

Análisis de Genomas Virales a través de la Plataforma Galaxy

BU-ISCIII



Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

11 al 15 Noviembre 2024

4ª Edición

Programa Formación Continua, ISCIII



Análisis de Genomas Virales a través de la
Plataforma Galaxy

Learning aims and outcomes

1. Understand some principles behind NGS and its applications to whole genome sequencing.
2. Know the format files generated in NGS data analysis and the workflow analysis.
3. Understand the uses of NGS in viral variants identification and consensus genome reconstruction.
4. Learning to use galaxy platform.

Teachers

- Sara Monzón Fernández, Dra Biociencias Moleculares, Biotecnóloga y Bioinformática (Analista de datos). Titulado Superior Especialista OPIS. Responsable técnico BU-ISCIII
- Sarai Varona Fernández, Bioquímica y Bioinformática (Analista de Datos). Contrato Titulado Superior asociado a proyecto (2021-2026)
- Emilia Arjona Bolaños, Bióloga y Bioinformática (Analista de Datos). Contrato Titulado Superior asociado a proyecto (2024-2026)
- Isabel Cuesta, Dra Biología, Bioinformática (Científico de Datos). Científico Titular de OPIS. Coordinador BU-ISCIII

Session 1.1 – Uso de la secuenciación masiva en virología

BU-ISCIII

Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

11 al 15 Noviembre 2024

4ª Edición

Programa Formación Continua, ISCIII

Análisis de Genomas Virales a través de la
Plataforma Galaxy

Index

- BU-ISCIII, Bioinformatics and preparedness to be a service provider
- High throughput sequencing in clinical virology
 - Application of HTS or NGS
 - Nucleic acid amplification
 - Library preparation
 - High throughput sequencing platforms
 - Data analysis

BU-ISCIII, Bioinformatics preparedness to be a service provider

What is Bioinformatics?

Biological Problems



Data
Processing

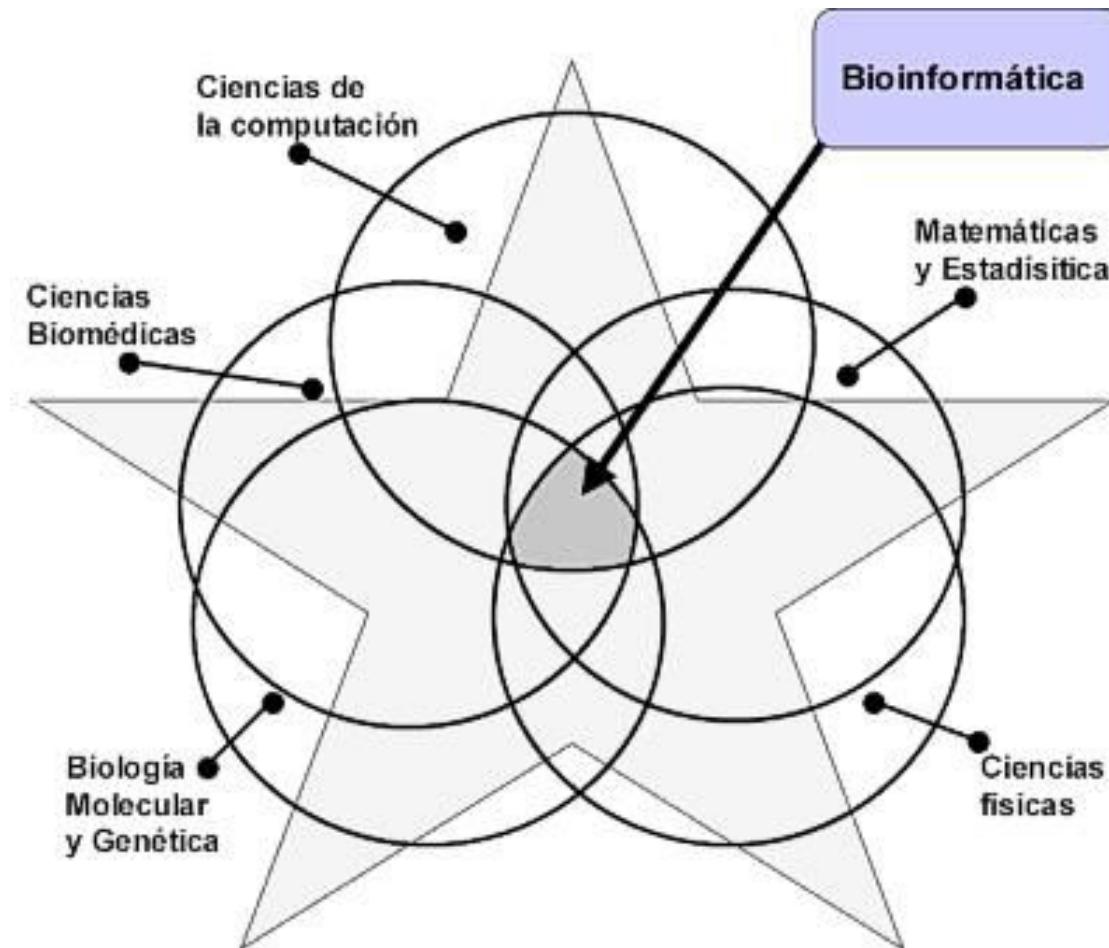


Computational Methods



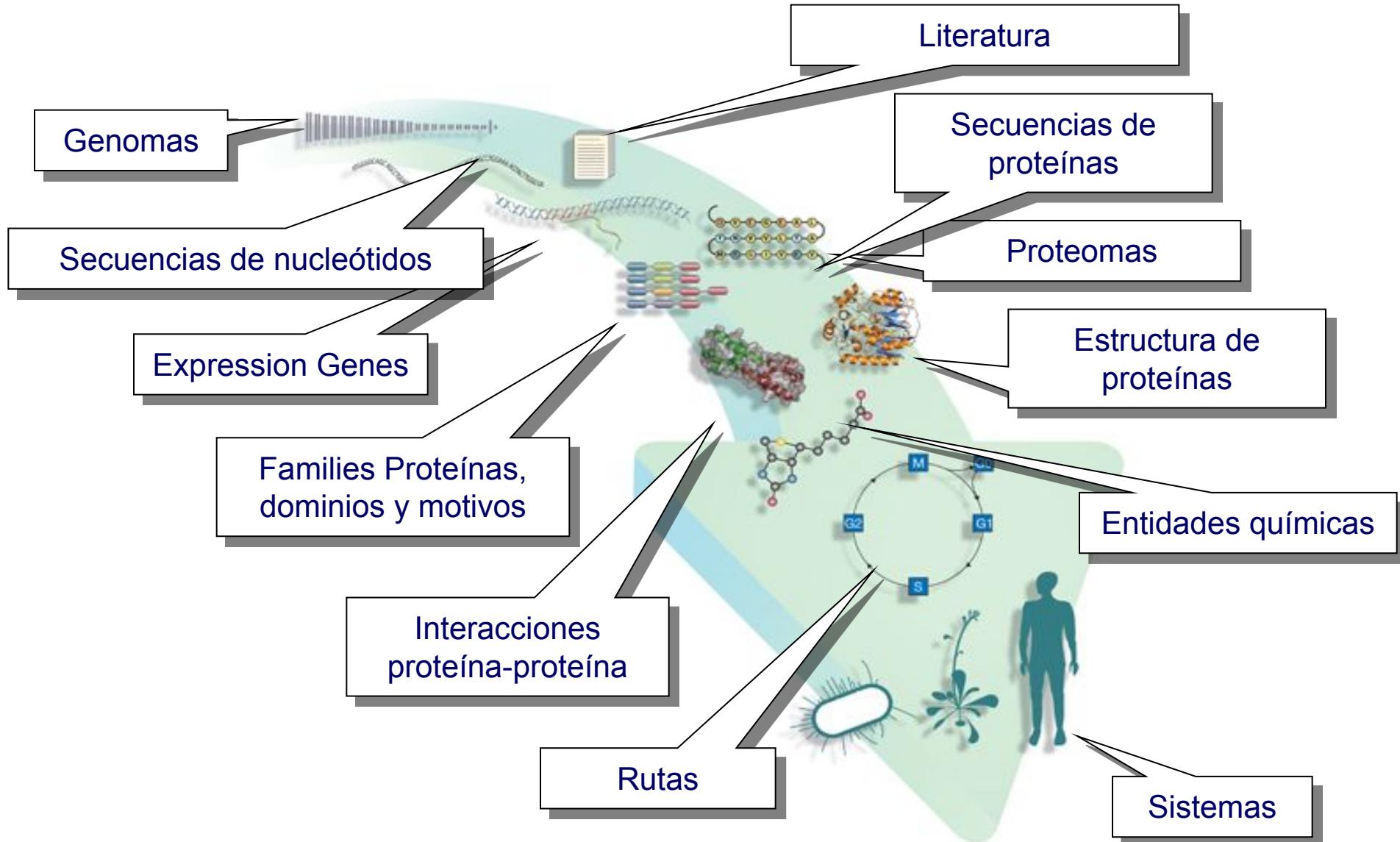
Bioinformatics is the science of storing, retrieving and analysing large amounts of biological information

Bioinformatics is a multidisciplinary field

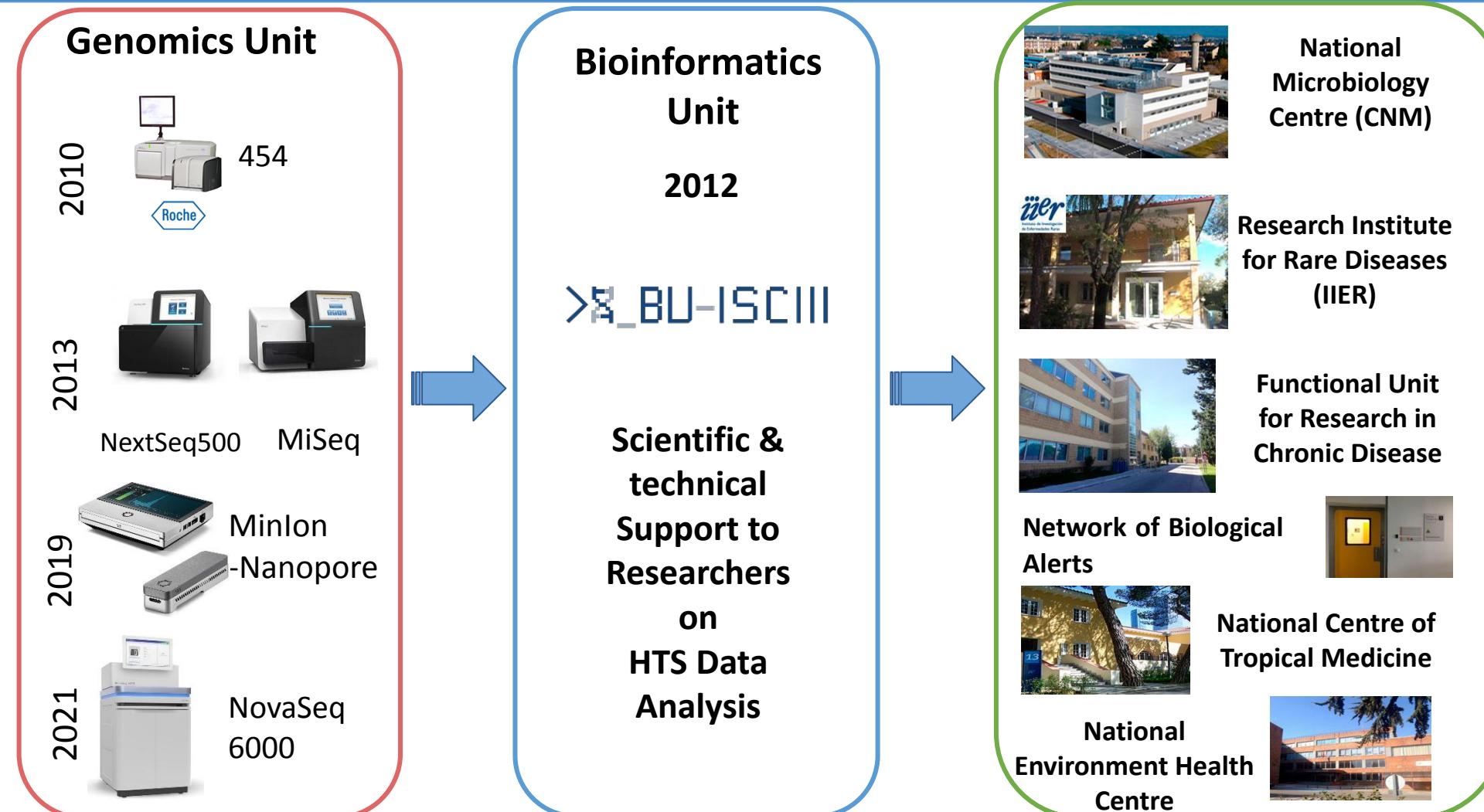


Análisis de Genomas Virales a través de la
Plataforma Galaxy

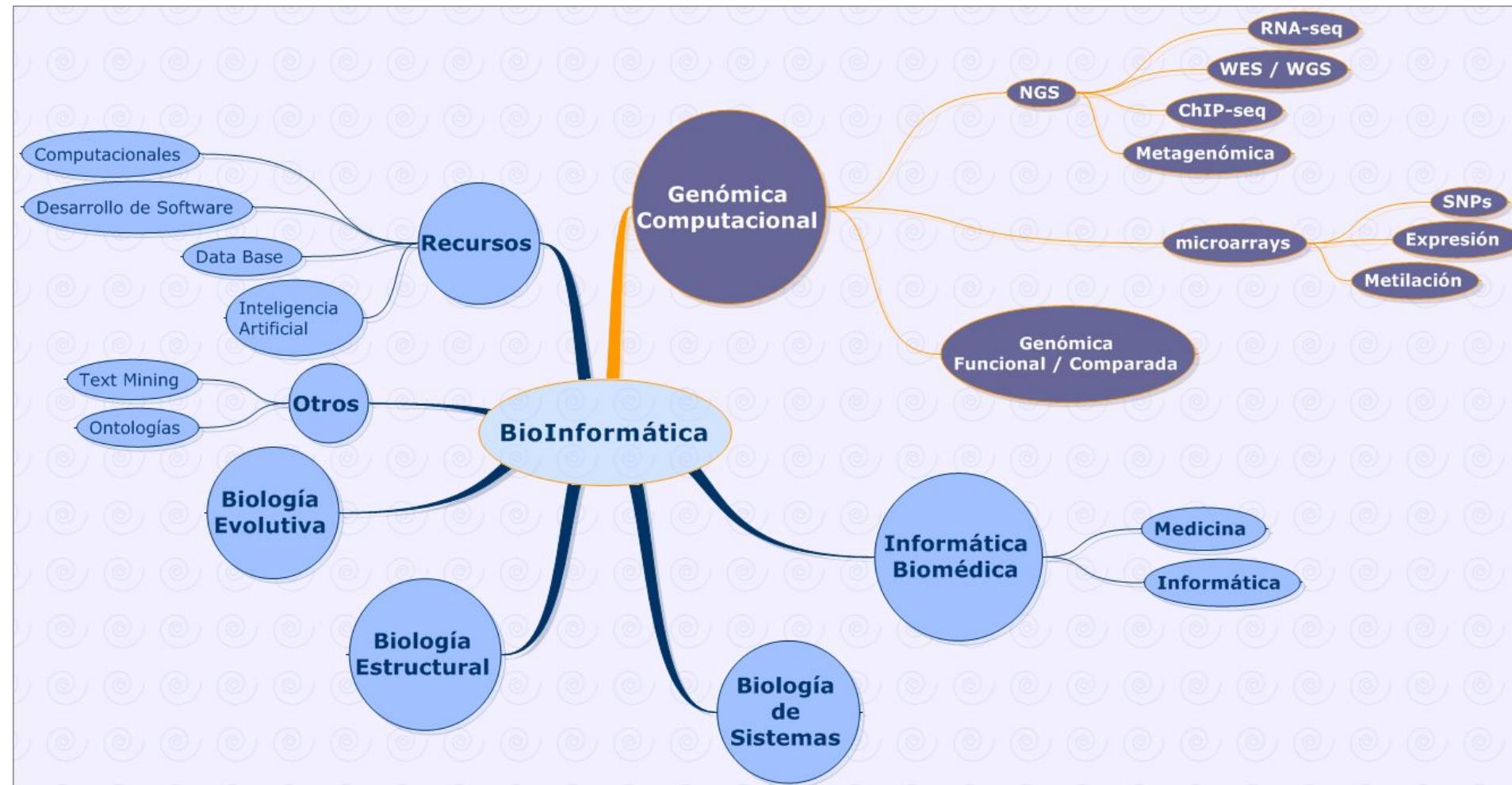
Dimension of Bioinformatics is given by data type



Why BU-ISCIII was founded



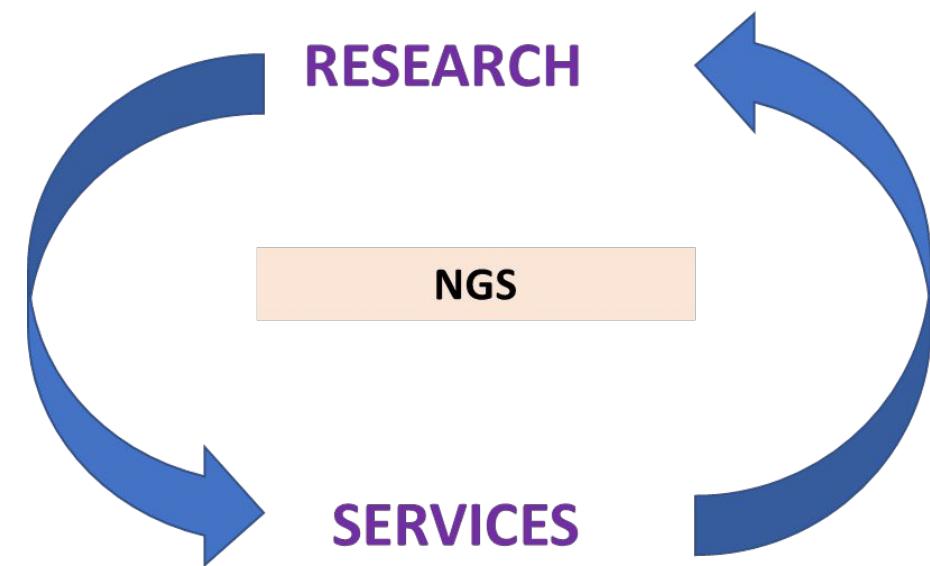
BU-ISCIII Mission - Activities



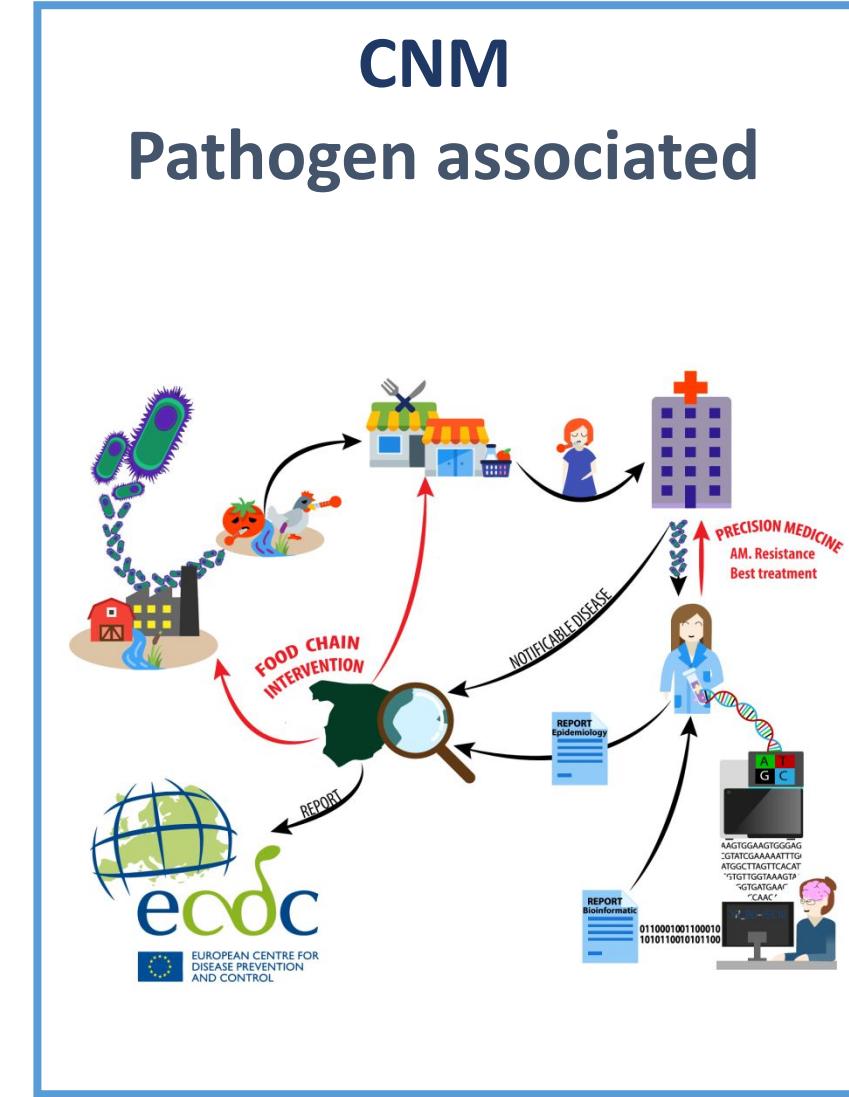
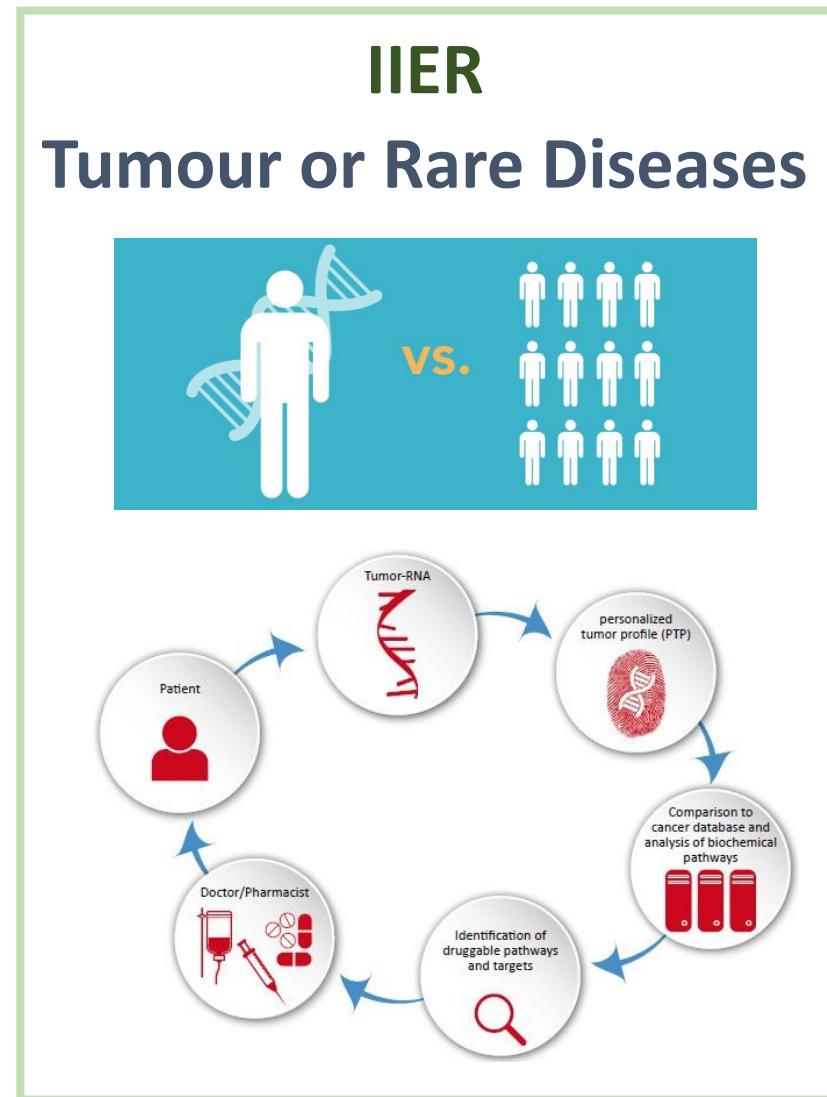
Bioinformatics Unit Activities

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

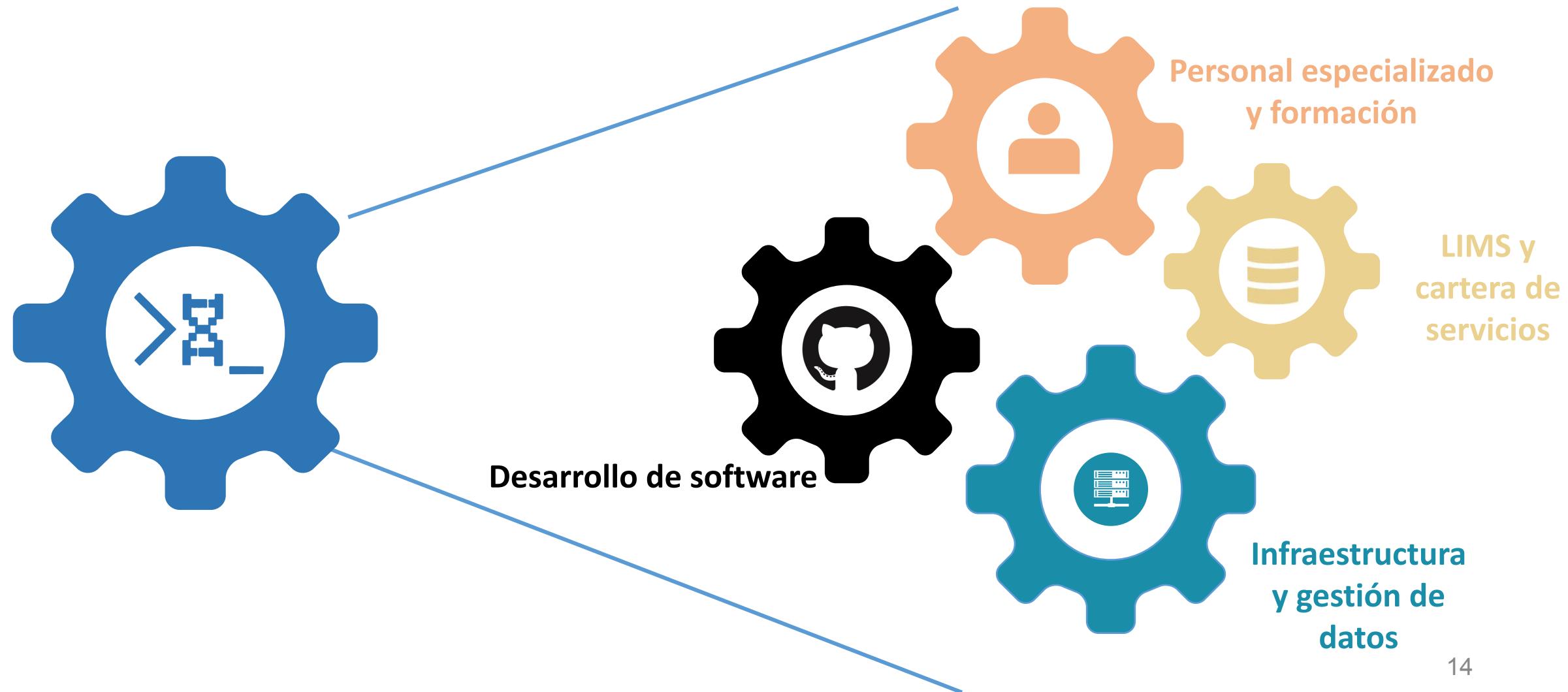
Provide scientific and technical solutions for using NGS in the diagnostic routine or research activity from different ISCIII labs



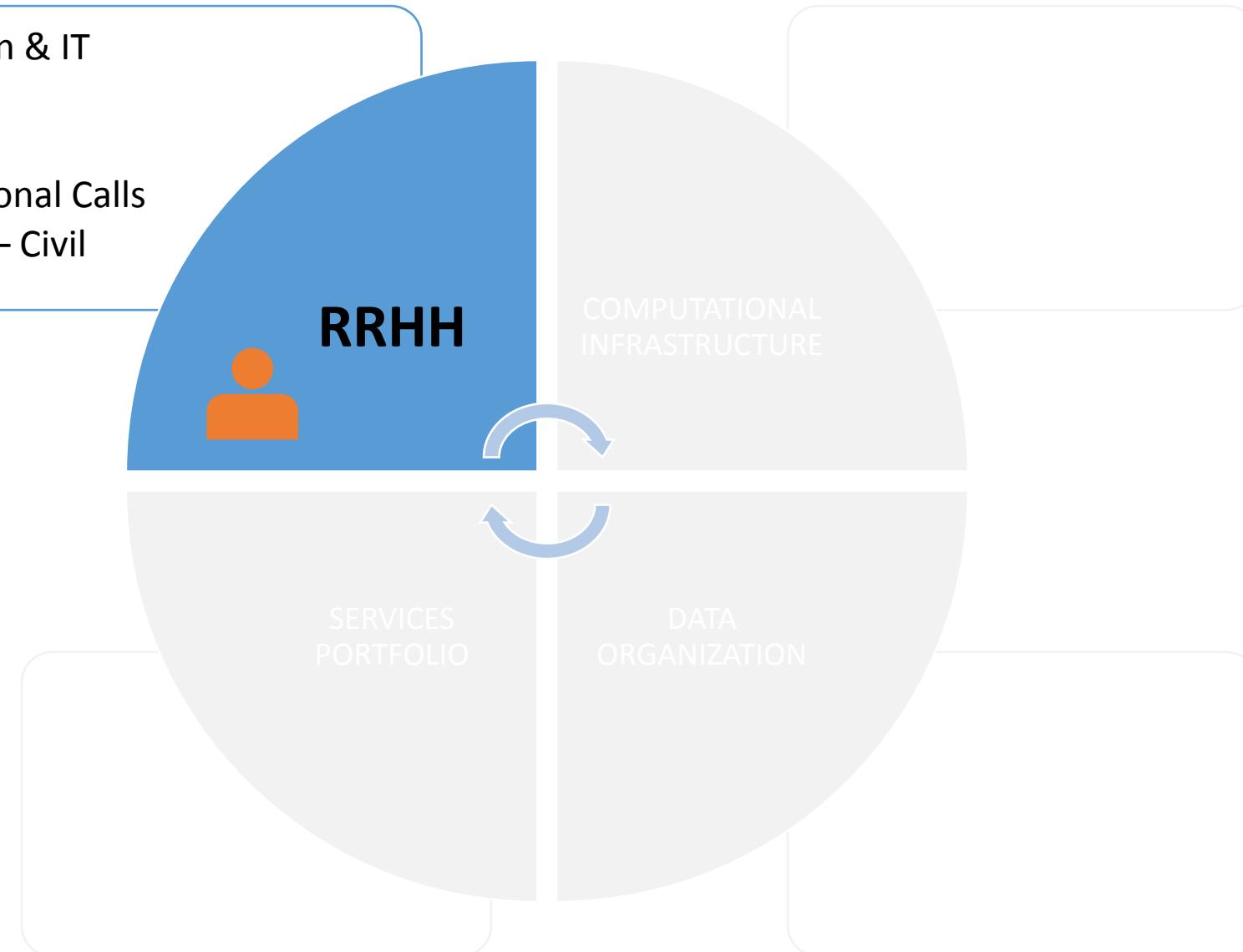
Clinical Bioinformatics - Precision Medicine



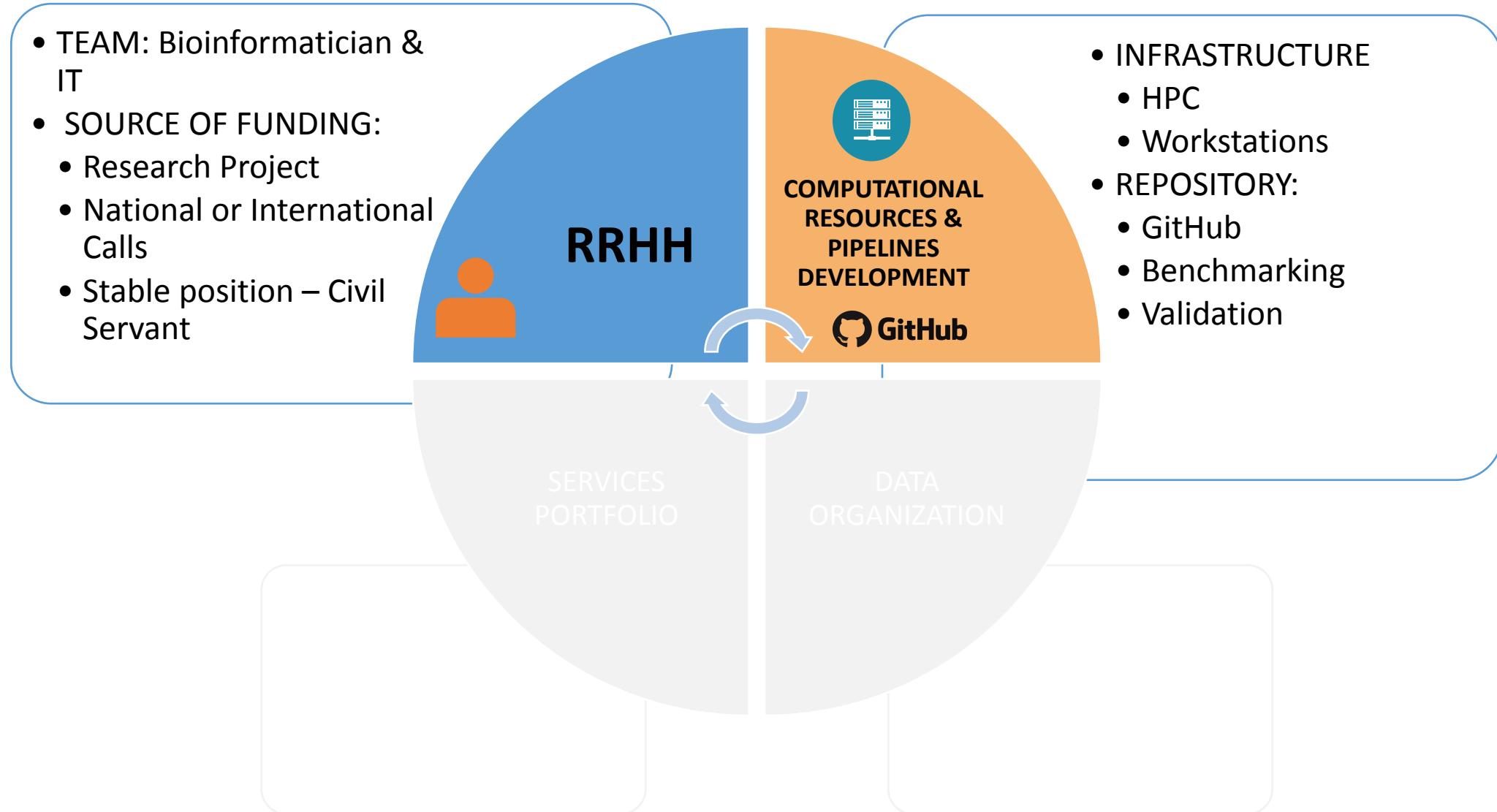
Organización de una Unidad de Bioinformática para labores de Salud Pública



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



	Disciplina	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Isabel Cuesta	Dr. Biología Molecular							CIENTÍFICO TITULAR OPIS						
Sara Monzón	Biotecnología	TFM	CIBERER				PTA MINECO		T.SUPERIOR OPIS					
Bruno Lobo	Administrador Sistemas			PROYECTO		PTA MINECO		SISTEMAS						
Jorge de la Barrera	Informática			PTA FIS										
Miguel Juliá	Matemáticas							PTA MINECO						
José Luis García	Telecomunicaciones							PROYECTO						
Pedro Sola	Biología					ANTIBIÓTICOS.			ANTIBIÓTICOS.					
Sarai Varona	Bioquímica									PROYECTO (2024)				
Luis Chapado	Telecomunicaciones							COLABORADOR						
Erika Kvalem	Biotecnología													
Alberto Lema	Biología								Proyecto Covid					
Luis Aranda	FP Informática													
Guillermo Gorines	Biología								Estancia	Proyecto				
Pablo Mata	Farmacia									Proyecto				
Jaime Ozález	Biología										Proyecto			
Daniel Valle- Millares	Biología									Proyecto				
Emilia Arjona	Biología									Proyecto				
Víctor López	Biología									Proyecto				
Juan Ledesma	Biología									Científico Titular OPIS				
Alejandro Bernabeu	Biología										Proyecto			
Alba Talavera	Biología										Proyecto			
Personal (número)		2	4	4	5	5	4	6	6	6	8	8	10	12



Computational Resources

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



Workstations (12), 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

- Reproducibility of in-silico pipelines analysis

nextflow



Singularity containers
Admin support & environment independency
Sharing code easier

 GitHub

<https://github.com/BU-ISCIII>

Bioinformatic Analysis: Software validation: ECDC EQAs

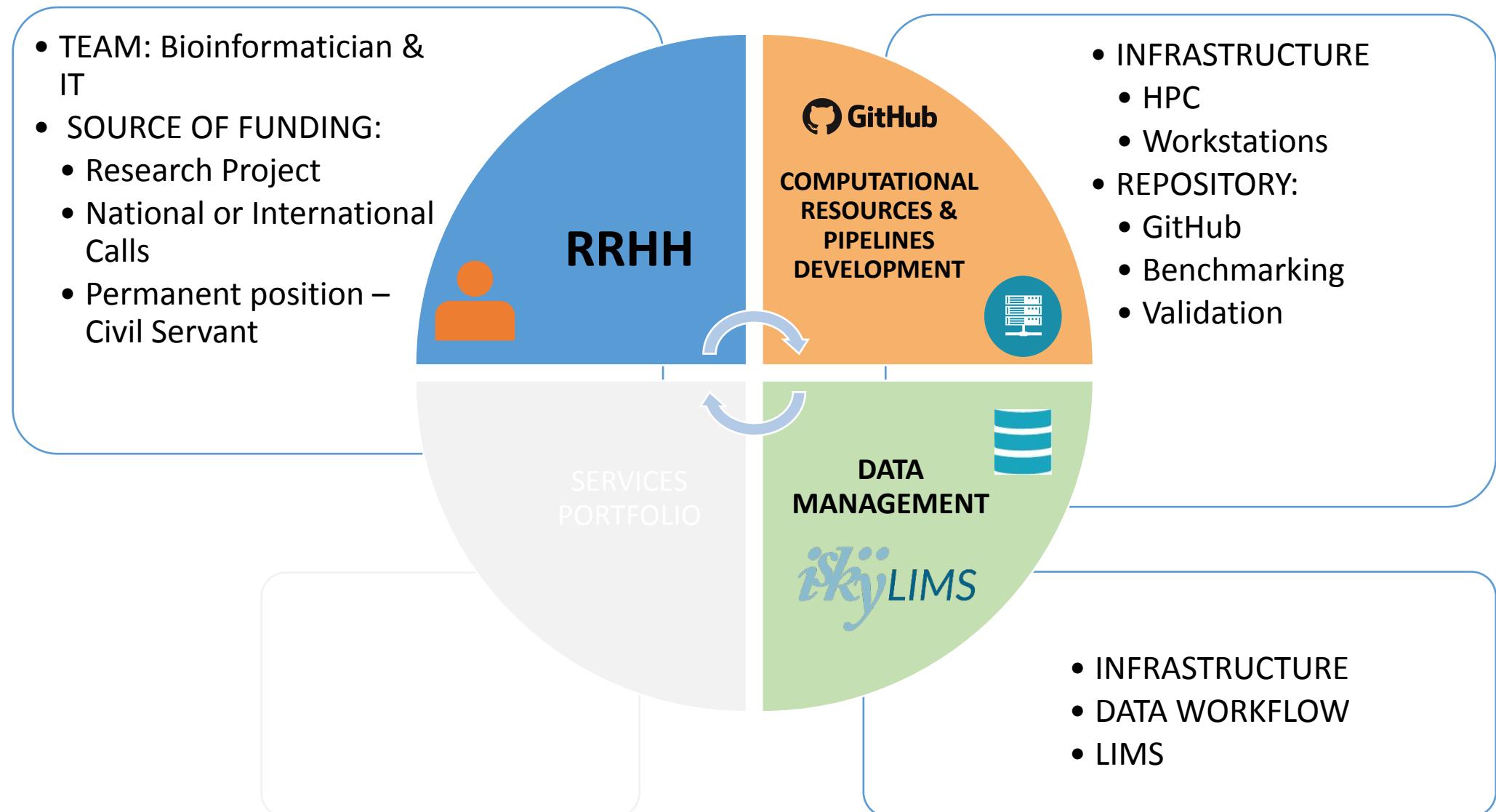
Table 5. Results of allele-based cluster analysis

Lab ID	Approach	Allelic calling method	Allele based analysis			
			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	SeqSphere	Assembly-based only	Velvet	Ruppitsch (cgMLST)	0-2	16-1065
70	SeqSphere	Assembly-based only	Velvet	Ruppitsch (cgMLST)	0-2	16-1062
105*	SeqSphere	Assembly-based only	SPAdes v 3.80	Ruppitsch (cgMLST)	0-1*	23-812
129	SeqSphere	Assembly-based only	Velvet			
135	SeqSphere	Assembly-based only	CLC Genomic Workbench 10			
141	SeqSphere	Assembly-based only	SPAdes 3.9.0			
142	Inhouse	Assembly-based only	SPAdes			
144	SeqSphere	Assembly-based only	Velvet			

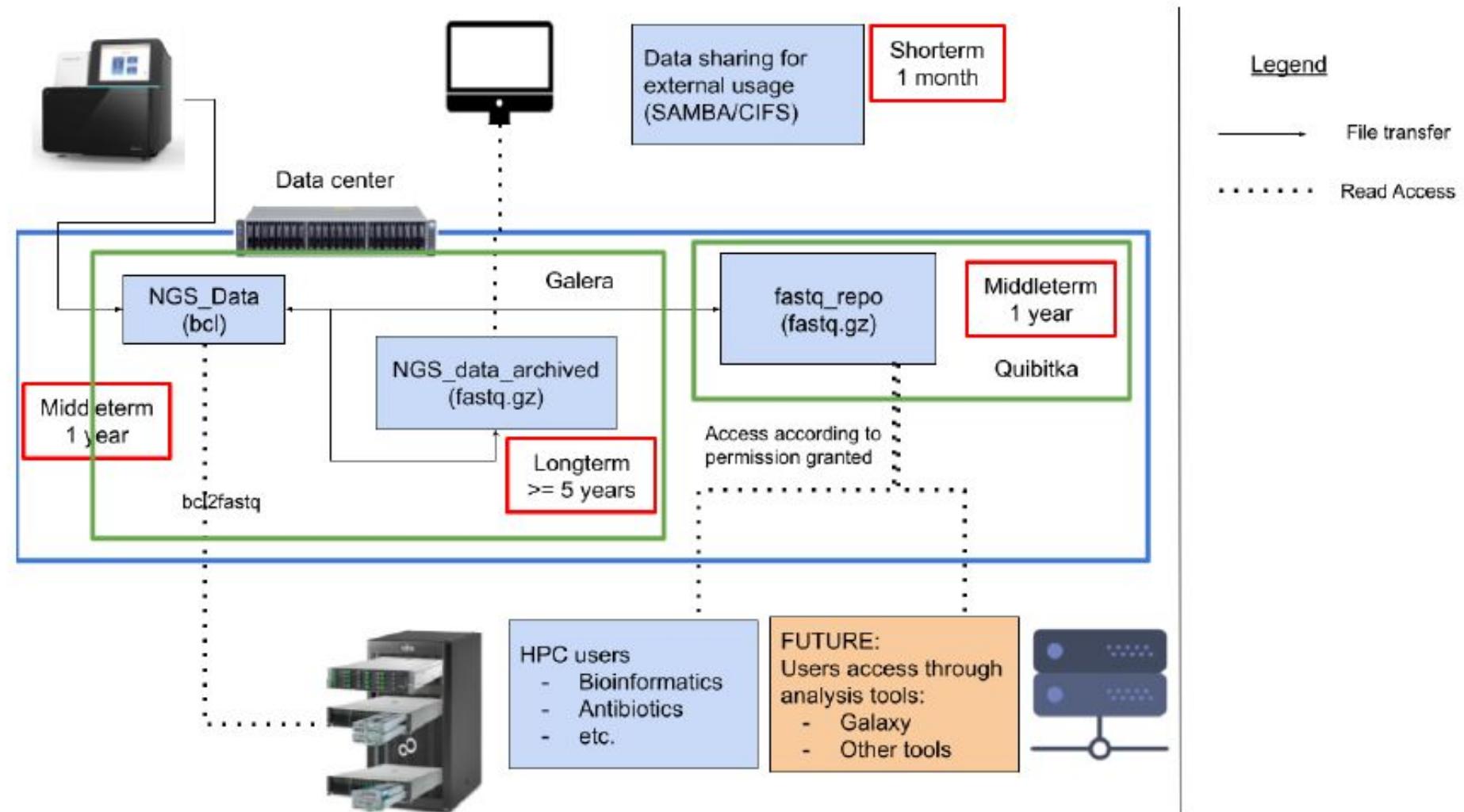
Table 4. Results of SNP-based cluster analysis

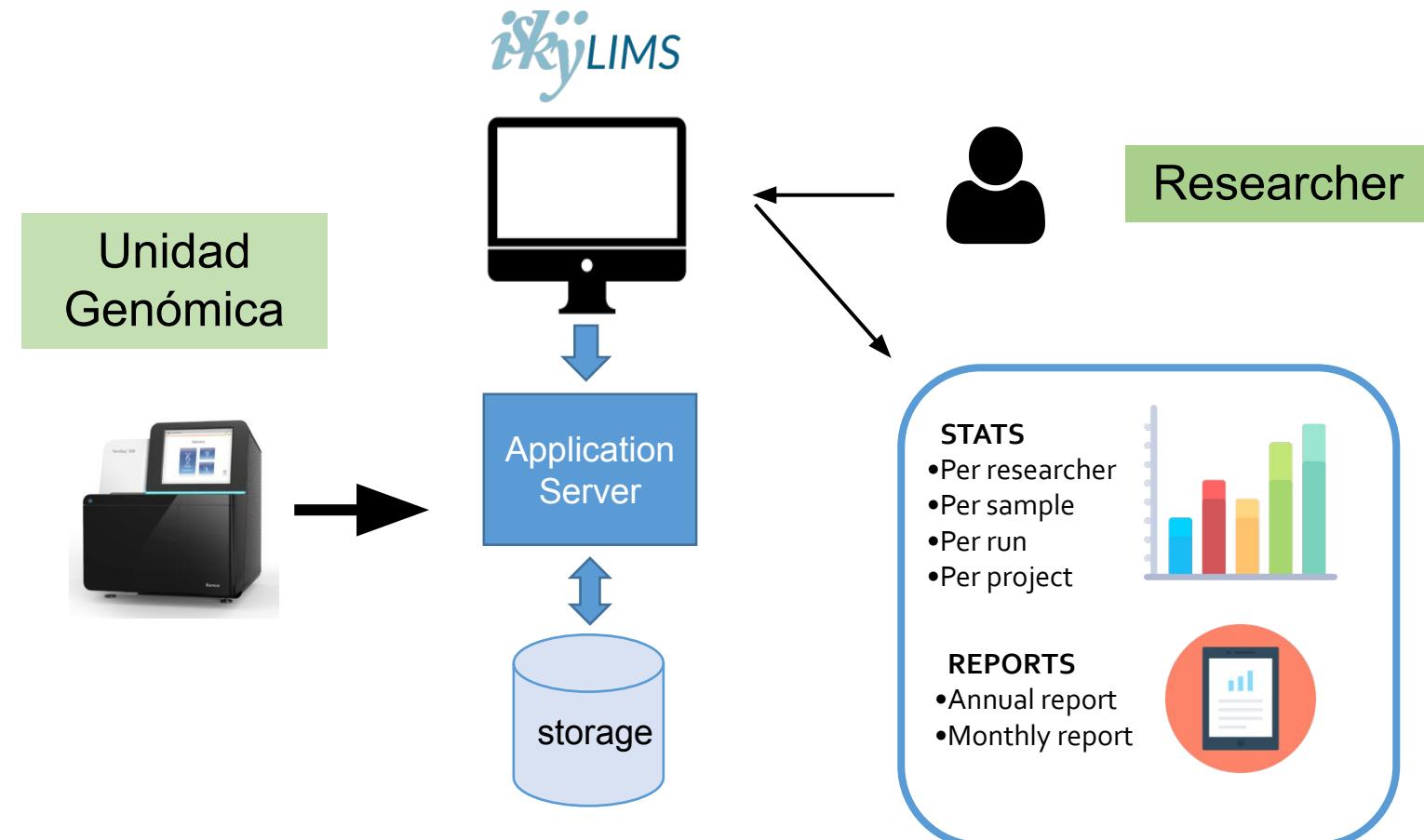
Lab ID	SNP-based						
	Approach	Reference	Read mapper	Variant caller	Assembler	Distance within cluster	Distance outside cluster
Provider	Reference-based	ST6 (REF4)	BWA	GATK		0-3	38-71
19*	Reference-based	ST6 ID 2362	BWA	GATK		0-4	43-81
56	Assembly-based			ksnp3	SPAdes	0-57*	561-591 (6109)
105	Reference-based	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

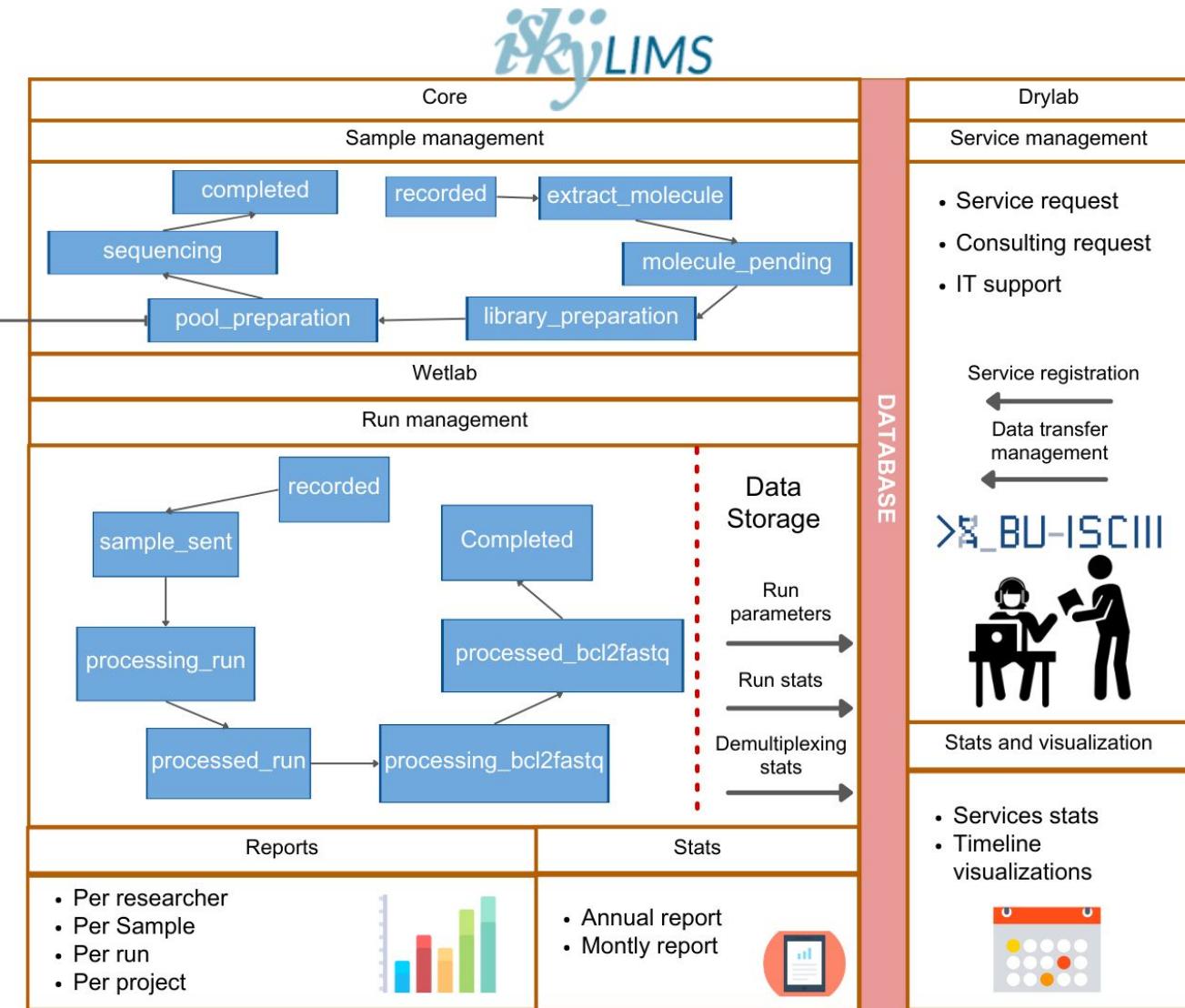
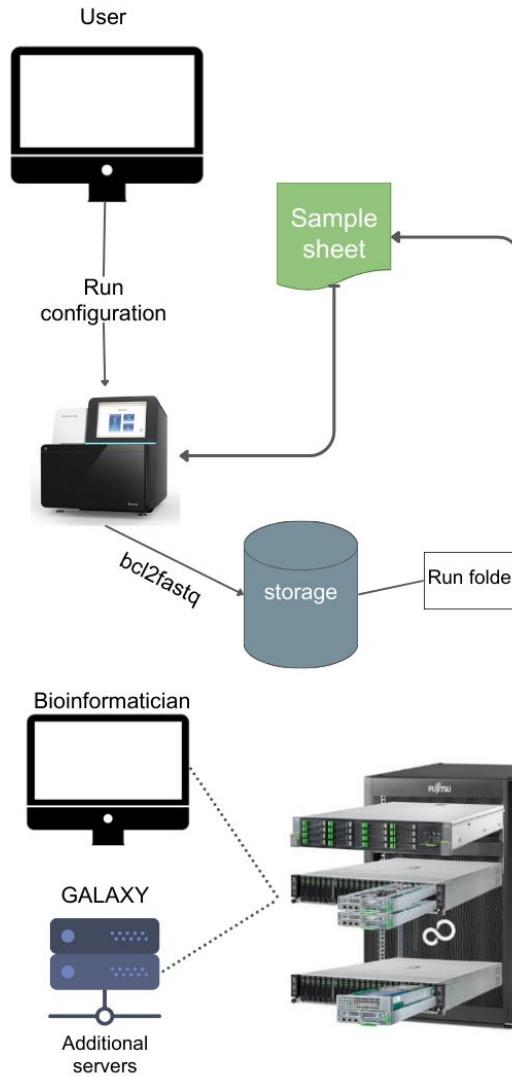


Infrastructure and data management





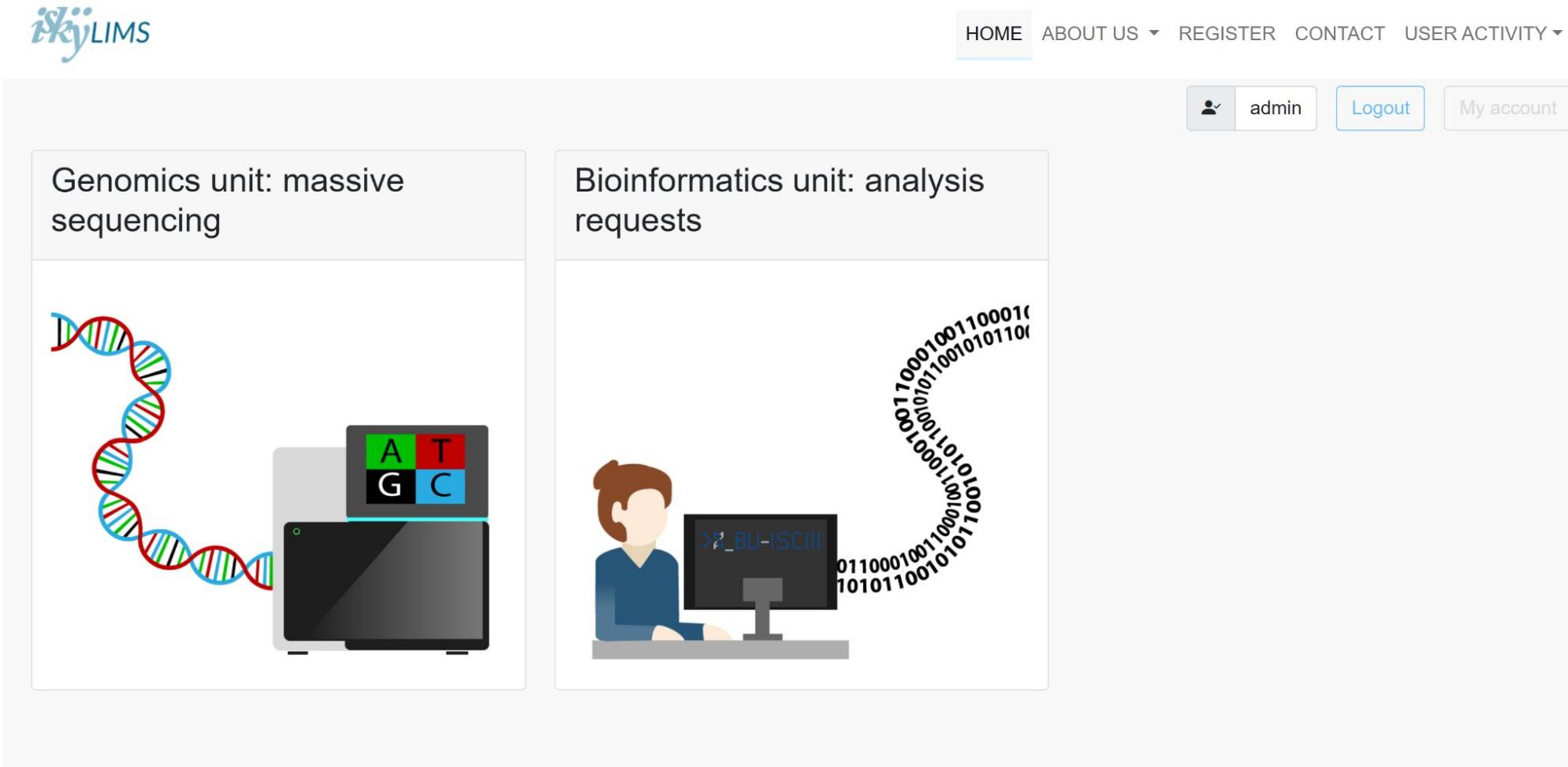
Infrastructure and data management: LIMS



Solicitud de servicio BU-ISCIII

iskyLIMS

<https://iskylims.isciii.es/>



Contact us



Social networks



Powered by



Solicitud de servicio BU-ISCIII



HOME CONFIGURATION ▾ SERVICES REQUEST ▾ PIPELINES ▾ MANAGE SERVICES ▾

admin Logout My account

iSkyLIMS: DryLab

Welcome

This section will allow you to check service activity. Available processes are:

- Sequencing
- Counseling
- Infrastructure

Email us for requesting detailed information about available services and conditions.



Recorded services

Show entries

Search:

Service Number

Requested by

SRVCNM1254

svaldezate

Search Service Number

Search Requested by

Showing 1 to 1 of 1 entries

Previous Next

In progress services

Show entries

Search:

Service number

Expected delivery date

SRVCNM026

2016-09-12

SRVCNM116

None

SRVCNM1242

2024-11-04

SRVCNM1243

None

SRVCNM491

2021-09-29

Search Service number

Search Expected delivery date

Showing 1 to 5 of 7 entries

Previous Next



<https://iskylims.isciii.es/>

Select samples sequenced in-house Select samples recorded but sequenced elsewhere

Search:

	Run Name	Project Name	Sample Name	Run finish date	Select Sampl
1	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0156	13/10/2023	<input type="checkbox"/>
2	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0157	13/10/2023	<input type="checkbox"/>
3	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0158	13/10/2023	<input type="checkbox"/>
4	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0159	13/10/2023	<input type="checkbox"/>
5	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0160	13/10/2023	<input type="checkbox"/>
6	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0161	13/10/2023	<input type="checkbox"/>
7	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0162	13/10/2023	<input type="checkbox"/>
8	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0163	13/10/2023	<input type="checkbox"/>
9	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0164	13/10/2023	<input type="checkbox"/>
10	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0165	13/10/2023	<input type="checkbox"/>
11	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0166	13/10/2023	<input type="checkbox"/>
12	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0167	13/10/2023	<input type="checkbox"/>
13	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0168	13/10/2023	<input type="checkbox"/>
14	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0169	13/10/2023	<input type="checkbox"/>
15	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0170	13/10/2023	<input type="checkbox"/>
16	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0171	13/10/2023	<input type="checkbox"/>
17	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0172	13/10/2023	<input type="checkbox"/>
18	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0173	13/10/2023	<input type="checkbox"/>
19	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0174	13/10/2023	<input type="checkbox"/>
20	MiSeq_GEN_398	MiSeq_GEN_398_20231011_MPerez	23pae0175	13/10/2023	<input type="checkbox"/>

Showing page 1 of 2005 entries

1 2 3 4 5 6 7 8 9 10 >

Service selection

Genomic Data Analysis

- Pre-processing and quality analysis
 - Sequencing quality analysis and host genome removal
- Next Generation Sequencing data analysis
 - DNAseq / cDNAseq: Exome sequencing (WES) / Genome sequencing (WGS) / Targeted sequencing
 - Low-frequency variants detection and annotation for whole genome or sequencing panel (e.g. retinoblastoma gene panel)
 - Eukaria: Variant calling and annotation for a sequencing panel (e.g. epidermophytosis gene panel, mouse or rat gene panel)
 - Human: Exome sequencing, variant calling, annotation and inheritance filtering (e.g. Exome sequencing of a human trio (two parents and one child))
 - Human: Whole genome sequencing for SNPs variant calling, annotation and inheritance filtering (e.g. WGS of a human)
 - Eukaria (non-human): Variant calling, annotation and SNP-based outbreak analysis (e.g. diploid fungal outbreak)
 - Fungal / bacteria / virus : Variant calling, annotation and SNP-based outbreak analysis (e.g. haploid fungal outbreak)
 - Bacteria: De novo genome assembly and annotation
 - Bacteria: In-depth analysis of Mycobacterium species genomes (e.g. M. tuberculosis, M. bovis)
 - Bacteria: Plasmid analysis and characterization
 - Bacteria: Multi-Locus Sequence Typing (MLST), analysis of virulence factors, antimicrobial resistance, and plasmids characterization
 - Bacteria: Core genome or whole genome Multi-Locus Sequence Typing analysis (cg/wgMLST)
 - Viral: Genomic reconstruction, variant calling and de novo assembly
 - Viral Flu: Influenza fragment reconstruction and variant detection
- RNAseq: Transcriptome sequencing
 - Differential Gene Expression (DEG)
 - Differential transcript expression (DET)
 - Differential miRNA expression (DEM)
 - Gene expression changes over a series of time points
- Metagenomics and targeted metagenomics
 - Taxonomic based identification and classification of organisms in complex communities
 - De novo assembly contigs' alignment to database
 - Bacteria: 16S rRNA gene analysis to assess bacterial diversity
 - Viral: Detection and characterization of viral genomes within metagenomic data

Additional Information

Add file... Start upload Cancel upload Delete

Upload Files

You can drag & drop files from your desktop on this webpage (if Browser supports).

The Maximum file size to upload is 5MB. For including files bigger than 5MB contact Bioinformatic Unit Responsible

Service Notes

Ongoing services

Services lists My services

Graphs Recorded Queued In Progress On hold

content

Show 10 entries

Search:

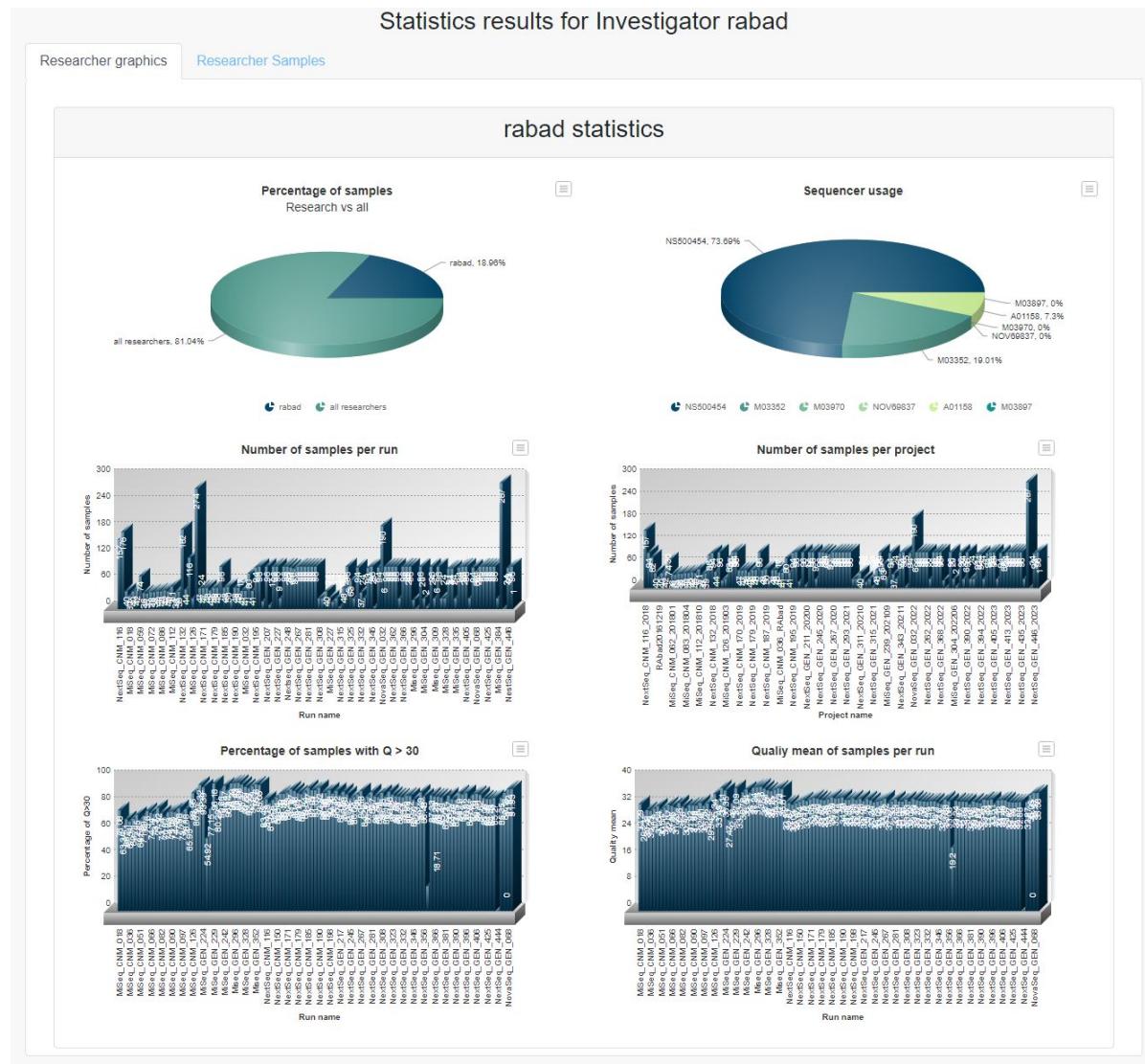
Service ID	User	Recorded date	Resolution date	Assigned to	On queue date	Estimated delivery date
SRVCNM559	smonzon	2024-01-25	--	--	--	--
SRVCNM560	a.vazquez	2024-02-08	--	--	--	--
SRVCNM979	jeecheva	2023-10-11	--	--	--	--
SRVCNM980	rabad	2023-10-13	--	--	--	--
SRVCNM981	sherrera	2023-10-16	--	--	--	--

Search Service ID Search User Search Recorded date Search Resolution date Search Assigned to Search On queue Search Estimated delivery

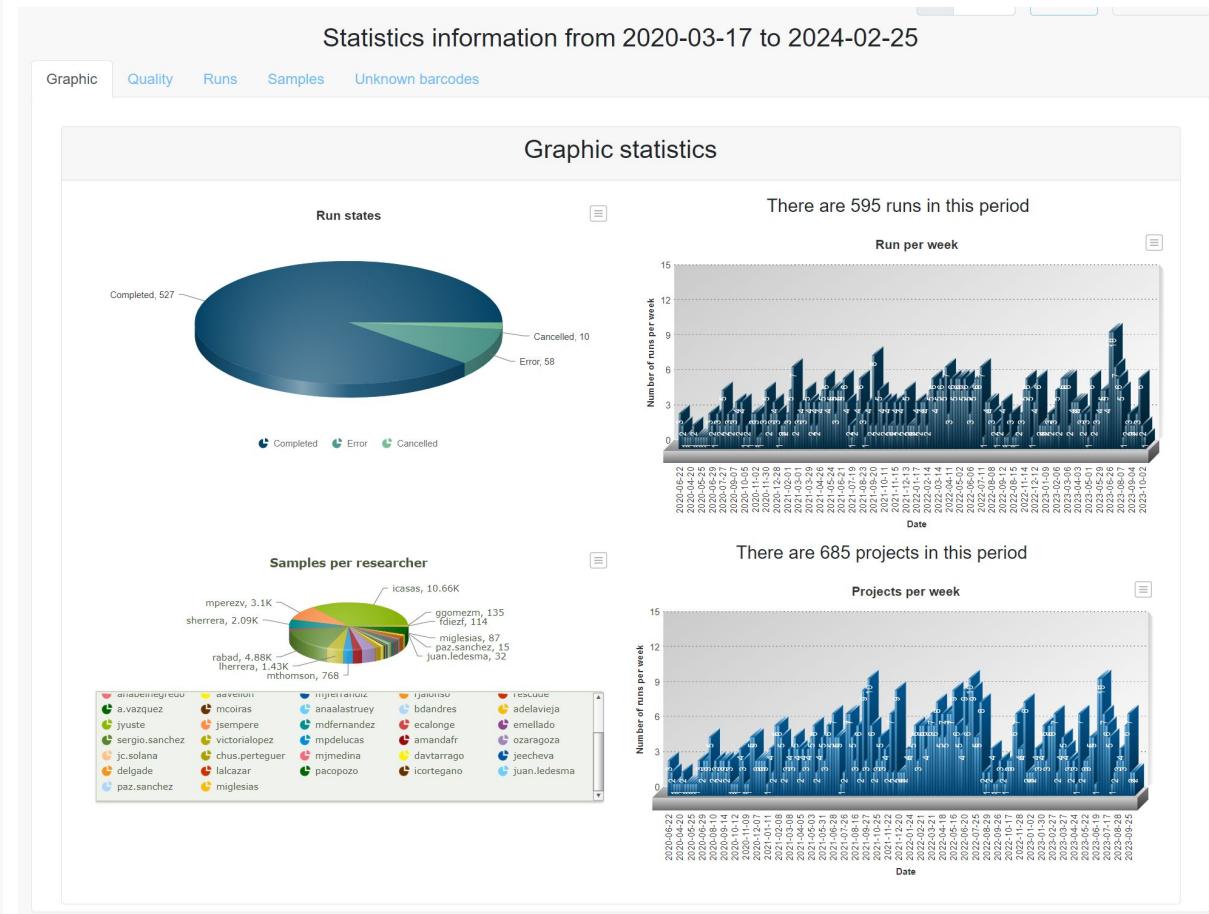
Showing 1 to 5 of 5 entries

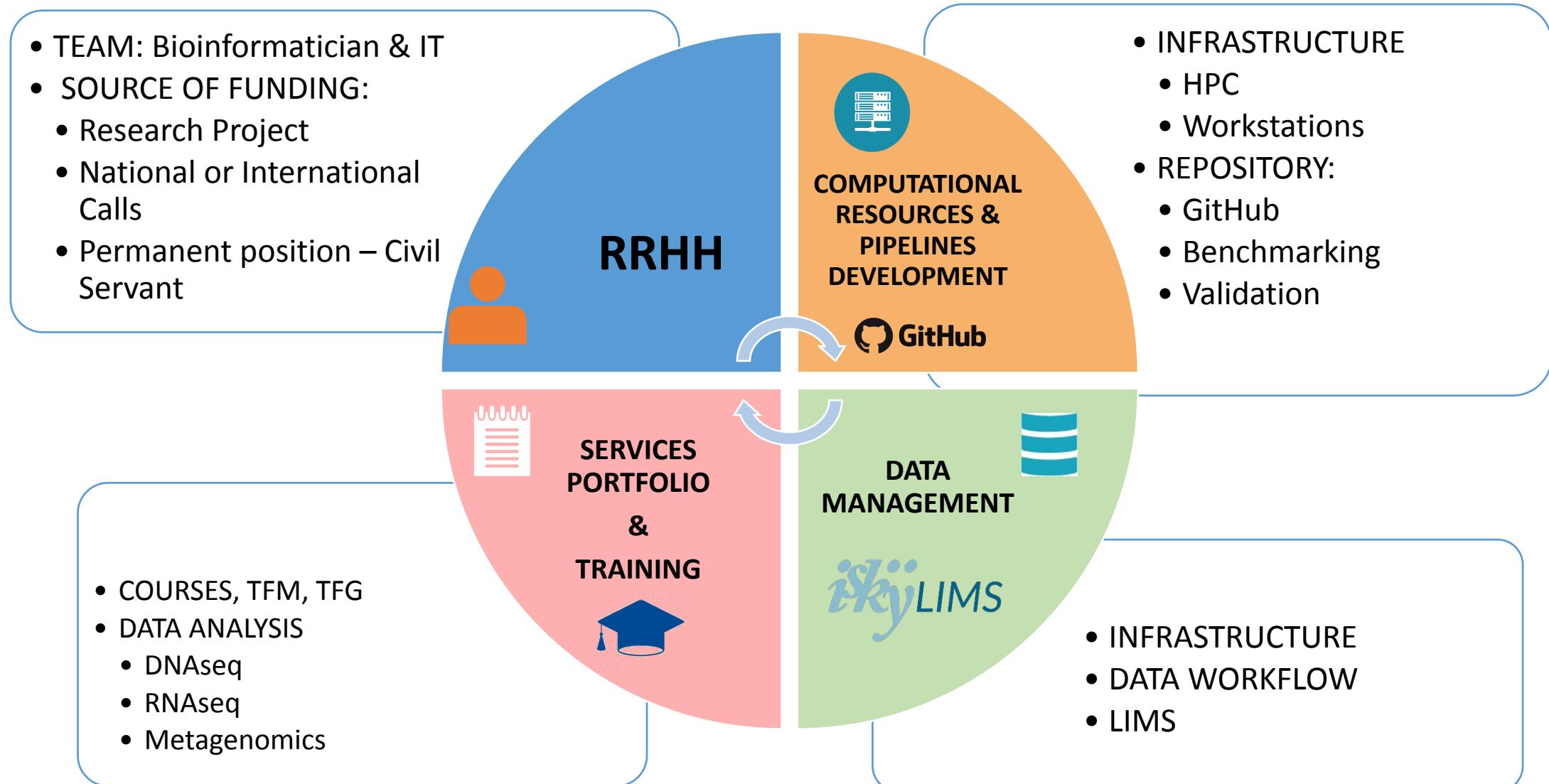
Previous 1 Next

Estadísticas por investigador y por carrera



595 carreras de secuenciación
685 proyectos
30852 muestras de 47 investigadores





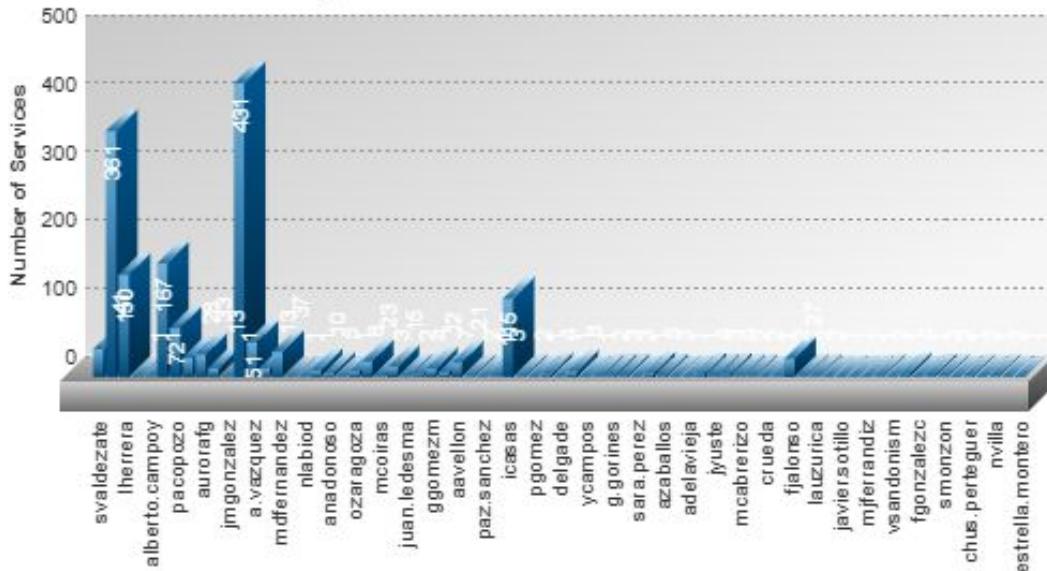
- Genomic Data Analysis
 - Pre-processing and quality analysis (if you ONLY want preprocesing analysis and/or Host removal)
 - Host genome removal
 - Next Generation Sequencing data analysis (all this include QC and preprocessing)
 - DNAseq / cDNAseq: Exome sequencing (WES) / Genome sequencing (WGS) / Targeted sequencing
 - Low-frequency variants detection and annotation for whole genome or sequencing panel (e.g. retinoblastoma gene panel)
 - Eukaria: Variant calling and annotation for a sequencing panel (e.g. epidermolysis gene panel, mouse or rat gene panel)
 - Human: Exome sequencing for variant calling, annotation and inheritance filtering (e.g. Exome sequencing of a human trio (two parents and one child))
 - Human: Whole genome sequencing for SNPs variant calling, annotation and inheritance filtering (e.g.WGS of a human trio)
 - Eukaria (non-human): Variant calling, annotation and SNP-based outbreak analysis (e.g. diploid fungal outbreak)
 - Fungal / bacteria / virus : Variant calling, annotation and SNP-based outbreak analysis (e.g. haploid fungal outbreak)
 - Bacteria: De novo genome assembly and annotation
 - Bacteria: In-depth analysis of Mycobacterium species genomes (e.g. M. tuberculosis. M. bovis)
 - Bacteria: Plasmid analysis and characterization
 - Bacteria: Multi-Locus Sequence Typing (MLST), analysis of virulence factors, antimicrobial resistance, and plasmids characterization
 - Bacteria: Core genome or whole genome Multi-Locus Sequence Typing analysis (cg/wgMLST)
 - Viral: Genomic reconstruction, reference mapping and variant calling and/or de novo assembly
 - Viral Flu: Influenza fragment reconstruction and variant detection
- RNAseq: Transcriptome sequencing
 - Differential Gene Expression (DEG)
 - Differential transcript expression (DET)
 - Differential miRNA expression (DEM)
 - Gene expression changes over a series of time points
- Metagenomics and targeted metagenomics
 - Taxonomic based Identification and classification of organisms in complex communities
 - Alignment of de novo assembly contigs to database
 - Bacteria: 16S rRNA gene analysis to assess bacterial diversity
 - Viral: Detection and characterization of viral genomes within metagenomic data

Number of services: 2012 – 2024

12 AÑOS

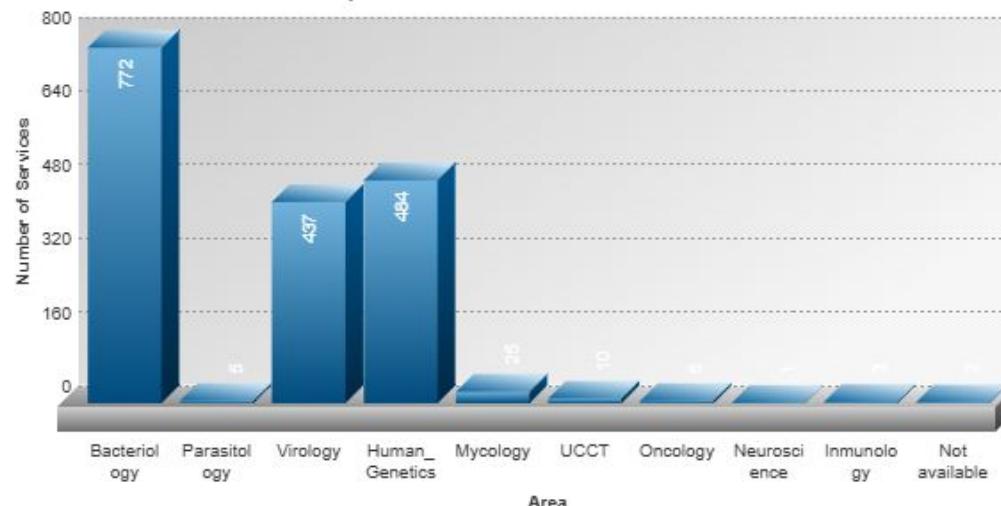
Requested Services by:

For the period between 2012-01-01 and 2024-11-10



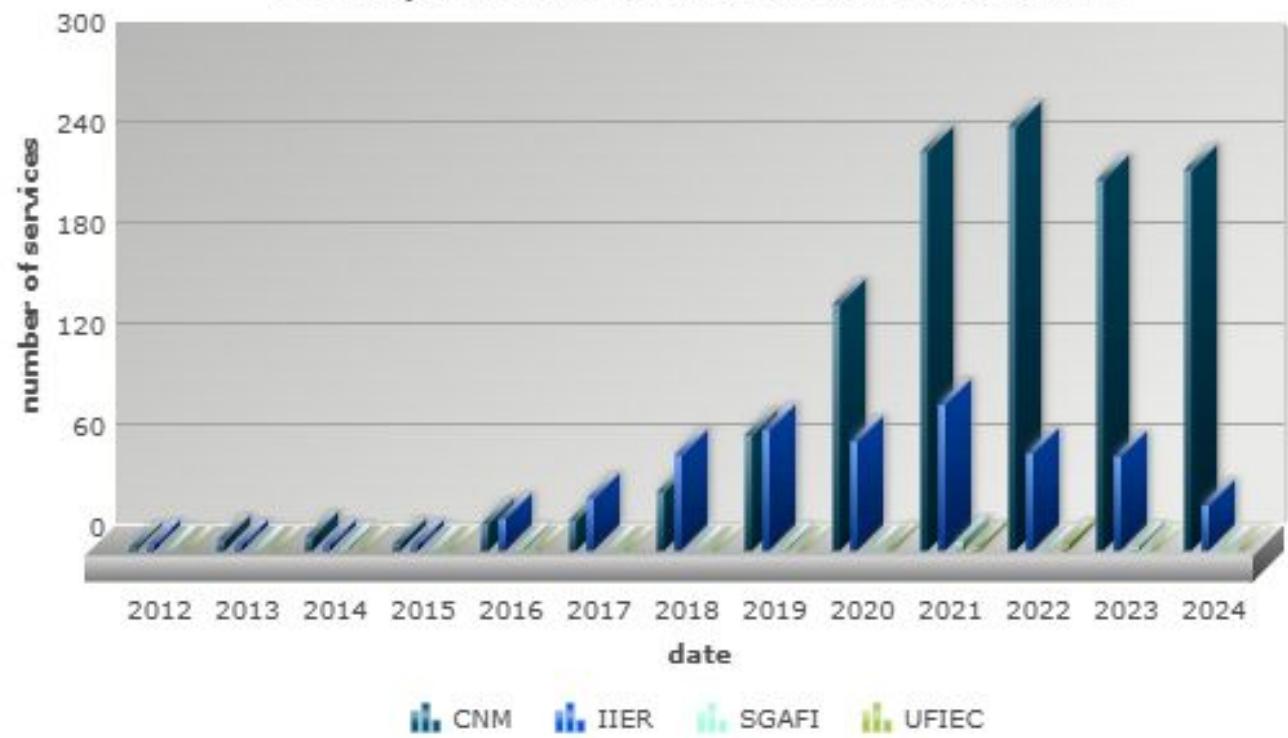
Services requested per Area

For the period between 2012-01-01 and 2024-11-10



Services requested by center

For the period between 2012-01-01 and 2024-11-10



Training

Courses

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

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

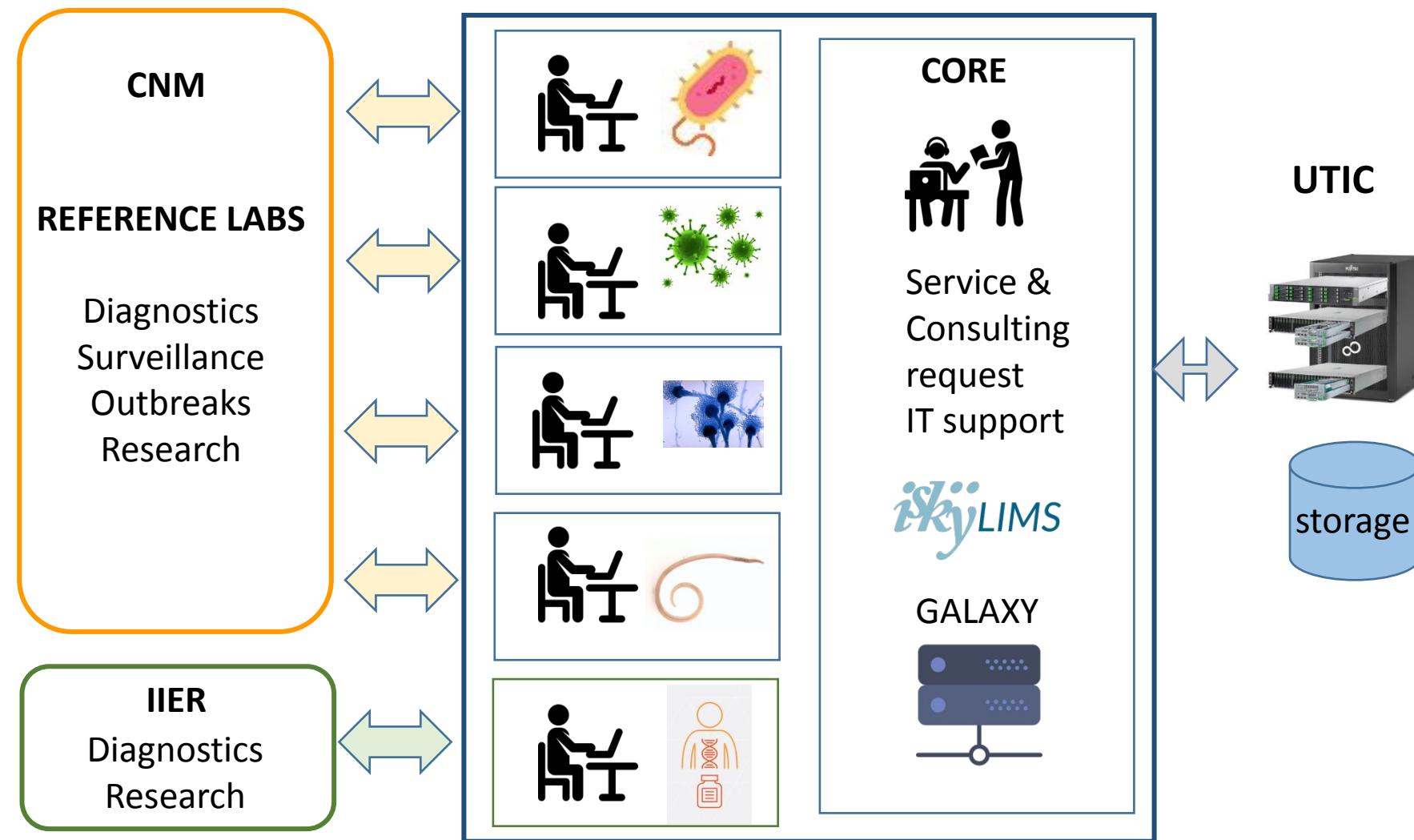
Análisis de genomas virales a través de la plataforma Galaxy, 2024 (3 edition)

Master & Grade Students

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

Hospitals Students

Roadmap: BU-ISCIII Model



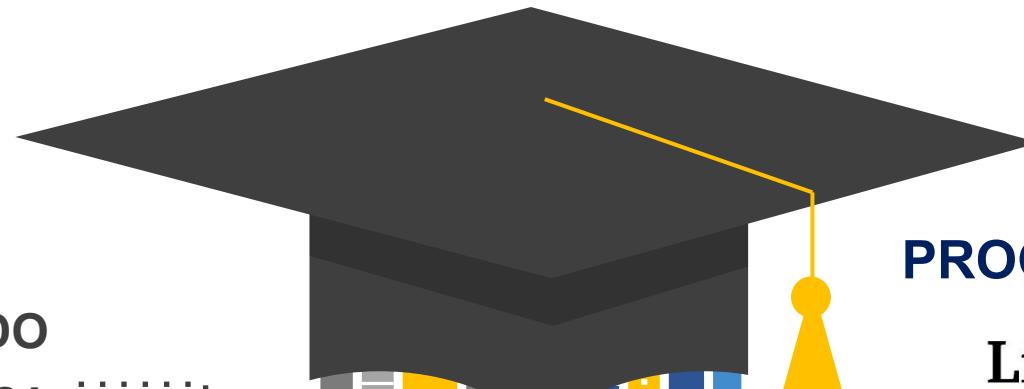
FORMACIÓN EN BIOINFORMÁTICA

Universidad
Barcelona.

GRADO

BIOINFORMÁTICA

- Biología
- Biotecnología
- Bioquímica
- Medicina
- Matemáticas
- Química
- Física
- Informática
- Telecomunicaciones



PROGRAMACIÓN



MASTER EN BIOINFORMÁTICA

DOCTORADO

BIOLOGÍA – ANÁLISIS DE DATOS

¿Dónde trabaja un Bioinformático?



UNIVERSIDAD

Biociencias
Informática

CENTRO DE INVESTIGACIÓN



EMPRESA

Bioinformática
Genética
Genómica

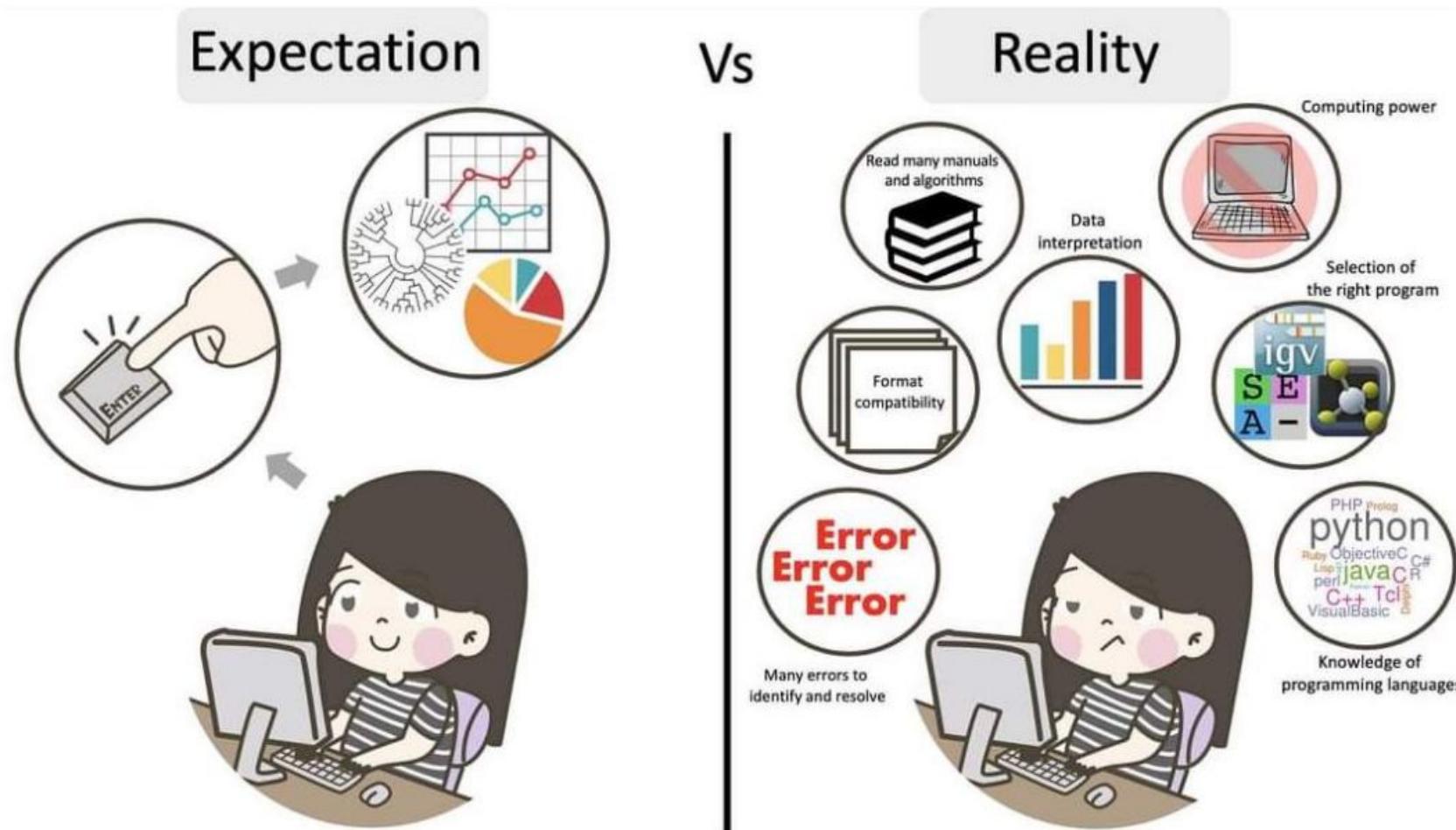
Biomedicina
Agricultura
Alimentación

HOSPITAL

BIOINFORMÁTICO CLÍNICO
Genética
Oncología
Cardiología

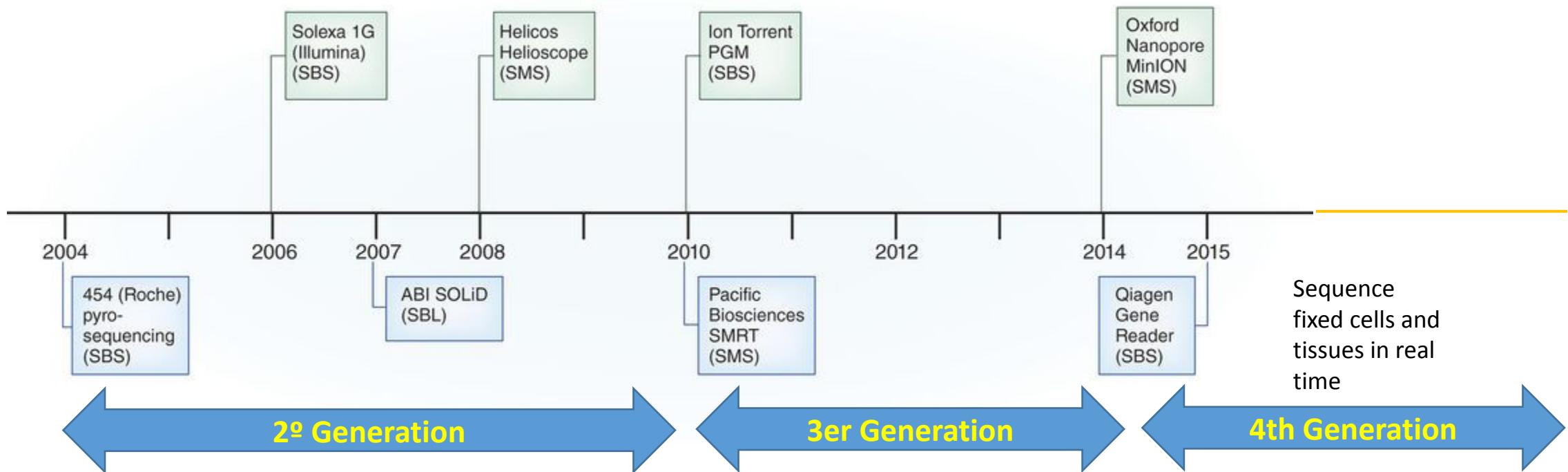
The truth about bioinformatics

.image-100[

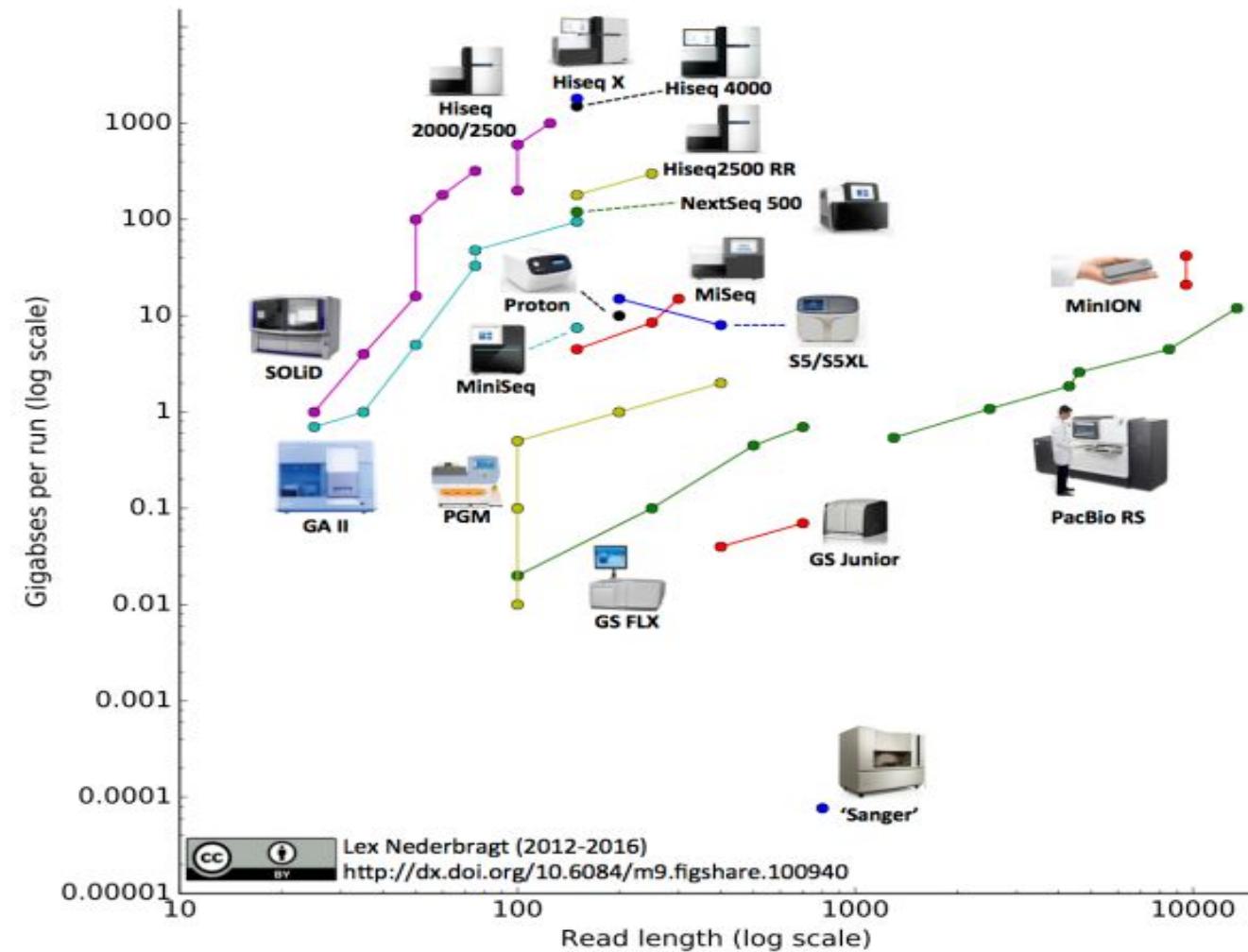


High throughput sequencing in clinical virology, Applications

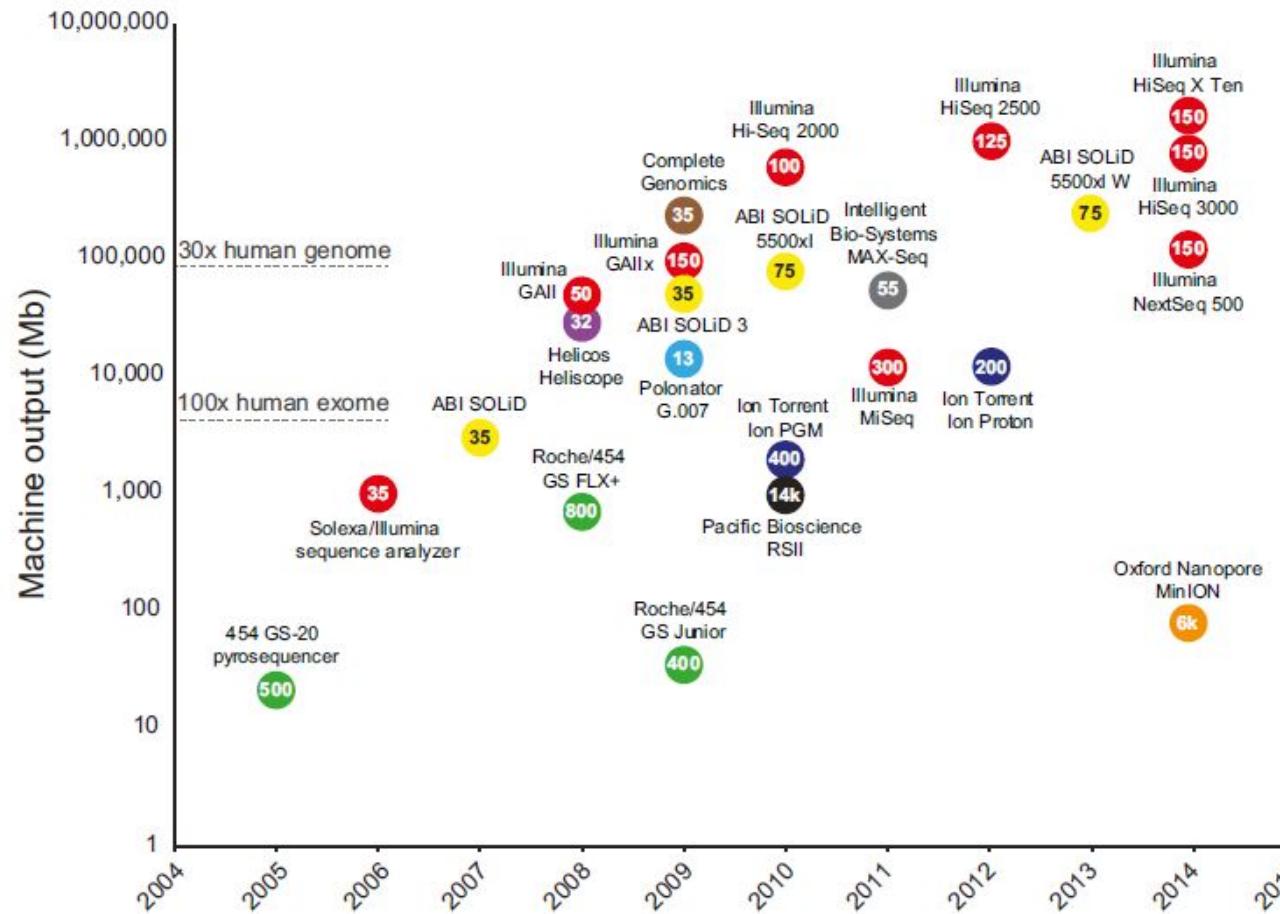
NGS Platforms - Timeline



High-Throughput Sequencing Technologies



High-Throughput Sequencing Technologies

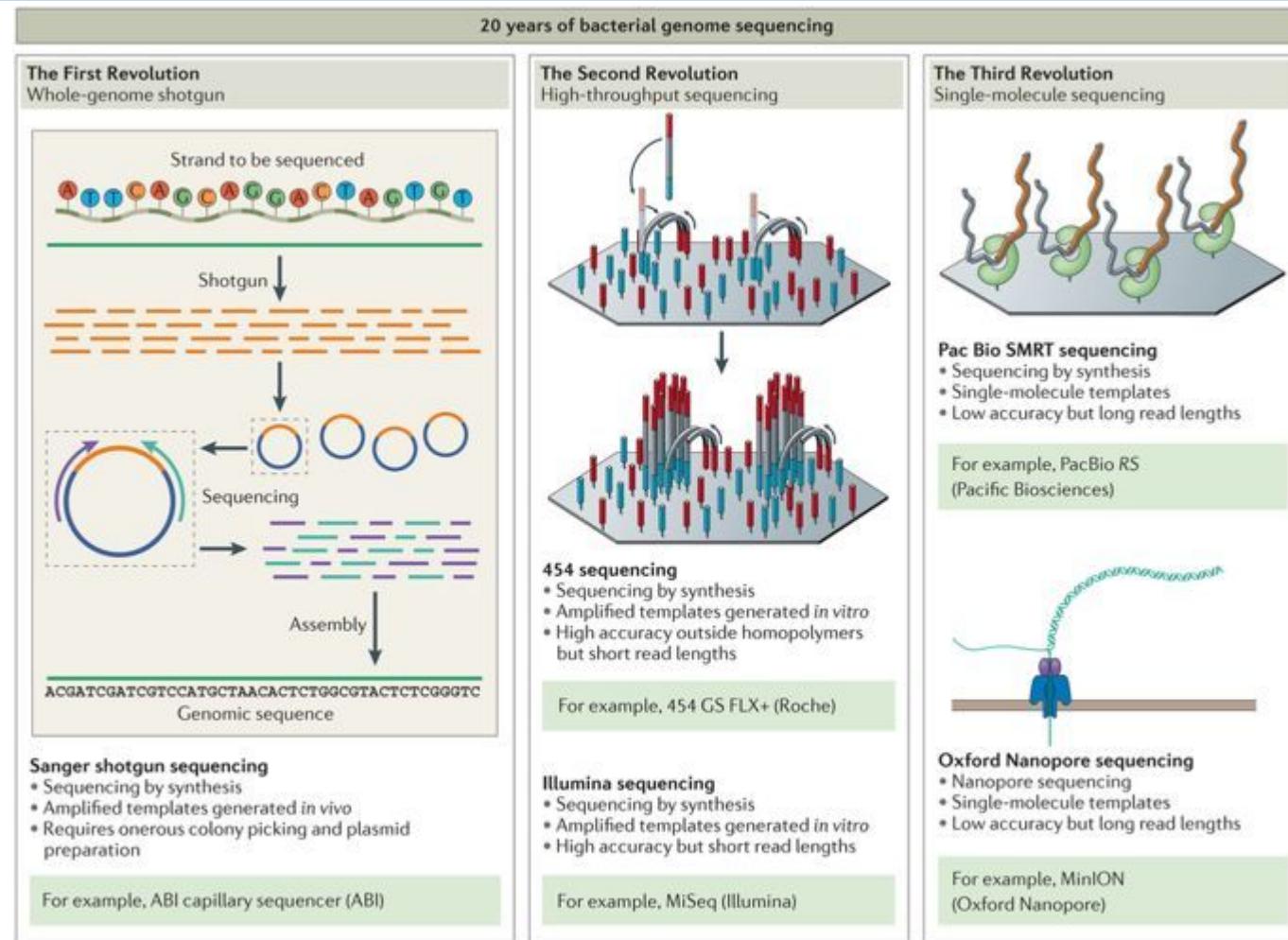


Numbers inside data points denote current read lengths.
Sequencing platforms are color coded.

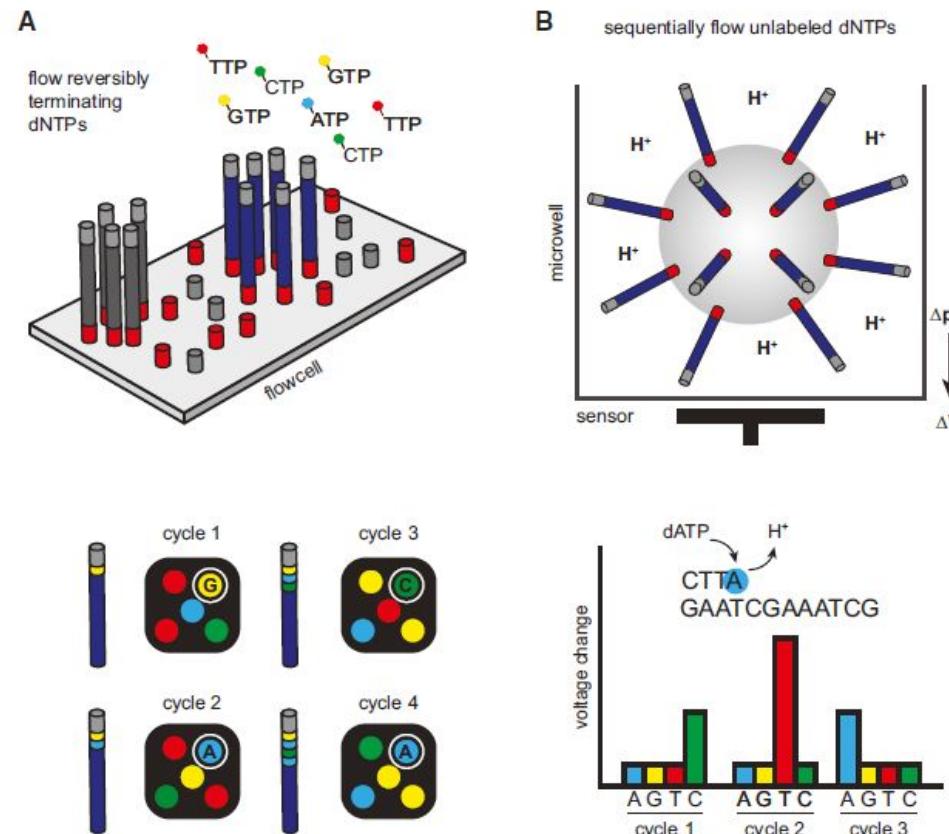
Reuter et al., Mol Cell 2015

High-Throughput Sequencing Technologies

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



The Second-generation Sequencing Technologies

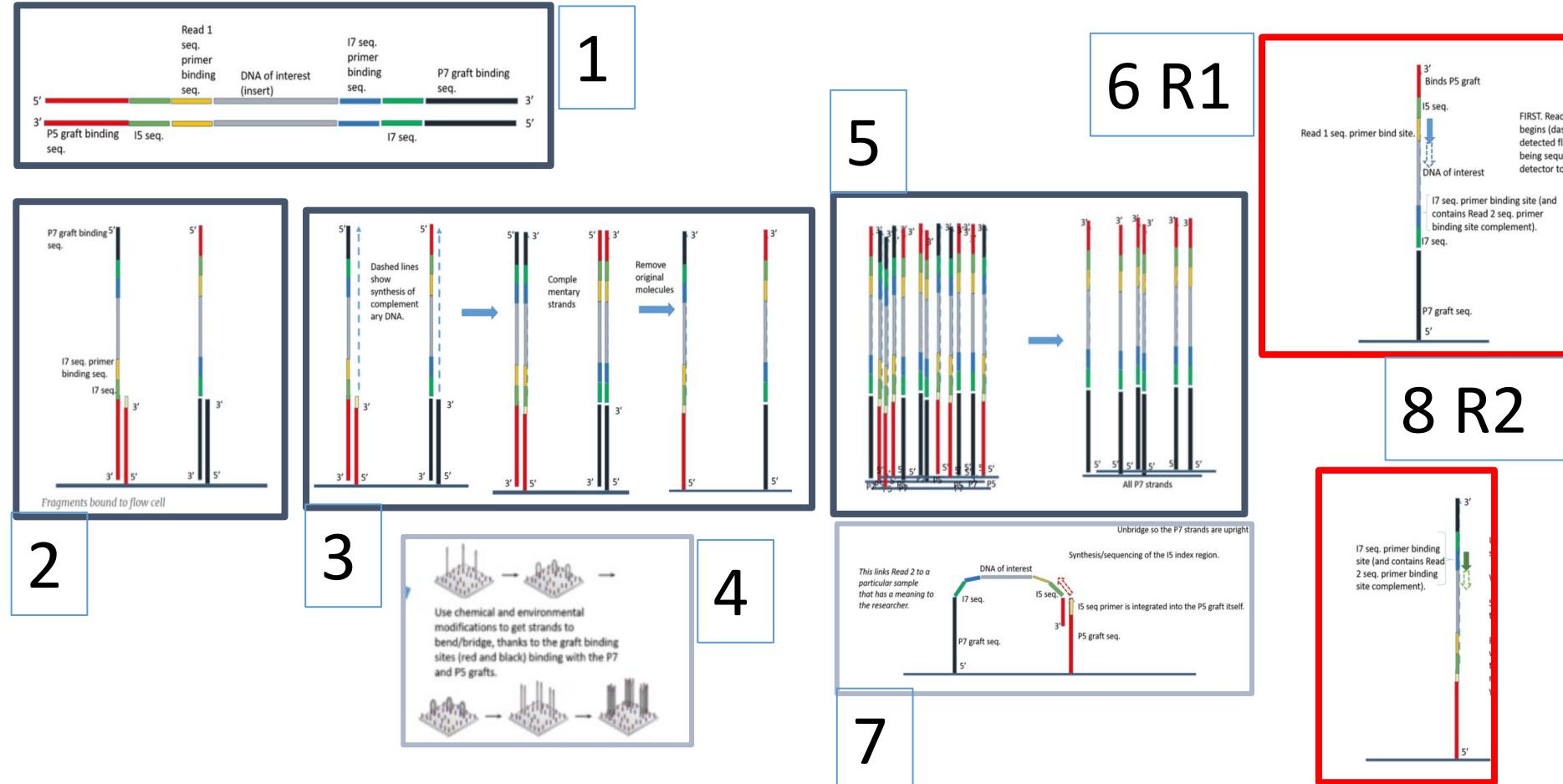
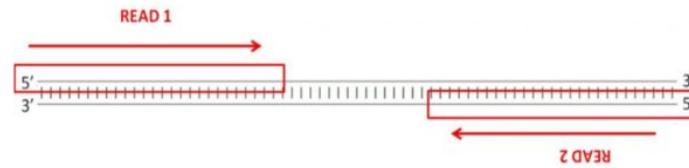


Clonal Amplification-Based Sequencing Platforms

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

(B) Ion Torrent's semiconductor sequencing method.

Illumina sequencing



Illumina benchtop sequencers

Pervez et al., BioMed Research International 2022

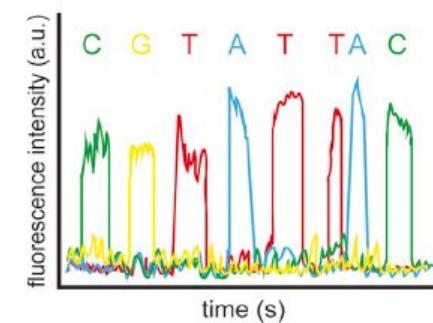
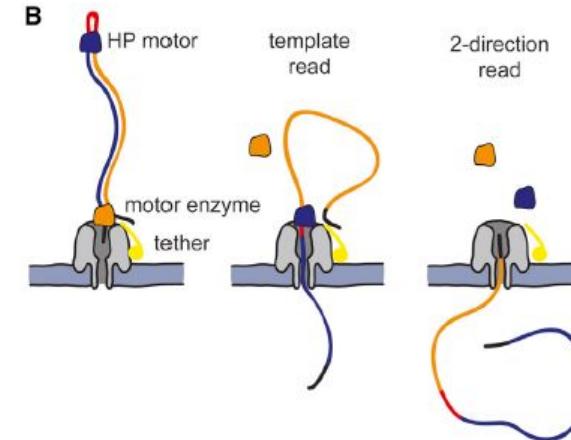
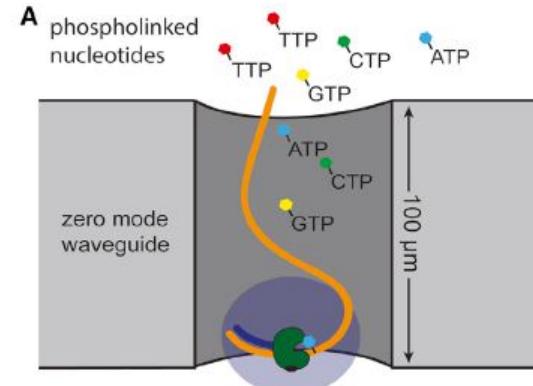
Methods/applications	iSeq 100	MiniSeq	MiSeq series	NextSeq 550 series	Next Seq 1000 & 2000
Ideal for	Every size lab	TG sequencing	Long read applications	Exome and transcriptome sequencing	miRNA and sRNA analysis
Major applications	sWGS (microbes) and TGS	iSeq 100+TG EP and 16S MS	iSeq 100+16S MGS	iSeq 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

Illumina Production Scale Sequencers

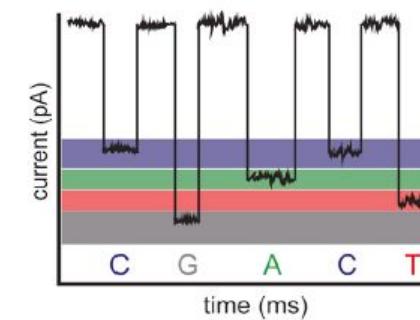
Pervez et al., BioMed Research International 2022

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+IWGS
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

The Third-generation Sequencing Technologies



Pacific Bioscience's SMRT sequencing



Oxford Nanopore's sequencing strategy

Reuter et al., Mol Cell 2015

Single Molecule Sequencing Platforms

PacBio sequencing and its applications

Rhoads & Au, Gen Prot Bioinf 2015



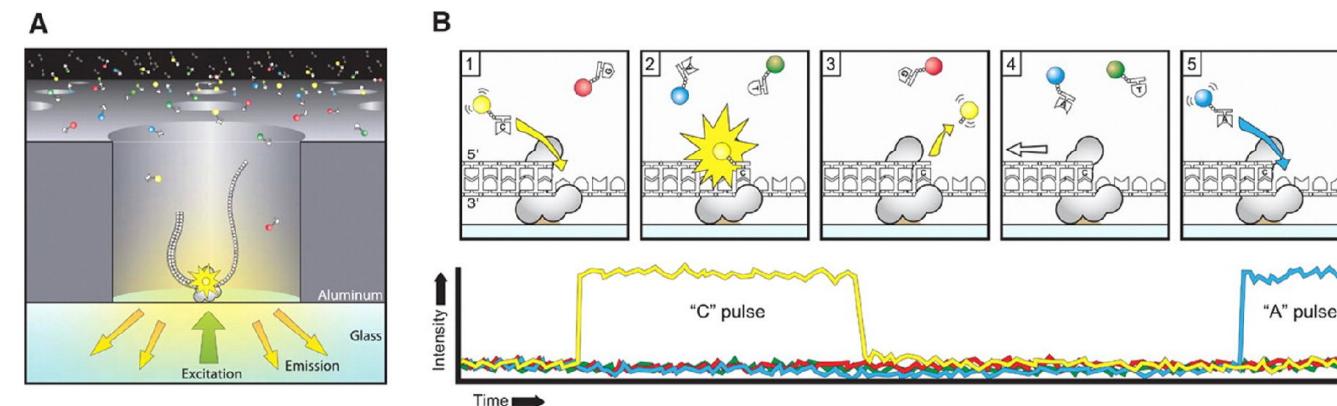
SMRTbell template: is a closed, single-stranded circular DNA that is created by ligating hairpin adaptors to both ends of a target dsDNA

Sequencing by light pulses: The replication processes in all ZMWs of a SMRTcell are recorder by a movie of light pulses, and the pulses corresponding to each ZMW can be interpreted to be a sequence of bases (**continuous long read, CLR**).

Both strands can be sequenced multiple times (passes) in a single CLR. CLR can be split to multiple reads (subreads) and CCS is the consensus sequence of multiple subreads



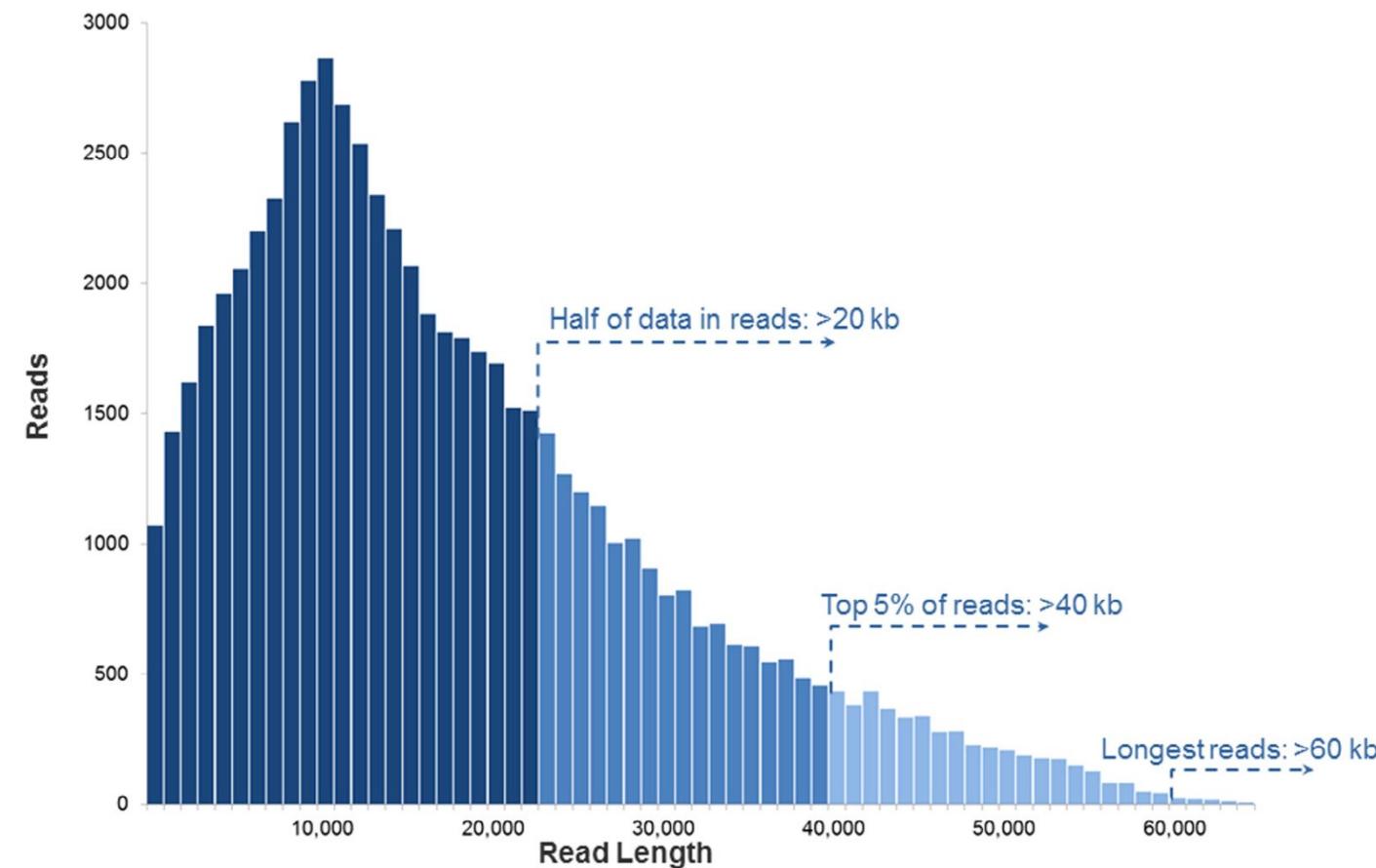
A single SMRT cell: this contains 150000 ZMWs (zero-mode waveguide). A SMRTbell diffuses into a ZMW. Approx 35000 -75000 ZMWs produce a read in a run lasting 0,5-4h resulting in 0,5-1Gb.



PacBio sequencing and its applications

Rhoads & Au, Gen Prot Bioinf 2015

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.



PacBio sequencing and its applications

Rhoads & Au, Gen Prot Bioinf 2015

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

Species	Method	Tools	SMRT cells	Coverage	Contigs	Achievements	Ref.
<i>Clostridium autoethanogenum</i>	PacBio	HGAP	2	179×	1	21 fewer contigs than using SGS; no collapsed repeat regions (≥ 4 using SGS)	[7]
<i>Potentilla micrantha</i> (chloroplast)	PacBio	HGAP, Celera, minimus2, SeqMan	26	320×	1	6 fewer contigs than with Illumina; 100% coverage (Illumina: 90.59%); resolved 187 ambiguous nucleotides in Illumina assembly; unambiguously assigned small differences in two > 25 kb inverted repeats	[33]
<i>Escherichia coli</i>	PacBio	PBcR, MHAP, Celera, Quiver	1	85×	1	4.6 CPU hours for genome assembly (10× improvement over BLASR)	[31]
<i>Saccharomyces cerevisiae</i>	PacBio	PBcR, MHAP, Celera	1p	117×	21	27 CPU hours for genome assembly (8× improvement over BLASR); improved current reference of telomeres	[31]
<i>Arabidopsis thaliana</i>	PacBio	PBcR, MHAP, Celera	46	144×	38	1896 CPU hours for genome assembly	[31]
<i>Drosophila melanogaster</i>	PacBio	PBcR, MHAP, Celera, Quiver	42	121×	132	1060 CPU hours for genome assembly (593× improvement over BLASR); improved current reference of telomeres	[31]
<i>Homo sapiens</i> (CHM1hert)	PacBio	PBcR, MHAP, Celera	275	54×	3434	262,240 CPU hours for genome assembly; potentially closed 51 gaps in GRCh38; assembled MHC in 2 contigs (60 contigs with Illumina); reconstructed repetitive heterochromatic sequences in telomeres	[31]
<i>Homo sapiens</i> (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]
<i>Melopsittacus undulatus</i>	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]
<i>Vibrio cholerae</i>	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]
<i>Helicobacter pylori</i>	PacBio	HGAP, Quiver, PGAP	8 per strain	446.5× average among strains	1 per strain	1 complete contig for each of 8 strains; methylation analysis associated motifs with genotypes of virulence factors	[35]

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

PacBio sequencing and its applications

Rhoads & Au, Gen Prot Bioinf 2015

Advantage

Closes gaps and completes genomes due to longer reads

Identifies non-SNP SVs

Achievements

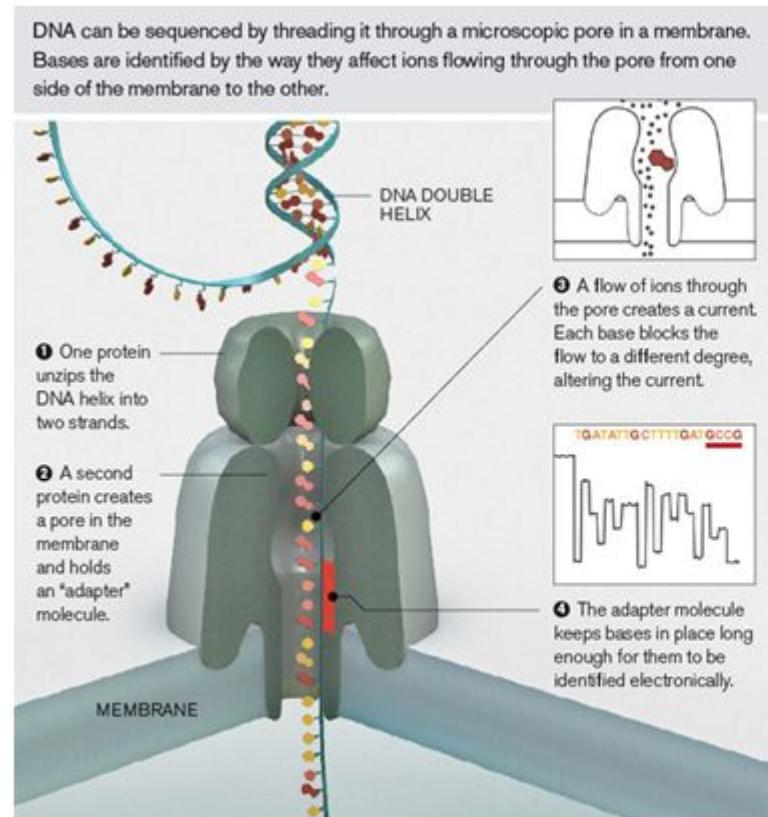
Produced highly-contiguous assemblies of bacterial and eukaryotic genomes

Discovered STRs (short tandem repeats)

Limitations

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

Nanopore-based fourth-generation DNA sequencing technology. ONT, Oxford Nanopore Technologies



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

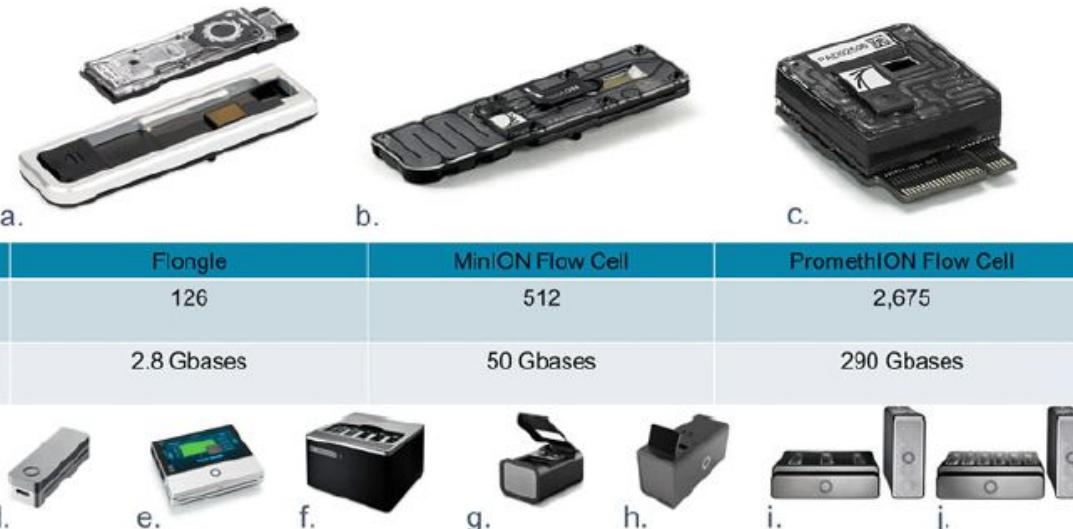
Nanopore sequencing also offers, for the first time, direct RNA sequencing, as well as PCR or PCR-free cDNA sequencing.

<https://nanoporetech.com/applications/dna-nanopore-sequencing>

Feng et al , Gen Prot Bioinf 2015



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



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.

Nanopore 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.

Sequencing technologies and its applications

Rhoads & Au, Gen Prot Bioinf 2015

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 × 10 ⁴ on average	13	$3.5\text{--}7.5 \times 10^4$	0.5–4 h	0.40–0.80	[5,12,15]
Oxford Nanopore MinION	3rd	2–5 × 10 ³ on average	38	$1.1\text{--}4.7 \times 10^4$	50 h	6.44–17.90	[22,23]

Comparison of various high-performing sequencing instruments

Pervez et al., BioMed Research International 2022

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

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
<i>Sanger sequencing</i>				
ABI 3500/3730	0.0003	Up to 1 kb	Read accuracy and length	Cost and throughput
<i>Illumina</i>				
MiniSeq	1.7–7.5	1×75 to ×150	Low initial investment	Run and read length
MiSeq	0.3–15	1×36 to 2×300	Read length, scalability	Run length
NextSeq	10–120	1×75 to 2×150	Throughput	Run and read length
HiSeq (2500)	10–1000	×50 to ×250	Read accuracy, throughput,	High initial investment, run
NovaSeq 5000/6000	2000–6000	2×50 to ×150	Read accuracy, throughput	High initial investment, run
<i>IonTorrent</i>				
PGM	0.08–2	Up to 400	Read length, speed	Throughput, homopolymers ^c
S5	0.6–15	Up to 400	Read length, speed,	Homopolymers ^c
Proton	10–15	Up to 200	Speed, throughput	Homopolymers ^c
<i>Pacific BioSciences</i>				
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
<i>Oxford Nanopore</i>				
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.

Advantages & disadvantages of sequencing platforms

Pervez et al., BioMed Research International 2022

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 3 rd GS+	High sequencing error rate 10–15% in the PacBio and 5–20% in the ONT Fresh DNA required 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	

Advantages & disadvantages for short vs. Long read sequencing

	Advantages	Limitations
Short-read sequencing	<ul style="list-style-type: none">Higher sequence fidelityCheapCan sequence fragmented DNA	<ul style="list-style-type: none">Not able to resolve structural variants, phasing alleles or distinguish highly homologous genomic regionsUnable to provide coverage of some repetitive regions
Long-read sequencing	<ul style="list-style-type: none">Able to sequence genetic regions that are difficult to characterize with short-read seq due to repeat sequencesAble to resolve structural rearrangements or homologous regionsAble to read through an entire RNA transcript to determine the specific isoformAssists <i>de novo</i> genome assembly	<ul style="list-style-type: none">Lower per read accuracyBioinformatic challenges, caused by coverage biases, high error rates in base allocation, scalability and limited availability of appropriate pipelines

Advantages 3rd GS over 2nd GS

Pervez et al., BioMed Research International 2022

- 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

2nd GS and 3rd GS

Pervez et al., BioMed Research International 2022

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

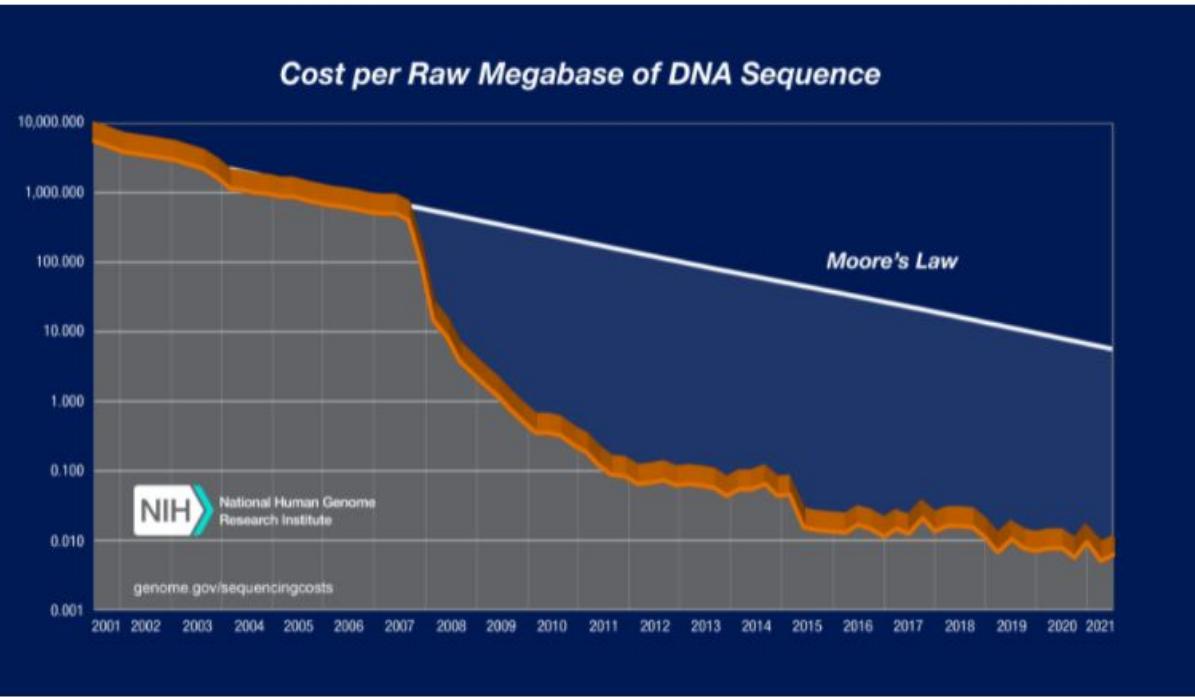
TABLE 6: Overall costs for sequencing a human genome [49].

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

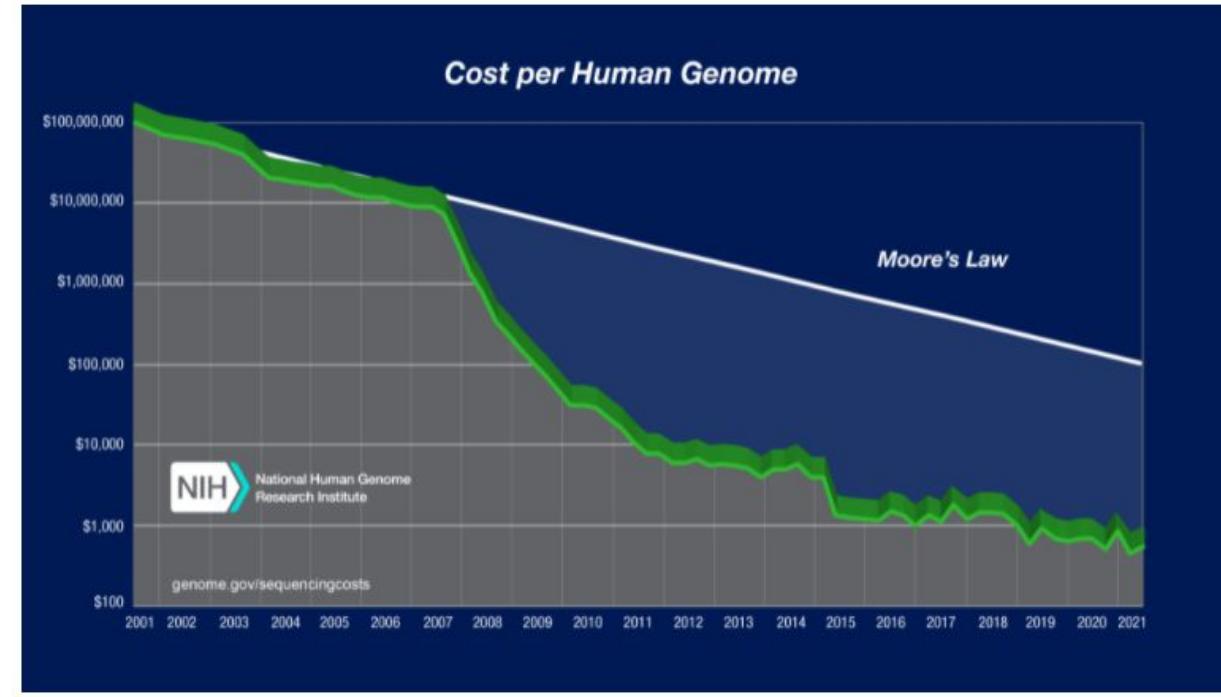
NGS PLATFORMS, main characteristics

- Número de bases que secuencia
- Número 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

Sequencing costs



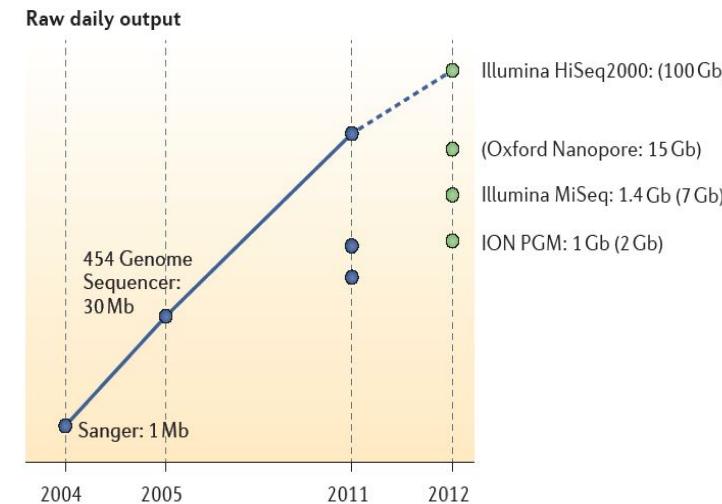
Sequencing cost per megabase - 2021



Cost per genome data - 2021

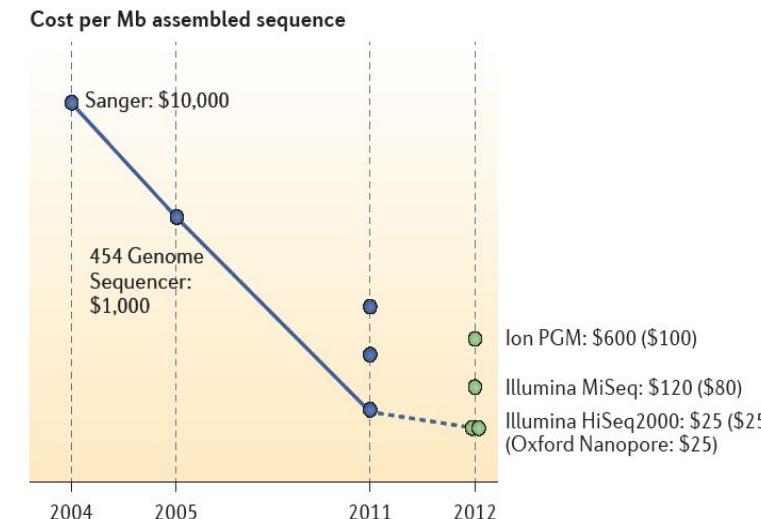
Sequencing platforms in Microbiology

Raw daily output

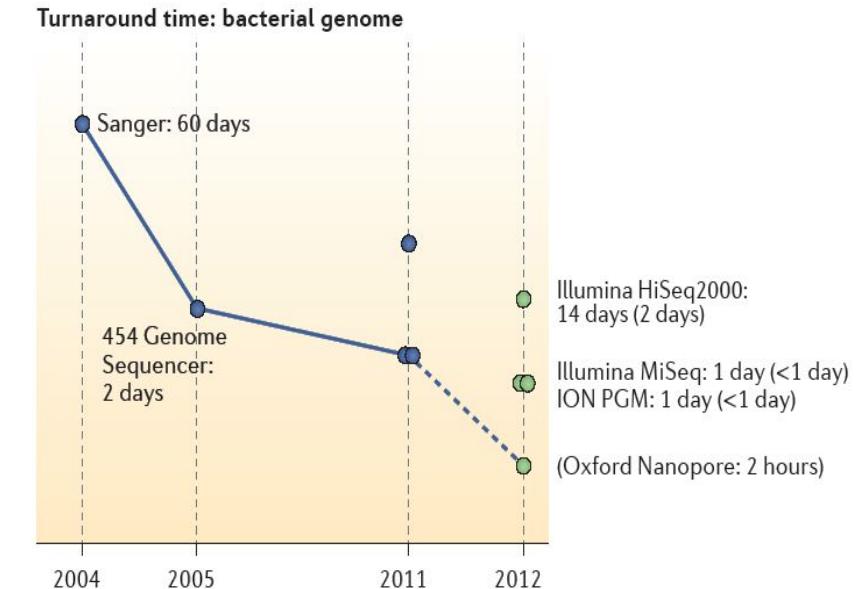


Didelot et al., Nature Genet Review 2012, 13:601-612

Cost per Mb assembled sequence



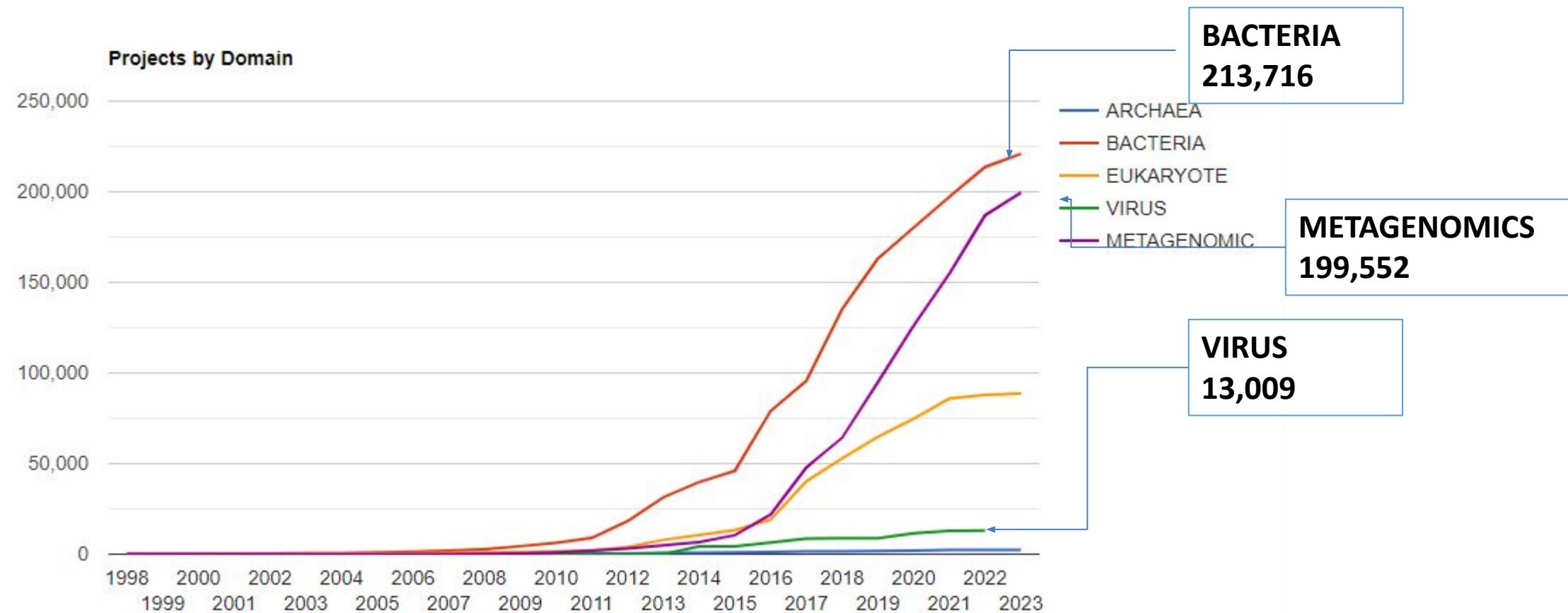
Turnaround time: bacterial genome



Sequencing projects

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

GOLD, Genome Online DataBase



APPLICATIONS OF NGS IN VIROLOGY

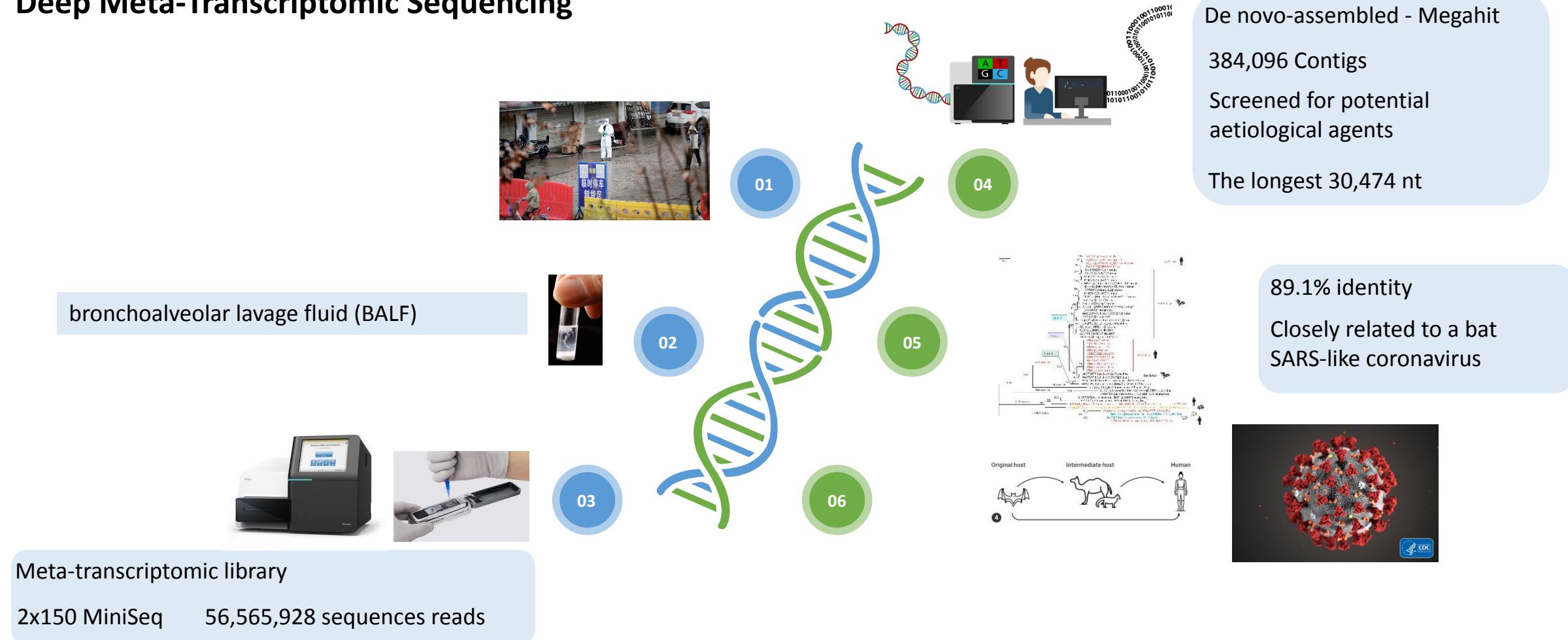
- ◆ Detección de patógenos virales no identificados en la muestra
- ◆ Identificación de virus nuevos
- ◆ Identificación de virus en muestras tumorales
- ◆ Caracterización del Viroma de un organismo
- ◆ Secuenciación del genoma viral completo
- ◆ Estudio de la Variabilidad genómica (Quasispecies)
- ◆ Monitorización de la resistencia a los Antivirales
- ◆ Vigilancia epidemiológica de las infecciones virales
- ◆ Estudio de la Evolución viral
- ◆ Control de calidad de las vacunas virales vivas atenuadas

High throughput sequencing in clinical virology, Applications

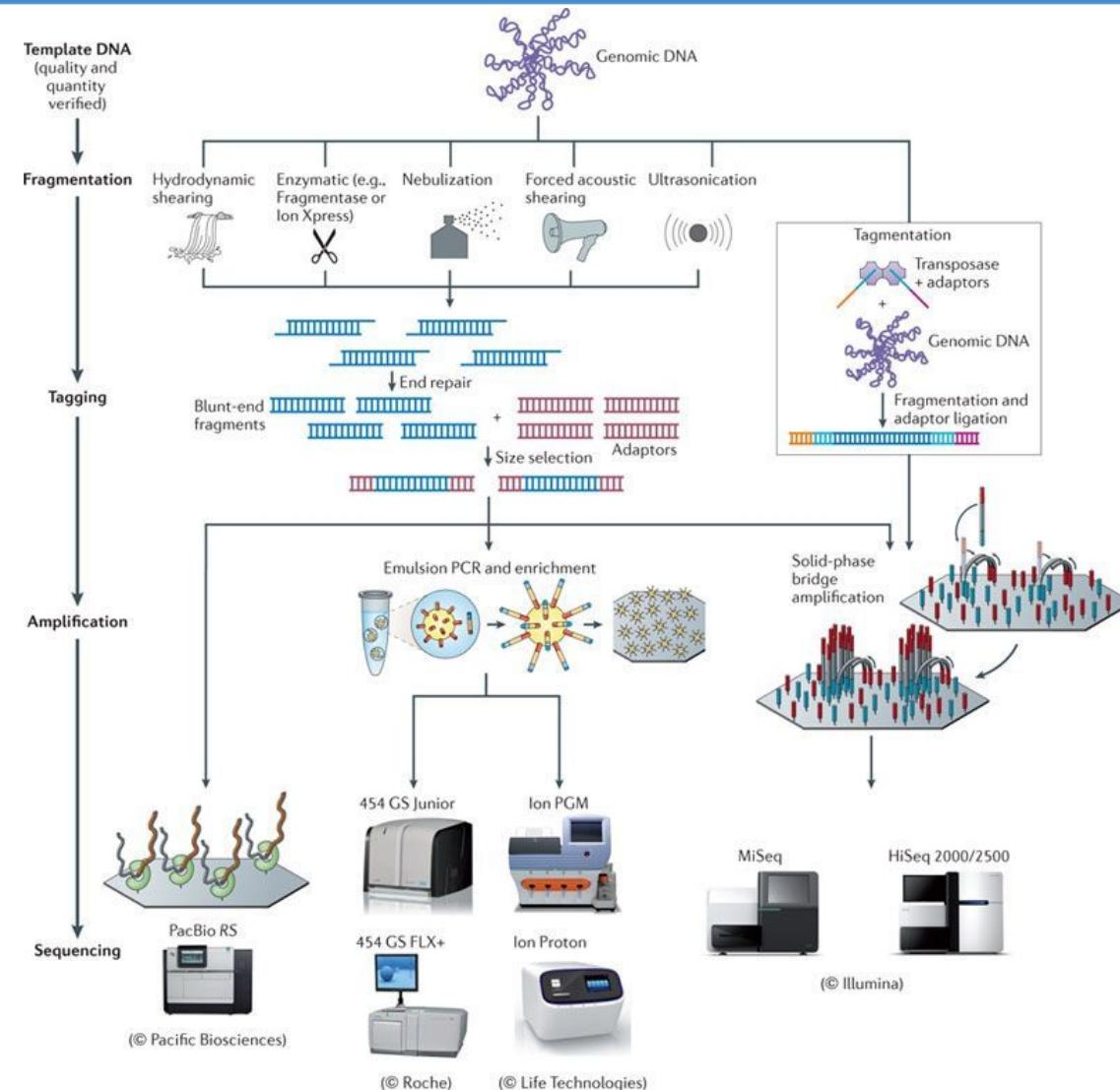
- Virus discovery: Metagenomics, sequence-independent amplification
- Viral genome reconstruction
- Variants identification & quasispecies

Pathogen discovery: new virus – SARS-CoV-2

Deep Meta-Transcriptomic Sequencing



High-throughput sequencing platforms



Loman et al, 2012

Main Steps of Viral Genome Sequencing by NGS or HTS

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

General features of viral genomes

S. No.	Class	Sequenced genomes	Size (Nt)	Proteins
1	DsDNA	414	4697–335,593	6–240
2	SsDNA	230	1360–10,958	6–11
3	DsRNA	61	3090–29,174	2–13
4	SsRNA (+)	421	2343–31,357	1–11
5	SsRNA (-)	81	8910–25,142	5–6

K. V. Chaitanya, Genome and Genomics,
https://doi.org/10.1007/978-981-15-0702-1_1

Genomes of Single Stranded DNA Viruses and their Mosaicism

> BU-ISCIII

Table 1.2 Morphological diversity of single stranded DNA viruses

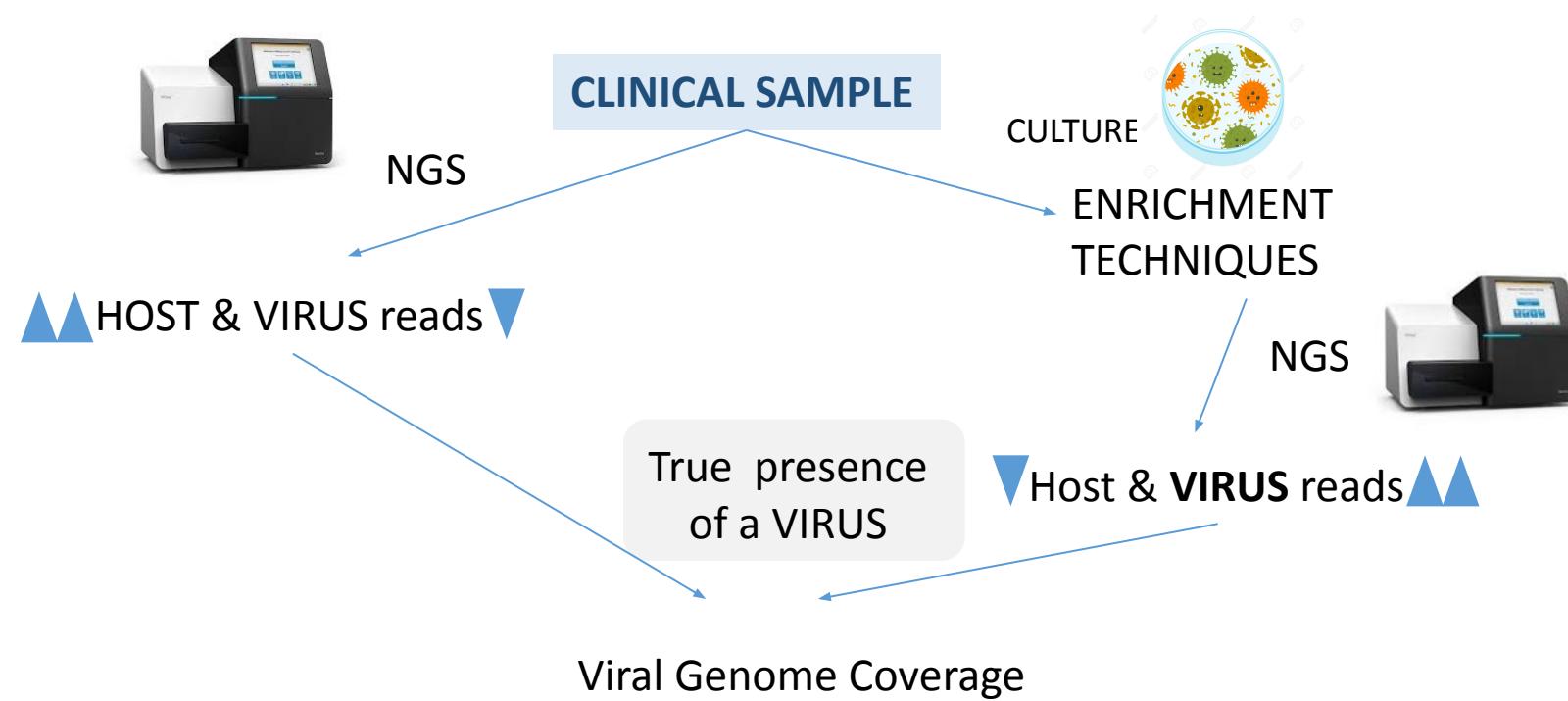
Host virus taxon	Virion morphology	Genome topology	Genome size
Microviridae	Icosahedral	Circular	4.4–6.1
Inoviridae			
Inovirus	Filamentous		5.8–12.4
Plectrovirus	Rod-shaped		4.5–8.2
Pleolipoviridae	Pleomorphic	Circular	7–10.6
Spiraviridae	Coil-shaped	Circular	24.9
Anelloviridae	Icosahedral	Circular	2–4
Bidnaviridae	Icosahedral	Linear, segmented, 6.5 per segment	
Circoviridae	Icosahedral	Circular	1.7–2.3
Geminiviridae	Icosahedral	Circular, segmented 3 per segment	
Nanoviridae	Icosahedral	Circular, segmented 0.98–1.1 per segment	
Parvoviridae	Icosahedral	Linear	4–6.3

Table 1.3 Size of influenza virus segments and the proteins they encode

RNA segment	No. of nucleotides	Encoding protein	No. of amino acids
1	2341	Polymerase PB2	759
2	2341	Polymerase PB1	757
3	2233	Polymerase PA	716
4	1778	Haemagglutinin HA	566
5	1565	Nucleoprotein NP	498
6	1413	Neuraminidase NA	454
7	1027	Matrix protein M1	252
		Matrix protein M2	97
8	890	Non-structural protein NS1	230
		Non-structural protein NS2	121

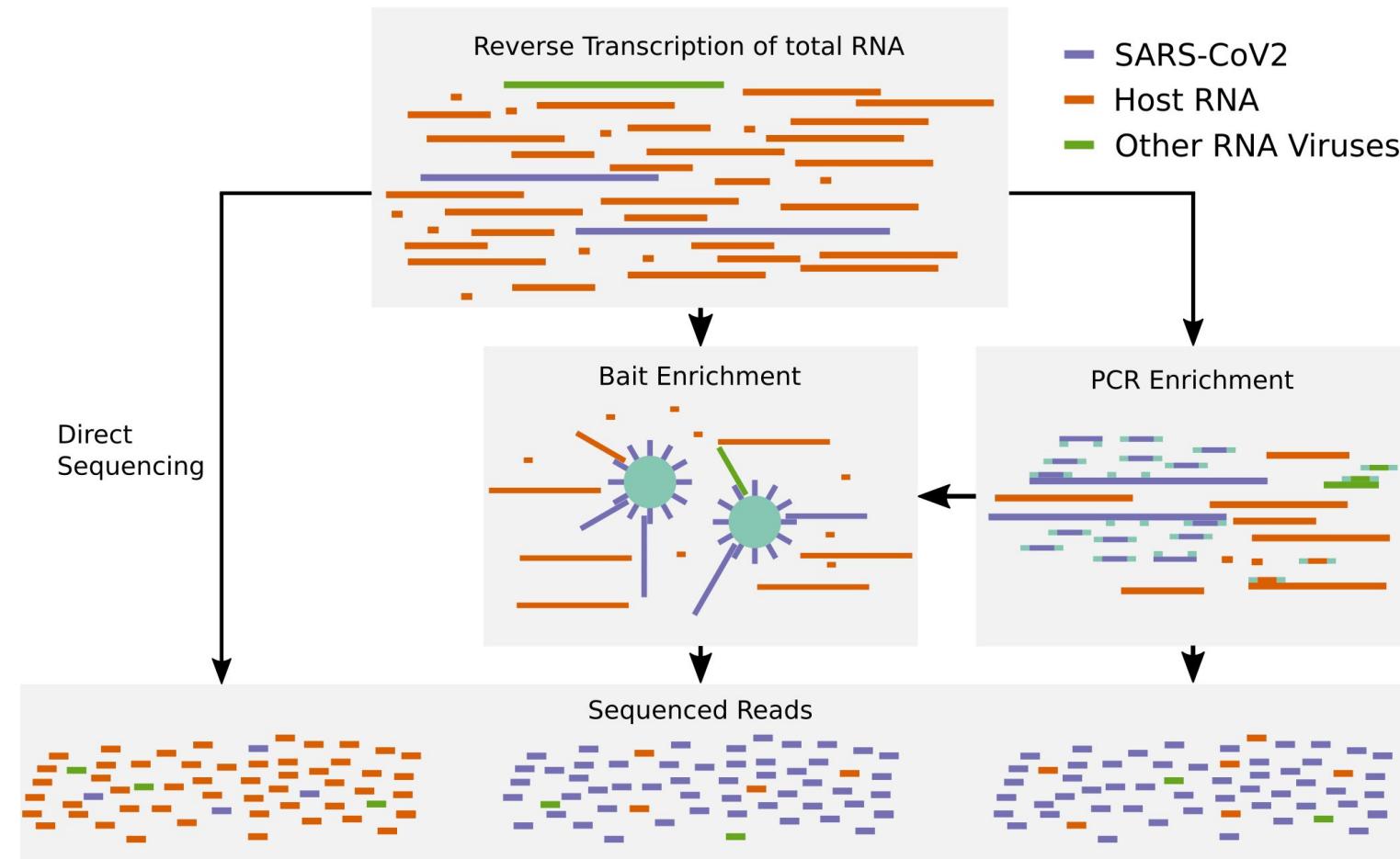
K. V. Chaitanya, Genome and Genomics,
https://doi.org/10.1007/978-981-15-0702-1_1

Viral Genome Sequencing

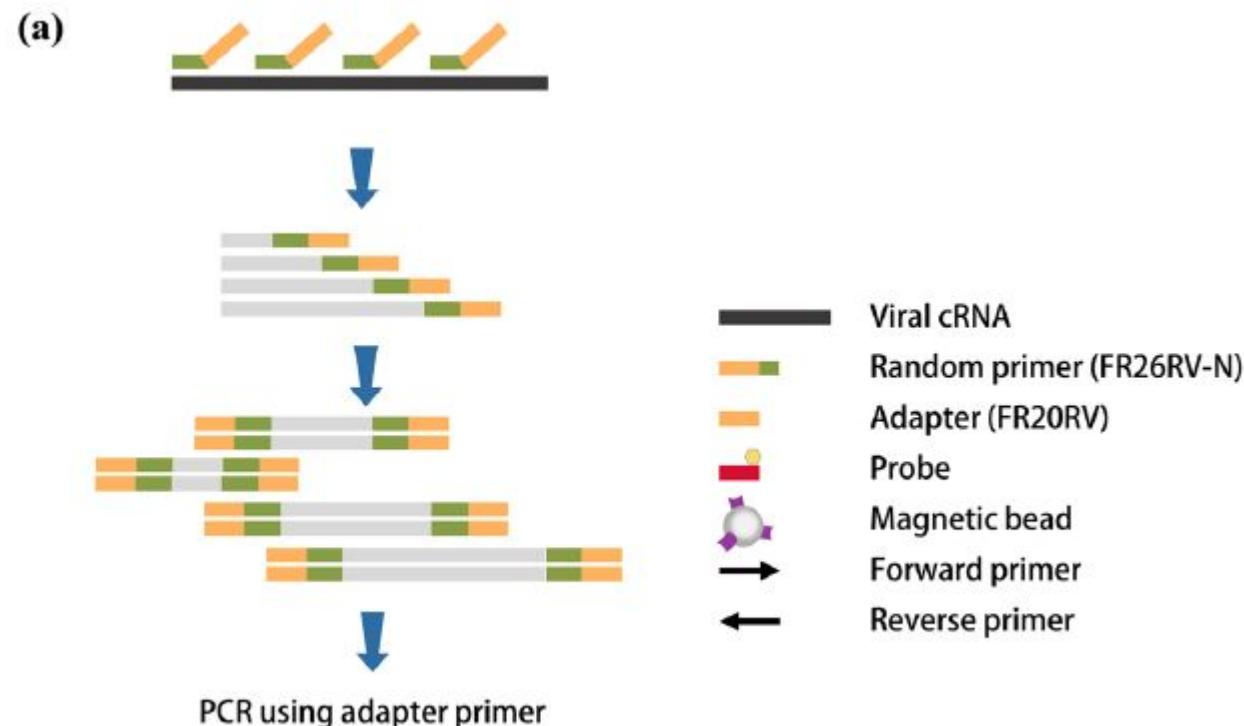


NGS needs a cutoff to determine the true presence of a pathogen versus carry-over or contamination between specimens or other non-specific reads.

Enrichment Techniques



Nucleic Acid Extraction: SISPA NGS method

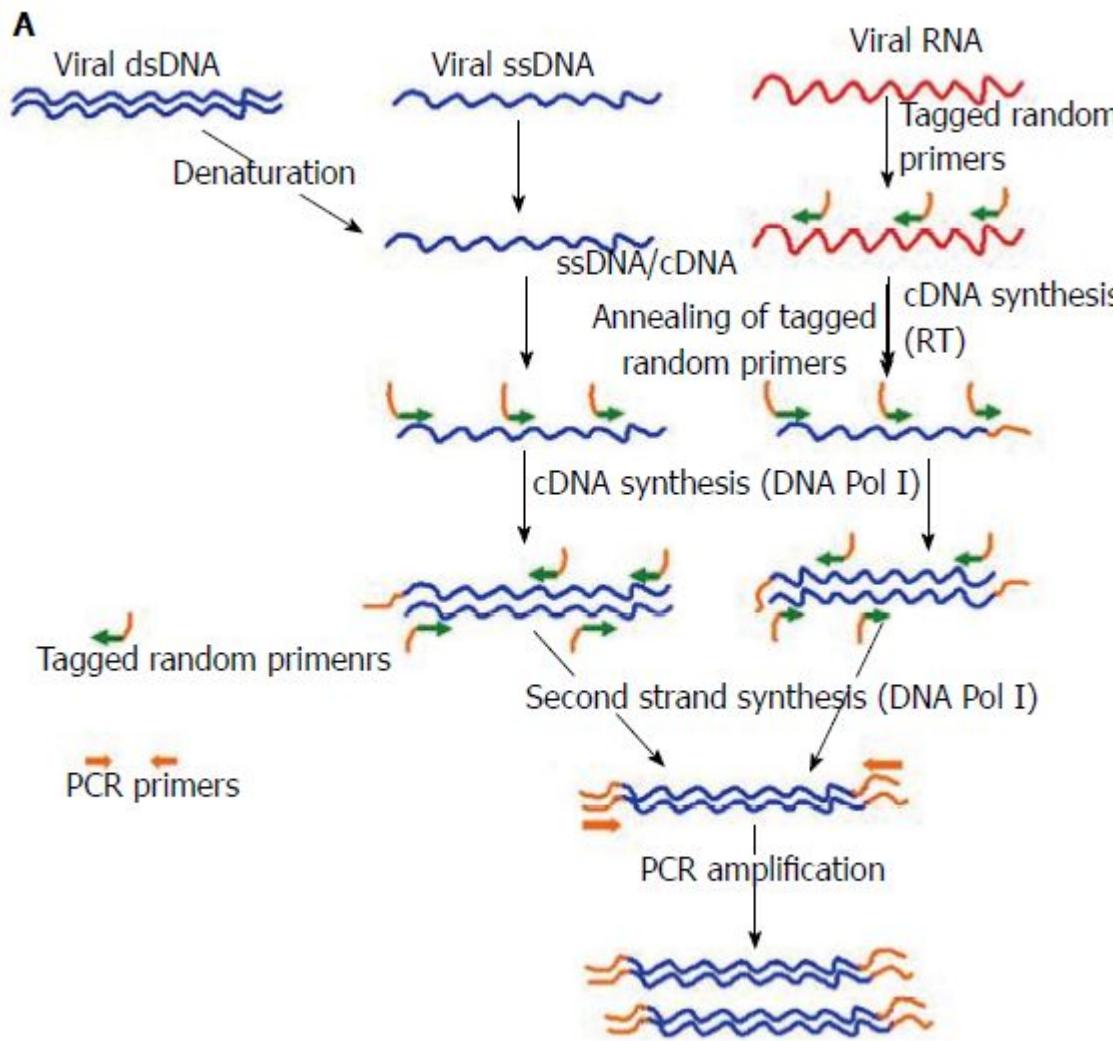


VIRUS DISCOVERY

- Nucleic acid **sequence-independent** amplification approaches
- Next-generation sequencers-based **metagenomic** approaches

RNA was reverse-transcribed using a random primer (FR26RV-N) and then cDNA was amplified using a single primer (FR20RV)

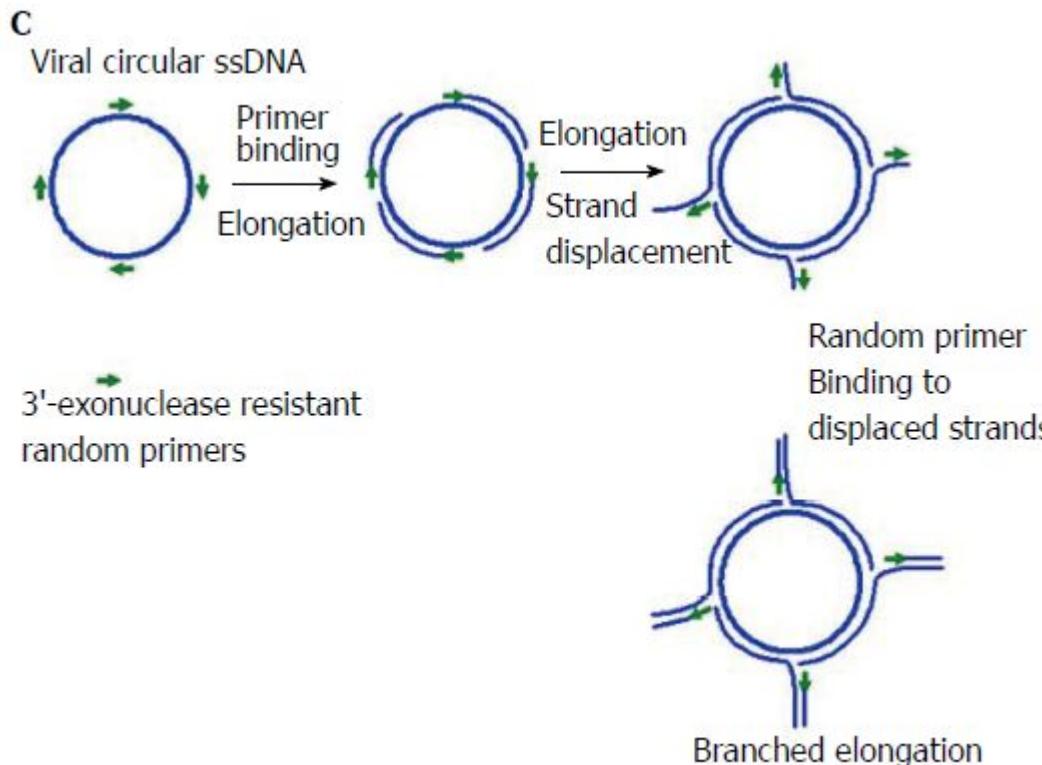
Sequence-independent single-primer amplification



Initially viral RNA and ssDNA is transcribed into complementary DNA (cDNA) using reverse transcriptase (RT) and DNA Pol I respectively, with the help of **tagged-primers having defined sequence at the 5' end while random nucleotides at the 3' end**. Subsequently, second strand synthesis is performed using DNA Pol I (Klenow) to make the cDNA double stranded (dsDNA). Now all the nucleic acids present in the reaction are dsDNA fragments have tagged sequence at their ends. Finally, **anchored dsDNA is amplified with primers annealing to the adapter specific sequences**, PCR product are checked and ready for analysis through cloning-sequencing or direct sequencing through next-generation sequencers (NGS);

Datta et al., World J Virol 2015
DOI: 10.5501/wjv.v4.i3.265

Rolling circle amplification

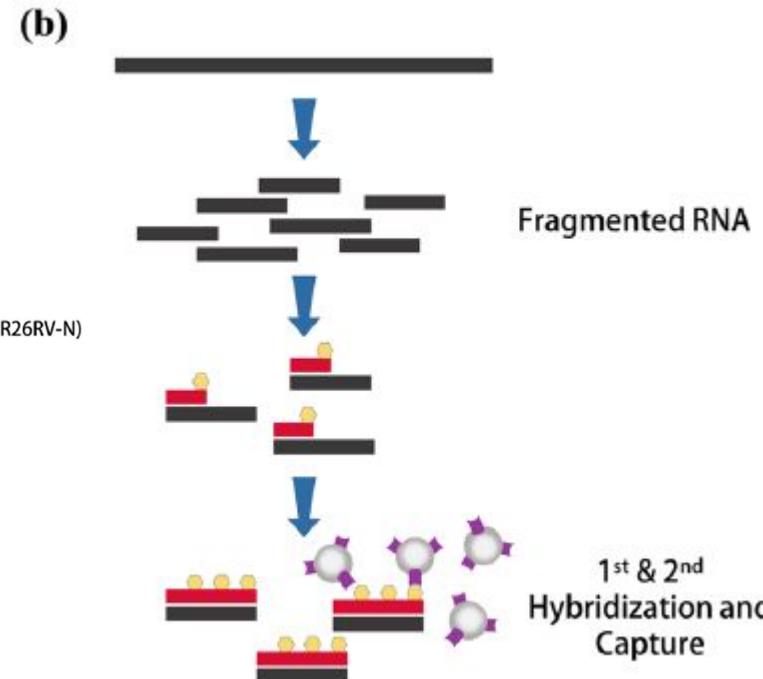


Amplification of multiply primed single stranded circular viral genomes. 3'-exonuclease resistant **primers randomly bind the genome and are elongated by the Phi29 polymerase**. The growing strand subsequently displaces the preceding strand of the DNA, making the strand available for binding of random primers and further elongation. This cyclic displacement and elongation leads to a highly branched structure of growing DNA, which is linear in topology.

Rolling circle amplification has the capability to specifically enrich the circular ssDNA genomes in an environment of other genetic materials, and could then be characterized by NGS.

Datta et al., World J Virol 2015
DOI: 10.5501/wjv.v4.i3.265

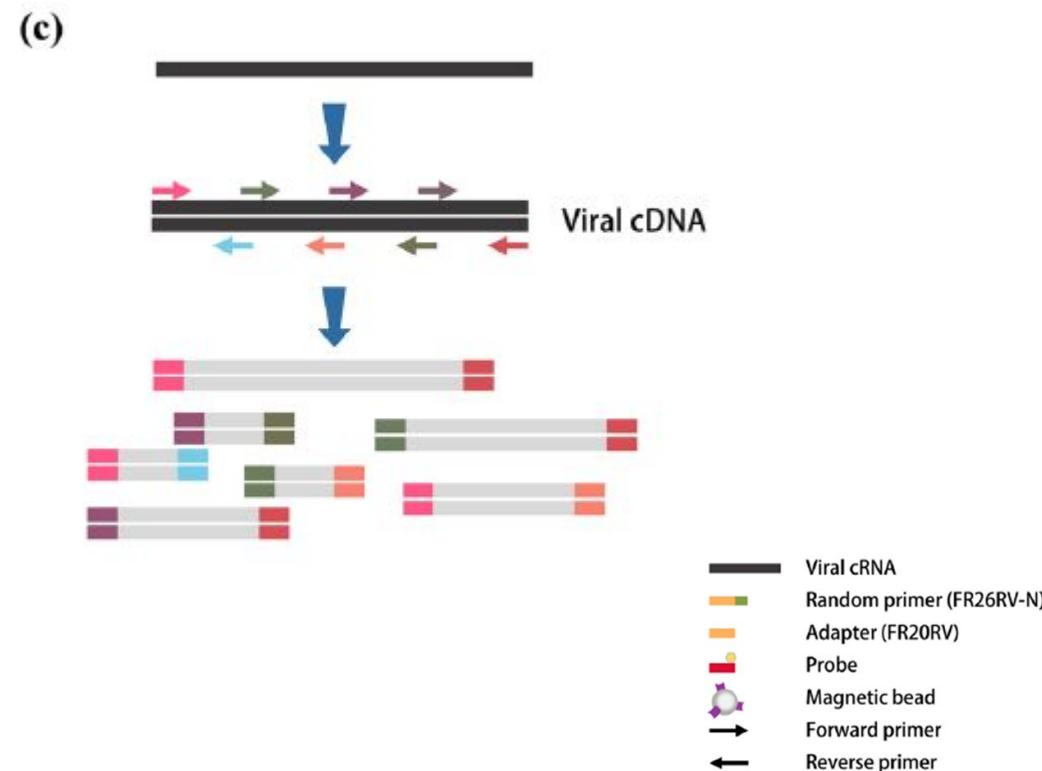
Nucleic Acid Extraction: Target capture NGS



WHOLE VIRAL GENOME RECONSTRUCTION

Target capture NGS captures viral genomes by using **virus-specific probes**. RNA samples were fragmented and then cDNA was synthesized using random hexamer. To enrich the target interest, first & second hybridization and capture were performed using virus-specific probes.

Nucleic Acid Extraction: Amplicon NGS



WHOLE VIRAL GENOME RECONSTRUCTION CHARACTERIZATION OF INTRA-HOST VARIABILITY

Amplicon NGS was applied to enrich viral genomes. Viral cDNA was amplified using **VIRUS-specific multiple primer mixture**

Nucleic Acid Extraction COMPARISON, Hantaan orthohantavirus genome sequencing

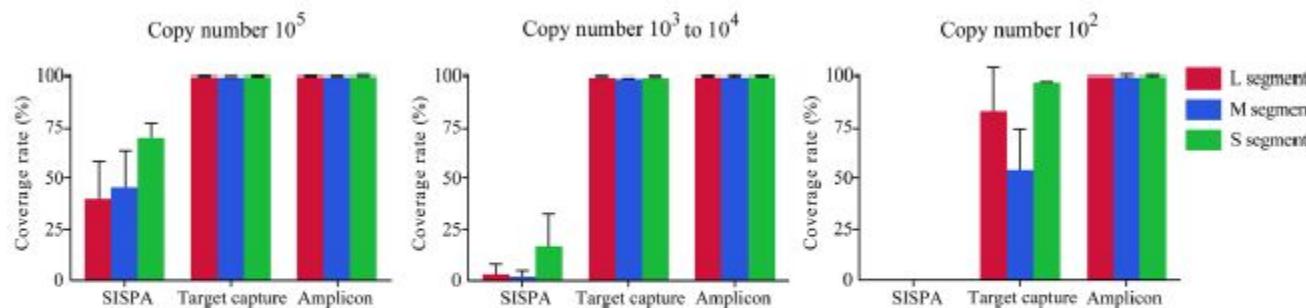


Figure 3. Coverage rate of consensus sequences for Hantaan orthohantavirus (HTNV) by viral copy number. The percentage coverage rate of consensus sequences by viral copy number among the three next-generation sequencing methods. Coverage rate was calculated by matching the consensus sequences with the sequence of HTNV 76–118 strain (GenBank accession number, NC_005222, NC_005219, NC_005218).

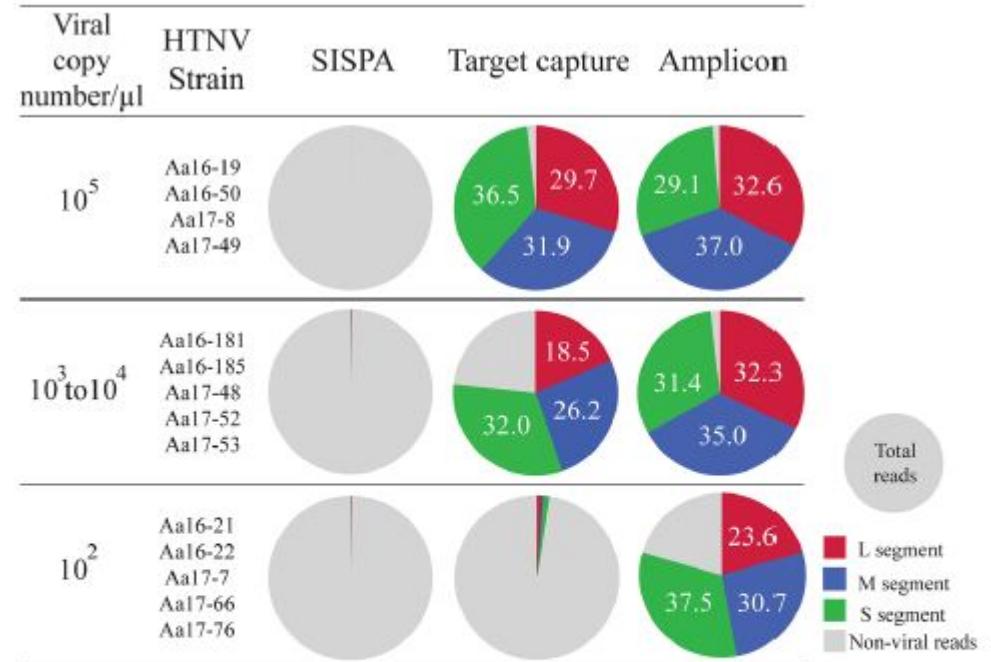
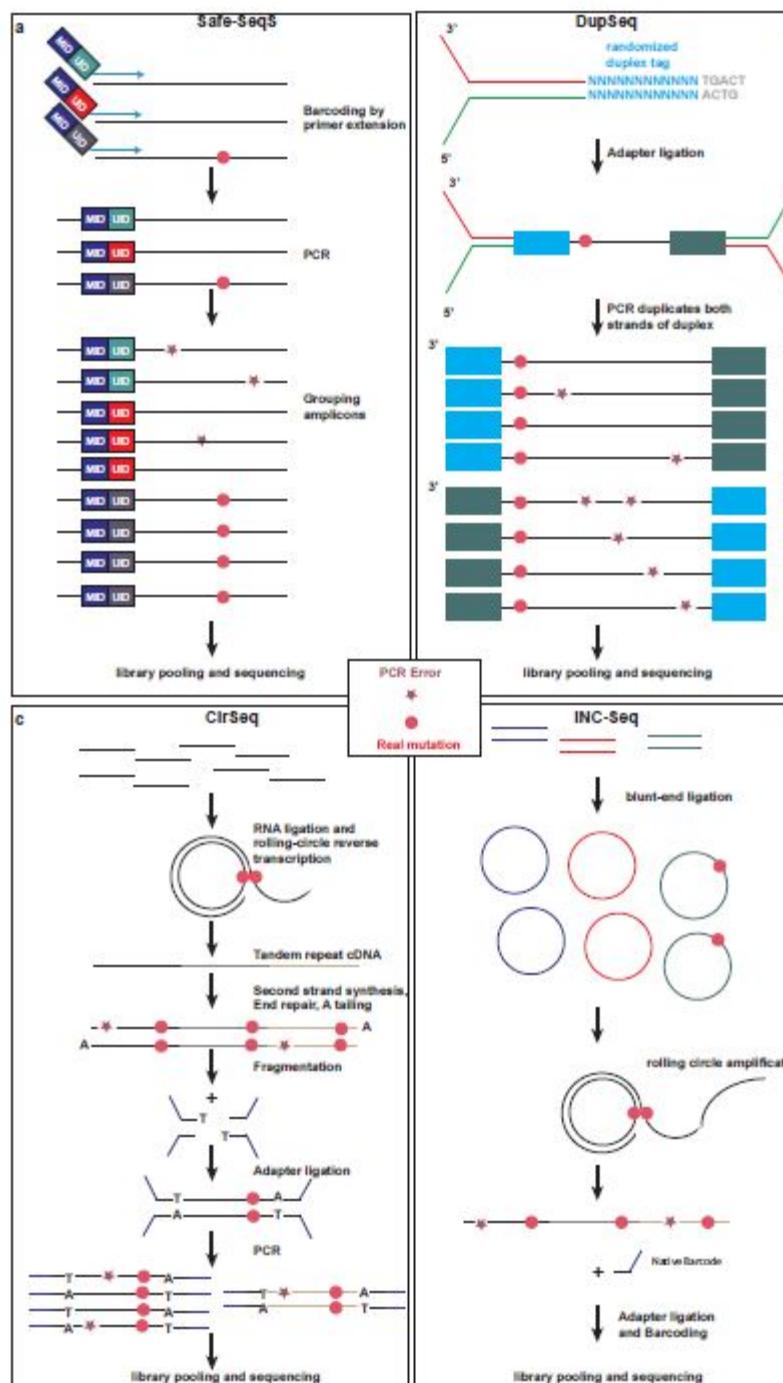


Figure 4. The composition of Hantaan orthohantavirus (HTNV) genomic reads in the total reads generated by three next-generation sequencing methods. The composition of viral reads mapped to HTNV was exhibited in the total number of reads. The total reads were produced by performing SISPA, target capture, and amplicon NGS. These reads were mapped to HTNV 76–118 tripartite genomes (GenBank accession number: L segment, NC_005222; M segment, NC_005219; S segment, NC_005218). A circle represents for total reads obtained by SISPA, target capture, and amplicon NGS methods. Red, Blue, and Green colors indicate the composition of viral reads for HTNV L, M and S segments, respectively, over the total reads. Gray color indicates non-viral reads in the total reads. The HTNV reads were shown as a percentage (%) evaluated by the ratio of viral reads over the total reads.

Nucleic Acid Extraction low frequency variant identification, quasispecies



Lu et al., Virus Research 2020

Schema main steps of clinical virus Discovery by NGS

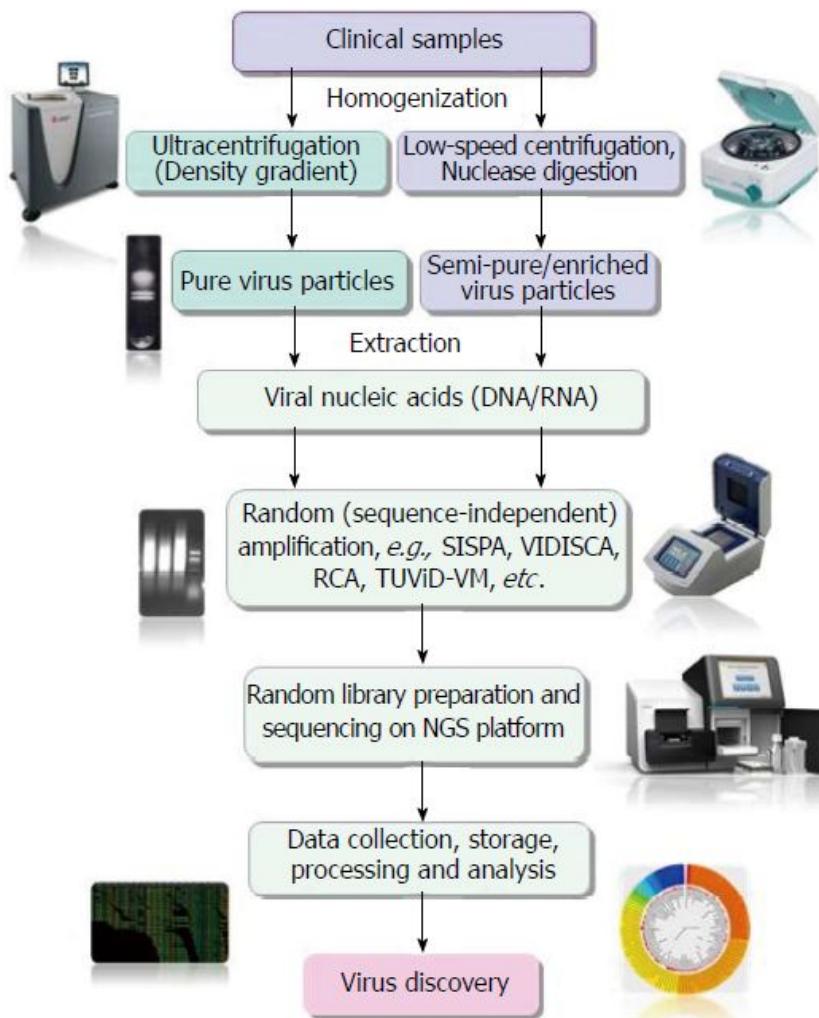


Figure 1 Diagrammatic representation of main steps of clinical virus discovery by next-generation sequencer based technologies.

Datta et al., World J Virol 2015
DOI: 10.5501/wjv.v4.i3.265

LIBRARY PREPARATION, strategies

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

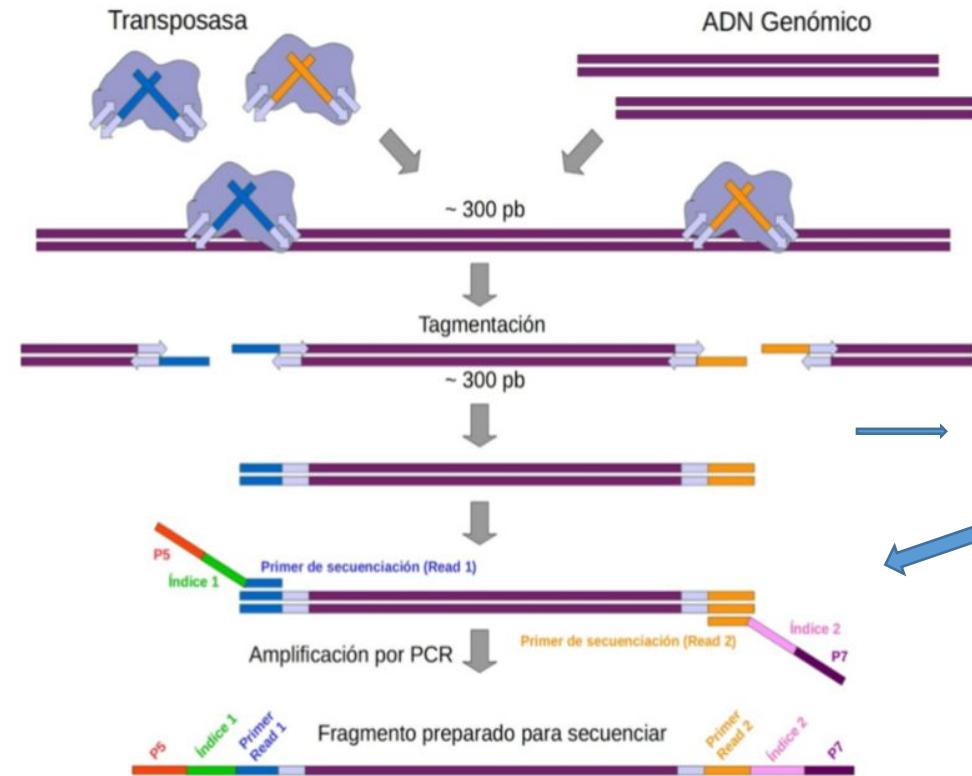
1. Metagenómica

IDENTIFICACIÓN MICROORGANISMOS

1. Metataxonomía o Targeted Metagenomics

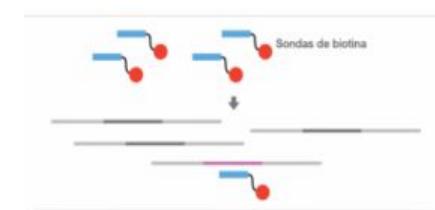
PREPARACIÓN LIBRERÍA

ENZIMÁTICA FÍSICA



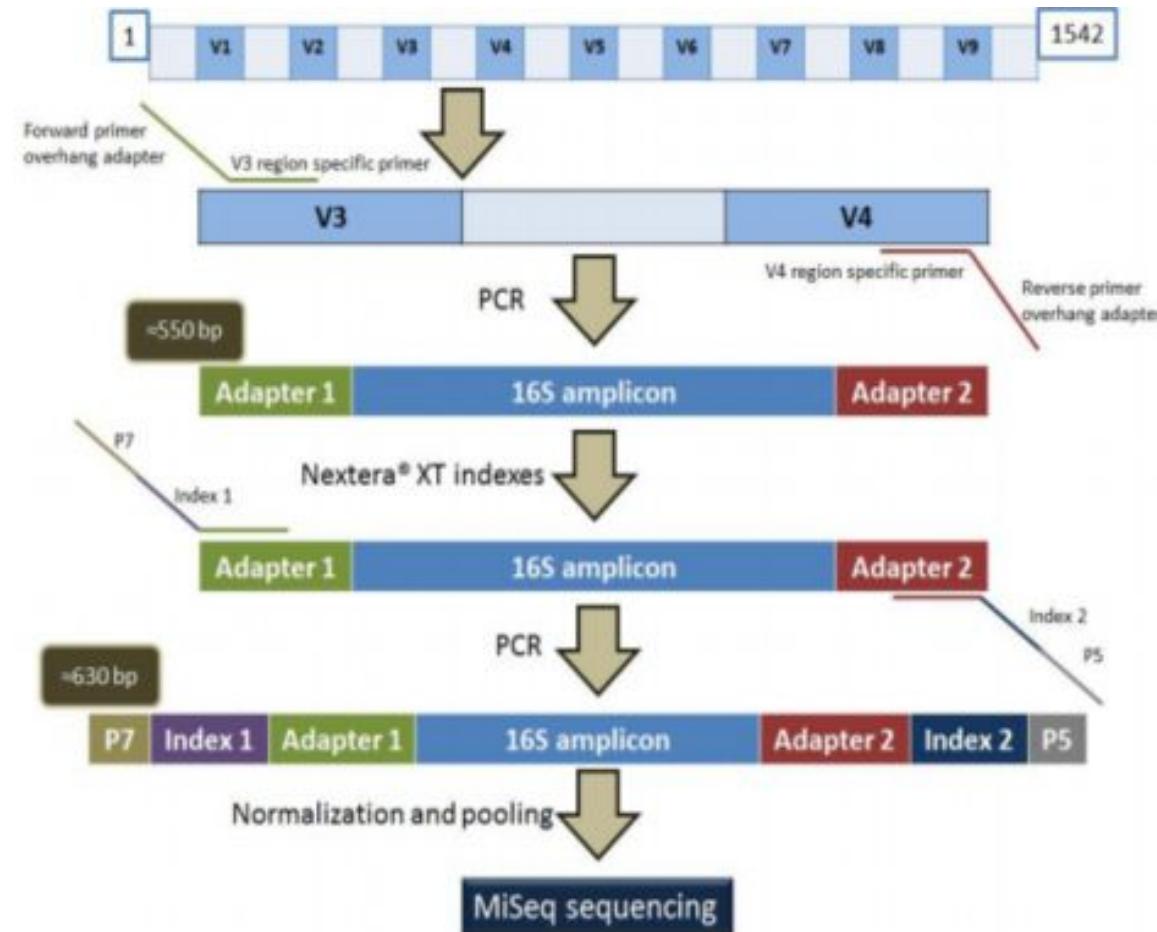
RNA -> cDNA

ENRIQUECIMIENTO:
PCR
CAPTURA SONDAS



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

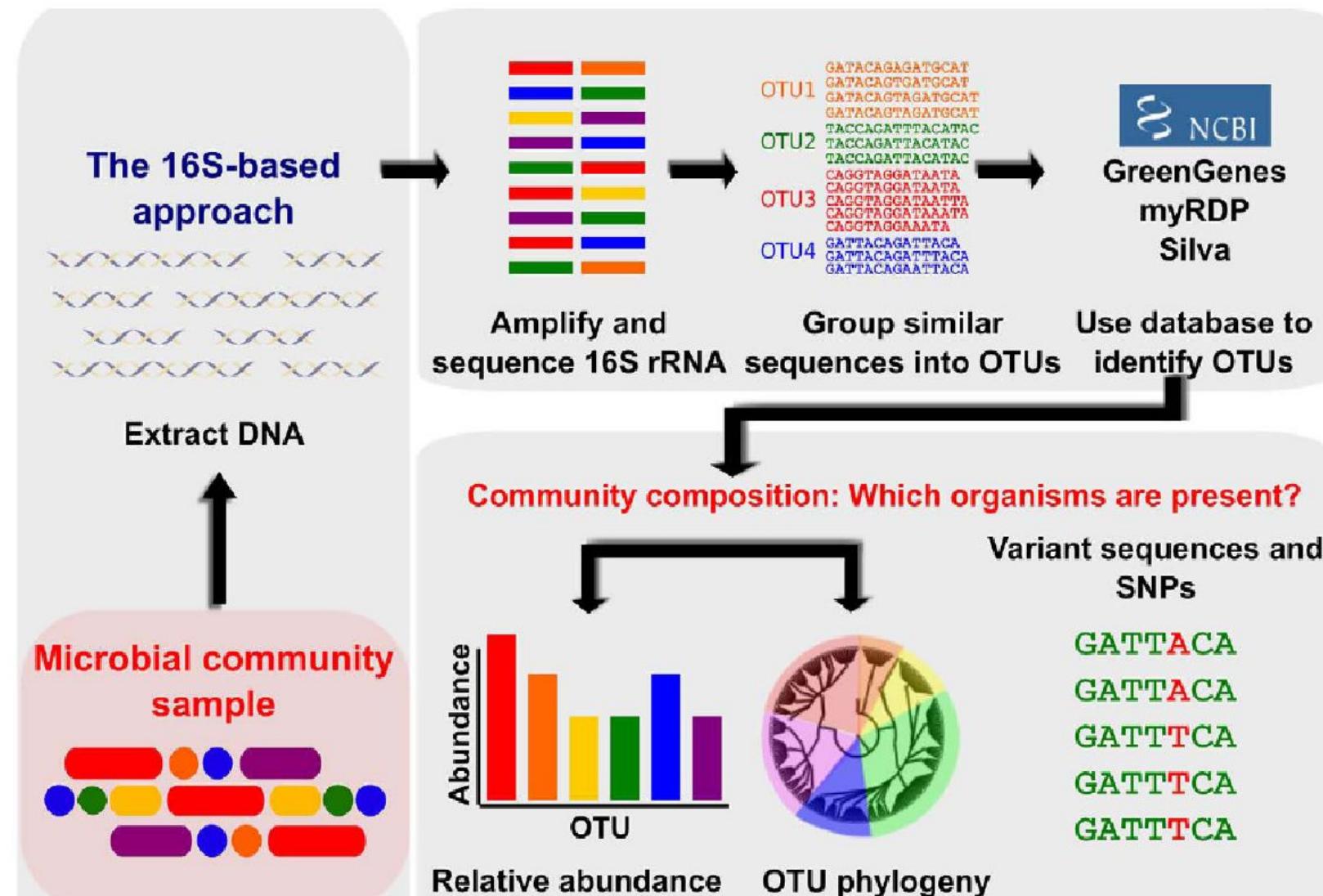
PREPARACIÓN LIBRERÍA, rRNA 16S, caracterización microbiota



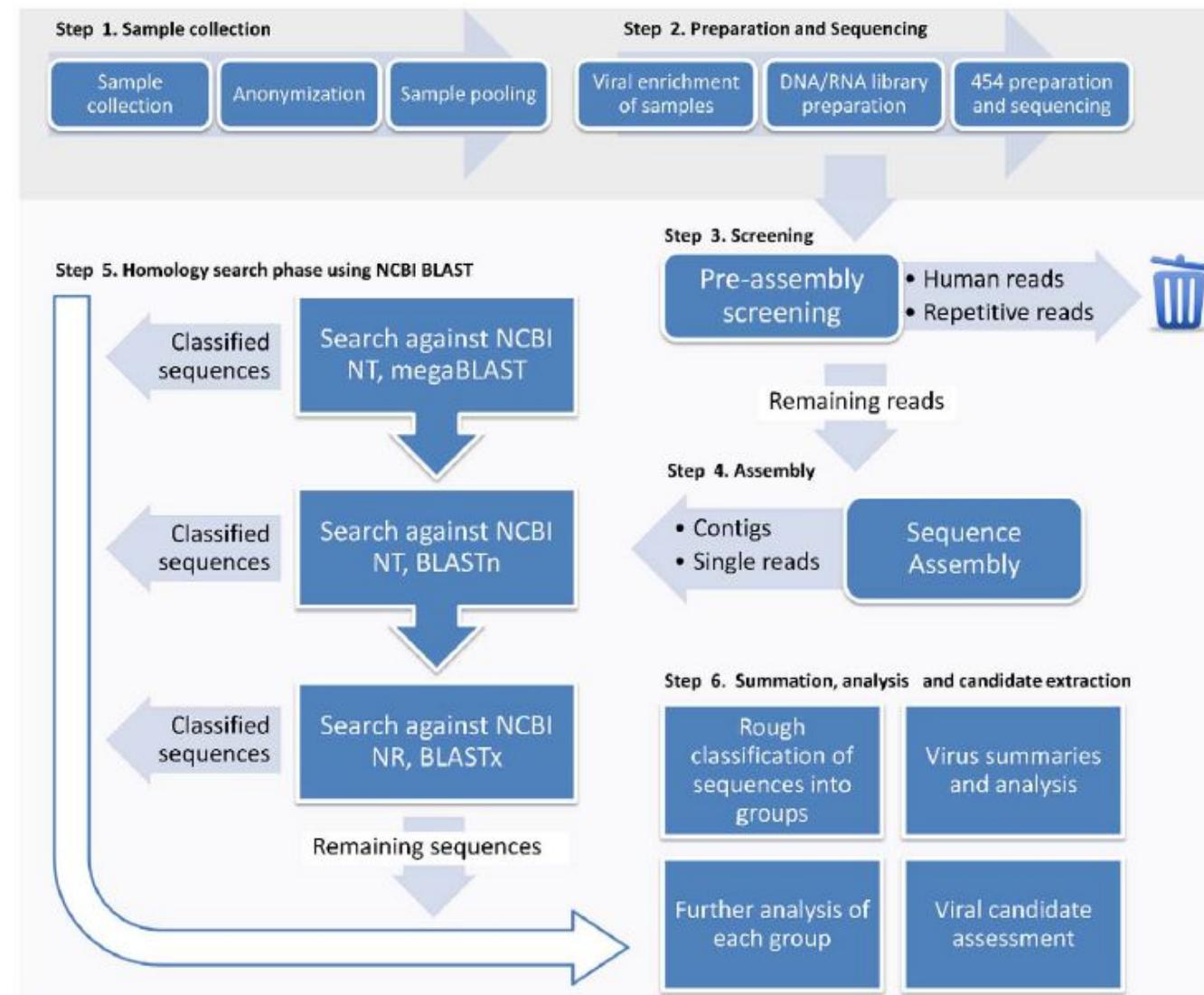
Targeted Metagenomics vs Metagenomics (16S vs Shotgun)

	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

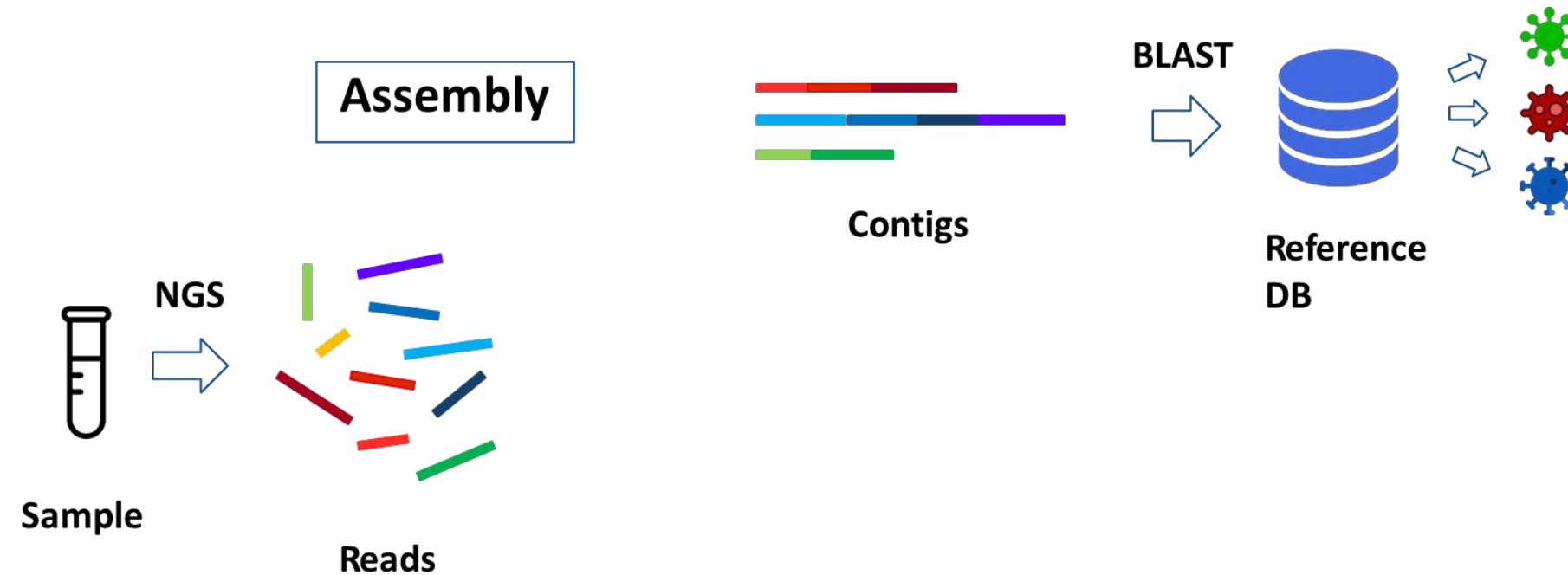
Targeted Metagenomics



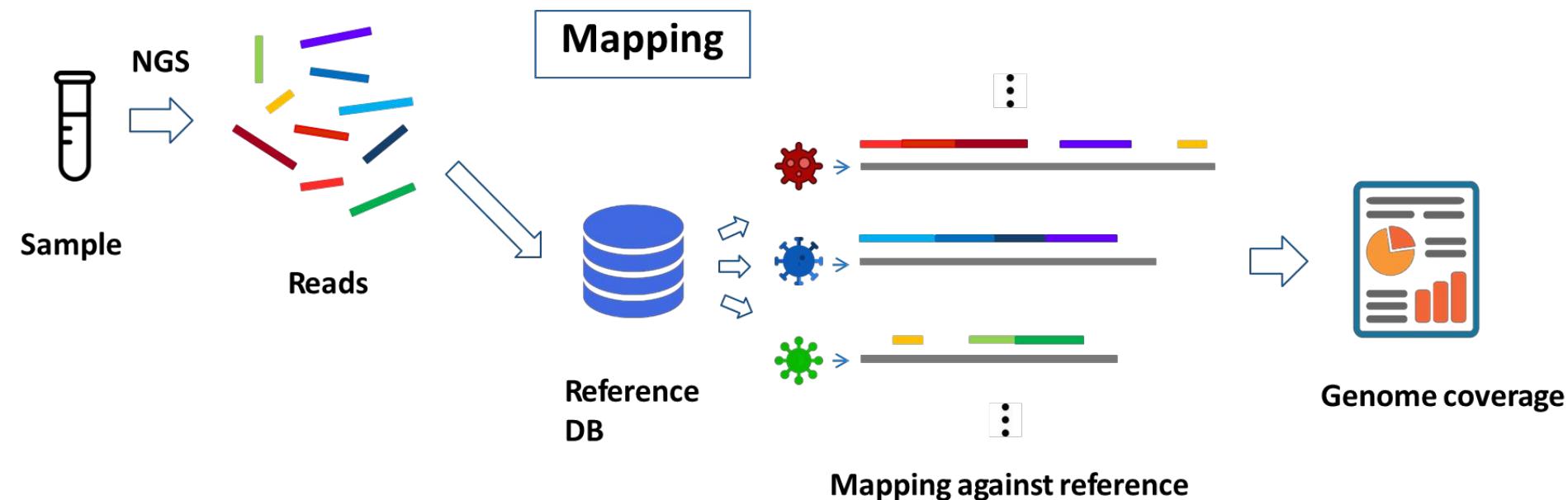
Metagenomics



Metagenomic analysis approaches



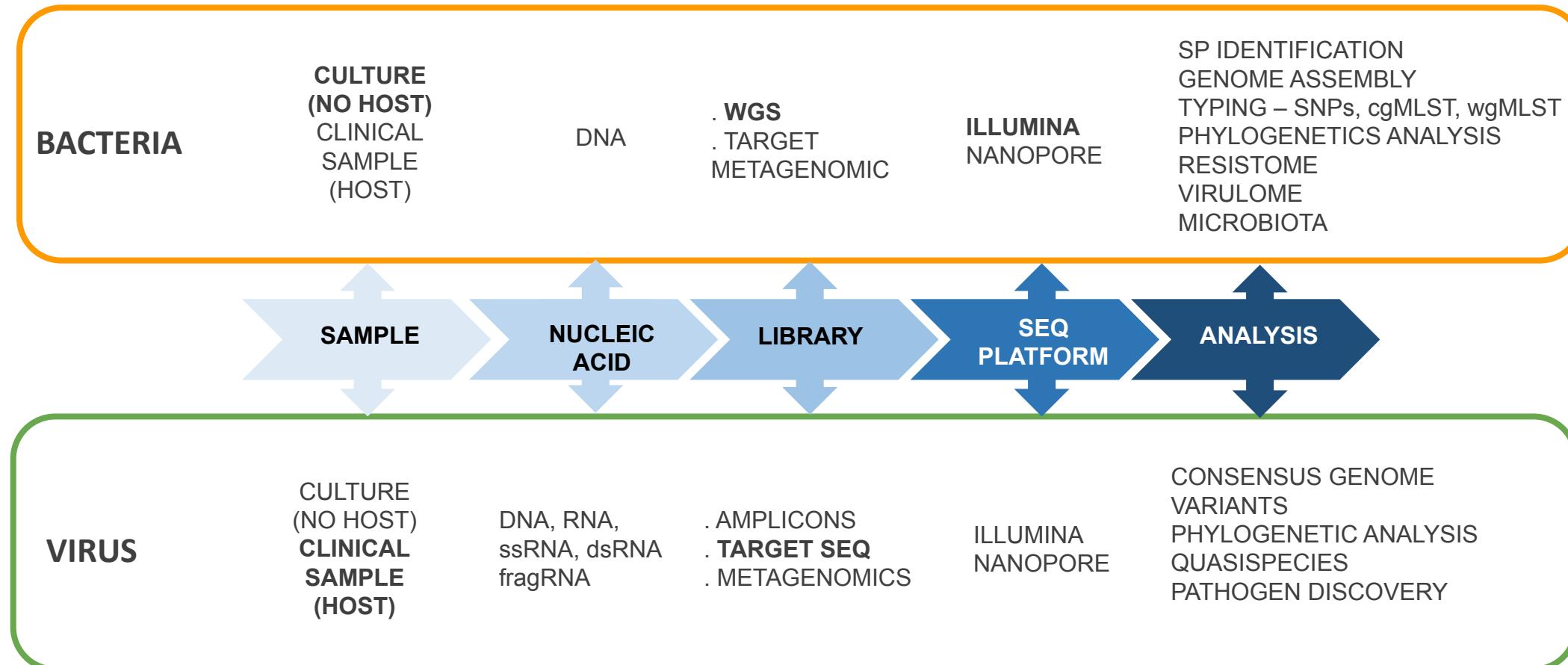
Metagenomic analysis approaches



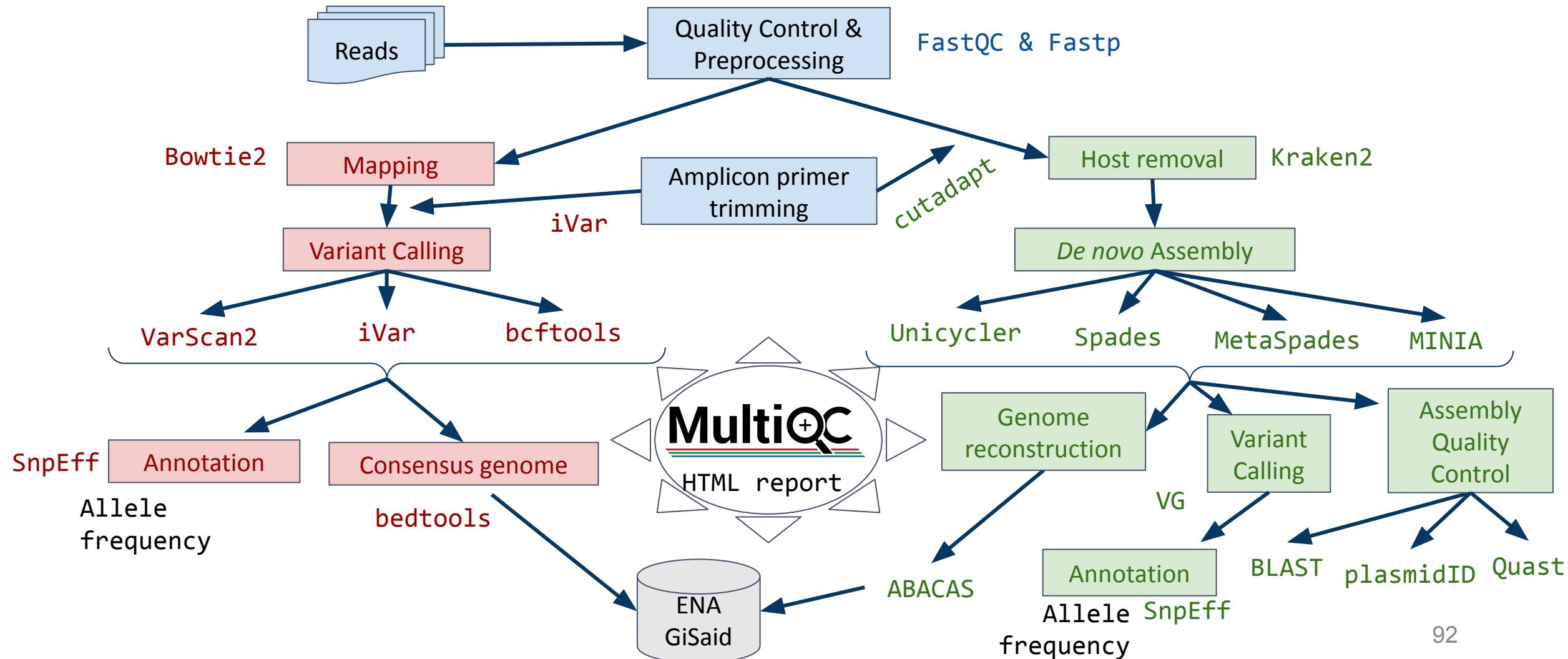
Metataxonomics vs Metagenomics (16S vs Shotgun)

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



Viralrecon nf-core



Thanks for your attention!

Questions???