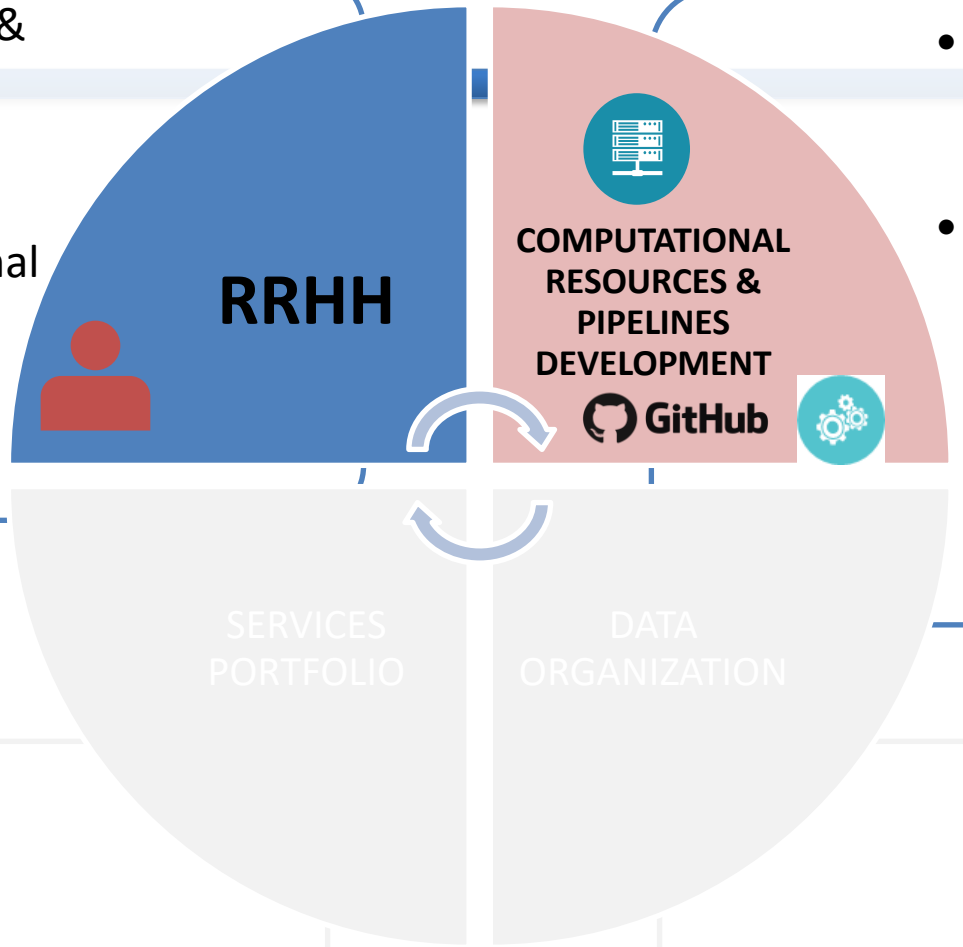
CNM

IIER

BU-ISCIII

FUNCIONARIO

- TEAM: Bioinformatician & IT
- SOURCE OF FUNDING:
  - Research Project
  - National or International Calls
  - Stable position – Civil Servant



- INFRASTRUCTURE
  - HPC
  - Workstations
- REPOSITORY:
  - GitHub
  - Benchmarking
  - Validation

Computational Resources

- IT support: establish agreement with IT department including permission for using Linux.

Workstations (5), 4cores, 64Gb, 8TB
Server, 4-quad, 120Gb, 16TB

Data Centre (CPD-ISCIII)

HPC 320 cores, 8TB RAM, 10Gbps.
2 flexible and scalable storages, NetApp, 70 TB and 250TB

- Analysis pipelines reproducibility

nextflow

Singularity containers
Admin support & environment independency
Sharing code easier

GitHub

https://github.com/BU-ISCIII

Bioinformatic Analysis: Software validation - ECDC EQAs

Table 5. Results of allele-based cluster analysis

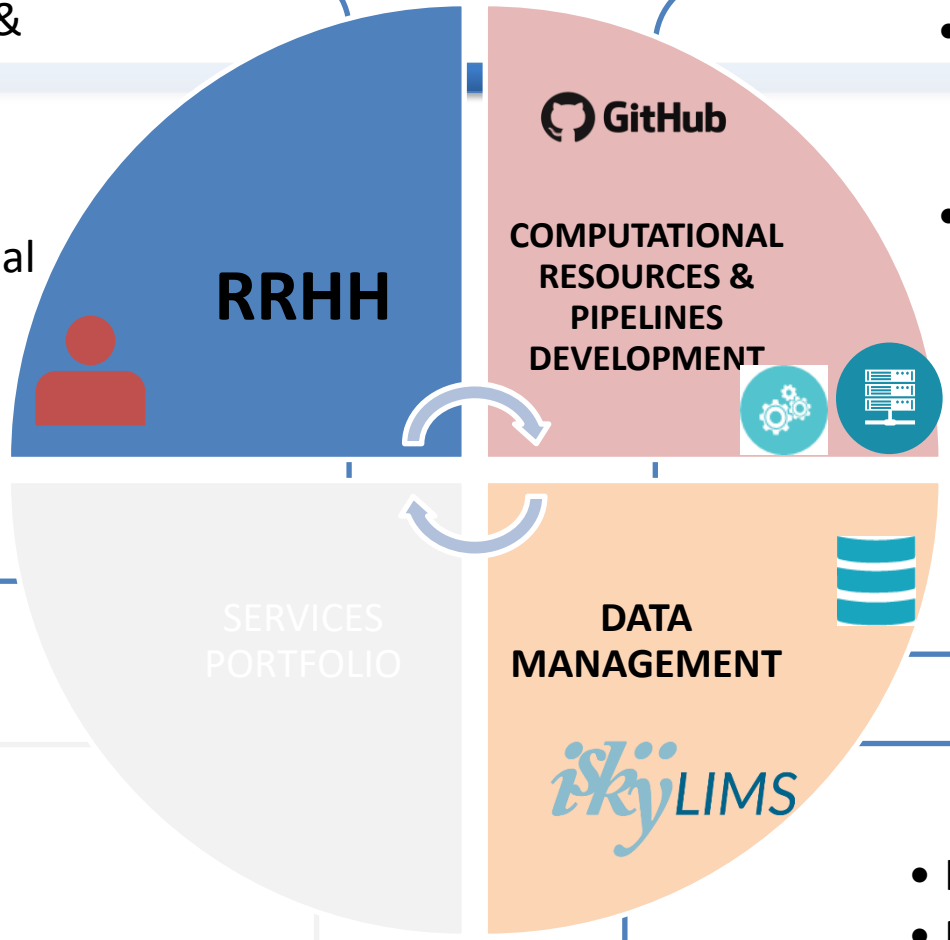
Lab ID	Allele based analysis					
	Approach	Allelic calling method	Assembler	Scheme	Difference within cluster	Difference outside cluster
EQA provider	BioNumerics	Assembly- and mapping-based	SPAdes	Applied Math (cgMLST/Pasteur)	0-3	24-1112
19	BioNumerics	Assembly- and mapping-based	SPAdes	Applied Math (cgMLST/Pasteur)	0-3	25-1120
35	SeqPhere	Assembly-based only	Velvet	Ruppitsch (cgMLST)	0-2	16-1065
70	SeqPhere	Assembly-based only	Velvet	Ruppitsch (cgMLST)	0-2	16-1062
105*	SeqPhere	Assembly-based only	SPAdes v 3.80	Ruppitsch (cgMLST)	0-1[#]	23-812
129	SeqPhere	Assembly-based only	Velvet			
135	SeqPhere	Assembly-based only	CLC Genomic Workbench 1			
141	SeqPhere	Assembly-based only	SPAdes 3.9.			
142	Inhouse	Assembly-based only	SPAdes			
144	SeqPhere	Assembly-based only	Velvet			

Table 4. Results of SNP-based cluster analysis

Lab ID	SNP-based						
	Approach	Reference	Read mapper	Variant caller	Assembler	Distance within cluster	Distance outside cluster
Provider	Reference-based	ST6 (REF4)	BWA	GATK		0-3	38-71
19*	Reference-based	ST6 ID 2362	BWA	GATK		0-4	43-81
56	Assembly-based			ksnp3	SPAdes	0-57[#]	561-591 (6109)
105	Reference-based	ST6 J1817	Bowtie2	VARSCAN 2		0-2[#]	22-42 (1049)
108	Reference-based	In-house strain resp ST	CLC assembly cell v4.4.2	CLC assembly cell v4.4.2		0-2	37-72
142*	Reference-based	*Listeria* EGDe (cc9)	CLC Bio	CLC Bio		0-1219	1223-2814 (8138)
146	Reference-based	ST6 ref. CP006046 ST1 ref. F2365 ST213/ST382 no ref.	BWA	In-house		0-358	

Fifth external quality assessment scheme for Listeria monocytogenes typing

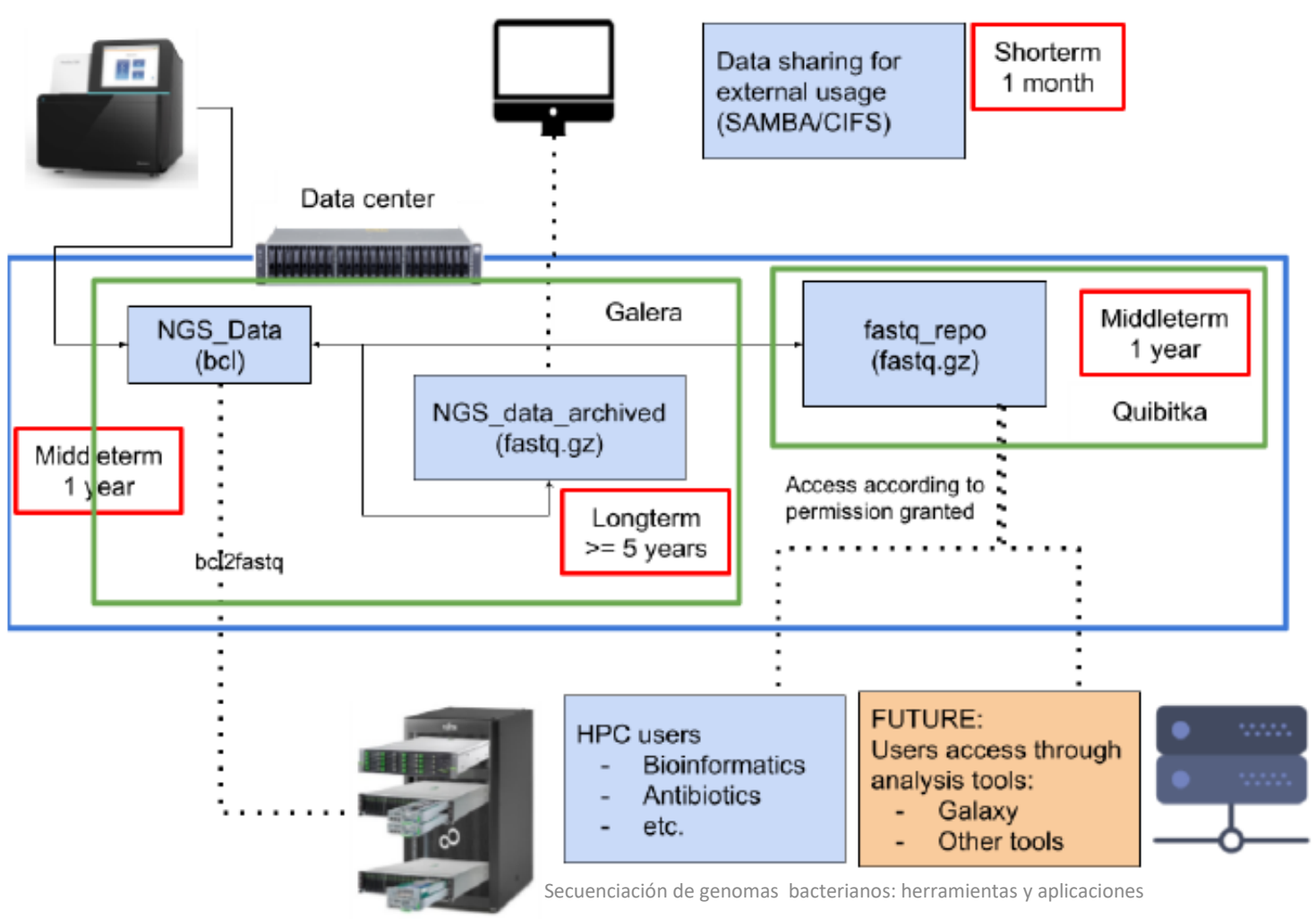
- TEAM: Bioinformatician & IT
- SOURCE OF FUNDING:
  - Research Project
  - National or International Calls
  - Permanent position – Civil Servant



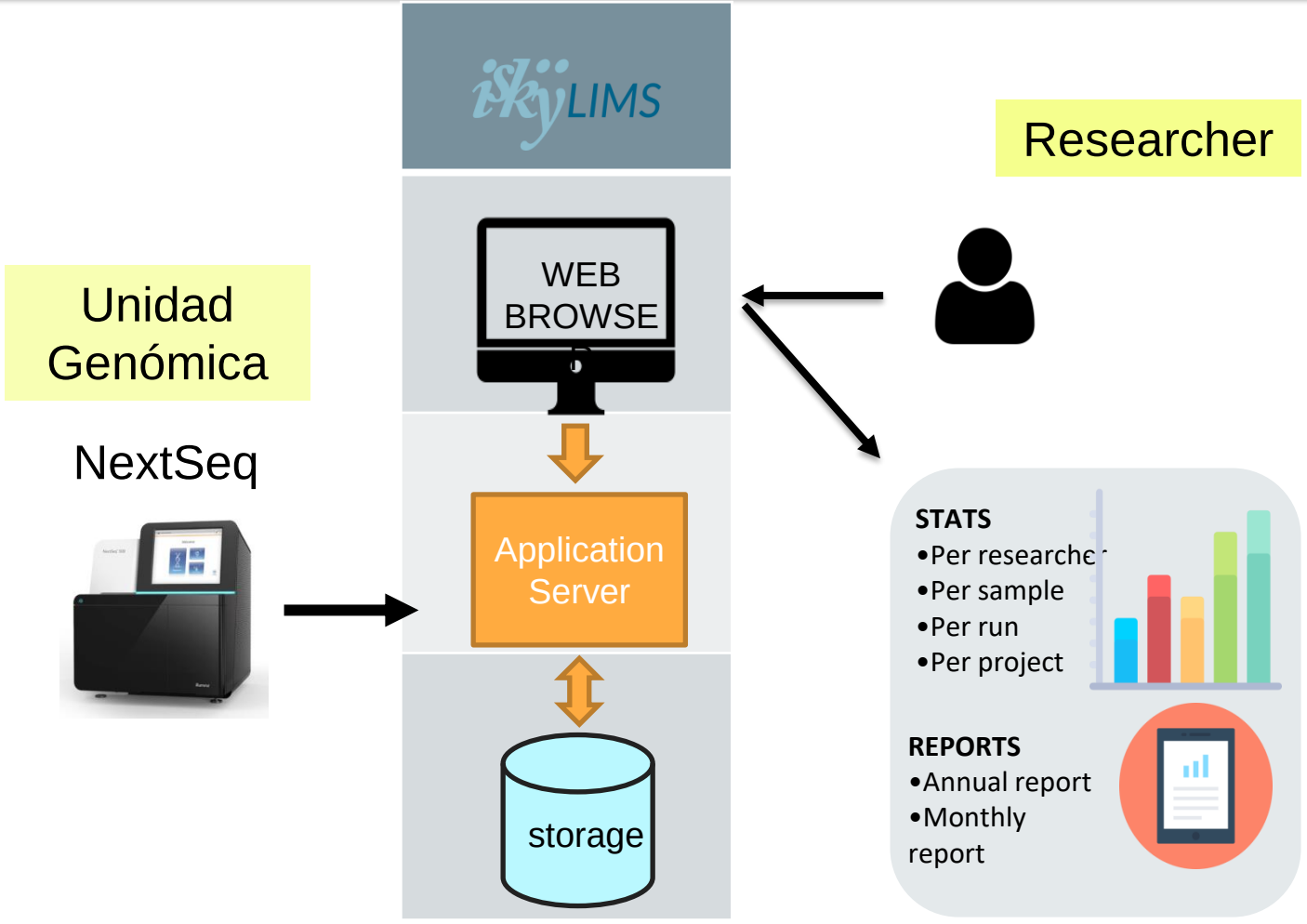
- INFRASTRUCTURE
  - HPC
  - Workstations
- REPOSITORY:
  - GitHub
  - Benchmarking
  - Validation

- INFRASTRUCTURE
- DATA WORKFLOW
- LIMS

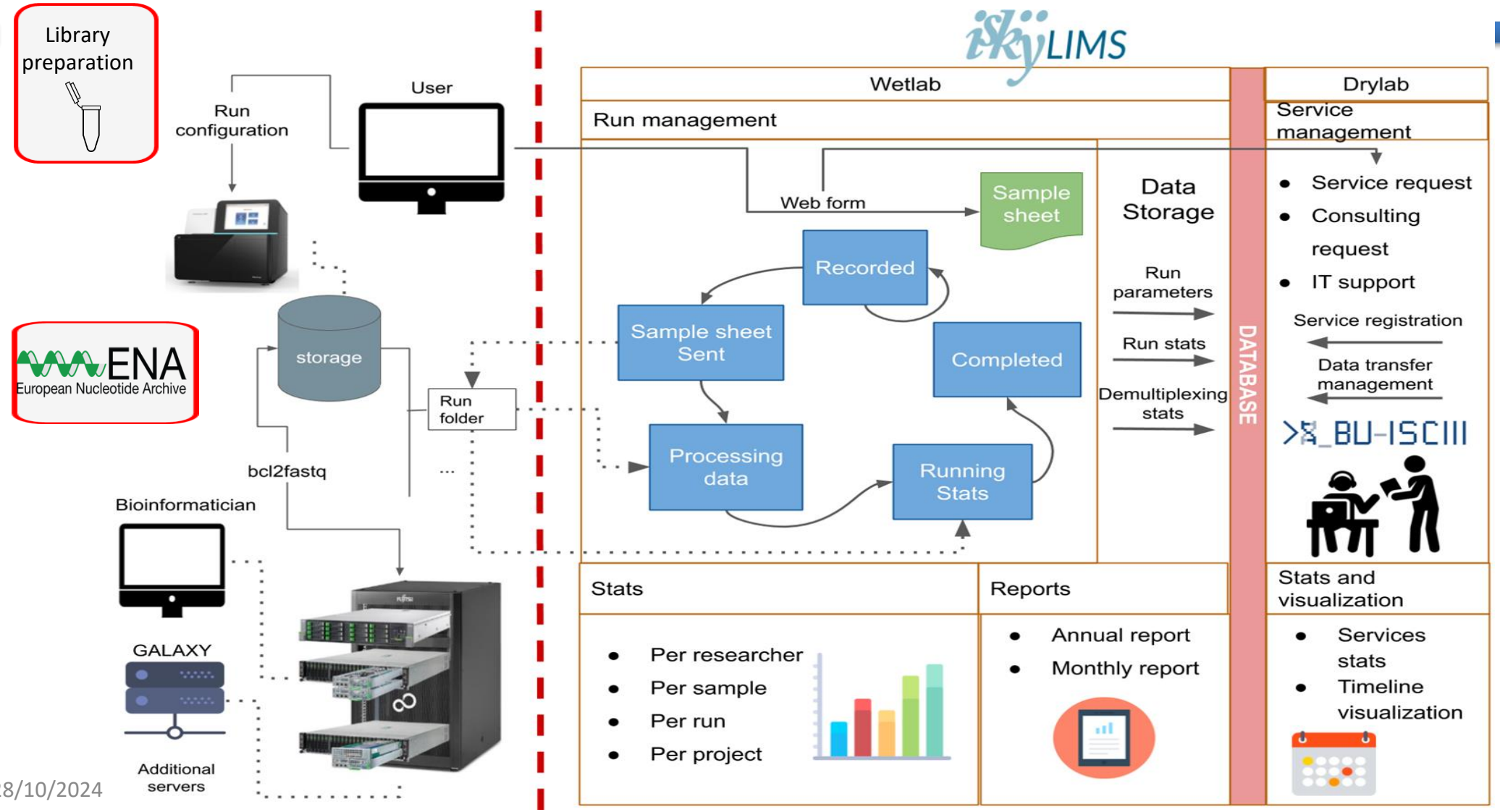
Infrastructure and data management



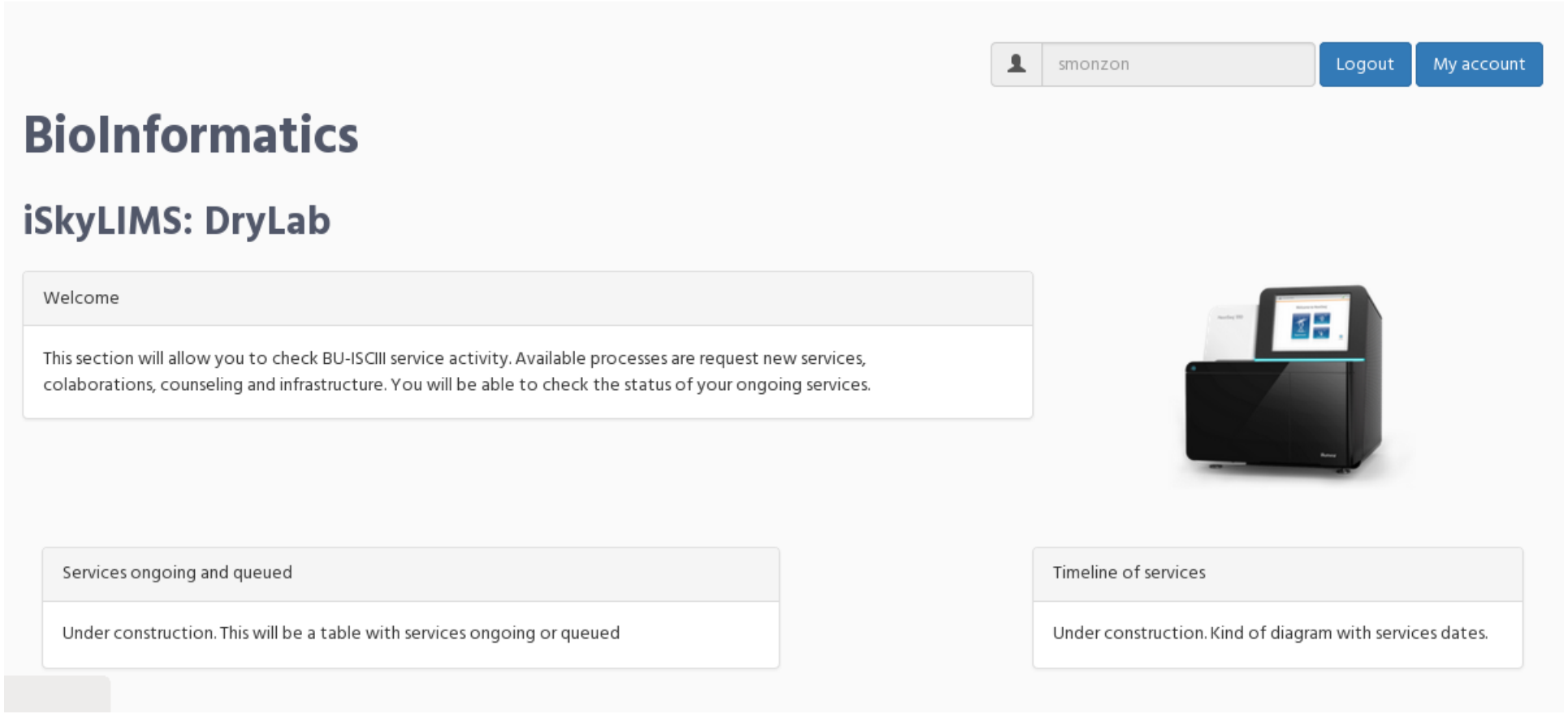
Infrastructure and data management - iSkyLIMS



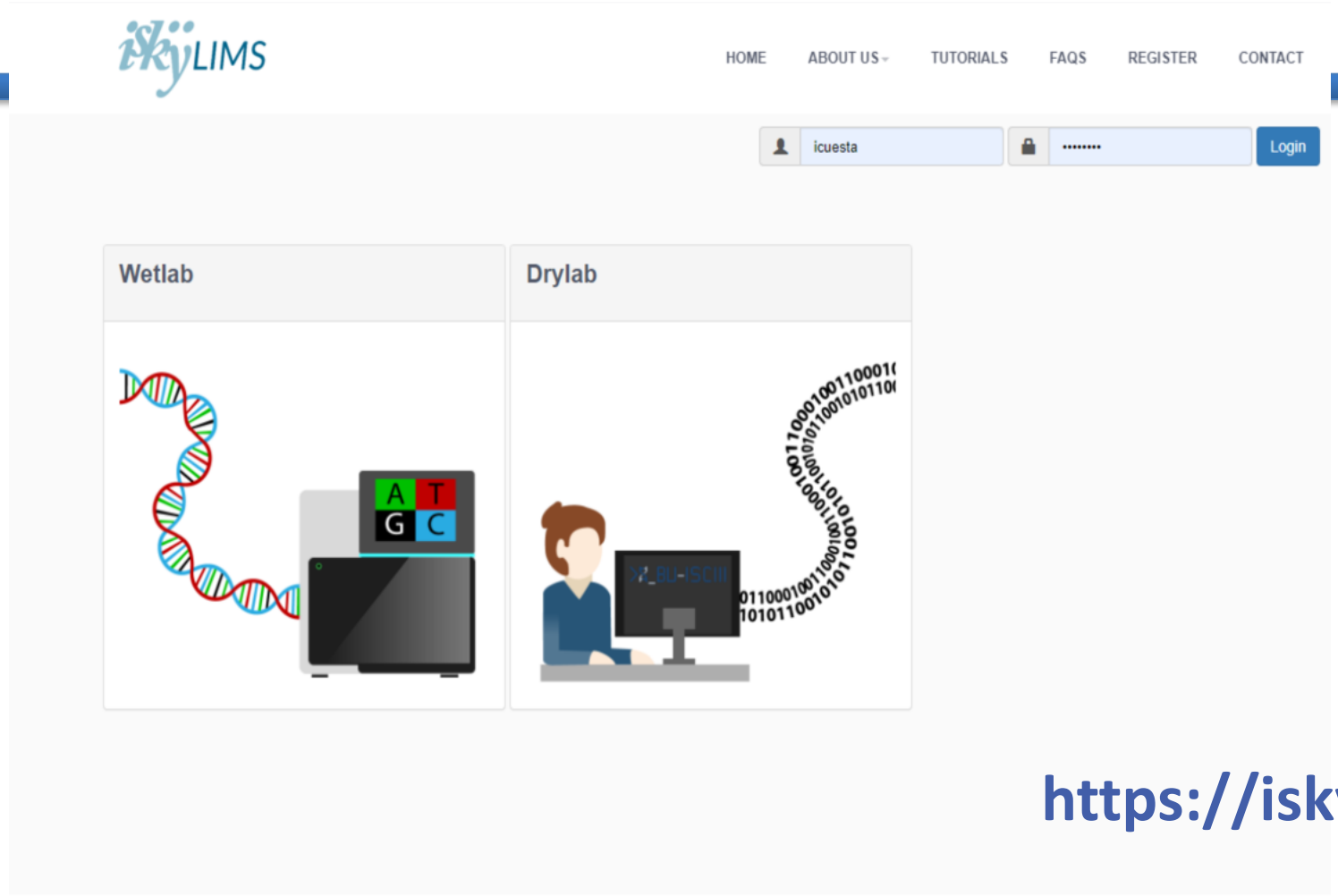
Infrastructure and data management: LIMS



Servicios de la Unidad

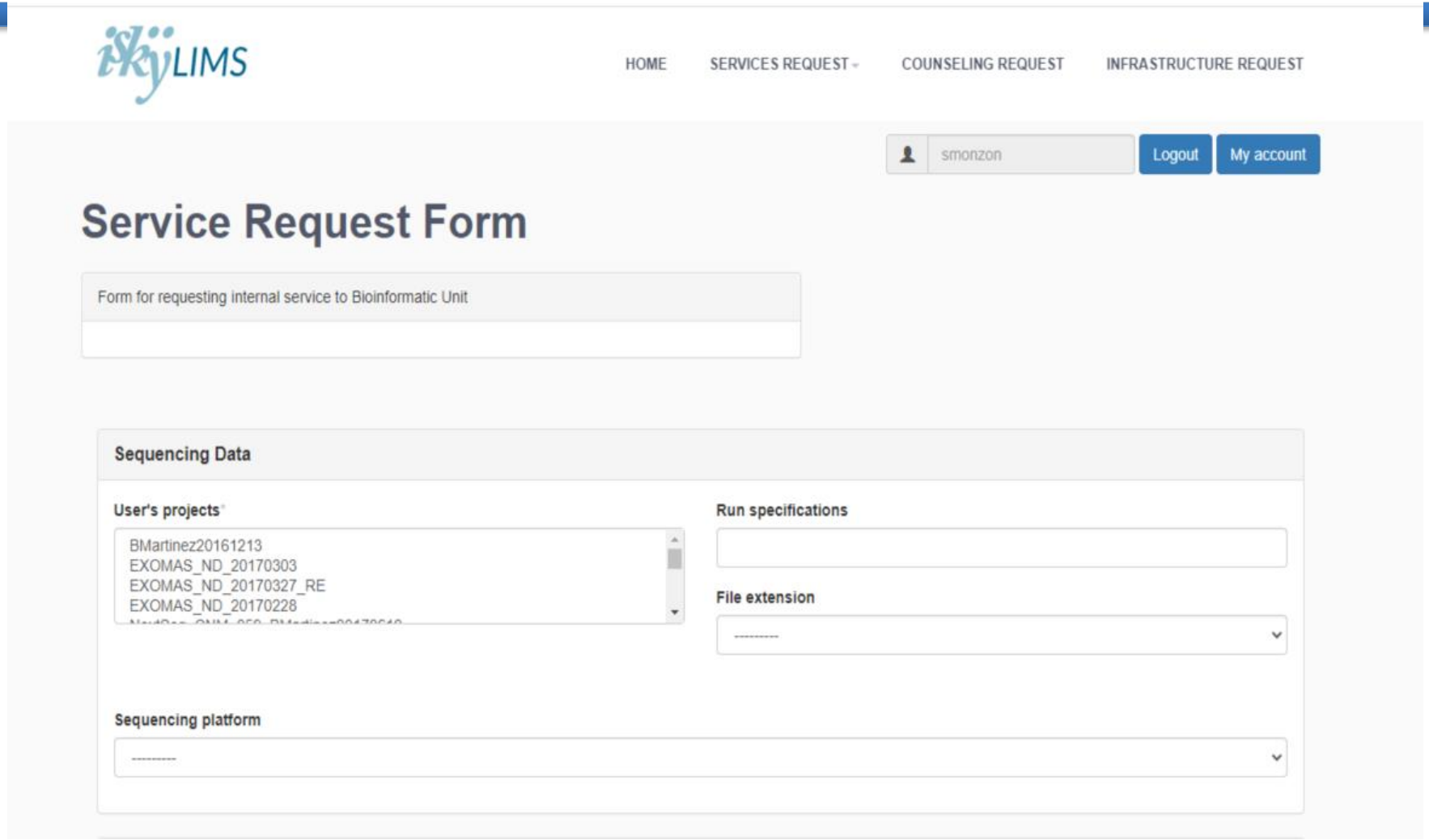


https://iskylims.isciii.es/

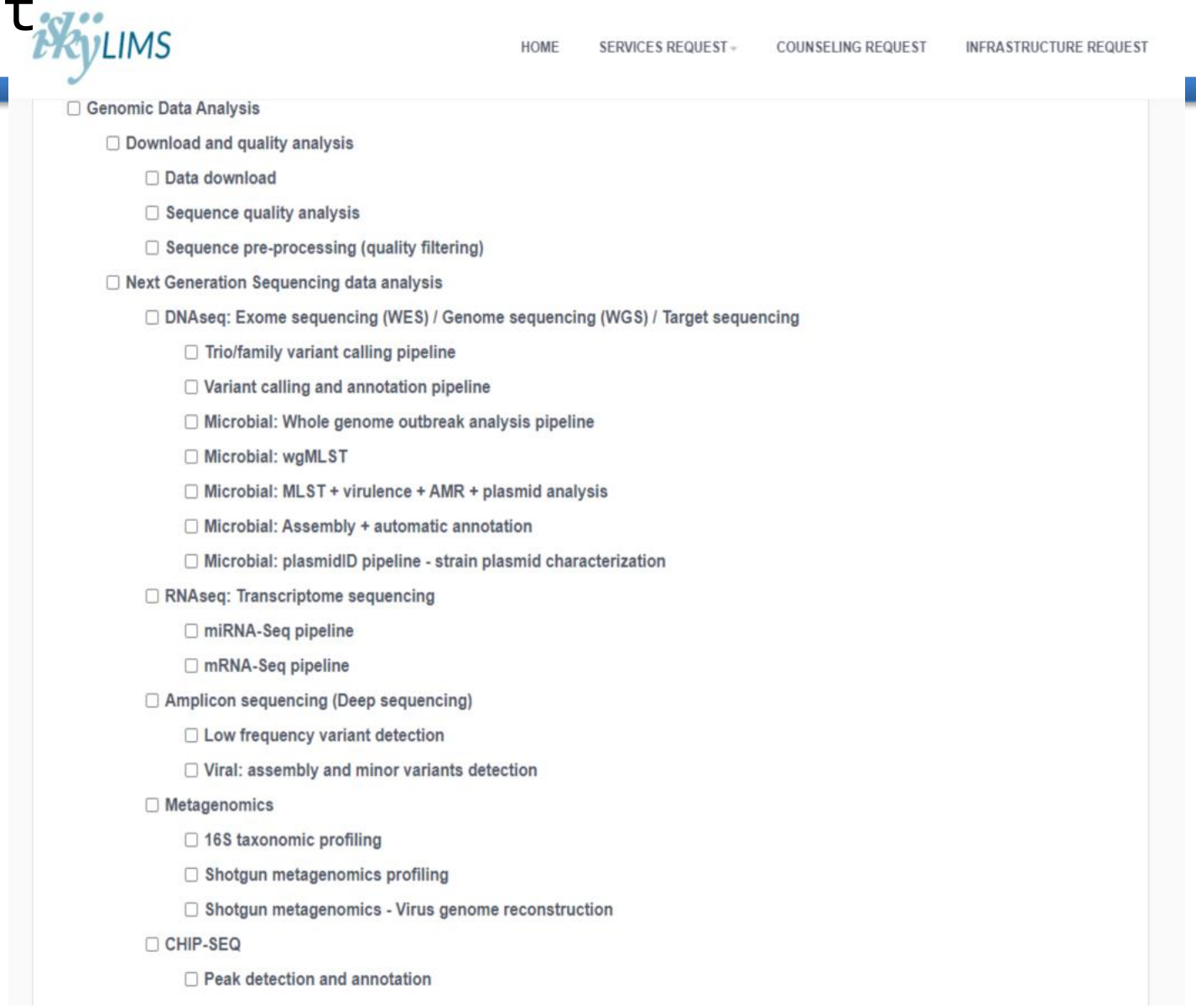


https://iskylims.isciii.es/

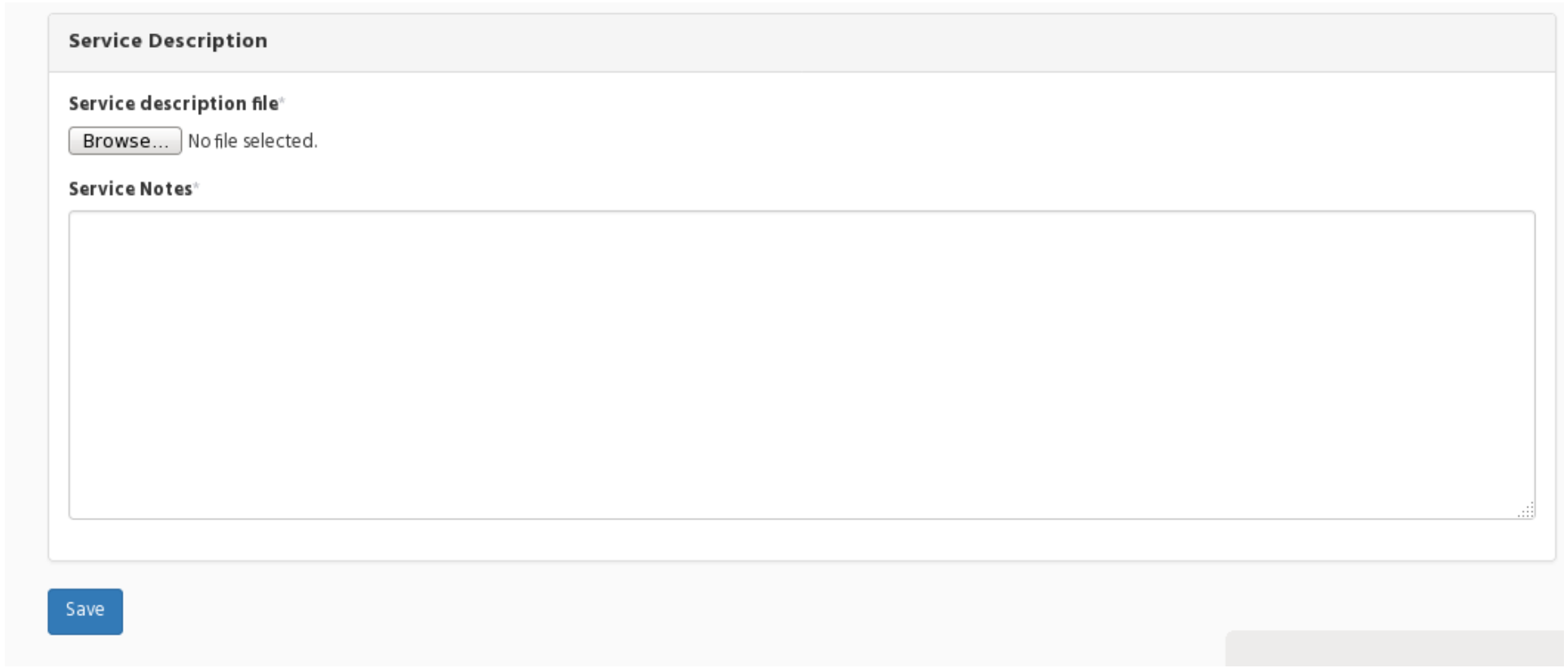
Service Request



Service Request



Service Request



Counseling Request

Service selection

AvailableServices *

- [ ] **Bioinformatics consulting and training**
  - [ ] **Bioinformatics analysis consulting**
  - [ ] **In-house and outer course organization**
  - [ ] **Student training in colaboration: Master thesis, research visit,...**

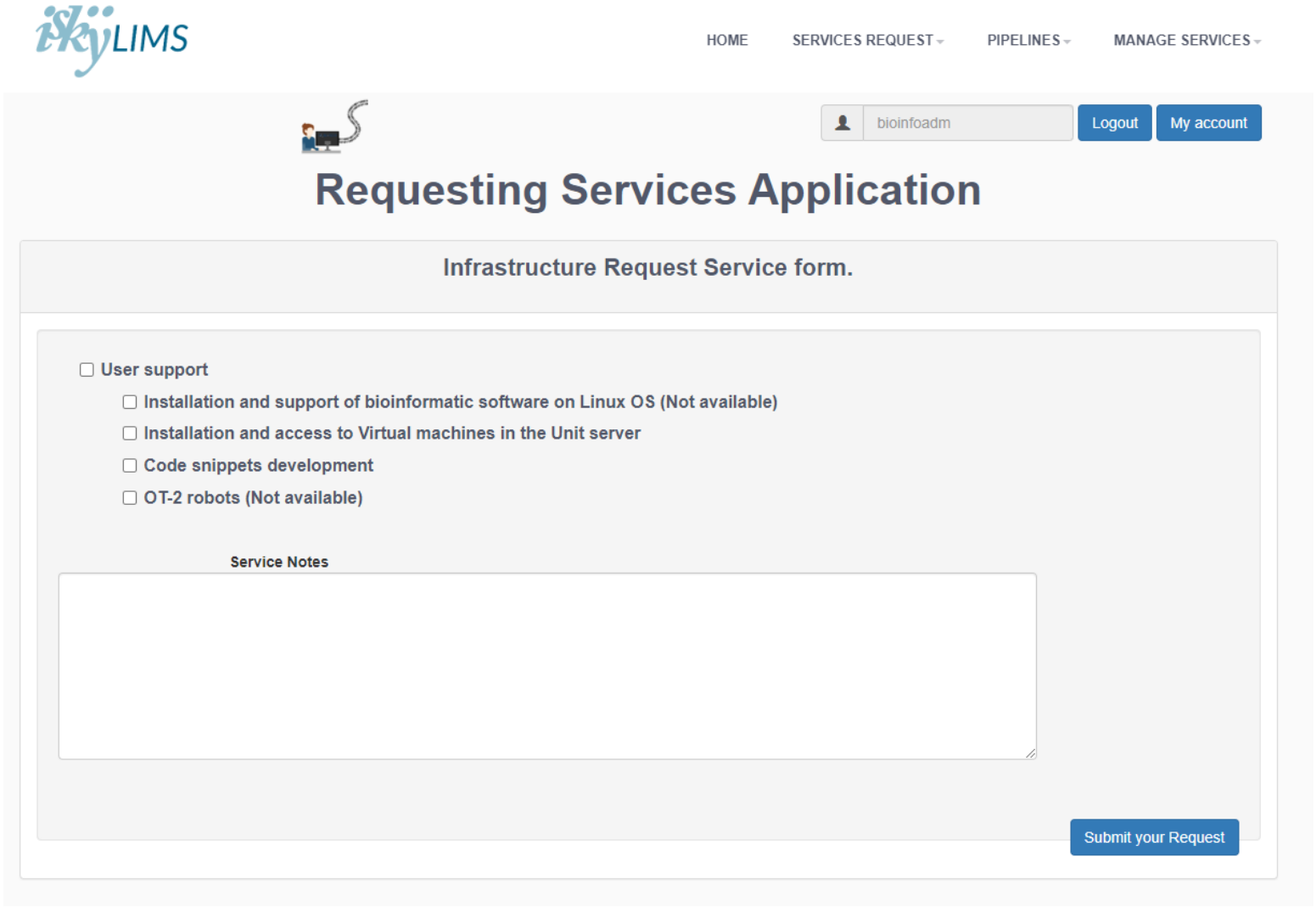
Service Description

Service description file*

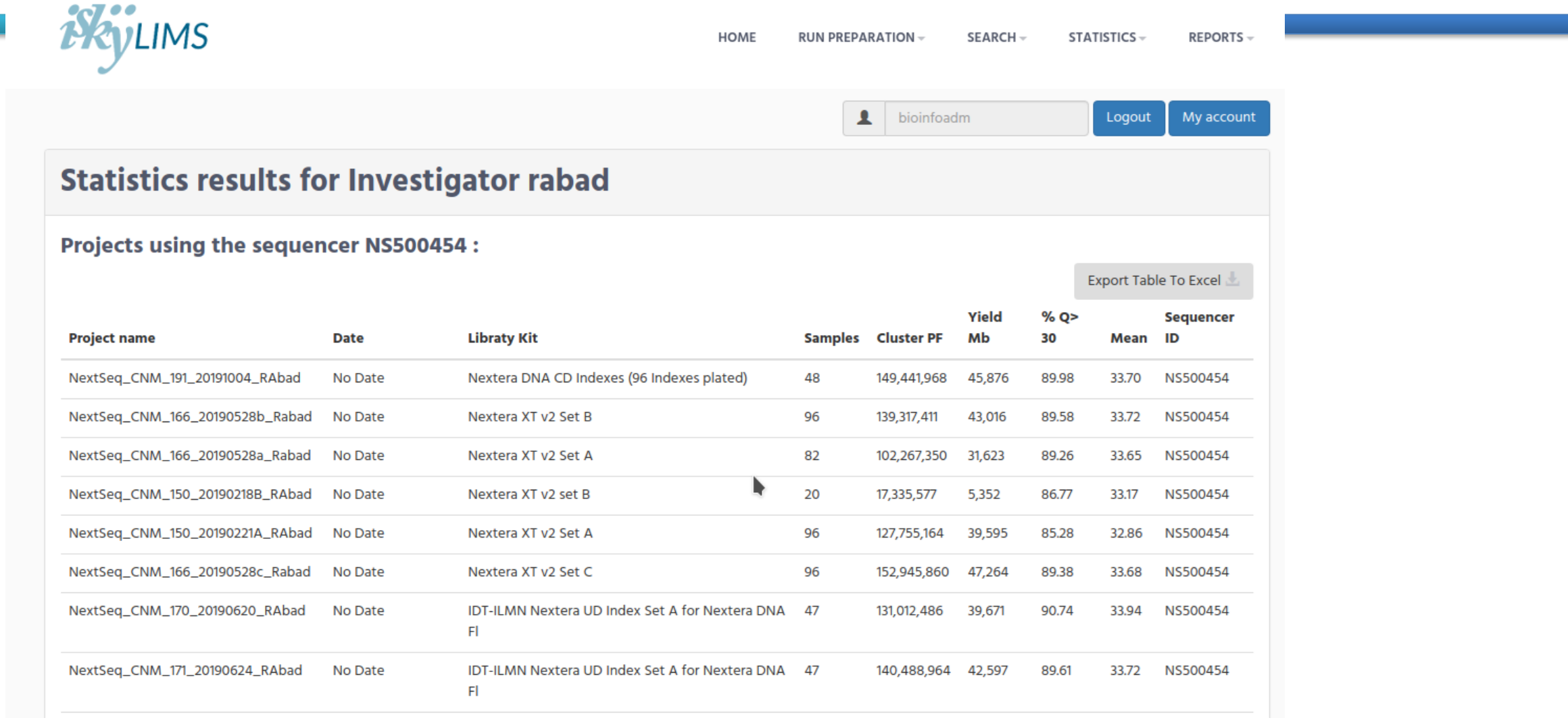
Browse... No file selected.

Service Notes*

Infrastructure Request

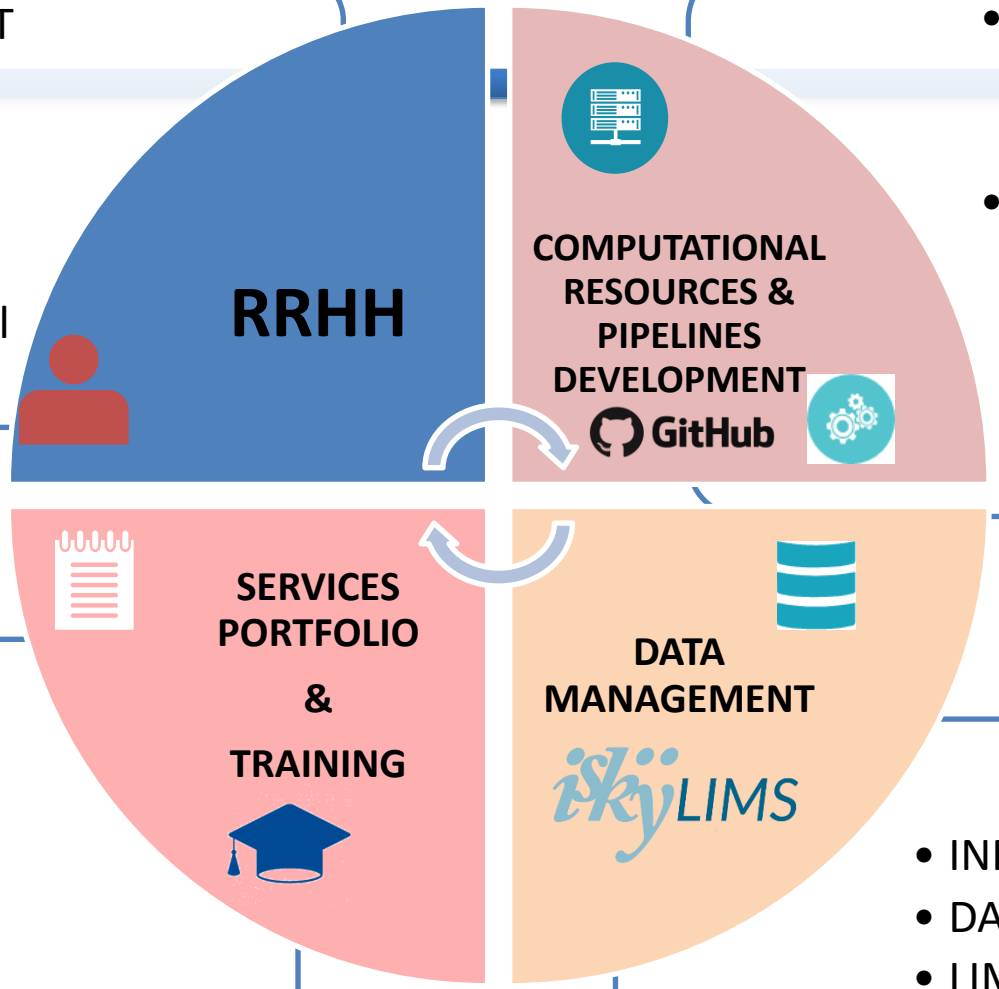


Infrastructure and data management



Project name	Date	Libraty Kit	Samples	Cluster PF	Yield Mb	% Q> 30	Mean	Sequencer ID
NextSeq_CNM_191_20191004_RAbad	No Date	Nextera DNA CD Indexes (96 Indexes plated)	48	149,441,968	45,876	89.98	33.70	NS500454
NextSeq_CNM_166_20190528b_Rabad	No Date	Nextera XT v2 Set B	96	139,317,411	43,016	89.58	33.72	NS500454
NextSeq_CNM_166_20190528a_Rabad	No Date	Nextera XT v2 Set A	82	102,267,350	31,623	89.26	33.65	NS500454
NextSeq_CNM_150_20190218B_RAbad	No Date	Nextera XT v2 set B	20	17,335,577	5,352	86.77	33.17	NS500454
NextSeq_CNM_150_20190221A_RAbad	No Date	Nextera XT v2 Set A	96	127,755,164	39,595	85.28	32.86	NS500454
NextSeq_CNM_166_20190528c_Rabad	No Date	Nextera XT v2 Set C	96	152,945,860	47,264	89.38	33.68	NS500454
NextSeq_CNM_170_20190620_RAbad	No Date	IDT-ILMN Nextera UD Index Set A for Nextera DNA Fl	47	131,012,486	39,671	90.74	33.94	NS500454
NextSeq_CNM_171_20190624_RAbad	No Date	IDT-ILMN Nextera UD Index Set A for Nextera DNA Fl	47	140,488,964	42,597	89.61	33.72	NS500454

- TEAM: Bioinformatician & IT
- SOURCE OF FUNDING:
  - Research Project
  - National or International Calls
  - Permanent position – Civil Servant



- INFRASTRUCTURE
  - HPC
  - Workstations
- REPOSITORY:
  - GitHub
  - Benchmarking
  - Validation

- COURSES, TFM, TFG
- DATA ANALYSIS
  - DNAseq
  - RNAseq
  - Metagenomics

- INFRASTRUCTURE
- DATA WORKFLOW
- LIMS

Cartera de Servicios

- **Genómica Computacional: Análisis de datos masivos**
  Técnicas de secuenciación masiva (NGS)

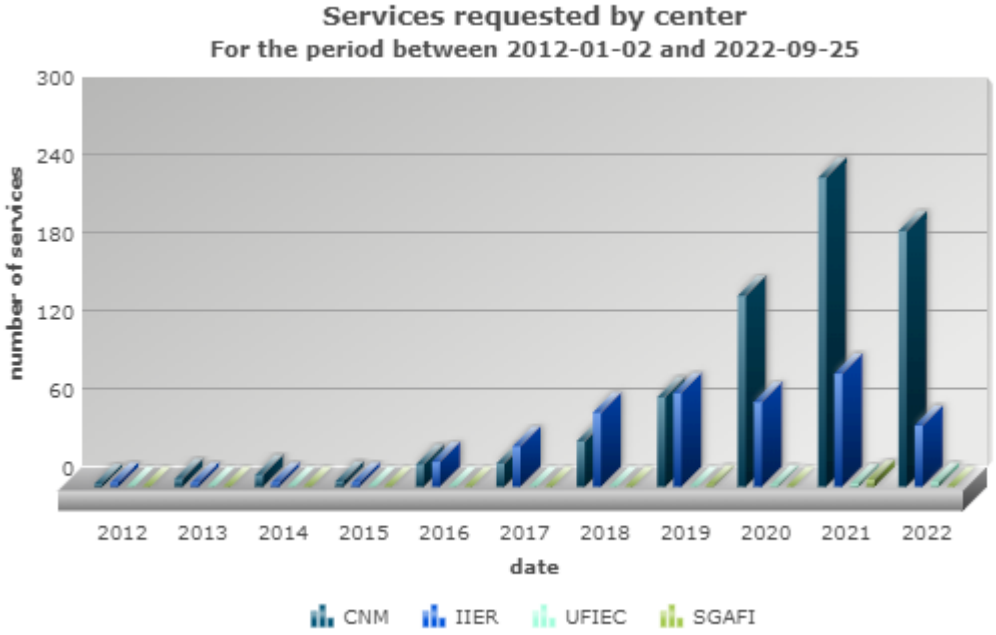
- **Asesoría y Formación en Bioinformática**
  Orientación en el análisis bioinformático
  Organización de cursos internos y externos

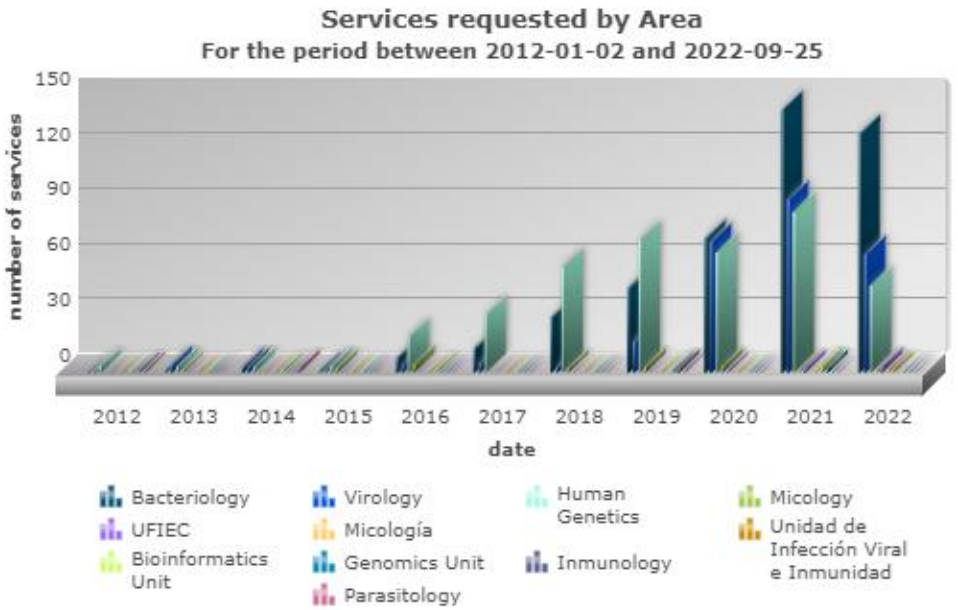
- **Soporte a Usuarios**
  Generación y acceso a máquinas virtuales que contienen software bioinformático, ubicadas en los servidores de la Unidad

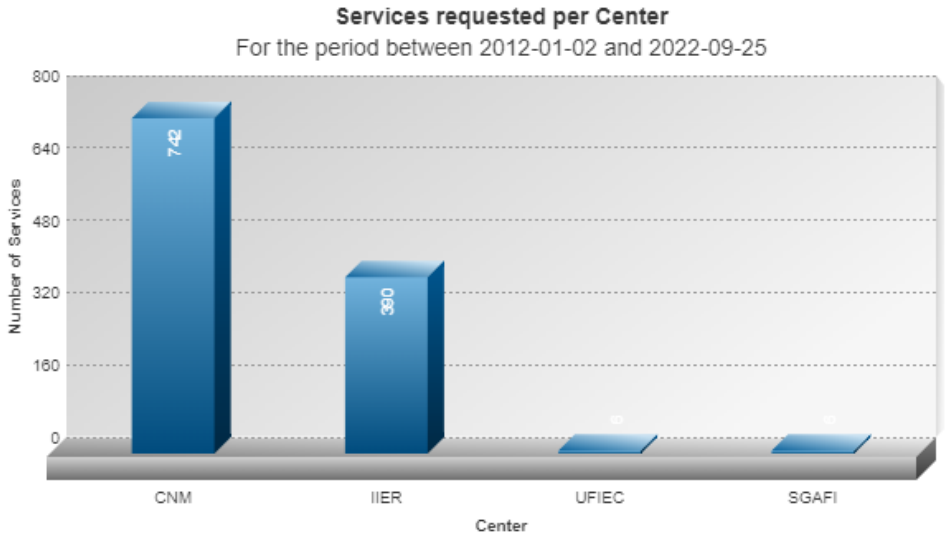
Services Portfolio

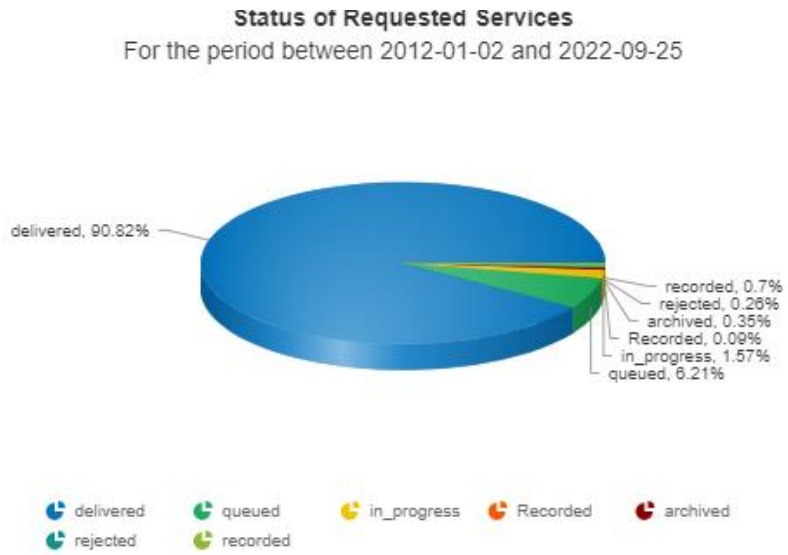
		QC	Assembly	Reference based Mapping	Variant calling	Annotation	Pipelines
DNAseq	**HUMAN**						
	WES Target –Panels	Report html		(Bam file)	(Vcf file)	Desease model (Vcf file annotated)	.Trio / family .Tumor .Pampu caller
	MICROBIAL						
	WGS Amplicon	Report html	*De novo* / Reference (fasta file)	MLST, Resistance g, Virulence g	SNPs Phylogenetic analysis	Structural Functional	.WGSOutbraker .Plasmid ID
RNAseq	**mRNA**	RSeQC Report html	*De novo* (fasta file)	Transcripts coverage / expression	Variants (Vcf file)	Transcripts annotation	mRNA seq
	miRNA						miRNA seq
Metagenomics	**16S taxonomic profile**	Report html	*De novo*	Green genes DB		species diversity	Qiime
	Shotgun			Genome Ref Seq		Pathogen / Genome coverage	PikaVirus

Number of services: 2012 – 2022









Training

Courses

ISCIII

Introduction to massive sequencing data analysis, 2013-2024 (11 editions)

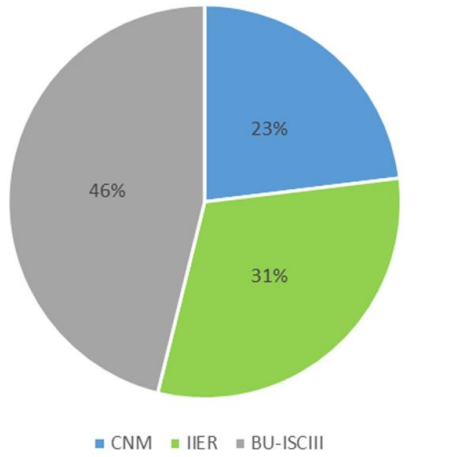
Secuenciación de genomas bacterianos: herramientas y aplicaciones, 2018-2024 (6 editions)

Análisis de genomas virales a través de la plataforma Galaxy, 2024 (4 edition - Noviembre)

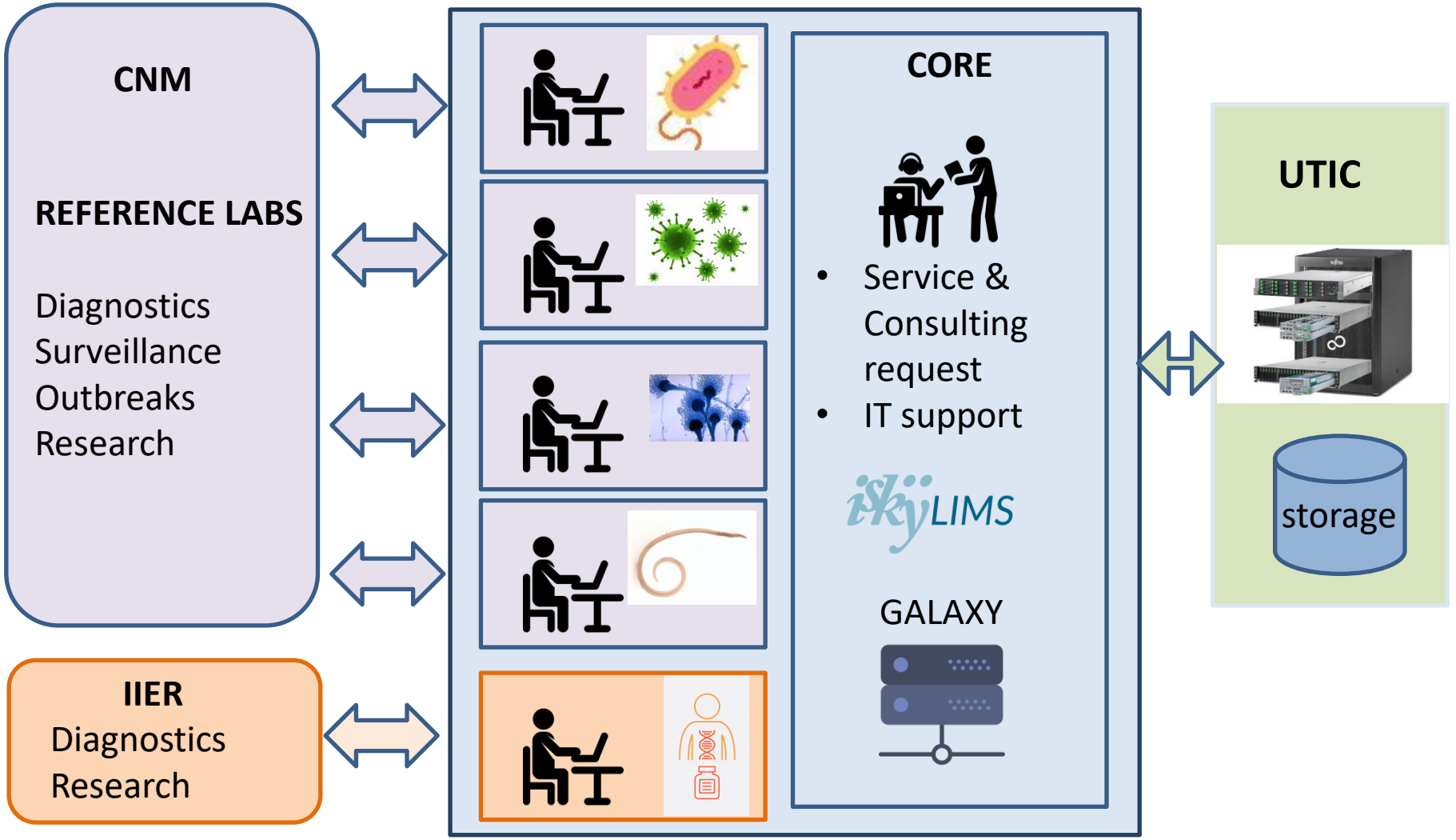
Master & Grade Students

- Bioinformatica y Biología Computacional ENS-ISCIII
- Bioinformática aplicada a la Medicina Personalizada ENS-ISCIII
- Bioinformatica UAM
- Genética y Biología Molecular UAM
- Microbiología aplicada a la salud pública e investigación en enfermedades infecciosas, U. Alcala de Henares
- Sciences in Omics Data Analysis, Universidad de VIC, U. Central de Cataluña
- Master Virología, Complutense University

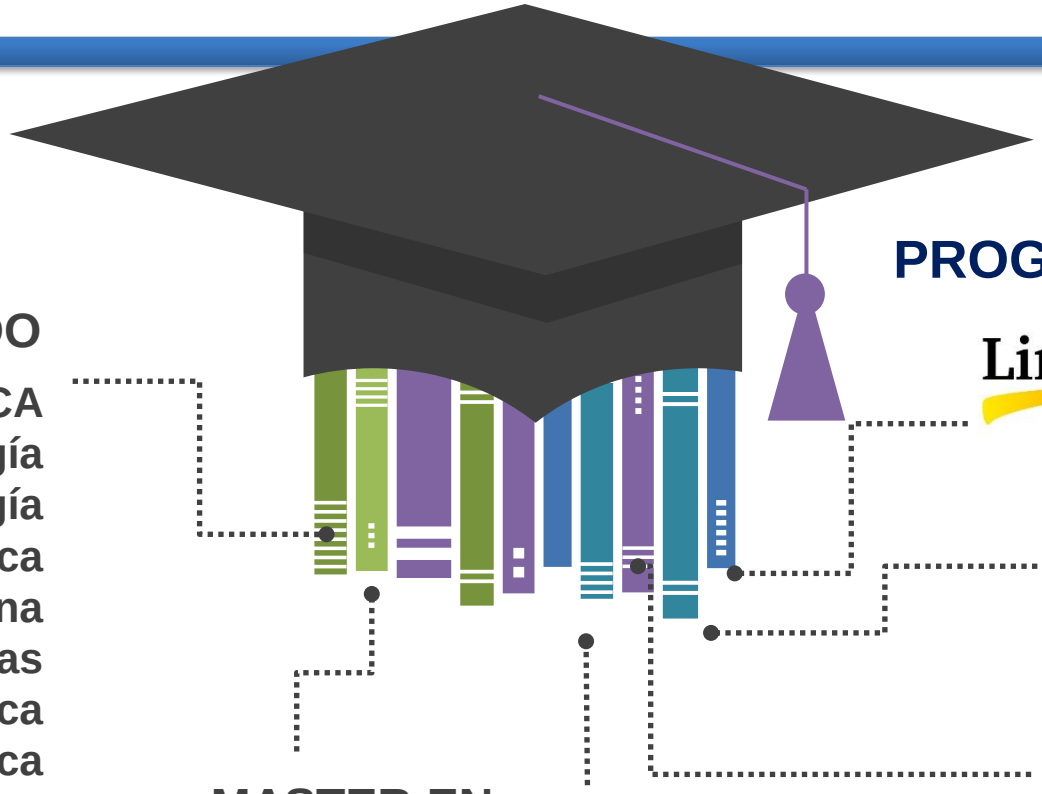
Hospitals Students



Roadmap: BU-ISCIII Model



Formación en Bioinformática



Universidad Barcelona.

GRADO

BIOINFORMÁTICA
Biología
Biotecnología
Bioquímica
Medicina
Matemáticas
Química
Fisica
Informática
Telecomunicaciones

MASTER EN BIOINFORMÁTICA

DOCTORADO

BIOLOGÍA – ANÁLISIS DE DATOS

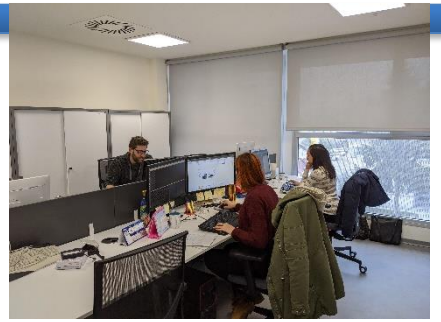
PROGRAMACIÓN

Linux

python

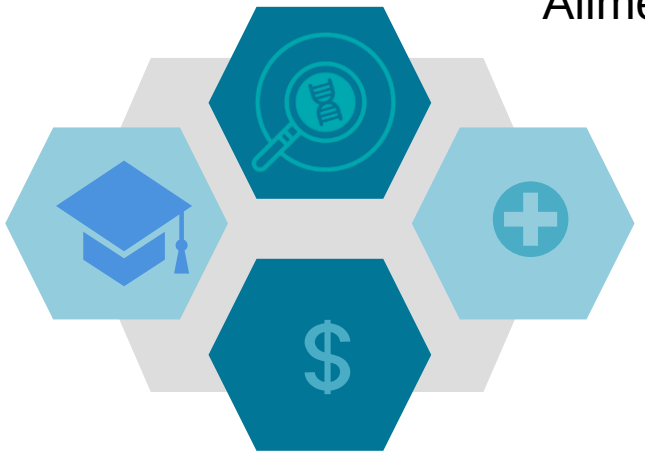
R

¿Dónde trabaja un Bioinformático?



CENTRO DE INVESTIGACIÓN

Biomedicina
Agricultura
Alimentación



UNIVERSIDAD

Biociencias
Informática

HOSPITAL

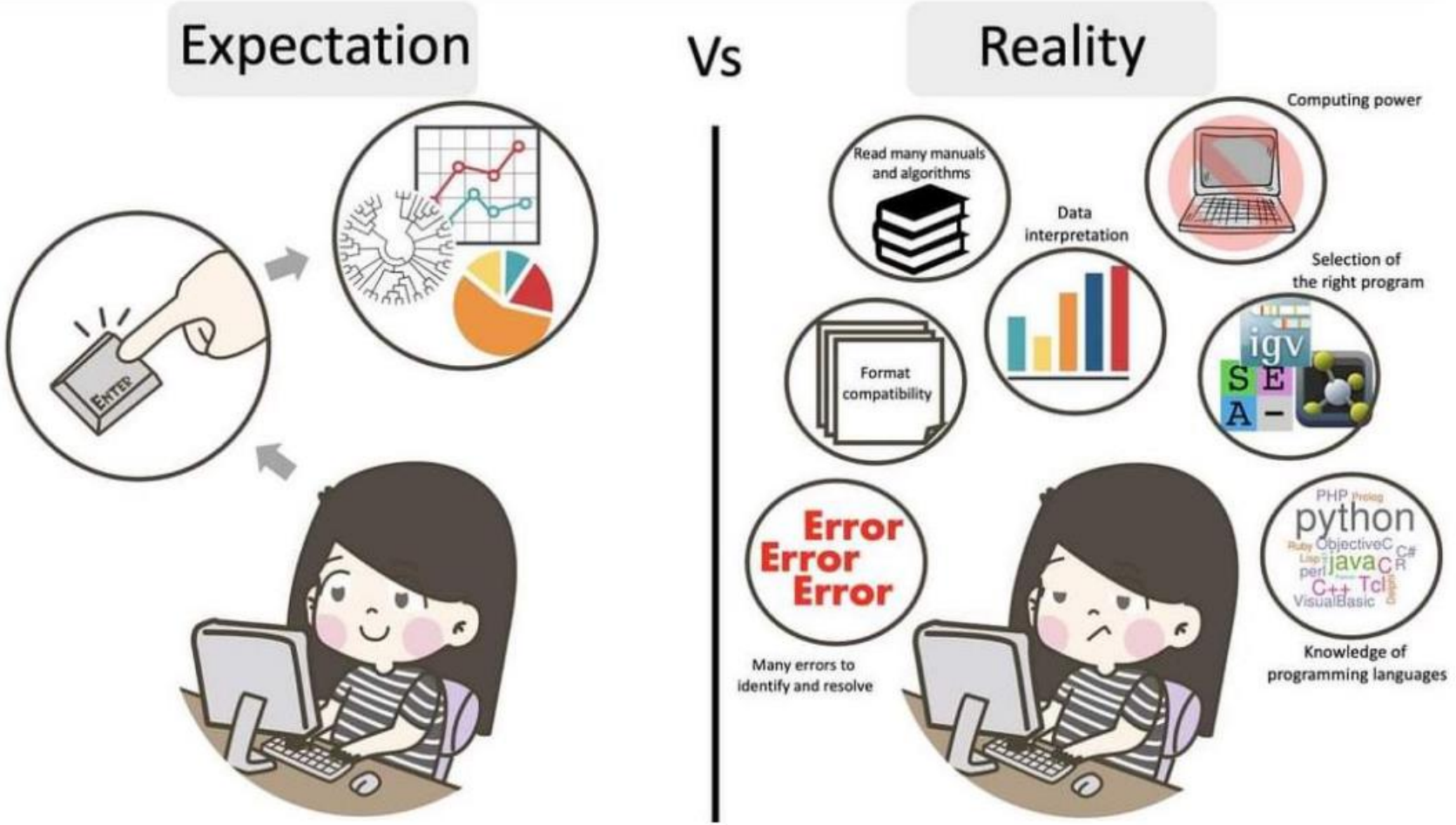
BIOINFORMÁTICO CLÍNICO
Genética
Oncología
Cardiología

EMPRESA

Bioinformática
Genética
Genómica

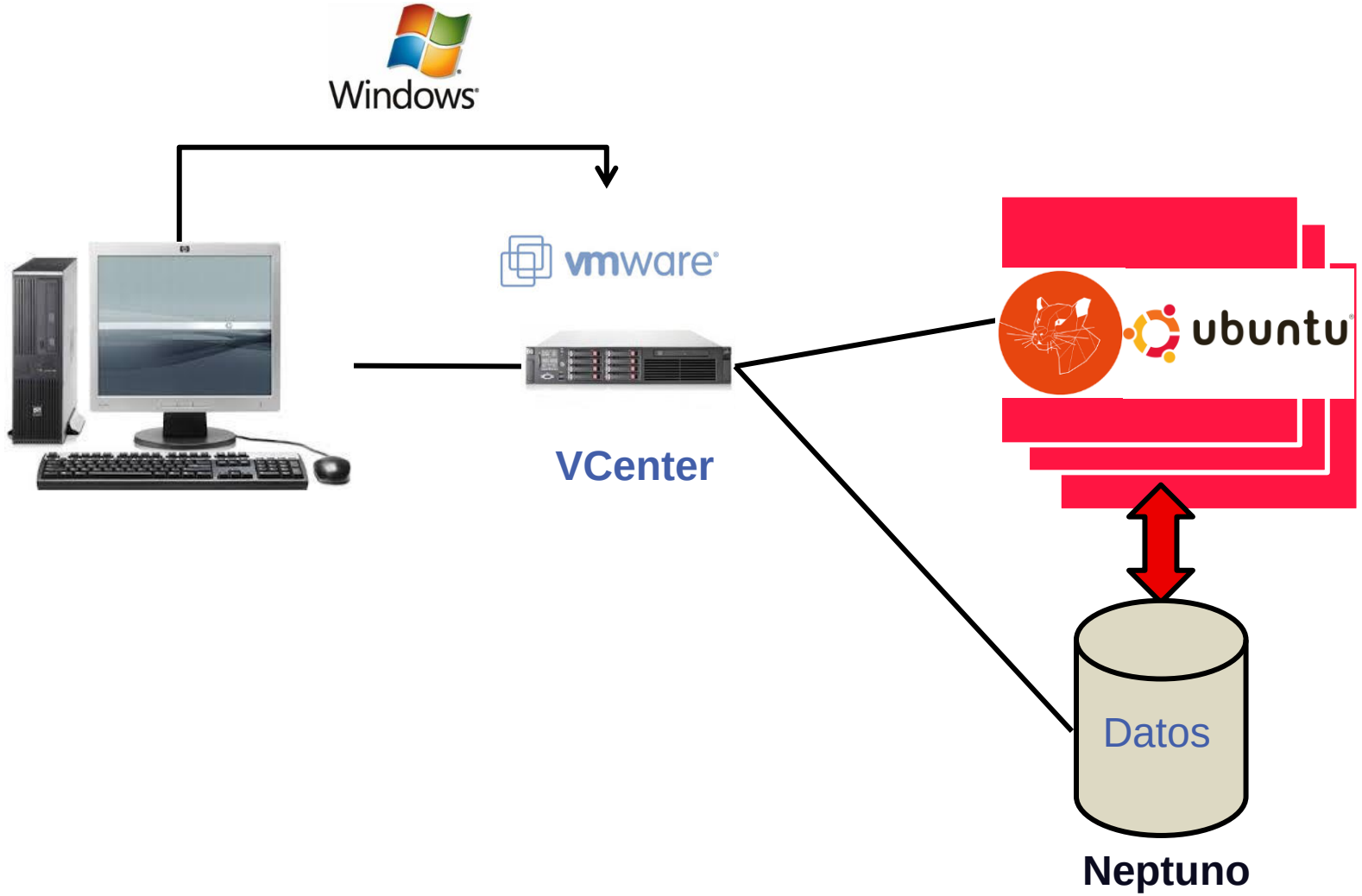
The truth about bioinformatics

.image-100[



https://training.galaxyproject.org/training-material/topics/assembly/tutorials/get-started-genome-assembly/slides-plain.html

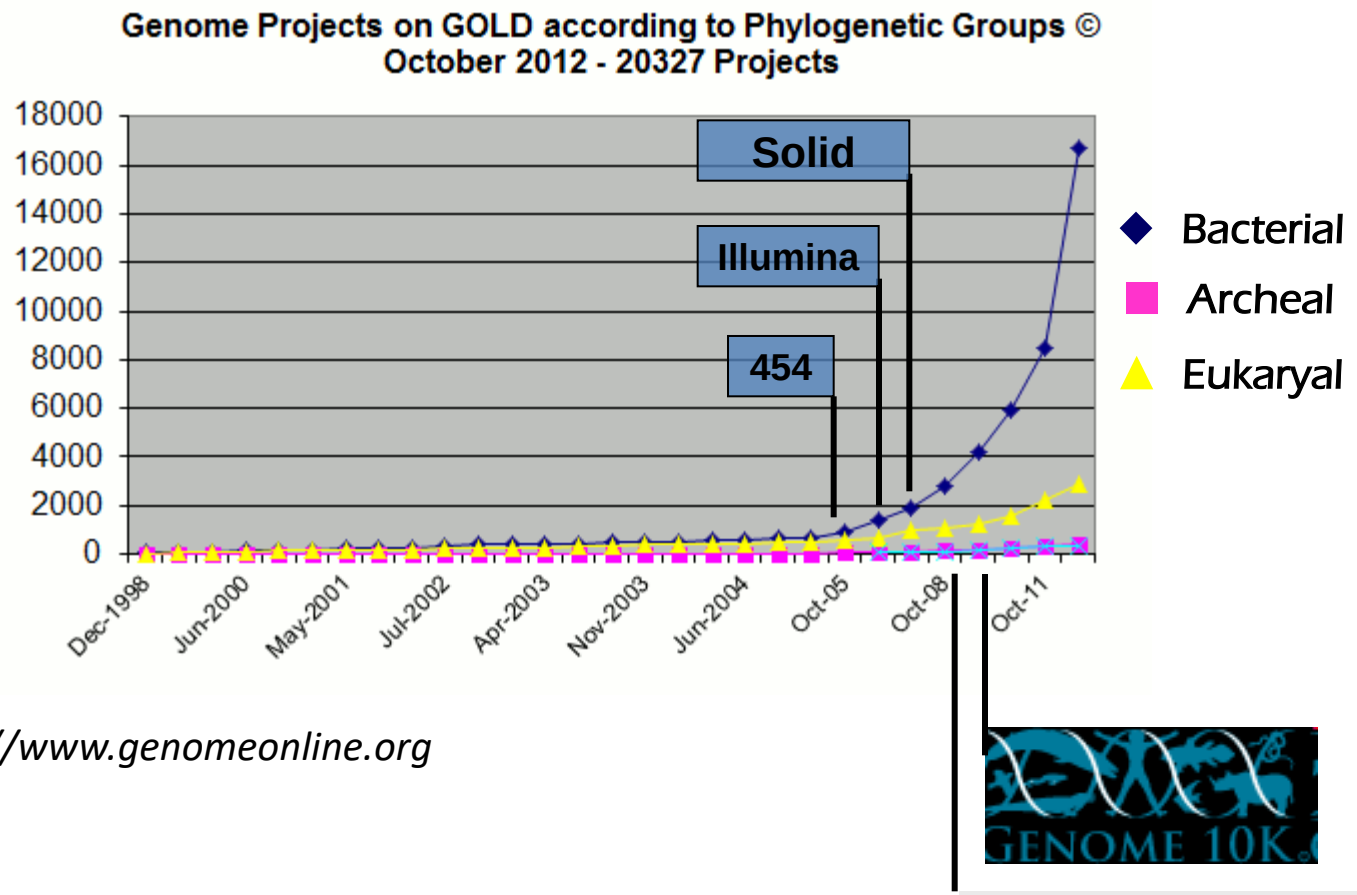
Recursos Informáticos para el curso



Index

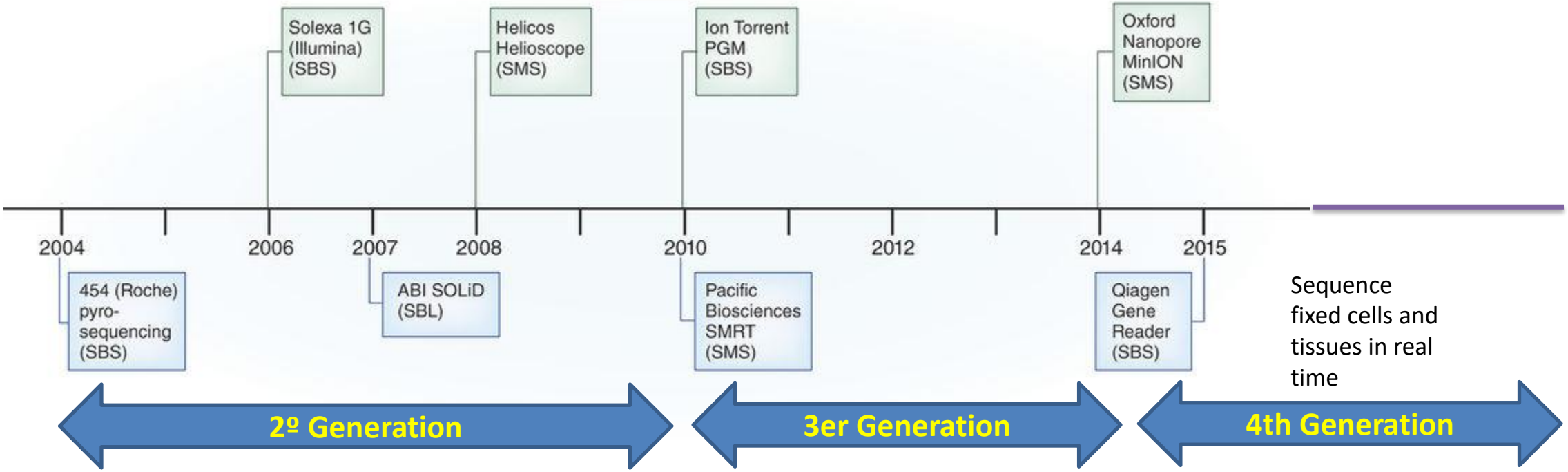
- BU-ISCIII
- High throughput sequencing platforms update
- Bacterial genome sequencing, brief history
- Advantages of WGS
- Use of WGS in Europe
- Library strategies
- Bioinformatics analysis

Genomics Revolution Era

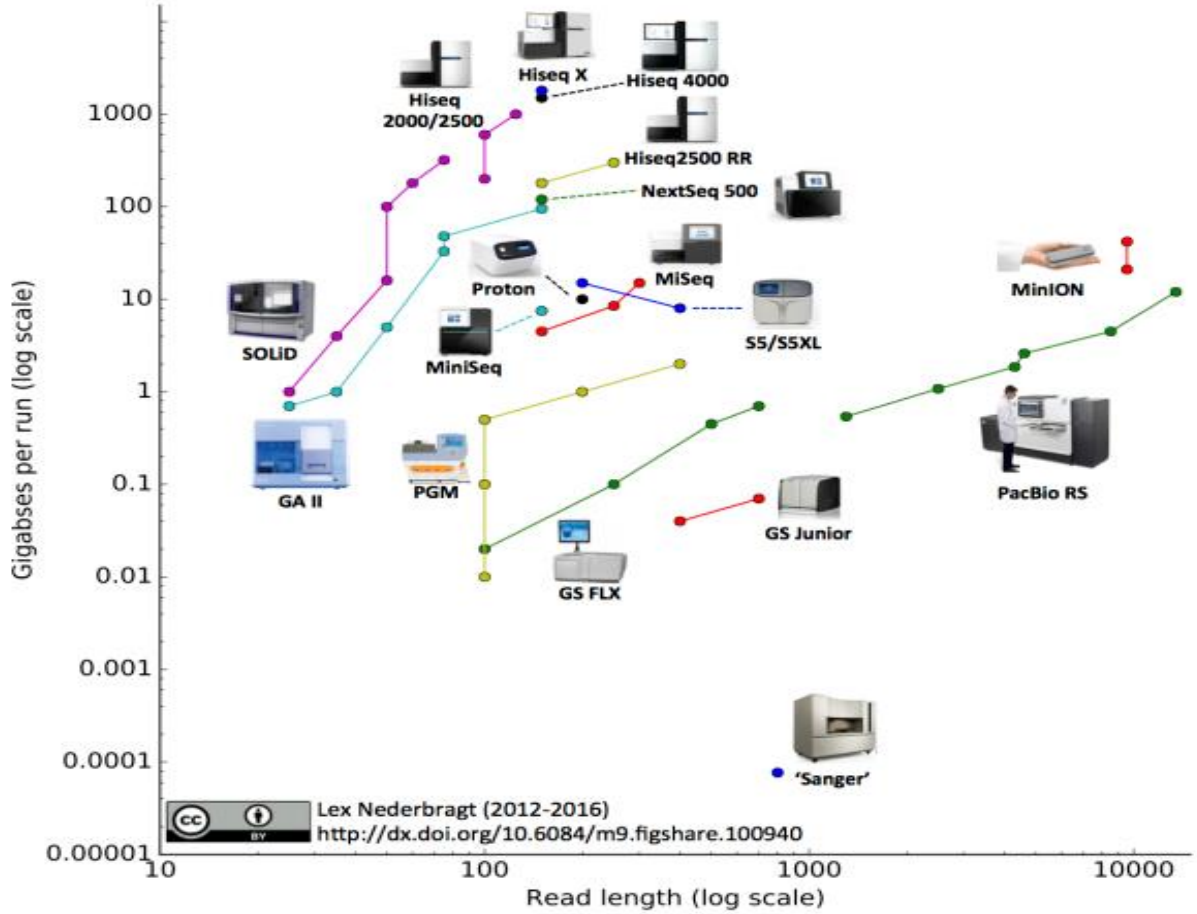


Source: http://www.genomeonline.org

NGS Platforms - Timelime

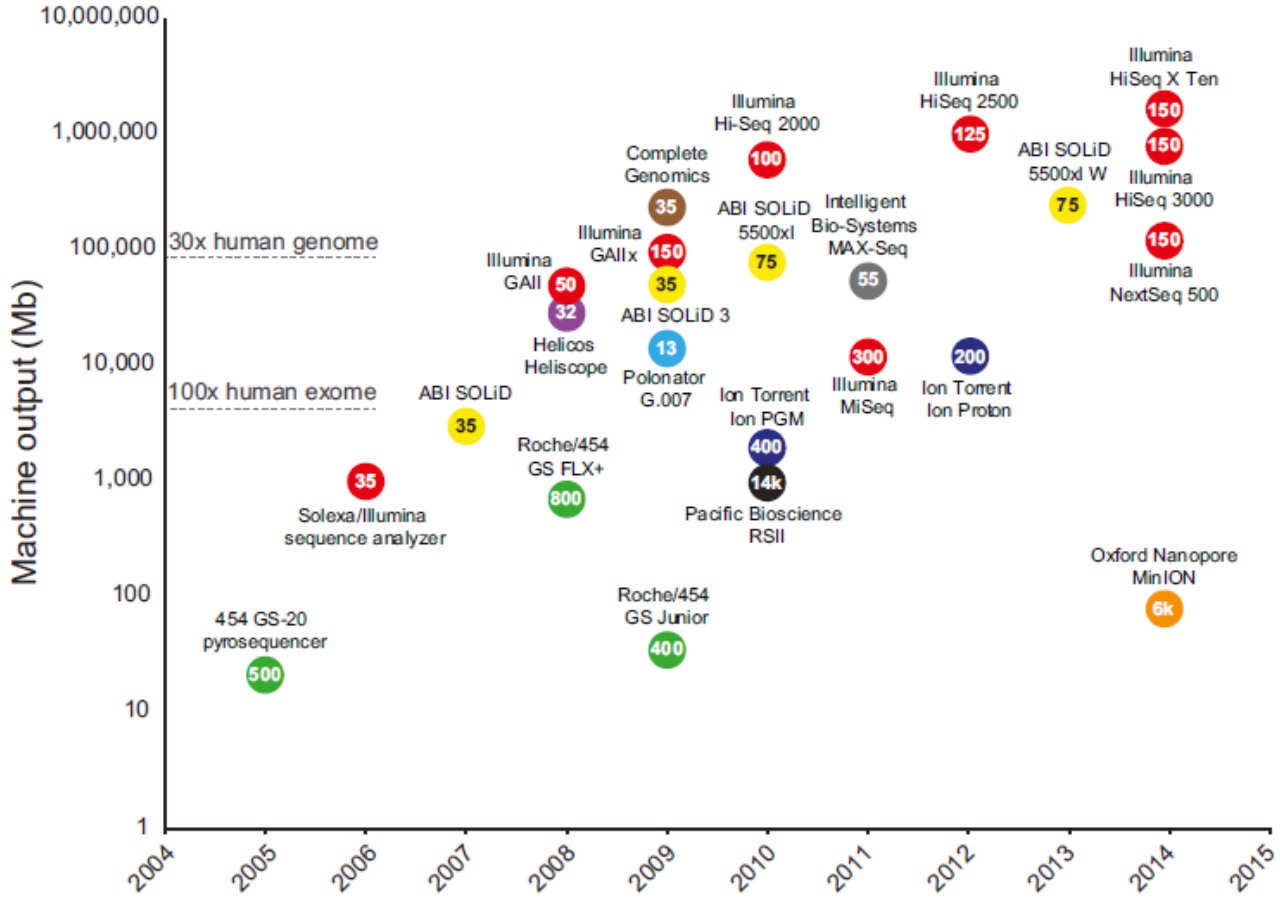


High-Throughput Sequencing Technologies



Reference

High-Throughput Sequencing Technologies

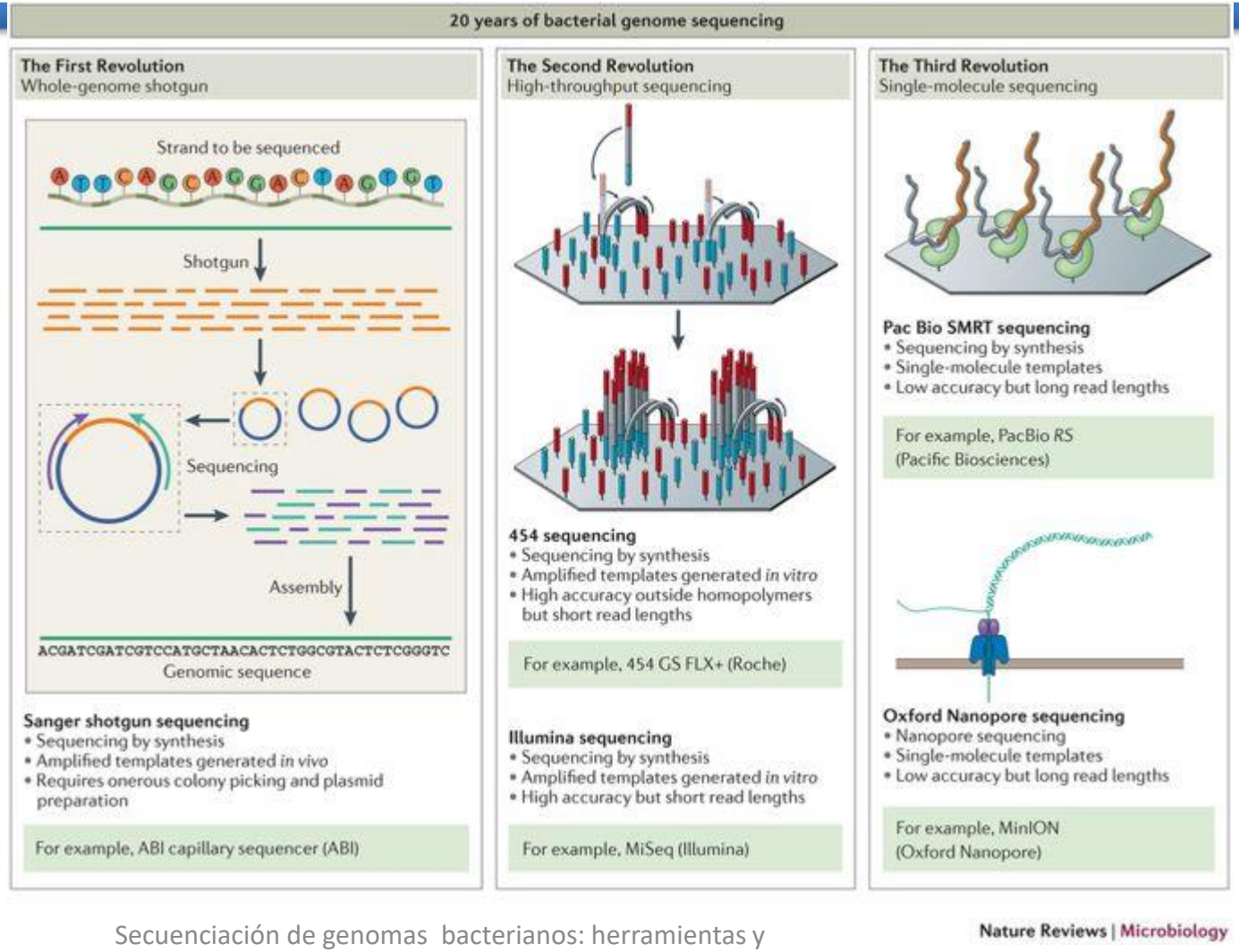


Numbers inside data points denote current read lengths. Sequencing platforms are color coded.

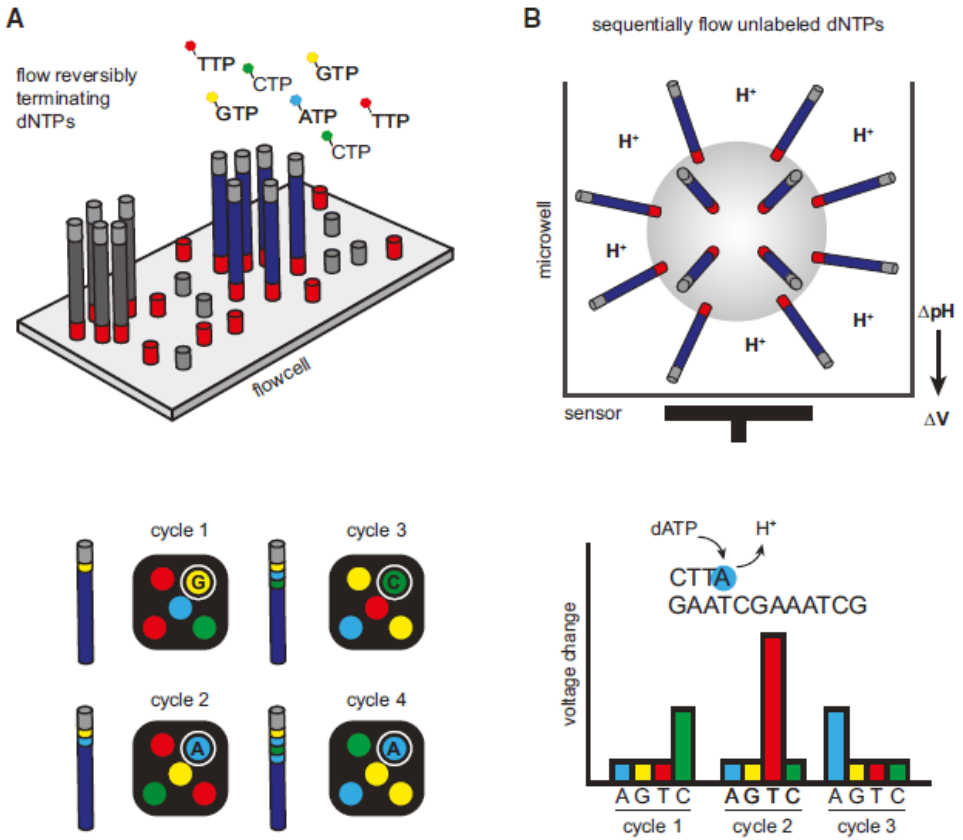
Reuter et al., Mol Cell 2015

High-Throughput Sequencing Technologies

The three revolutions in sequencing technology that have transformed the landscape of bacterial genome sequencing



The Second-generation Sequencing Technologies



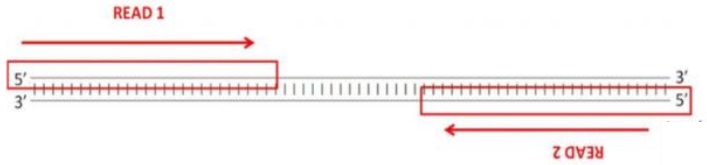
Clonal Amplification-Based Sequencing Platforms

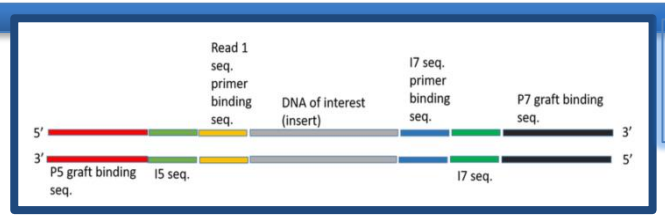
(A) Illumina's four-color reversible termination sequencing method.

(B) Ion Torrent's semiconductor sequencing method.

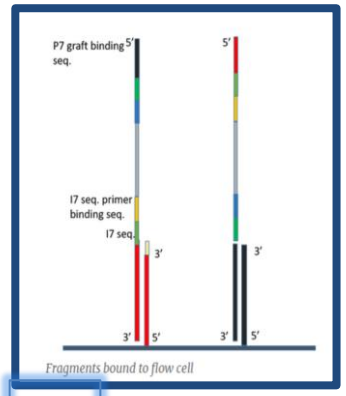
Reuter et al., Mol Cell 2015

Illumina sequencing

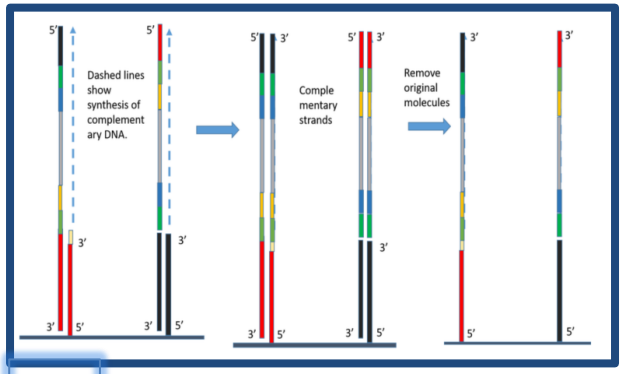




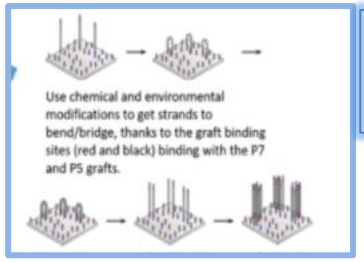
1



2

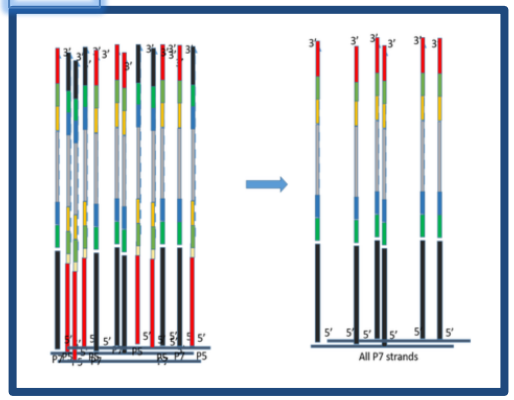


3

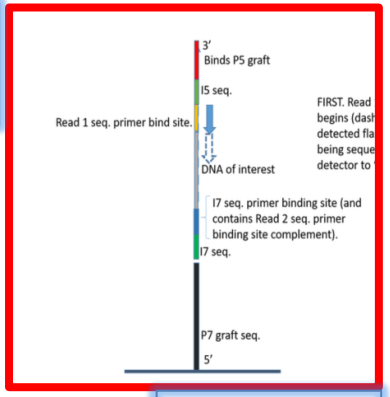


4

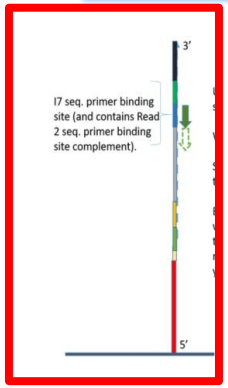
5

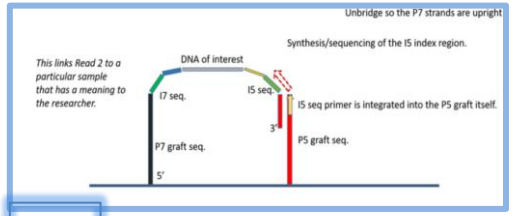


6 R1



8 R2





7

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

Illumina Benchtop Sequencers

	iSeq 100	MiniSeq	MiSeq Series	NextSeq 550 Series	NextSeq 1000 & 2000
Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)					
Small Whole-Genome Sequencing (microbe, virus)	●	●	●	●	●
Exome & Large Panel Sequencing (enrichment-based)				●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)				●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)				●	●
Targeted Gene Expression Profiling	●	●	●	●	●
miRNA & Small RNA Analysis	●	●	●	●	●
DNA-Protein Interaction Analysis (ChIP-Seq)			●	●	●
Methylation Sequencing				●	●
16S Metagenomic Sequencing		●	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)				●	●
Cell-Free Sequencing & Liquid Biopsy Analysis				●	●

Benchtop Sequencer Sheds Light on Ebola Outbreak

Local scientists use the iSeq 100 Sequencing System to analyze transmission patterns and trace the origin of an Ebola outbreak in the Democratic Republic of the Congo.

Read Article ›

https://emea.illumina.com/systems/sequencing-platforms.html

Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	11–48 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	330 Gb*
Maximum Reads Per Run	4 million	25 million	25 million †	400 million	1.1 billion*
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

Illumina Production-Scale Sequencers

NextSeq 550 Series

NextSeq 1000 & 2000

NovaSeq 6000

Popular Applications & Methods	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)			●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●
Exome & Large Panel Sequencing (enrichment-based)	●	●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)	●	●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)	●	●	●
Chromatin Analysis (ATAC-Seq, ChIP-Seq)	●	●	●
Methylation Sequencing	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)	●	●	●
Cell-Free Sequencing & Liquid Biopsy Analysis	●	●	●

Optimized NGS Sample Tracking and Workflows

See how a Laboratory Information Management System (LIMS) enabled this large genomics lab to standardize lab procedures and cope with increasing sample volumes from diverse clients.

Read Case Study >

Run Time	12–30 hours	11–48 hours	~13 - 38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)
Maximum Output	120 Gb	330 Gb*	6000 Gb
Maximum Reads Per Run	400 million	1.1 billion*	20 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 x 250**

Illumina Benchtop Sequencers

Methods/applications	iSeq 100	MiniSeq	MiSeq series	NextSeq 550 series	Next Seq 1000 & 2000
Ideal for	Every size lab	TG sequencing	Long read applications	Exome and transcriptome sequencing	miRNA and sRNA analysis
Major applications	sWGS (microbes) and TGS	iSeq 100+TG EP and 16S MS	iSeq 100+16S MGS	iSeq 100+TCS	sWGS (microbes), ES, SC profiling, TS, miRNA, and sRNA analysis
Max. data quality	>85% > Q30	>85% > Q30	>90% > Q30	>80% > Q30	>90% > Q30
Run time	9.5–19 h	4–24 hours	4–55 hours	12–30 hours	11-48 hours
Maximum output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	330 Gb*
Maximum reads per run	4 million	25 million	25 million	400 million	1.1 billion
Maximum read length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

Pervez et al., BioMed Research International 2022 QUITA

Illumina Production scale Sequencers

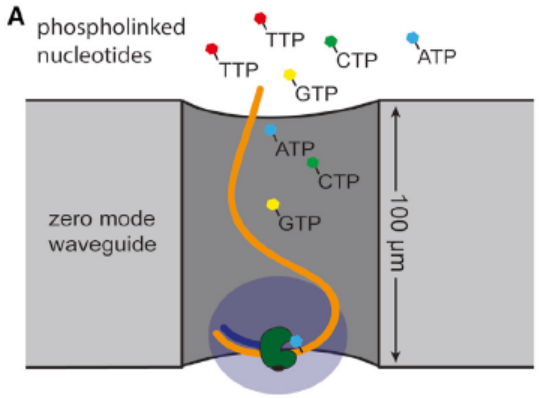
Methods/applications	NextSeq 550	NextSeq 550Dx	NextSeq 1000 & 2000	NovaSeq 6000
Ideal for	Research	Research+in vitro diagnostic	Targeted sequencing	Long read applications
Major applications	sWGS (microbes), TGS, and TCS	NextSeq 550+clinical NGS applications	NextSeq 550 series+SCP	NextSeq 550 series+NextSeq 1000 & 2000+lWGS
Max. data quality	>80% > Q30	>75% > Q30	>90% > Q30	>90% > Q30
Run time	12-30 hours	35 hours	11-48 hours	13-44 hours
Maximum output	120 Gb	90 Gb	360 Gb	6000 Gb
Maximum reads per run	400 million	300 million	1.2 billion	20 billion
Maximum read length	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 250 bp

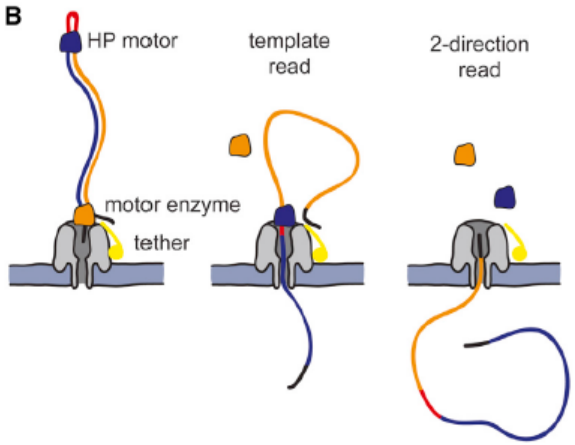
Pervez et al., BioMed Research International 2022 QUITA

3ª GENERACIÓN: LECTURAS MAS LARGAS Y MOLECULA ÚNICA

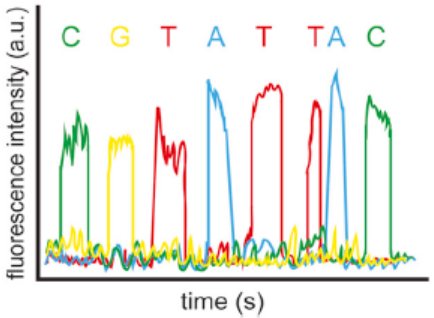
- PacBio, PACIFIC BIOSCIENCE
- MinION, GridION, OXFORD NANOPORE

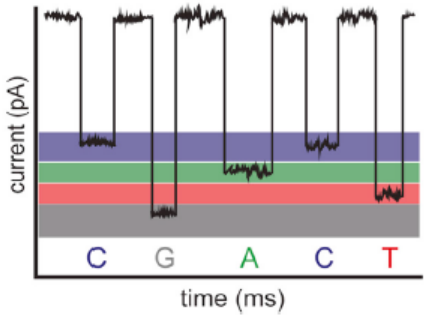
The Third-generation Sequencing Technologies





Single Molecule Sequencing Platforms





Pacific Bioscience's SMRT sequencing

Oxford Nanopore's sequencing

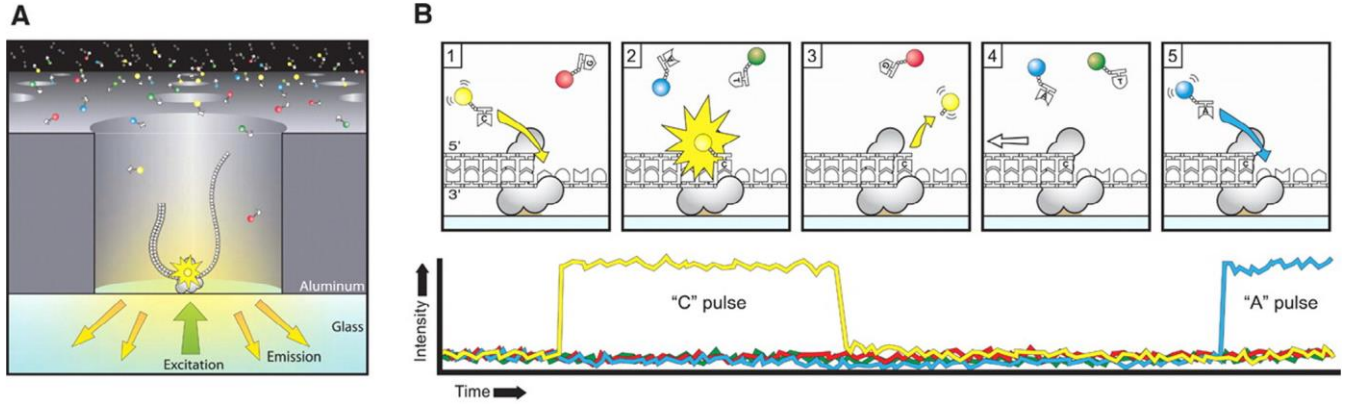
Reuter et al., Mol Cell 2015

PacBio sequencing and its applications

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

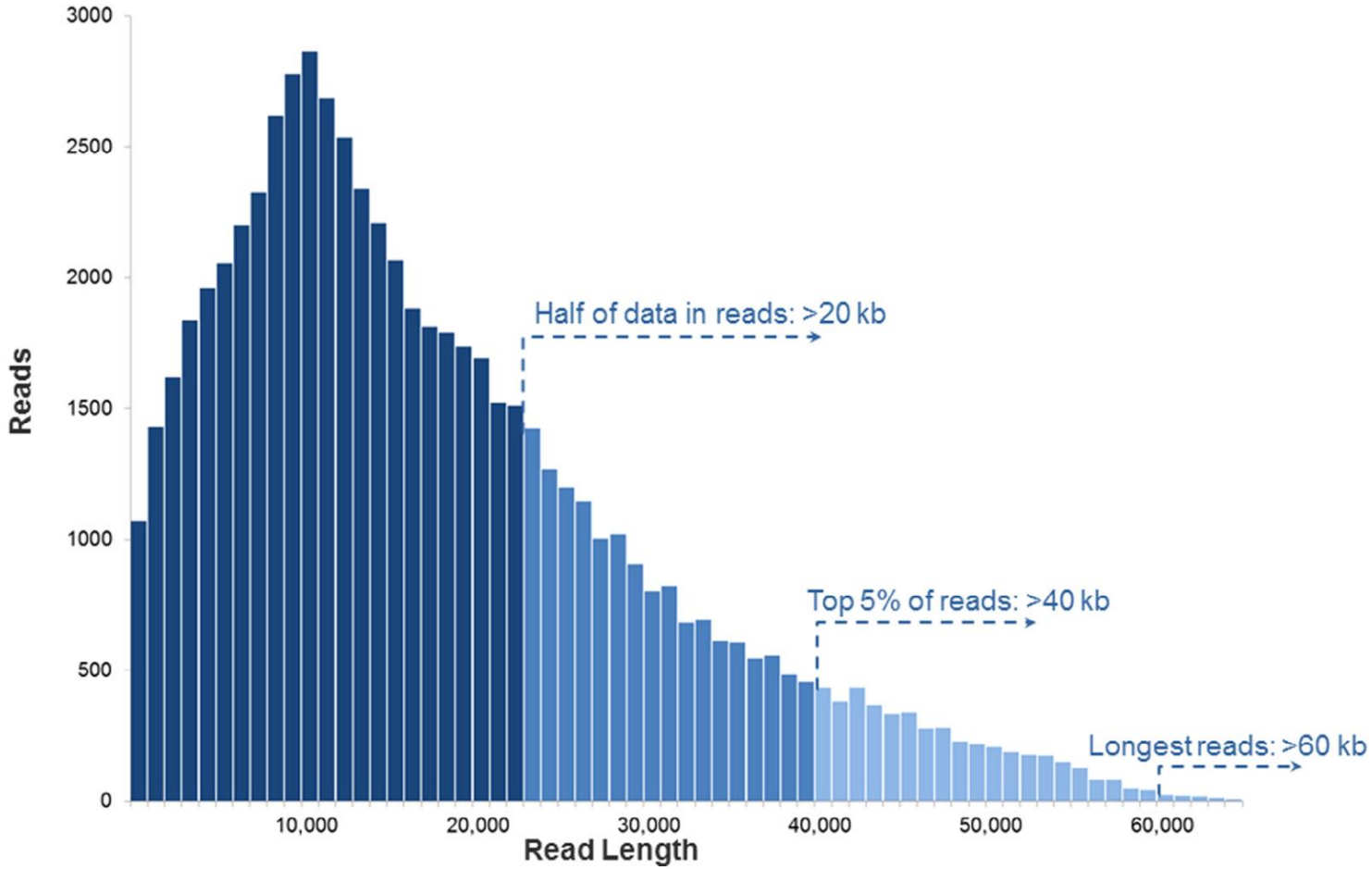
A single SMRT cell: this contains 150000 ZMWs (zero-mode waveguide). A SMRTbell diffuses into a ZMW.
Approx 35000 -75000 ZMWs produce a read in a run lasting 0,5-4h resulting in 0,5-1Gb.



Rhoads & Au, Gen Prot Bioinf 2015

PacBio sequencing and its applications

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



Rhoads & Au, Gen Prot Bioinf 2015

PacBio sequencing and its applications

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

Species	Method	Tools	SMRT cells	Coverage	Contigs	Achievements	Ref.
Clostridium autoethanogenum	PacBio	HGAP	2	179×	1	21 fewer contigs than using SGS; no collapsed repeat regions (≥4 using SGS)	[7]
Potentilla micrantha (choloroplast)	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]
Escherichia coli	PacBio	PBcR, MHAP, Celera, Quiver	1	85×	1	4.6 CPU hours for genome assembly (10× improvement over BLASR)	[31]
Saccharomyces cerevisiae	PacBio	PBcR, MHAP, Celera	1[2]	117×	21	27 CPU hours for genome assembly (8× improvement over BLASR); improved current reference of telomeres	[31]
Arabidopsis thaliana	PacBio	PBcR, MHAP, Celera	46	144×	38	1896 CPU hours for genome assembly	[31]
Drosophila melanogaster	PacBio	PBcR, MHAP, Celera, Quiver	42	121×	132	1060 CPU hours for genome assembly (593× improvement over BLASR); improved current reference of telomeres	[31]
Homo sapiens (CHM1htert)	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]
Homo sapiens (CHM1tert)	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]
Melopsittacus undulatus	Hybrid	PBcR, Celera	3	5.5× PacBio + 15.4× 454 = 3.83× corrected	15,328	1st assembly of >1 Gb parrot genome; N50 = 93,069	[34]
Vibrio cholerae	Hybrid	BLASR, Bambus, AHA	195	200× PacBio + 28× Illumina + 22× 454	2	No N's in contigs; 99.99% consensus accuracy; N50 = 3.01 Mb	[30]
Helicobacter pylori	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.

Rhoads & Au, Gen Prot Bioinf 2015

PacBio sequencing and its applications

Advantage

Closes gaps and completes genomes due to longer reads

Identifies non-SNP SVs

Achievements

Produced highly-contiguous assemblies of bacterial and eukaryotic genomes

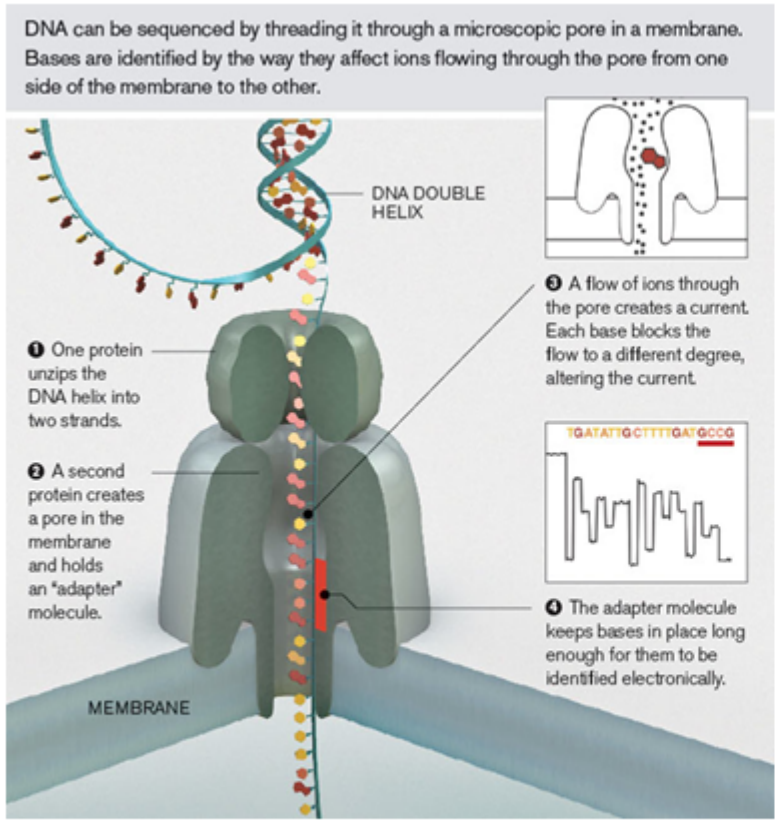
Discovered STRs (short tandem repeats)

Limitations

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

Rhoads & Au, Gen Prot Bioinf 2015

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

Nanopore sequencing

The current state of Nanopore Sequencing. Arakawa. Methods in Molecular Biology 2023

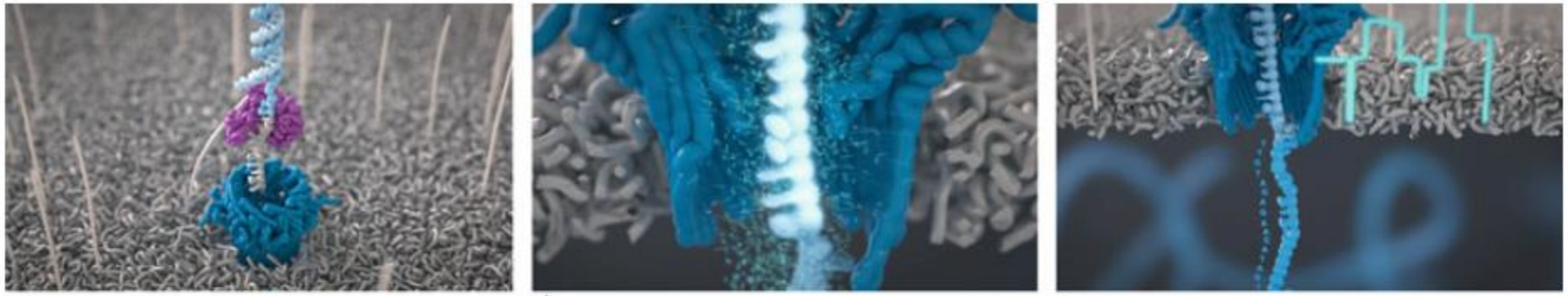
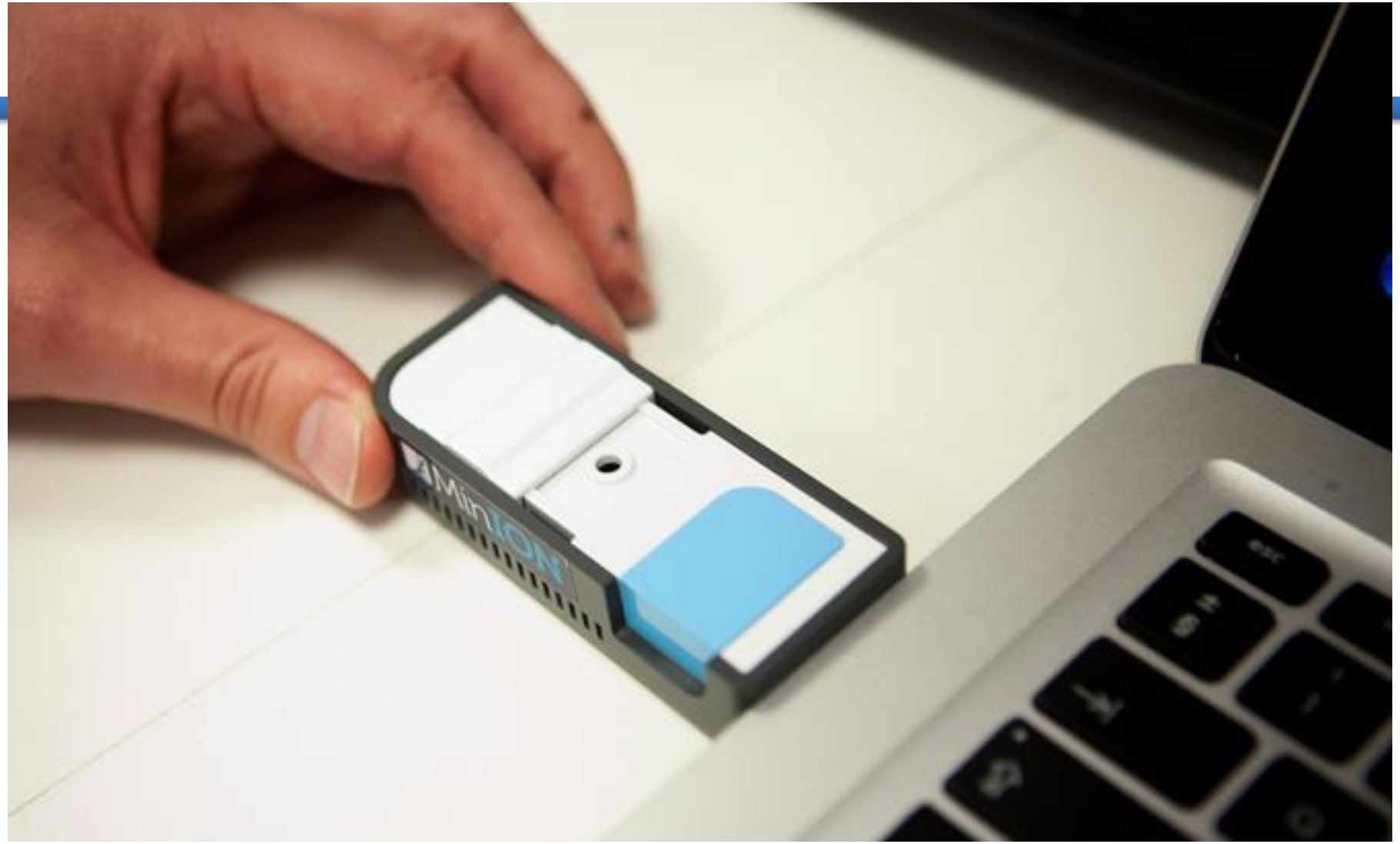


Fig. 1 The principle of nanopore sequencing. (**a**) A protein nanopore (blue) is imbedded into an electronically resistive lipid membrane (grey), before adapted DNA libraries containing a motor protein (purple) are introduced, and the motor feeds DNA progressively through the pore. (**b**) An ionic current (represented by light blue dots) is passed through the nanopore as the DNA translocates through the pore. (**c**) The bases within the nanopore block the current depending on their size and structure. As the strand moves progressively through the pore, a "squiggle" trace is produced, which is decoded into sequence data using artificial neural networks

MinIon, OXFORD NANOPORE



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

The current state of Nanopore Sequencing. Arakawa. Methods in Molecular Biology 2023

Flow cell name	Flongle	MinION Flow Cell	PromethION Flow Cell
Number of channels	126	512	2,675
Theoretical maximum output*	2.8 Gbases	50 Gbases	290 Gbases

Device name	MinION	MinION Mk1C	GridION	PromethION 2 Solo	PromethION 2	PromethION 24	PromethION 48
Flow cell compatibility	Flongle, MinION			PromethION			
Number of flow cells that can be run	1	1	5	2	2	24	48

Fig. 2 The flow cells and devices for nanopore sequencing. The Flongle (**a**) consists of two parts, a reusable adapter, and a single-use flow cell. It has the same footprint as the MinION Flow Cell (**b**) meaning both can be run on the MinION (**d**), MinION Mk1C (**e**), or GridION (**f**) devices. Any combination of Flongle or MinION can be run on the GridION device. The PromethION Flow Cell (**c**) is compatible with all PromethION devices (**g**–**j**). With capacity for different numbers of flow cells, total device yields vary in line with the number of flow cells they can run. Where multiple flow cells can be run, all are individually controllable, meaning no requirement exists to run all flow cells at once and as a result samples can be run on demand. *Theoretical maximum output when flow cell or device is run 72 h (16 h for Flongle) at 420 bases/second. For devices, this is when all flow cells are run at once and the highest yielding flow cell option is chosen. Outputs may vary according to library type, run conditions, etc.

Oxford Nanopore Technologies

	Flongle	MinION Mk1B	MinION Mk1C	GridION Mk1	PromethION 24	PromethION 48
Number of channels per flow cell	126	512	512	512	3000	3000
Number of flow cells per device	1	1	1	5	24	48
Price per flow cell	$90	$900 - $475	$900 - $475	$900 - $475	$2000 - $625	$2000 - $625
Run time	1 min - 16 hours	1 min - 72 hours	1 min - 72 hours	1 min - 72 hours	1 min - 72 hours	1 min - 72 hours
Yields in field are dependent on sample and preparation methods. Users can get outputs in the following ranges per flow cell when utilising the latest chemistries and protocols	1 - 2 Gb	10 - 30 - 50 Gb	10 - 30 - 50 Gb	10 - 30 - 50 Gb	100 - 200 - 300 Gb	100 - 200 - 300 Gb
Price per Gb for different flow cell yields (yields vary according to sample and preparation methods)	**@1 - 2 Gb** $90 per flow cell: $90 - 45	**@10 - 30 - 50 Gb** $900 per flow cell: $90 - 30 - 18 $790 per flow cell: $79 - 26 - 16 $675 per flow cell: $68 - 23 - 14 $500 per flow cell: $50 - 17 - 10 $475 per flow cell: $48 - 16 - 9.5	**@10 - 30 - 50 Gb** $900 per flow cell: $90 - 30 - 18 $790 per flow cell: $79 - 26 - 16 $675 per flow cell: $68 - 23 - 14 $500 per flow cell: $50 - 17 - 10 $475 per flow cell: $48 - 16 - 9.5	**@10 - 30 - 50 Gb** $900 per flow cell: $90 - 30 - 18 $790 per flow cell: $79 - 26 - 16 $675 per flow cell: $68 - 23 - 14 $500 per flow cell: $50 - 17 - 10 $475 per flow cell: $48 - 16 - 9.5	**@100 - 200 - 300 Gb** $1,600 per flow cell: $16 - 8 - 5 $1,120 per flow cell: $11 - 6 - 4 $940 per flow cell: $9 - 5 - 3.1 $680 per flow cell: $7 - 3.4 - 2.3 $625 per flow cell: $6 - 3 - 2	**@100 - 200 - 300 Gb** $1,600 per flow cell: $16 - 8 - 5 $1,120 per flow cell: $11 - 6 - 4 $940 per flow cell: $9 - 5 - 3.1 $680 per flow cell:

https://nanoporetech.com/products/comparison

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.

MinIT, Analysis

Eliminating the need for a dedicated laptop for nanopore sequencing with MinION.
$2400

MinIT Specifications:

Pre-installed software: Linux OS, MinKNOW, Guppy, EPI2ME
Bluetooth and Wi-Fi enabled; you can control your experiments using a laptop, tablet or smartphone
fastq or fast5 files are written to Onboard storage: 512 GB SSD
Processing: GPU accelerators (ARM processor 6 cores, 256 Core GPU), 8 GB RAM.
Small footprint, 290g
1 x USB 2.0 port, 1 x USB 3.0 port and 1 x Ethernet port (1 Gbit capacity)

MinIT has now been replaced by the MinION Mk1C, which combines the real-time, portable sequencing of MinION, with powerful integrated compute, a high-resolution touchscreen, and full connectivity.

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.

Long-read Sequencing Platforms characteristics

Company	Pacific Biosciences			Oxford Nanopore			
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 *de novo* 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
Dimensions	92.7 x 91.4 x 167.6 cm			105 x 23 x 8 mm	Mk1b: 105 x 23 x 33 mm; Mk1c: 140 x 30 x 116 mm	365 x 220 x 360 mm	Sequencer: 590 x 190 x 430 mm; Data Acquisition unit: 178 x 440 x 470 mm
Weight	362 kg			20 g	Mk1b: 87 g Mk1c: 450 g	11 kg	Sequencer: 28 kg; Data Acquisition unit: 25 kg
Advantages	Very long reads can help resolve ambiguities; no DNA amplification required, comparatively faster turnaround time			Fast-sequencing; Small instrument footprint; Portability; Real-time data analysis			
Disadvantages	Sequencing equipment is expensive, which could be cost-prohibitive for smaller clinical laboratories, large footprint of the equipment, historically higher error rate (continues to improve)			Historically higher error rate (continues to improve)			

PacBio sequencing and its applications

Performance comparison of sequencing platforms of various generations

Method	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (USD)	Refs.
Sanger ABI 3730×1	1st	600–1000	0.001	96	0.5–3 h	500	[14,18–21]
Ion Torrent	2nd	200	1	8.2×10^7	2–4 h	0.1	[15,25]
454 (Roche) GS FLX+	2nd	700	1	1×10^6	23 h	8.57	[14,17,27]
Illumina HiSeq 2500 (High Output)	2nd	2×125	0.1	8×10^9 (paired)	7–60 h	0.03	[9,16,26]
Illumina HiSeq 2500 (Rapid Run)	2nd	2×250	0.1	1.2×10^9 (paired)	1–6 days	0.04	[9,16,26]
SOLiD 5500×1	2nd	2×60	5	8×10^8	6 days	0.11	[14,24]
PacBio RS II: P6-C4	3rd	$1.0–1.5 \times 10^4$ on average	13	$3.5–7.5 \times 10^4$	0.5–4 h	0.40–0.80	[5,12,15]
Oxford Nanopore MinION	3rd	$2–5 \times 10^3$ on average	38	$1.1–4.7 \times 10^4$	50 h	6.44–17.90	[22,23]

Rhoads & Au, Gen Prot Bioinf 2015

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)[a]	Read length (bp)	Strength	Weakness
Sanger sequencing				
ABI 3500/3730	0.0003	Up to 1 kb	Read accuracy and length	Cost and throughput
Illumina				
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
IonTorrent				
PGM	0.08–2	Up to 400	Read length, speed	Throughput, homopolymers[c]
S5	0.6–15	Up to 400	Read length, speed,	Homopolymers[c]
Proton	10–15	Up to 200	Speed, throughput	Homopolymers[c]
Pacific BioSciences				
PacBio RSII	0.5–1[b]	Up to 60 kb	Read length, speed (Average 10 kb, N50 20 kb)	High error rate and initial
Sequel	5–10[b]	Up to 60 kb	Read length, speed (Average 10 kb, N50 20 kb)	High error rate
Oxford Nanopore				
MInION	0.1–1	Up to 100 kb	Read length, portability	High error rate, run length,

[a] 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.
[b] Per one single-molecule real-time cell.
[c] 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

Comparison of various high-performing sequencing instruments

Manufacturer	Read length	Data output	Max. run time (hours)	Chemistry	Key applications**
Illumina (NovaSeq 6000)	300 PE	6 Tb (6000 Gb)	44	Sequencing by synthesis	SS-WGS and TGS, TGEP, 16sMGS, WES, SCP, LS-WGS, CA, MS, MGP, CFS, LBA
Thermo Fisher Scientific Ion Torrent (Ion GeneStudio S5 Prime)	600 SE	50 Gb	12	Sequencing by synthesis	WGS, WES, TGS
GenapSys (16 chips)	150 SE	2 Gb	24	Sequencing by synthesis	TS, SS-WGS, GEV, 16S rRNA sequencing, sRNA sequencing, TSCAS
QIAGEN (GeneReader)	100 SE	Not available	Not available	Sequencing by synthesis	Cancer research and identifying mutations
BGI/Complete Genomics	400 SE	6 Tb (6000 Gb)	40	DNA nanoball	Small and large WGS, WES and TGS
PacBio (HiFi Reads)	25 Kb	66.5 Gb	30	Real-time sequencing	DN sequencing, FT, identifying ASI, mutations, and EPM
Nanopore (PromethION)	4 Mb	14 Tb (14000 Gb)	72	Real-time sequencing	SV, GS, phasing, DNA and RNA base modifications, FT, and isoform detection

Pervez et al., BioMed Research International 2022 QUITA

Advantages and disadvantages of sequencing generations

Sequencing generation	Advantages	Disadvantages
First generation	High accuracy Helps in validating findings of NGS	High cost Low throughput
Second generation	High throughput Low cost Have clinical applications Short run time	Short read length Difficult sample preparation PCR amplification Long run time
Third generation	No PCR amplification Require less starting material Longer read lengths Very low cost Low error rate during library preparation Advantages of 3rdGS+	High sequencing error rate 10–15% in the PacBio and 5–20% in the ONT Fresh DNA requires for ensuring quality of ultralong reads Database systems and algorithms/tools are rare for analyzing 3rd and 4th GS data
Fourth generation	Ultrafast: scan of whole genome in 15 minutes Spatial distribution of the sequencing reads over the sample can be seen	

Pervez et al., BioMed Research International 2022 QUITA

Advantages & disadvantages for short vs. Long read sequencing

https://www.technologynetworks.com/genomics/articles/an-overview-of-next-generation-sequencing-346532

	Advantages	Limitations
Short-read sequencing	· Higher sequence fidelity · Cheap · Can sequence fragmented DNA	· Not able to resolve structural variants, phasing alleles or distinguish highly homologous genomic regions · Unable to provide coverage of some repetitive regions
Long-read sequencing	· Able to sequence genetic regions that are difficult to characterize with short-read seq due to repeat sequences · Able to resolve structural rearrangements or homologous regions · Able to read through an entire RNA transcript to determine the specific isoform · Assists *de novo* genome assembly	· Lower per read accuracy · Bioinformatic challenges, caused by coverage biases, high error rates in base allocation, scalability and limited availability of appropriate pipelines

Advantages 3rd GS over 2nd GS

- Higher throughput
- Detecting haplotype directly
- Longer read lengths
- Better consensus accuracy to identify rare variants
- Whole chromosome phasing
- Small amount of sample are the salient features of the 3rd-generation sequencing which had it useful in clinical diagnostic

Pervez et al., BioMed Research International 2022 QUITA

2nd GS and 3rd GS

Table 5: An overview of human genome assembly quality metrics between PacBio system, Nanopore, and Illumina [49].

	Nanopore+Illumina	PacBio HiFi sequencing
Contiguity (N50)	32.3 Mb	98.7 Mb
Correctness (quality score)	Q34	Q51
Completeness (genome size)	2.8 Gb	3.1 Gb

Table 6: Overall costs for sequencing a human genome [49].

	Nanopore+Illumina (US $)	PacBio HiFi sequencing
Consumables	4800	3800
Compute	5050	3850
Data storage	5200	3900

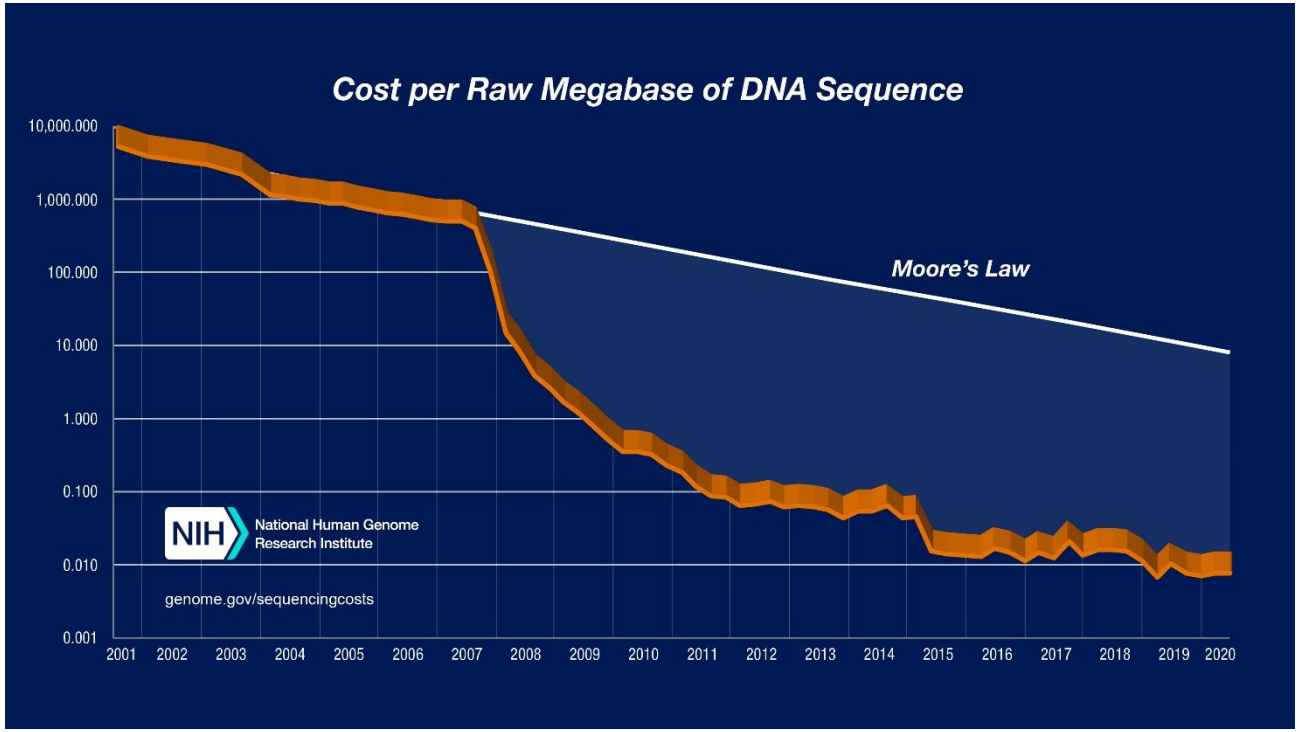
Pervez et al., BioMed Research International 2022 QUITA

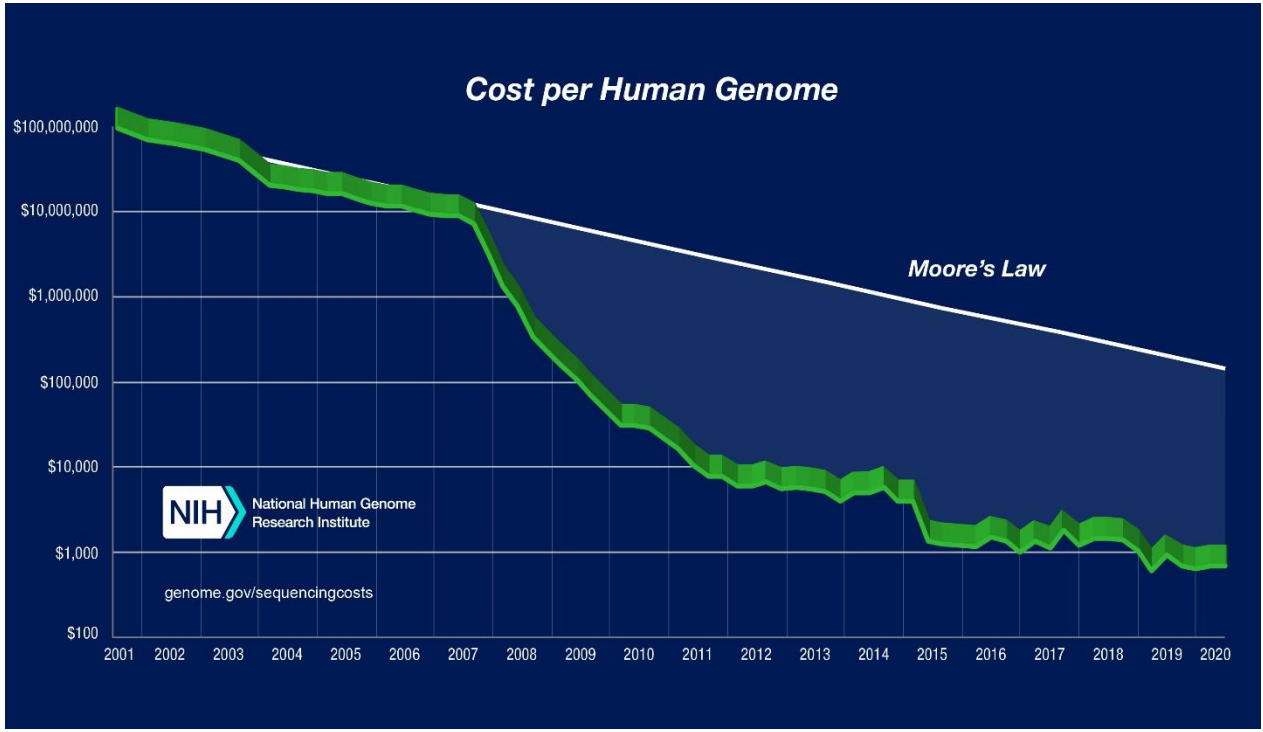
NGS PLATFORMS, main characteristics

- Numero de bases que secuencia
- Numero lecturas → aplicaciones
- Longitud de las lecturas -→ importante para las aplicaciones ensamblado genomas, de illumina a PacBio

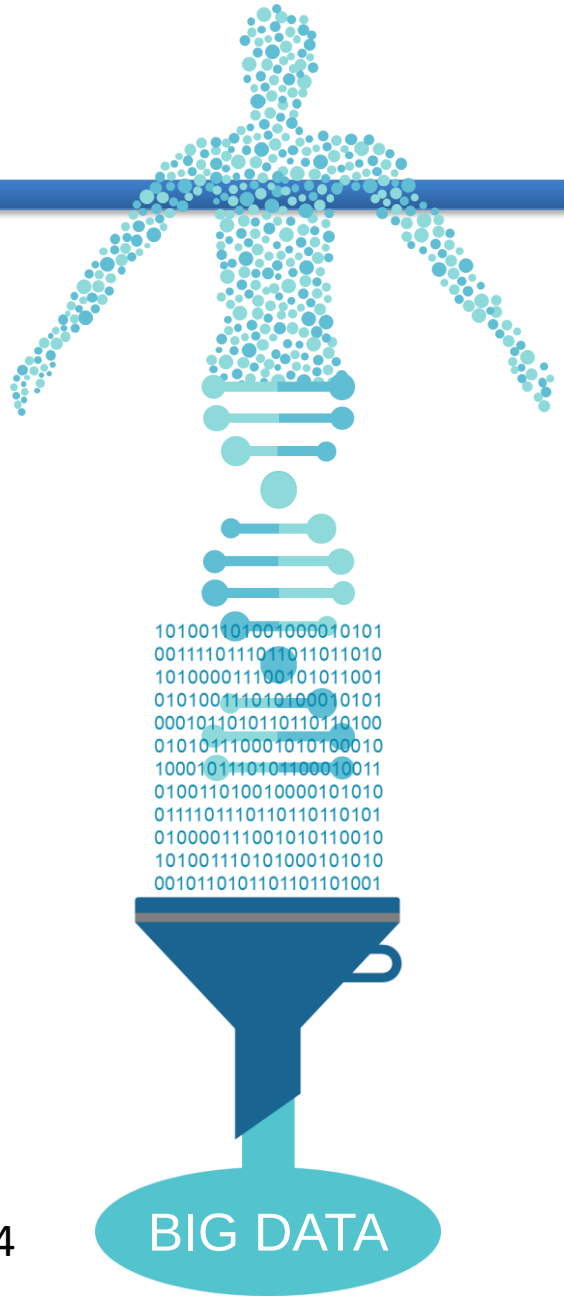
- Error de la base → Corrección con profundidad de lectura
- Formato fichero salida
- Software dedicado, universal fastq

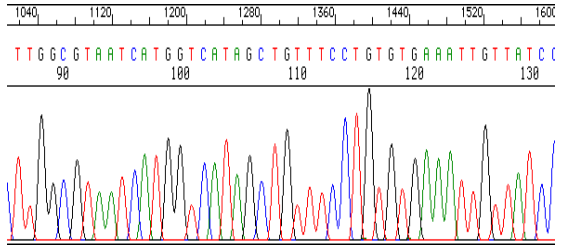
Coste actual de la secuenciación





BIG DATA


BIG DATA


1040
1120
1200
1280
1360
1440
1520
1600
TTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCC
90
100
110
120
130

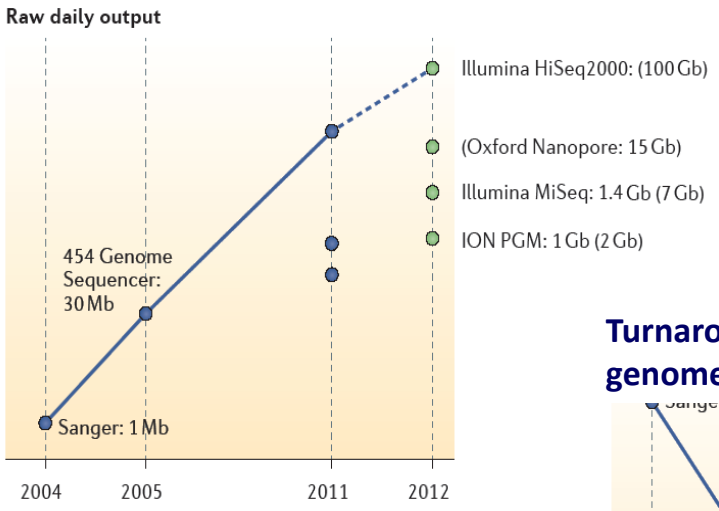

@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:2458:1027 1:N:0:ACAGTG
AGAAAAAACCTTGGANGGAAAAAAATCAGACATTTTCTAGAGGTGGAAGGCAAACTGAACAAAGAAATAATTCACA
+
DGGGEDHHHHGGGFE#CBACBCA<?HHHHBHHHHHHHHDHHHHEHEFEGGGGGG/GGDDDGHFHGFCHFHHEHEH8
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3082:1029 1:N:0:ACAGTG
GGTAATACAGACTGANATGATCAAAGGCATGCTGGAAACAAACCTATTAAAGATAAGCTTGGATCAAGCTTTCATT
+
B:B:?BB/:=55177#55877<775EDD>E=B?BBBBGGGDDAG@G>GGGGGG@)EEEEBEG>GGGGGGGAAA?<D
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3185:1033 1:N:0:ACAGTG
TCTGGGACATTGCTCNTGGCTGGGAGTCACCTGTCTGGGACATTGCTCAGGGCTGGGAGACACGTGTTGGAGGGAC
+
BC??A66;)74781<#7??;452.27'64(8,851DDG8GB?##############################
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3268:1033 1:N:0:ACAGTG
ATTCAAATTAGAAGANAGTTGATCGTTCTTCATGATGCCCAAAAATTTCACTGAGAAAACCCTTTTTTAAGCCCAC
+
IIIIIIIIIIFFFFE#ABACFEEFFIIGIIIFIHE@BIIIIIIIIIHHIIFIIF>HHIHIFGDIIIIIIGFHIEGH
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3400:1035 1:N:0:ACAGTG
TCCTGCTTTAGGAGANTCCTCATGCTCTGACAGGATGCTCTCTATGTGAGTTGAGCTGGTCTTCTCACTTTTATAG
+
IIIIIHIHIIGGEGG#AACA@?=?BHHIIIIIHHIHIIIHIHIHHGIHIHGHGIGIHGEGGGGHG@EFGGCEFAB
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3962:1033 1:N:0:ACAGTG
CCACCAACACAGTCTNCACCTTCTGTTGCTGGTGATAGATTTTTGCACCTTTCCATCCTCCAGGTTTCAAAATAGC
+
HHFHHDHDHH>C?CA#EEEE>?A?>HHDGHEGBGBCEEEEGHHF8HEHEEHECH,=>>==EAEE>BEBBAEAACAB
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:4491:1028 1:N:0:ACAGTG
AGAGAGAGAGAGAGANAGAGGACTCTGGAGATGCCGAAGCACAAGCCTGCAAGAGTCCCAGCAAAGAAAATAAAAA
+
GADGGEGGEGBBB?B#@=@@72;64GGGFGB>GGGBDG<DBGB<DA??/?##########################

Index

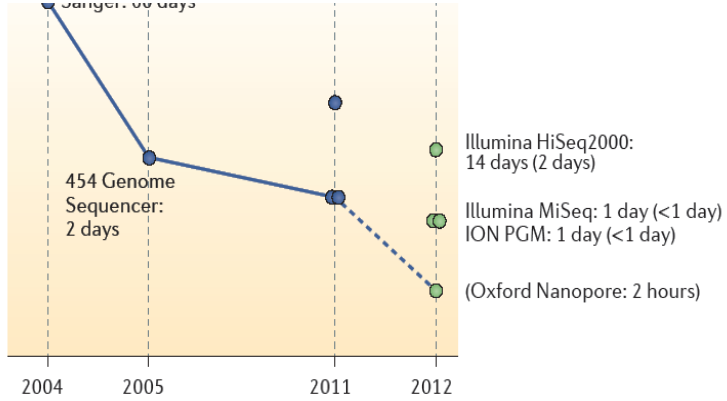
- BU-ISCIII
- High throughput sequencing platforms update
- Bacterial genome sequencing, brief history
- Advantages of WGS
- Use of WGS in Europe
- Library strategies
- Bioinformatics analysis

Sequencing platforms in Microbiology

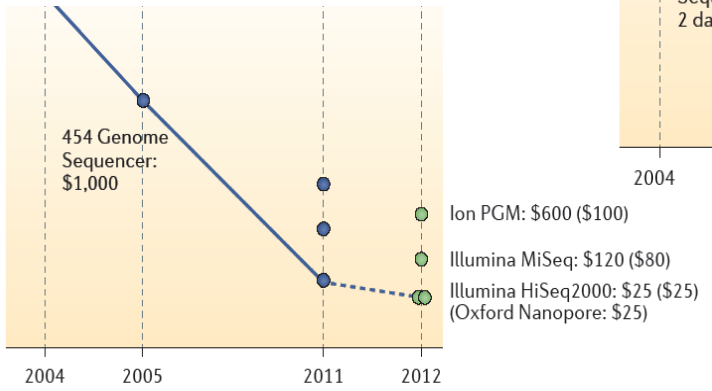
Raw daily output



Turnaround time: bacterial genome



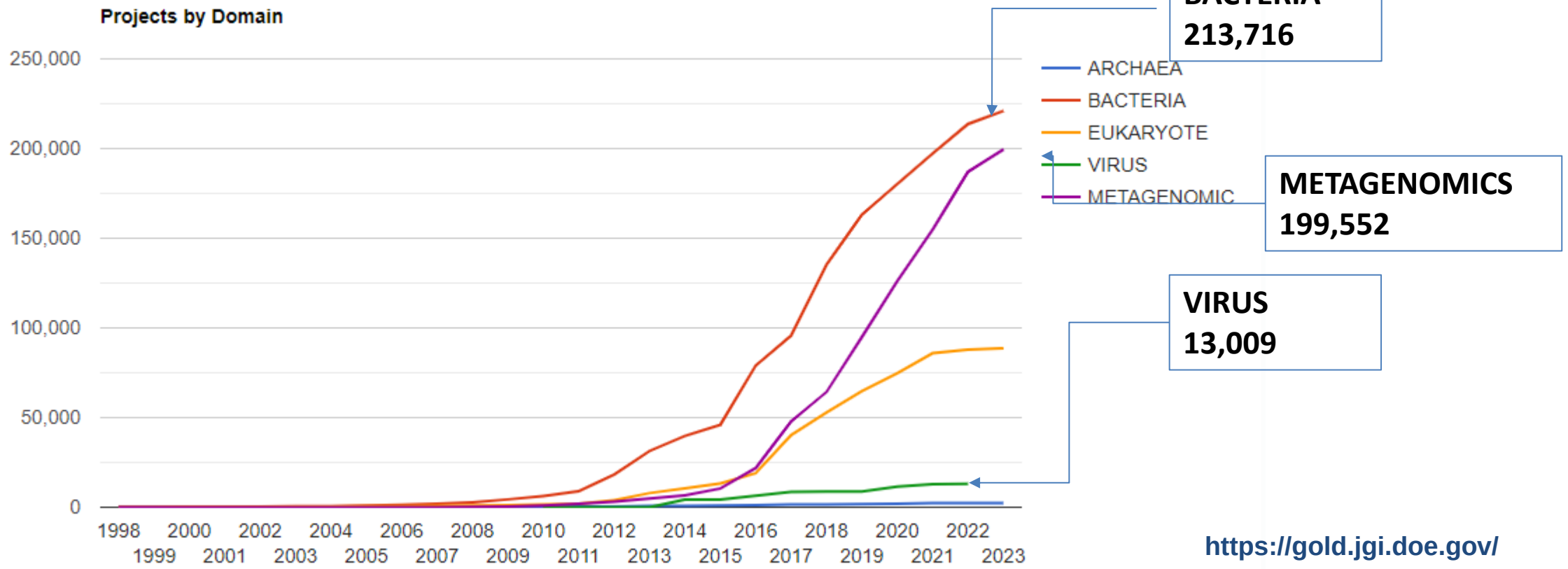
Cost per Mb assembled sequence



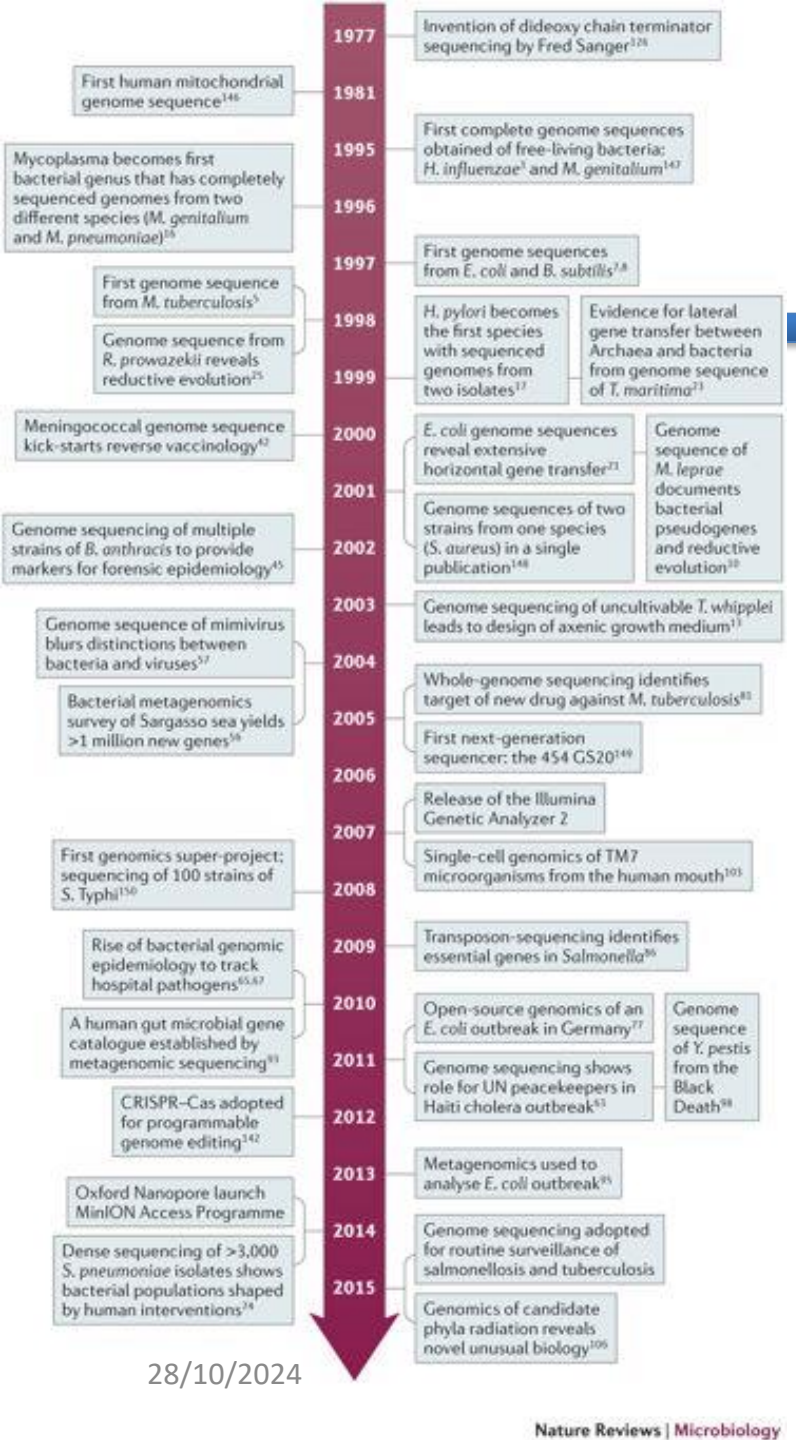
2

Sequencing projects

GOLD, Genome Online DataBase



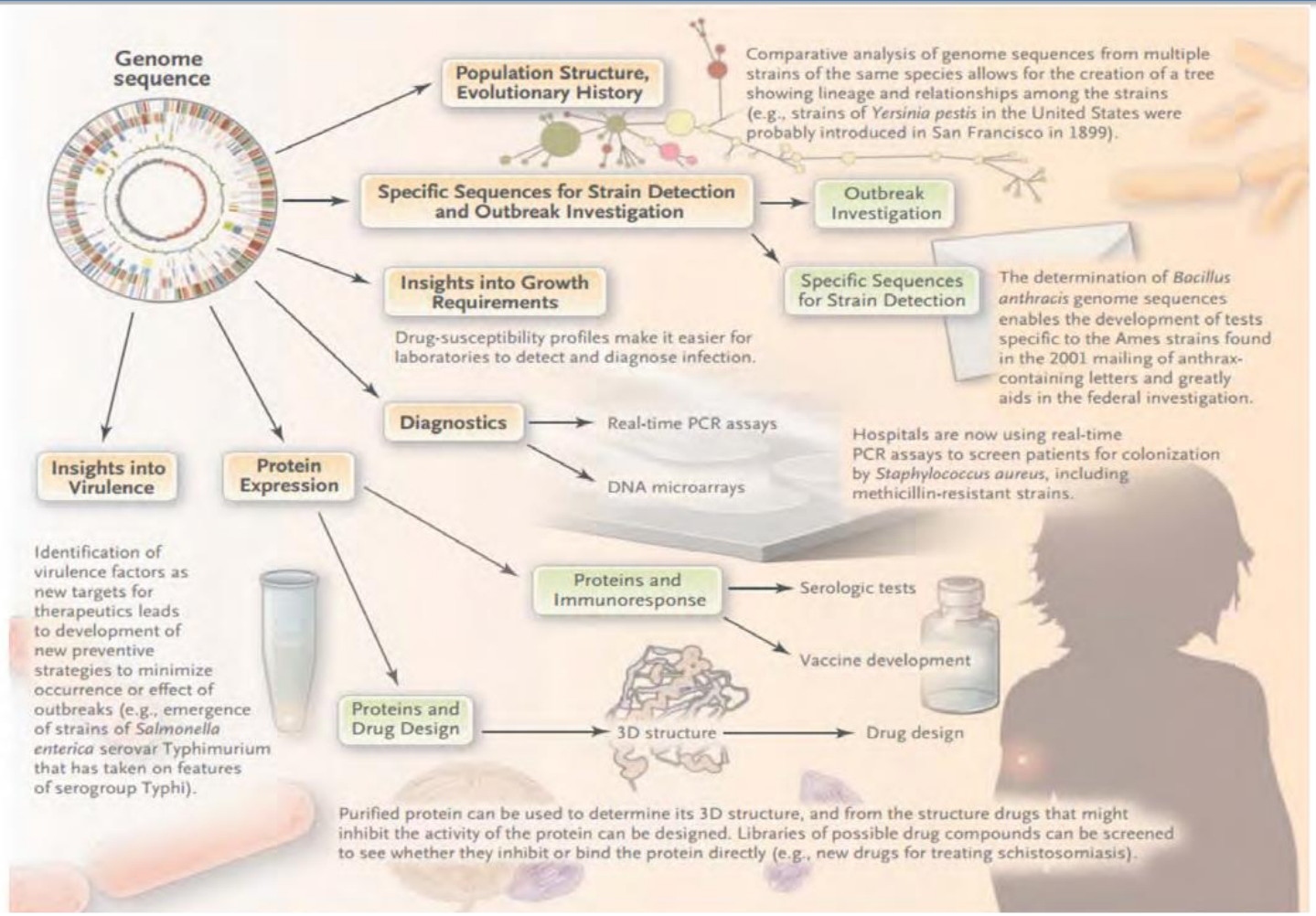
https://gold.jgi.doe.gov/



Brief history of the major events that have shaped the sequencing and analysis of bacterial genomes in the past two decades

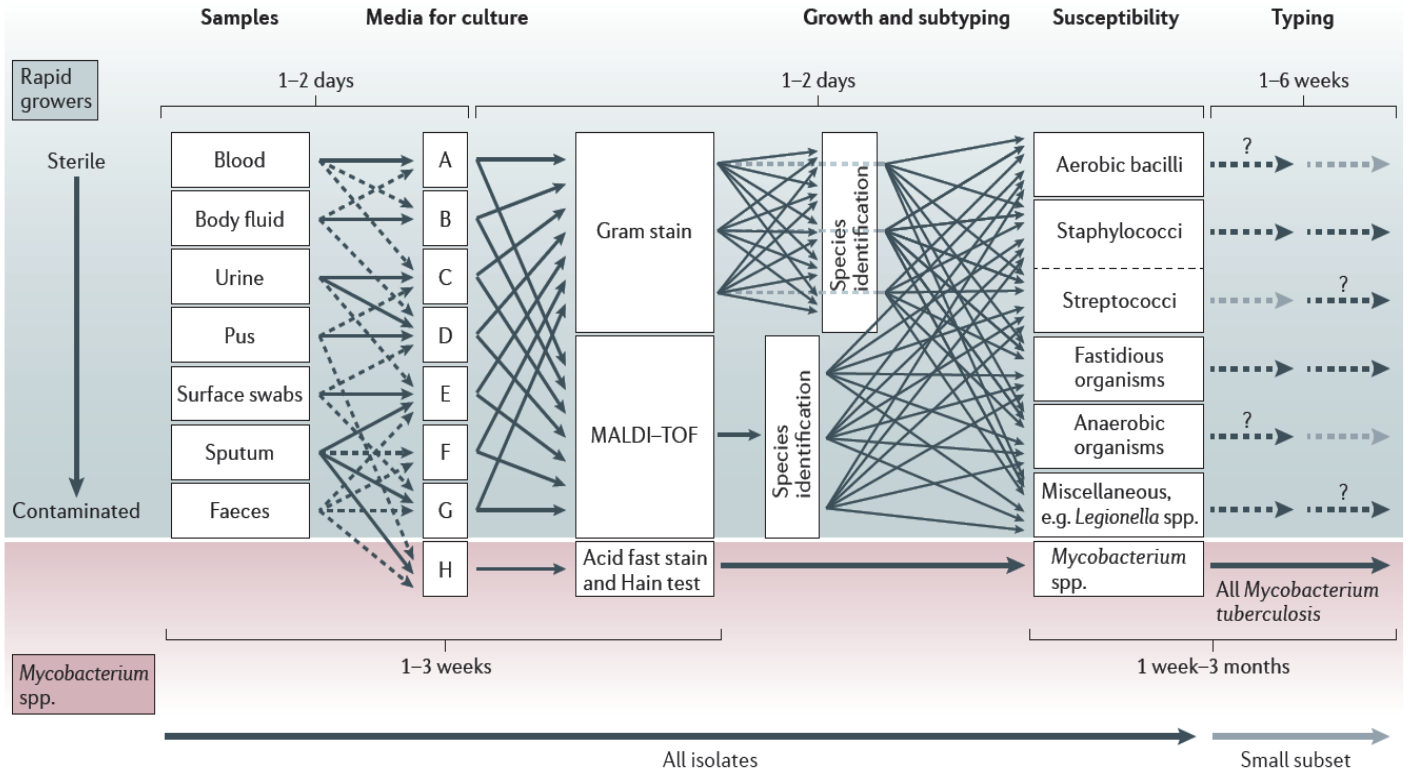
Loman et al., Nat Rev Microbiol 2015

Use of microbial genomics for tool development



Report from The American Academy of Microbiology, 2015

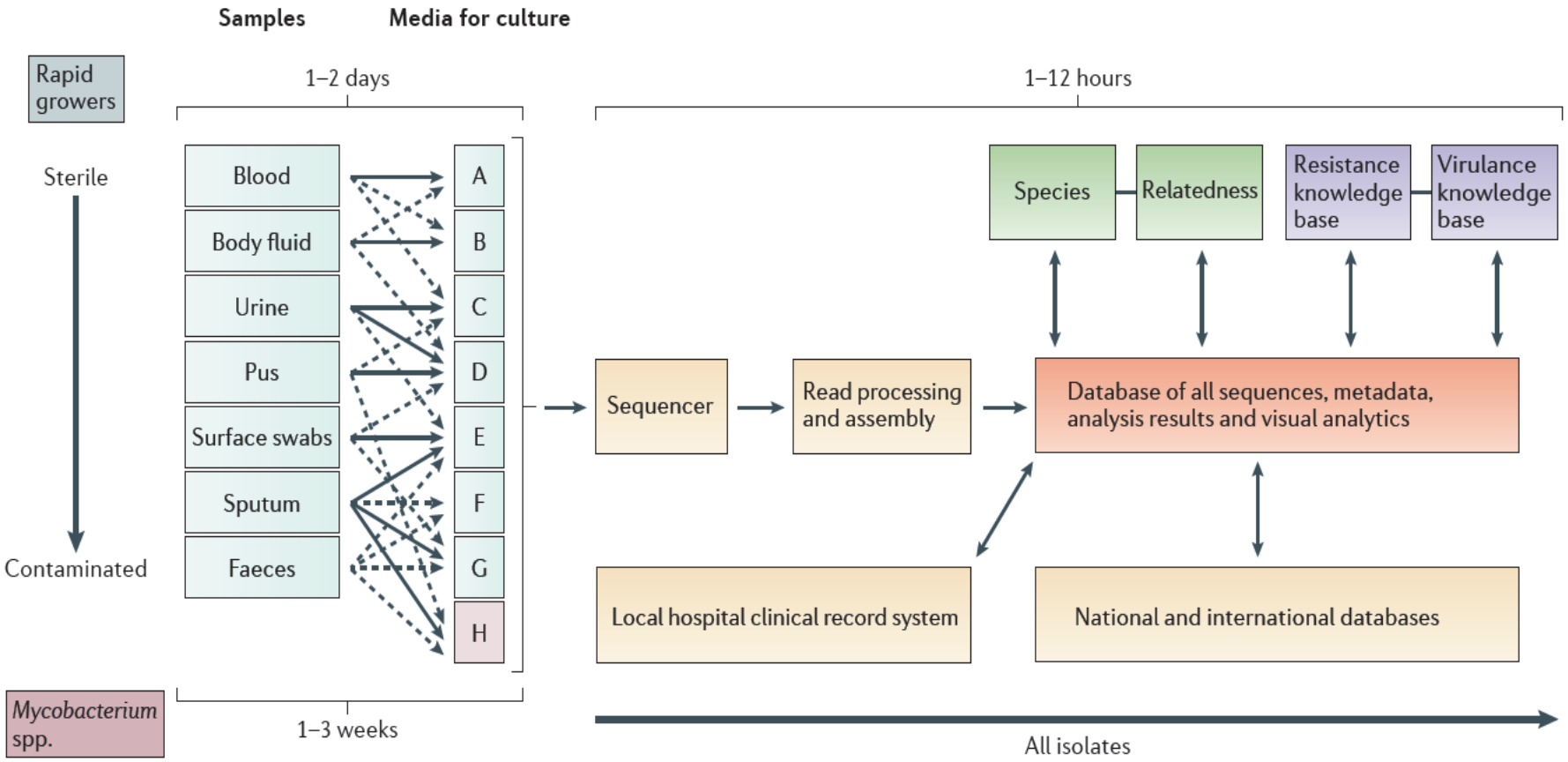
Workflow for processing samples for bacterial pathogens



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

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

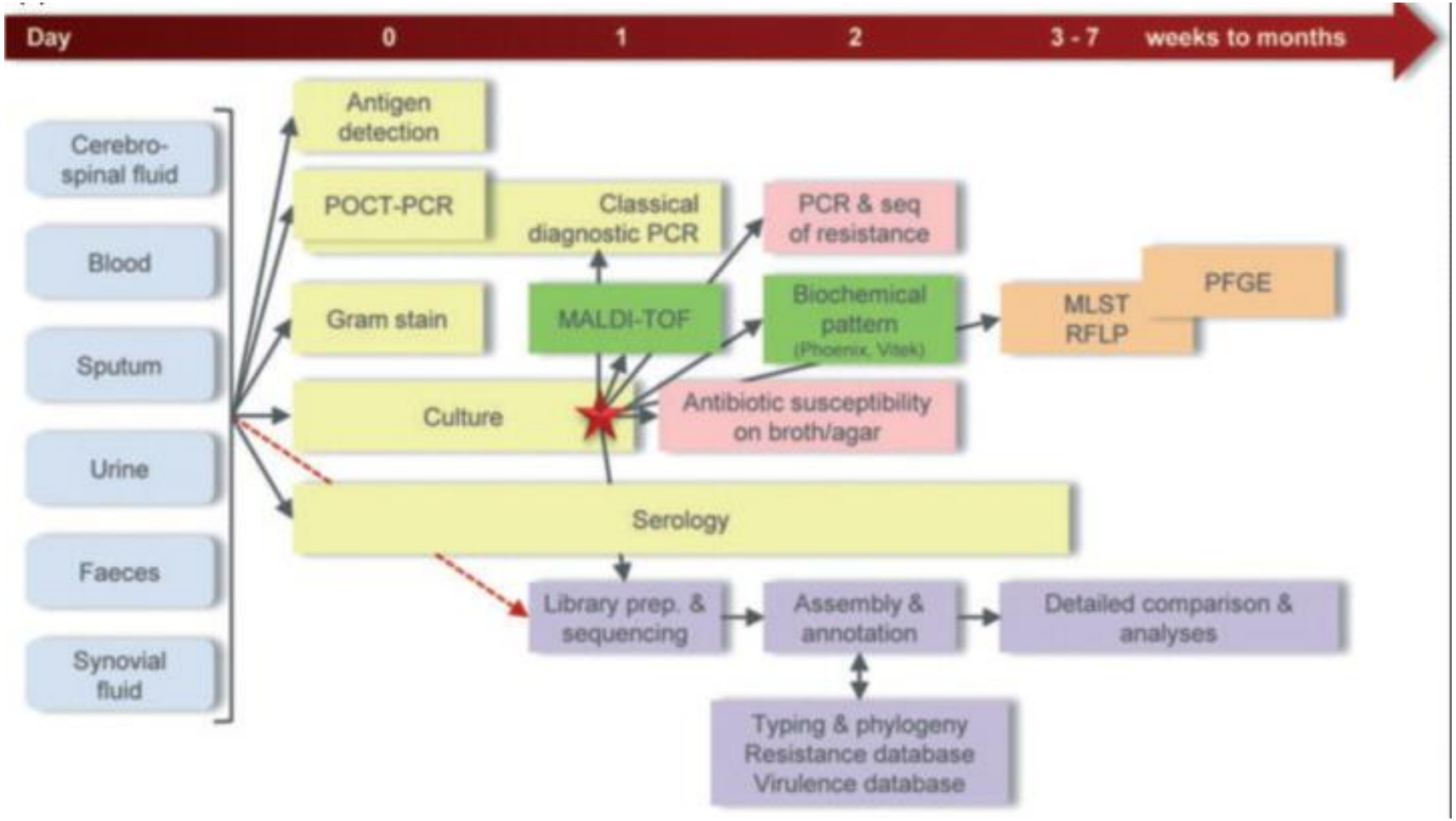
The diagnostic and clinical applications of bacterial WGS



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

Schematic representation of the timeline for the processing of clinical samples with classical pathogens

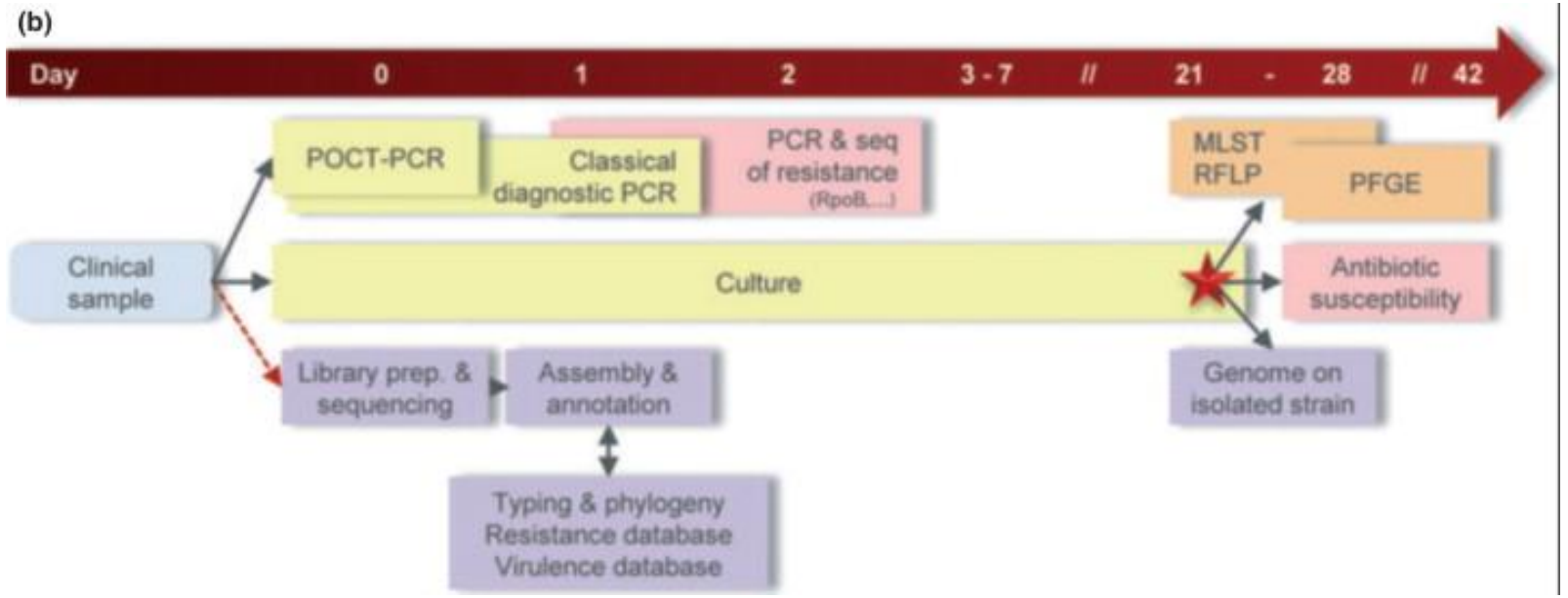
(a) classical pathogens



Bertelli and Greub, Clin Microb and Infect, 2013

Schematic representation of the timeline for the processing of clinical samples with classical pathogens

(b) slow-growing bacteria such as *Mycobacterium tuberculosis*



Bertelli and Greub, Clin Microb and Infect, 2013

Index

- BU-ISCIII
- High throughput sequencing platforms update
- Bacterial genome sequencing, brief history
- Advantages of WGS
- Use of WGS in Europe
- Library strategies
- Bioinformatics analysis

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

Bertelli and Greub, Clin Microb and Infect, 2013

Foodborne outbreak identification “Crisis del pepino”

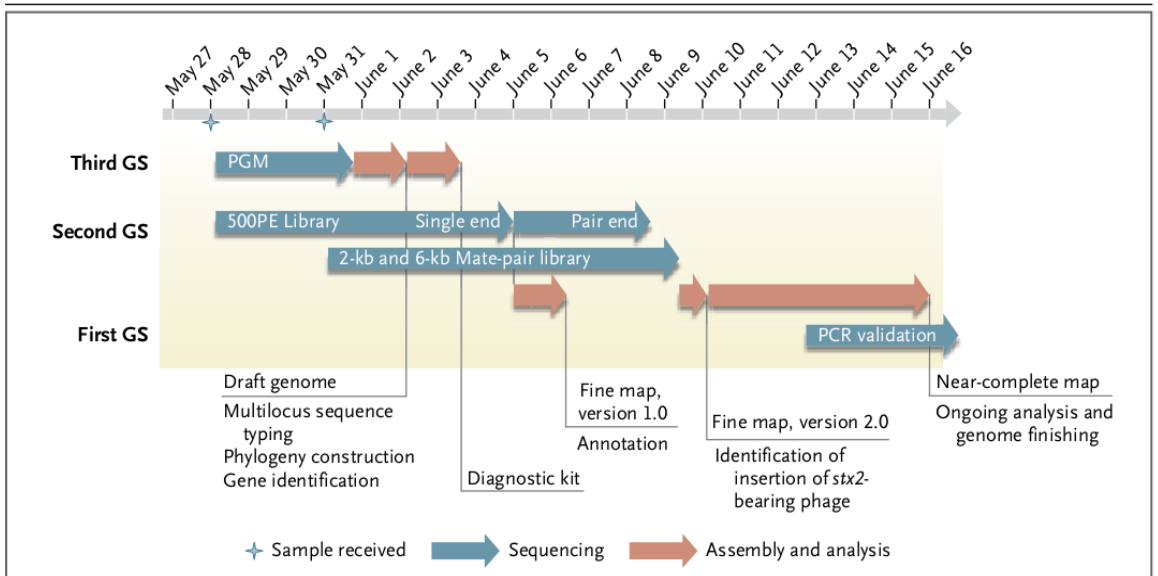


Figure 1. Timeline of the Open-Source Genomics Program.

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.

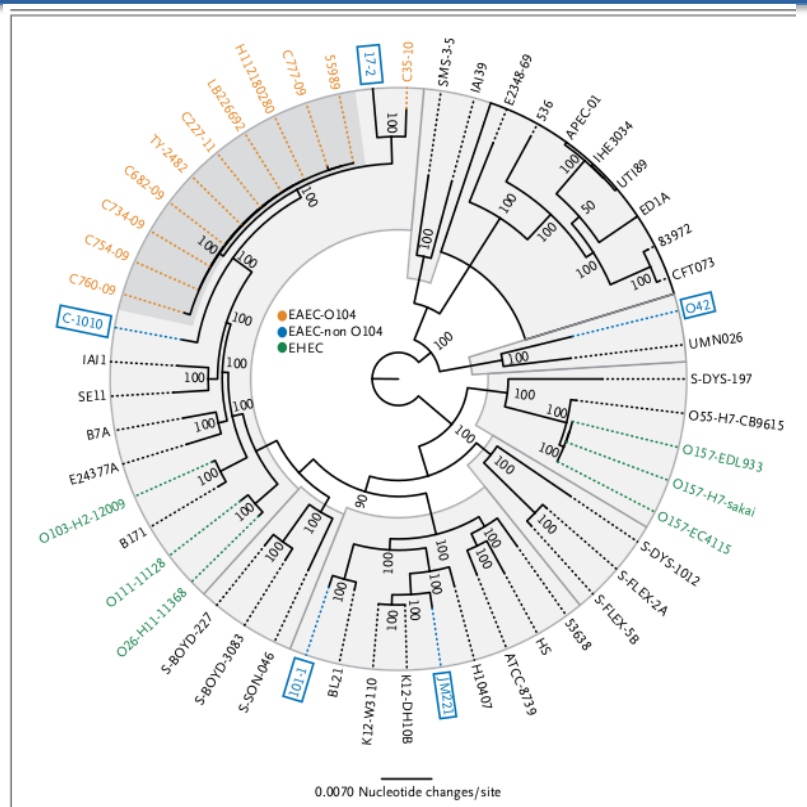


Figure 2. Phylogenetic Comparisons of 53 *Escherichia coli* and Shigella Isolates.

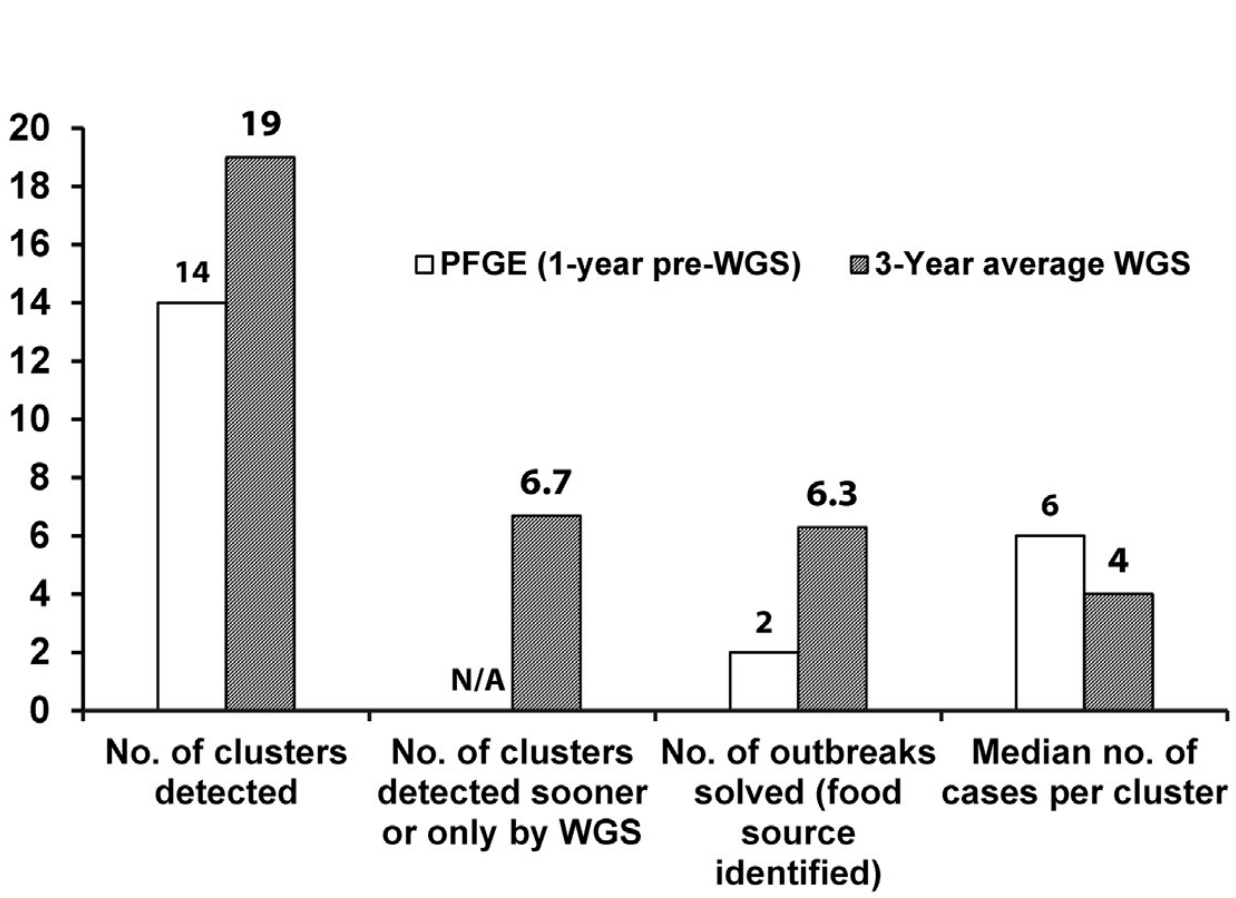
Genomic sequences were compared with the use of 100 bootstrap calculations, as described by Sahl et al.[35] The species-based phylogeny was inferred with the use of 2.56 Mbp of the conserved core genome. The O104:H4 isolates are shown in orange, the reference enteroaggregative *E. coli* (EAEC) isolates 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.

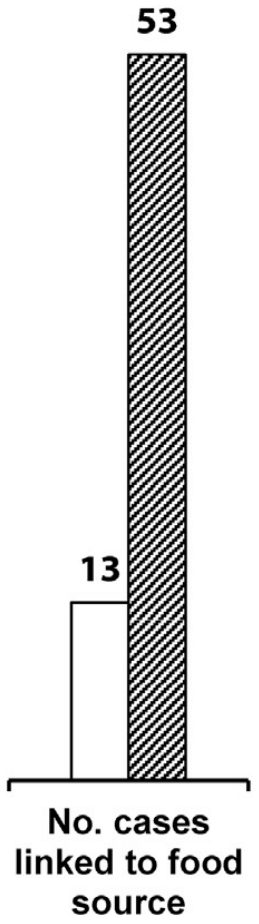
Rohde et al NEJM 2011, 365:718-24

The idea that via a ***One Health approach infectious diseases*** *could be better controlled and prevented*

Global Microbial Identifier

Early data from surveillance of listeriosis in the USA





The number of outbreaks detected increased 36% after implementation of real-time WGS based surveillance, and likewise the number of solved outbreaks increased more than three-fold

Besser et al., Clin Micr Infect, 2018

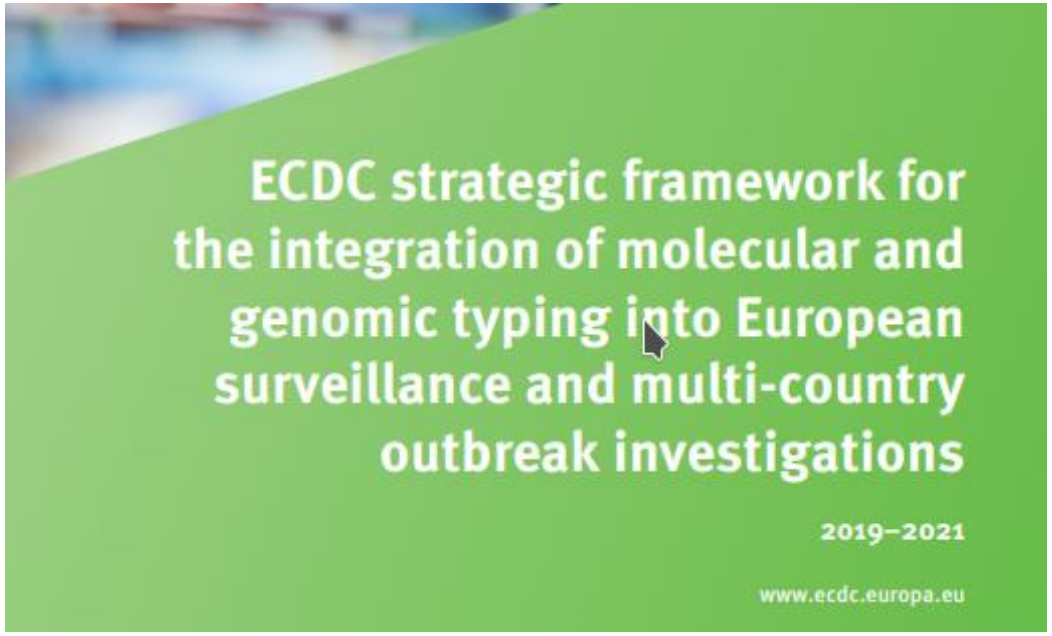
ECDC technical report: Monitoring the use of wgs in infectious disease surveillance in Europe 2015-2017

WGS provides higher resolution and accuracy than classical molecular typing methods, such as PFGE or MLVA, contributing to a better understanding of infectious disease and drug resistance transmission patterns and thereby improving the effectiveness of interventions for their control.

Index

- BU-ISCIII
- High throughput sequencing platforms update
- Bacterial genome sequencing, brief history
- Advantages of WGS
- Use of WGS in Europe
- Library strategies
- Bioinformatics analysis

ECDC roadmap and international commitment

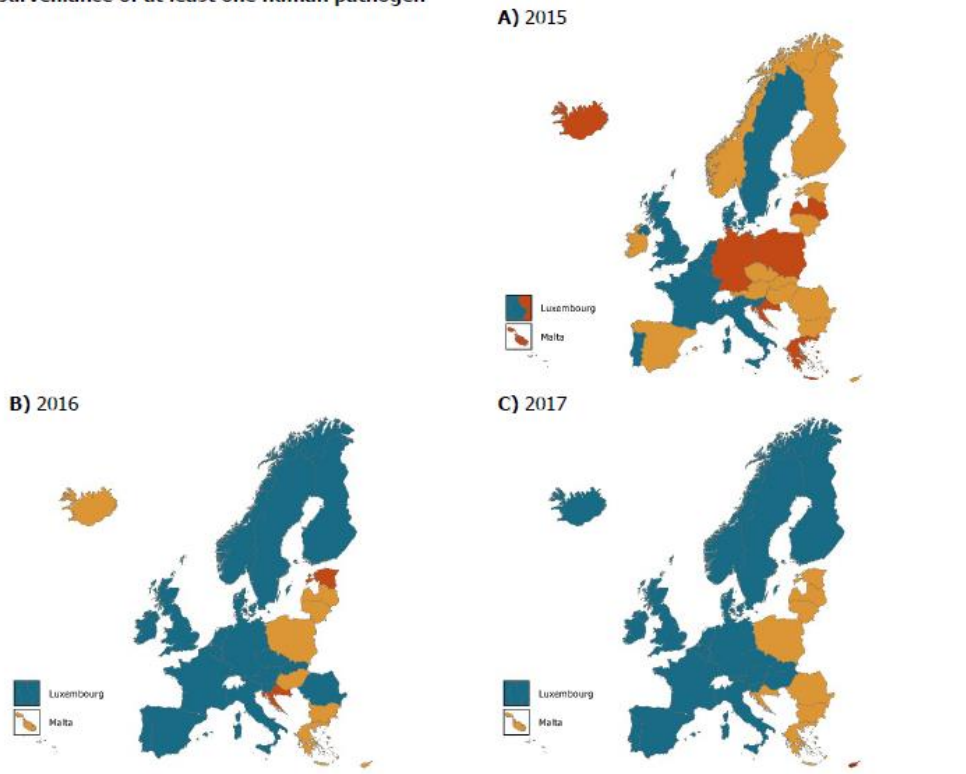


2018

- **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*;

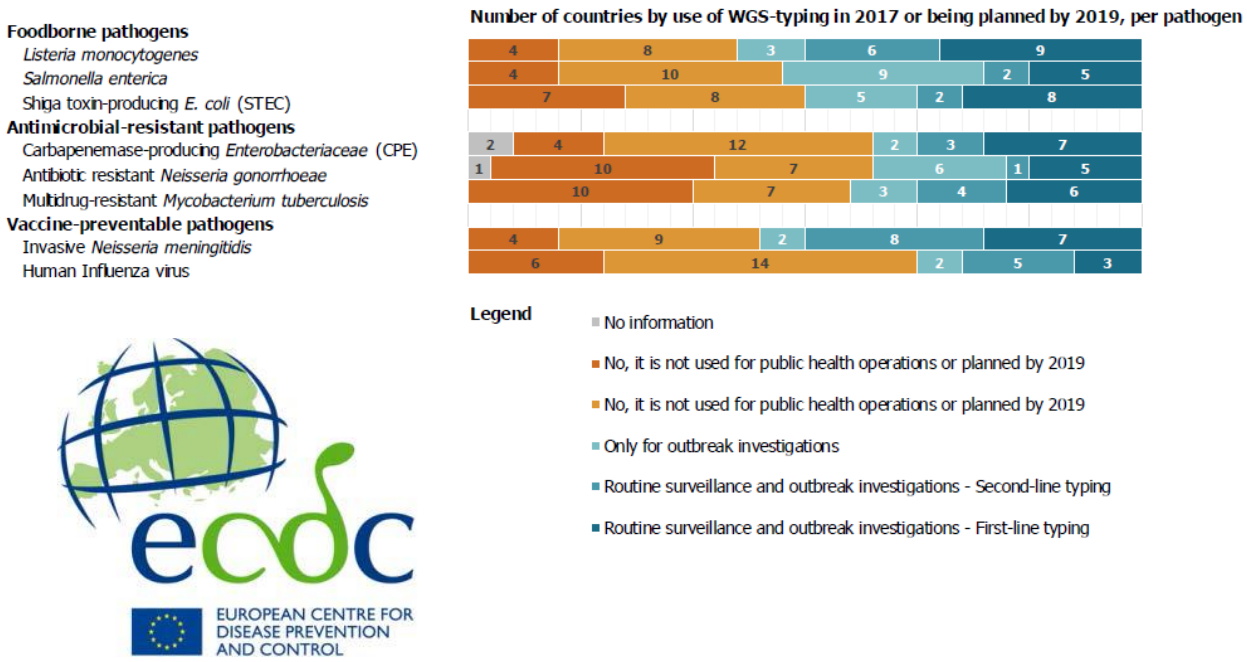
ECDC technical report: Monitoring the use of WGS in infectious disease surveillance in Europe 2015-2017

Figure 1. National public health reference laboratories use of WGS-based typing for national surveillance of at least one human pathogen



Legend: blue: in use; light orange: a plan in place/in progress within three years; dark orange: no use or national plan in place within three years. EU/EEA countries, 2015–2017

Figure 3. Number of EU/EEA countries using WGS-based typing as first or second-line method for routine surveillance and outbreak investigations in National Public Health Reference Laboratories by disease group and pathogen, 2017



ECDC technical report: Monitoring the use of WGS in infectious disease surveillance in Europe 2015-2017

		Foodborne pathogens			Antimicrobial-resistant pathogens			Vaccine-preventable diseases	
NGS technology	**Instrument**	*L. monocytogenes*	*S. enterica*	STEC	CPE	AR-*N. gonorrhoeae*	MDR-TB	*N. meningitidis*	Human Influenza virus
Illumina	HiSeq series	3	3	3	2	1	2	3	1
	HiSeq X series						1		
	MiniSeq	1	2	3	2		4		1
	MiSeq series	12	10	7	7	10	7	13	7
	NextSeq	2	2	2	3	1	3	2	2
Ion Torrent	S4								1
	S5	1	1	1	1		1		
	S5 XL								1
	PGM		1	1		1		1	1
	Proton							1	1
Oxford Nanopore Technologies	MinION		1	1	2		1	1	1
Pacific Biosciences-PacBio	PacBio RS II		1		2				
	Sequel		1	1					
Other not specified	-	3	2	2	1	1	1	1	

Table 1. Number of EU/EEA countries with one or more national public health reference laboratories having access to next-generation sequencing (NGS) technologies for routine public health operations, by technology and instruments used, 2017

ECDC technical report: Monitoring the use of WGS in infectious disease surveillance in Europe 2015-2017

Table 2. Bioinformatics tools used by the National Public Health Reference Laboratories using WGS-based typing for surveillance and outbreak investigations of foodborne pathogens, July 2017*

	Number of EU/EEA countries		
Tools used for sequence analysis	*L. monocytogenes* (n=14)	*S. enterica* (n=7)	STEC (n=9)
Commercial software	9	4	4
Open source software	4	3	5
In-house suite of customised tools	4	2	2

* Not mutually exclusive

ECDC technical report: Monitoring the use of WGS in infectious disease surveillance in Europe 2015-2017

	2017	Foodborne pathogens			Antimicrobial resistant pathogens			Vaccine preventable pathogens	
		L. monocytogenes	*S. enterica*	STEC	CPE	AR-*N. gonorrhoeae*	MDR TB	Human influenza virus	*N. meningitidis*
Number of countries using WGS for routine surveillance and outbreak investigations		14	7	9	10	6	10	8	15
Typing scheme	First-line WGS	9	5	8	7	5	6	3	7
	Second-line WGS	5	2	1	3	1	4	5	8
Sampling frame	Continuous comprehensive	12	6	8	7	2	9	-	15
	Sentinel/ subset of case samples	2	1	1	3	4	1	8	-
Bioinformatic analysis *	cgMLST	12	6	5	6	4	5	-	12
	SNP	7	5	5	5	2	7	-	5
	Resistome prediction	4	5	7	8	4	6	-	3
	wgMLST	5	3	3	2	2	2	-	2
	Virulome/ mobilome prediction	4	2	9	5	1	-	-	1
	MLST prediction	12	6	8	3	-	-	-	2
	Serogroup prediction	7	6	9	1	-	-	-	2
	NG-MAST	-	-	-	-	3	-	-	-
	Speciation	-	1	1	-	-	3	-	1
	Hemagglutinin and neuraminidase sequence prediction	-	-	-	-	-	-	4	-
	Phylogenetic relationship	-	1	1	1	-	-	7	1
	Identification of specific point mutations	-	1	1	-	-	-	6	1
	rMLST	-	-	-	-	-	-	-	5
	MLST+*porA* VR1 and VR2+*fetA*	-	-	-	-	-	-	-	12
	Vaccine antigen prediction	-	-	-	-	-	-	-	9
	Other not specified	-	-	-	1	-	3	3	1
Raw sequence data storage *	Dedicated closed database(s)	13	5	7	10	6	10	6	12
	Publicly available database(s)	1	2	2	-	1	1	2	3

** Not mutually exclusive*

Table 3. Number of EU/EEA countries using WGS-based typing for surveillance and outbreak investigations in the national public health reference laboratories and respective typing scheme, sampling frame, bioinformatics analysis, and raw data storage practice by pathogen 2017

ECDC technical report: Monitoring the use of WGS in infectious disease surveillance in Europe 2015-2017

Conclusions

This emerging mainstream practice should enable **pan-European WGS-derived data exchange in the medium-term**, subject to **harmonisation** of sequence analysis pipelines for output compatibility, agreement on international WGS derived type nomenclature and development of **secure and efficient international data sharing** and management platforms.

Current bottlenecks mainly relate to development of expertise in **epidemiological-WGS data integrative analysis** and access to user-friendly international nomenclature

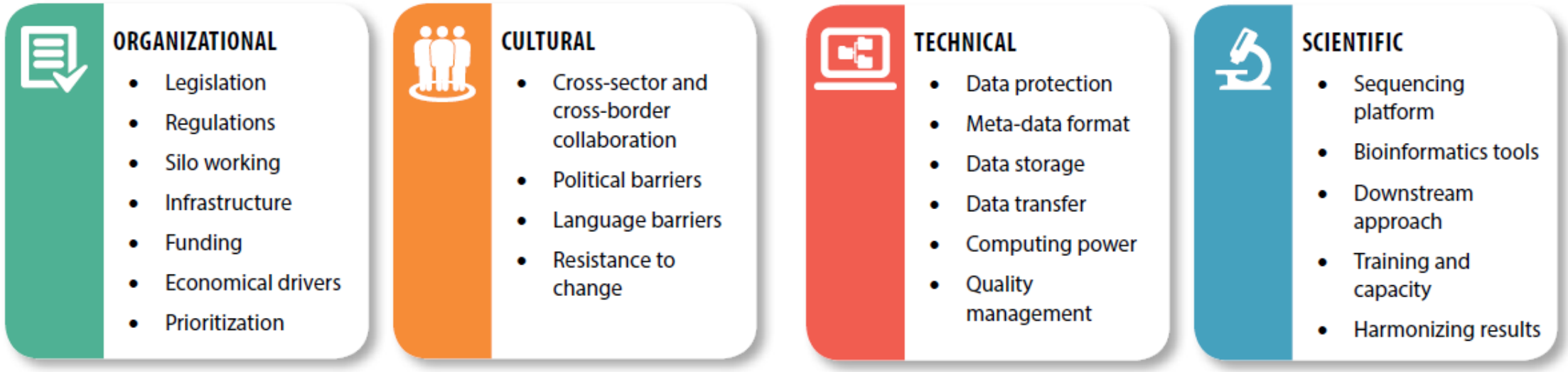
Skills needed to translate WGS data into public health action

Bioinformatician	Epidemiologist	Microbiologist
Algorithms for genome mapping, assembly and comparisons	Epidemiology of communicable diseases	Microbiological diagnostics
Inferences from genomic data	Statistical analysis	Subtyping of pathogens
Genomic data handling and processing	Case-control studies	Pathogen genomics and evolution
Genome data visualization and integration	Health data linkage	Access to culture collections with epidemiological context
	Risk assessment and communication	

WHO, landscape paper: WGS for foodborne disease surveillance

FIGURE 2.1

Challenges of coordinating WGS for integrated food chain surveillance



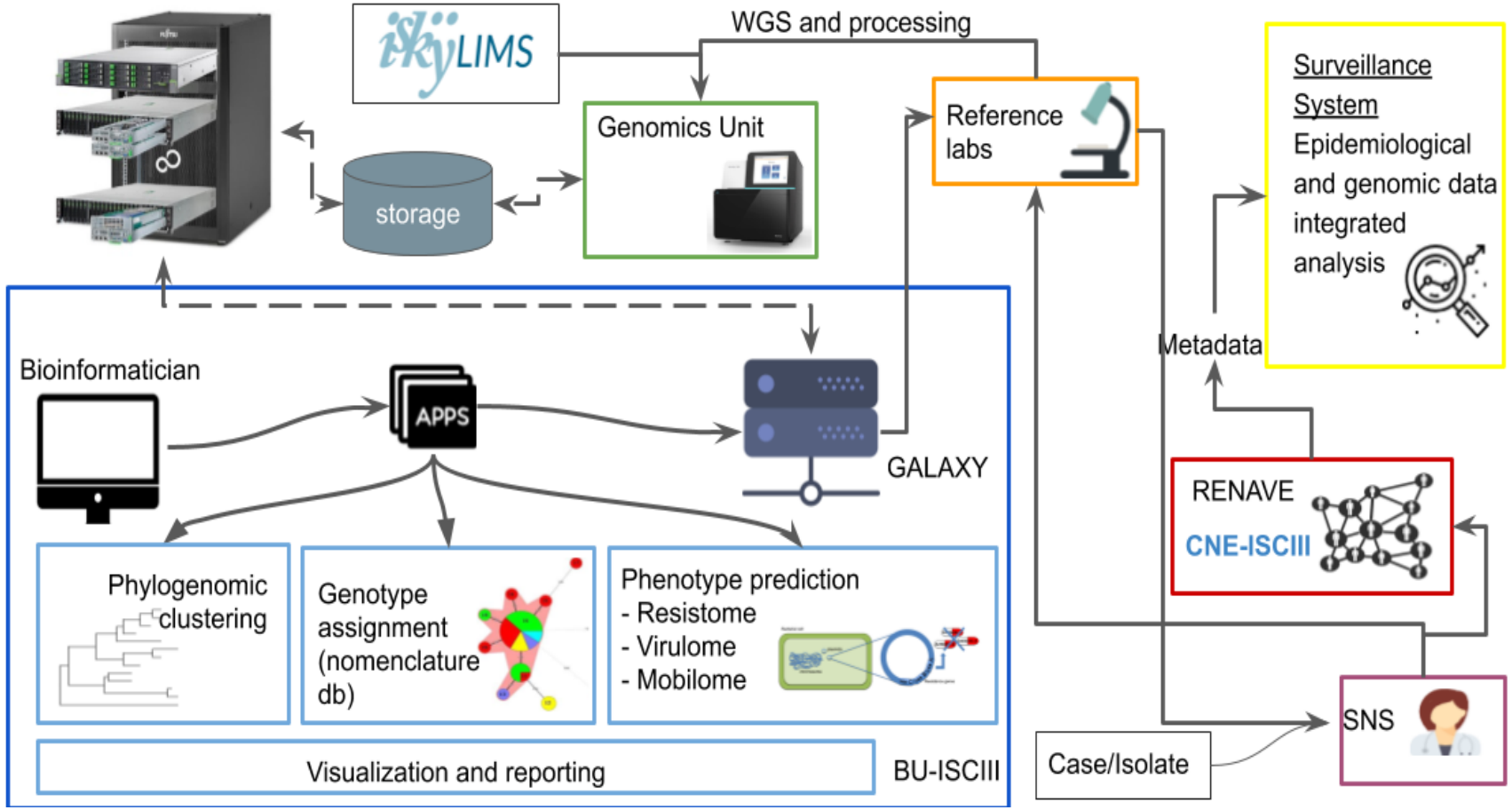
World Health Organization

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

Bioinformatics Unit preparedness



- Infrastructure
- Data
- Sample tracking
- Sharing and intercommunication
- Bioinformatics analysis

Adapted and extended from Expert opinion on whole genome sequencing for public health surveillance. ECDC

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

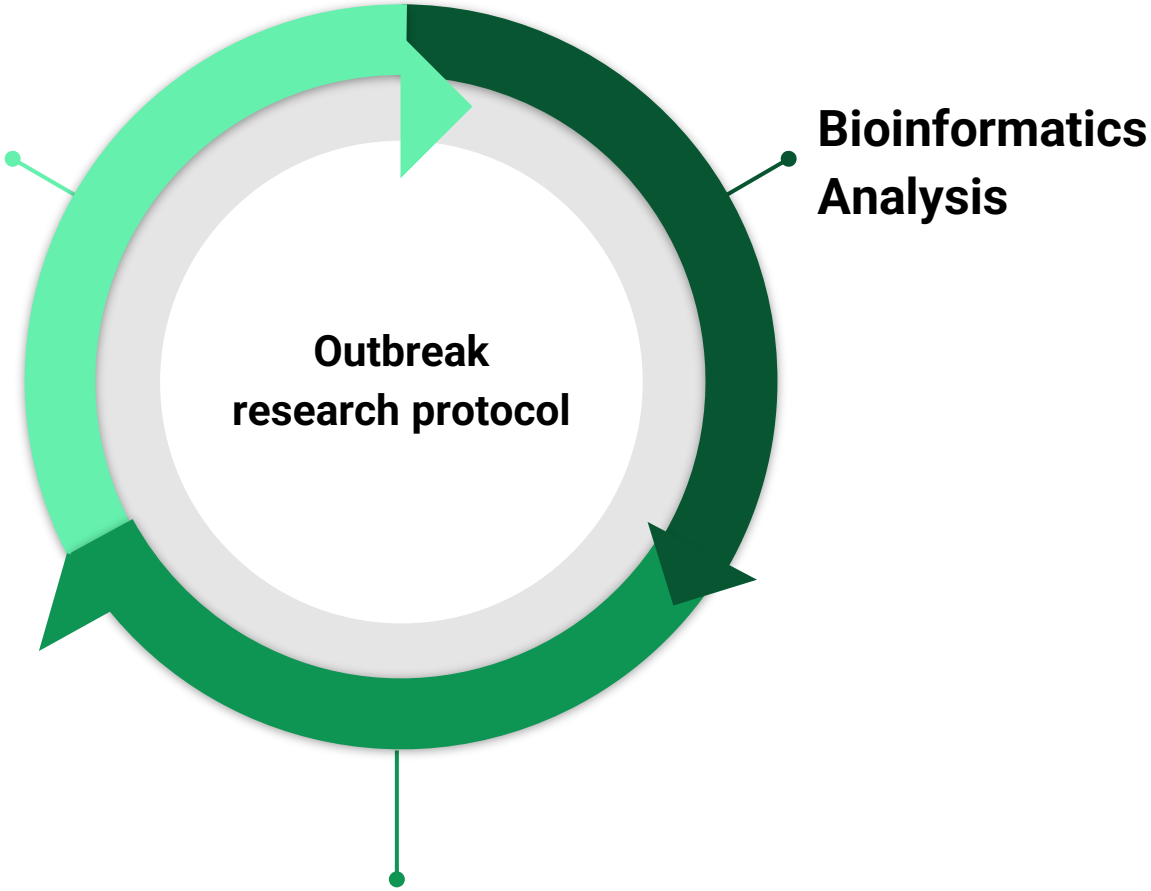
- Tres nuevos casos, en Sevilla y Cádiz, dejan el número de personas afectadas en Andalucía en 192.
- La carne con listeria de la marca blanca se vendió en los municipios de Sevilla.
- La empresa que vendió la marca blanca de Magrudis dice que cumple los protocolos.

- 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

Transversal:

- Data management
- Sample tracking
- Infrastructure

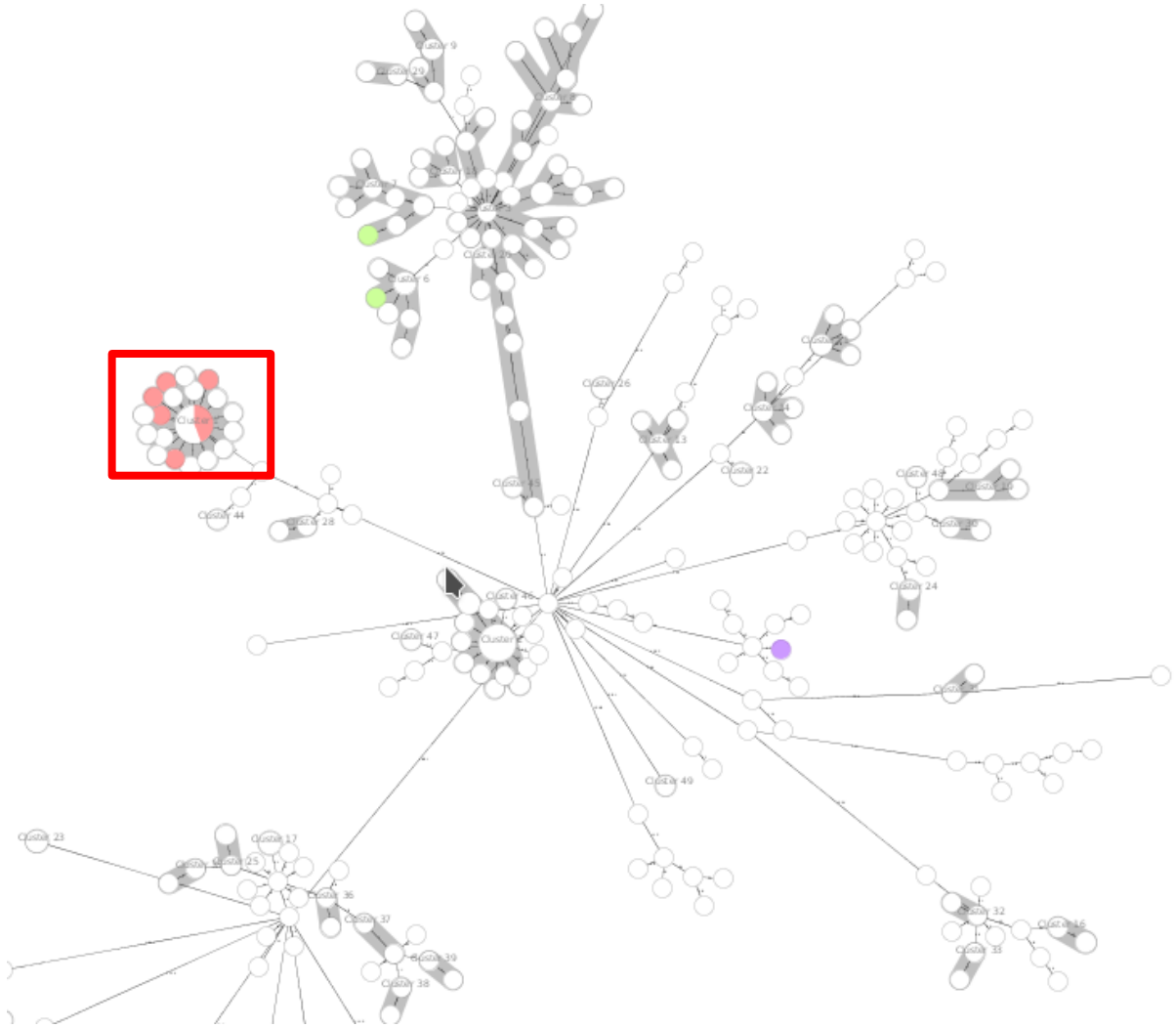


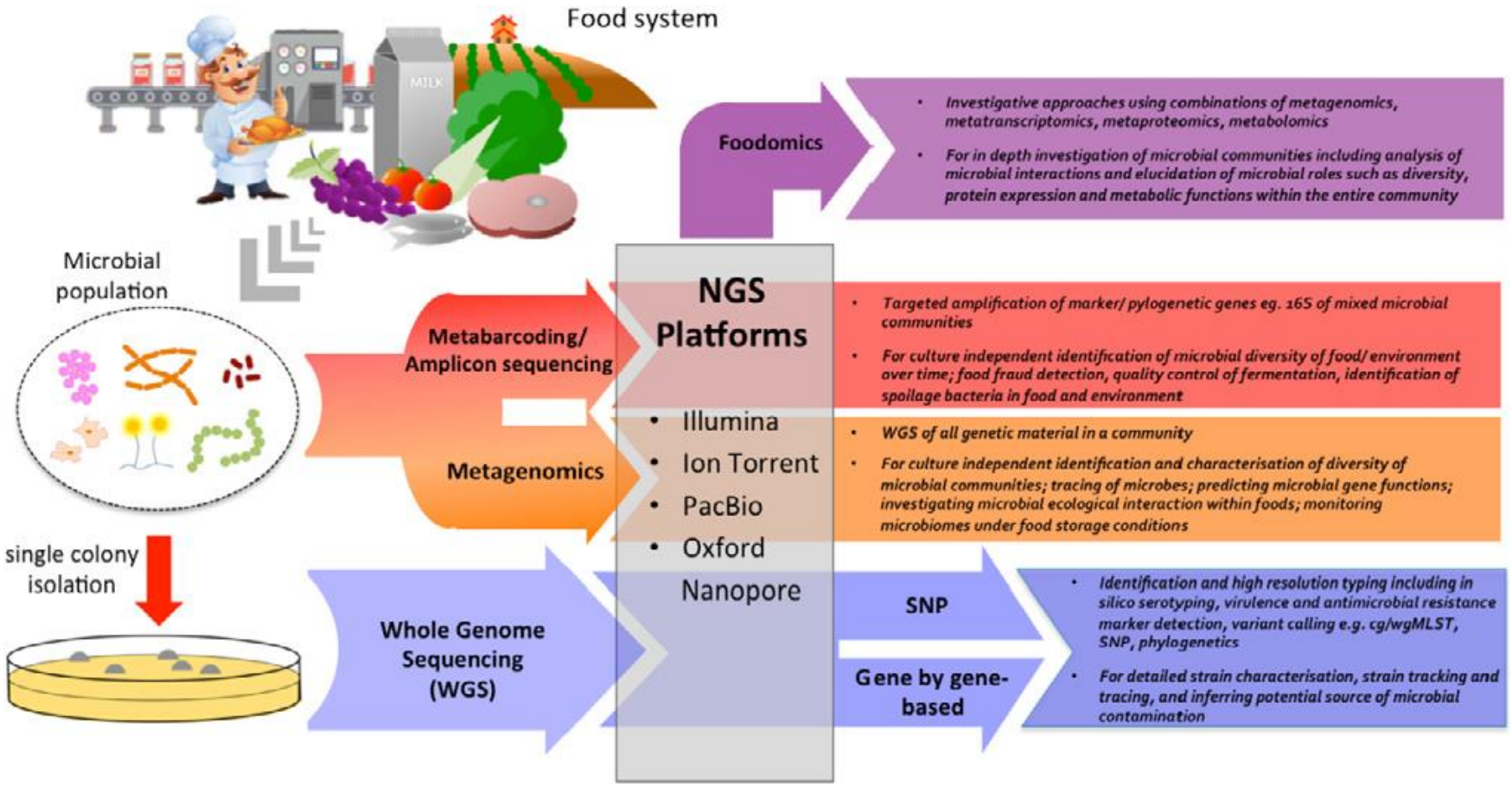
Andalusian Listeria Outbreak

- 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.

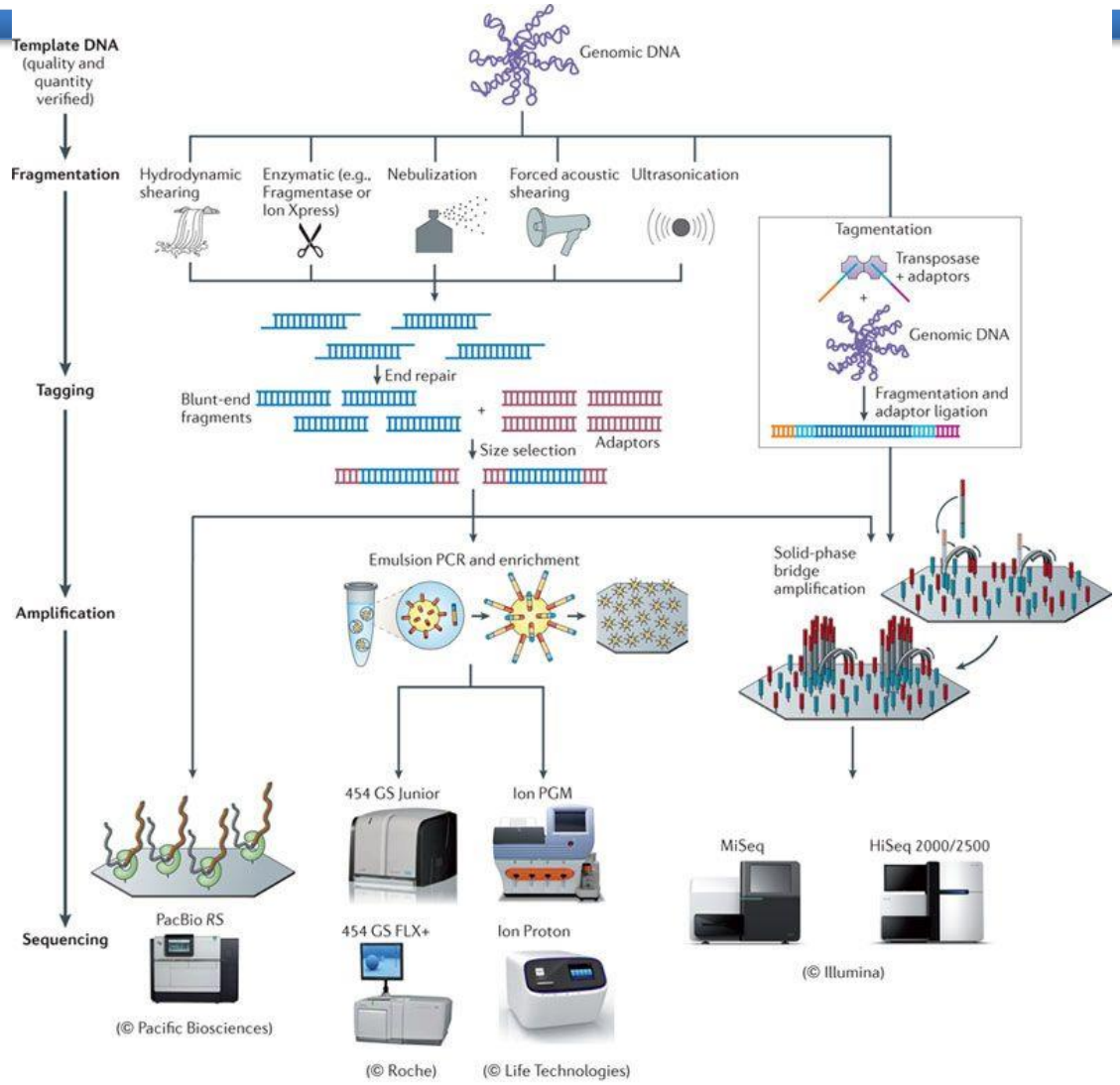



Food system
Microbial population
single colony isolation
Foodomics
Investigative approaches using combinations of metagenomics, metatranscriptomics, metaproteomics, metabolomics
For in depth investigation of microbial communities including analysis of microbial interactions and elucidation of microbial roles such as diversity, protein expression and metabolic functions within the entire community
Metabarcoding/ Amplicon sequencing
Metagenomics
NGS Platforms
Illumina
Ion Torrent
PacBio
Oxford Nanopore
Targeted amplification of marker/ pylogenetic genes eg. 16S of mixed microbial communities
For culture independent identification of microbial diversity of food/ environment over time; food fraud detection, quality control of fermentation, identification of spoilage bacteria in food and environment
WGS of all genetic material in a community
For culture independent identification and characterisation of diversity of microbial communities; tracing of microbes; predicting microbial gene functions; investigating microbial ecological interaction within foods; monitoring microbiomes under food storage conditions
Whole Genome Sequencing (WGS)
SNP
Gene by gene-based
Identification and high resolution typing including in silico serotyping, virulence and antimicrobial resistance marker detection, variant calling e.g. cg/wgMLST, SNP, phylogenetics
For detailed strain characterisation, strain tracking and tracing, and inferring potential source of microbial contamination

Index

- BU-ISCIII
- High throughput sequencing platforms update
- Bacterial genome sequencing, brief history
- Advantages of WGS
- Use of WGS in Europe
- **Library strategies**
- Bioinformatics analysis

High-throughput sequencing platforms



Loman et al, 2012

Preparación Librerías: 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

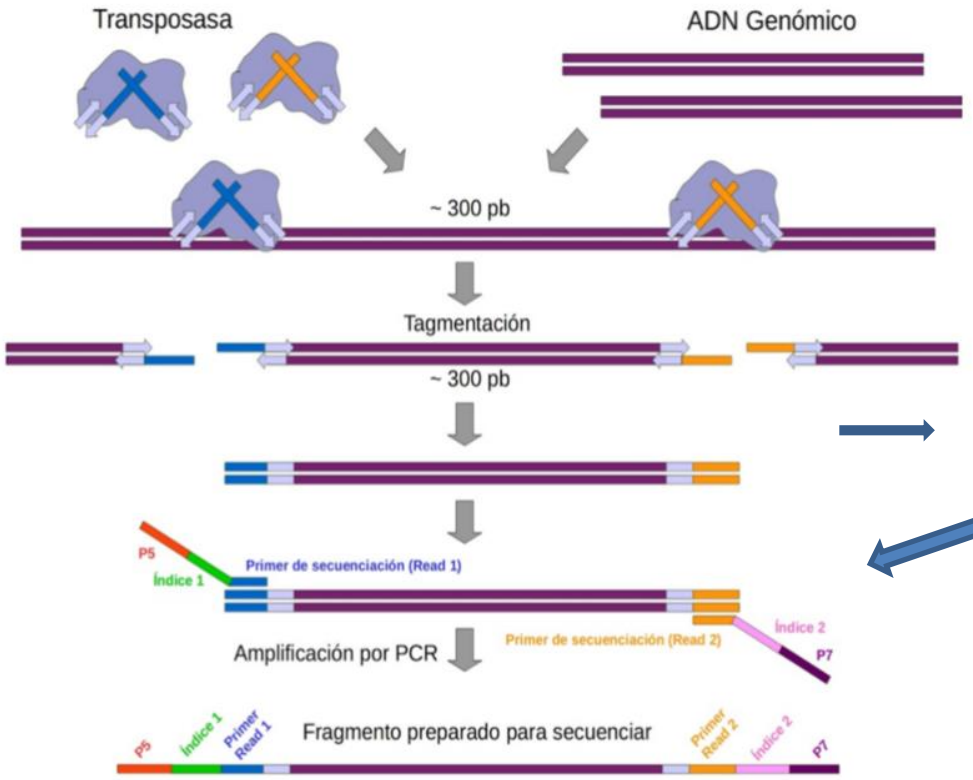
1. Metagenómica

Identificación de microorganismos

1. Metataxonomía

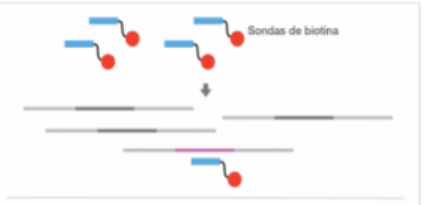
Preparación Librerías

ENZIMÁTICA FÍSICA



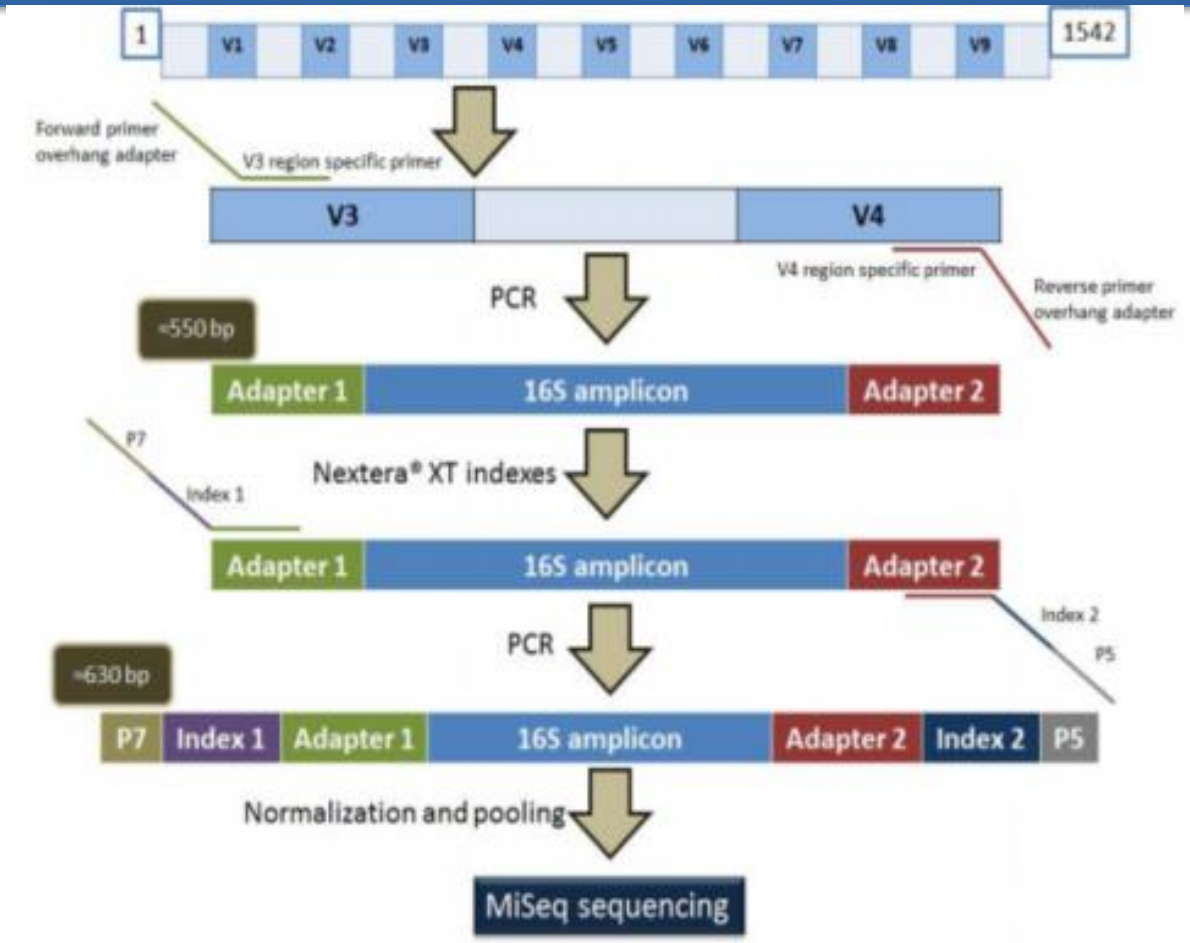
RNA -> cDNA

ENRIQUECIMIENTO:
PCR
CAPTURA SONDAS



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

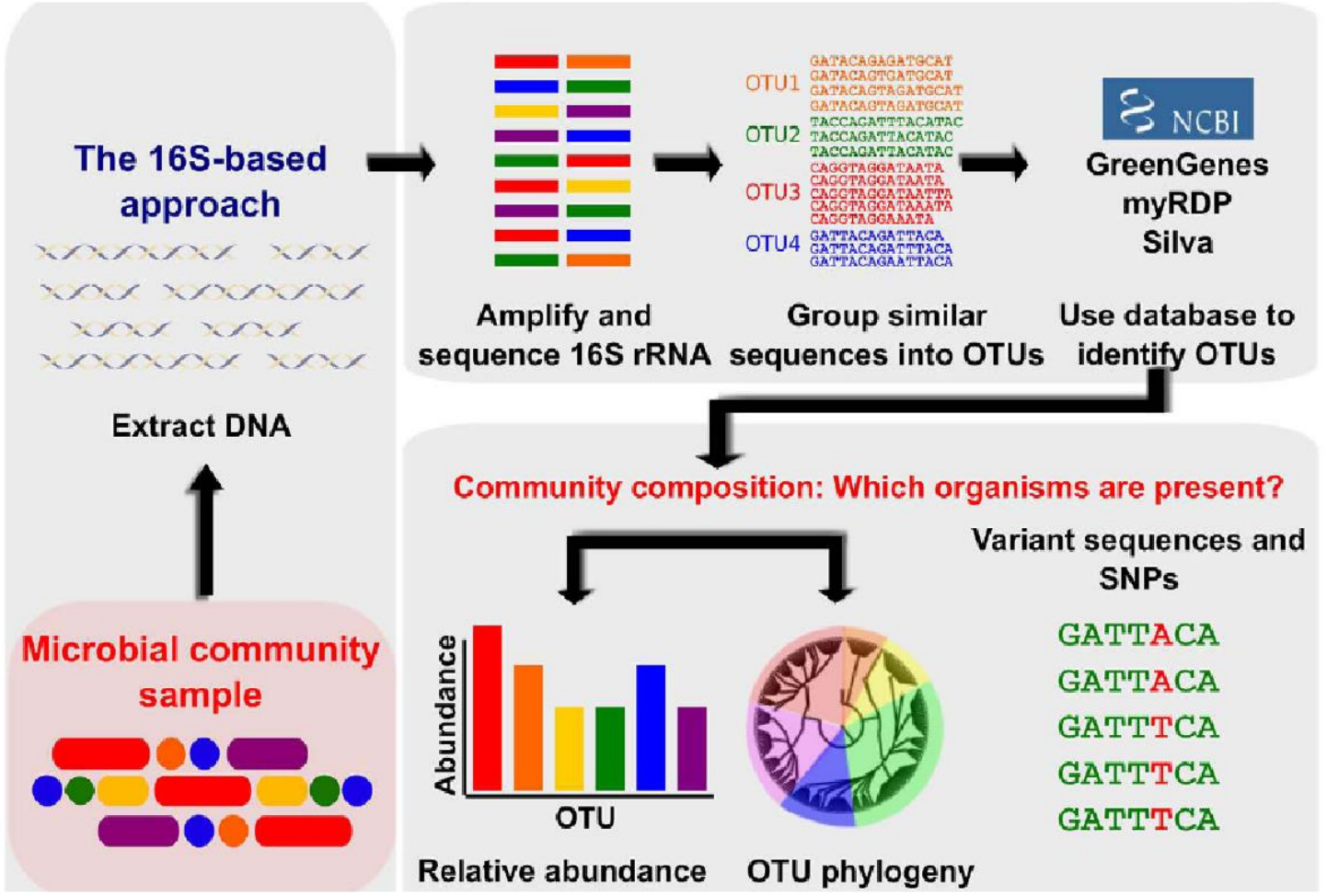
Preparación LibreríAs, rRNA 16S, Caracterización microbiota



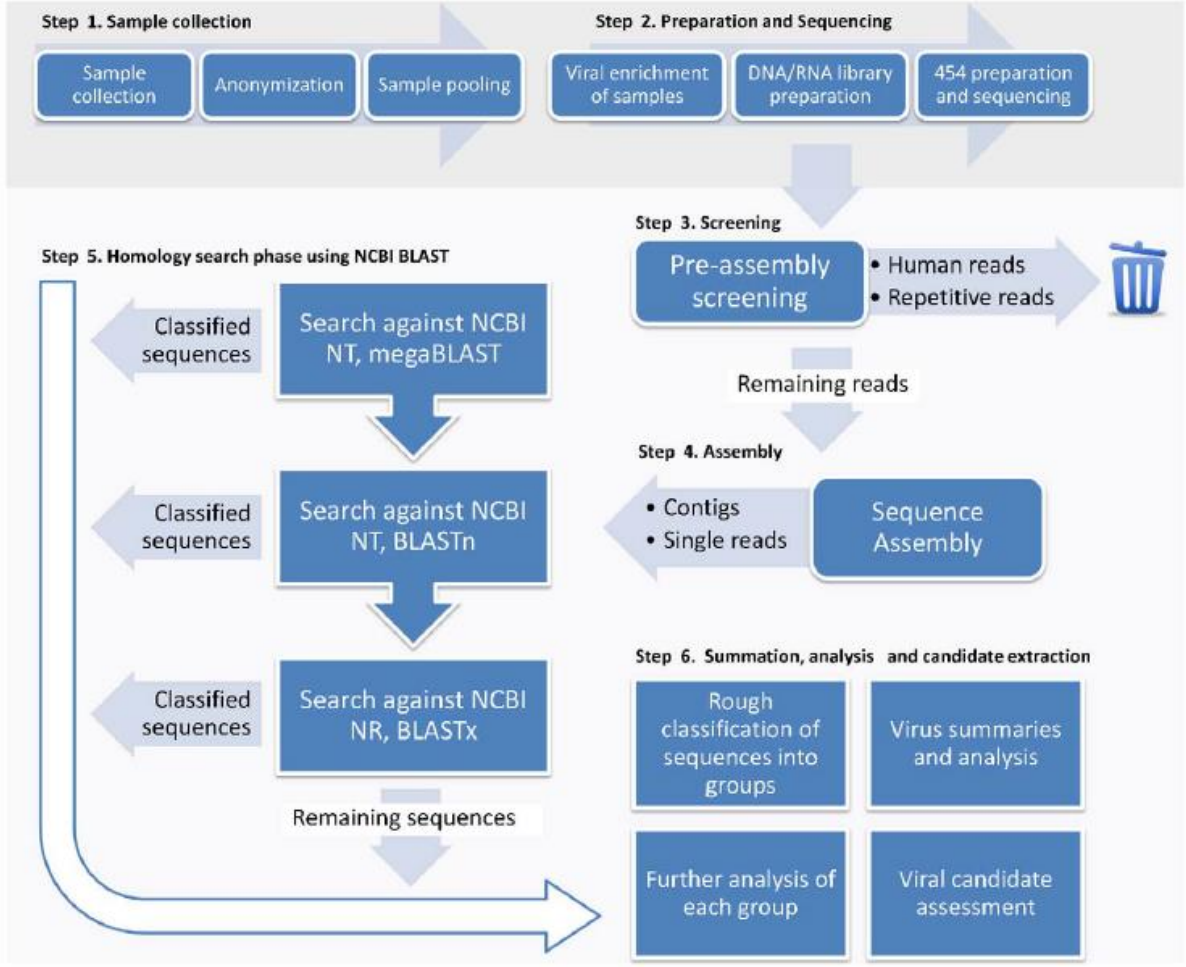
Targeted Metagenomics vs Metagenomics (16S vs Shotgun)

	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

Targeted metagenomics

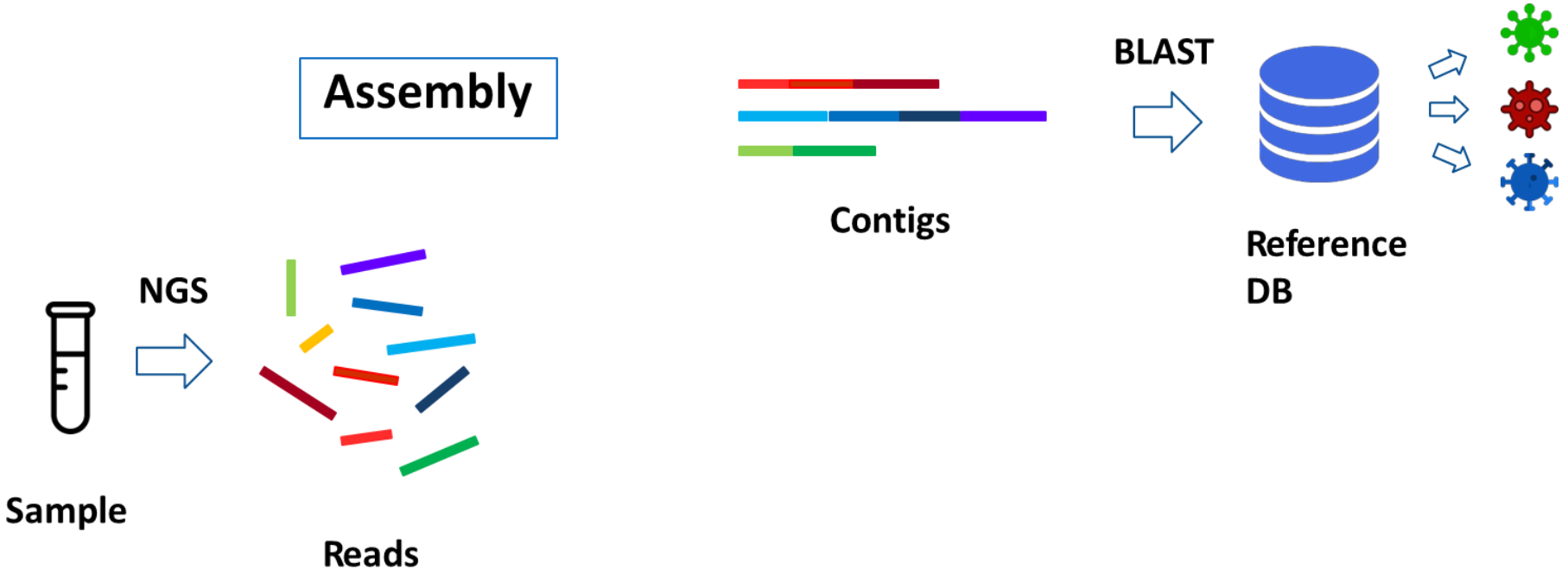


Metagenómica, pipeline de análisis

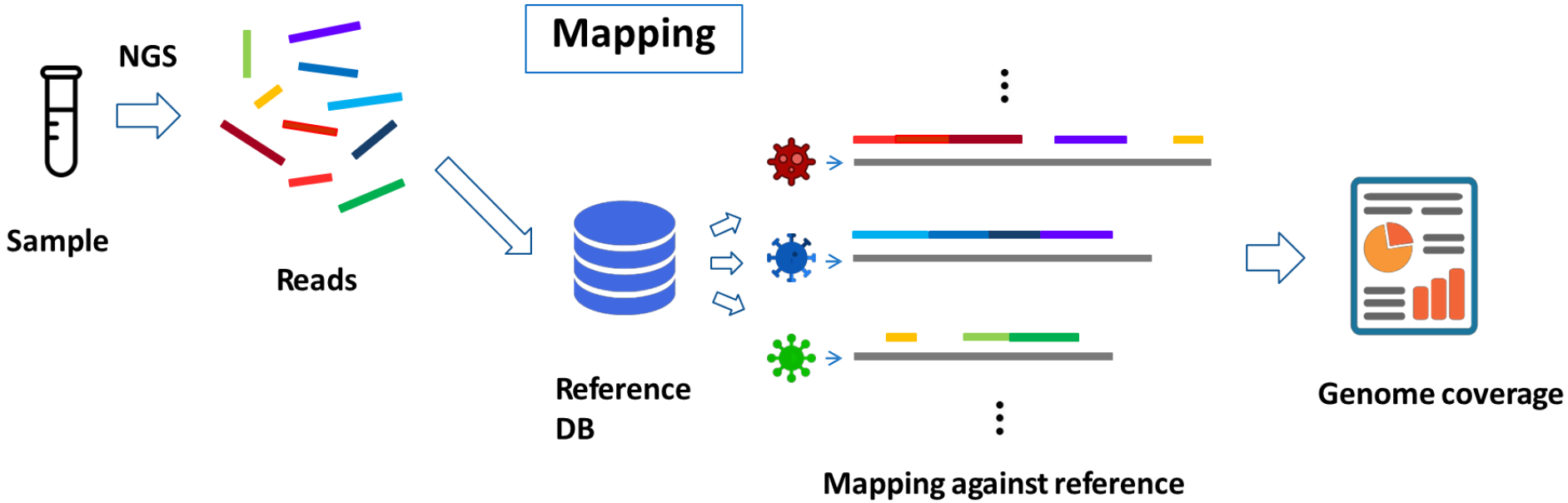


Lysholm et al., Plos One 2012:7,2, e30875

Metagenomic analysis approaches



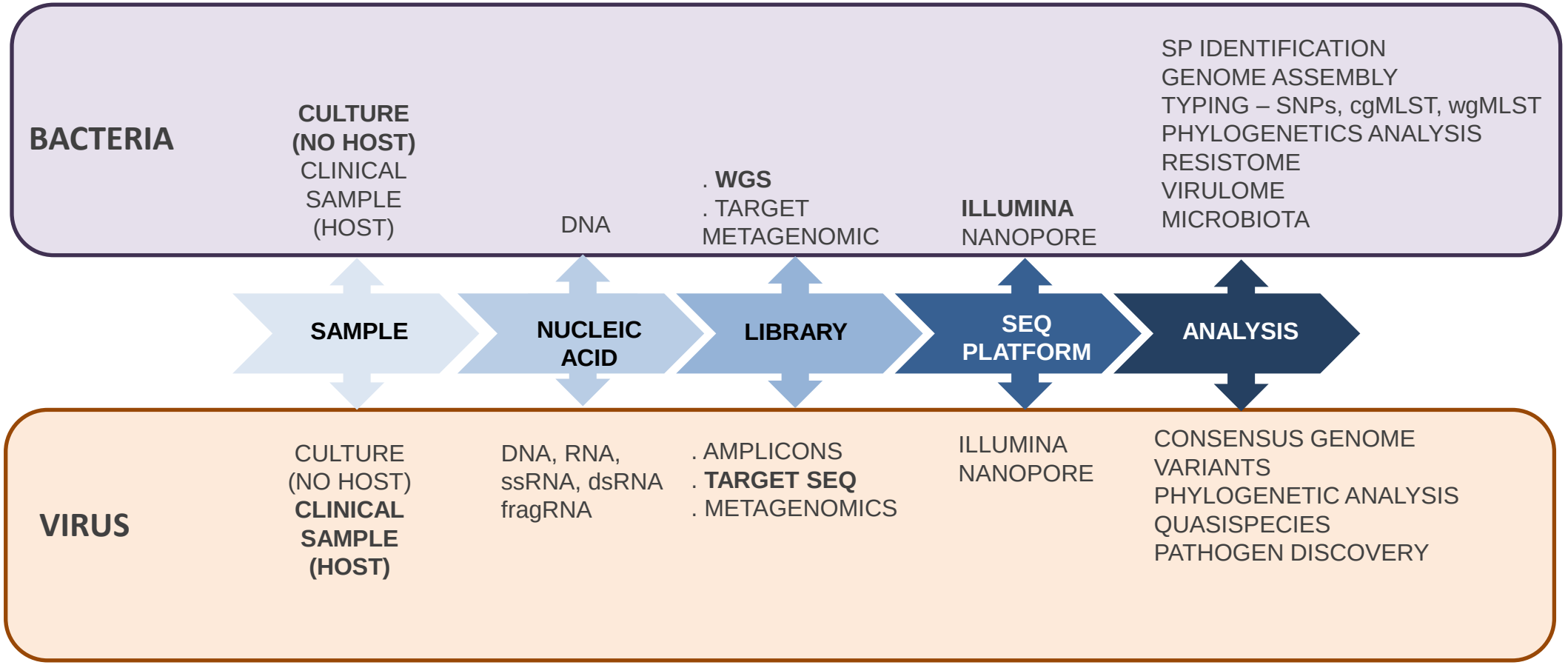
Metagenomic analysis approaches



Targeted Metagenomics vs Metagenomics (16S vs Shotgun)

Software	Organism	Genetic portion used		Binning algorithm used			Genome coverage	Novel pathogen discovery
		Genetic markers	Whole Genome	Clustering	Mapping	Assembly		
Mothur	Bacteria	X		X			No	No
QIIME	Bacteria	X		X		X	No	No
MEGAN	Bacteria		X			X	No	No
Platypus	Bacteria		X		X		No	No
SURPI	Virus		X			X	No	Yes
Virus-TAP	Virus		X			X	No	Yes
VIP	Virus		X		X		No	Yes
Pathosphere	Virus, Bacteria, Eukarya		X			X	No	Yes

Bacterial and Viral Genome Sequencing



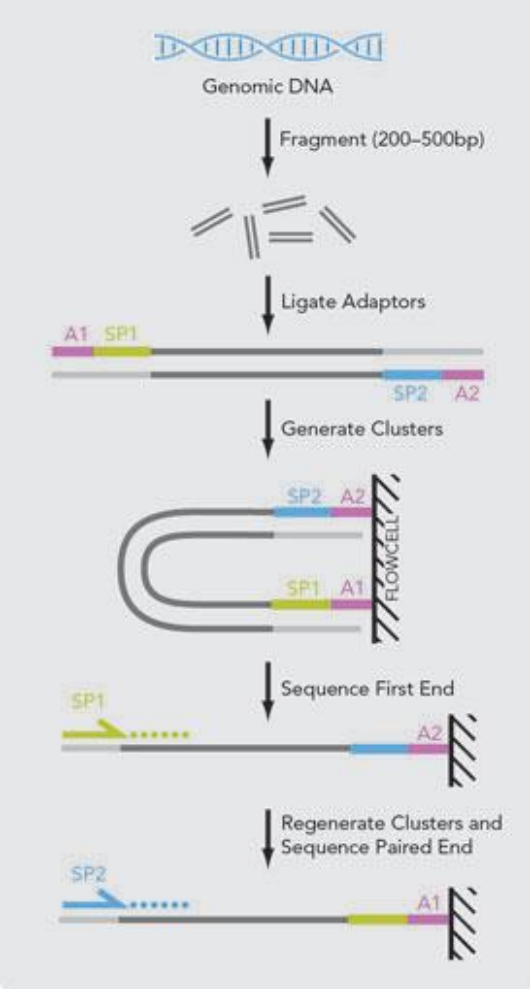
Index

- BU-ISCIII
- High throughput sequencing platforms update
- Bacterial genome sequencing, brief history
- Advantages of WGS
- Use of WGS in Europe
- **Library strategies**
- **Bioinformatics analysis**

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

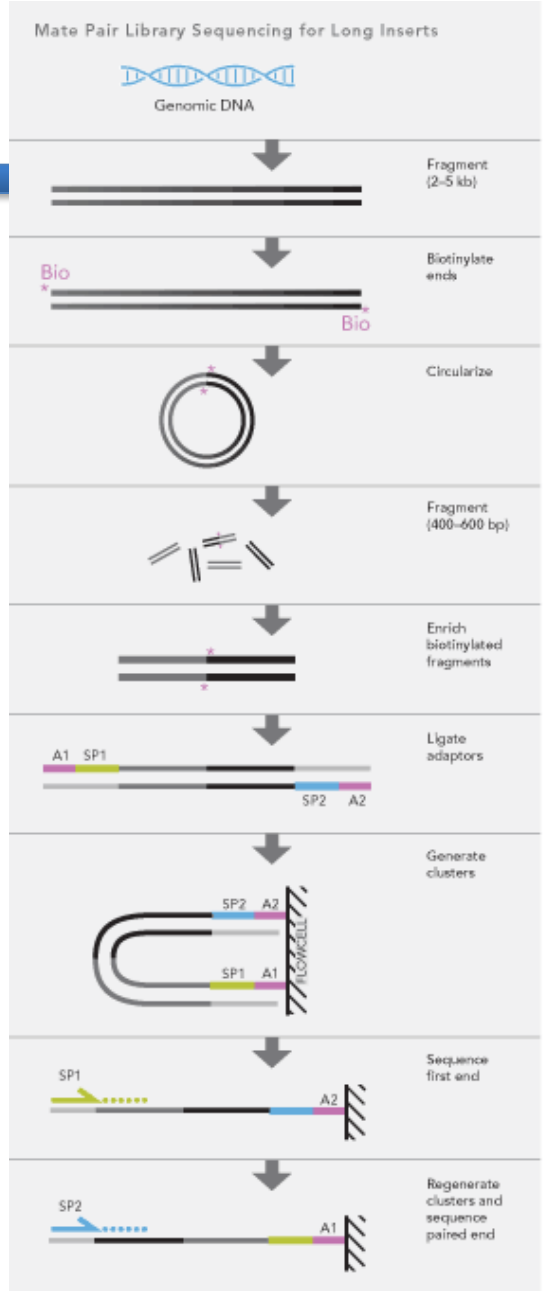
¿Qué es Pair-end?



Secuenciación de un fragmento (bp)

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

¿Qué 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

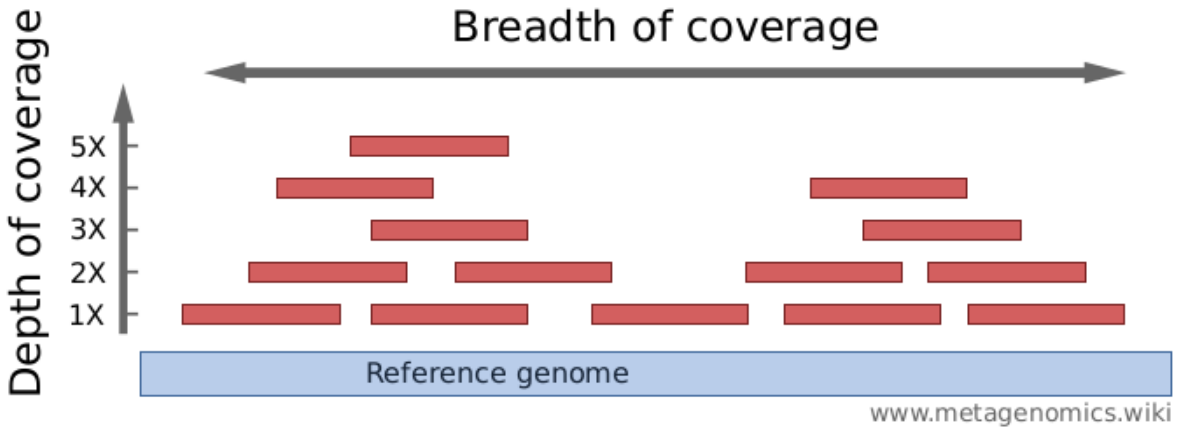
How much of a genome is "covered" by short reads? Are there regions that are not covered, even not by a single read?

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).



Cálculo 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:

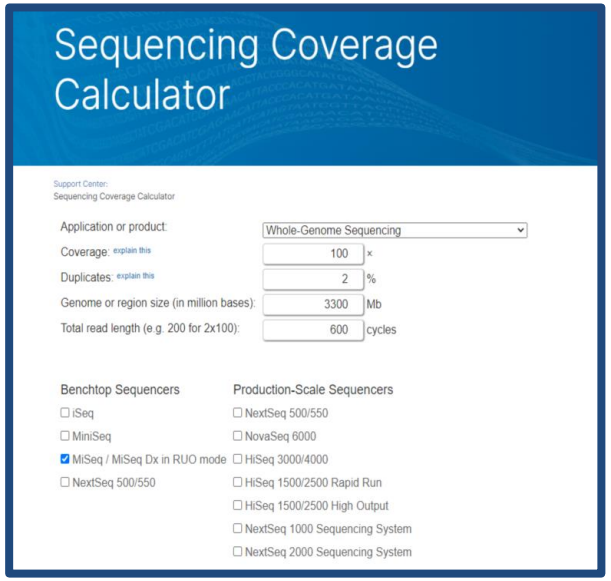
$$C = LN / G$$

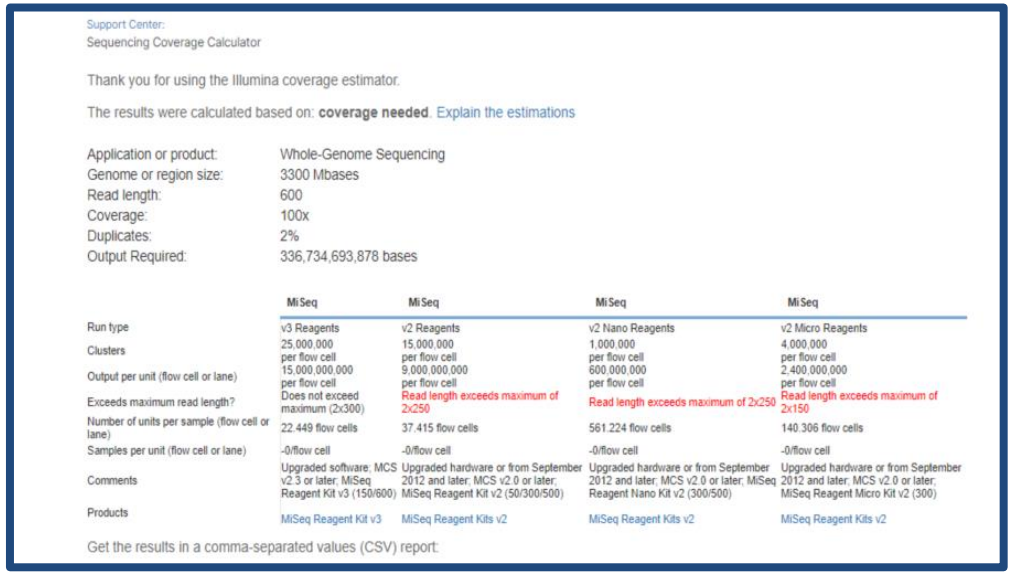
- **C stands for coverage**
- **G is the haploid genome length**
- **L is the read length**
- **N is the number of reads**

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

$$C = (100\ bp)*(189\times10^{6})/(3\times10^{9}\ bp) = 6.3$$

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





https://emea.support.illumina.com/downloads/sequencing_coverage_calculator.html

Thanks for your attention!

Questions???