

Análisis de Genomas Virales a través de la Plataforma Galaxy

BU-ISCIII

Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

14 al 18 Noviembre 2022

2^a 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 NGS in viral variants identification and consensus genome reconstruction.
- Learning to use galaxy platform.

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). Contrato Titulado Superior asociado a proyecto (2021-2022)
- 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

Isabel Cuesta

BU-ISCIII

Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

14 al 18 Noviembre 2022

2^a Edición

Programa Formación Continua, ISCIII

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

Index

- BU-ISCIII, Bioinformatics and preparedness to be a service provider

What is Bioinformatics?

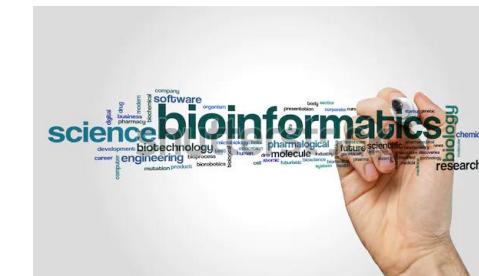
Biological Problems



Data
Processing

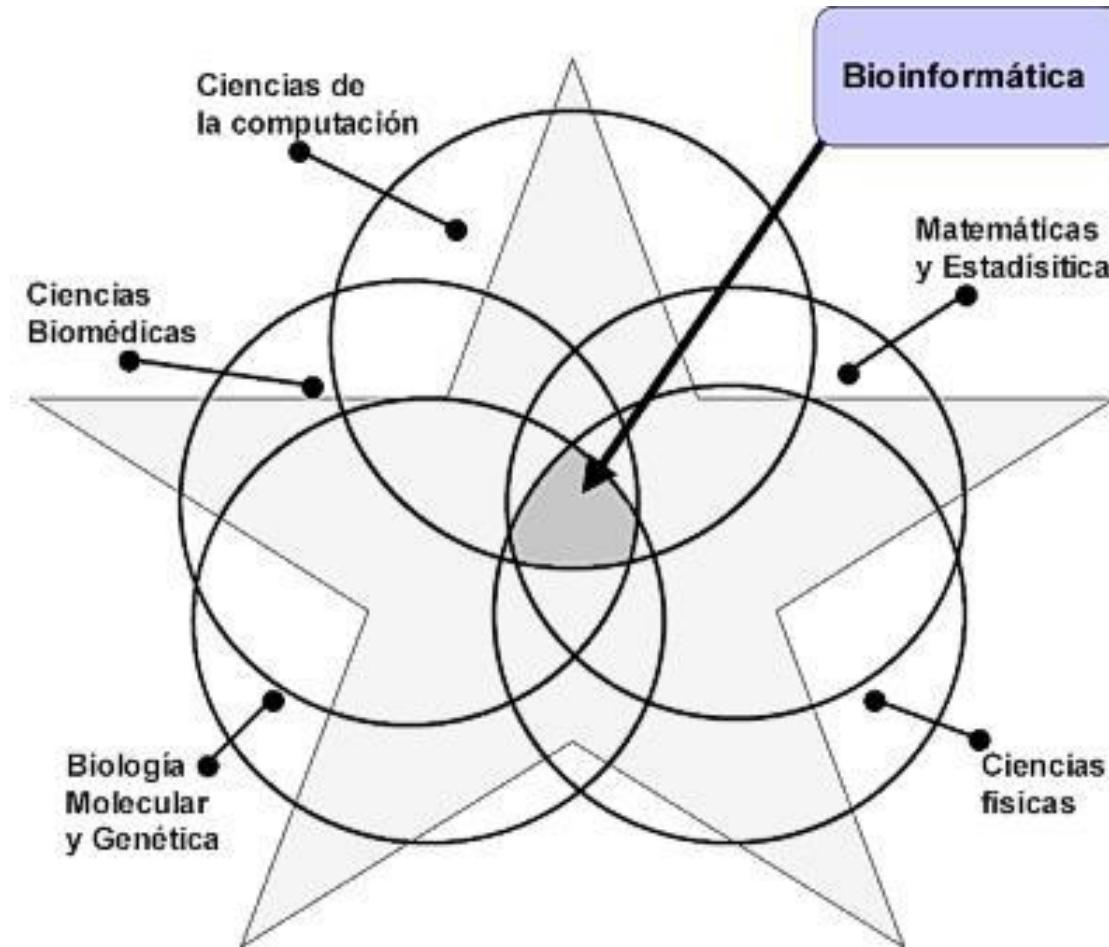


Computational Methods

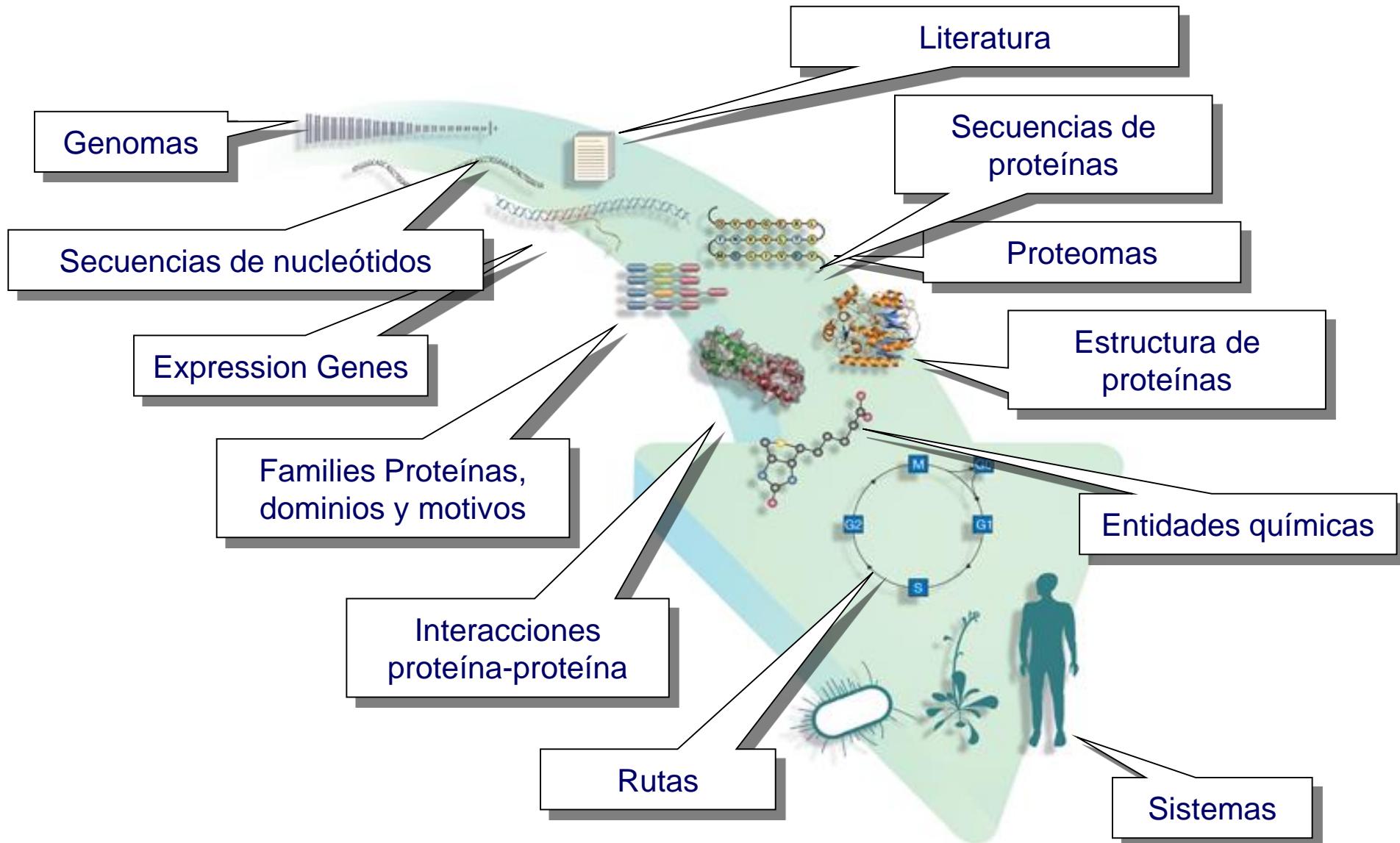


Bioinformatics is the science of storing, retrieving and analysing large amounts of biological information

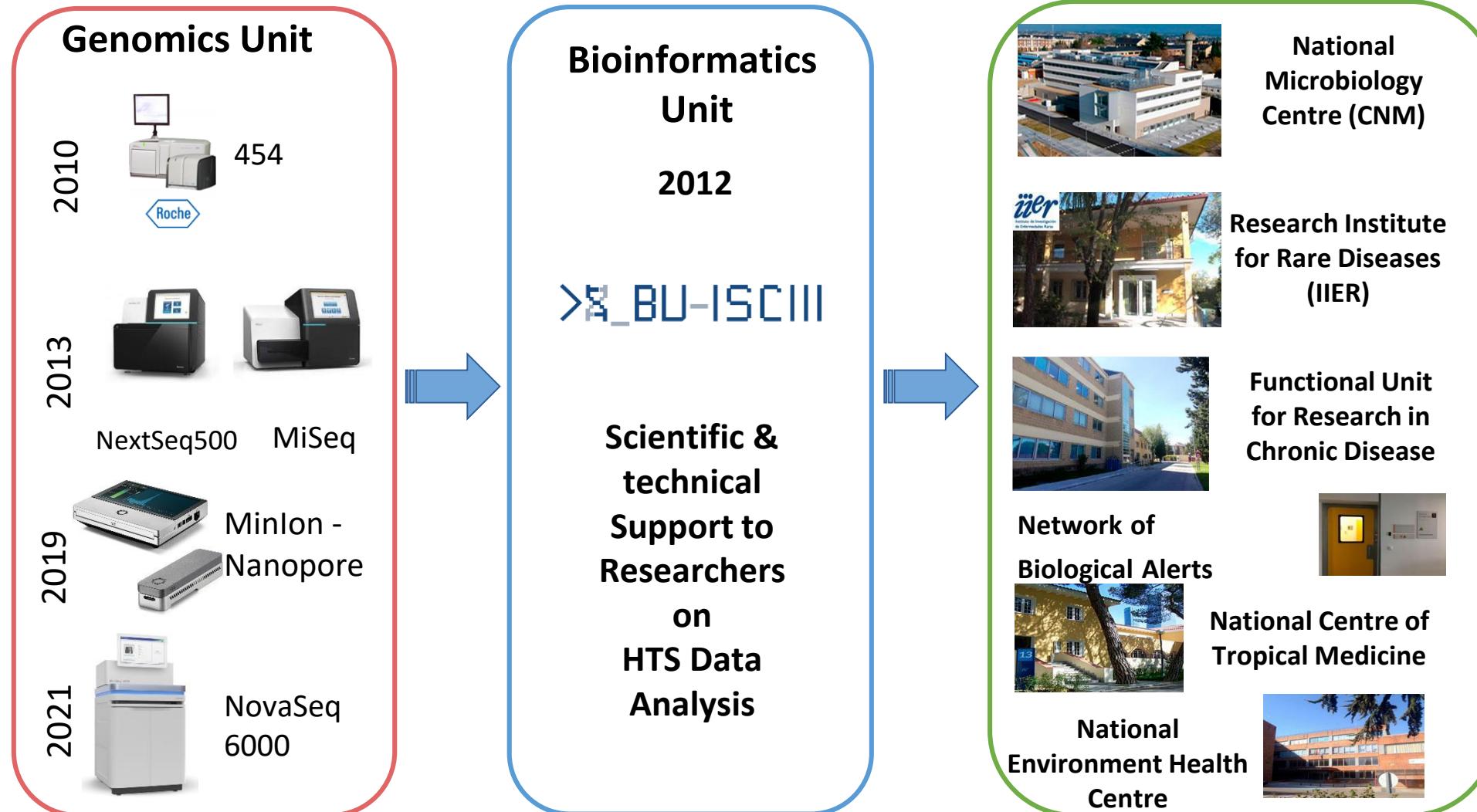
Bioinformatics is a multidisciplinary field



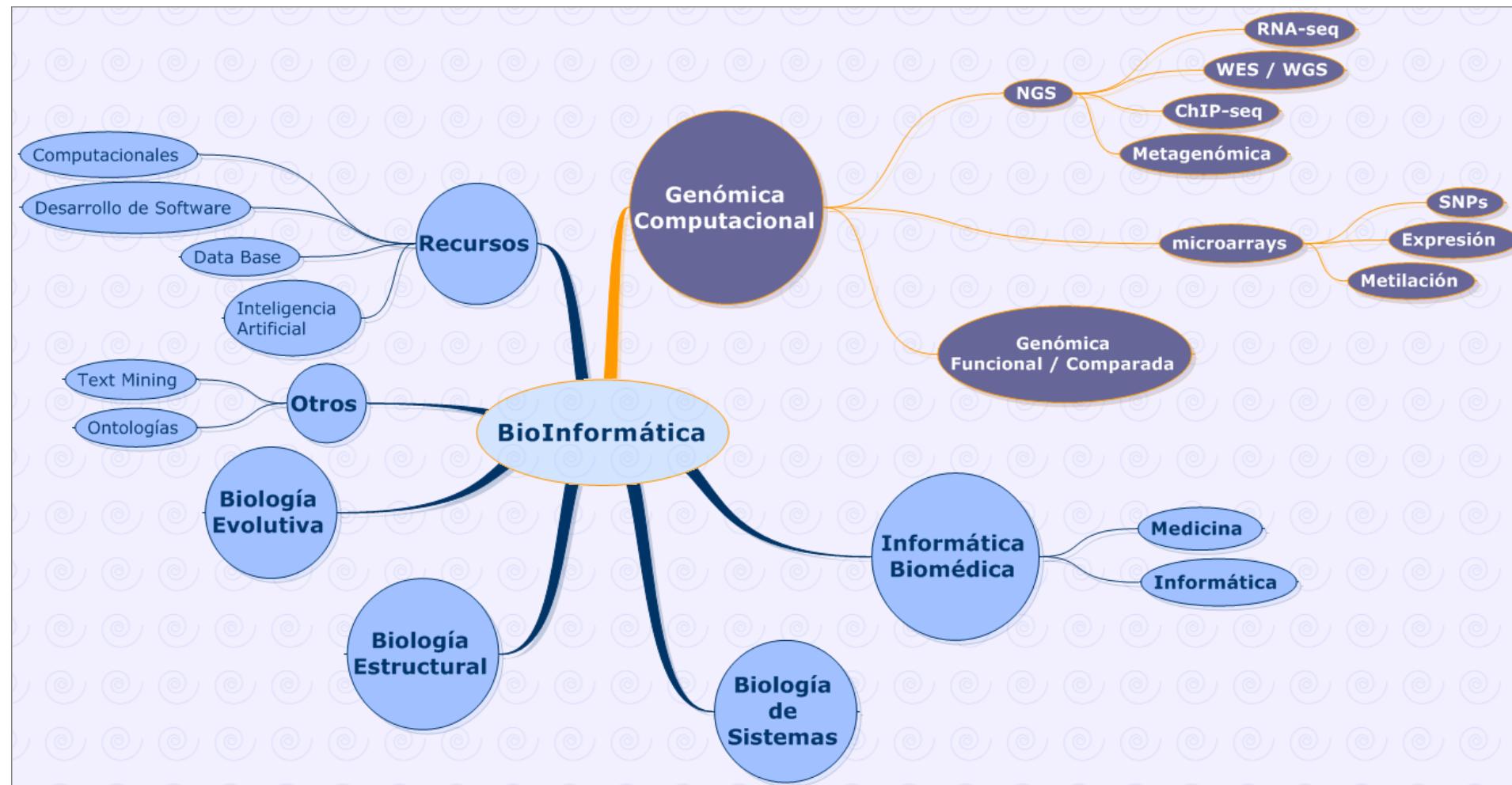
Dimension of Bioinformatics is given by data type



Why BU-ISCIII was founded



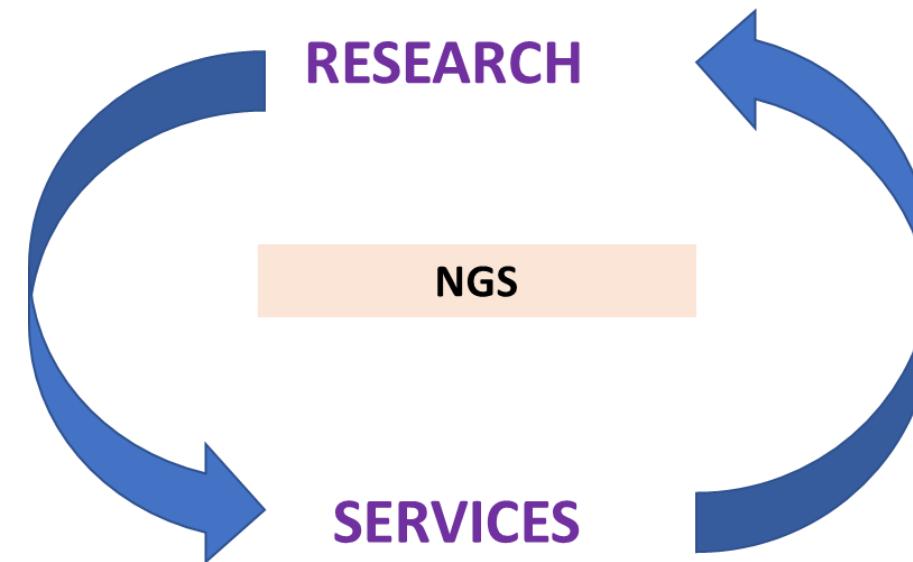
BU-ISCIII Mission - Activities



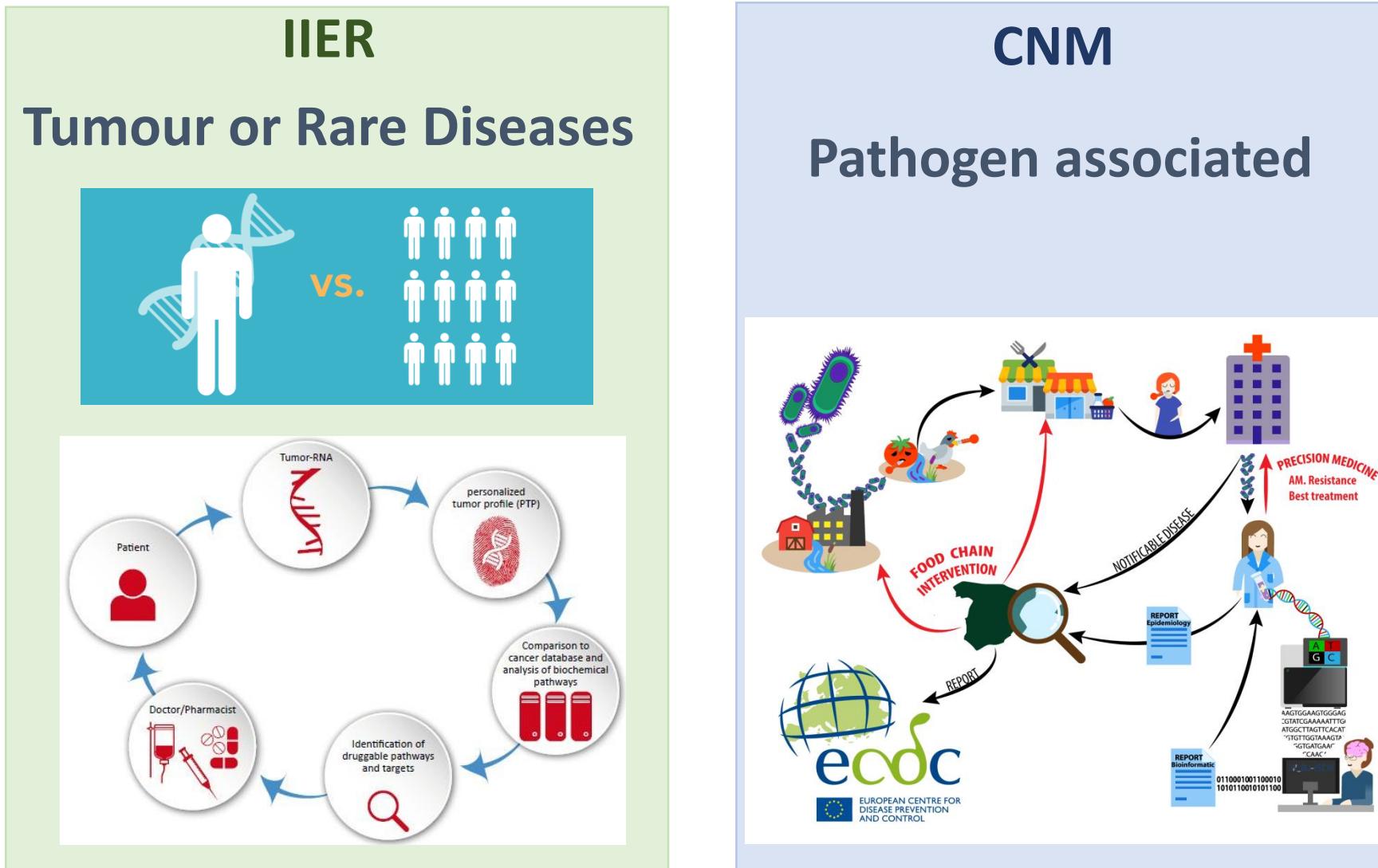
Bioinformatics Unit Activities

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

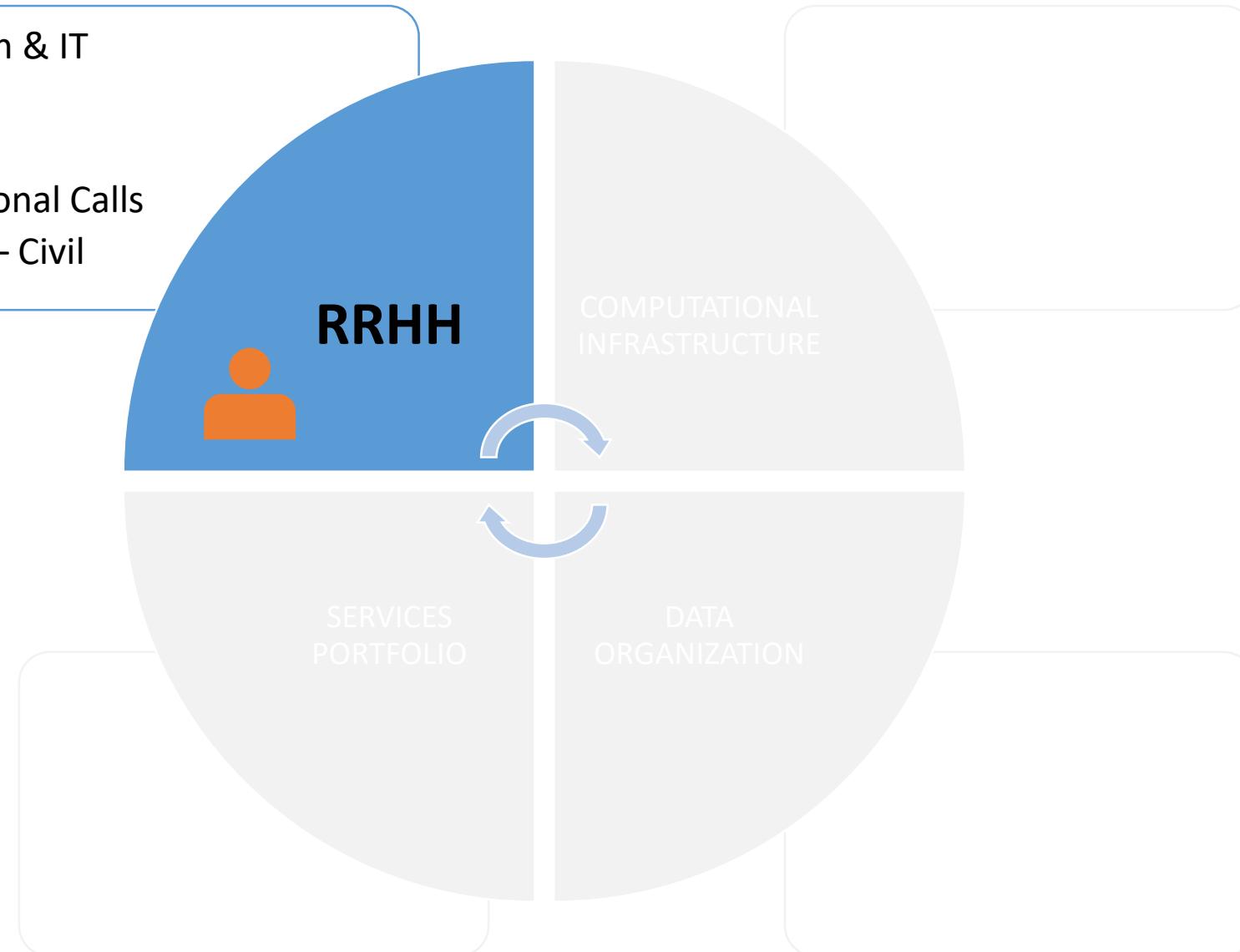
Provide scientific and technical solutions for using NGS in the diagnostic routine or research activity from different ISCIII labs



Clinical Bioinformatics - Precision Medicine

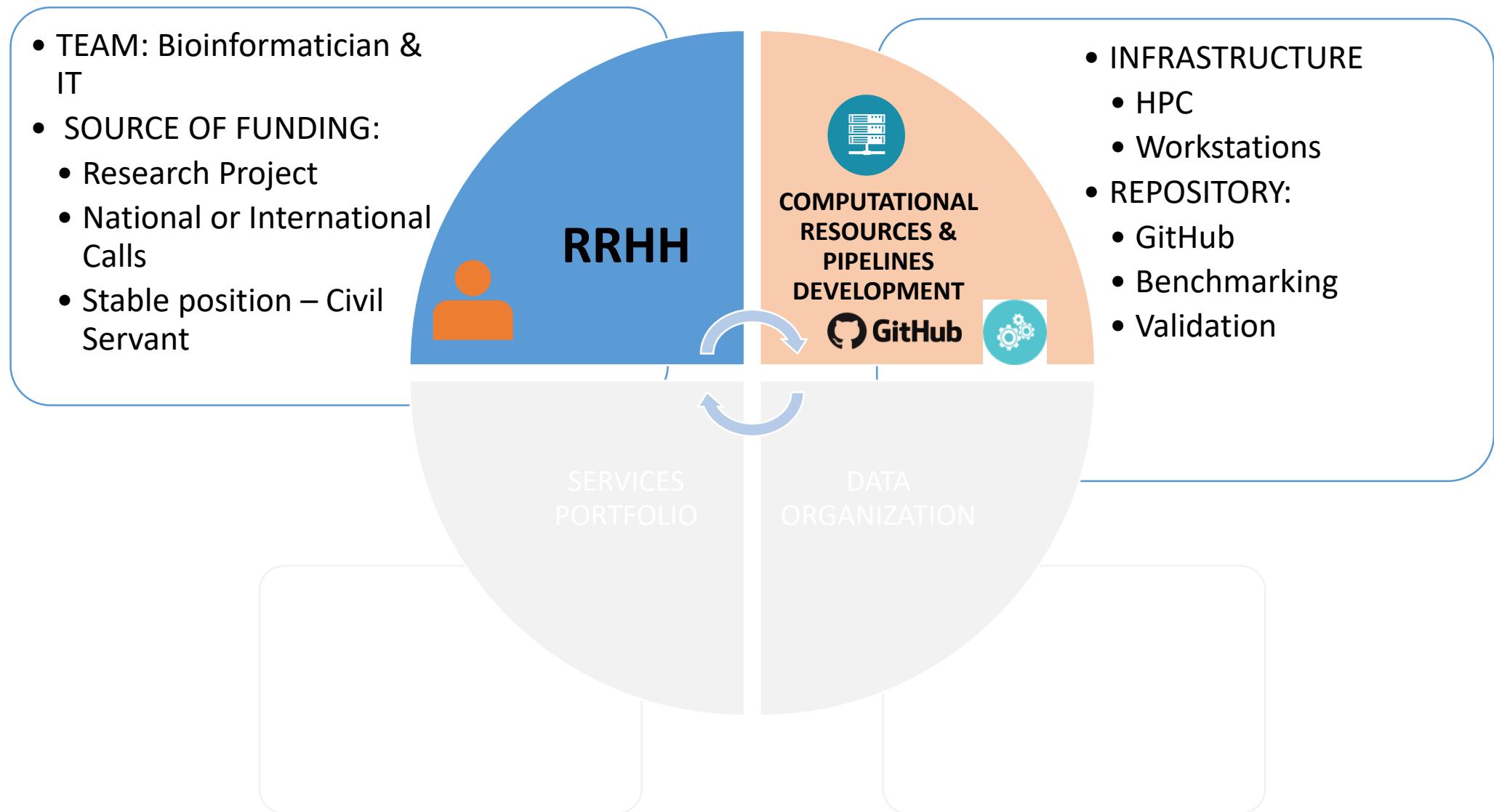


- TEAM: Bioinformatician & IT
- SOURCE OF FUNDING:
 - Research Project
 - National or International Calls
 - Permanent position – Civil Servant



Human resources

	Disciplina	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Isabel Cuesta	Dr. Biología Molecular												
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					ANTIBIOTICOSU.				ANTIBIOTICOSU.			
Sarai Varona	Bioquímica									PROYECTO (2022)			
Luis Chapado	Telecomunicaciones						COLABORADOR						
Erika Kvalem	Biotecnología										PROYECTO COVID19		
Alberto Lema	Biología										1 año		
Luis Aranda	FP Informática										oct 2021 al 2022		
Guillermo Gorines	Biología										ESTANCIA		
Personal (número)		2	4	4	5	5	4	6	6	6	8	8	2



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

- 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

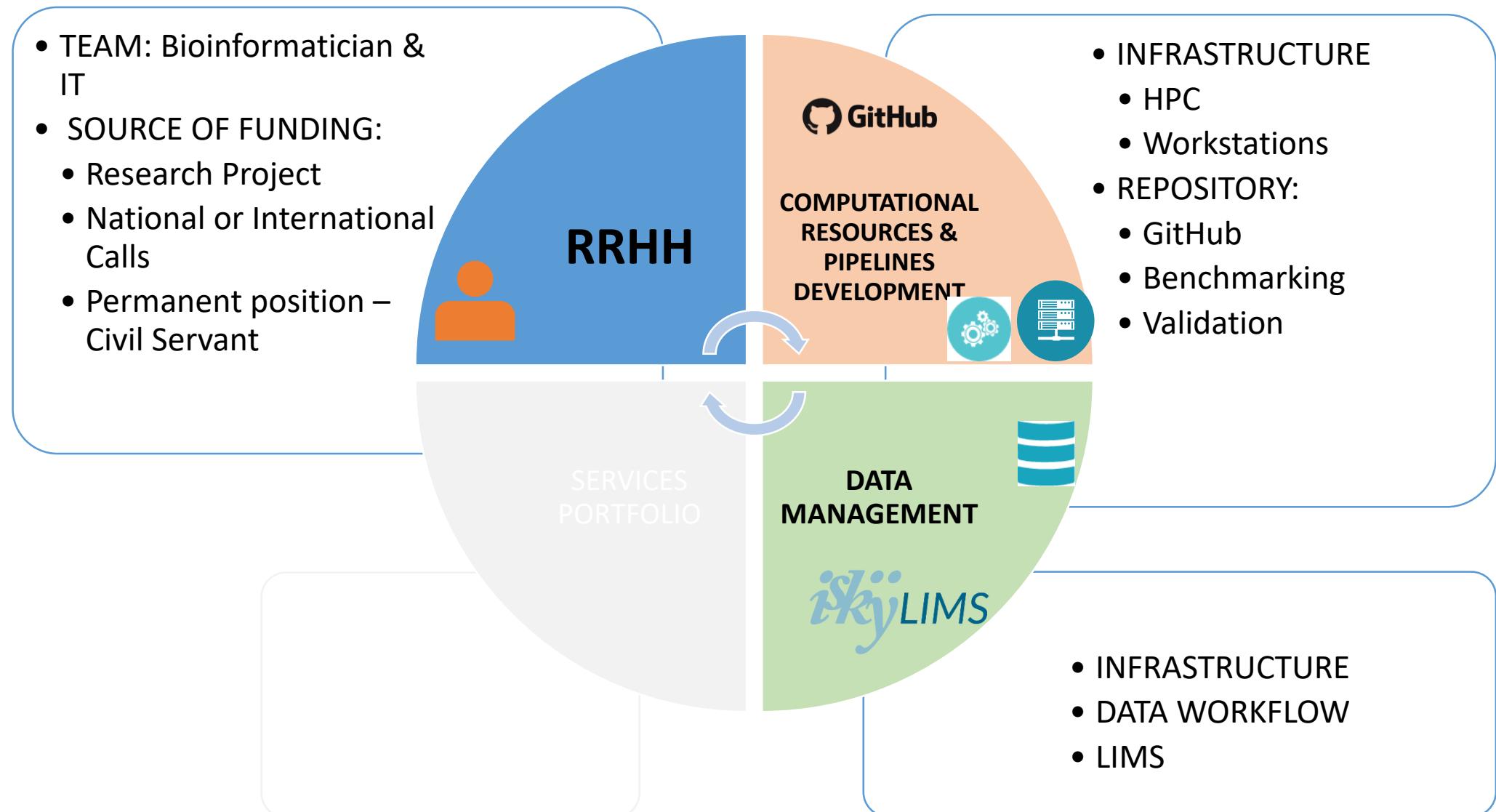
Table 5. Results of allele-based cluster analysis

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			

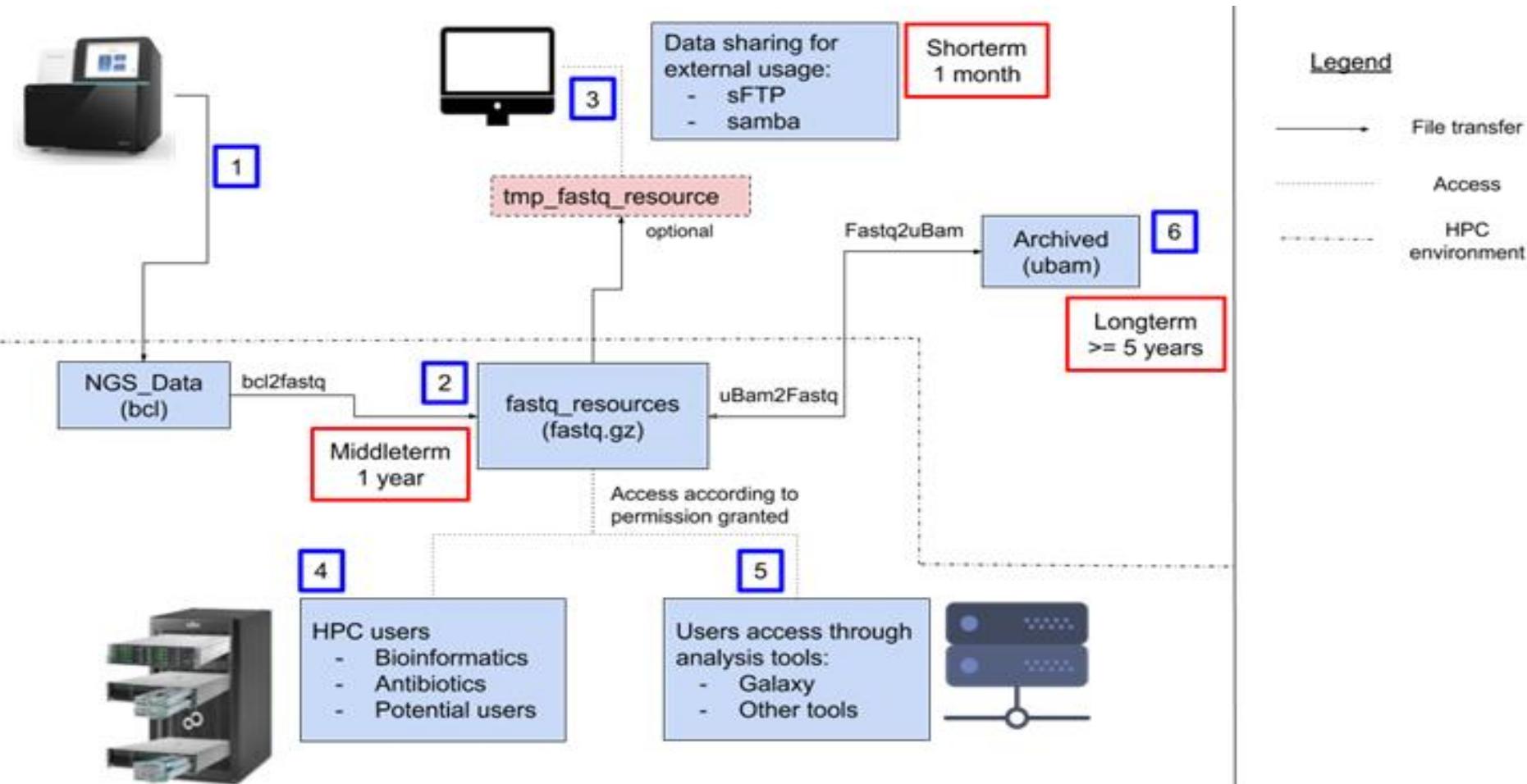
Table 4. Results of SNP-based cluster analysis

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	

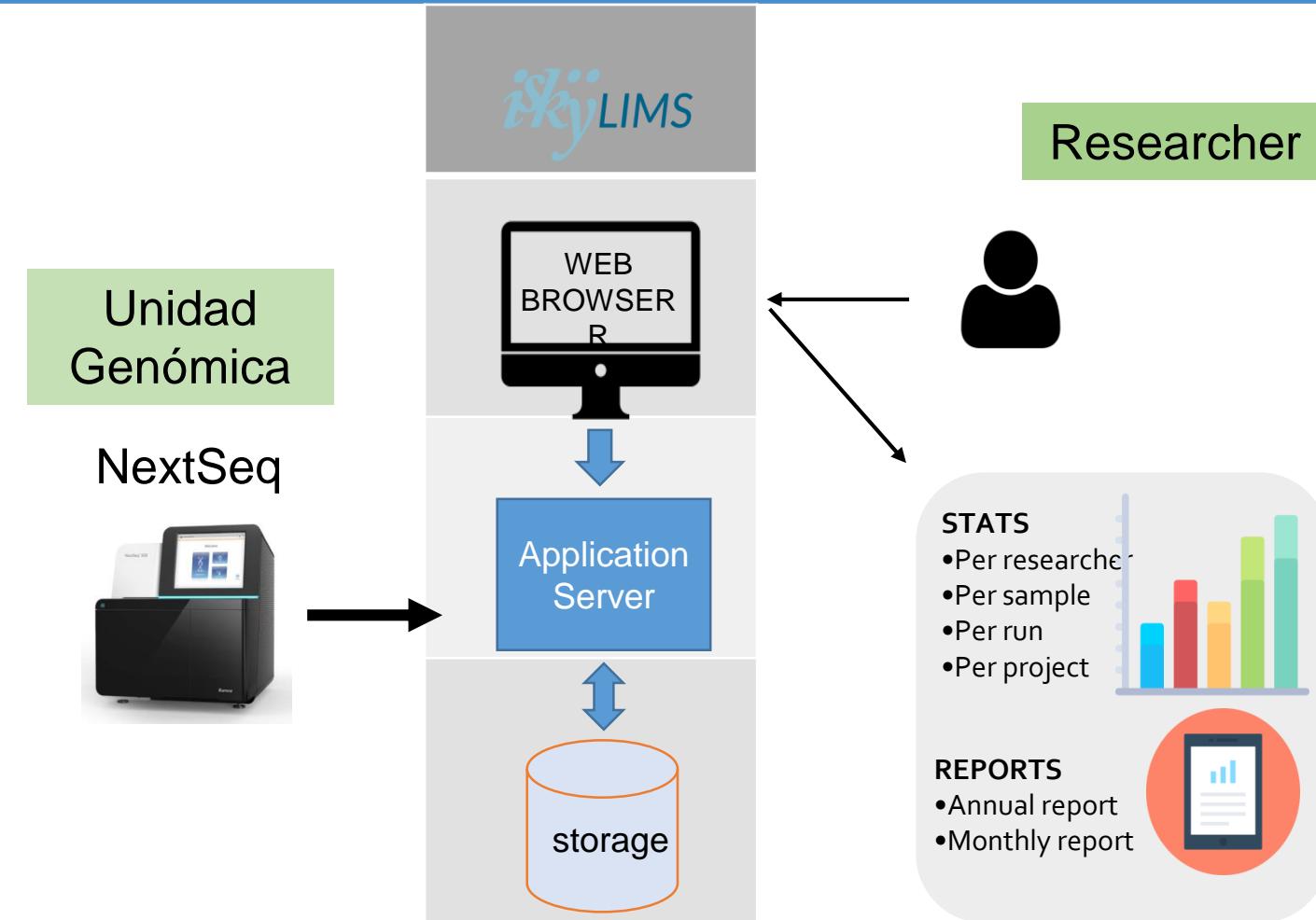
Fifth external quality assessment scheme for Listeria monocytogenes typing



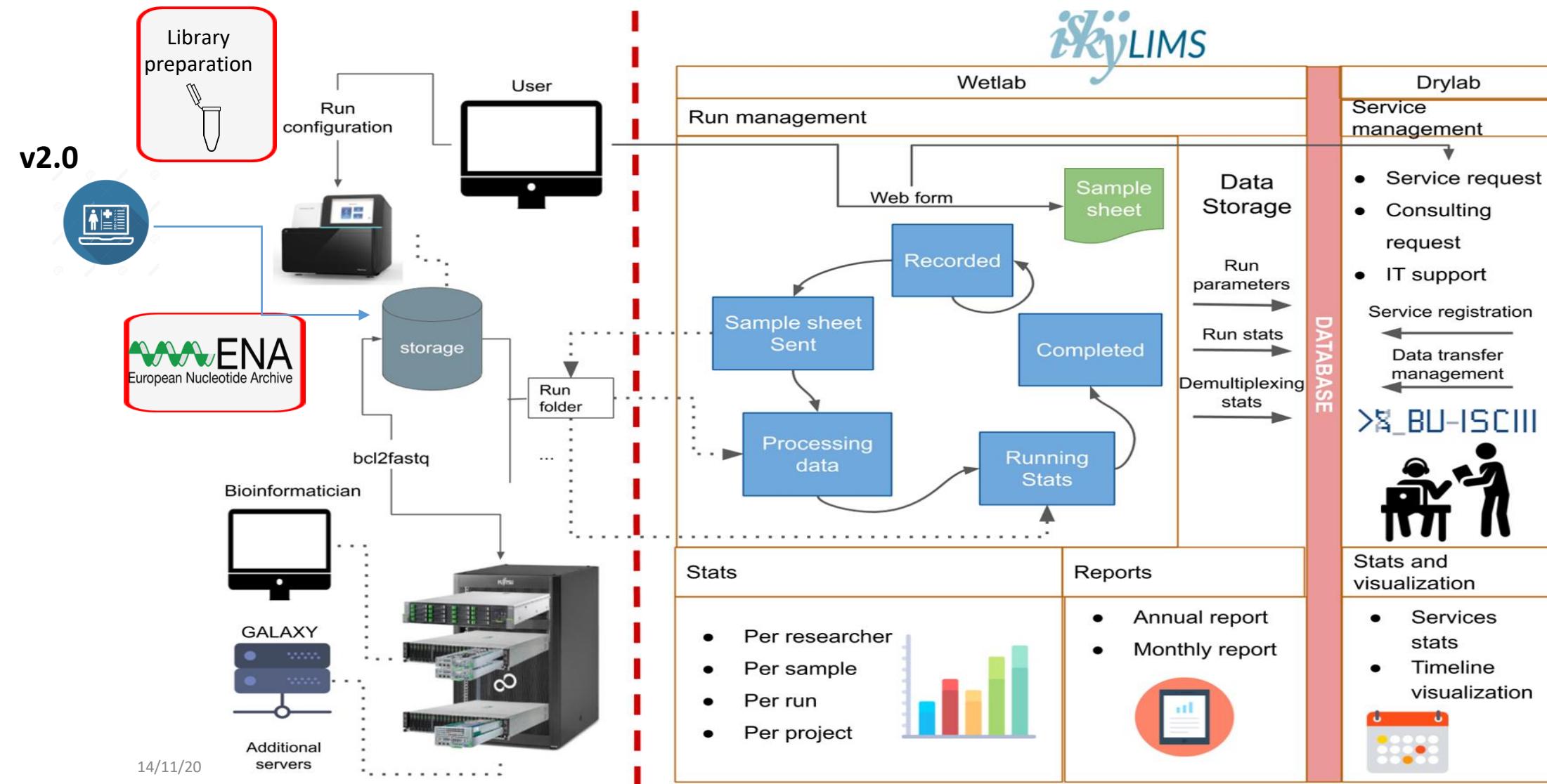
Infrastructure and data management



- Minimize I/O issues
- Maximize storage uses



Infrastructure and data management: LIMS



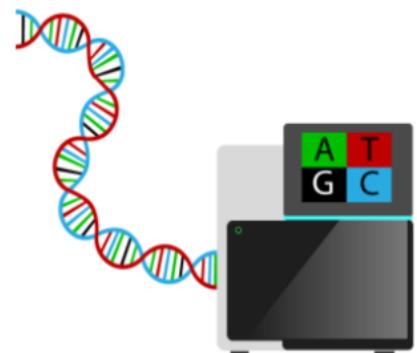
SERVICIOS DE LA UNIDAD



HOME ABOUT US TUTORIALS FAQS REGISTER CONTACT

icuesta

Wetlab



Drylab



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

Logos



Connect



Links

- Contact
- Getting started
- FAQs

Sitemap

- iSkyLIMS home
- Drylab page
- Wetlab page

SERVICIOS DE LA UNIDAD



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

smonzon [Logout](#) [My account](#)

BioInformatics

iSkyLIMS: DryLab

Welcome

This section will allow you to check BU-ISCIII service activity. Available processes are request new services, colaborations, counseling and infrastructure. You will be able to check the status of your ongoing services.

Services ongoing and queued

Under construction. This will be a table with services ongoing or queued



Timeline of services

Under construction. Kind of diagram with services dates.



Service Request Form

Form for requesting internal service to Bioinformatic Unit

Sequencing Data

User's projects*

BMartinez20161213
EXOMAS_ND_20170303
EXOMAS_ND_20170327_RE
EXOMAS_ND_20170228
Mardoc_Sequencing_20170210

Run specifications**File extension****Sequencing platform**

SERVICES REQUEST



HOME SERVICES REQUEST COUNSELING REQUEST INFRASTRUCTURE REQUEST



- Genomic Data Analysis
 - Download and quality analysis
 - Data download
 - Sequence quality analysis
 - Sequence pre-processing (quality filtering)
 - Next Generation Sequencing data analysis
 - DNAseq: Exome sequencing (WES) / Genome sequencing (WGS) / Target sequencing
 - Trio/family variant calling pipeline
 - Variant calling and annotation pipeline
 - Microbial: Whole genome outbreak analysis pipeline
 - Microbial: wgMLST
 - Microbial: MLST + virulence + AMR + plasmid analysis
 - Microbial: Assembly + automatic annotation
 - Microbial: plasmidID pipeline - strain plasmid characterization
 - RNAseq: Transcriptome sequencing
 - miRNA-Seq pipeline
 - mRNA-Seq pipeline
 - Amplicon sequencing (Deep sequencing)
 - Low frequency variant detection
 - Viral: assembly and minor variants detection
 - Metagenomics
 - 16S taxonomic profiling
 - Shotgun metagenomics profiling
 - Shotgun metagenomics - Virus genome reconstruction
 - CHIP-SEQ
 - Peak detection and annotation

SERVICES REQUEST



Service Description

Service description file*

No file selected.

Service Notes*

COUNSELING REQUEST



Service selection
Available Services *
<input type="checkbox"/> Bioinformatics consulting and training
<input type="checkbox"/> Bioinformatics analysis consulting
<input type="checkbox"/> In-house and outer course organization
<input type="checkbox"/> Student training in collaboration: Master thesis, research visit,...
Service Description
Service description file*
<input type="button" value="Browse..."/> No file selected.
Service Notes*
<div style="border: 1px solid #ccc; height: 150px; width: 100%;"></div>

INFRASTRUCTURE REQUEST

[HOME](#)[SERVICES REQUEST](#) ▾[COUNSELING REQUEST](#)[INFRASTRUCTURE REQUEST](#)

Form for requesting Infrastructure service to Bioinformatic Unit

Service selection

Available Services *

- User support
 - Installation and support of bioinformatic software on Linux OS
 - Installation and access to Virtual machines in the Unit server containing bioinformatic software
 - Code snippets development
 - OT-2 robots

Service Description

Service description file Ningún archivo seleccionado**Service Notes** *

Análisis de Genomas Virales a través de la Plataforma Galaxy

Infrastructure and data management

skyLIMS

HOME RUN PREPARATION SEARCH STATISTICS REPORTS

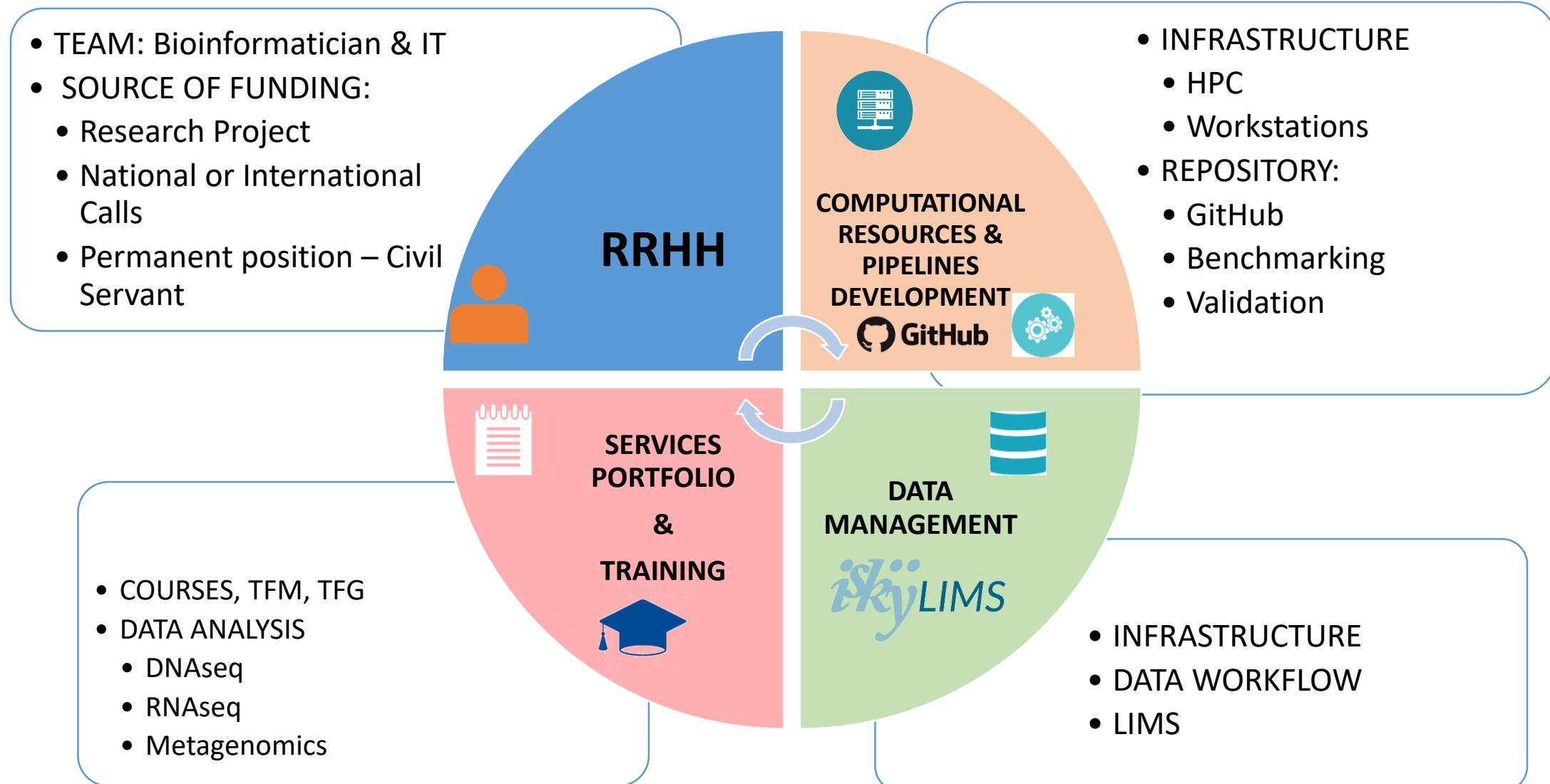
bioinfoadm Logout My account

Statistics results for Investigator rabad

Projects using the sequencer NS500454 :

[Export Table To Excel](#)

Project name	Date	Library 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 FI	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 FI	47	140,488,964	42,597	89.61	33.72	NS500454



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
RNaseq	MICROBIAL						
	WGS Amplicon	Report html	<i>De novo</i> / Reference (fasta file)	MLST, Resistance g, Virulence g	SNPs Phylogenetic analysis	Structural Functional	.WGSOutbraker .Plasmid ID
Metagenomics	mRNA	RSeQC Report html	<i>De novo</i> (fasta file)	Transcripts coverage / expression	Variants (Vcf file)	Transcripts annotation	mRNA seq
	miRNA						miRNA seq
	16S taxonomic profile	Report html	<i>De novo</i>	Green genes DB		species diversity	Qlime
	Shotgun			Genome Ref Seq		Pathogen / Genome coverage	PikaVirus

Galaxy

(  | 172.23.2.60)

Galaxy

Analyze Data Workflow Shared Data Visualization Help Login or Register Using 0 bytes

Tools

- search tools
- [Get Data](#)
- [Send Data](#)
- [Collection Operations](#)
- [Text Manipulation](#)
- [Filter and Sort](#)
- [Join, Subtract and Group](#)
- [Convert Formats](#)
- [Extract Features](#)
- [Fetch Sequences](#)
- [Fetch Alignments](#)
- [Statistics](#)
- [Graph/Display Data](#)
- [MyTools](#)
- [IRMA](#)
- [NGS Data Quality Check](#)

Workflows

- All workflows

Welcome to our Galaxy platform!

This galaxy server has been built and is maintained by the Bioinformatics Unit of Instituto de Salud Carlos III in order to give an user friendly enviroment to run limited bioinformatic tools and data analysis. Contact us if you are interested in the service and want to take an introductory course to the use this platform.

>X_BU-ISCIII

Instituto de Salud Carlos III

THIS IS A PROTOTYPE. If you find any bugs please report them to mjuliam@isciii.es

Take an interactive tour: [Galaxy UI](#) [History](#) [Scratchbook](#)

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by [The Galaxy Team](#) with the support of many contributors.

If you use this platform to analyse your data, remember to cite both Galaxy Project and this server.

The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

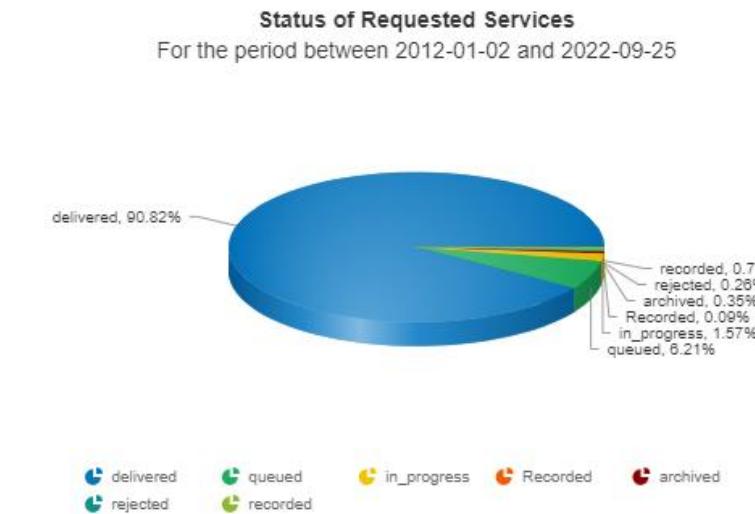
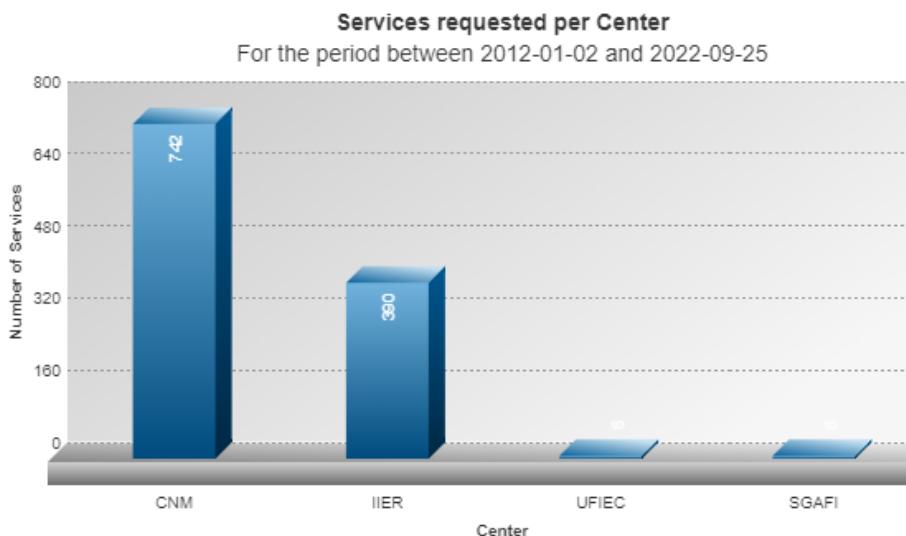
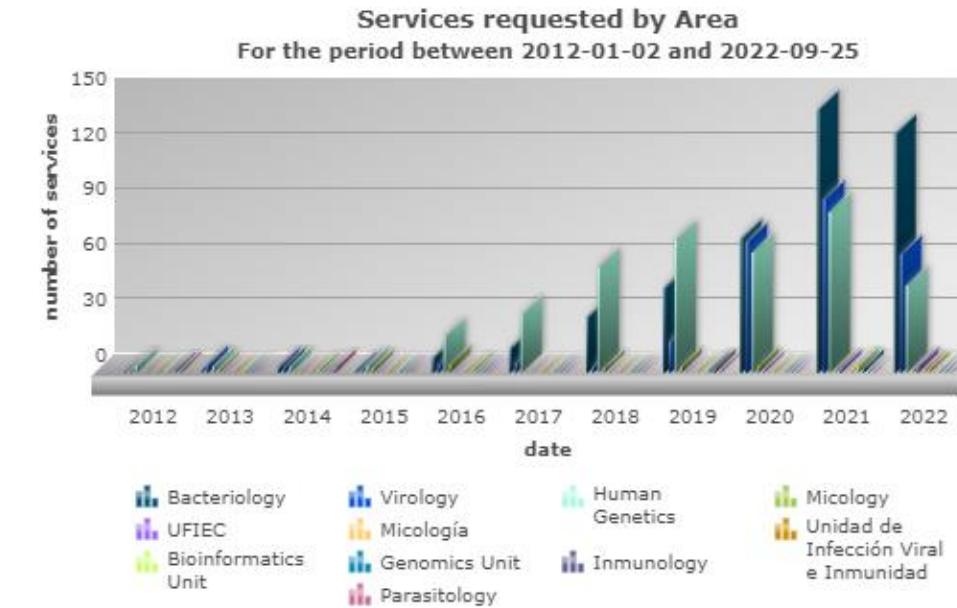
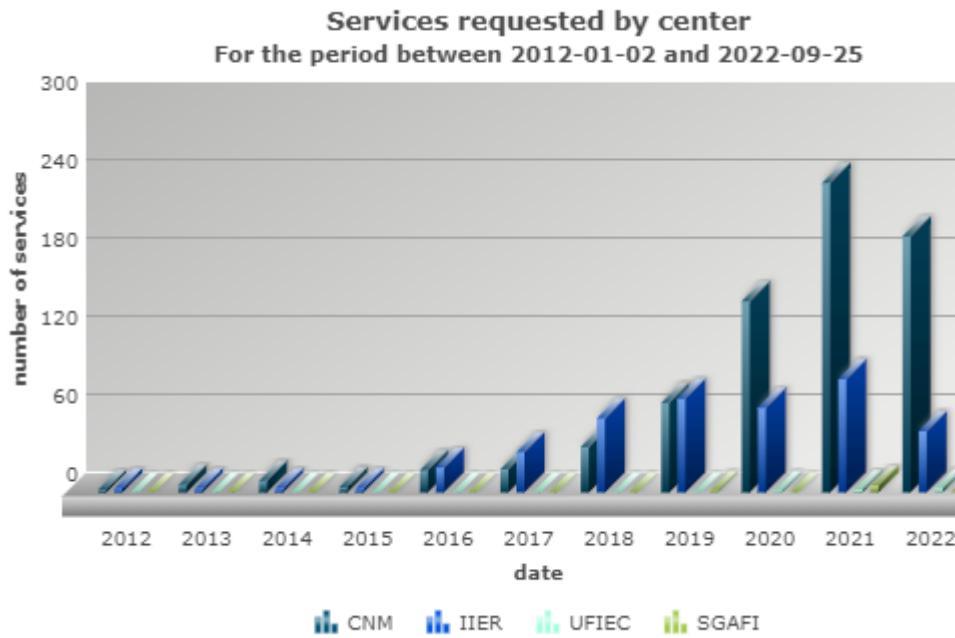
History

search datasets

Unnamed history (empty)

This history is empty. You can load your own data or get data from an external source

Number of services: 2012 – 2022


 10
AÑOS

Training

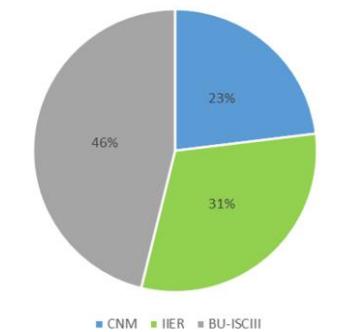
Courses

ISCIII

Introduction to massive sequencing data analysis, 2013-2021 (8 editions)

Secuenciación de genomas bacterianos: herramientas y aplicaciones, 2018-2021 (3 editions)

Análisis de genomas virales a través de la plataforma Galaxy, 2021 (1 edition)

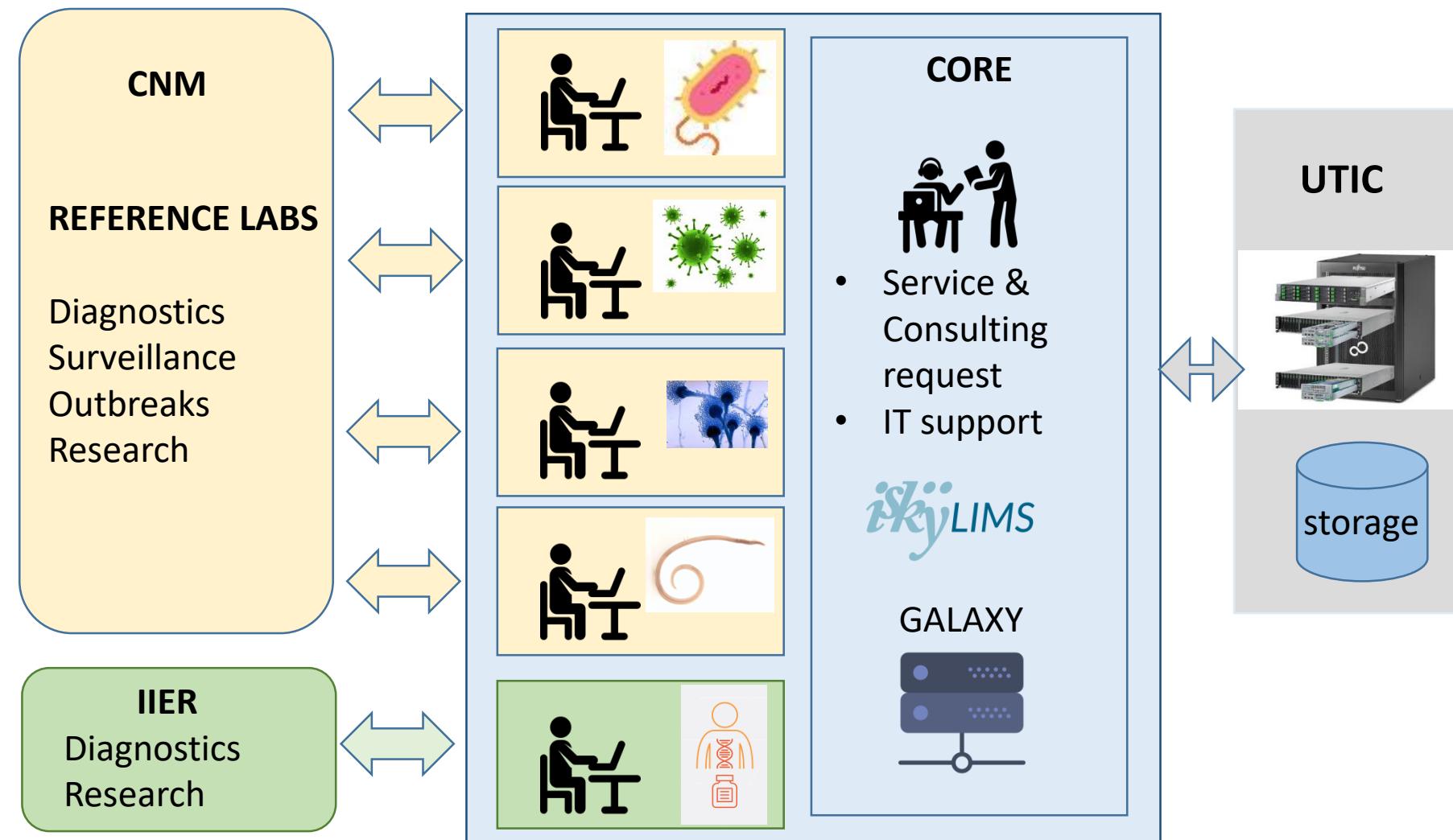


Master & Grade Students

- 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

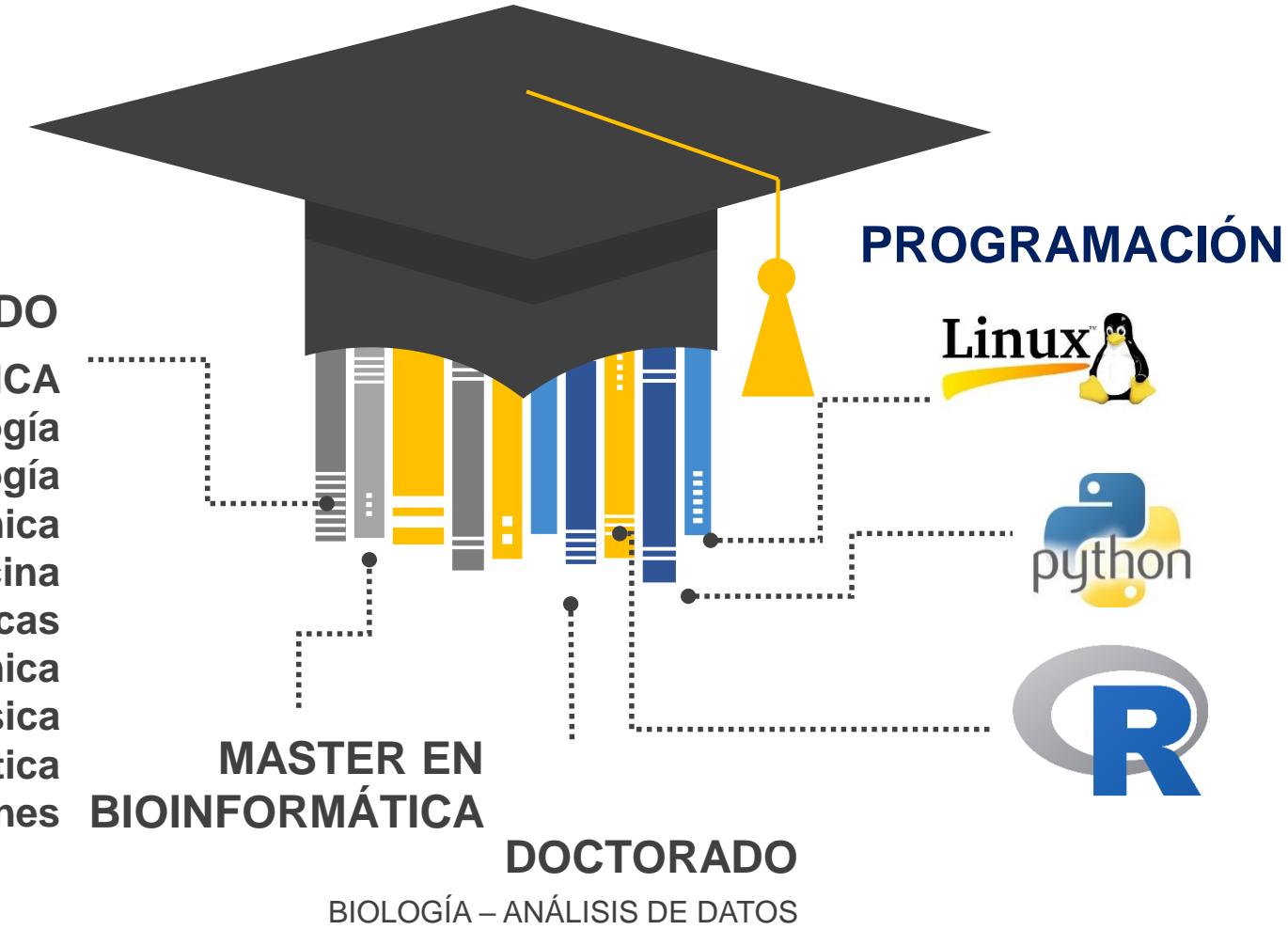
Roadmap: BU-ISCIII Model



FORMACIÓN EN BIOINFORMÁTICA

Universidad
Barcelona.

GRADO
BIOINFORMÁTICA
Biología
Biotecnología
Bioquímica
Medicina
Matemáticas
Química
Física
Informática
Telecomunicaciones



¿Dónde trabaja un Bioinformático?



UNIVERSIDAD

Biociencias
Informática

CENTRO DE INVESTIGACIÓN



EMPRESA

Bioinformática
Genética
Genómica

Biomedicina
Agricultura
Alimentación

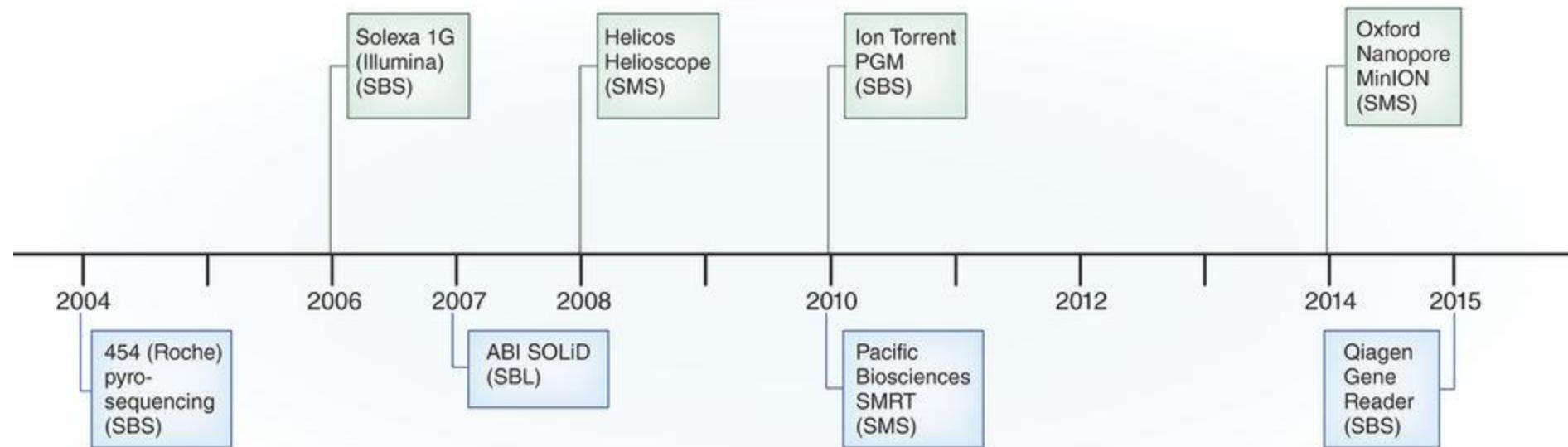
HOSPITAL

BIOINFORMÁTICO
CLÍNICO
Genética
Oncología
Cardiología

Index

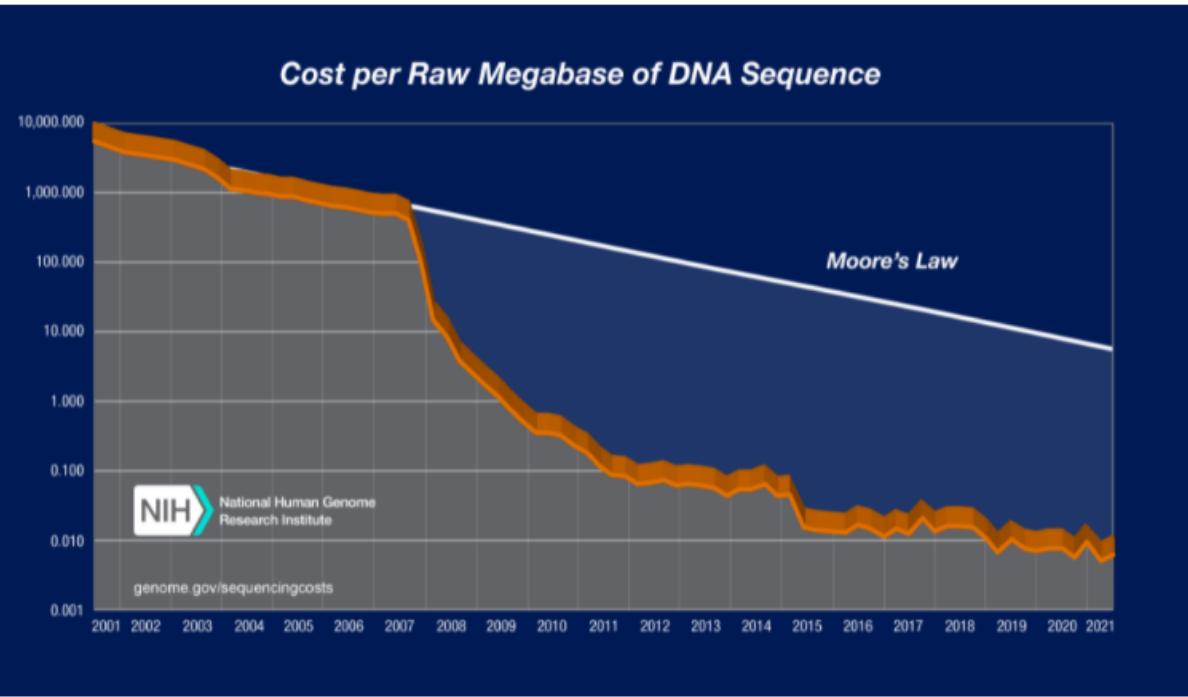
- **High throughput sequencing in clinical virology, Applications**

DNA sequencing technologies 2006-2016

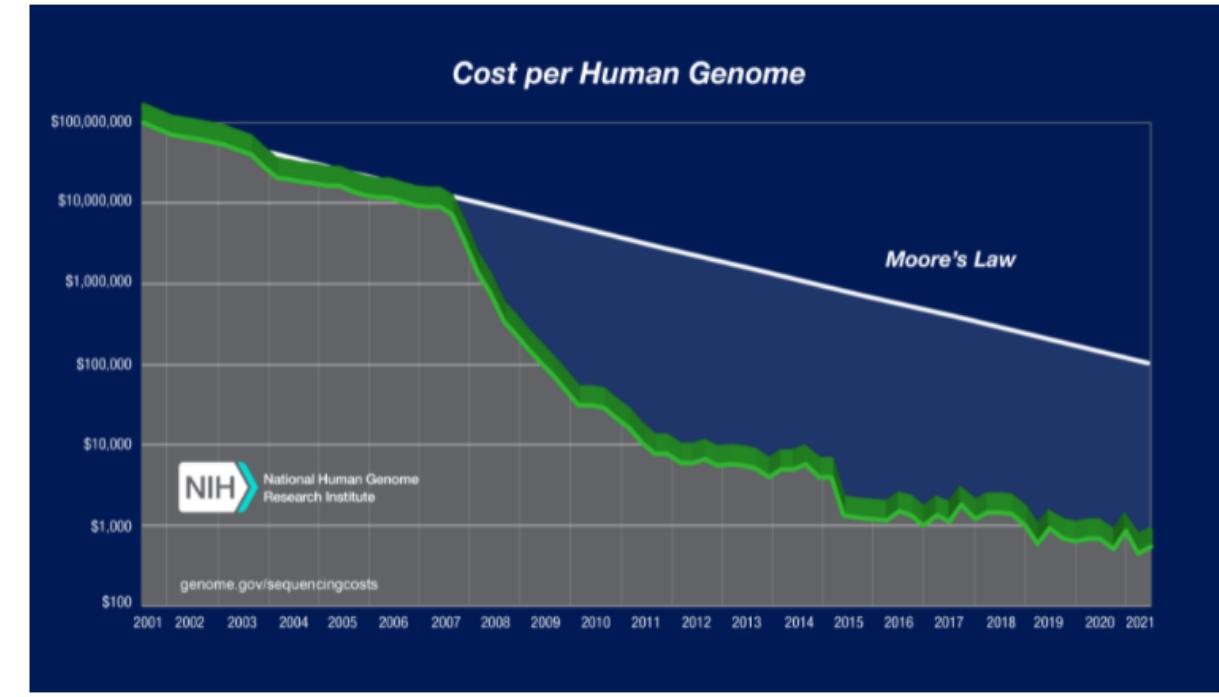


Mardis, Nature Protocols 2017

Secuencing cost



Sequencing cost per megabase - 2021

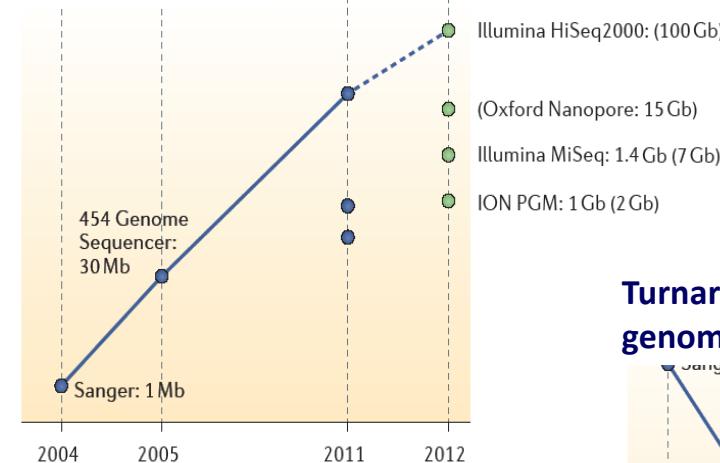


Cost per genome data - 2021

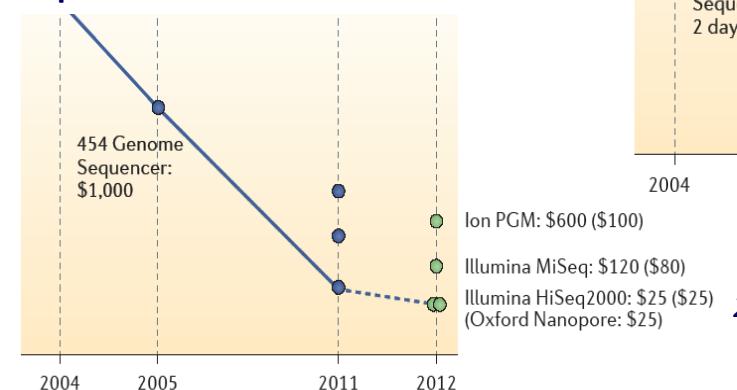
<https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>

Sequencing platforms in Microbiology

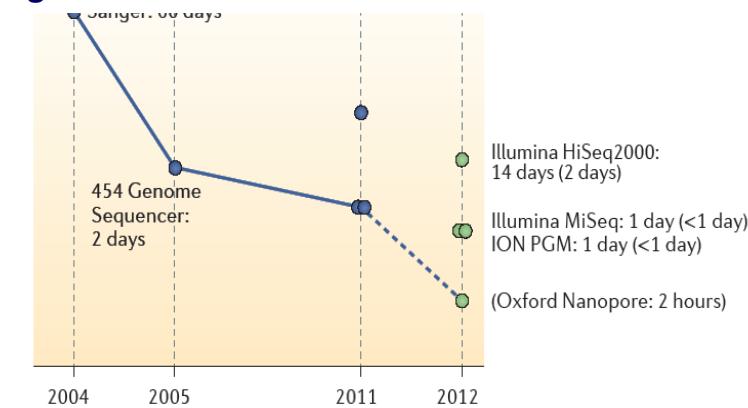
Raw daily output



Cost per Mb assembled sequence



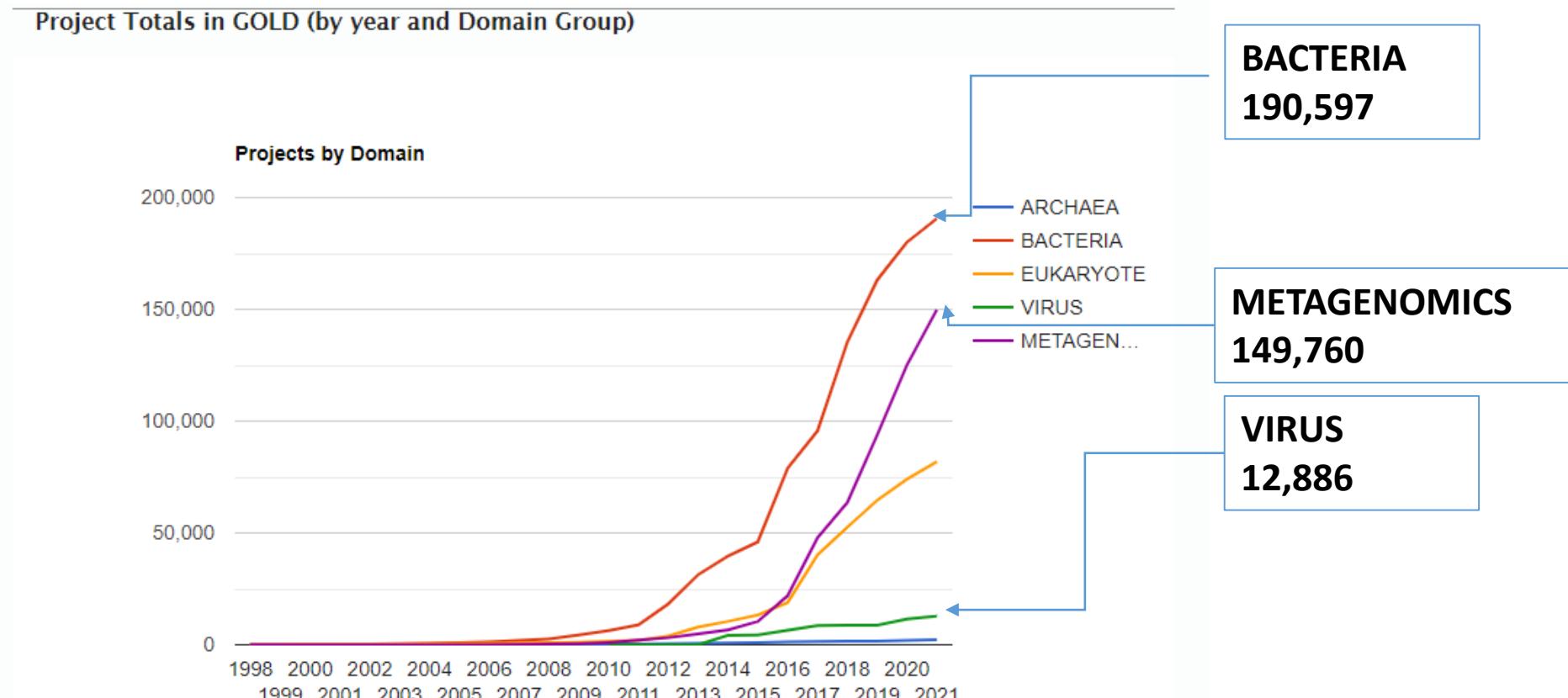
Turnaround time: bacterial genome



Sequencing projects

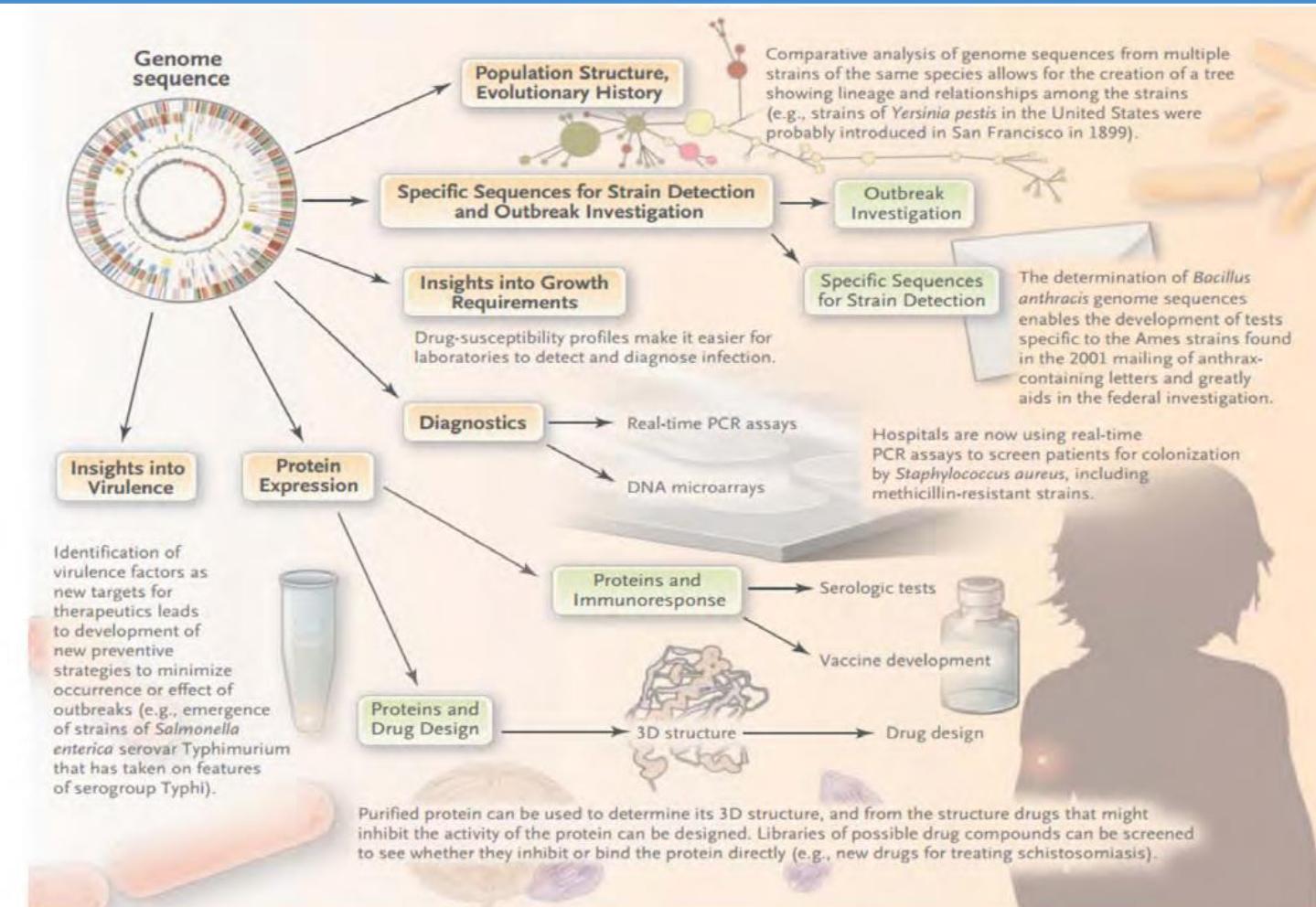
<https://gold.jgi.doe.gov/>

GOLD, Genome Online DataBase



Use of microbial genomics for tool development

Report from The American Academy of Microbiology, 2015



APPLICATIONS OF NGS IN VIROLOGY

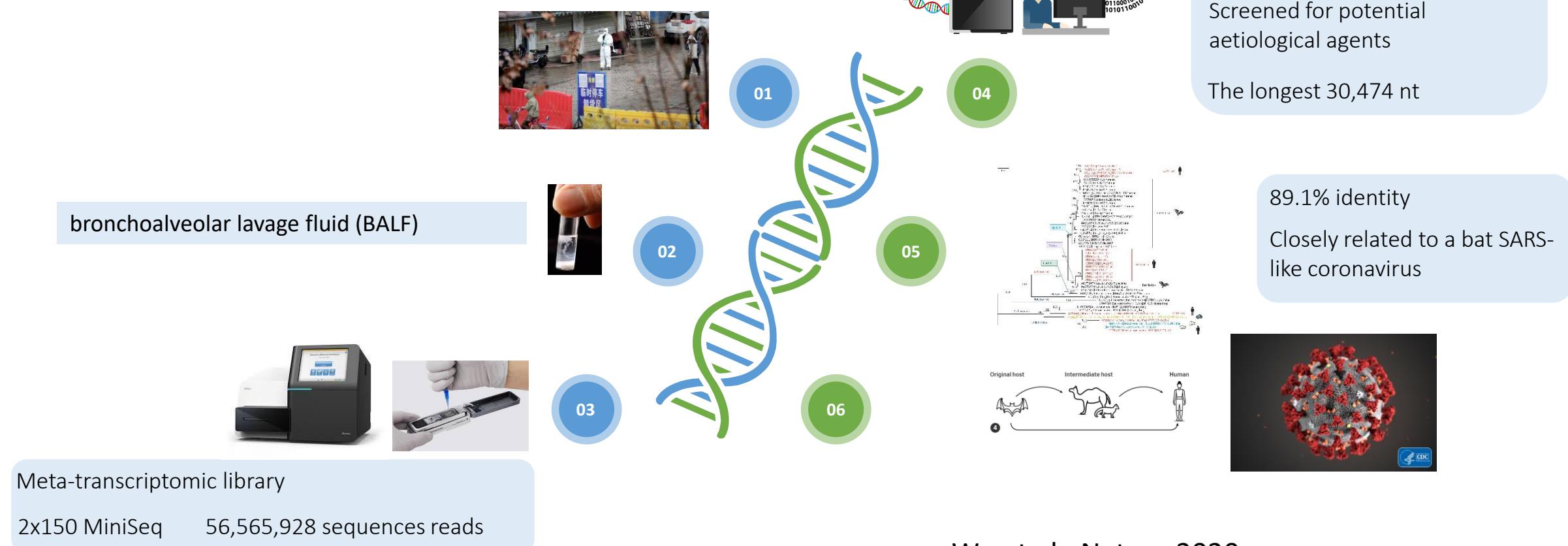
- ◆ Detección de patógenos virales no identificados en la muestra
- ◆ Identificación de virus nuevos
- ◆ Identificación de virus en muestras tumorales
- ◆ Caracterización del Viroma de un organismo
- ◆ Secuenciación del genoma viral completo
- ◆ Estudio de la Variabilidad genómica (Quasispecies)
- ◆ Monitorización de la resistencia a los Antivirales
- ◆ Vigilancia epidemiología 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

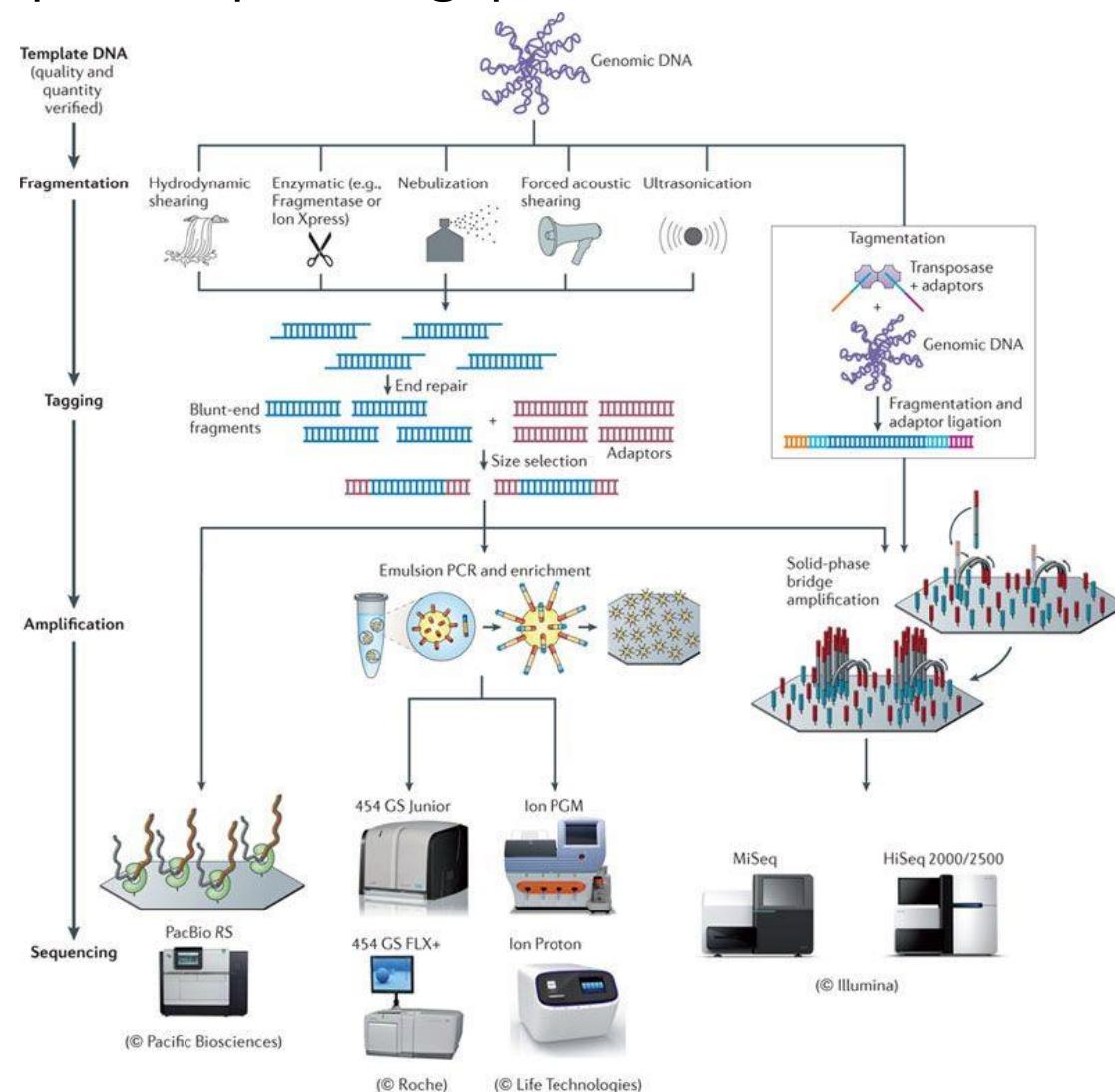
Pathogen discovery: new virus – SARS-CoV-2

Deep Meta-Transcriptomic Sequencing



Wu et al., Nature 2020

High-throughput sequencing process



Nature Reviews | Microbiology Loman et al, 2012

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

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

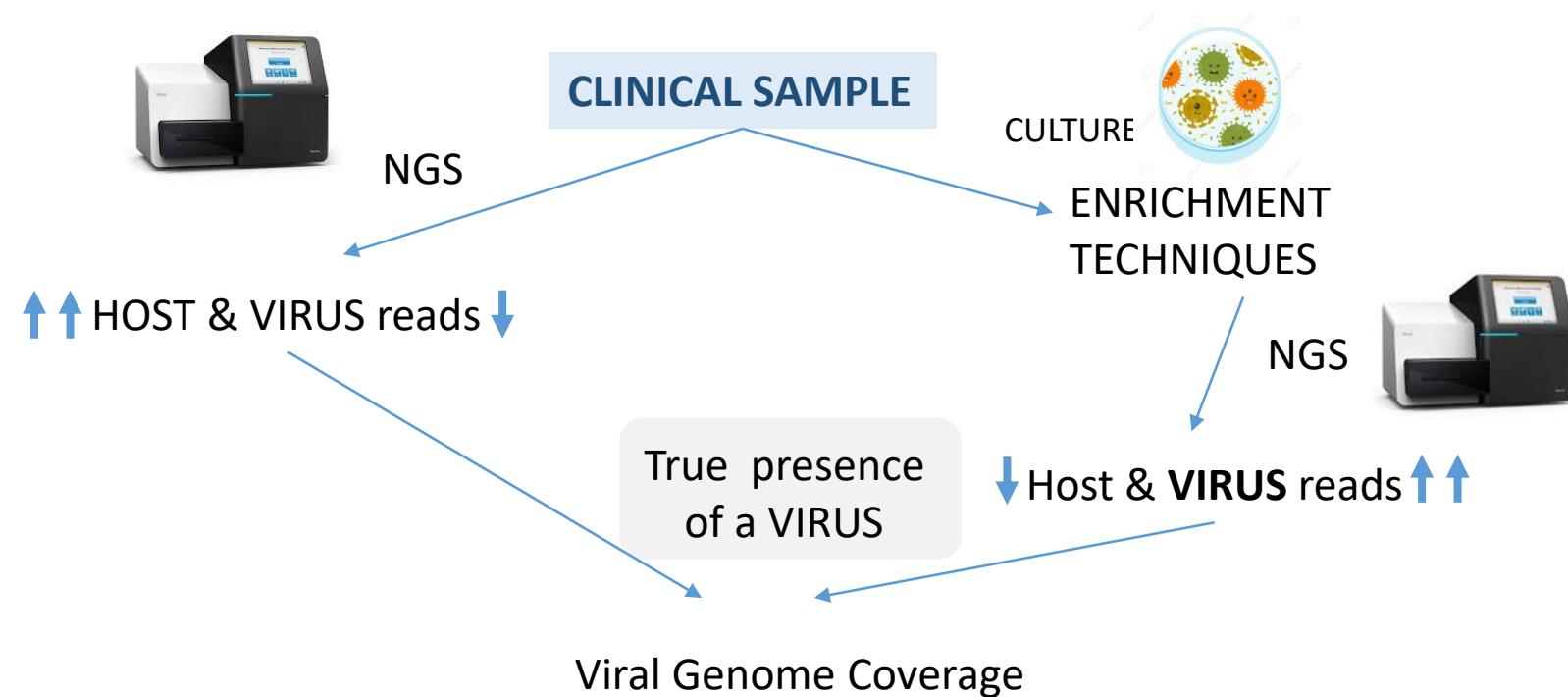
RNA segment	No. of nucleotides	Encoding protein	No. of amino acids
1	2341	Polymerase PB2	759
2	2341	Polymerase PB1	757
3	2233	Polymerase PA	716
4	1778	Haemagglutinin HA	566
5	1565	Nucleoprotein NP	498
6	1413	Neuraminidase NA	454
7	1027	Matrix protein M1	252
		Matrix protein M2	97
8	890	Non-structural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

Main Steps of Viral Genome Sequencing by NGS or HTS

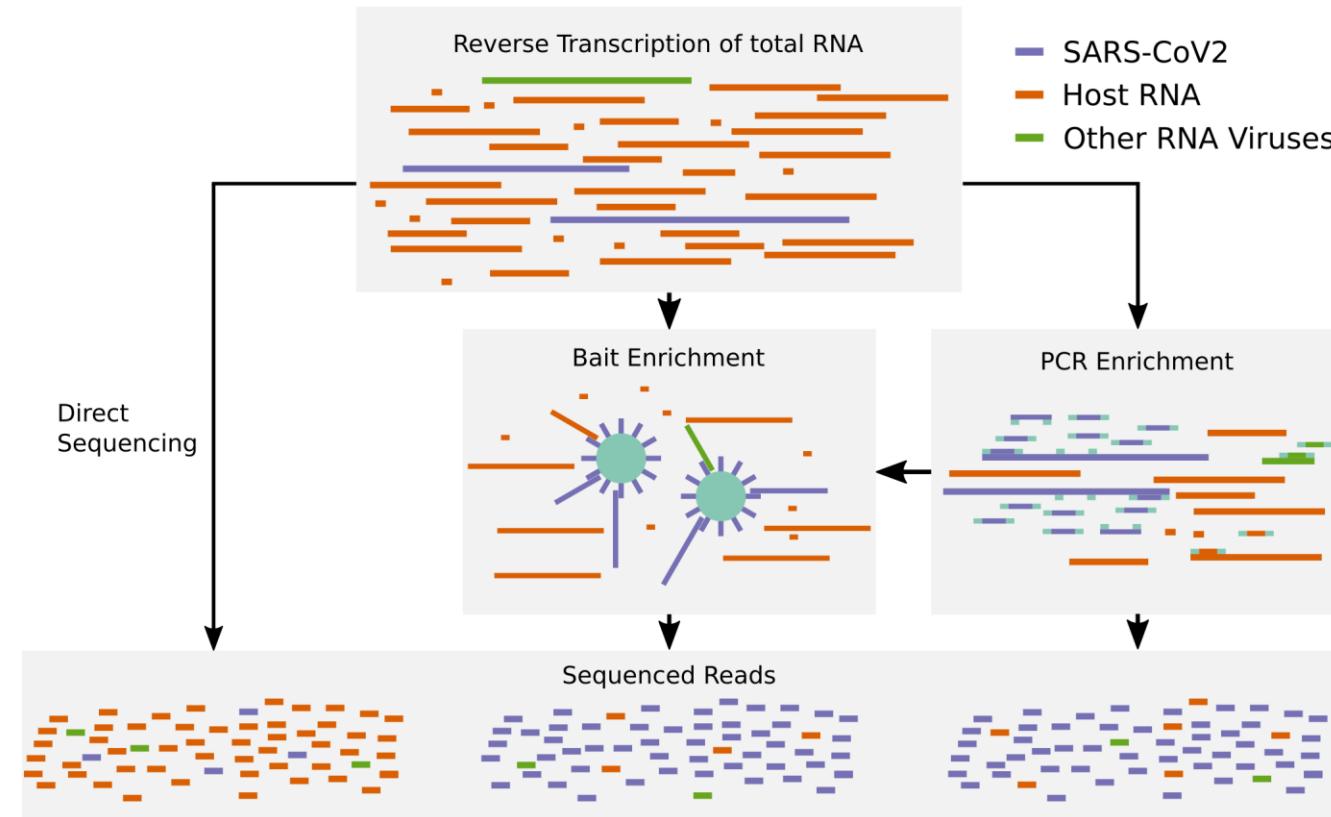
- Nucleic acid amplification
- Library preparation
- High throughput sequencing platforms
- Data analysis

Viral Genome Sequencing

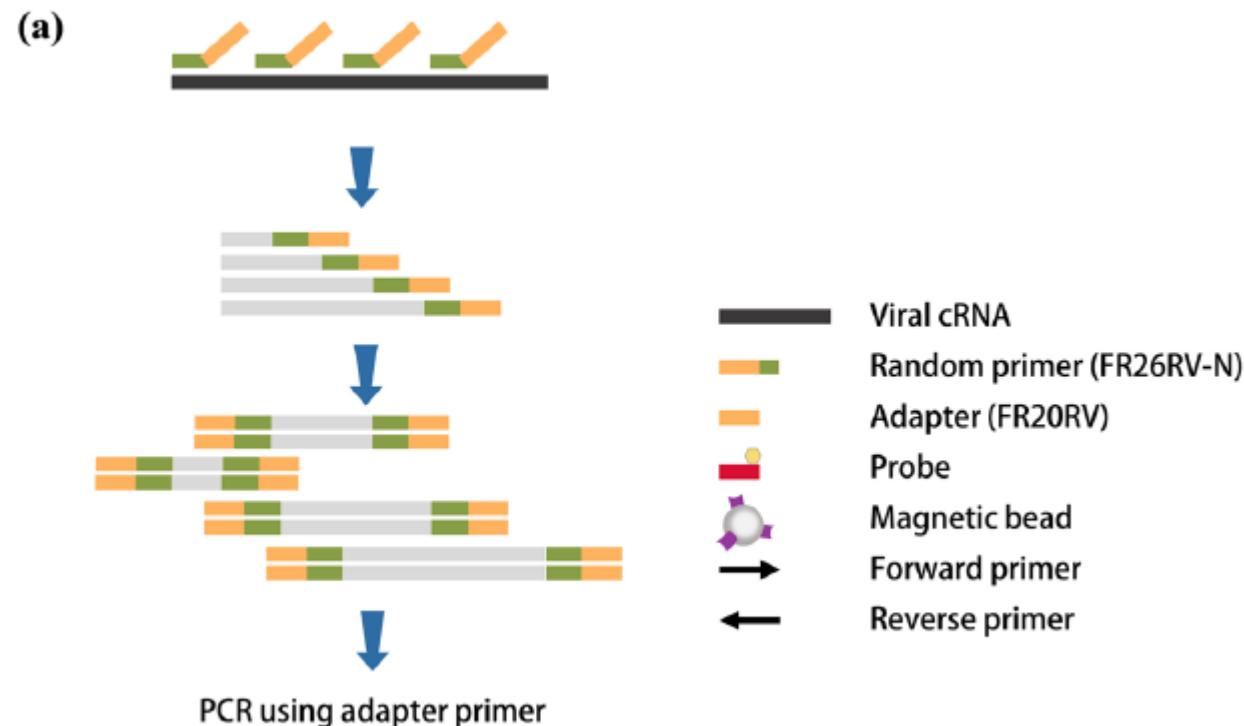


NGS needs a cutoff to determine the true presence of a pathogen versus carry-over or contamination between specimens or other non-specific reads.

Enrichment Techniques



Nucleic Acid Extraction: SISPA NGS method



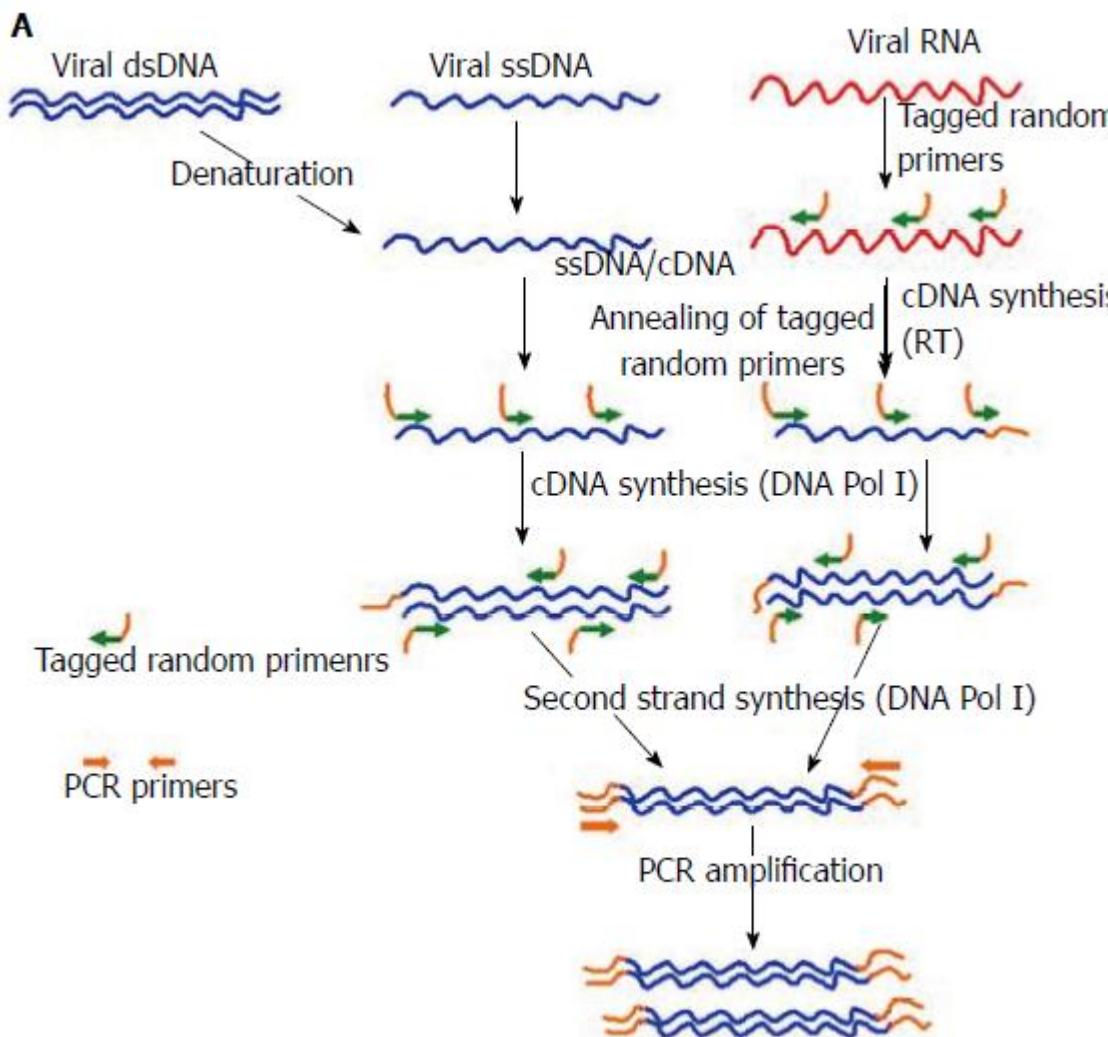
VIRUS DISCOVERY

- 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)

Jin Sun No et al., Scientific Reports (2019) 9:16631 | <https://doi.org/10.1038/s41598-019-53043-2>

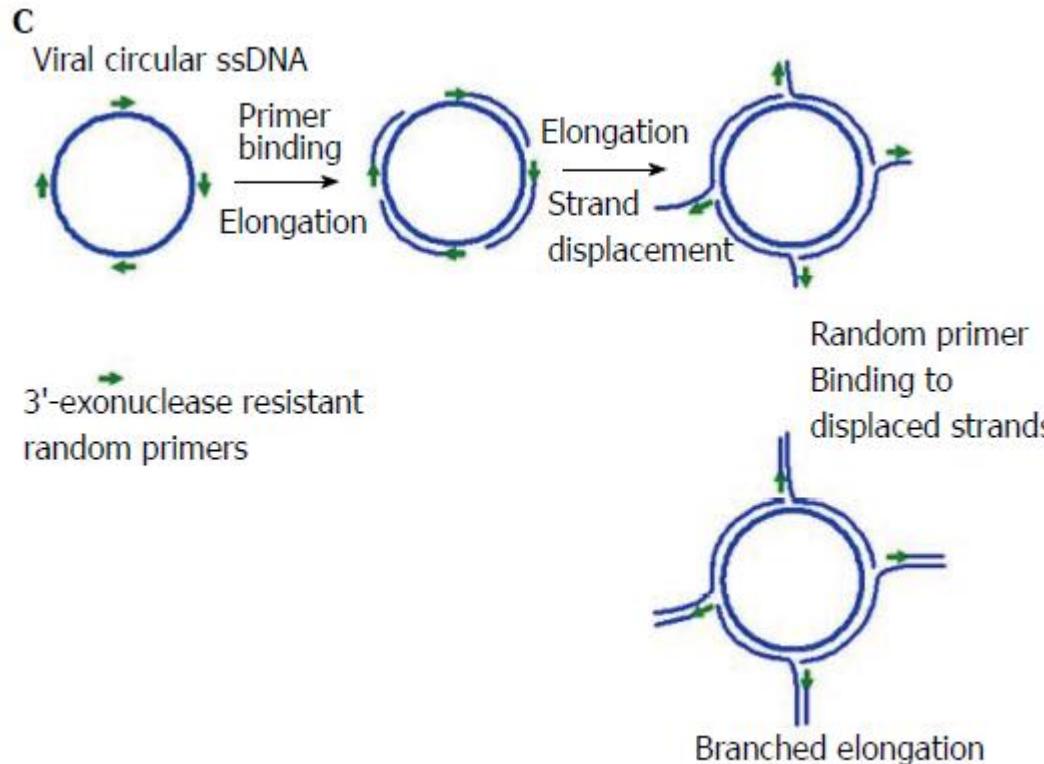
Sequence-independent single-primer amplification



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

Schema main steps of clinical virus Discovery by NGS

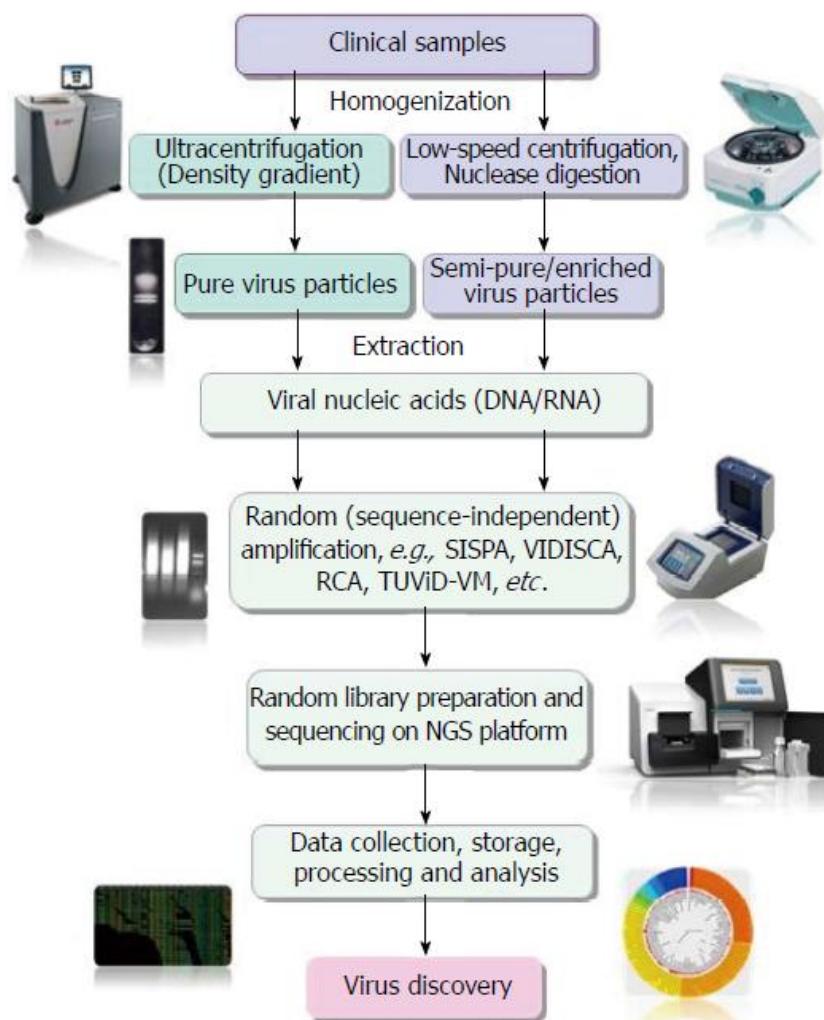
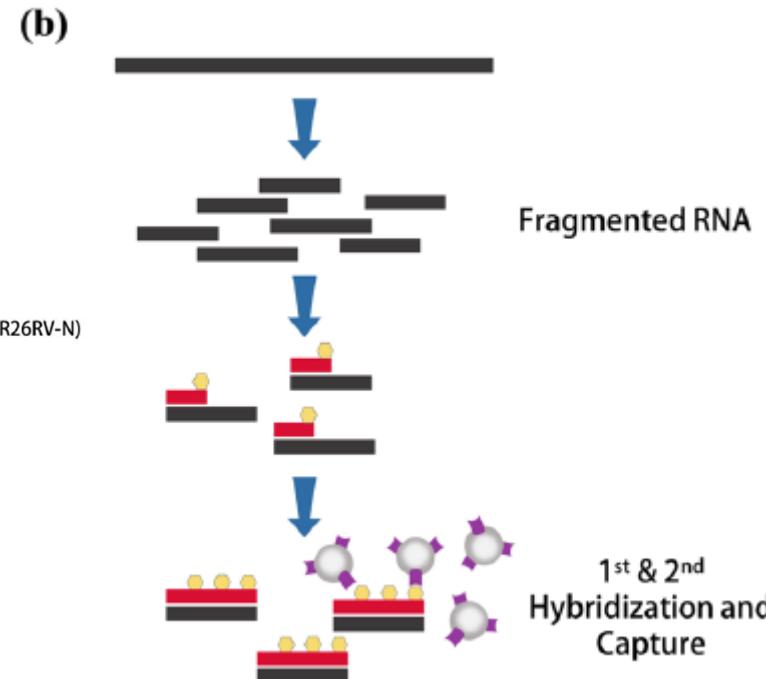


Figure 1 Diagrammatic representation of main steps of clinical virus discovery by next-generation sequencer based technologies.

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

Nucleic Acid Extraction: Target capture NGS

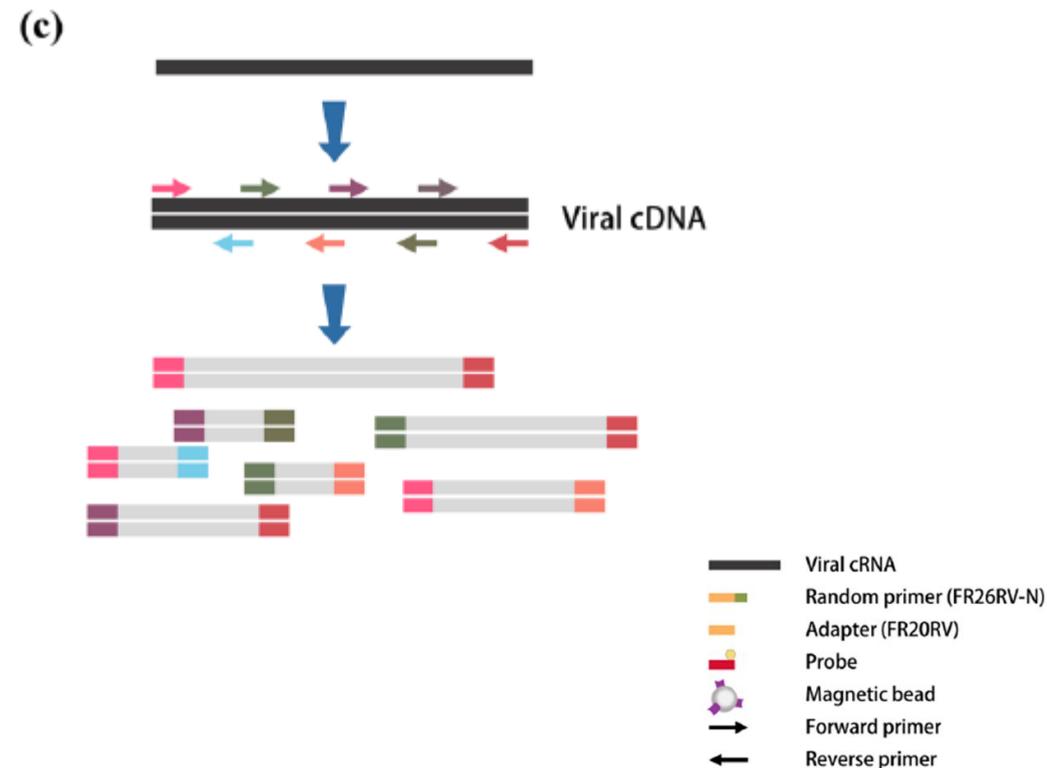


WHOLE VIRAL GENOME RECONSTRUCTION

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.

Jin Sun No et al., Scientific Reports (2019) 9:16631 | <https://doi.org/10.1038/s41598-019-53043-2>

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**

Jin Sun No et al., Scientific Reports (2019) 9:16631 | <https://doi.org/10.1038/s41598-019-53043-2>

Nucleic Acid Extraction COMPARISON, Hantaan orthohantavirus genome sequencing

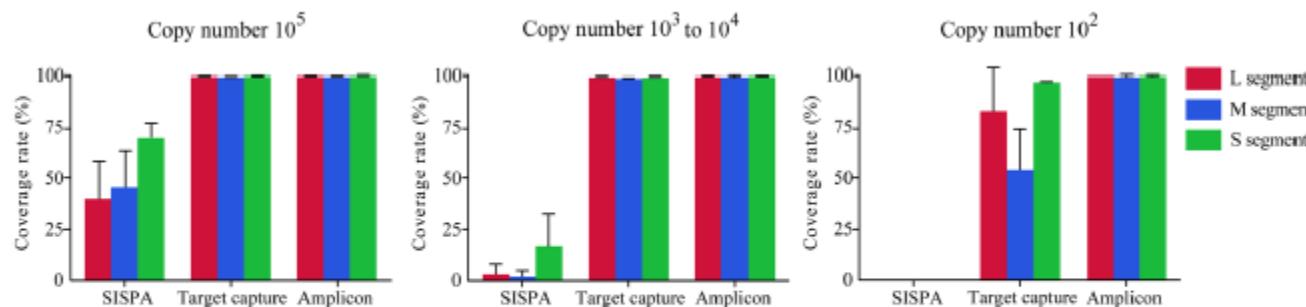


Figure 3. Coverage rate of consensus sequences for Hantaan orthohantavirus (HTNV) by viral copy number. The percentage coverage rate of consensus sequences by viral copy number among the three next-generation sequencing methods. Coverage rate was calculated by matching the consensus sequences with the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

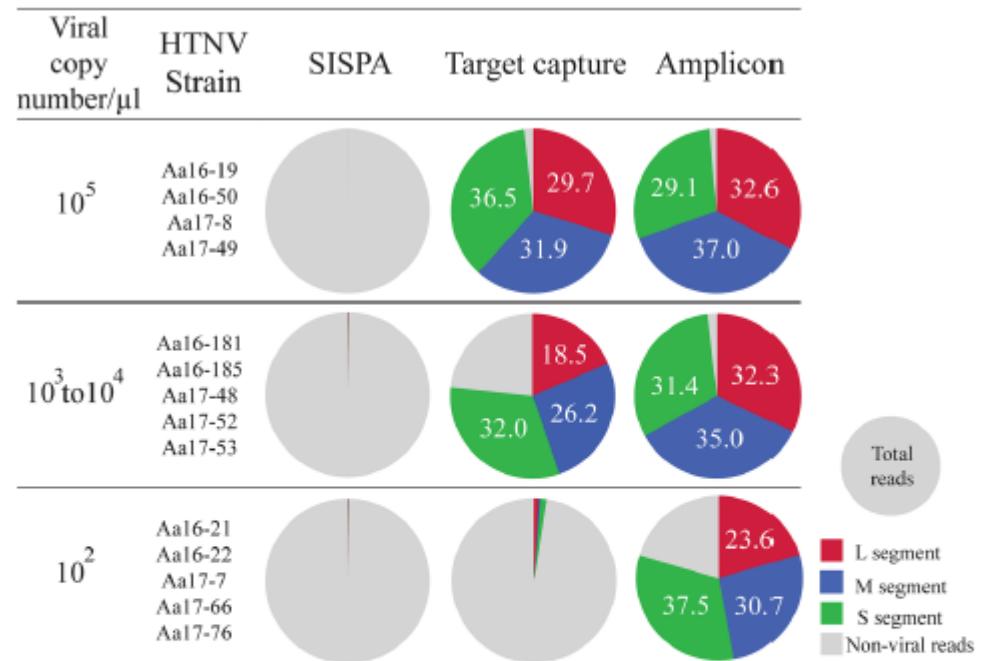
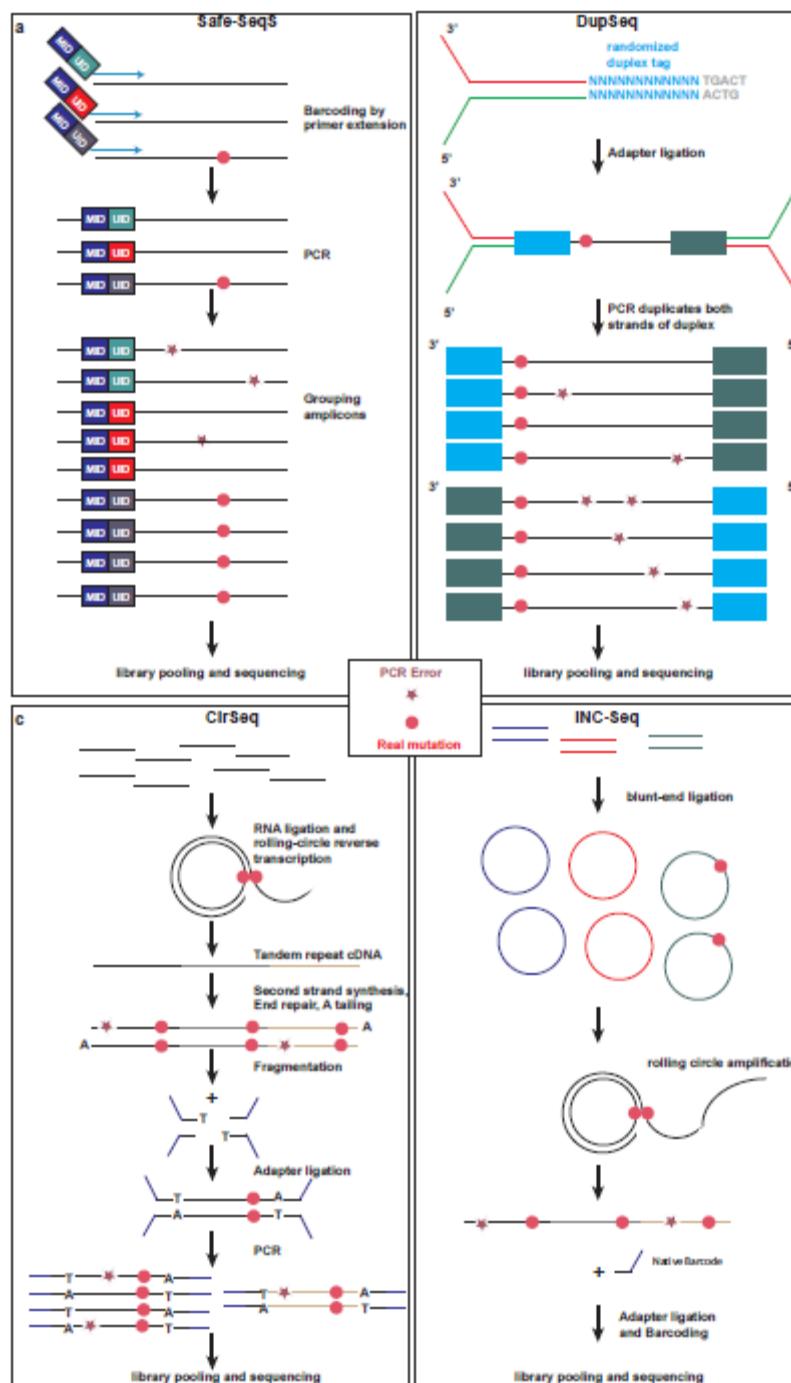


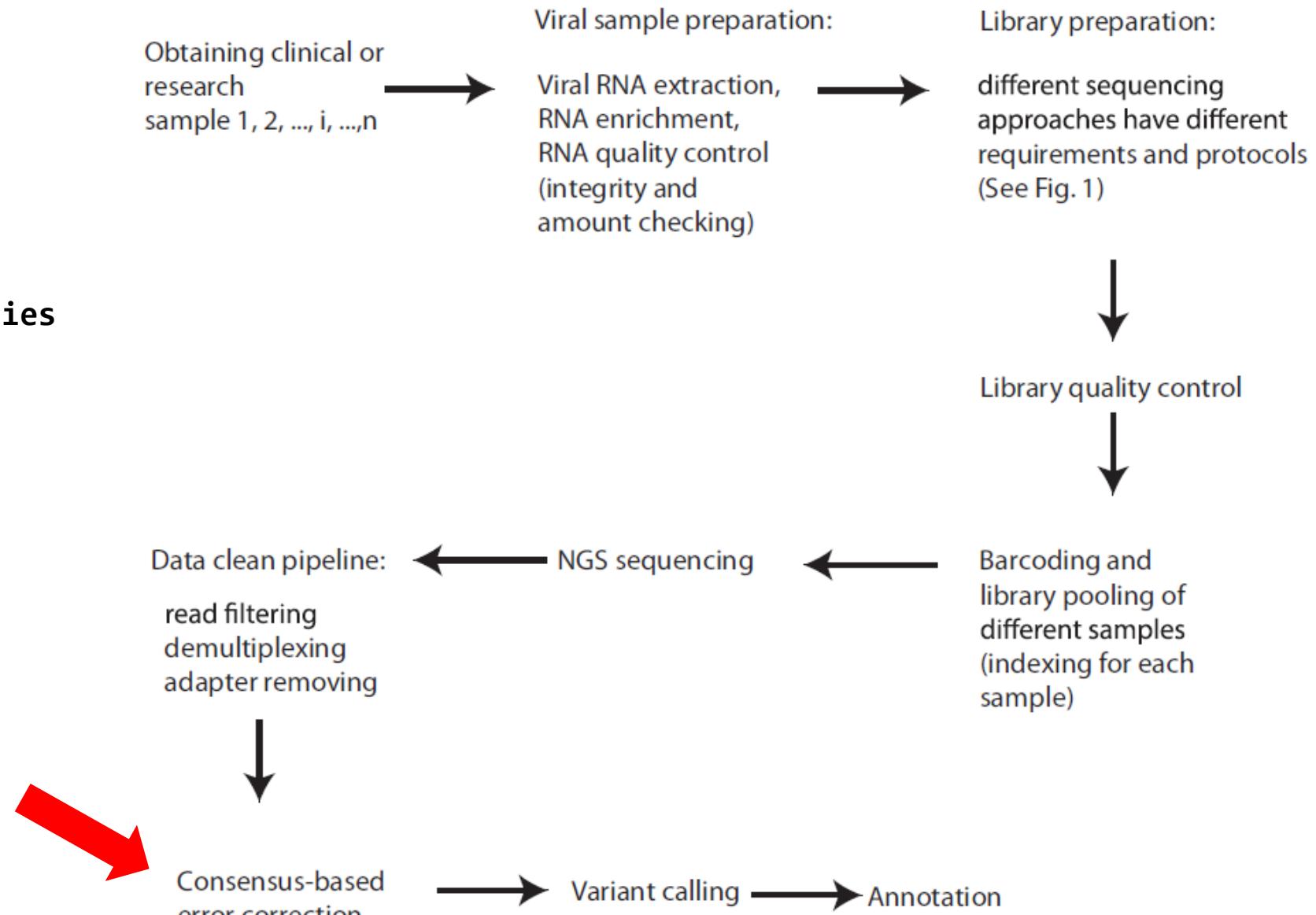
Figure 4. The composition of Hantaan orthohantavirus (HTNV) genomic reads in the total reads generated by three next-generation sequencing methods. The composition of viral reads mapped to HTNV was exhibited in the total number of reads. The total reads were produced by performing SISPA, target capture, 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

**Nucleic Acid Extraction
low frequency variant
identification, quasispecies**



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

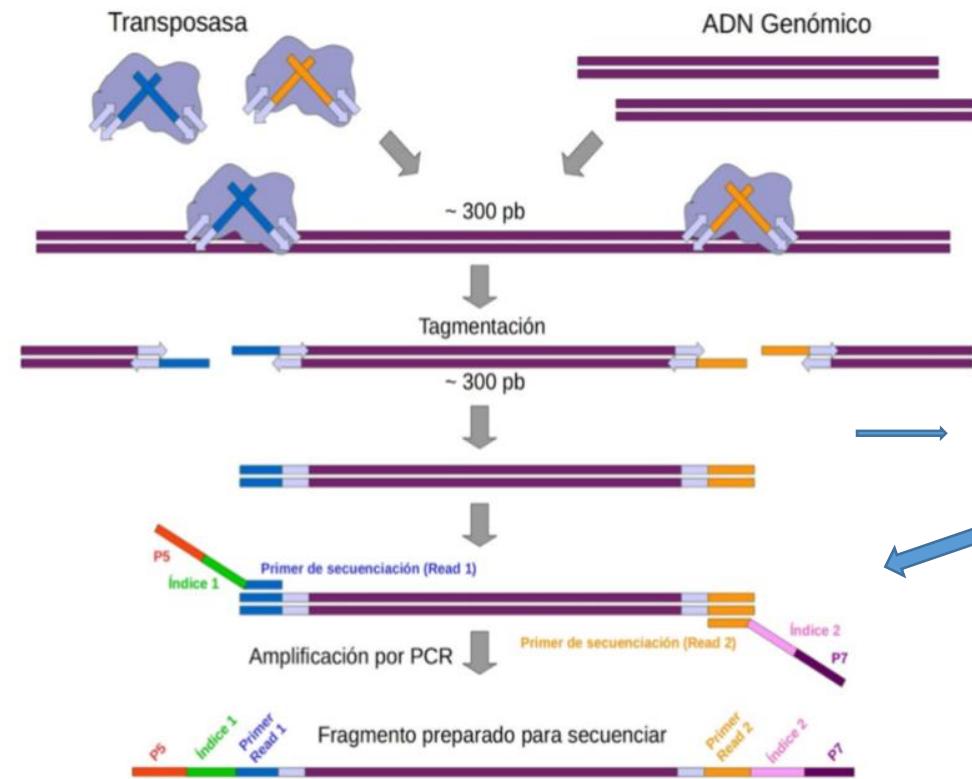
1. Metagenómica

IDENTIFICACIÓN MICROORGANISMOS

1. Metataxonomía o Targeted Metagenomics

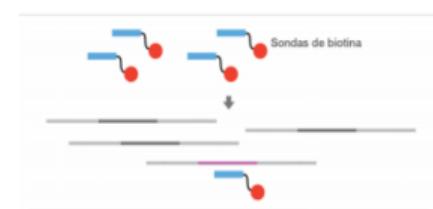
PREPARACIÓN LIBRERÍA

ENZIMÁTICA FÍSICA



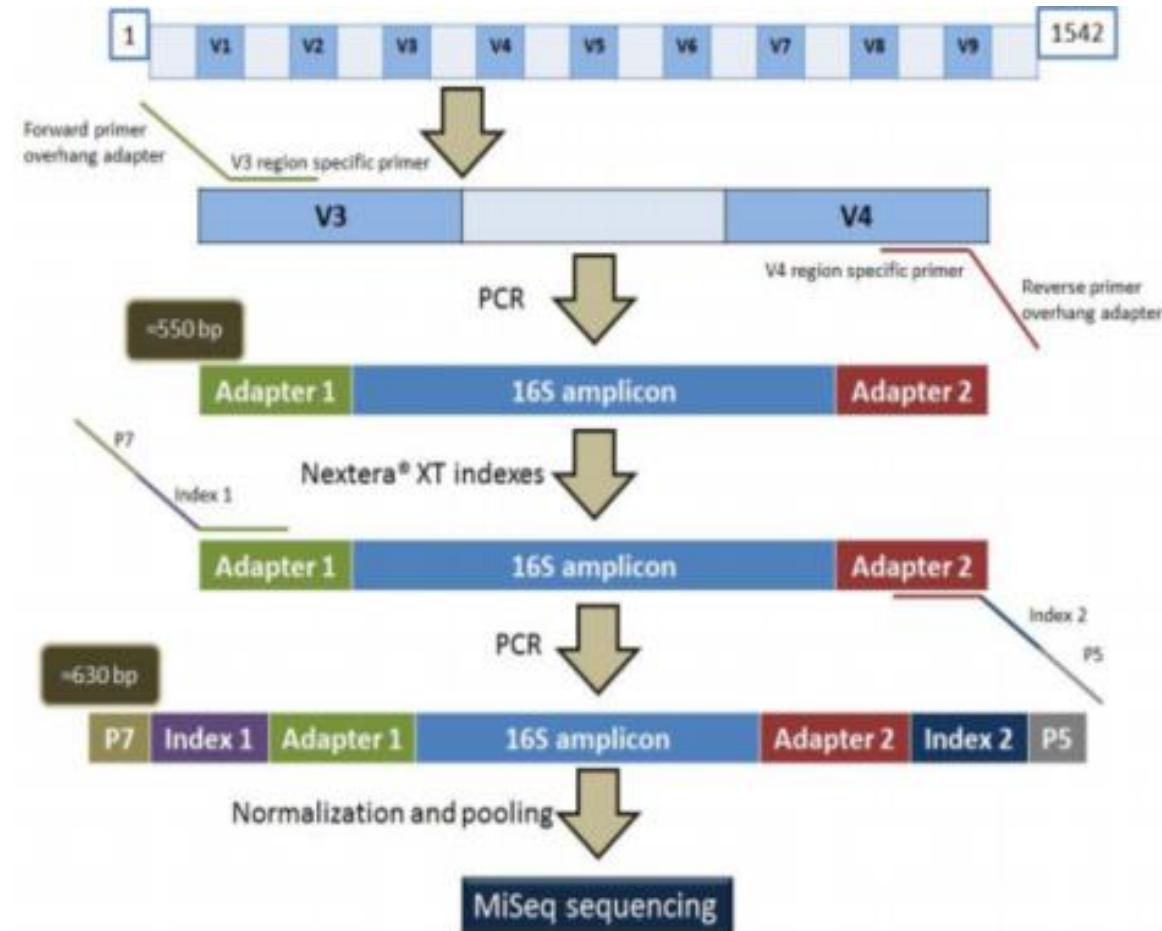
RNA → cDNA

ENRIQUECIMIENTO:
PCR
CAPTURA SONDAS



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

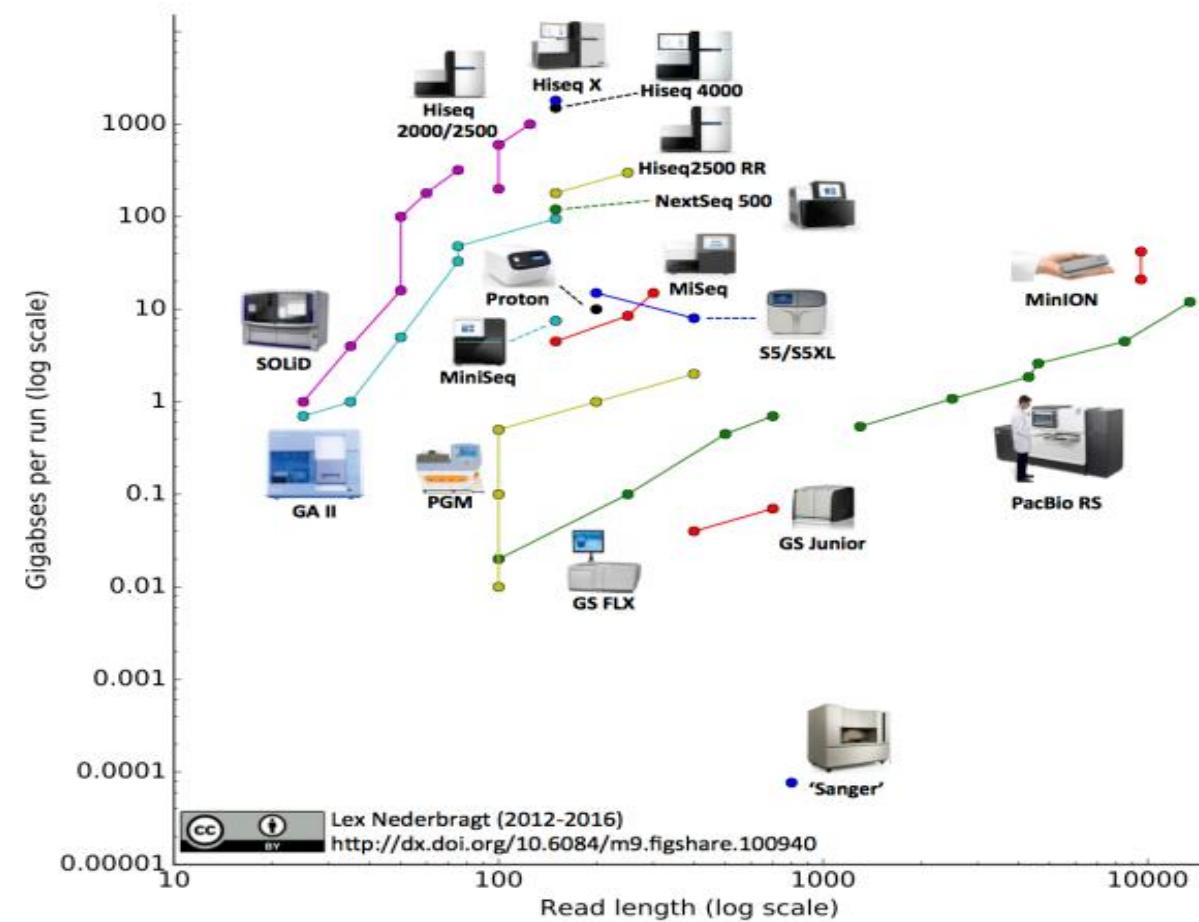
PREPARACIÓN LIBRERÍA, rRNA 16S, caracterización microbiota



NGS PLATFORMS, main characteristics

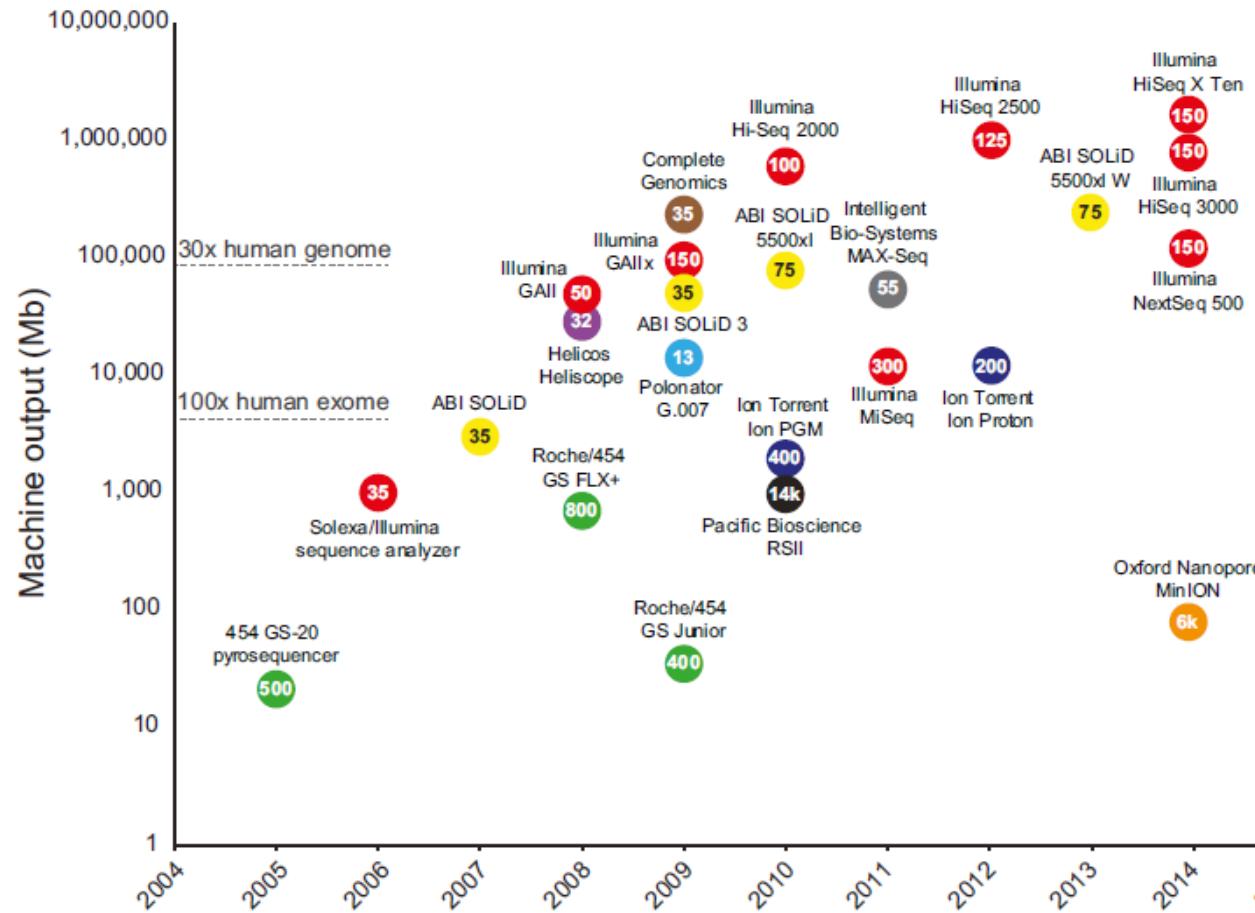
- Numero de bases que secuencia
- Numero lecturas → aplicaciones
- Longitud de las lecturas -→ importante para las aplicaciones ensamblado genomas, de illumina a PacBio
- Error de la base → Corrección con profundidad de lectura
- Formato fichero salida
- Software dedicado, universal fastq

High-Throughput Sequencing Technologies



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

High-Throughput Sequencing Technologies

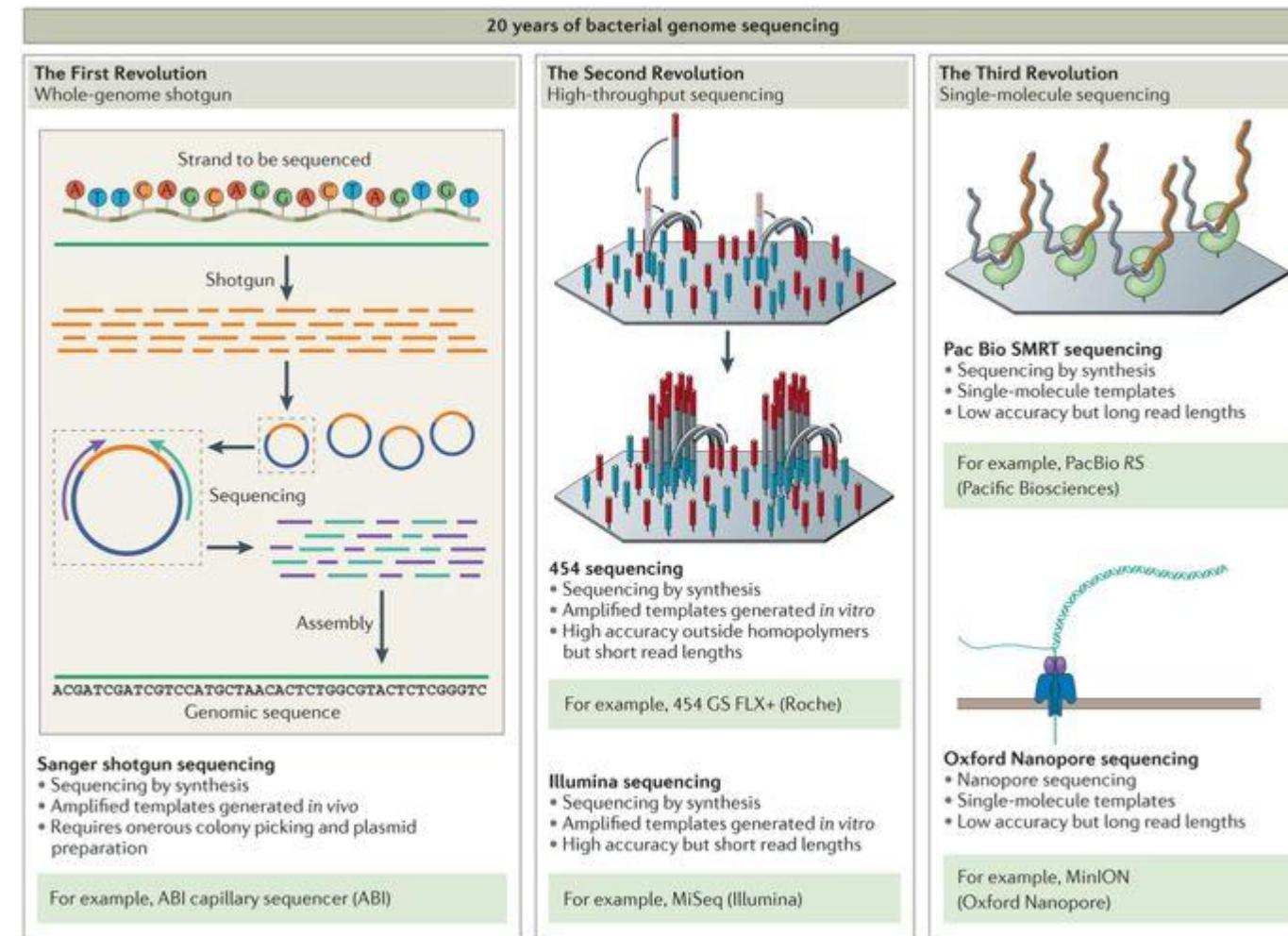


Numbers inside data points denote current read lengths.
Sequencing platforms are color coded.

Reuter et al., Mol Cell 2015

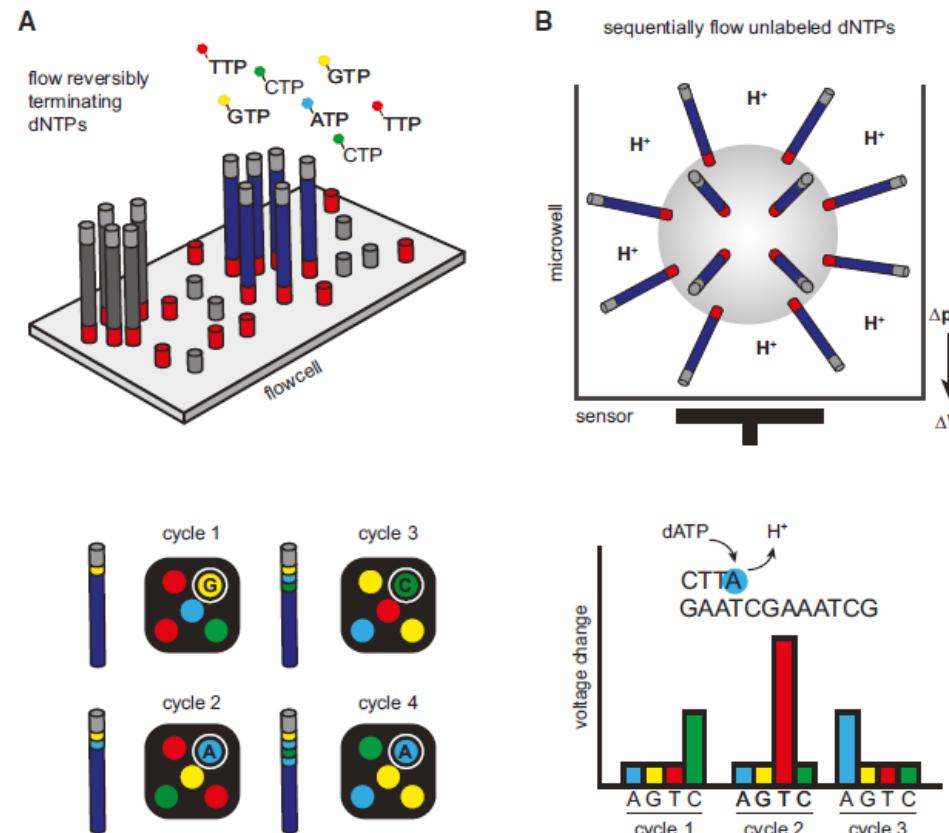
High-Throughput Sequencing Technologies

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



Nature Reviews | Microbiology

The Second-generation Sequencing Technologies



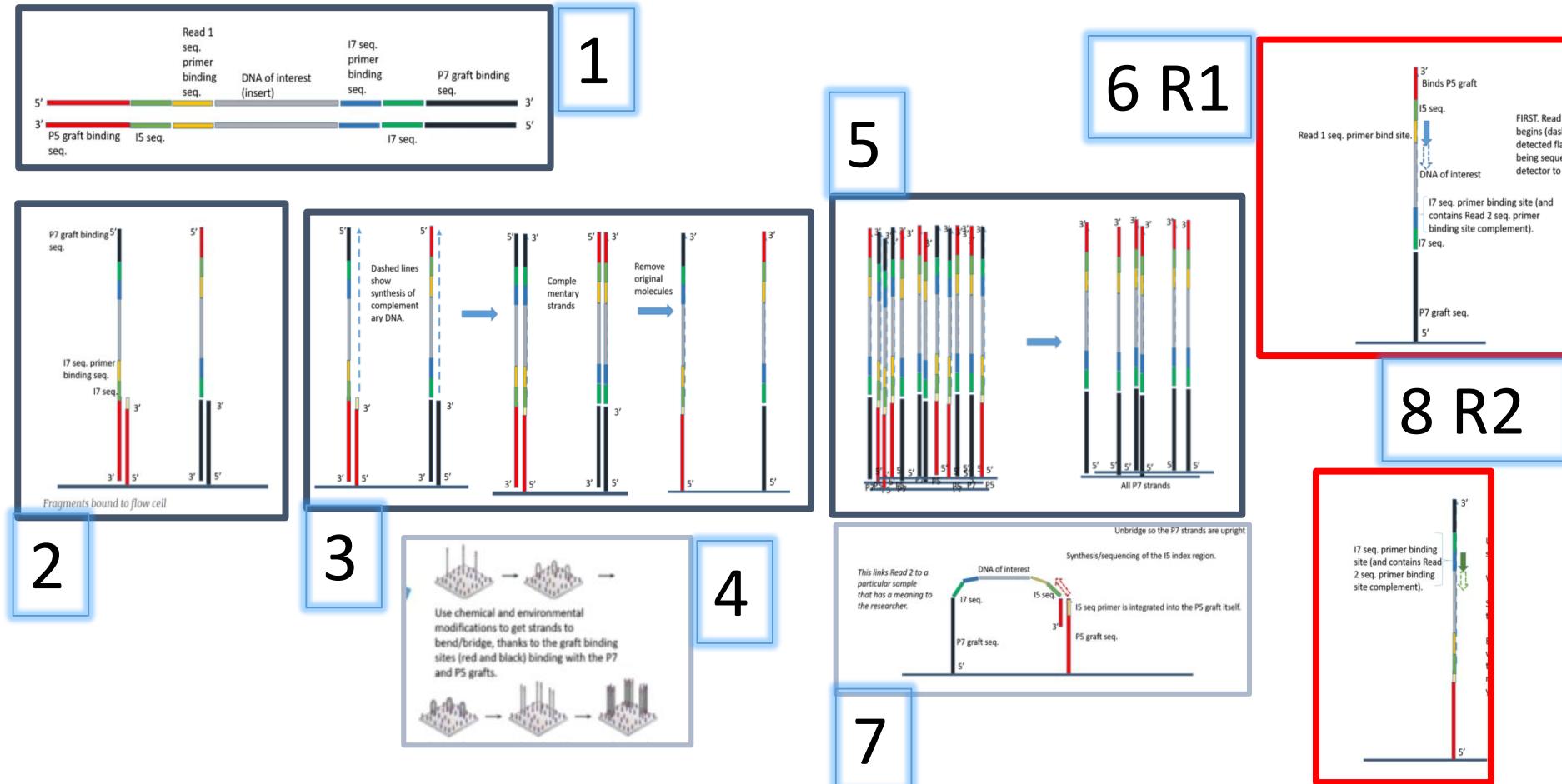
Clonal Amplification-Based Sequencing Platforms

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

(B) Ion Torrent's semiconductor sequencing method.

Reuter et al., Mol Cell 2015

Illumina sequencing



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

Illumina benchtop sequencers

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



	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 >					
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



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

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, miRNA-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

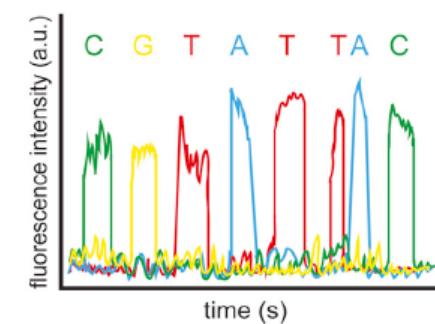
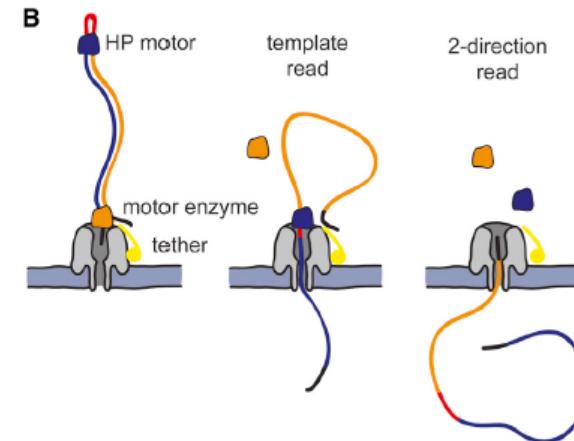
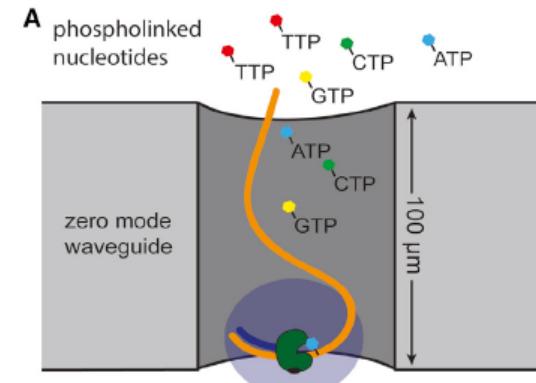
See how a Laboratory Information Management System (LIMS) enabled this large genomics lab to standardize lab procedures and cope with increasing sample volumes from diverse clients.

[Read Case Study >](#)

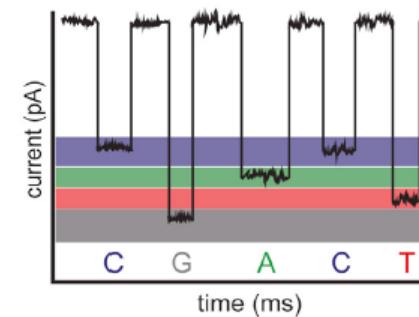
Run Time	12-30 hours	11-48 hours	-13 - 38 hours (dual SP flow cells) -13-25 hours (dual S1 flow cells) -16-36 hours (dual S2 flow cells) -44 hours (dual S4 flow cells)
Maximum Output	120 Gb	330 Gb*	6000 Gb
Maximum Reads Per Run	400 million	1.1 billion*	20 billion
Maximum Read Length	2 x 150 bp	2 x 150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

The Third-generation Sequencing Technologies

Single Molecule Sequencing Platforms



Pacific Bioscience's SMRT sequencing



Oxford Nanopore's sequencing strategy

Reuter et al., Mol Cell 2015

PacBio sequencing and its applications

Rhoads & Au, Gen Prot Bioinf 2015



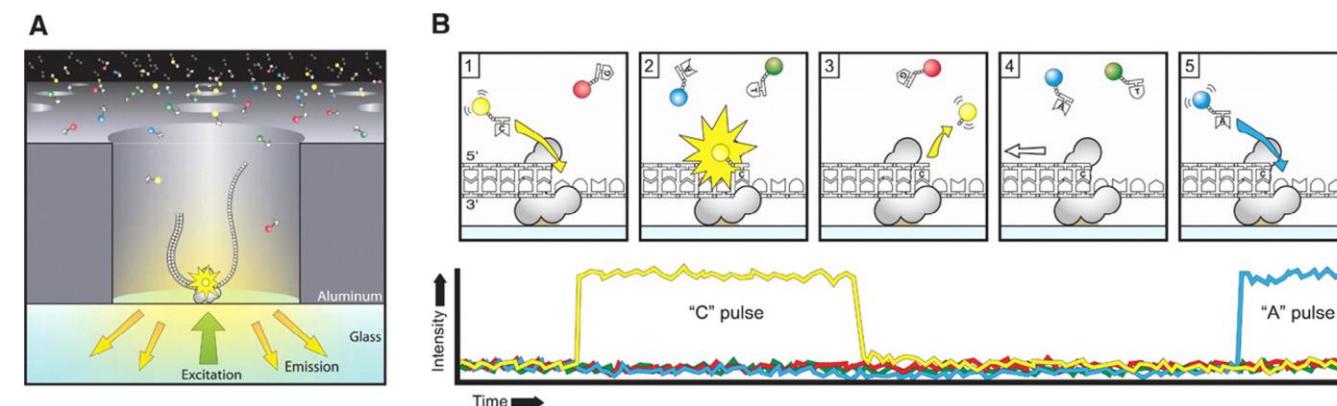
SMRTbell template: is a closed, single-stranded circular DNA that is created by ligating hairpin adaptors to both ends of a target dsDNA

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



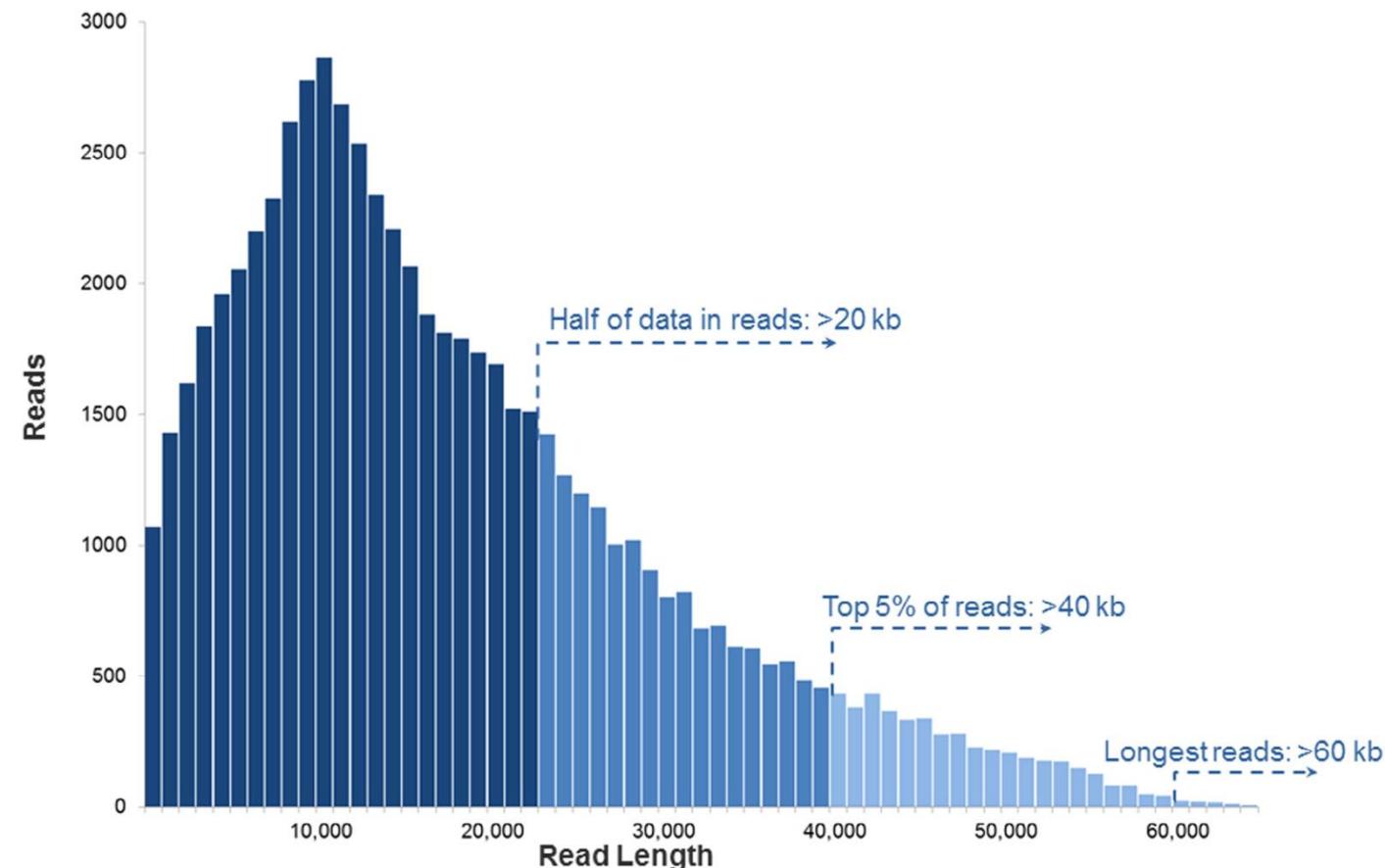
A single SMRT cell: this contains 150000 ZMWs (zero-mode waveguide). A SMRTbell diffuses into a ZMW. Approx 35000 -75000 ZMWs produce a read in a run lasting 0,5-4h resulting in 0,5-1Gb.



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	12	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

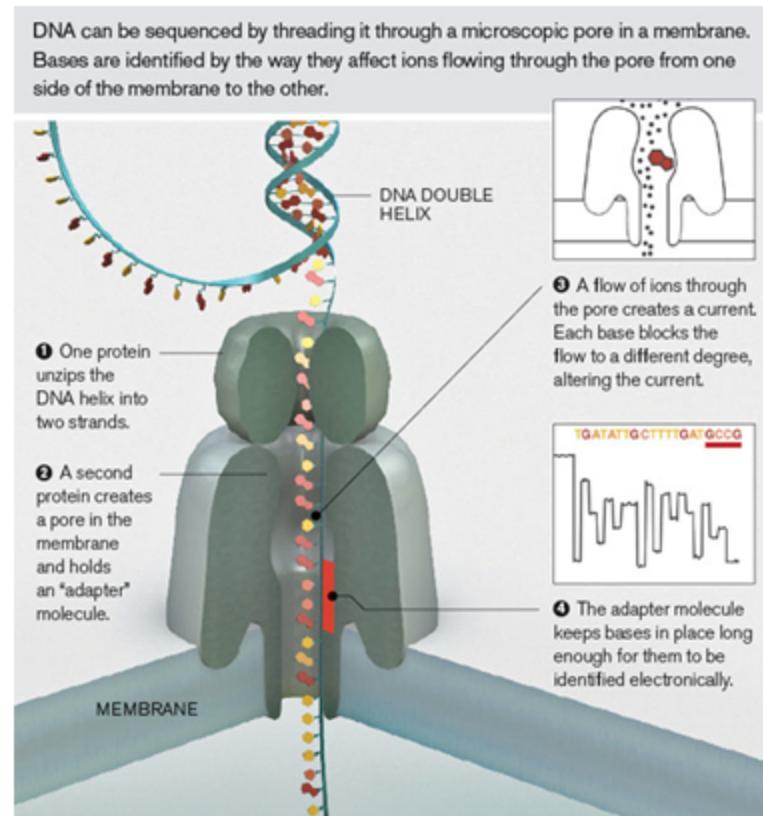
Produced highly-contiguous assemblies of bacterial and eukaryotic genomes

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>

Oxford Nanopore Technologies, MinION



The MinION is a portable sequencer; flow cells contain up to 512 nanopore sensors.

The Oxford Nanopore system processes the reads that are presented to it rather than generating read lengths. Sample-prep dependent, the longest read reported by a MinION user to date is >1 Mb.

Long reads confer many advantages, including simpler assembly and in the analysis of repetitive regions, phasing or CNVs.



For MinION / GridION
Flongle

Adapter to enable small, rapid nanopore sequencing tests, for mobile or desktop sequencers



MinION Mk1B

Your personal nanopore sequencer, putting you in control



MinION Mk1C

Your personal nanopore sequencer including compute and screen, putting you in control



GridION Mk1

Higher-throughput, on demand nanopore sequencing at the desktop, for you or as a service



PromethION 24/48

Ultra-high throughput, on-demand nanopore sequencing, for you or as a service

Oxford Nanopore Technologies

<https://nanoporetech.com/products/comparison>

	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: \$7 - 3.4 - 2.3 \$625 per flow cell: \$6 - 3 - 2

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

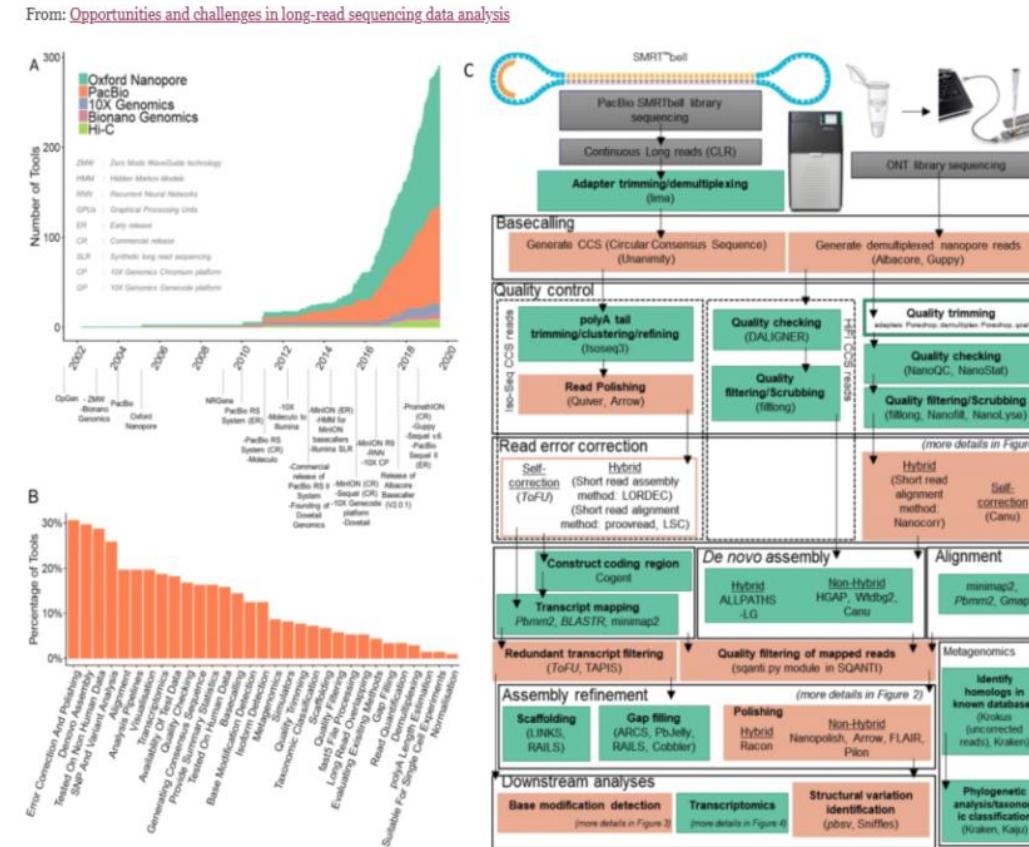


Oxford Nanopore has now started developing an even smaller device, SmidgION.

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 data analysis

Amarasinghe et al., Genome Biology 2020, 21:30



Overview of long-read analysis tools and pipelines. **a** Release of tools identified from various sources and milestones of long-read sequencing. **b** Functional categories. **c** Typical long-read analysis pipelines for SMRT and nanopore data. Six main stages are identified through the presented workflow (i.e. basecalling, quality control, read error correction, assembly/alignment, assembly refinement, and downstream analyses). The green-coloured boxes represent processes common to both short-read and long-read analyses. The orange-coloured boxes represent the processes unique to long-read analyses. Unfilled boxes represent optional steps. Commonly used tools for each step in long-read analysis are within brackets. Italics signify tools developed by either PacBio or ONT companies, and non-italics signify tools developed by external parties. Arrows represent the direction of the workflow.

PacBio sequencing 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	$1.0\text{--}1.5 \times 10^4$ 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	$2\text{--}5 \times 10^3$ on average	38	$1.1\text{--}4.7 \times 10^4$	50 h	6.44–17.90	[22,23]

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>Ion Torrent</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.

Besser et al., Clin Micr Infect, 2018

Advantages & disadvantages for short vs. Long read sequencing

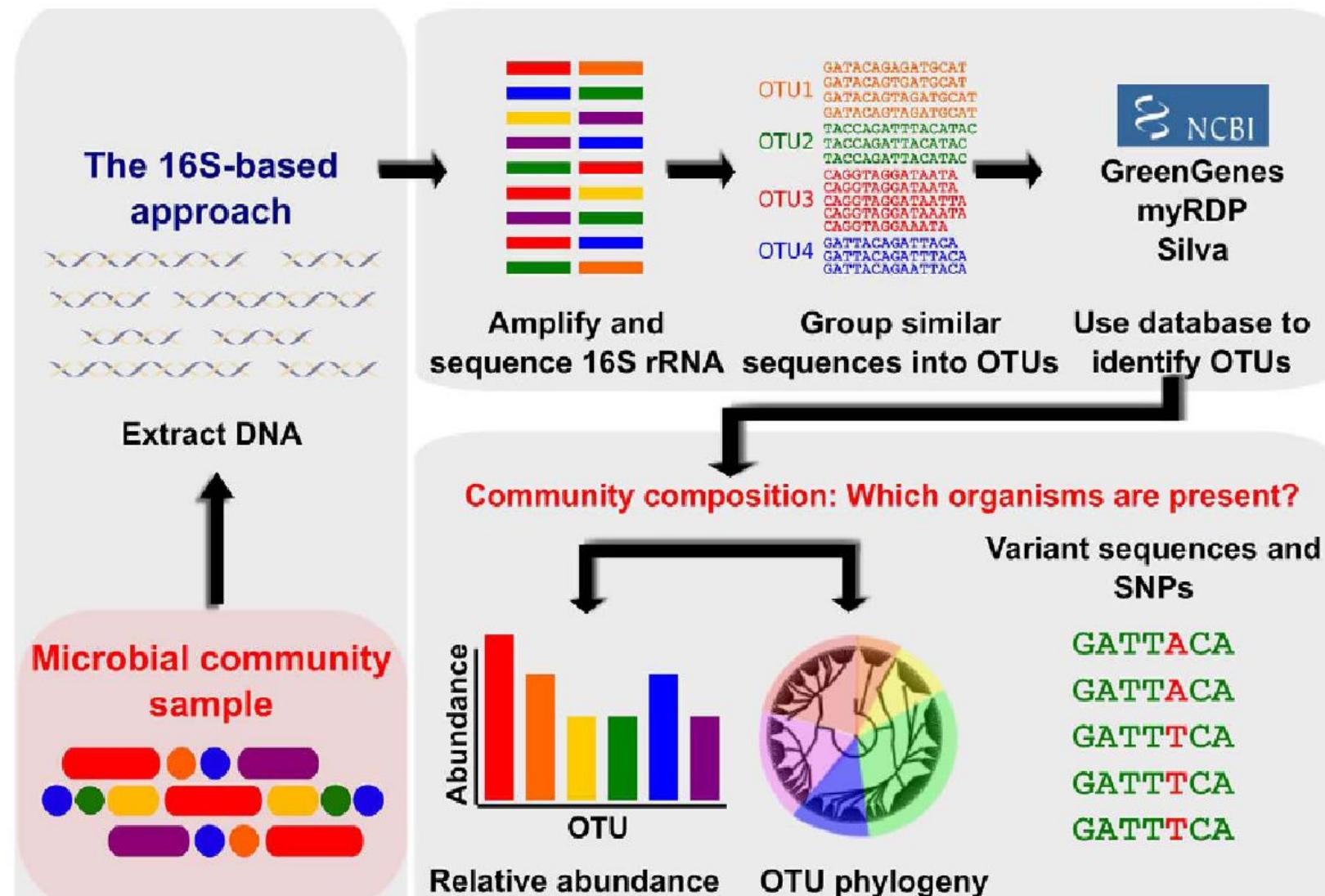
	Advantages	Limitations
Short-read sequencing	<ul style="list-style-type: none"> Higher sequence fidelity Cheap Can sequence fragmented DNA 	<ul style="list-style-type: none"> Not able to resolve structural variants, phasing alleles or distinguish highly homologous genomic regions Unable to provide coverage of some repetitive regions
Long-read sequencing	<ul style="list-style-type: none"> 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 <i>de novo</i> genome assembly 	<ul style="list-style-type: none"> Lower per read accuracy Bioinformatic challenges, caused by coverage biases, high error rates in base allocation, scalability and limited availability of appropriate pipelines

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

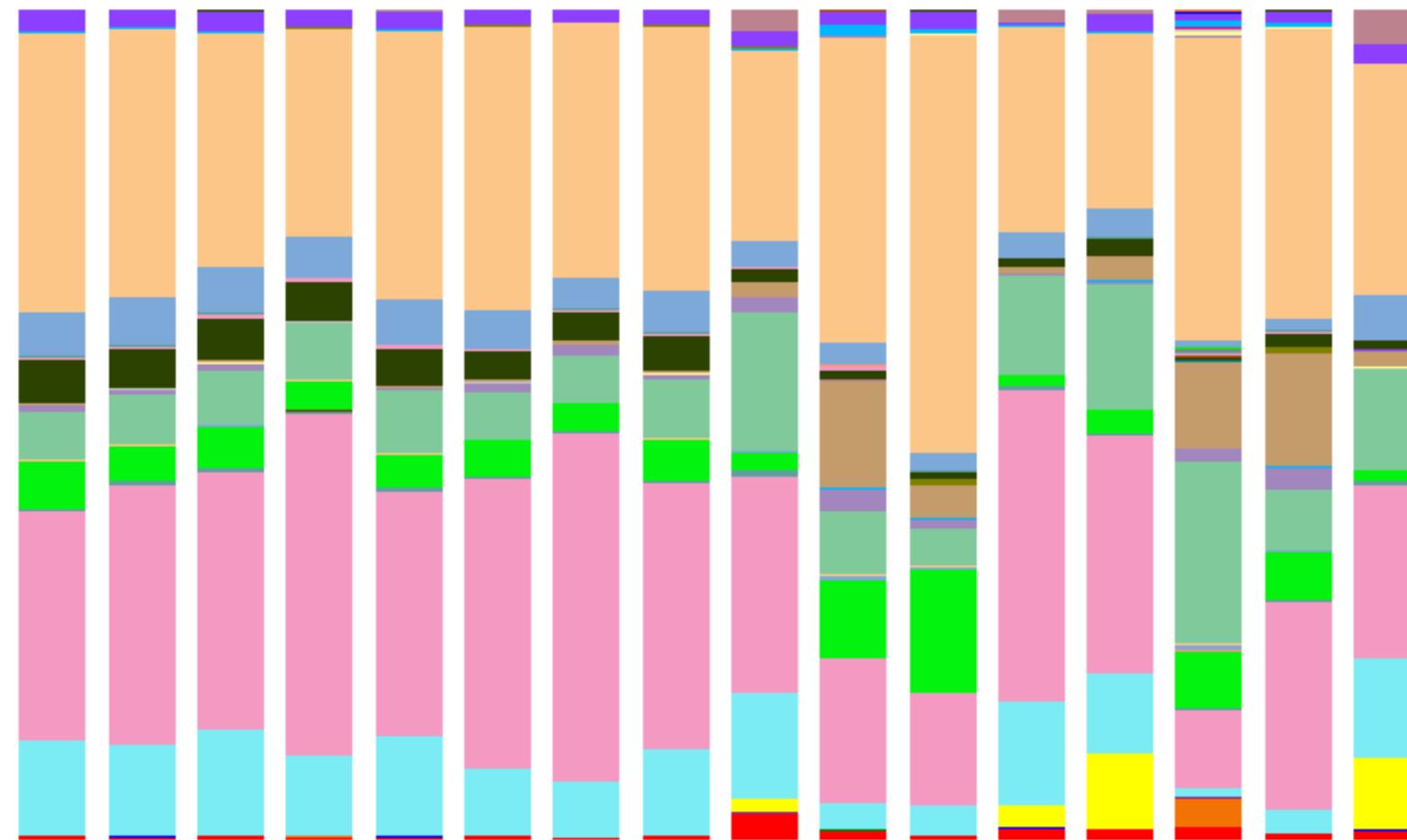
Targeted Metagenomics vs Metagenomics (16S vs Shotgun)

	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



Taxonomy summary (i.e. phylum level)

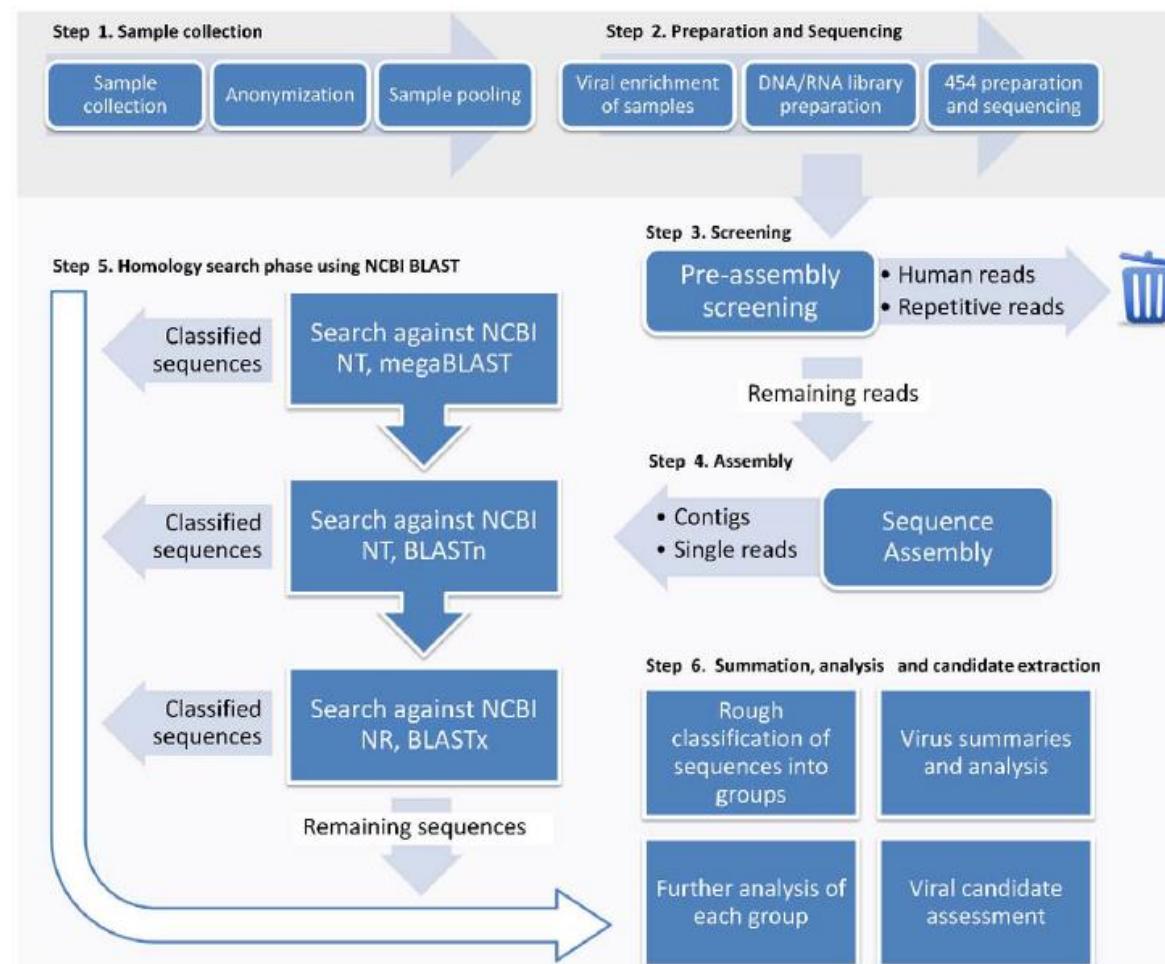


Targeted Metagenomics

Problemas:

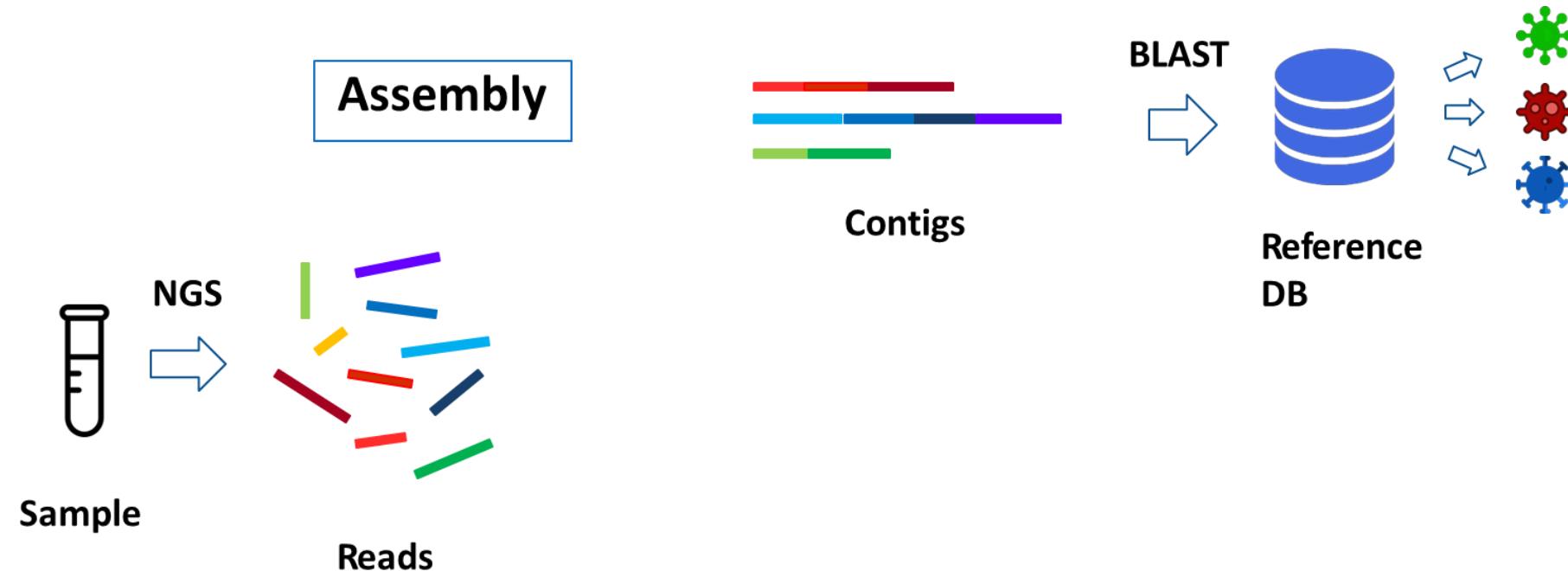
- Raros en el genoma (< 0.1%)
- Los trozos similares dificultan el ensamblado correcto de lecturas pequeñas
- No todos los rRNA se amplifican en la misma medida con los *primers* universales
- Especies con diversas copias de sus genes rRNA
- **No se conoce un umbral fijo de similitud que separe especies**
- **Tendencia a producirse quimeras en la PCR**

Metagenomics

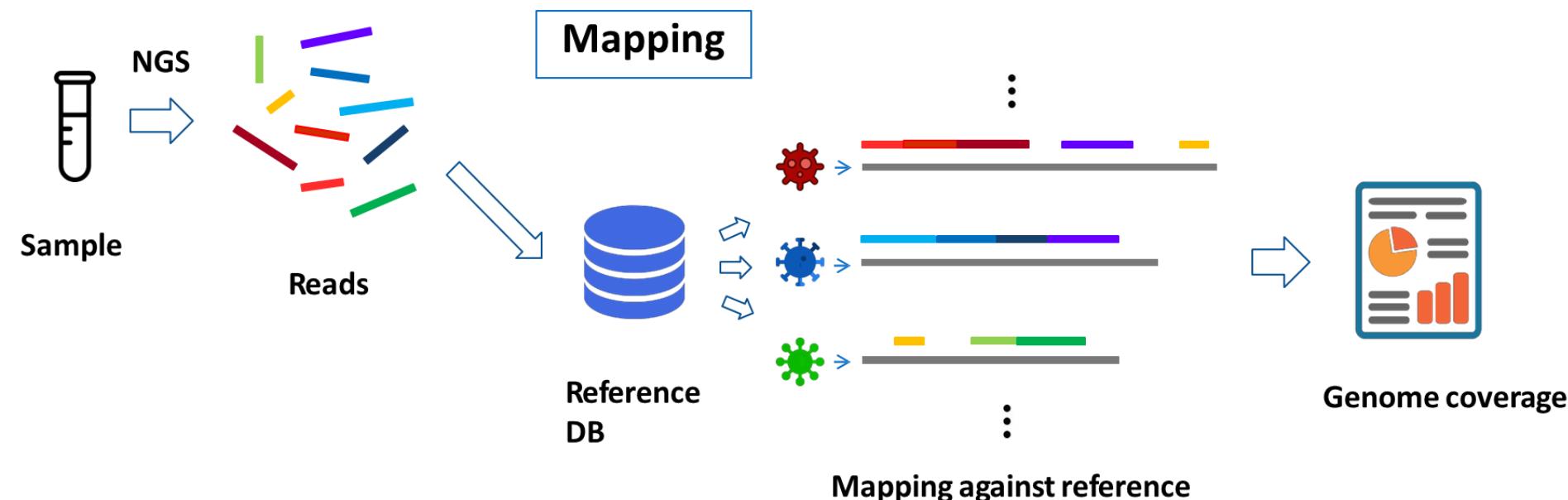


Lysholm et al., Plos One 2012:7,2, e30875

Metagenomic analysis approaches



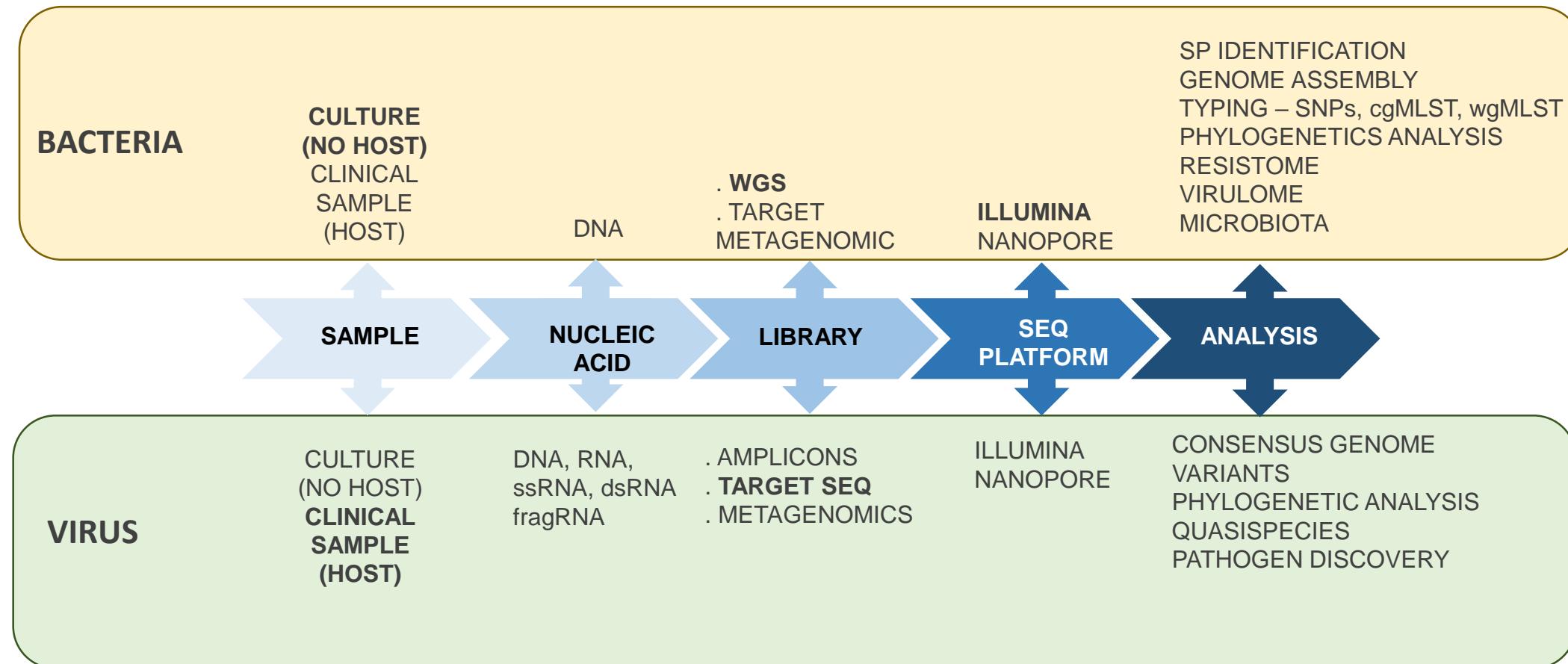
Metagenomic analysis approaches



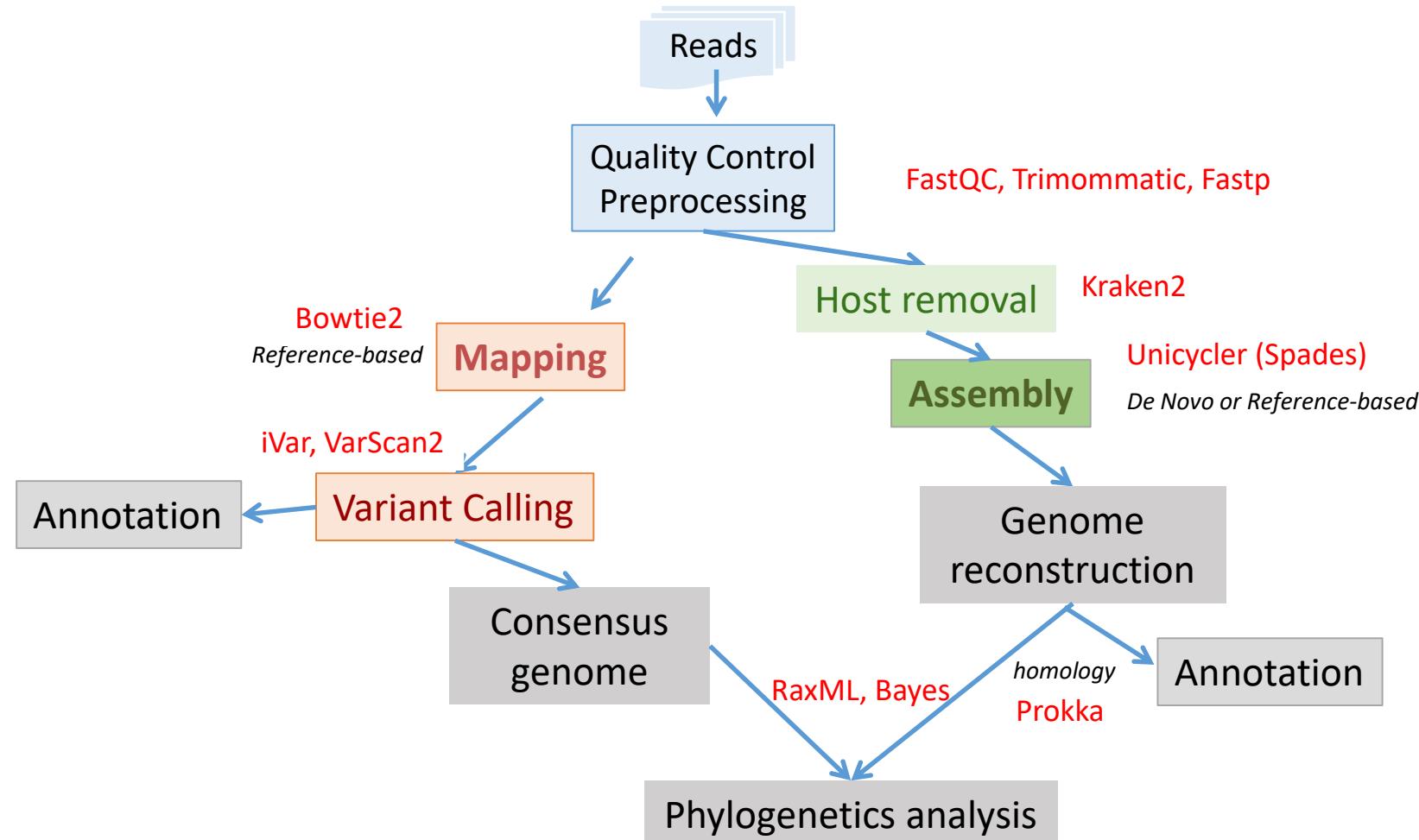
Metataxonomics vs Metagenomics (16S vs Shotgun)

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

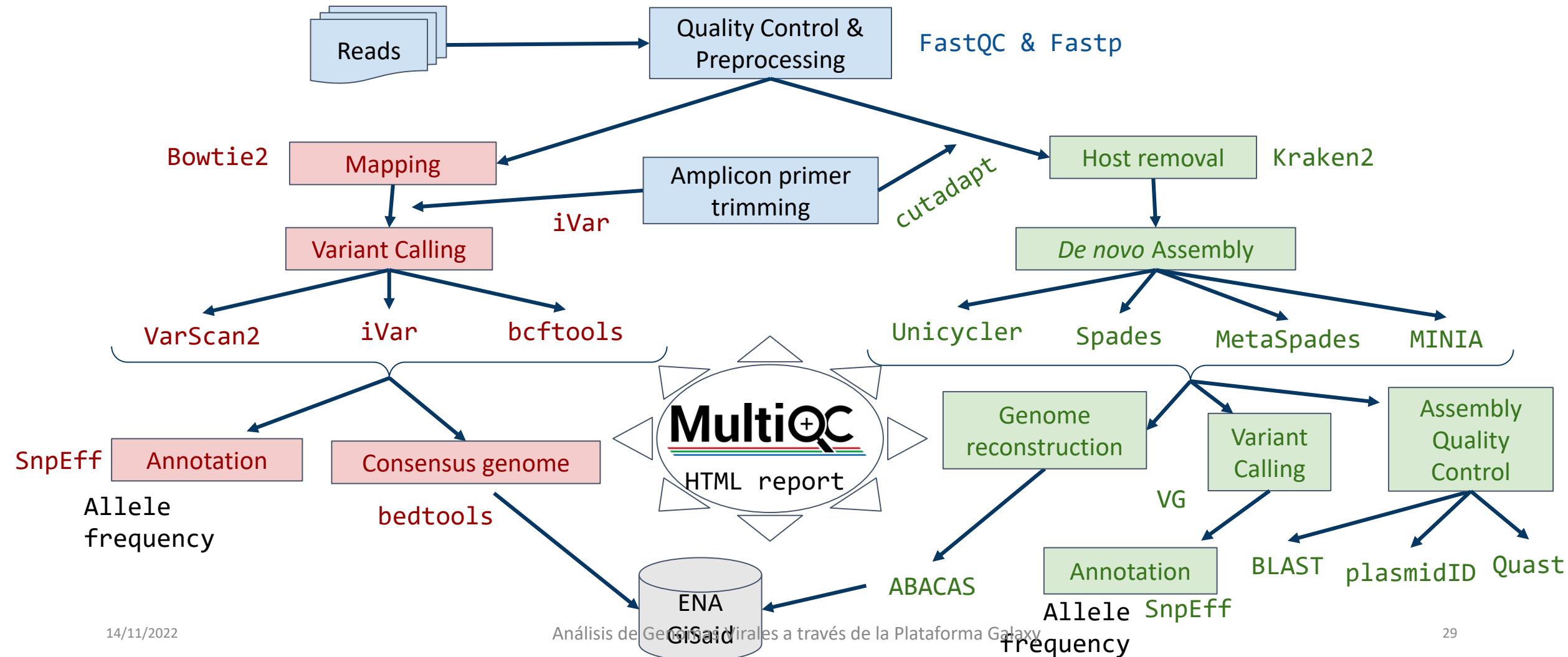
Bacterial and Viral Genome Sequencing



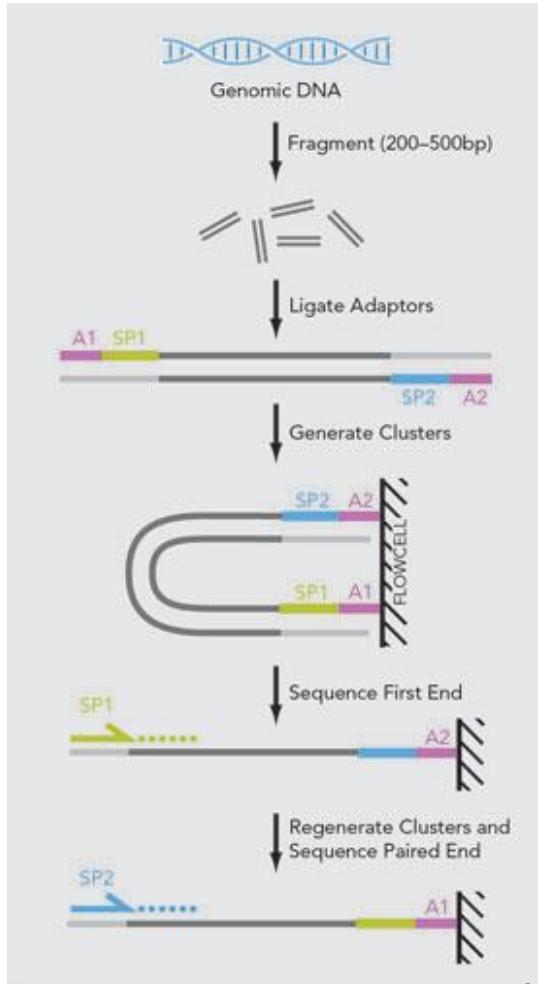
DATA ANALYSIS: Workflow example for virus genome analysis



Viralrecon



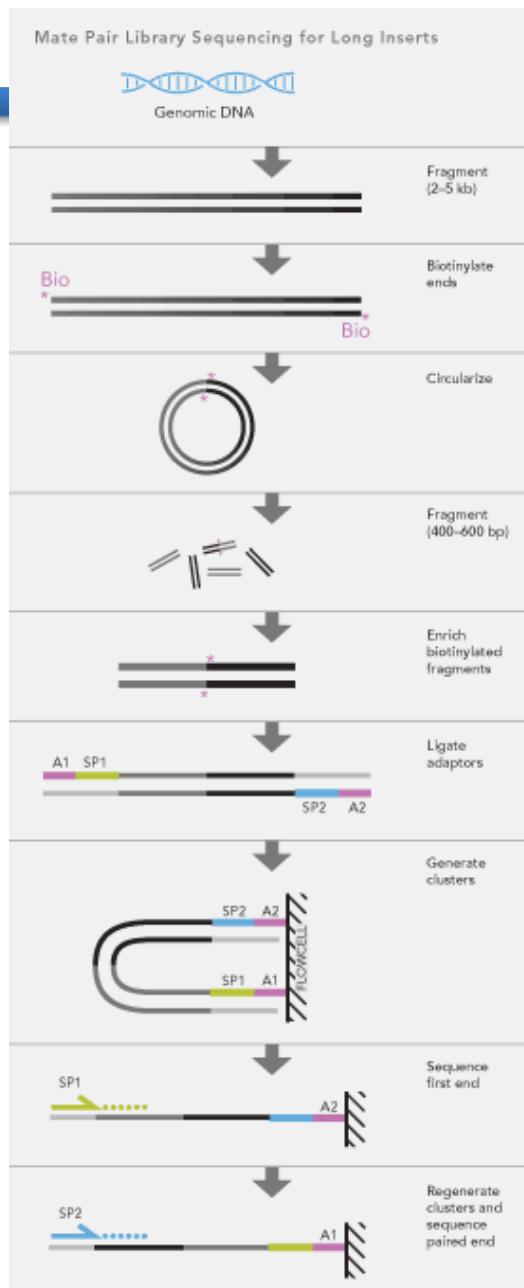
Que es Pair-end?



Secuenciación de un fragmento (bp)

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

Que es Mate-pair?



Mate Pair library preparation is designed to generate short fragments that consist of two segments that originally had a separation of several kilobases in the genome. Fragments of sample genomic DNA are end-biotinylated to tag the eventual mate pair segments. Self-circularization and refragmentation of these large fragments generates a population of small fragments, some of which contain both mate pair segments with no intervening sequence. These Mate Pair fragments are enriched using their biotin tag. Mate Pairs are sequenced using a similar two-adapter strategy as described for paired-end sequencing.

Secuenciación de dos fragmentos separados kb.

Util:
Secuenciación de un Genoma de novo
Finalizar un genoma
Detección de variantes estructurales

Análisis de Genomas Virales a través de la Plataforma Galaxy

> BU-ISCIII

Sequencing terms

Depth of coverage

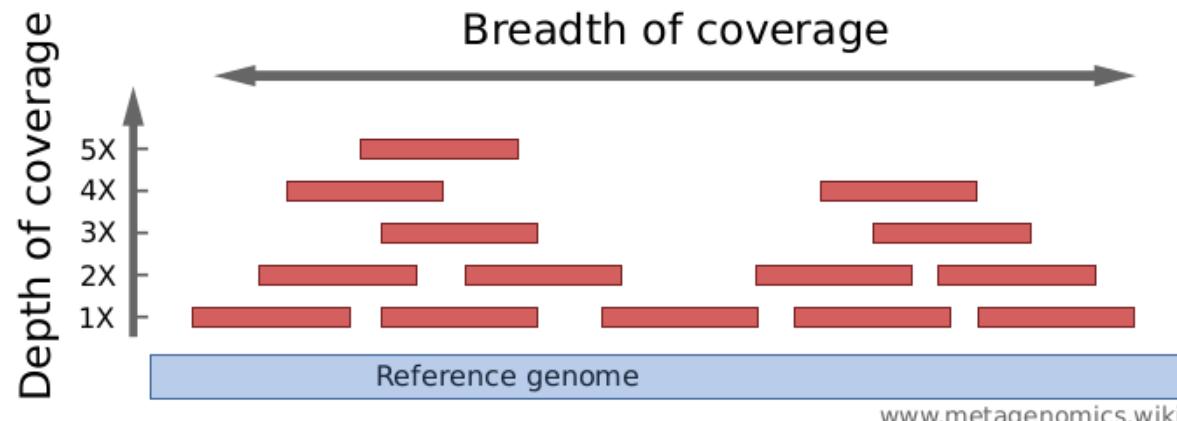
How strong is a genome "covered" by sequenced fragments (short reads)?

Per-base coverage is the average number of times a base of a genome is sequenced. The coverage depth of a genome is calculated as the number of bases of all short reads that match a genome divided by the length of this genome. It is often expressed as 1X, 2X, 3X,... (1, 2, or, 3 times coverage).

Breadth of coverage

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

Breadth of coverage is the percentage of bases of a reference genome that are covered with a certain depth. For example: 90% of a genome is covered at 1X depth; and still 70% is covered at 5X depth.



Calculo de cobertura: número de lecturas

Estimating Sequencing Runs

Coverage Equation

The Lander/Waterman equation is a method for computing coverage¹.

The general equation is:

$$C = LN / G$$

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

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

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

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

Sequencing Coverage Calculator

Support Center: Sequencing Coverage Calculator

Application or product: Whole-Genome Sequencing

Coverage: 100 x

Duplicates: 2 %

Genome or region size (in million bases): 3300 Mb

Total read length (e.g. 200 for 2x100): 600 cycles

Benchtop Sequencers Production-Scale Sequencers

iSeq NextSeq 500/550

MiSeq NovaSeq 6000

MiSeq / MiSeq Dx in RUO mode HiSeq 3000/4000

NextSeq 500/550 HiSeq 1500/2500 Rapid Run

 HiSeq 1500/2500 High Output

 NextSeq 1000 Sequencing System

 NextSeq 2000 Sequencing System

Support Center: Sequencing Coverage Calculator

Thank you for using the Illumina coverage estimator.

The results were calculated based on: **coverage needed**. Explain the estimations

Application or product:	Whole-Genome Sequencing			
Genome or region size:	3300 Mbases			
Read length:	600			
Coverage:	100x			
Duplicates:	2%			
Output Required:	336,734,693,878 bases			
	MiSeq	MiSeq	MiSeq	MiSeq
Run type	v3 Reagents	v2 Reagents	v2 Nano Reagents	v2 Micro Reagents
Clusters	25,000,000 per flow cell	15,000,000 per flow cell	1,000,000 per flow cell	4,000,000 per flow cell
Output per unit (flow cell or lane)	15,000,000,000 per flow cell	9,000,000,000 per flow cell	600,000,000 per flow cell	2,400,000,000 per flow cell
Exceeds maximum read length?	Does not exceed maximum (2x300)	Read length exceeds maximum of 2x250	Read length exceeds maximum of 2x250	Read length exceeds maximum of 2x150
Number of units per sample (flow cell or lane)	22,449 flow cells	37,415 flow cells	561,224 flow cells	140,306 flow cells
Samples per unit (flow cell or lane)	-0/flow cell	-0/flow cell	-0/flow cell	-0/flow cell
Comments	Upgraded software: MCS v2.3 or later; MiSeq	Upgraded hardware or from September 2012 and later: MCS v2.0 or later; Reagent Kit v3 (150/600)	Upgraded hardware or from September 2012 and later: MCS v2.0 or later; MiSeq Reagent Kit v2 (50/300/500)	Upgraded hardware or from September 2012 and later: MCS v2.0 or later; MiSeq Reagent Nano Kit v2 (300/500)
Products	MiSeq Reagent Kit v3	MiSeq Reagent Kits v2	MiSeq Reagent Kits v2	MiSeq Reagent Kits v2

Get the results in a comma-separated values (CSV) report:

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

Depth of coverage and genome coverage

Depth of coverage

the sequencing coverage =
$$\frac{\text{the number of total reads} \times \text{the read length}}{\text{the length of target sequence or genome}}$$

Genome coverage

% length sequence genome

Increase number of raw reads

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

Thanks for your attention!

Questions???