Iniciación al análisis de datos procedentes de técnicas de secuenciación masiva (NGS)

Unidad de Bioinformática (BU-ISCIII)
Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

22-26 Mayo 2023, 10ª Edición
Programa Formación Continua, ISCIII

OBJETIVOS DEL CURSO

- Aproximación a las técnicas de secuenciación masiva (NGS) y a sus aplicaciones
- Adquirir conocimientos básicos del entorno linux
- Familiarizarse con los formatos de ficheros generados en el análisis de datos procedentes de la SM
- Conocer el flujo del análisis de los datos procedentes de la SM

Sesión 1 - Secuenciación Masiva Plataformas de Secuenciación

Isabel Cuesta

Unidad de Bioinformática
Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

22-26 Mayo 2023, 10ª Edición
Programa Formación Continua, ISCIII

INDICE

- Unidad de Bioinformática
  Servicios ofertados

- Evolución de la secuenciación

- Plataformas de secuenciación masiva (NGS o HTS)

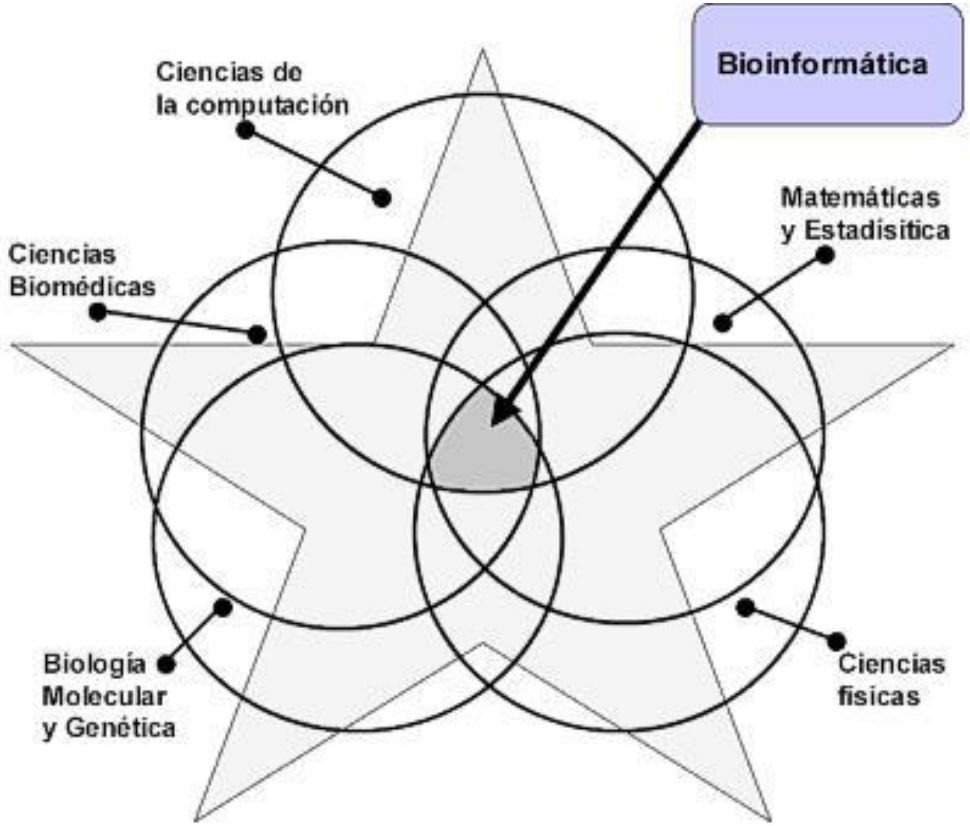
Qué es la Bioinformática?

PROBLEMAS BIOLÓGICOS

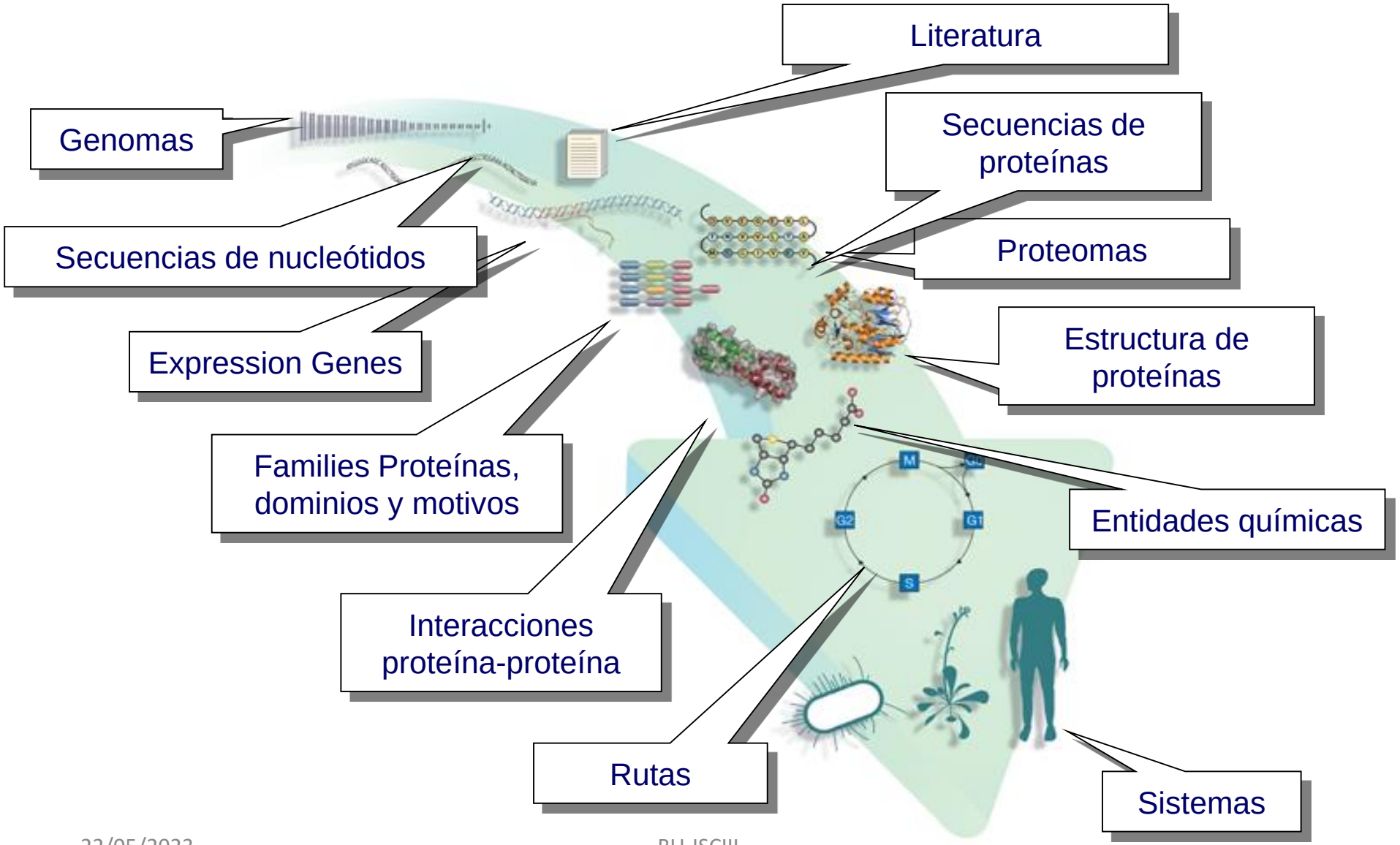
Procesamiento de datos

MÉTODOS COMPUTACIONALES

Bioinformática es multidisciplinar



Tipos de datos dan idea de la dimensión de la Bioinformática



Why BU-ISCIII was founded

Genomics Unit

2010

454

2013

NextSeq500 MiSeq

2019

MinIon - Nanopore

2021

NovaSeq 6000

Bioinformatics Unit

2012

Scientific & technical Support to Researchers on HTS Data Analysis

National Microbiology Centre (CNM)

Research Institute for Rare Diseases (IIER)

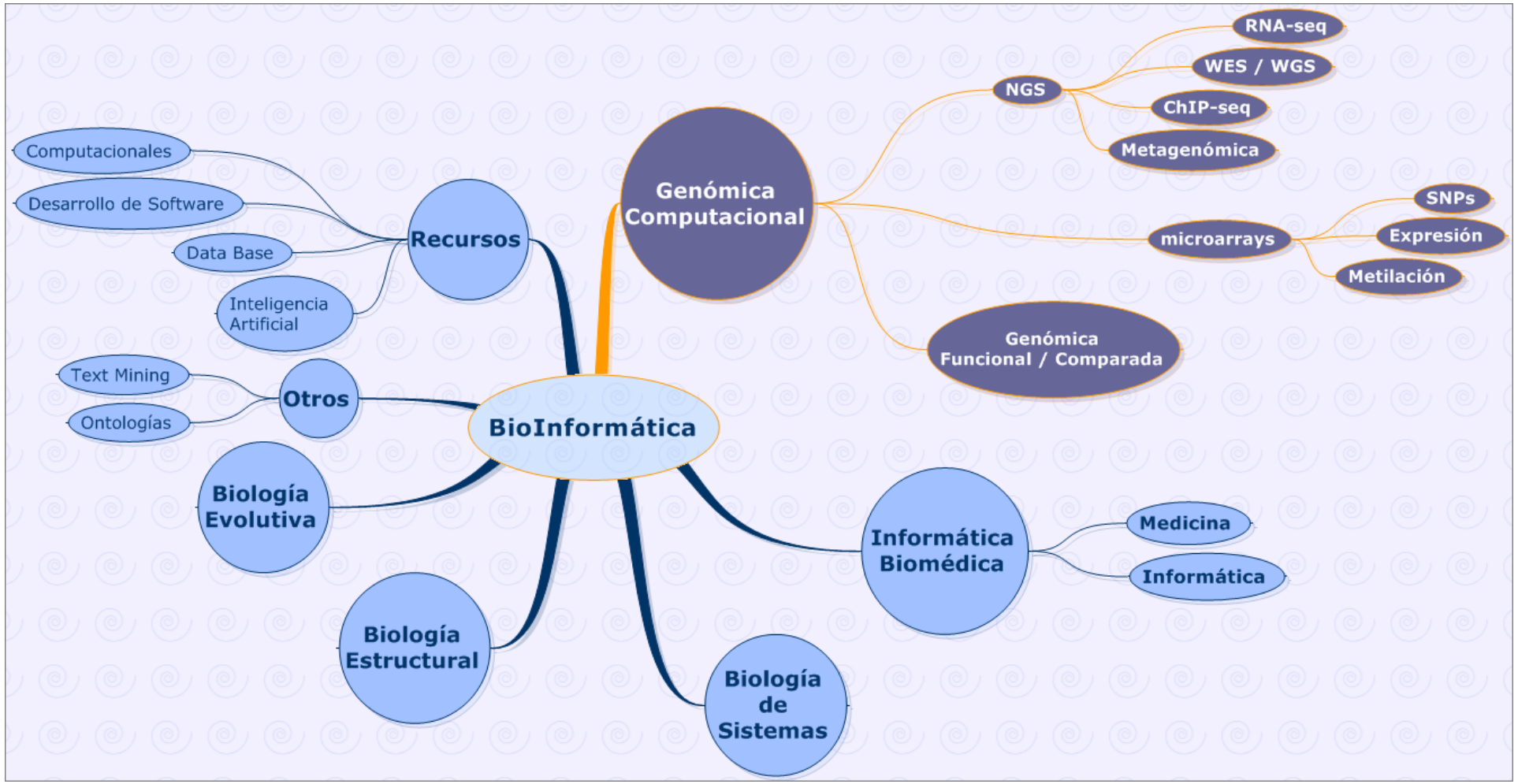
Functional Unit for Research in Chronic Disease

Network of Biological Alerts

National Centre of Tropical Medicine

National Environment Health Centre

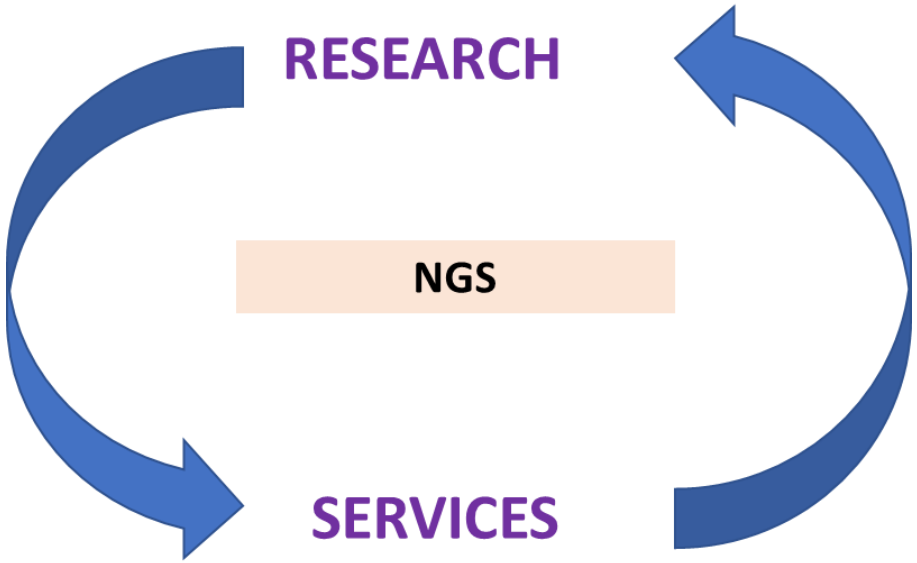
BU-ISCIII Mission - Activities



Bioinformatics Unit Activities

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

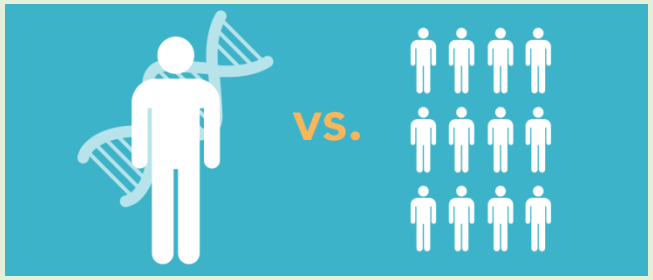
Provide scientific and technical solutions for using NGS in the diagnostic routine or research activity from different ISCIII labs

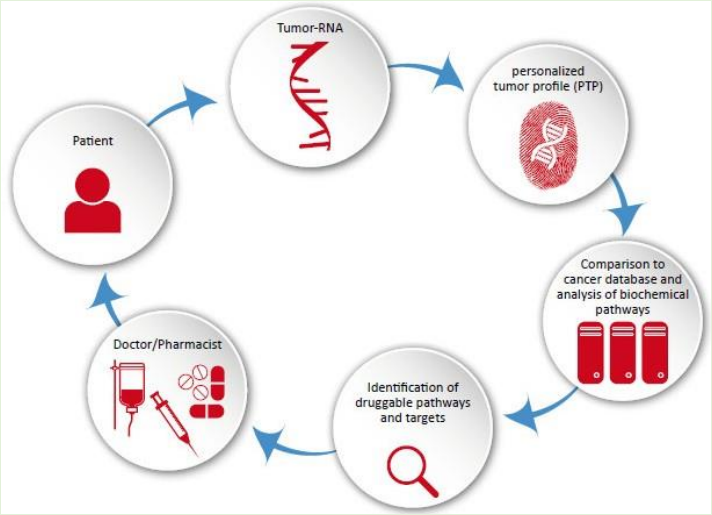


Clinical Bioinformatics - Precision Medicine

IIER

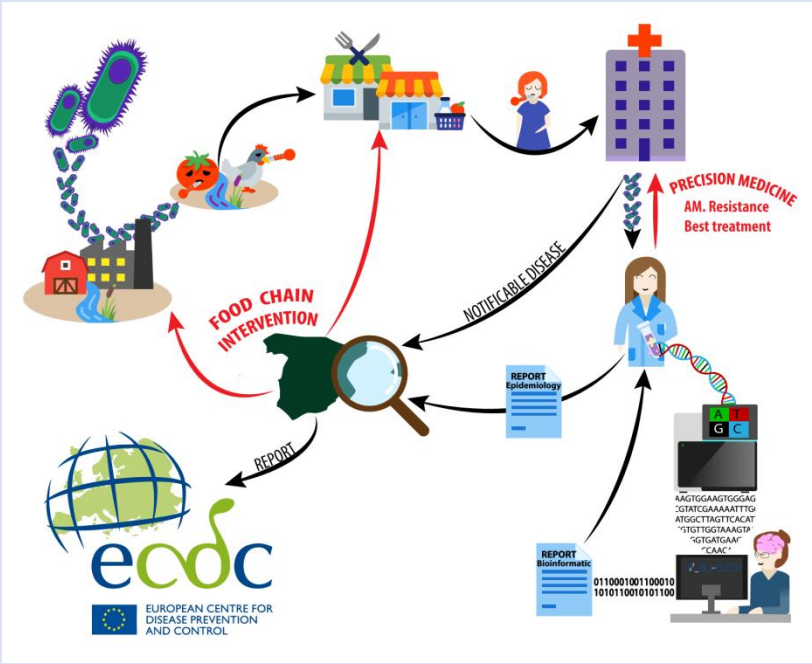
Tumour or Rare Diseases



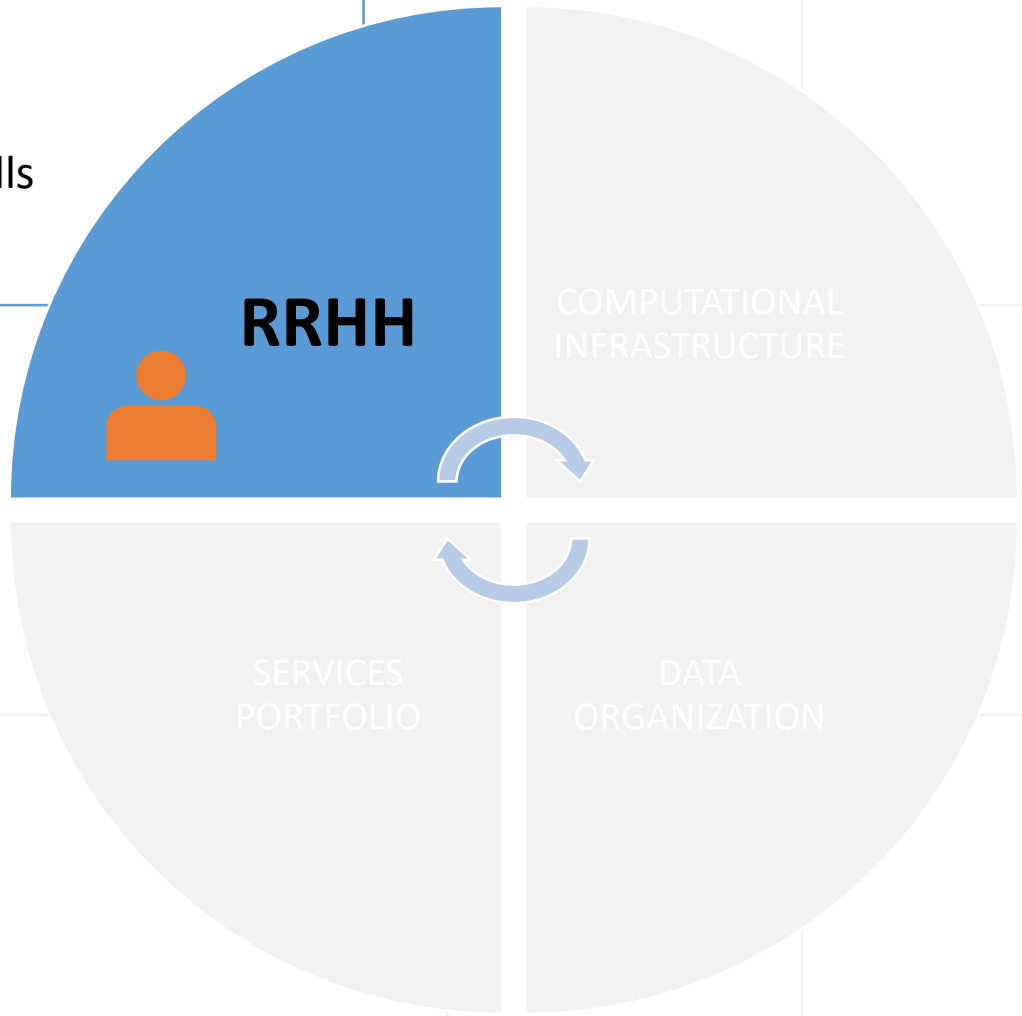


CNM

Pathogen associated



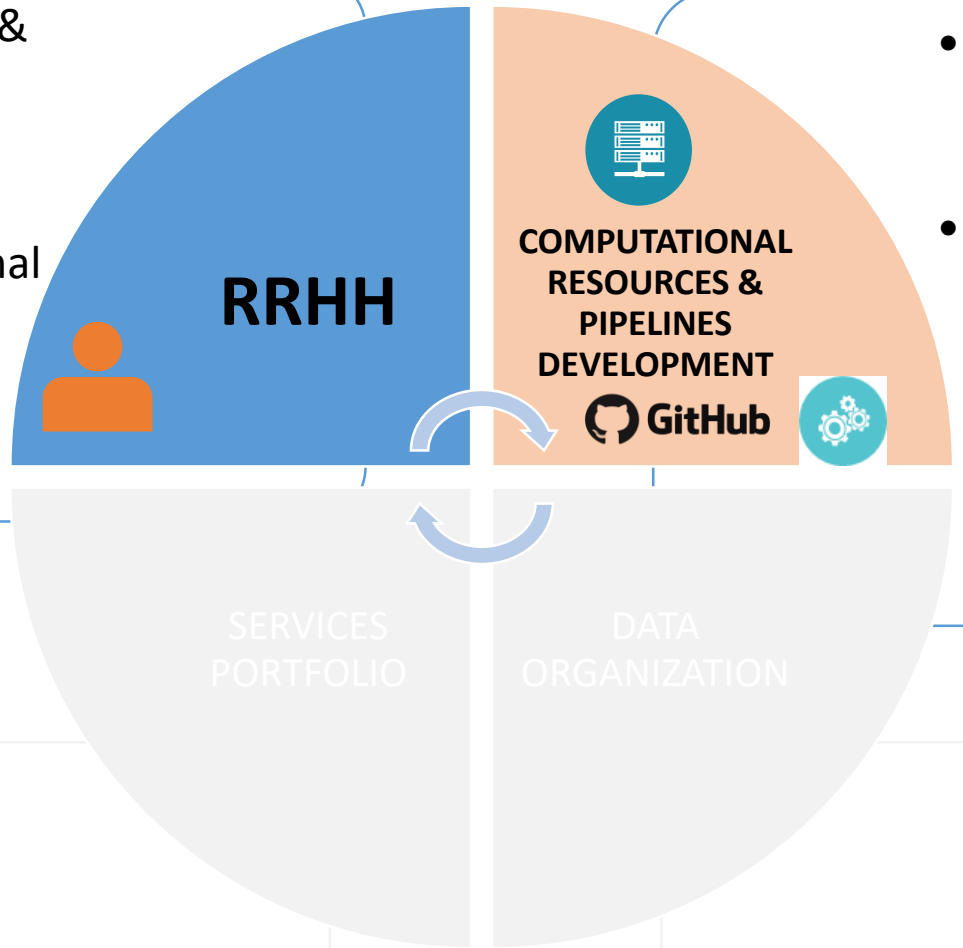
- TEAM: Bioinformatician & IT
- SOURCE OF FUNDING:
  - Research Project
  - National or International Calls
  - Permanent position – Civil Servant



Human resources

	Disciplina	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Isabel Cuesta	Dr. Biología Molecular	CIENTIFICO 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				ANTIBIOTICOSU.					ANTIBIOTICOSU.			
Sarai Varona	Bioquímica								PROYECTO (2024)				
Luis Chapado	Telecomunicaciones						COLABORADOR						
Erika Kvalem	Biotecnología										PROYECTO COVID19 1 año oct 2021 al 2022		
Alberto Lema	Biología												
Luis Aranda	FP Informática												
Guillermo Gorines	Biología										ESTANCIA		
Contratos (proyectos)													5
Personal (número)		2	4	4	5	5	4	6	6	6	8	8	3
	Master en Bioinformática	CNM		IIER		BU-ISCIII	BU-ISCIII		FUNCIONARIO				

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



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

Computational Resources

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

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

Data Centre (CPD-ISCIII)

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

- 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

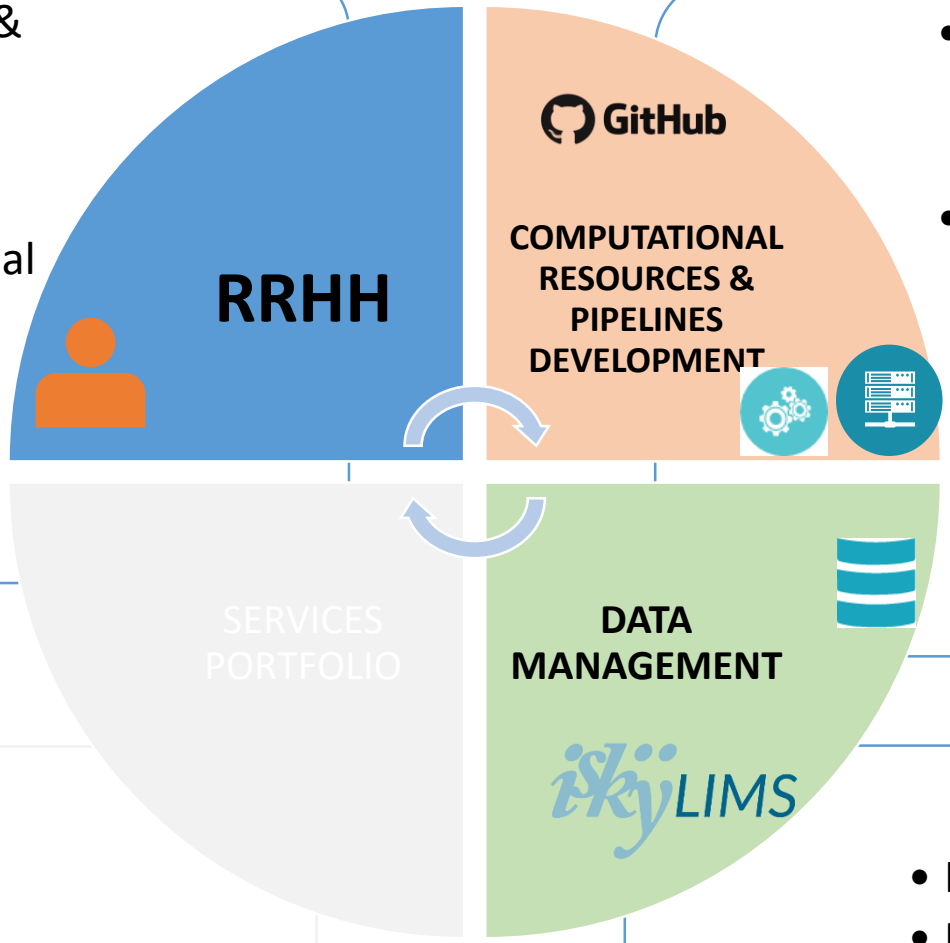
Lab ID	Allele based analysis					
	Approach	Allelic calling method	Assembler	Scheme	Difference within cluster	Difference outside cluster
EQA provider	BioNumerics	Assembly- and mapping-based	SPAdes	Applied Math (cgMLST/Pasteur)	0-3	24-1112
19	BioNumerics	Assembly- and mapping-based	SPAdes	Applied Math (cgMLST/Pasteur)	0-3	25-1120
35	SeqPhere	Assembly-based only	Velvet	Ruppitsch (cgMLST)	0-2	16-1065
70	SeqPhere	Assembly-based only	Velvet	Ruppitsch (cgMLST)	0-2	16-1062
105*	SeqPhere	Assembly-based only	SPAdes v 3.80	Ruppitsch (cgMLST)	0-1#	23-812
129	SeqPhere	Assembly-based only	Velvet			
135	SeqPhere	Assembly-based only	CLC Genomics Workbench 10			
141	SeqPhere	Assembly-based only	SPAdes 3.9.0			
142	Inhouse	Assembly-based only	SPAdes			
144	SeqPhere	Assembly-based only	Velvet			

Fifth external quality assessment scheme for Listeria monocytogenes typing

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	

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



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

- INFRASTRUCTURE
- DATA WORKFLOW
- LIMS

Infrastructure and data management

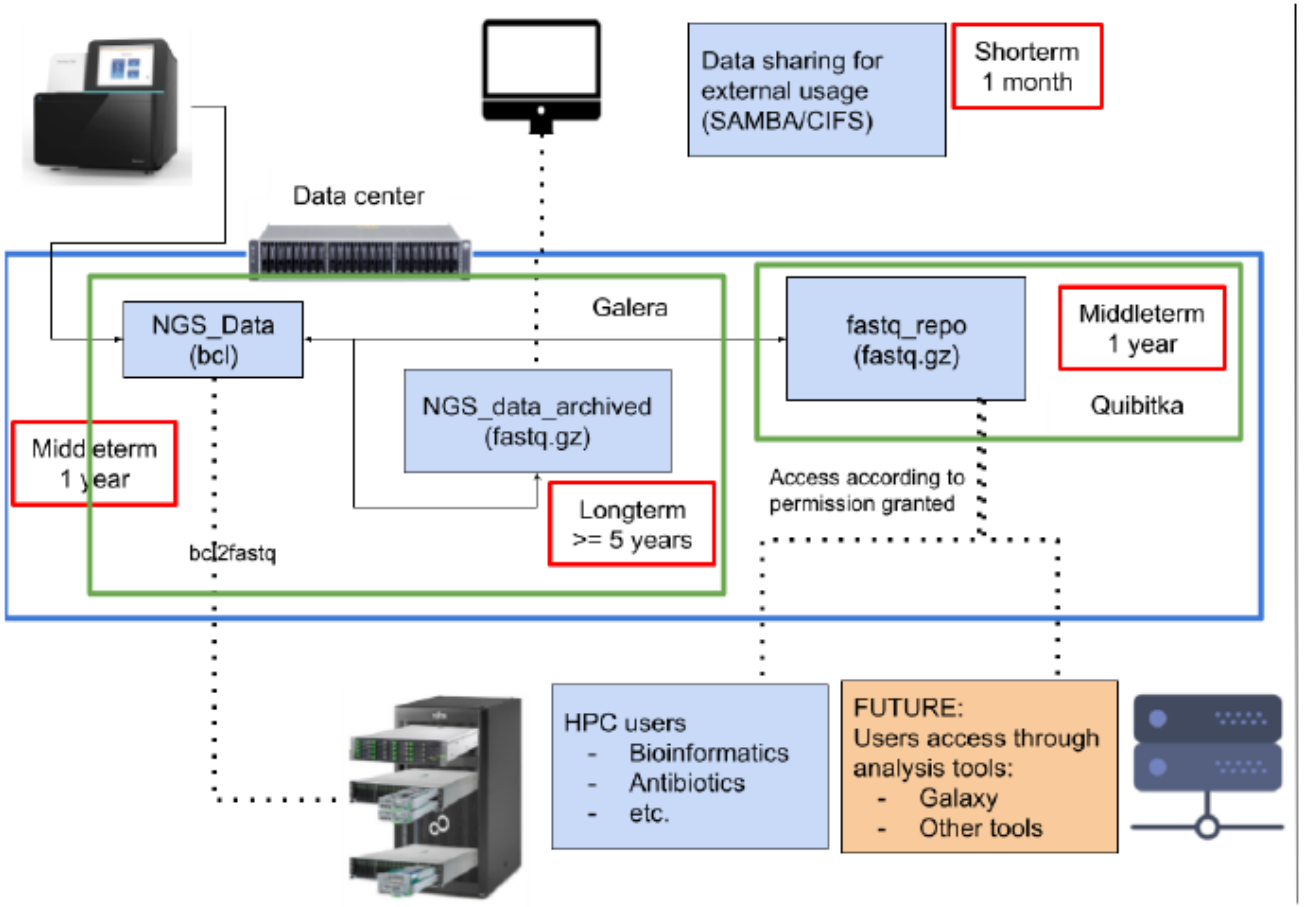
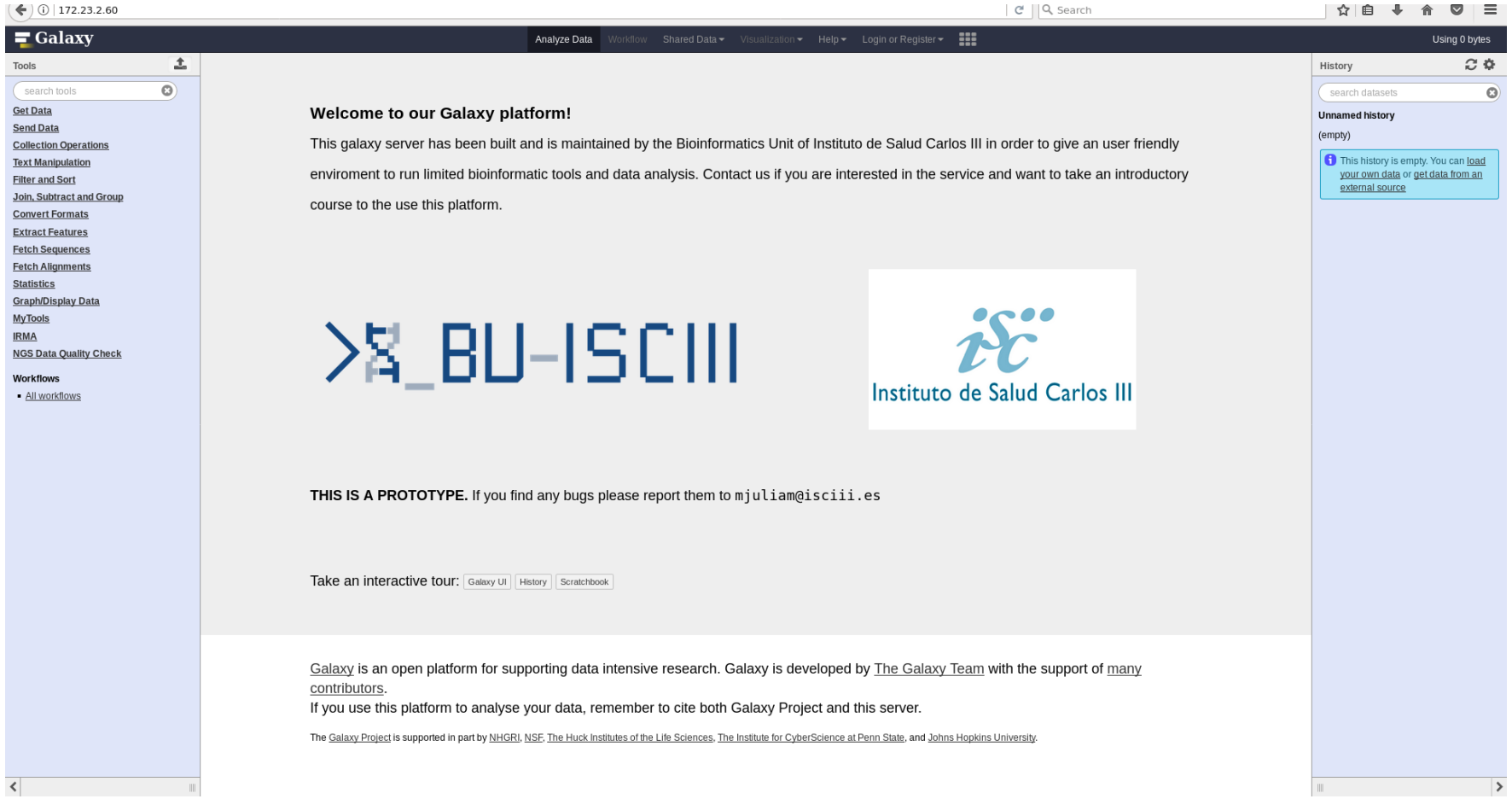


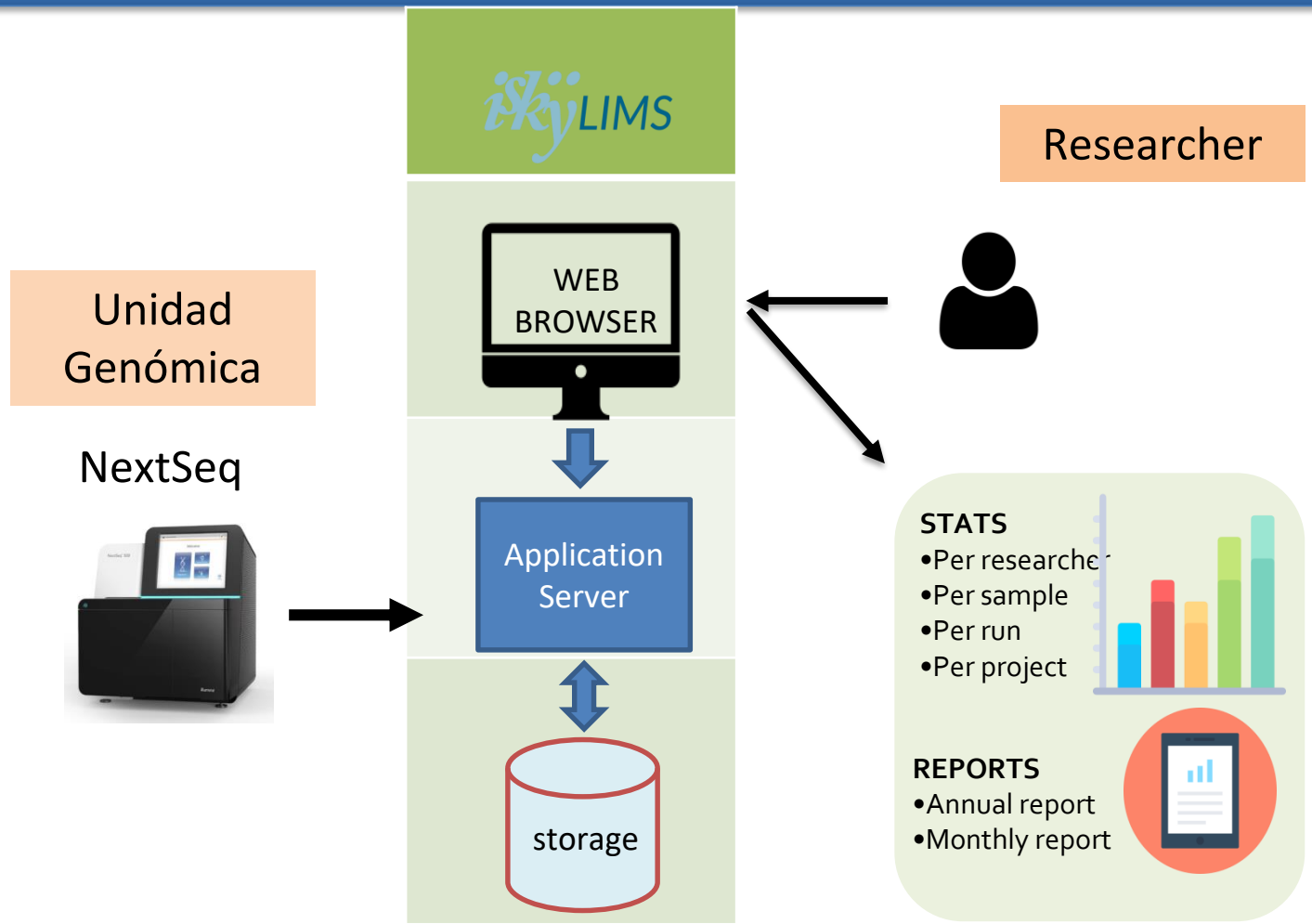
Fig. 3: *Esquema de organización datos genómicos*

Minimize I/O issues
Maximize storage uses

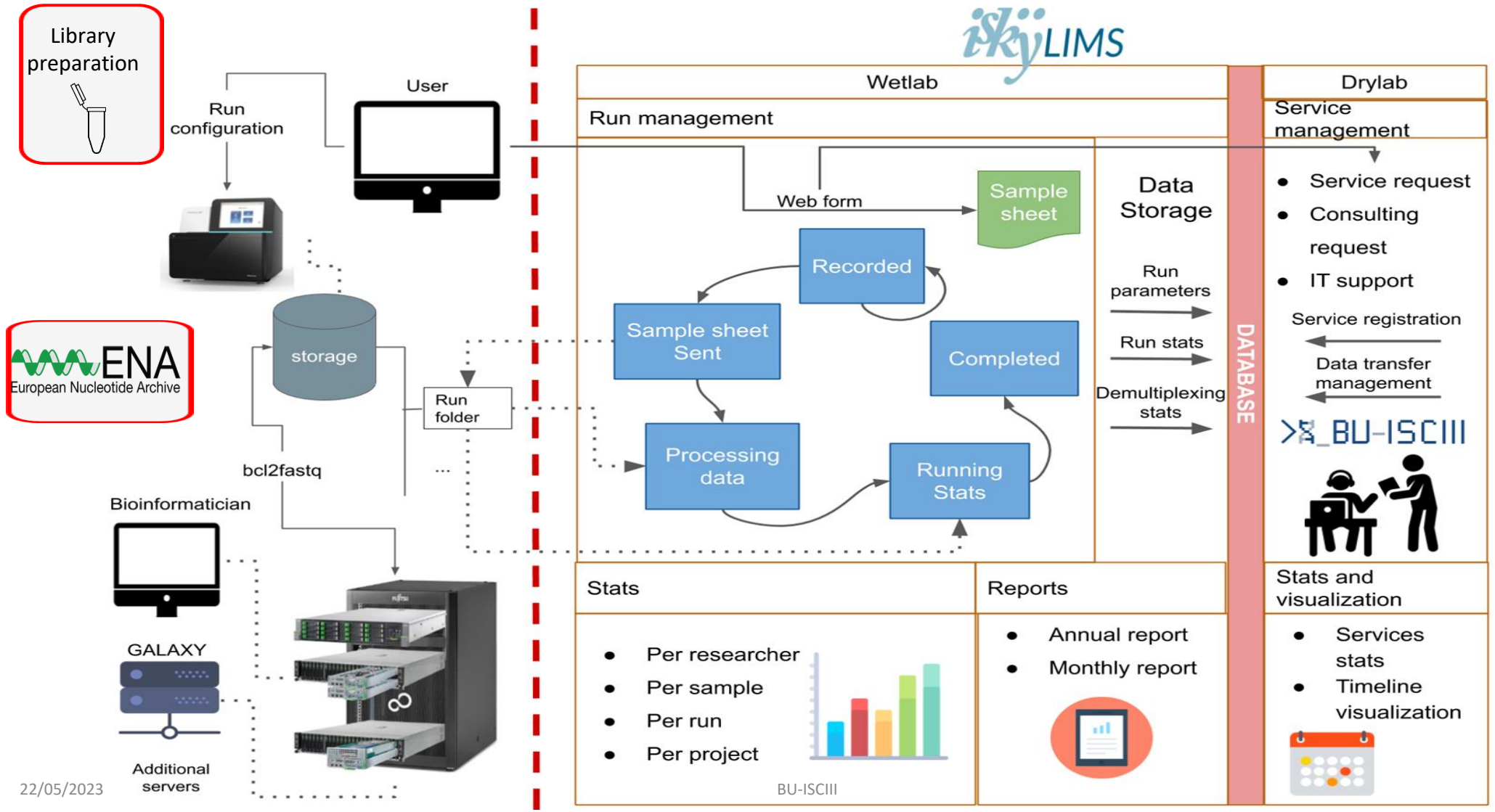
Galaxy – not maintained



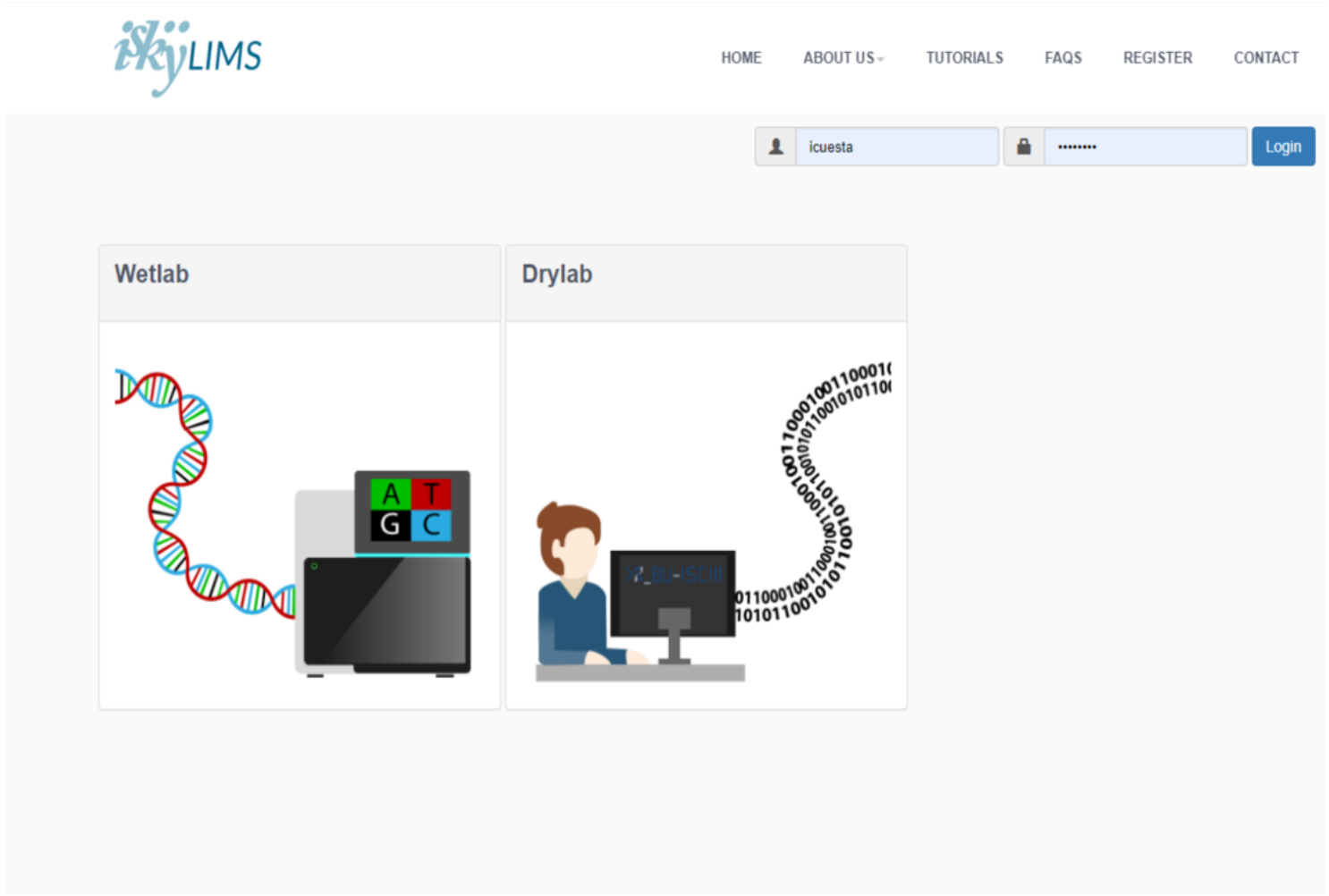
iSkyLIMS https://iskylims.isciii.es/



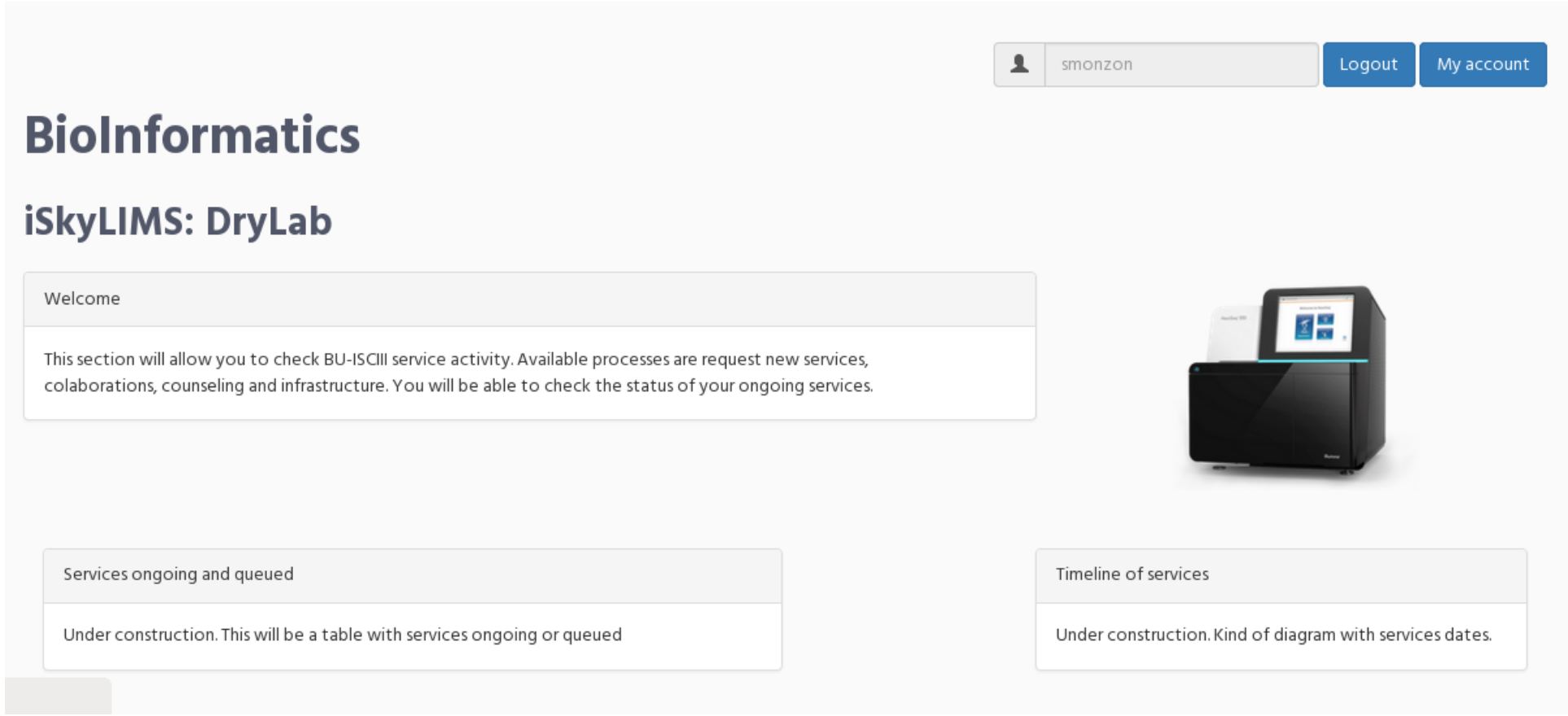
Infrastructure and data management: LIMS



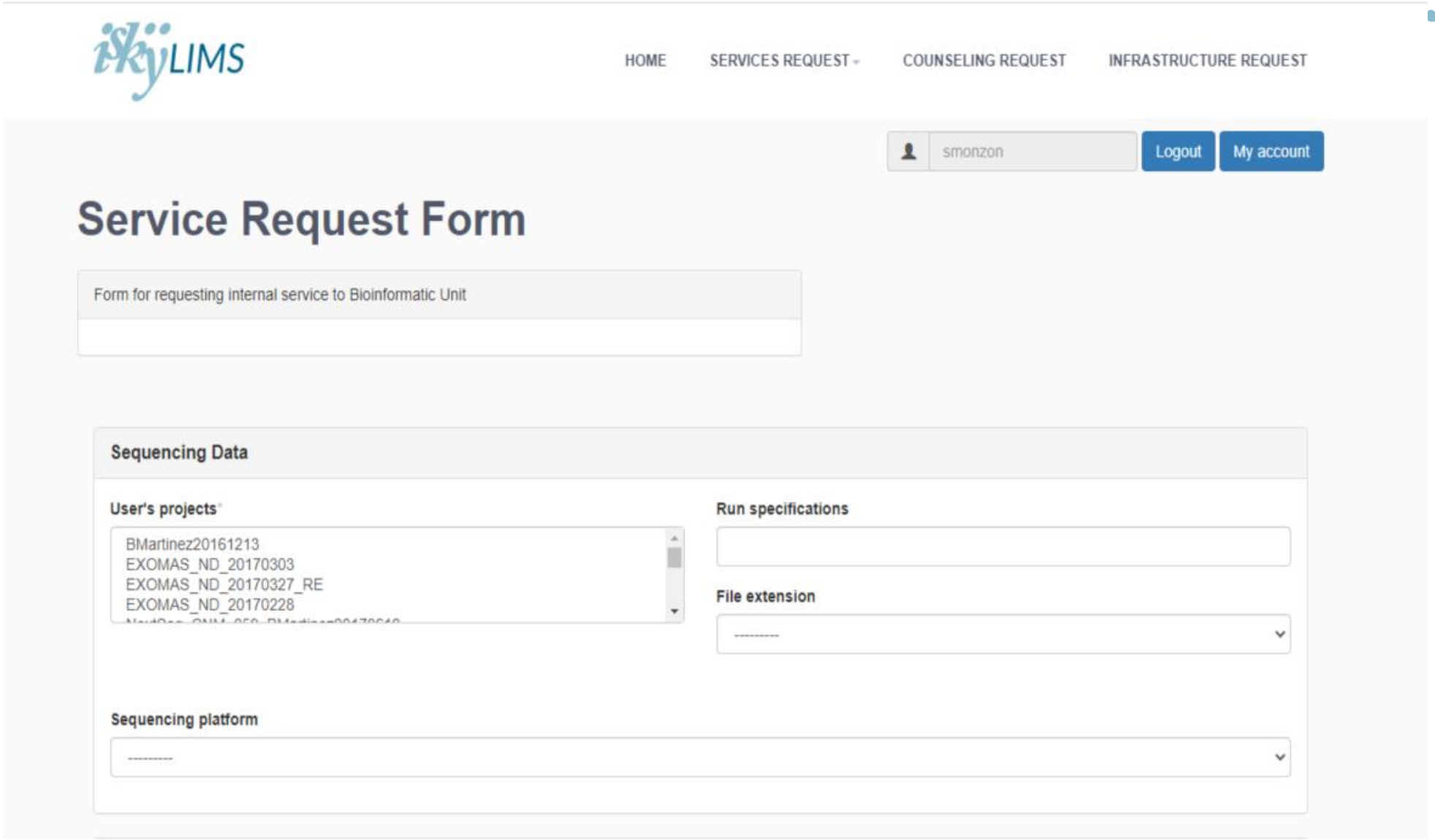
SOLICITUD DE SERVICIO https://iskylims.isciii.es/



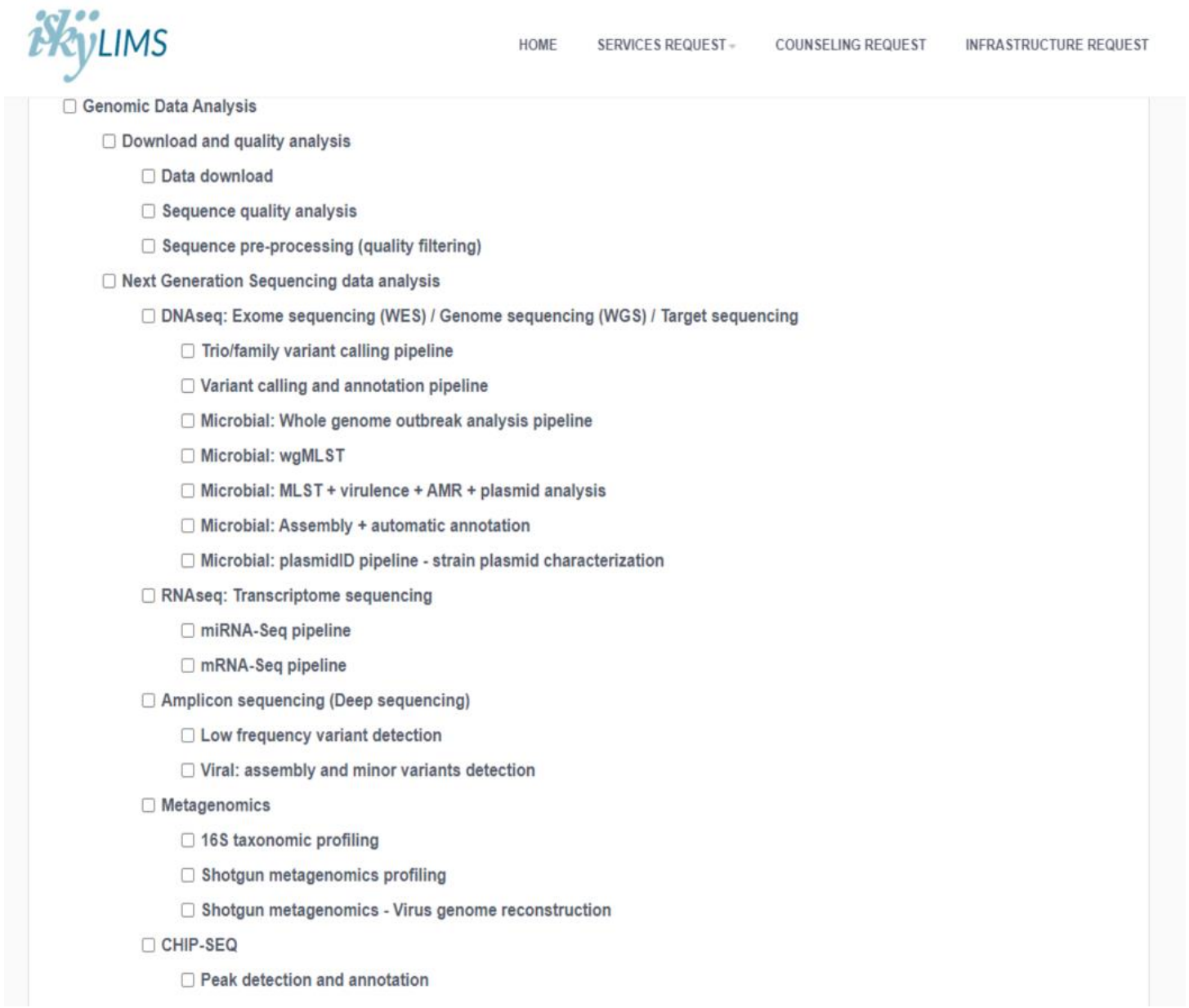
SOLICITUD DE SERVICIO https://iskylims.isciii.es/



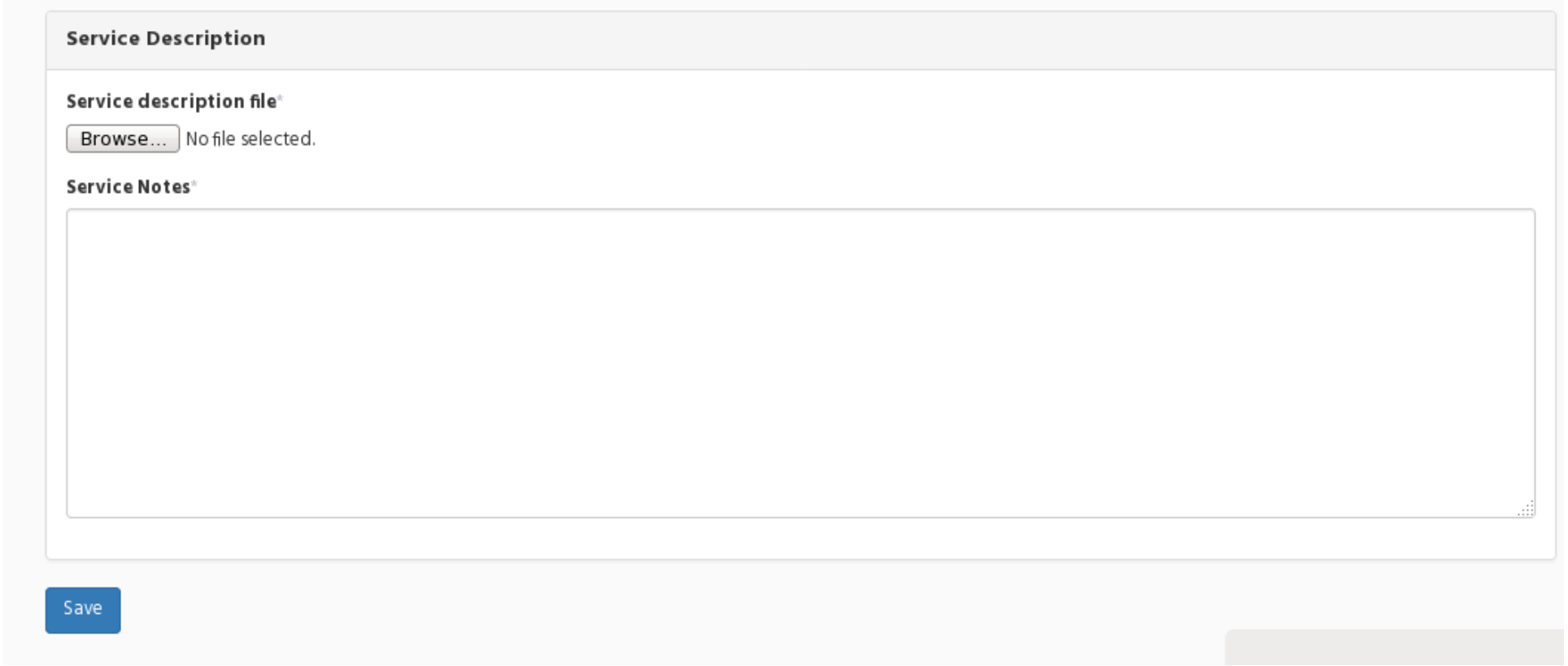
SOLICITUD DE SERVICIO https://iskylims.isciii.es/



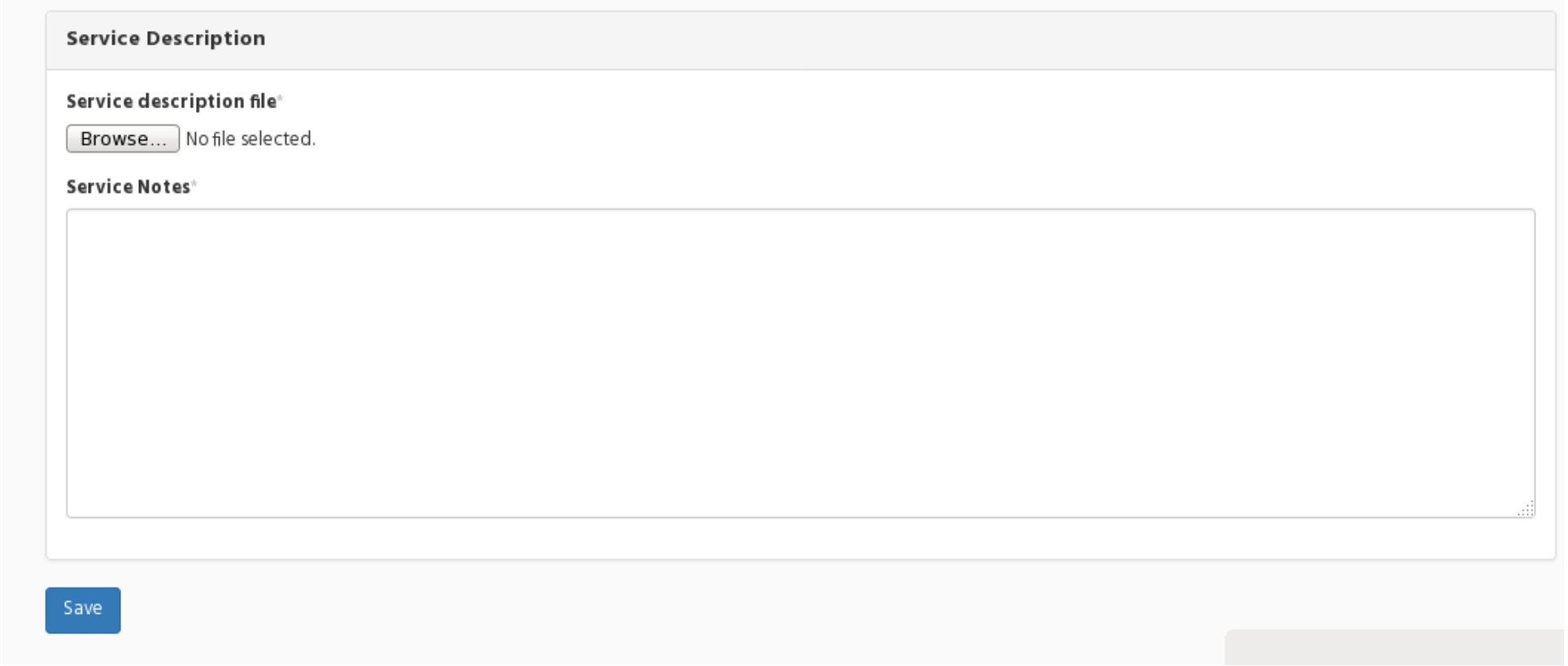
SOLICITUD DE SERVICIO https://iskylims.isciii.es/



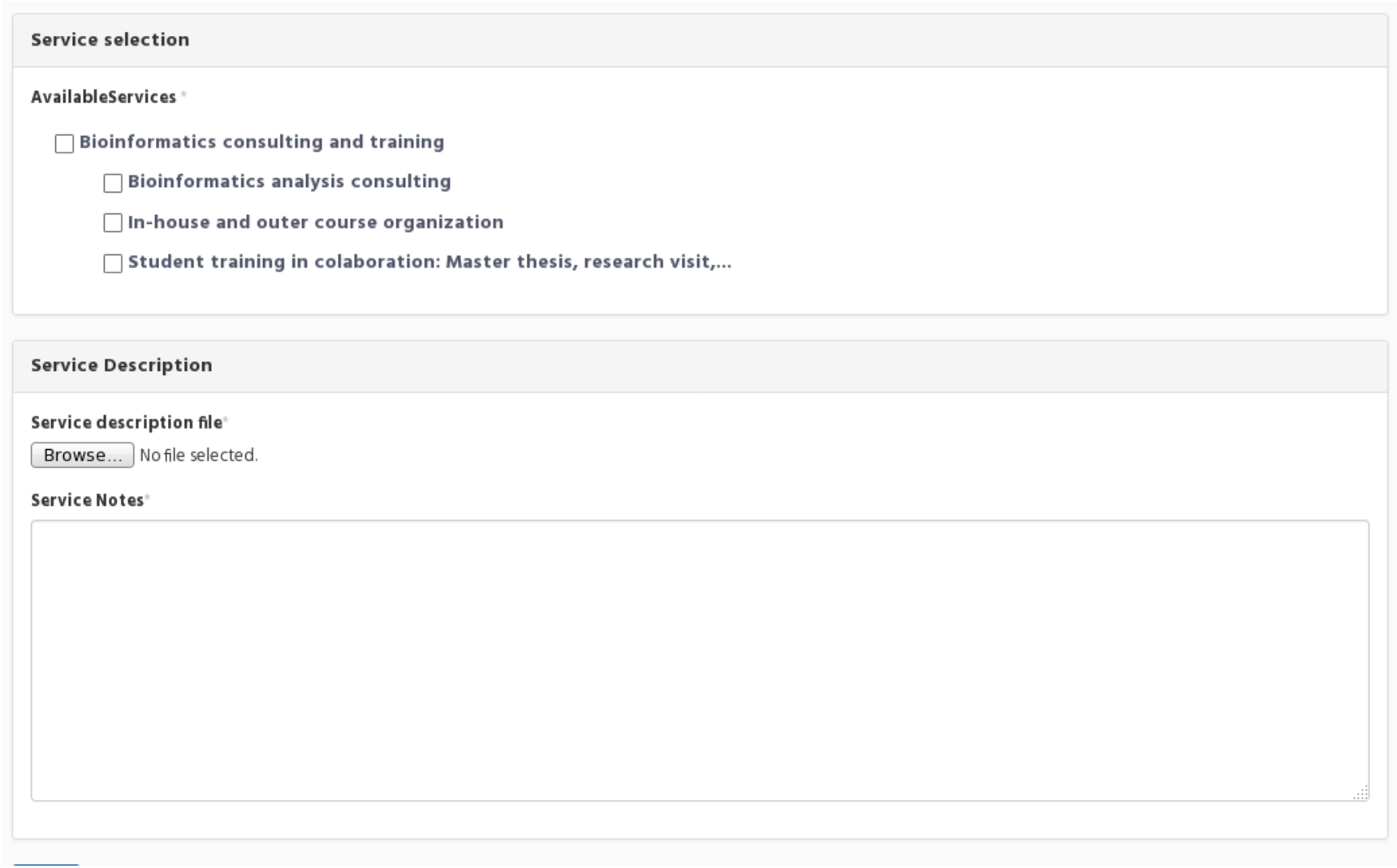
SOLICITUD DE SERVICIO https://iskylims.isciii.es/



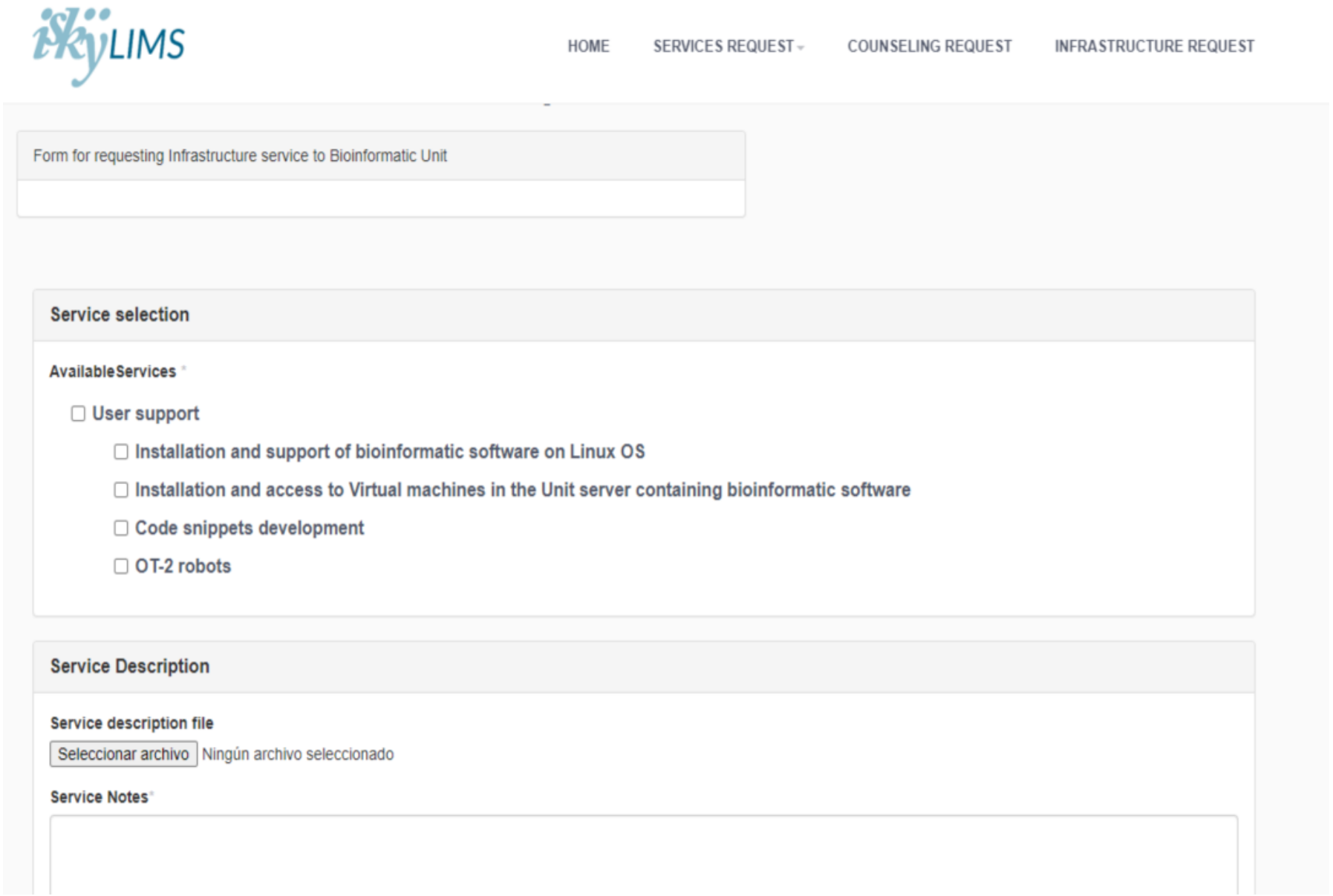
SOLICITUD DE SERVICIO https://iskylims.isciii.es/



SOLICITUD DE SERVICIO https://iskylims.isciii.es/



SOLICITUD DE SERVICIO https://iskylims.isciii.es/



Infrastructure and data management

iSkyLIMS

HOME | RUN PREPARATION | SEARCH | STATISTICS | REPORTS

bioinfoadm | Logout | My account

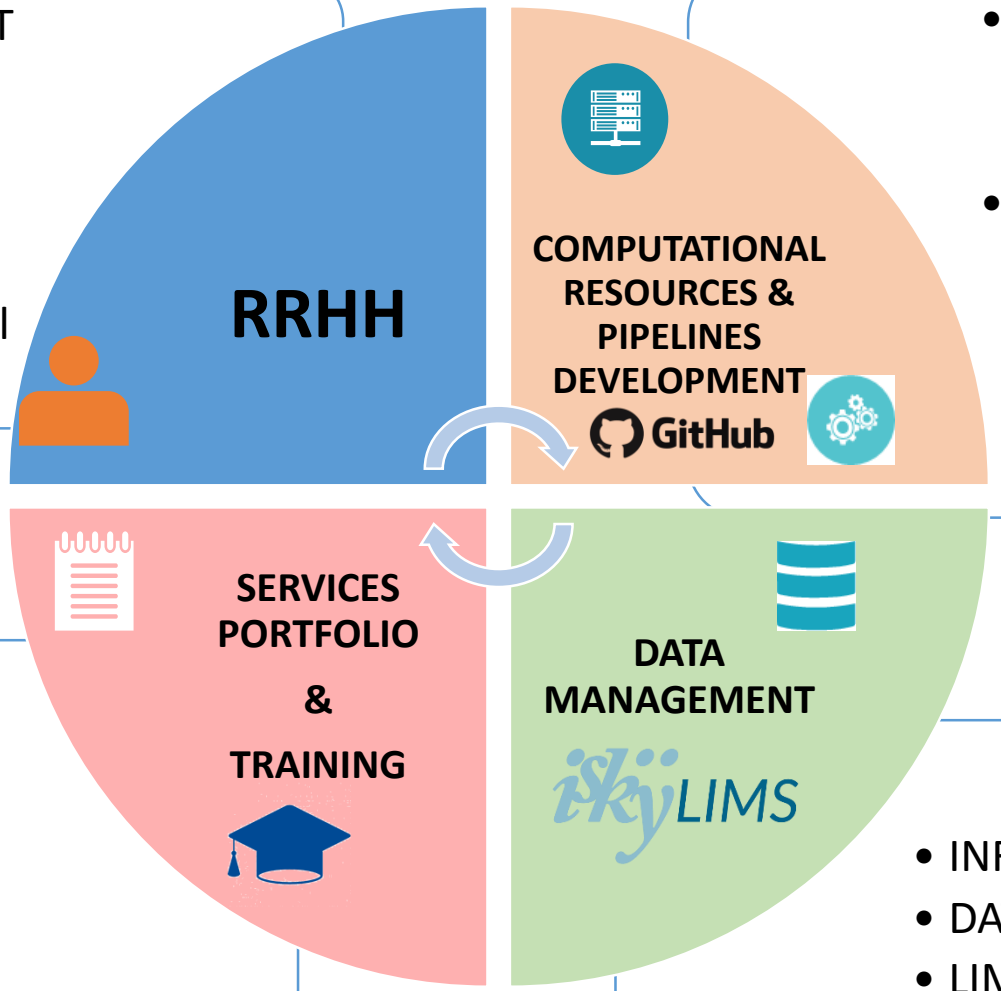
Statistics results for Investigator rabad

Projects using the sequencer NS500454 :

Export Table To Excel

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

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



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

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

- INFRASTRUCTURE
- DATA WORKFLOW
- LIMS

Cartera de Servicios

- **GENÓMICA COMPUTACIONAL: ANÁLISIS DE DATOS MASIVOS**
  Técnicas de secuenciación masiva (NGS)

- **ASESORIA Y FORMACIÓN EN BIOINFORMÁTICA**
  Orientación en el análisis bioinformático
  Organización de cursos internos y externos

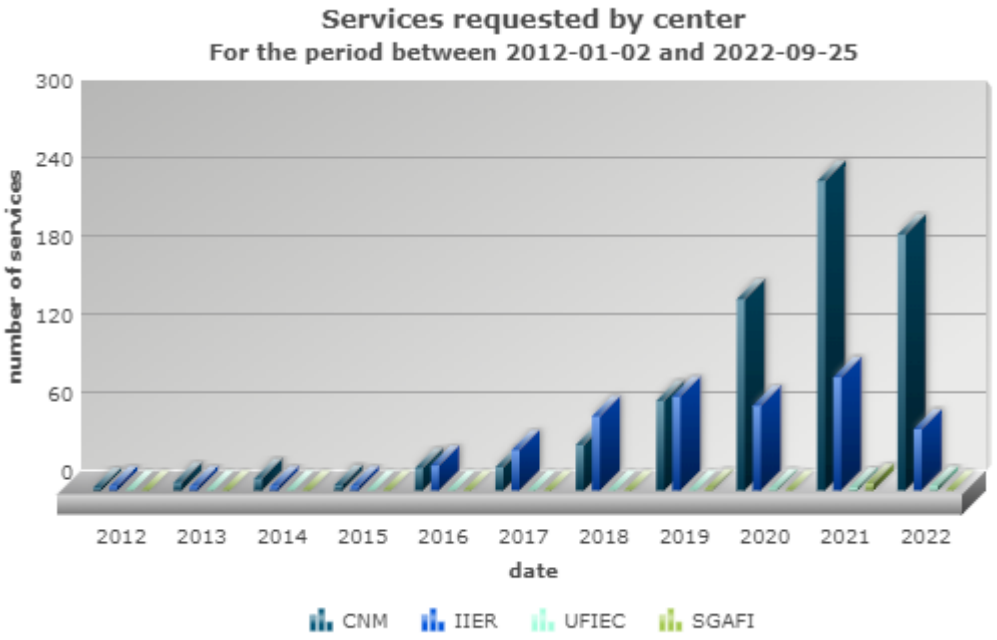
- **SOPORTE A USUARIOS**
  Generación y acceso a máquinas virtuales que contienen software bioinformático, ubicadas en los servidores de la Unidad

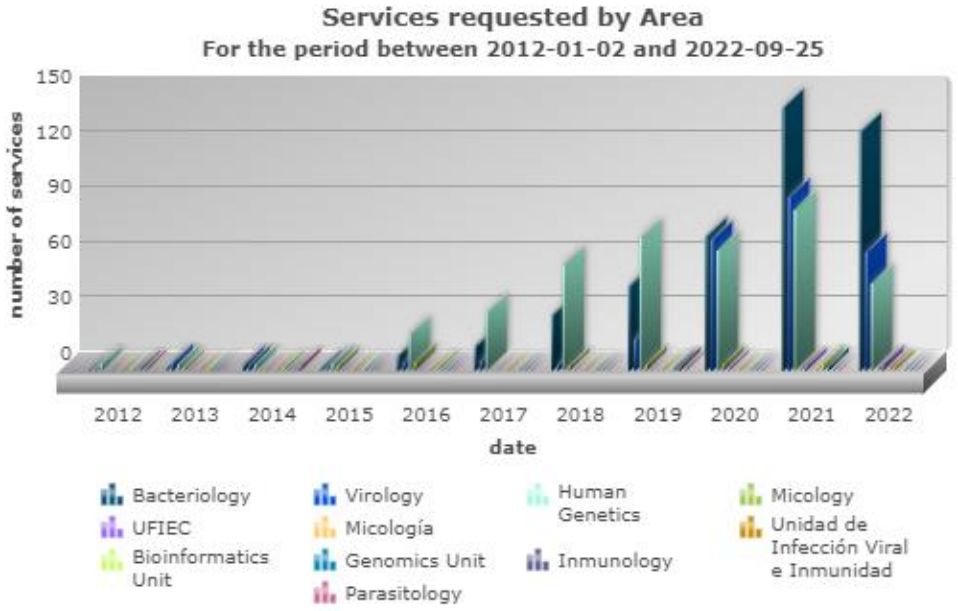
Services Portfolio

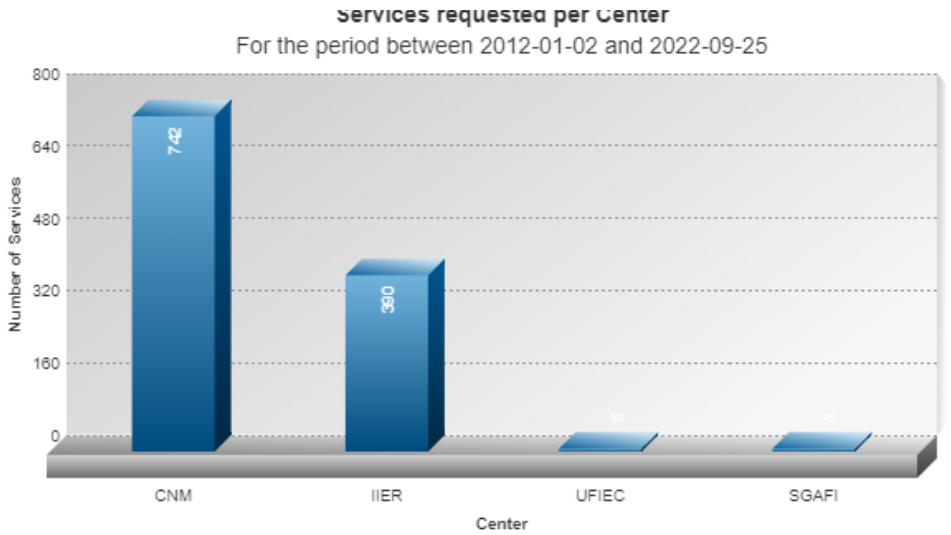
		QC	Assembly	Reference based Mapping	Variant calling	Annotation	Pipelines
DNAseq	**HUMAN**						
	WES Target –Panels	Report html		(Bam file)	(Vcf file)	Desease model (Vcf file annotated)	.Trio / family .Tumor .Pampu caller
	MICROBIAL						
	WGS Amplicon	Report html	*De novo* / Reference (fasta file)	MLST,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

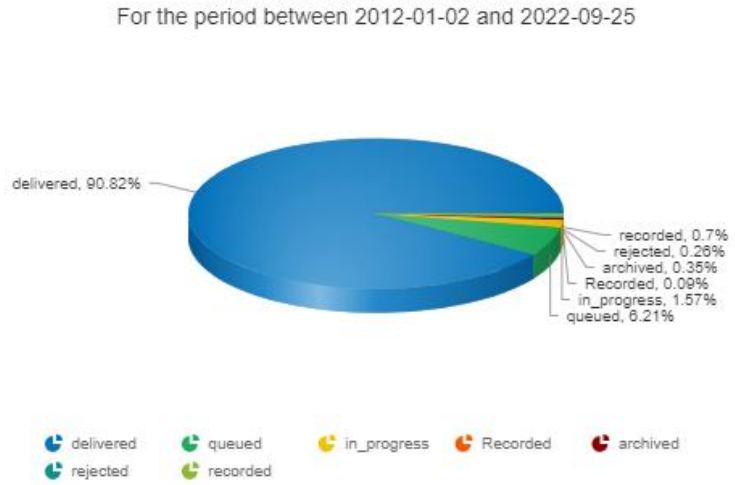
Number of services: 2012 – 2022

10 AÑOS









>1000 Services per center / CNM – IIER / 22 Researchers

Training

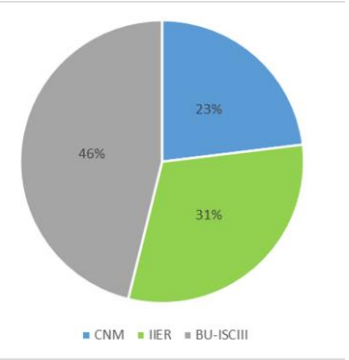
Courses

ISCIII

Introduction to massive sequencing data analysis, 2013-2021 (10 editions)

Secuenciación de genomas bacterianos: herramientas y aplicaciones, 2018-2022 (4 editions)

Análisis de genomas virales a través de la plataforma Galaxy, 2021 -2022 (2 editions)

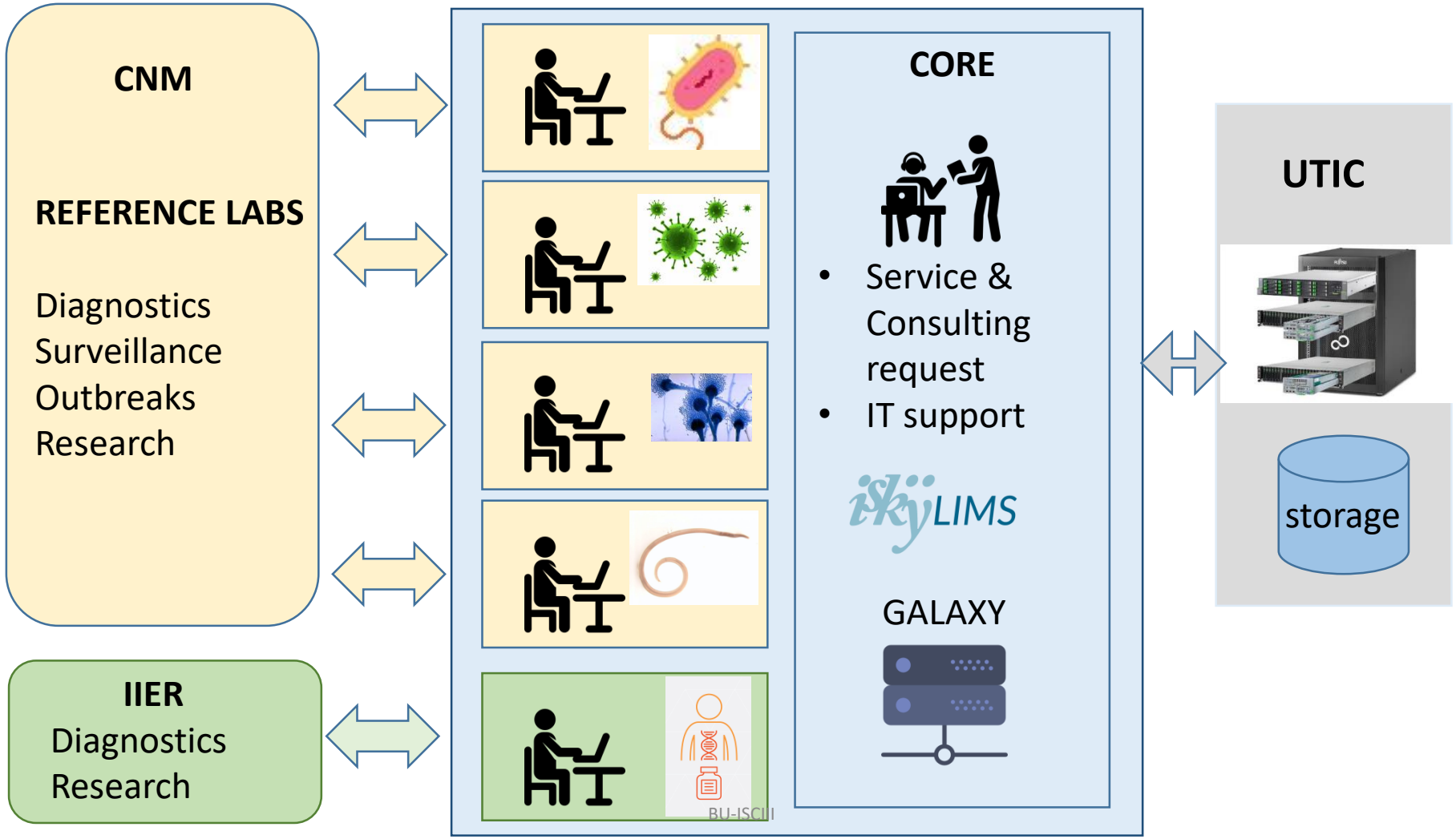


Master & Grade Students

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

Hospitals Students

Roadmap: BU-ISCIII Model



FORMACIÓN EN BIOINFORMÁTICA

Universidad Barcelona.

GRADO

BIOINFORMÁTICA
Biología
Biotecnología
Bioquímica
Medicina
Matemáticas
Química
Fisica
Informática
Telecomunicaciones

MASTER EN BIOINFORMÁTICA

PROGRAMACIÓN

Linux

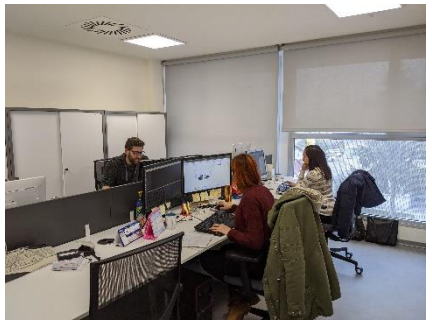
python

R

DOCTORADO

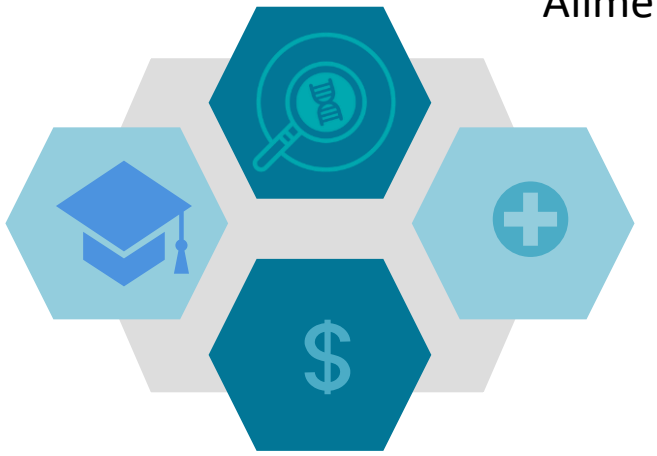
BIOLOGÍA – ANÁLISIS DE DATOS

¿Dónde trabaja un Bioinformático?



CENTRO DE INVESTIGACIÓN

Biomedicina
Agricultura
Alimentación



UNIVERSIDAD

Biociencias
Informática

HOSPITAL

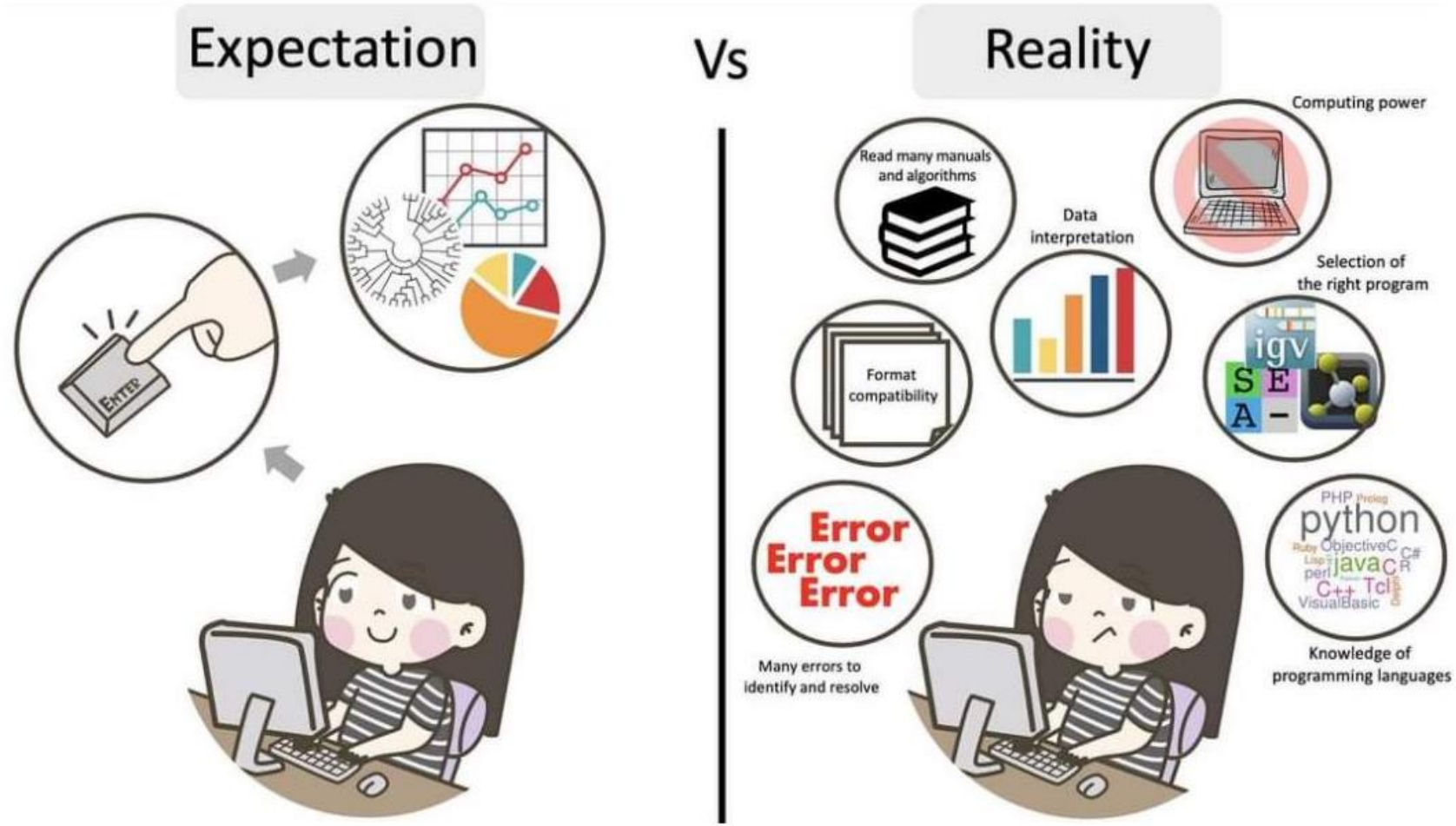
BIOINFORMÁTICO CLÍNICO
Genética
Oncología
Cardiología

EMPRESA

Bioinformática
Genética
Genómica

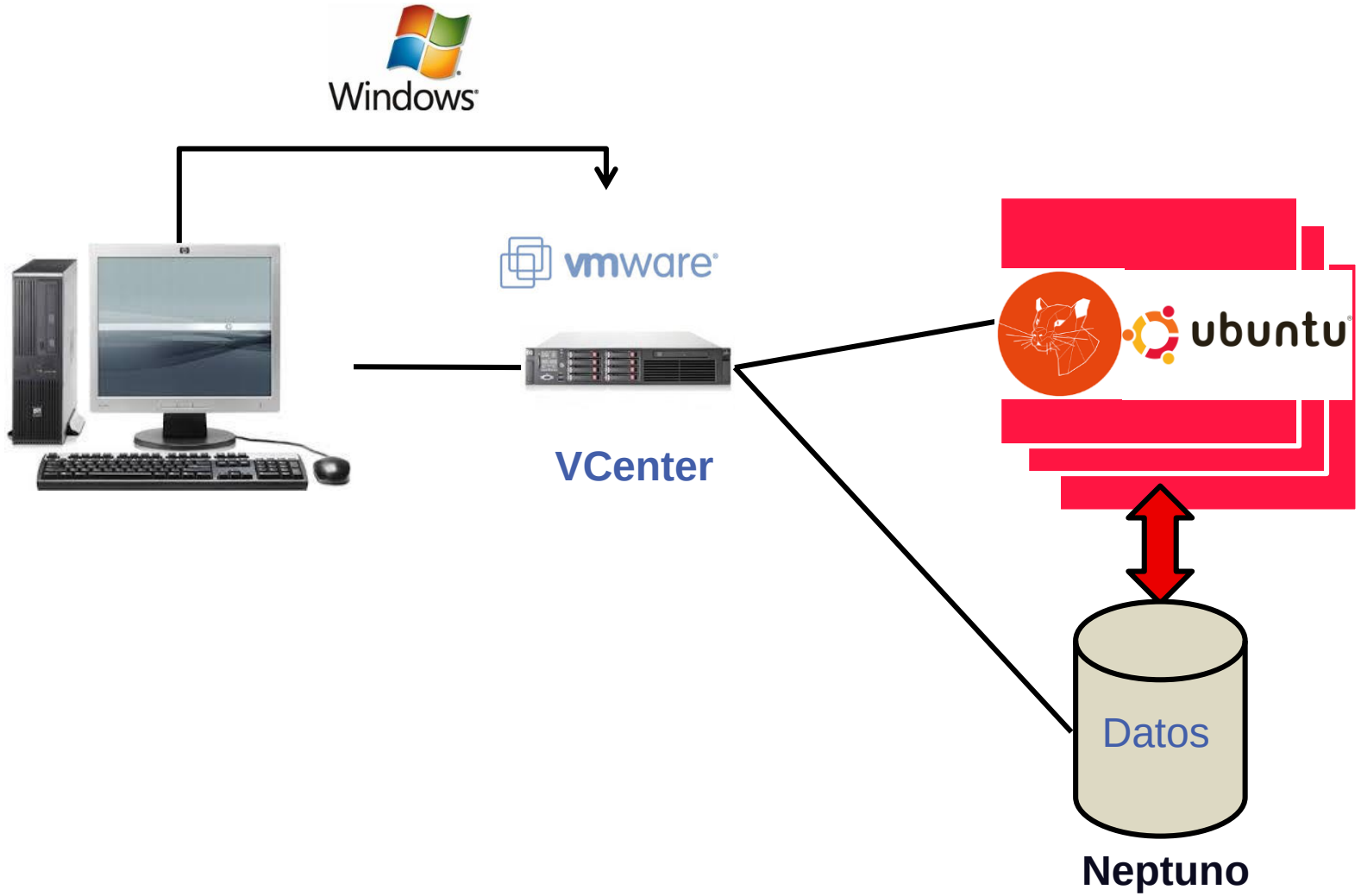
The truth about bioinformatics

.image-100[



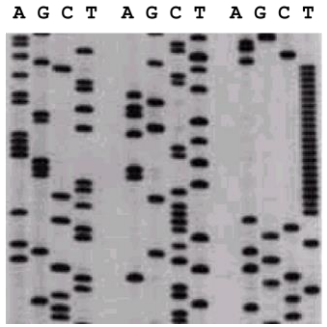
https://training.galaxyproject.org/training-material/topics/assembly/tutorials/get-started-genome-assembly/slides-plain.html

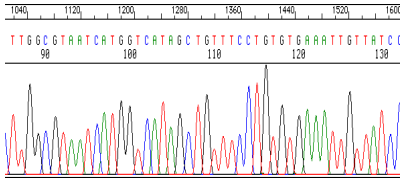
Recursos Informáticos para el curso

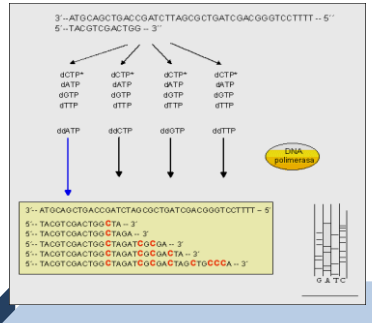


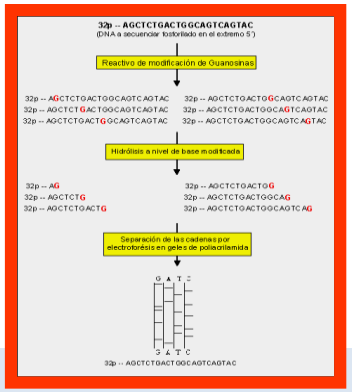
INDICE

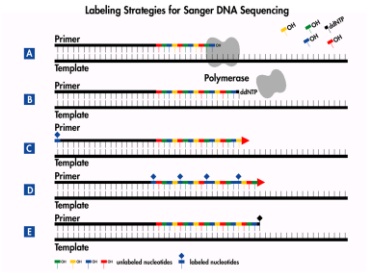
Métodos de secuenciación de DNA











MAXAM Y GILBERT

SANGER

SANGER
Secuenciadores Automáticos

1º GENERACIÓN

Evolution of DNA Revolution

A walk through the biological history: from Sanger to NGS

1953

Watson & Crick: The discovery of the molecular structure of DNA: the double helix *(Nature, 171, 1953).*

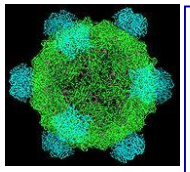
1972

Paul Berg: The first recombinant DNA molecule is build (*PNAS 69, 1972*).

Development of recombinant genetic engineering

1977

Gilbert & Maxam Sanger Developed new techniques for rapid DNA sequencing.



Bacteriofago ΦX174 5386nt plus and minus method

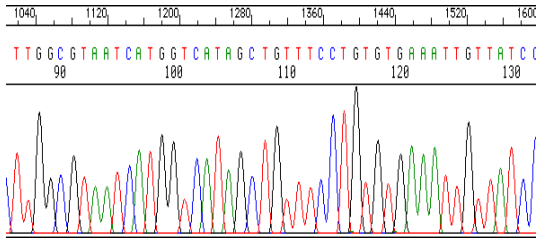
Evolution of DNA Revolution

A walk through the biological history: from Sanger to NGS

1986

1990

Hood (Applied Biosystems Incorporated, ABI) Developed the automatic sequencer (Nature 321, 1986; Science 254 1991)



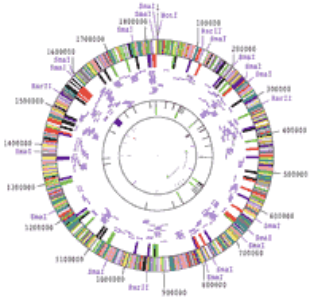
BLAST: Myers & Lipman publish the first algorithm to align DNA sequences (JMB, 215 1990).

Evolution of DNA Revolution

A walk through the biological history: from Sanger to NGS

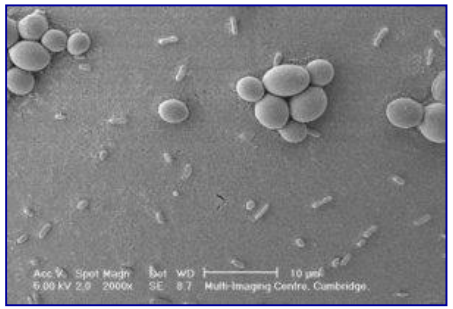
1995

Venter (TIGR)
Haemophilus influenzae
Genome (1,8Mb),
Mycoplasma genitalium
(0,5Mb)
(Science 269, 1995)



1996

Consortium of 600 Scientist from Europe, North America and Japan Saccharomyces cerevisiae genome (Science).



2003

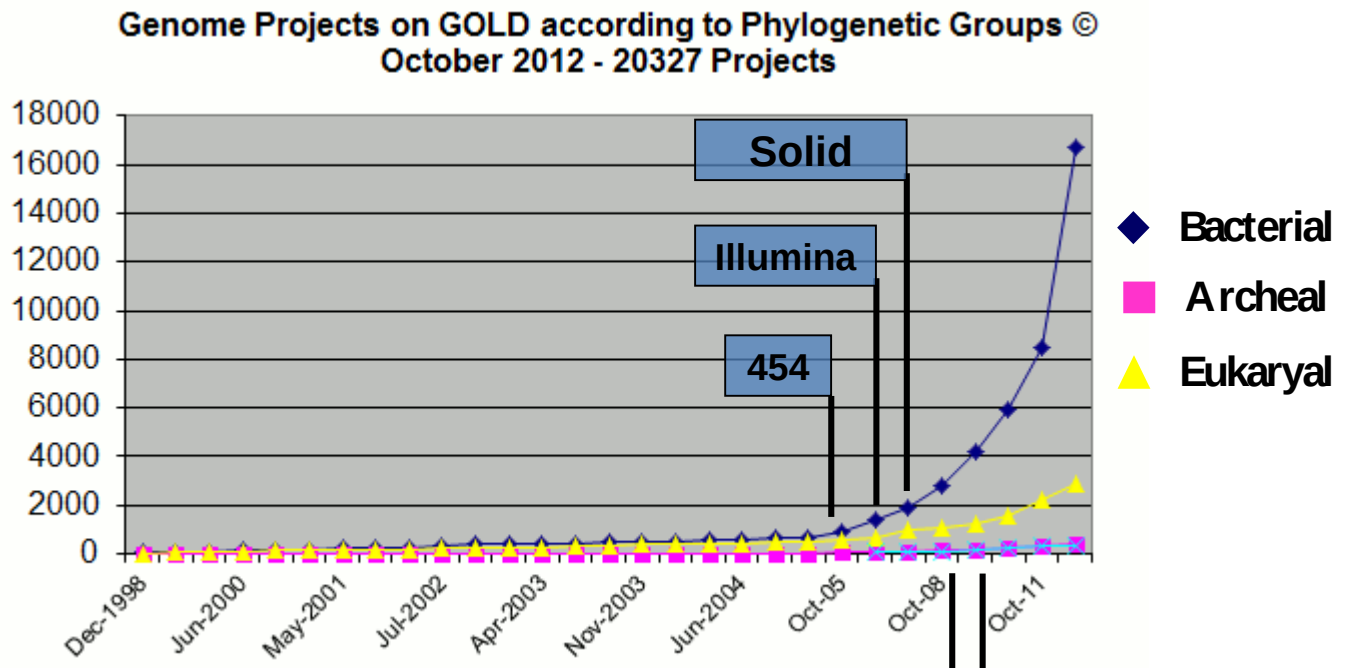
Bacterial Genomes

Bacillus subtilis
Escherichia coli

Eukaryotic Genomes

Caenorhabditis elegans
Arabidopsis thaliana
Drosophila melanogaster
Human genome **(1986-2003)**

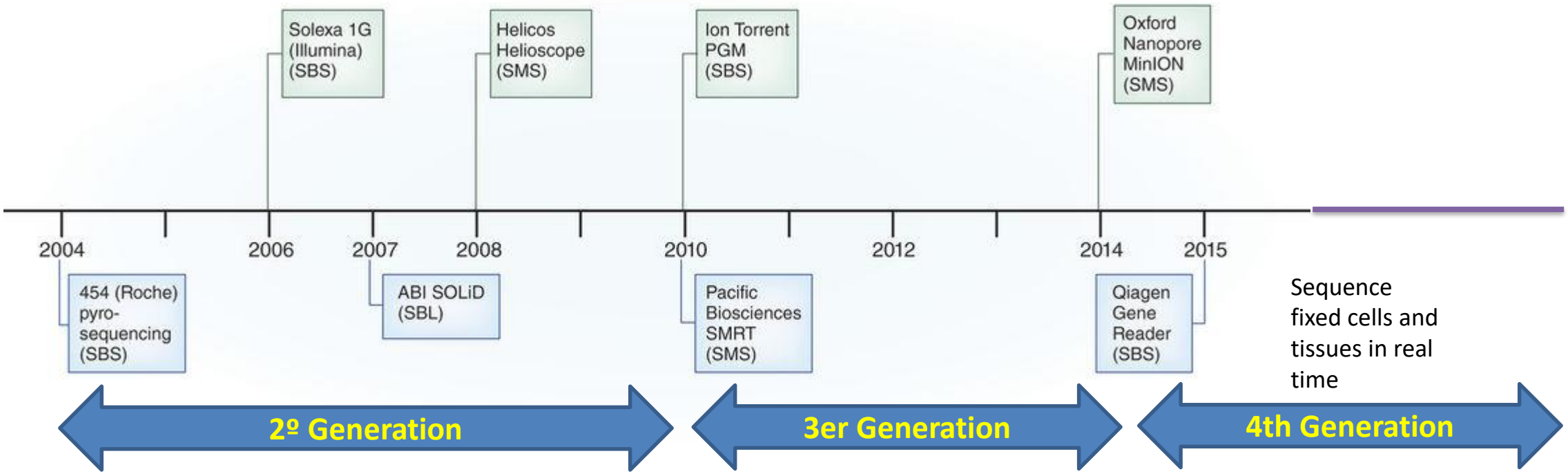
Genomics Revolution Era



Source: http://www.genomeonline.org

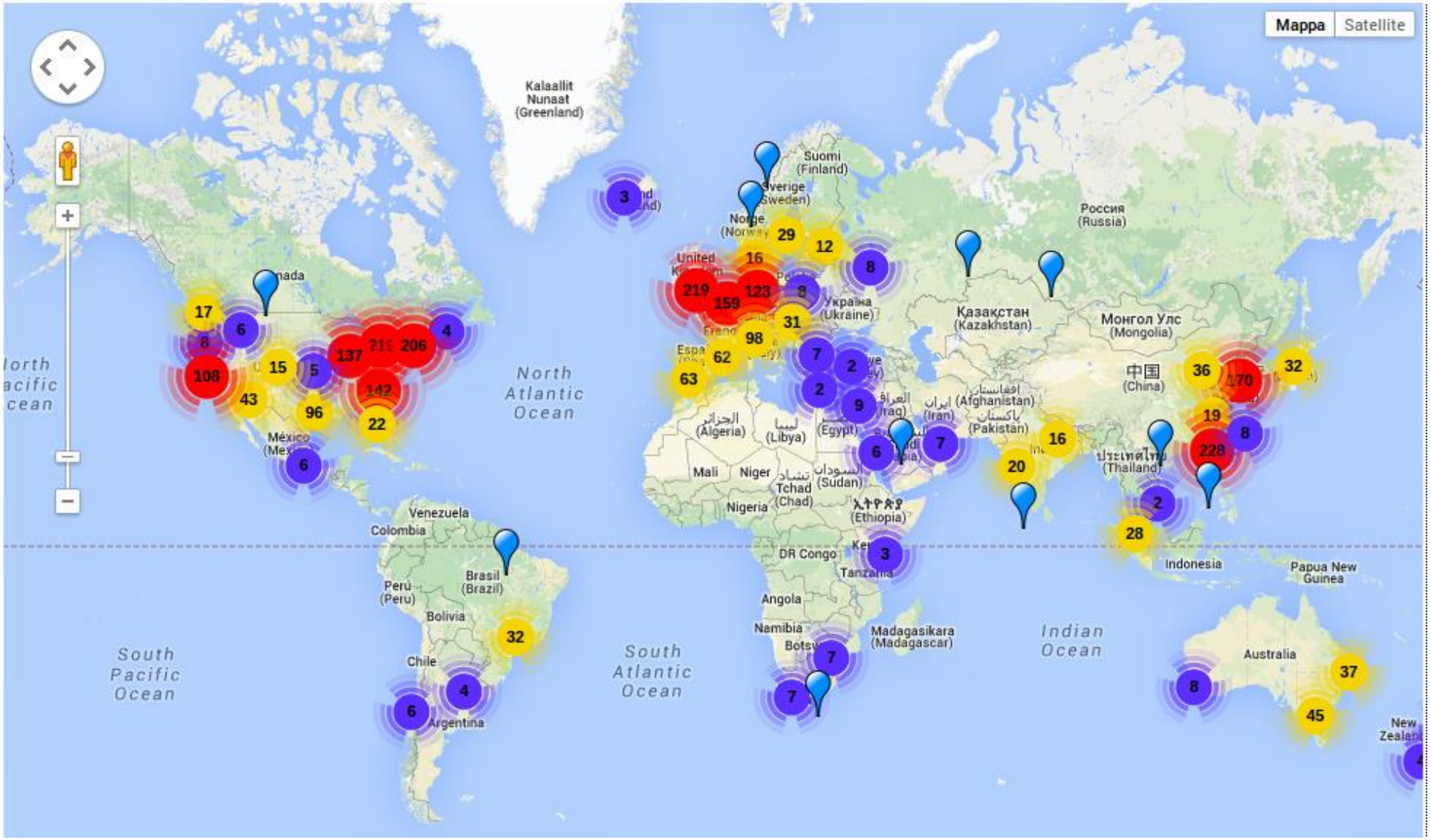
1000 Genomes Project

NGS Platforms - Timelime



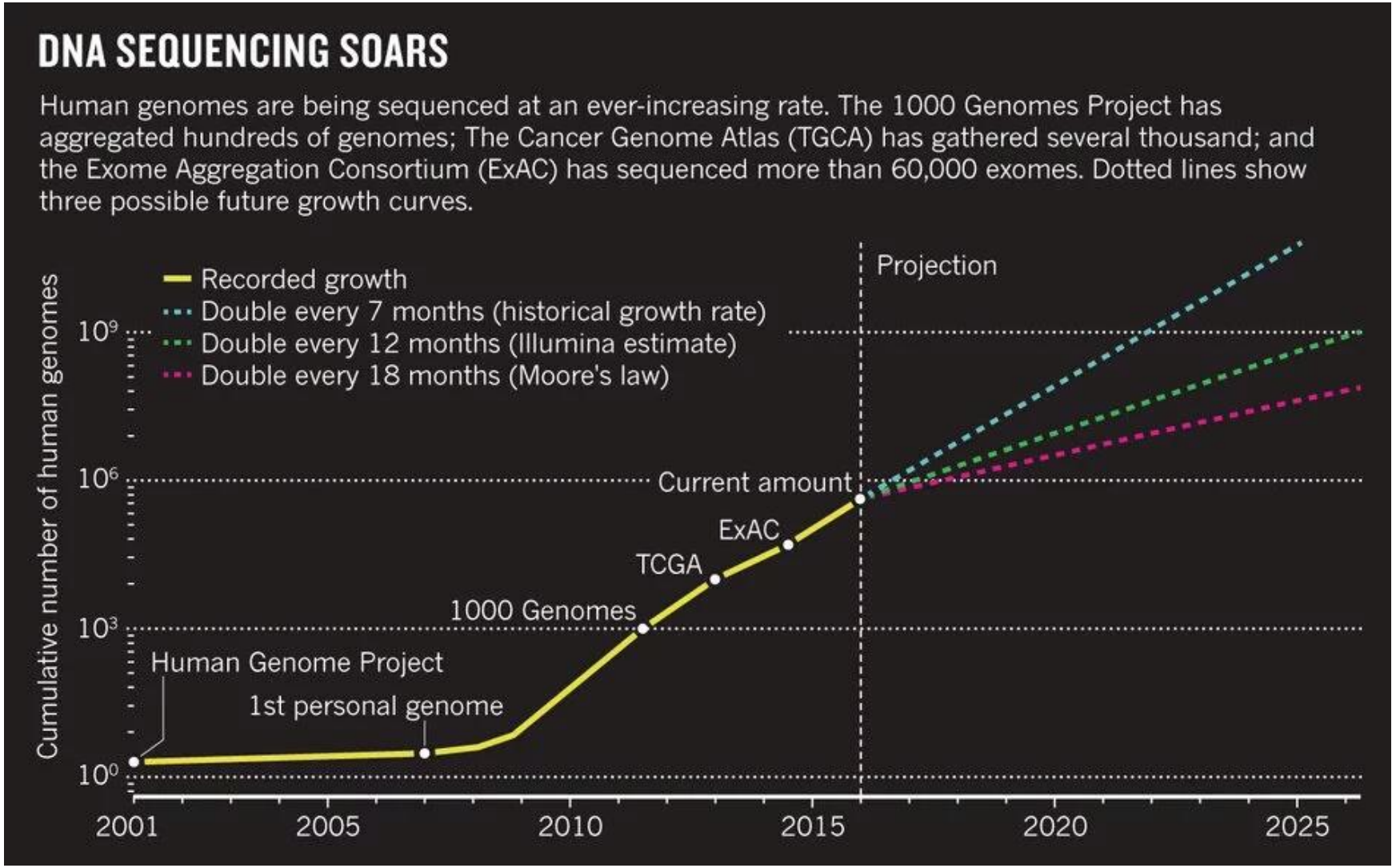
Democratización de la secuenciación.

Mapa de secuenciadores de alto rendimiento



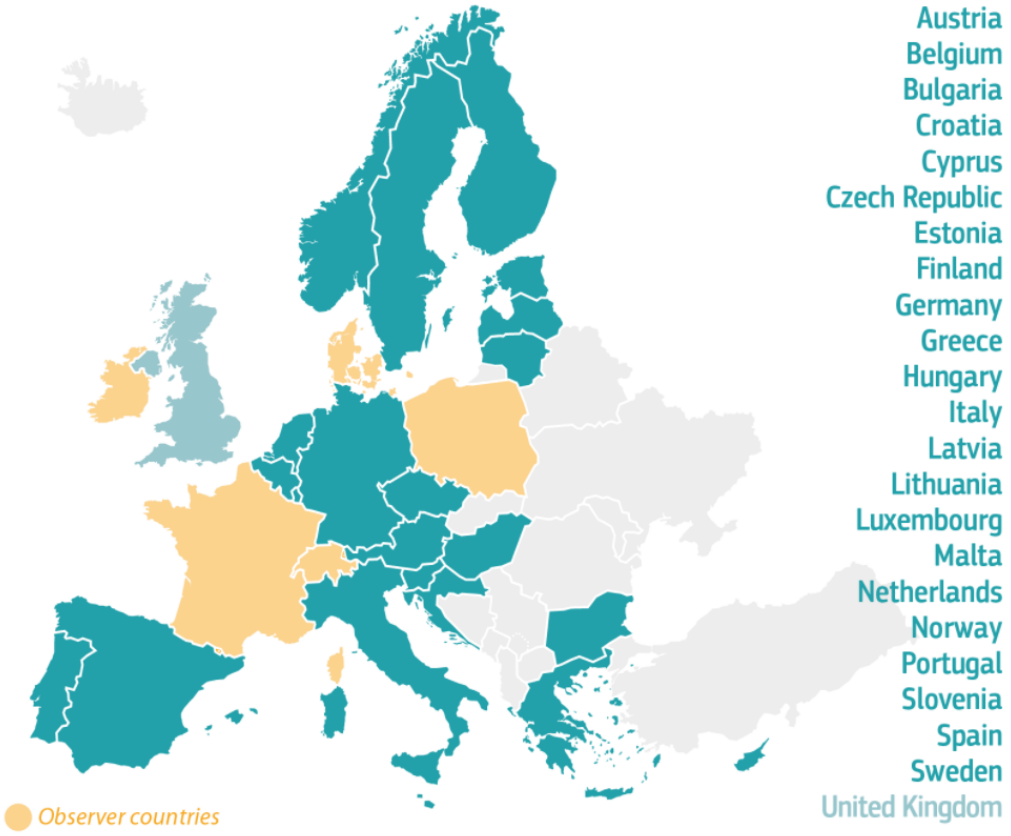
http://omicsmap.com

Sequencing projects



Sequencing projects

Countries that have signed the 1+MG Declaration since 2018



Austria
Belgium
Bulgaria
Croatia
Cyprus
Czech Republic
Estonia
Finland
Germany
Greece
Hungary
Italy
Latvia
Lithuania
Luxembourg
Malta
Netherlands
Norway
Portugal
Slovenia
Spain
Sweden
United Kingdom

The vision

Designing and testing

Scaling up and sustaining

1+MG

B1MG

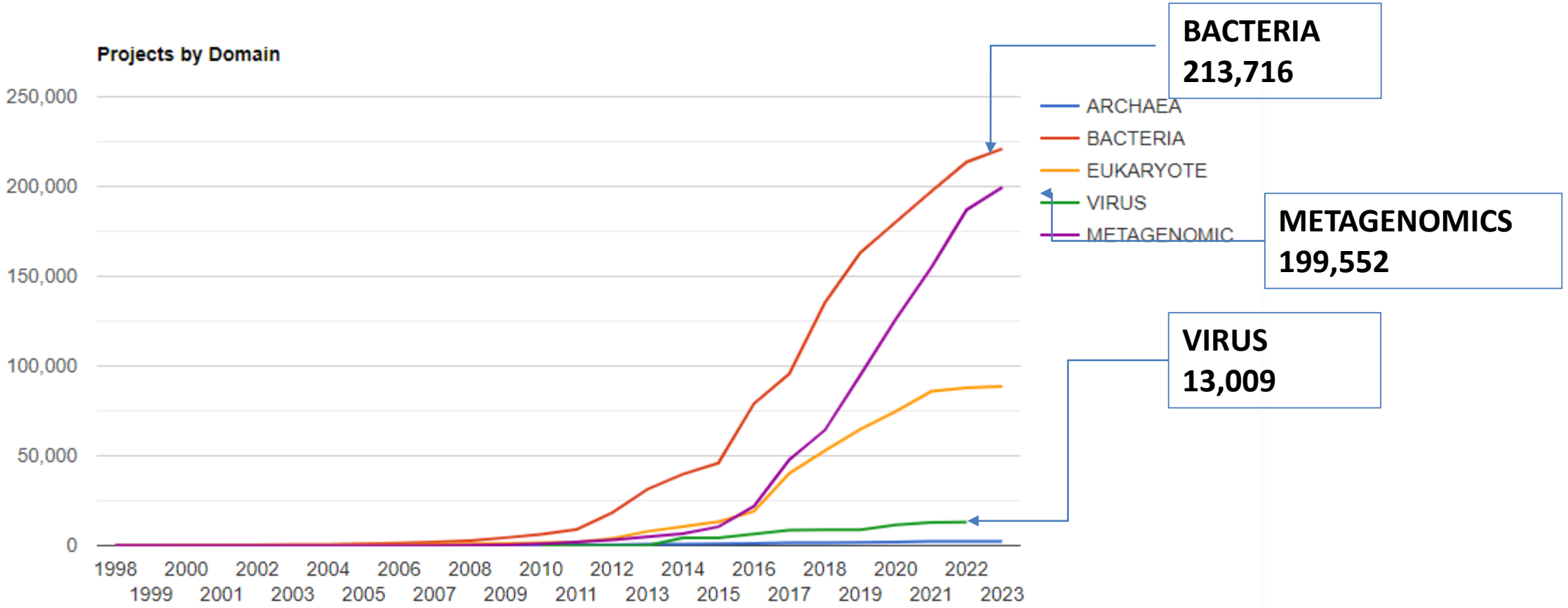
European Genomic Data Infrastructure

https://gdi.onemilliongenomes.eu/

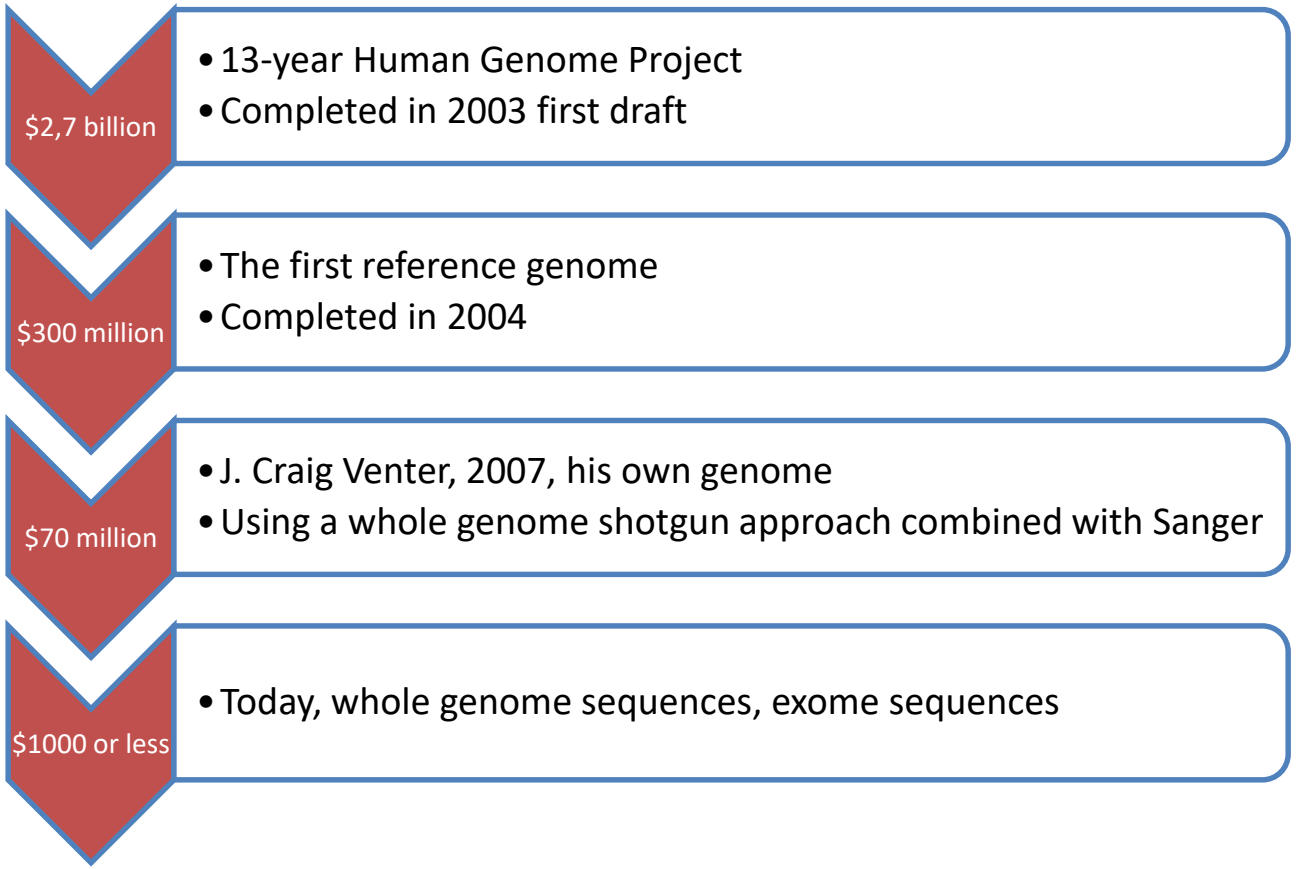
Sequencing projects

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

GOLD, Genome Online DataBase

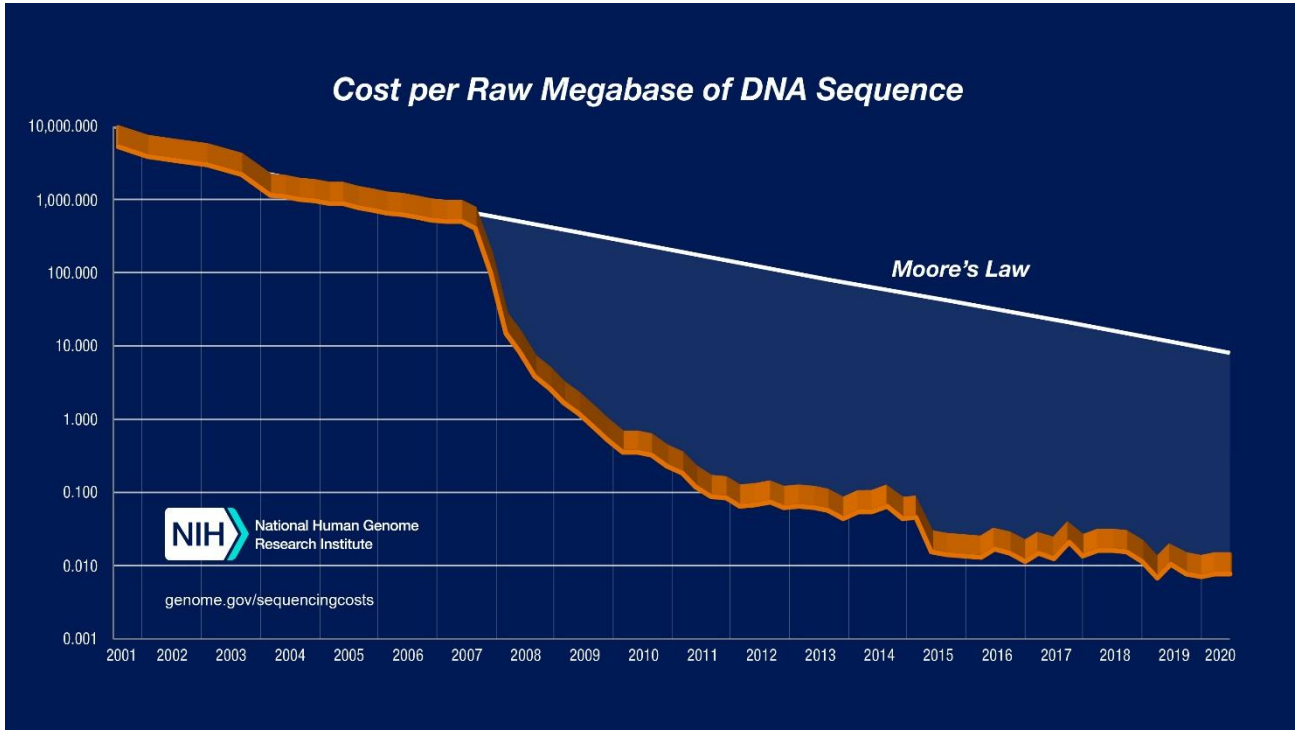


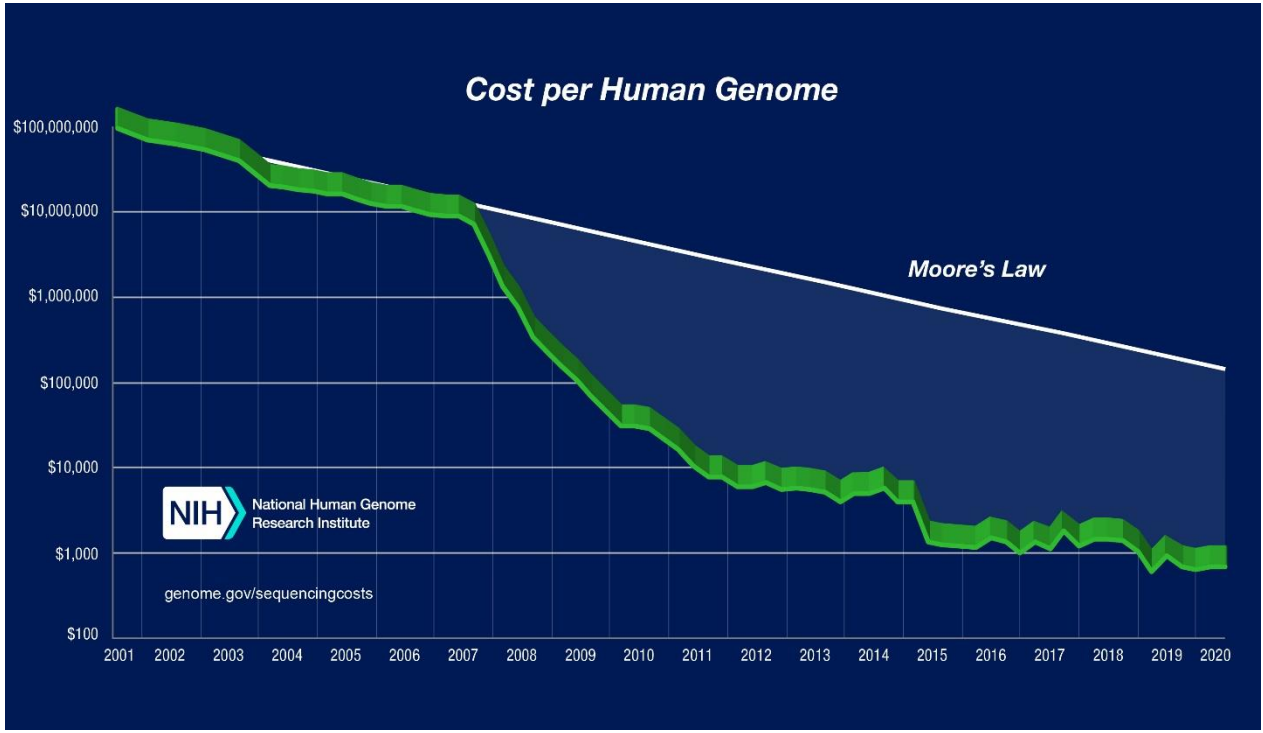
Evolución del coste de la secuenciación de un genoma humano



https://www.aacc.org/publications/cln/articles/2012/april/sequencing

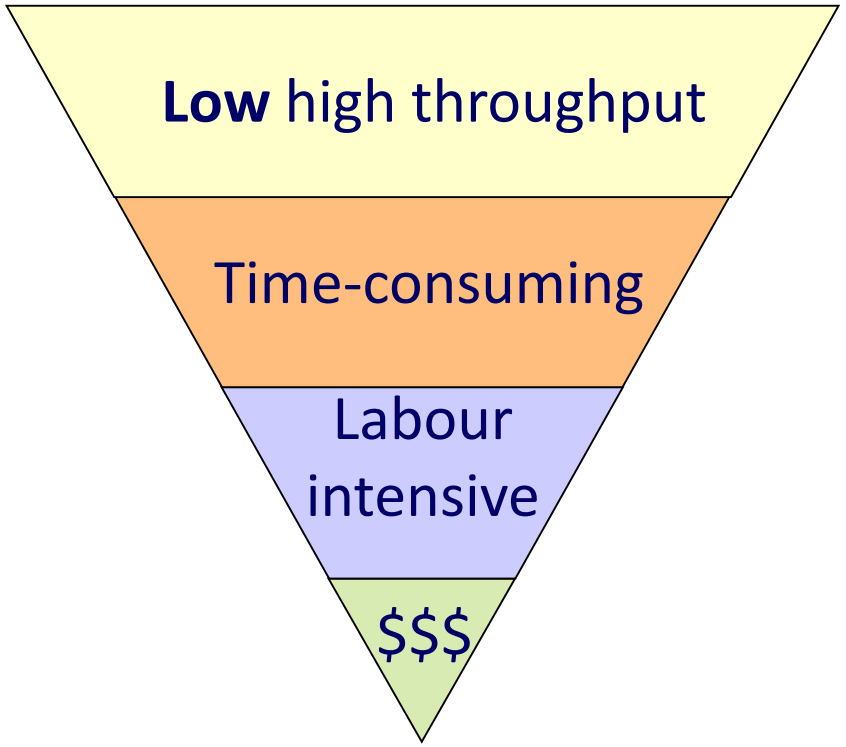
Coste actual de la secuenciación

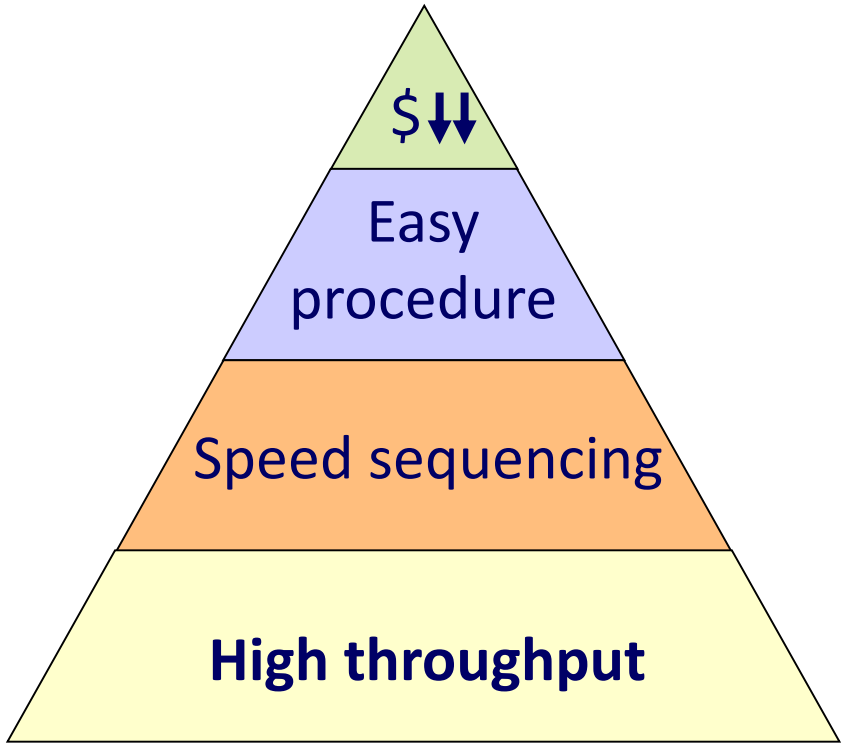




Sanger vs NGS

advantages of new technologies

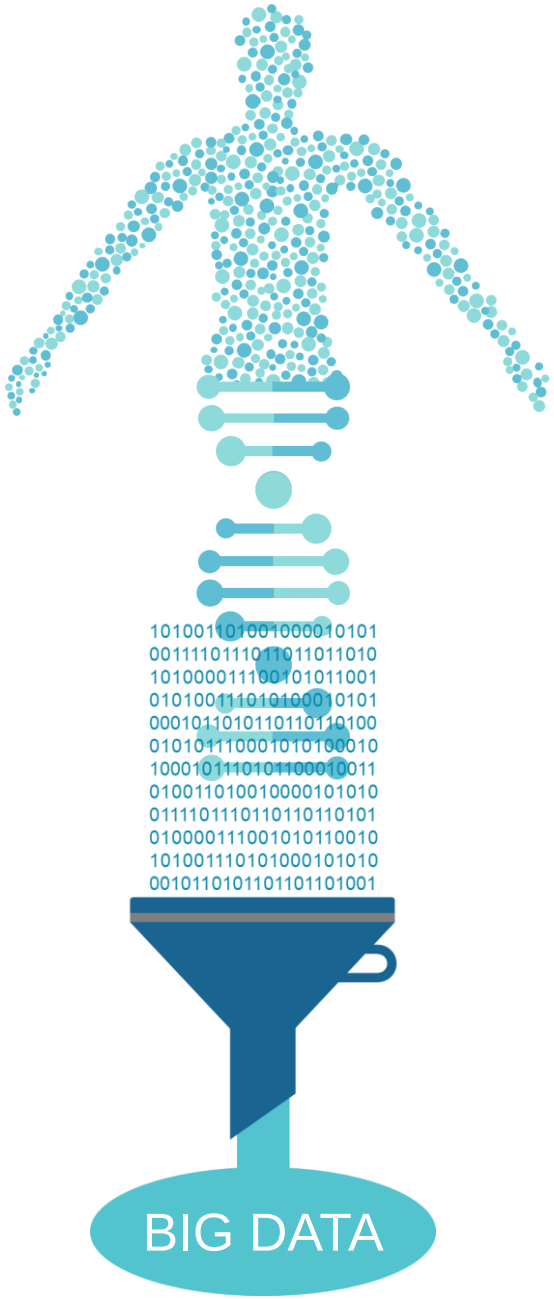


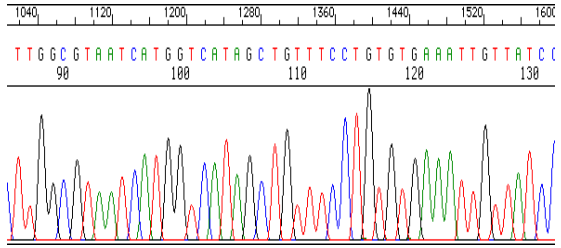


Semiautomatic **Sanger** capillary-based sequencing technology

NGS
Next Generation Sequencing =
Now Generation Sequencing

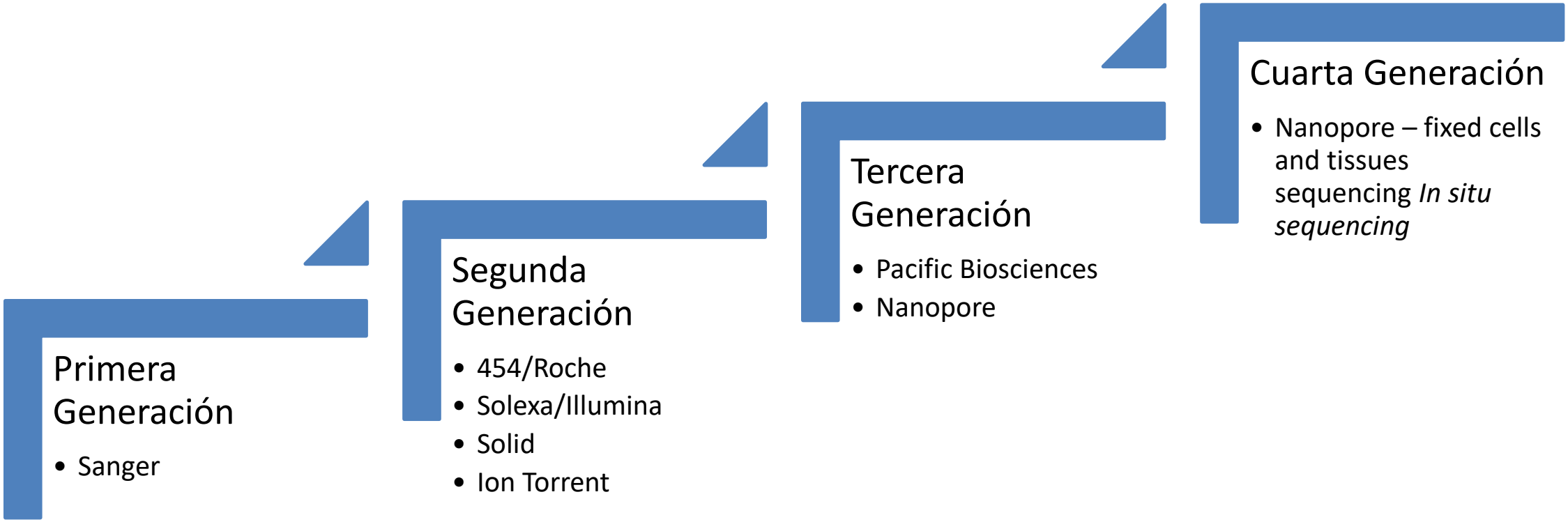
BIG DATA



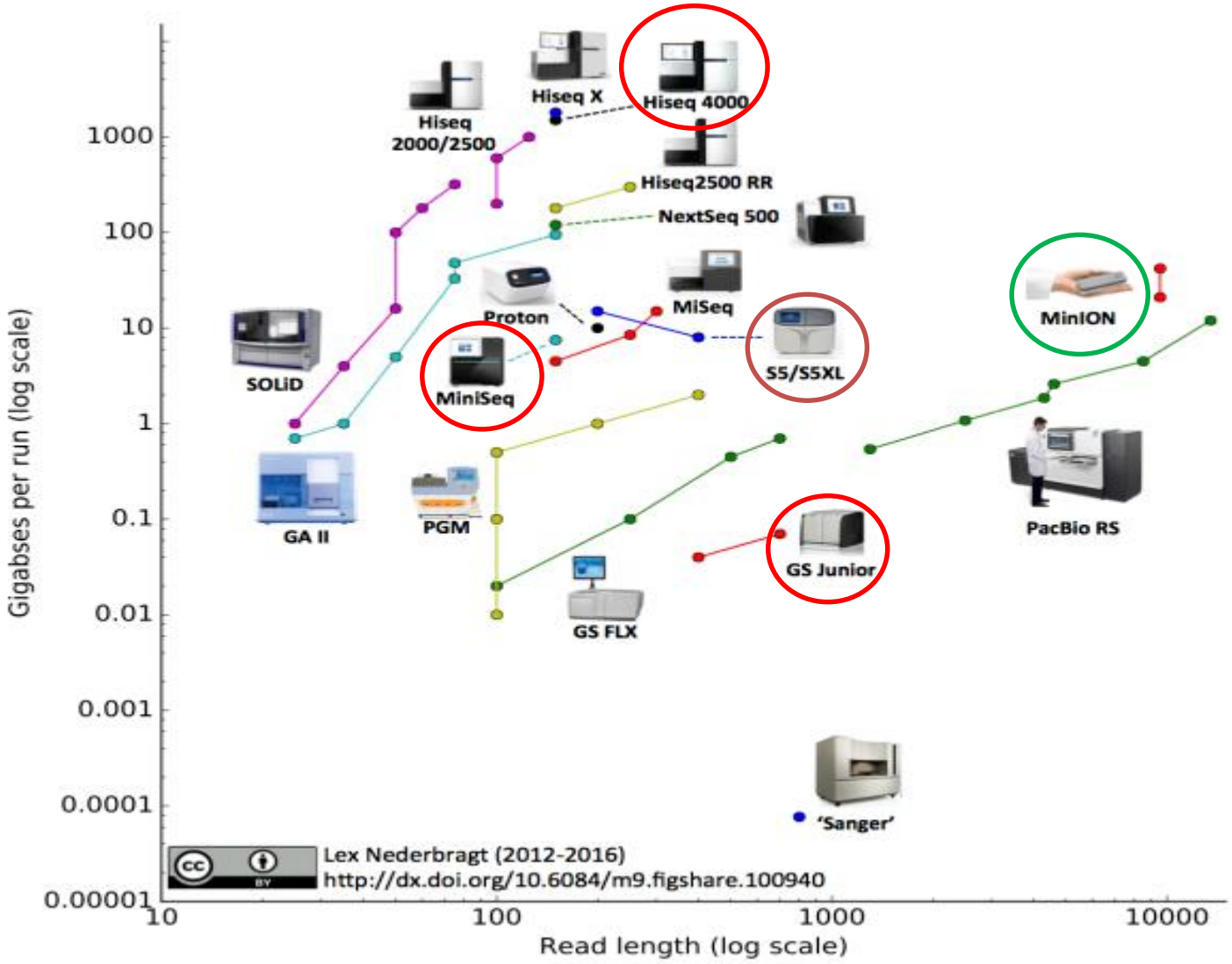


```
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:2458:1027 1:N:0:ACAGTG
AGAAAAAACCTTGGANGGAAAAAAATCAGACATTTTCTAGAGGTGGAAGGCAAACTGAACAAAGAAATAATTCACA
+
DGGGEDHHHHGGGFE#CBACBCA<?HHHHBHHHHHHHHDHHHHEHEFEGGGGGG/GGDDDGHFHGFCHFHHEHEH8
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3082:1029 1:N:0:ACAGTG
GGTAATACAGACTGANATGATCAAAGGCATGCTGGAAACAAACCTATTAAAGATAAGCTTGGATCAAGCTTTCATT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

Generaciones de Secuenciadores

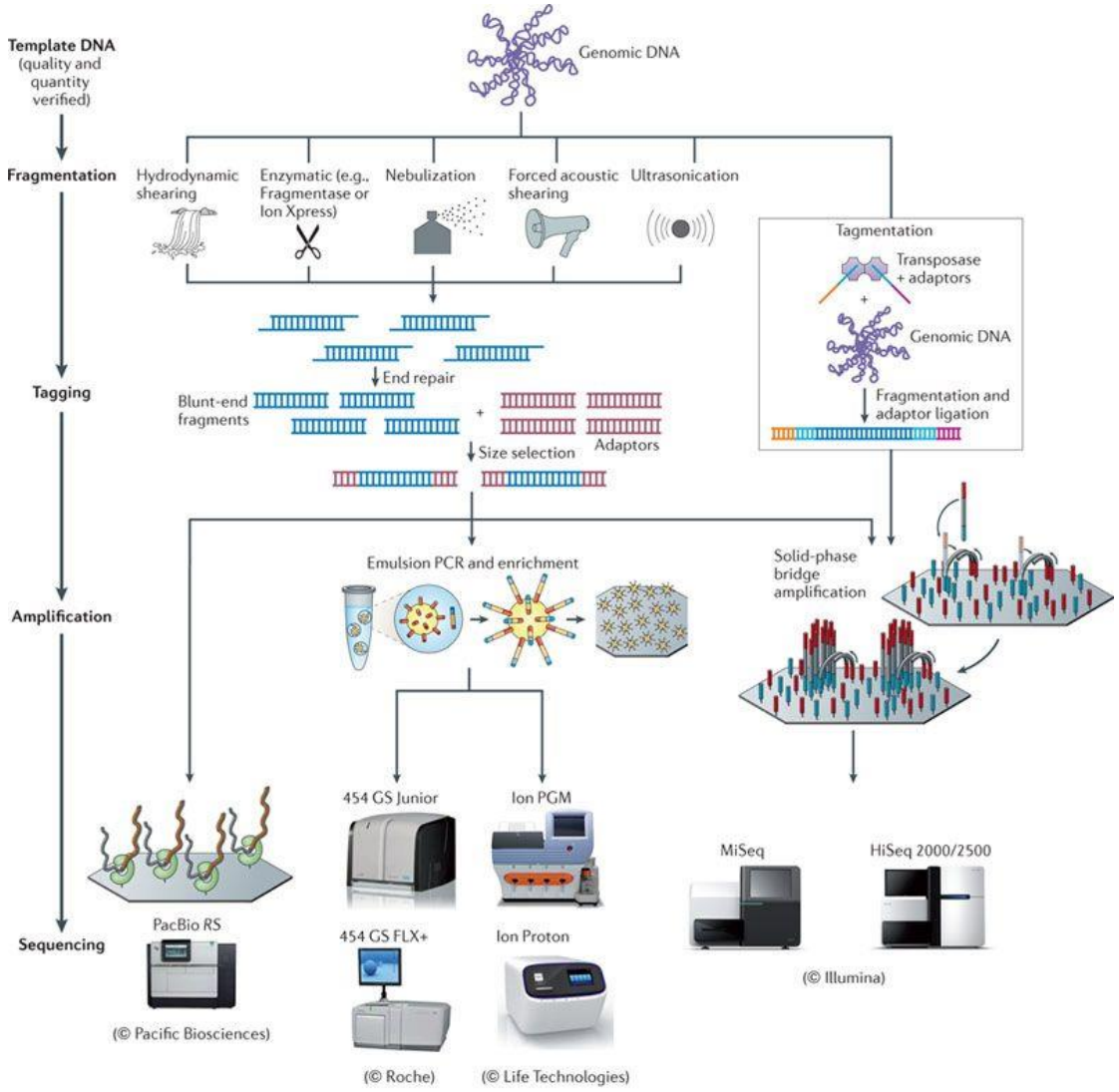


PLATAFORMAS DE SECUENCIACIÓN. 2016 Edition



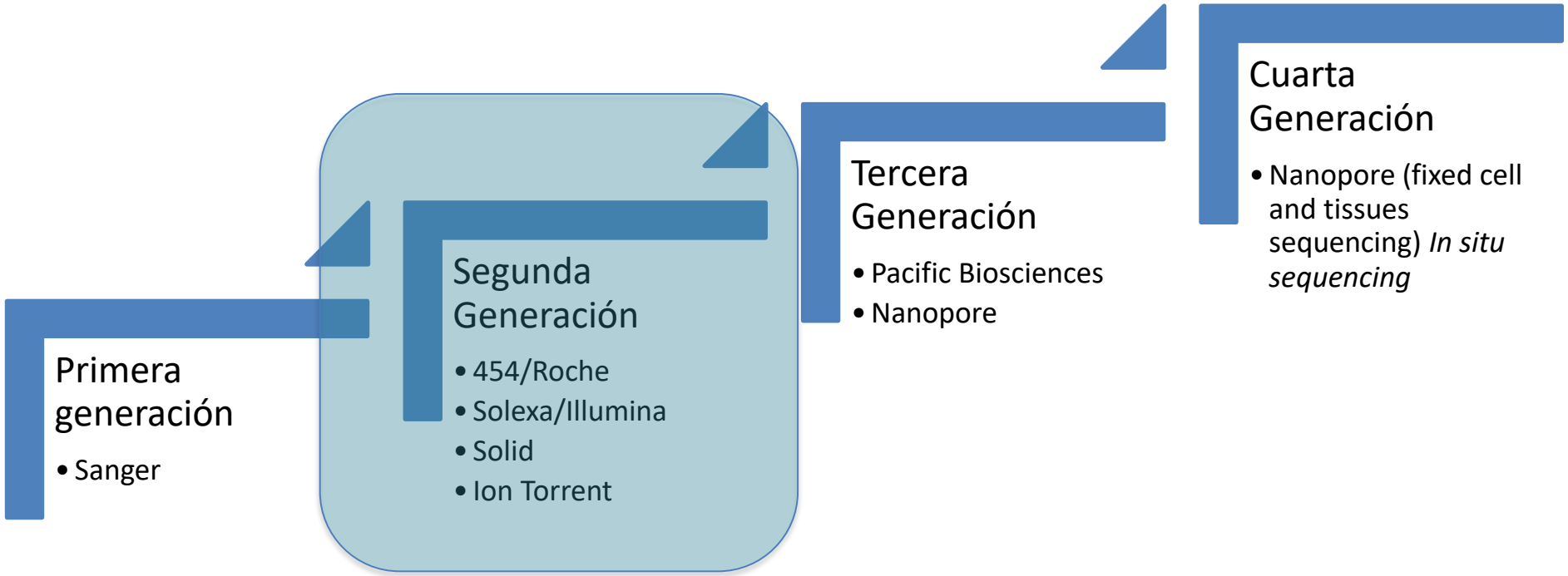
https://flxlexblog.wordpress.com/

High-throughput sequencing platforms



Nature Reviews | Microbiology Loman et al, 2012

Secuenciadores



High-throughput sequencing platforms 2nd GS

GS-FLX System

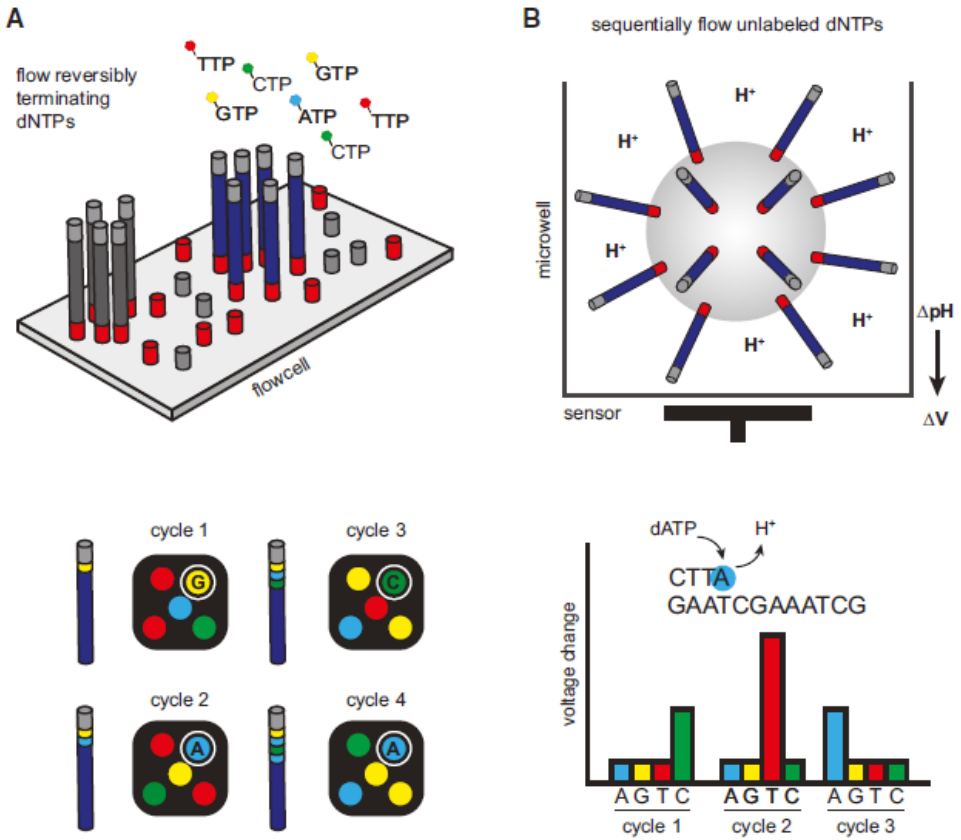
Genome Analyzer IIx

SOLID 3 Plus/4

Sequencing Chemistry	Sequencing by synthesis, pyrosequencing	Sequencing by synthesis with reversible terminators	Sequencing by ligation
Amplification approach	Emulsion PCR	Cluster amplification	Emulsion PCR
DNA support	25-35 µm bead	Flow cell surface	Bead (Solid 3 Plus/4) Flow cell surface (GA5500w)

Mardis et al., Trends in Genetics 2008, 24:3

The Second-generation Sequencing Technologies



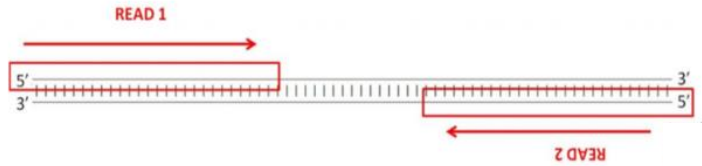
Clonal Amplification-Based Sequencing Platforms

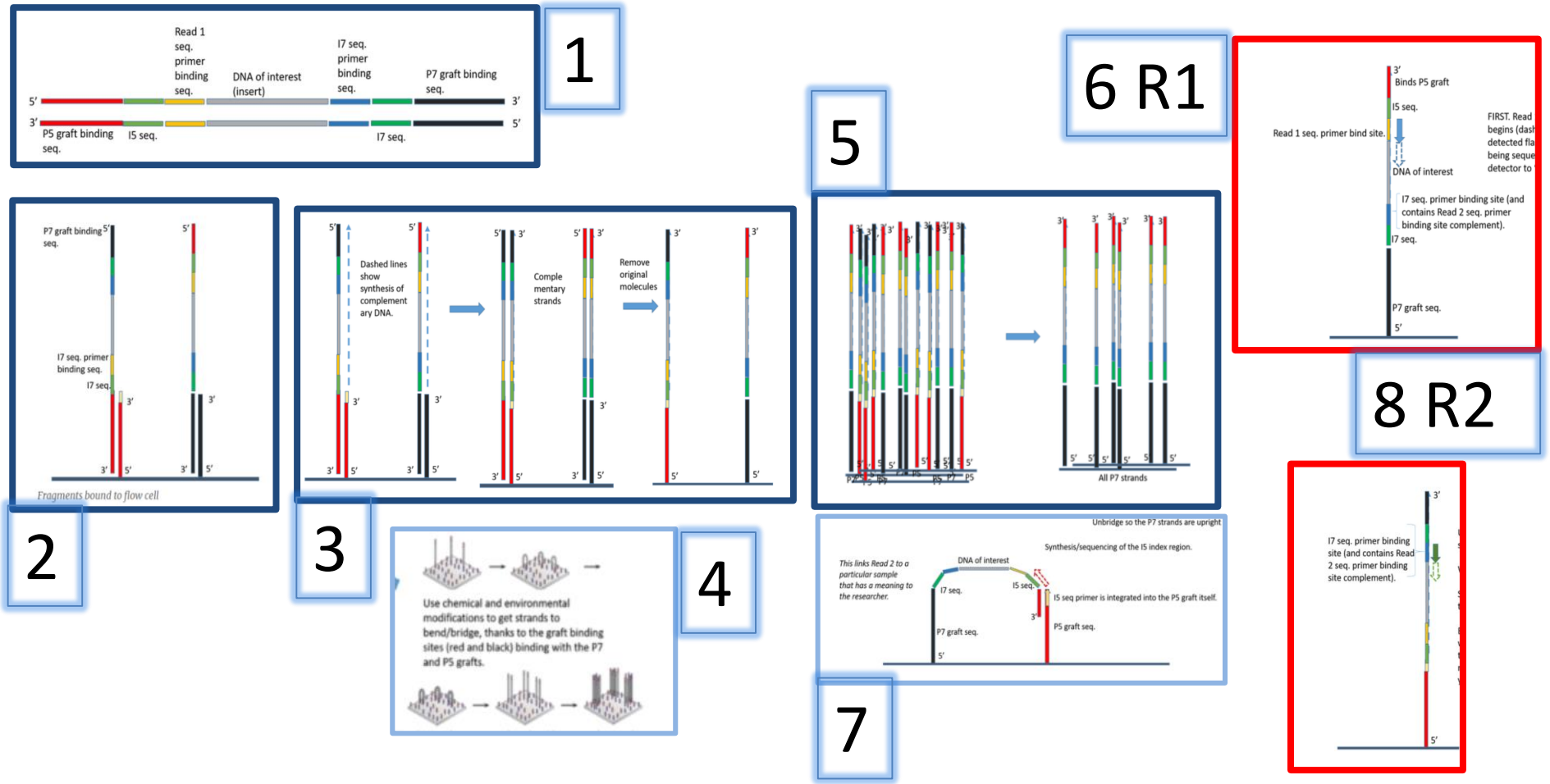
(A) Illumina's four-color reversible termination sequencing method.

(B) Ion Torrent's semiconductor sequencing method.

Reuter et al., Mol Cell 2015

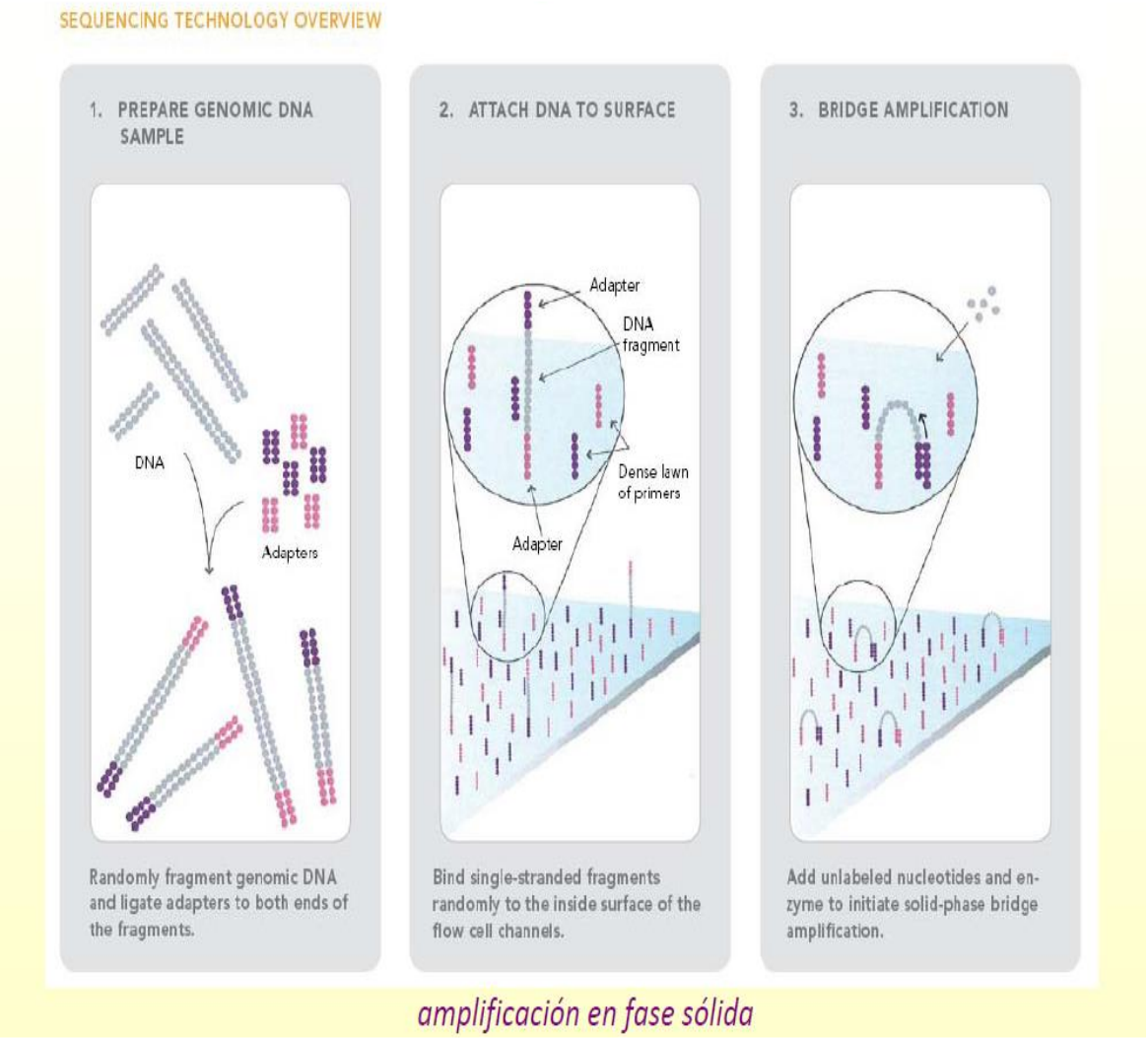
Illumina sequencing



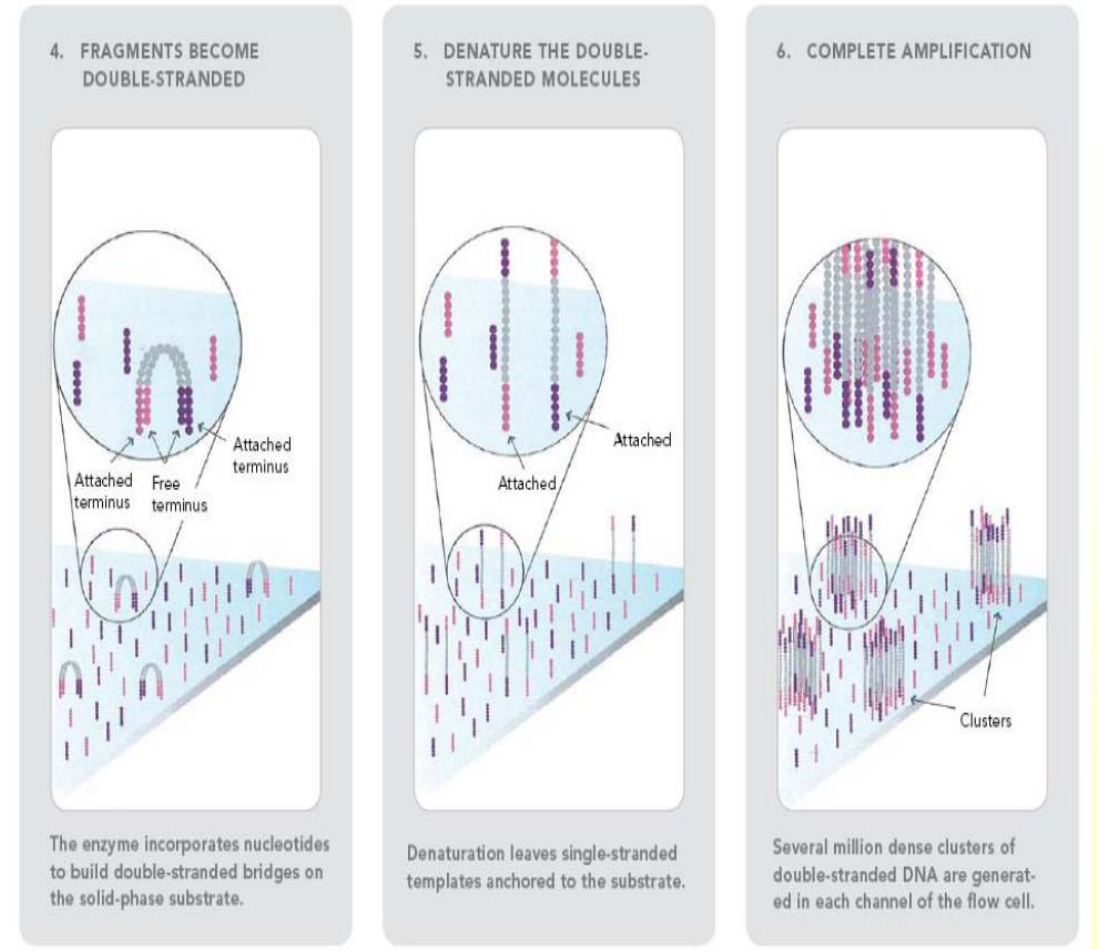


https://kscbioinformatics.wordpress.com/2017/02/13/illumina-sequencing-for-dummies-samples-are-sequenced/

Illumina sequencing

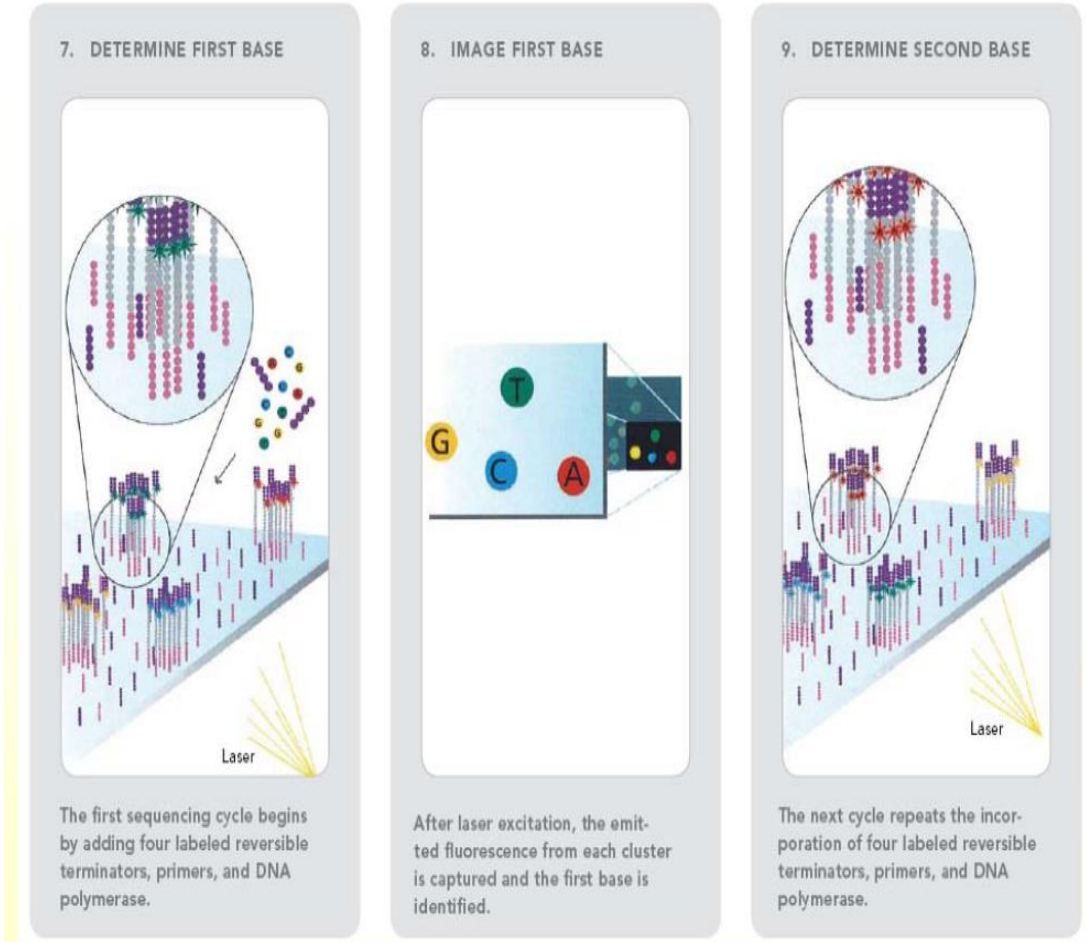


Illumina sequencing



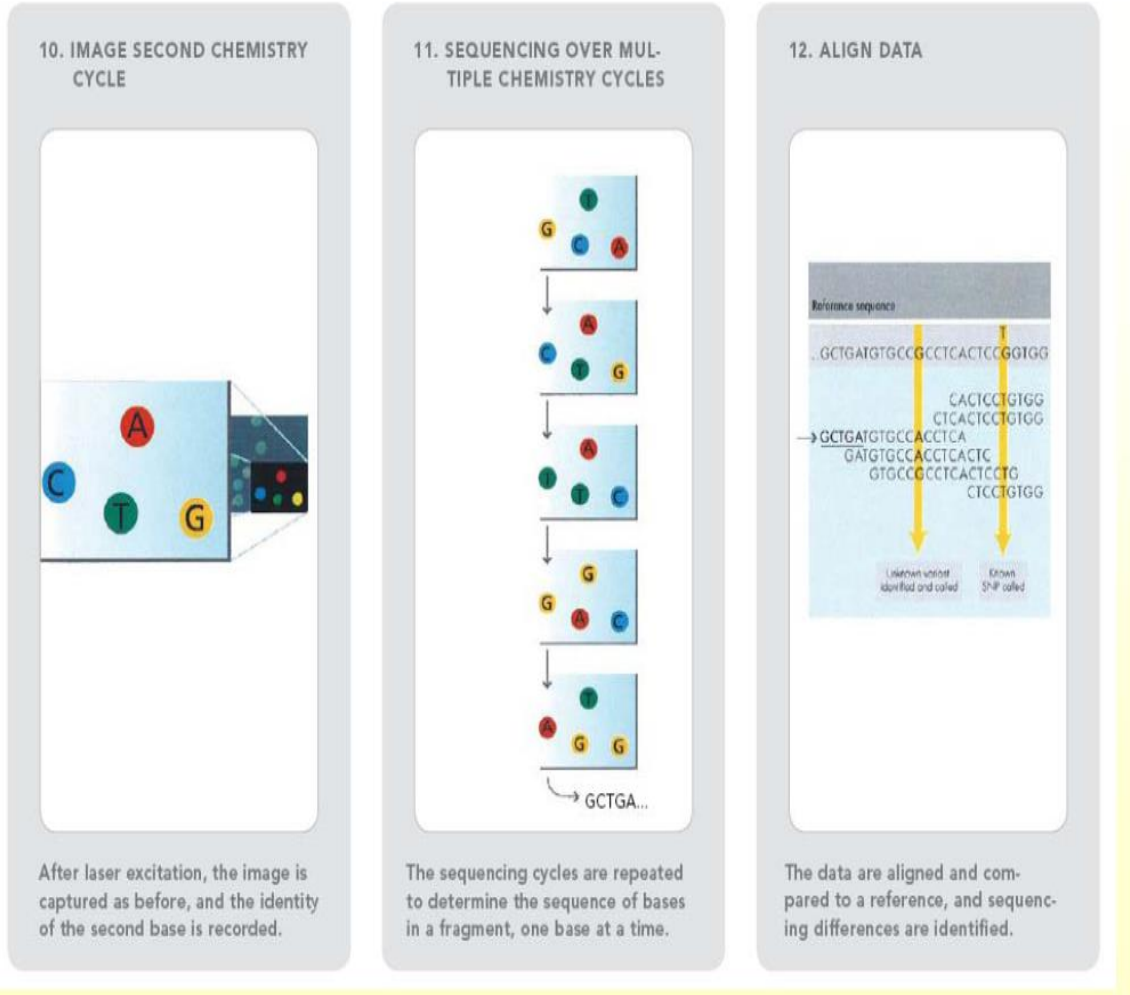
> 1000 copies in ≤ 1 µm; 10^7 clusters per cm^2

Illumina sequencing



Secuenciación con nucleótidos terminadores reversibles

Illumina sequencing



High-throughput sequencing platforms 2nd GS

Roche 454 SEQUENCING

GS-FLX System

illumina

Genome Analyzer IIx

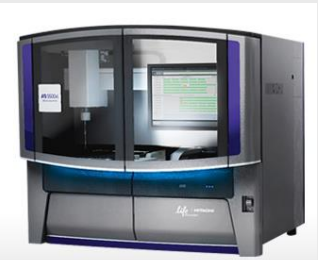
Hiseq 2000/1000

Hiseq 2500/1500

NextSeq

life technologies™

SOLID 3 Plus/4



Genetic Analyzers 5500W Series

Benchtop High-throughput sequencing platforms 2nd GS

Roche 454 SEQUENCING

Roche 454 GS Junior

illumina

MiSeq

iSeq 100

MiniSeq

life technologies™

Ion Proton™ System

Ion PGM™ System

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+lWGS
Max. data quality	>80% > Q30	>75% > Q30	>90% > Q30	>90% > Q30
Run time	12-30 hours	35 hours	11-48 hours	13-44 hours
Maximum output	120 Gb	90 Gb	360 Gb	6000 Gb
Maximum reads per run	400 million	300 million	1.2 billion	20 billion
Maximum read length	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 250 bp

Illumina Benchtop Sequencers

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

Benchtop Sequencer Sheds Light on Ebola Outbreak

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

Read Article ›

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

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

Illumina Production-Scale Sequencers



	NextSeq 550 Series	NextSeq 1000 & 2000	NovaSeq 6000
Popular Applications & Methods	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)			●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●
Exome & Large Panel Sequencing (enrichment-based)	●	●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)	●	●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)	●	●	●
Chromatin Analysis (ATAC-Seq, ChIP-Seq)	●	●	●
Methylation Sequencing	●	●	●
Metagenomic Profiling (shotgun metagenomics, metatranscriptomics)	●	●	●
Cell-Free Sequencing & Liquid Biopsy Analysis	●	●	●

Optimized NGS Sample Tracking and Workflows

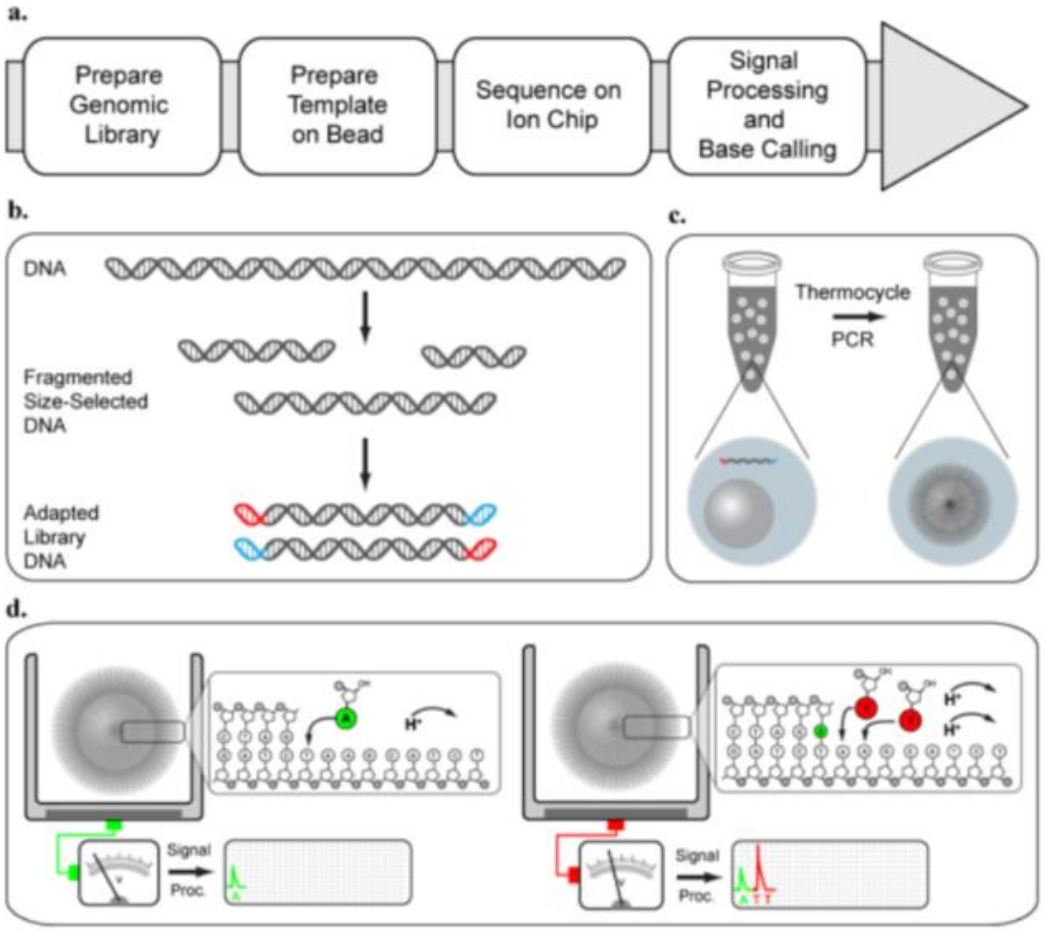
See how a Laboratory Information Management System (LIMS) enabled this large genomics lab to standardize lab procedures and cope with increasing sample volumes from diverse clients.

Read Case Study ›

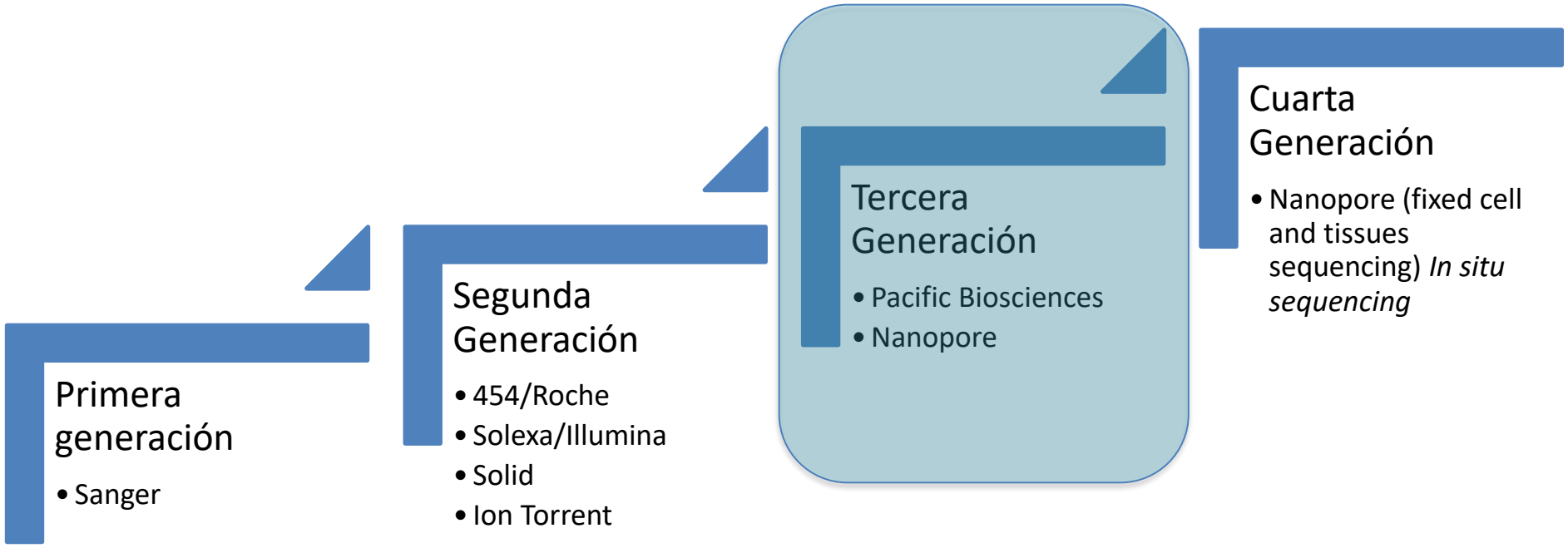
Run Time	12–30 hours	11–48 hours	~13 - 38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)
Maximum Output	120 Gb	330 Gb*	6000 Gb
Maximum Reads Per Run	400 million	1.1 billion*	20 billion
Maximum Read Length	2 × 150 bp	2 × 150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

Ion Torrent PGM

Personal Genome Machine



Secuenciadores



High-throughput single molecule sequencing platforms 3rd GS

Helicos BioSciences Corporation

Helicos Genetic Analysis System

PACIFIC BIOSCIENCES®

PacBio RS II is a Single Molecule, Real-Time (SMRT®) DNA Sequencing System

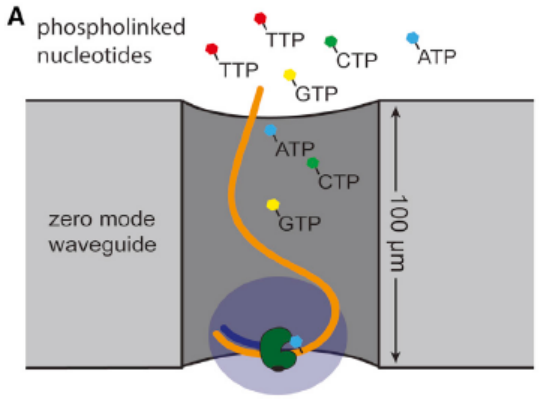
Oxford Nanopore

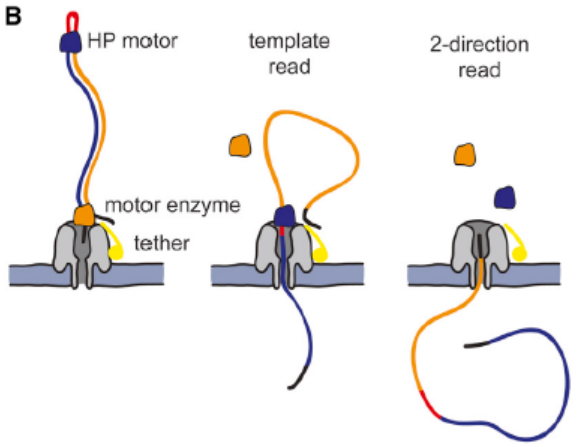
MinIon™ Device

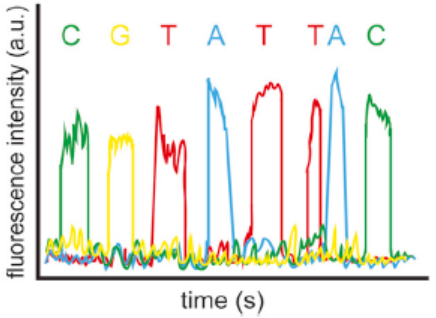
3ª GENERACIÓN: LECTURAS MAS LARGAS Y MOLECULA ÚNICA

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

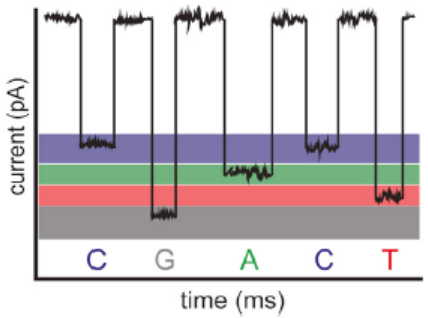
The Third-generation Sequencing Technologies







Pacific Bioscience's SMRT sequencing



Oxford Nanopore's sequencing strategy

Reuter et al., Mol Cell 2015

Single Molecule Sequencing Platforms

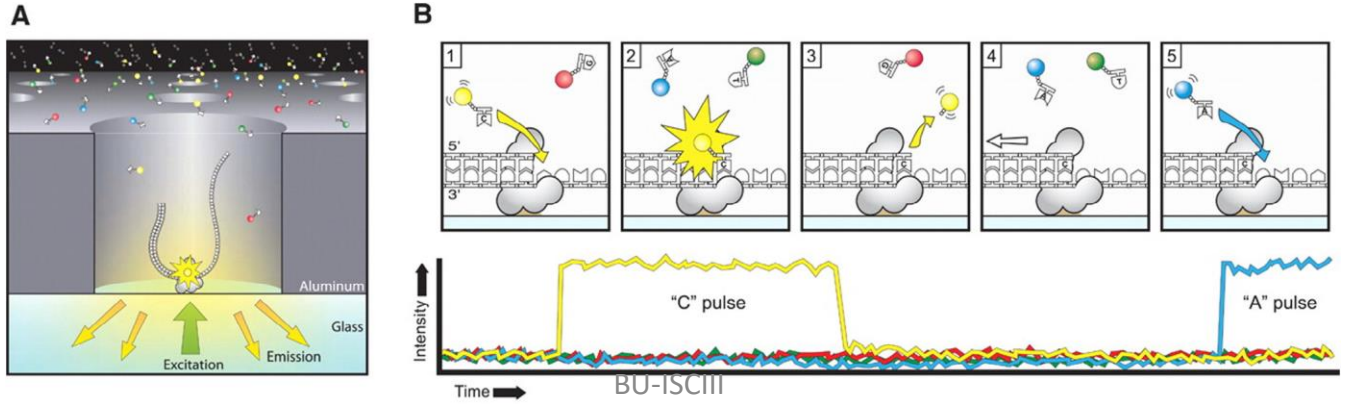
PacBio sequencing and its applications

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

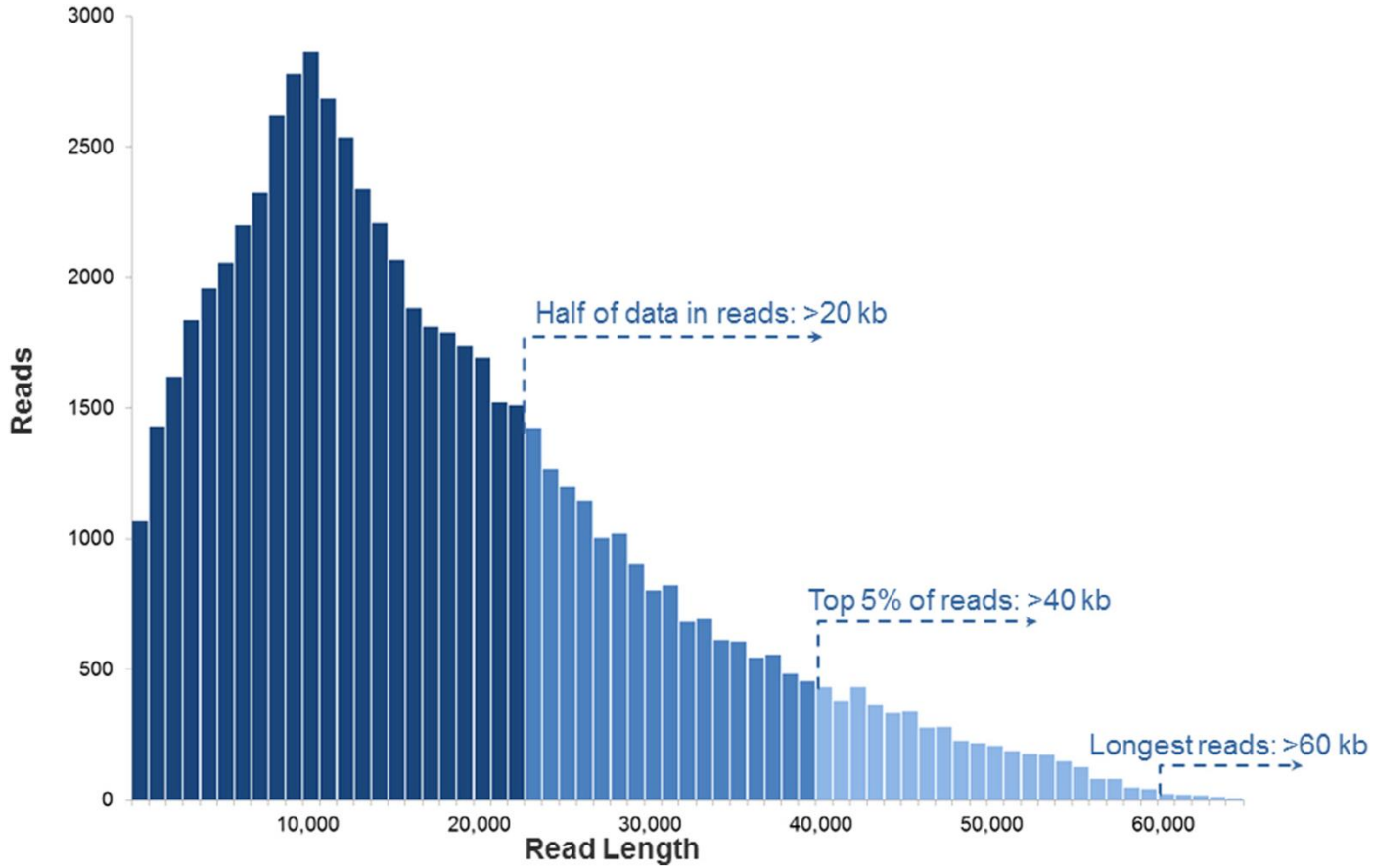
A single SMRT cell: this contains 150000 ZMWs (zero-mode waveguide). A SMRTbell diffuses into a ZMW.
Approx 35000 -75000 ZMWs produce a read in a run lasting 0,5-4h resulting in 0,5-1Gb.



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.
Clostridium autoethanogenum	PacBio	HGAP	2	179×	1	21 fewer contigs than using SGS; no collapsed repeat regions (≥4 using SGS)	[7]
Potentilla micrantha (choloroplast)	PacBio	HGAP, Celera, minimus2, SeqMan	26	320×	1	6 fewer contigs than with Illumina; 100% coverage (Illumina: 90.59%); resolved 187 ambiguous nucleotides in Illumina assembly; unambiguously assigned small differences in two >25 kb inverted repeats	[33]
Escherichia coli	PacBio	PBcR, MHAP, Celera, Quiver	1	85×	1	4.6 CPU hours for genome assembly (10× improvement over BLASR)	[31]
Saccharomyces cerevisiae	PacBio	PBcR, MHAP, Celera	1	117×	21	27 CPU hours for genome assembly (8× improvement over BLASR); improved current reference of telomeres	[31]
Arabidopsis thaliana	PacBio	PBcR, MHAP, Celera	46	144×	38	1896 CPU hours for genome assembly	[31]
Drosophila melanogaster	PacBio	PBcR, MHAP, Celera, Quiver	42	121×	132	1060 CPU hours for genome assembly (593× improvement over BLASR); improved current reference of telomeres	[31]
Homo sapiens (CHM1htert)	PacBio	PBcR, MHAP, Celera	275	54×	3434	262,240 CPU hours for genome assembly; potentially closed 51 gaps in GRCh38; assembled MHC in 2 contigs (60 contigs with Illumina); reconstructed repetitive heterochromatic sequences in telomeres	[31]
Homo sapiens (CHM1tert)	PacBio	BLASR, Celera, Quiver	243	41×	N/A (local assembly)	Closed 50 gaps and extended into 40 additional gaps in GRCh37; added over 1 Mb of novel sequence to the genome; identified 26,079 indels at least 50 bp in length; cataloged 47,238 SV breakpoints	[32]
Melopsittacus undulatus	Hybrid	PBcR, Celera	3	5.5× PacBio + 15.4× 454 = 3.83× corrected	15,328	1st assembly of >1 Gb parrot genome; N50 = 93,069	[34]
Vibrio cholerae	Hybrid	BLASR, Bambus, AHA	195	200× PacBio + 28× Illumina + 22× 454	2	No N's in contigs; 99.99% consensus accuracy; N50 = 3.01 Mb	[30]
Helicobacter pylori	PacBio	HGAP, Quiver, PGAP	8 per strain	446.5× average among strains	1 per strain	1 complete contig for each of 8 strains; methylation analysis associated motifs with genotypes of virulence factors	[35]

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

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

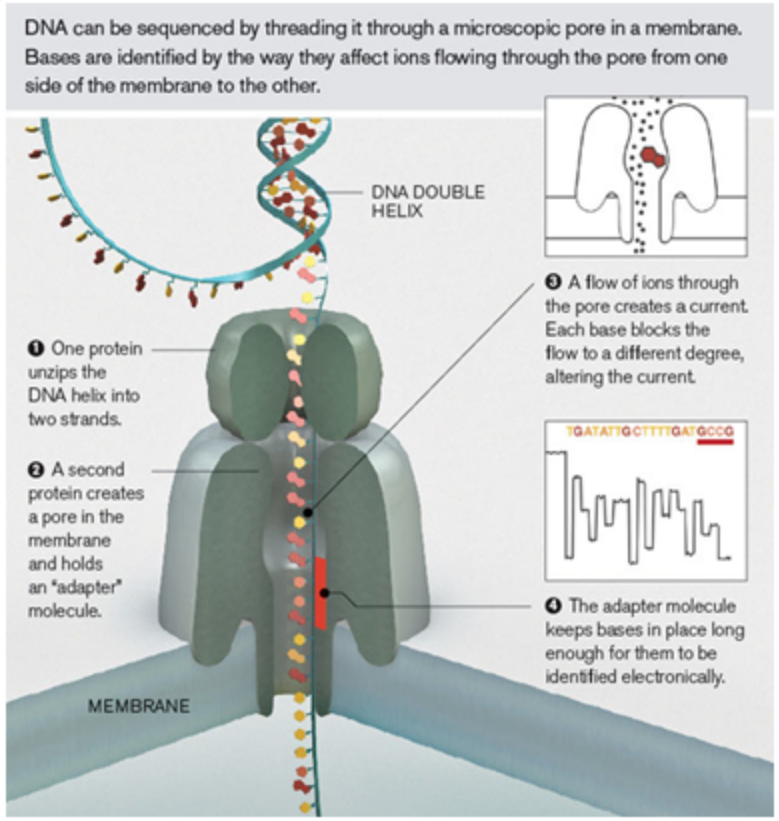
Produced highly-contiguous assemblies of bacterial and eukaryotic genomes

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

Nanopore sequencing

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

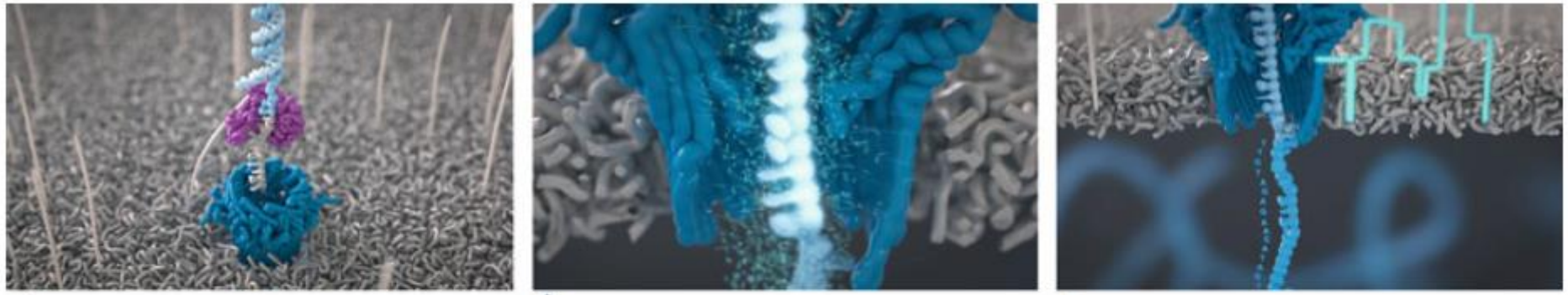
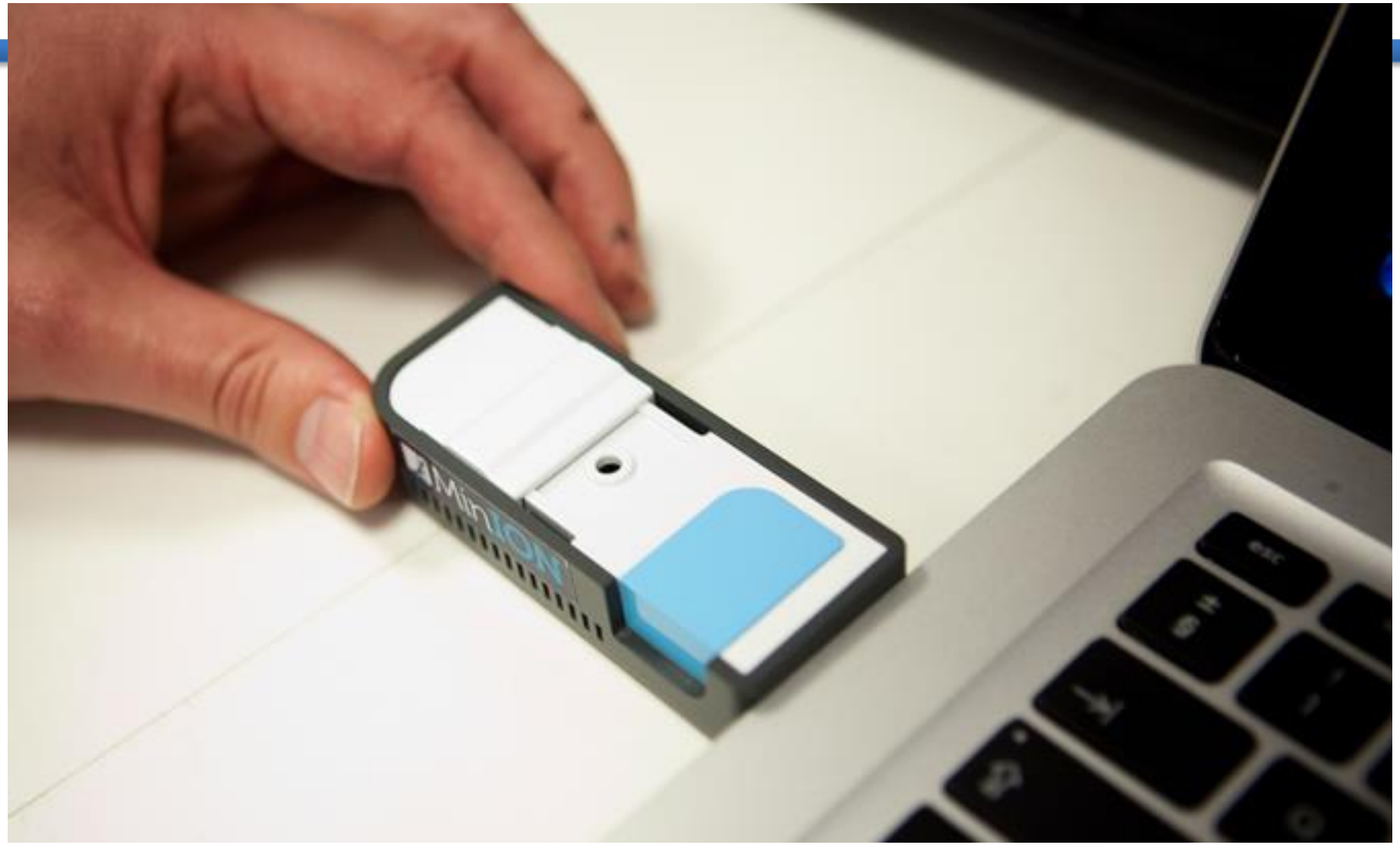


Fig. 1 The principle of nanopore sequencing. (**a**) A protein nanopore (blue) is imbedded into an electronically resistive lipid membrane (grey), before adapted DNA libraries containing a motor protein (purple) are introduced, and the motor feeds DNA progressively through the pore. (**b**) An ionic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

MinIon, OXFORD NANOPORE



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

22/05/2023

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

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

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

Oxford Nanopore Technologies

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:

Library preparation

Oxford Nanopore has developed VolTRAX – a small device designed to perform library preparation automatically, so that a user can get a biological sample ready for analysis, hands-free. VolTRAX is designed as an alternative to a range of lab equipment, to allow consistent and varied, automated library prep options.

VolTRAX V2 Starter Pack

$8,000.00

VolTRAX V2 is designed to automate all laboratory processes associated with Nanopore Sequencing from sample extraction to library preparation.

MinIT, Analysis

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

MinIT Specifications:

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

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

SmidgION, Mobile analysis

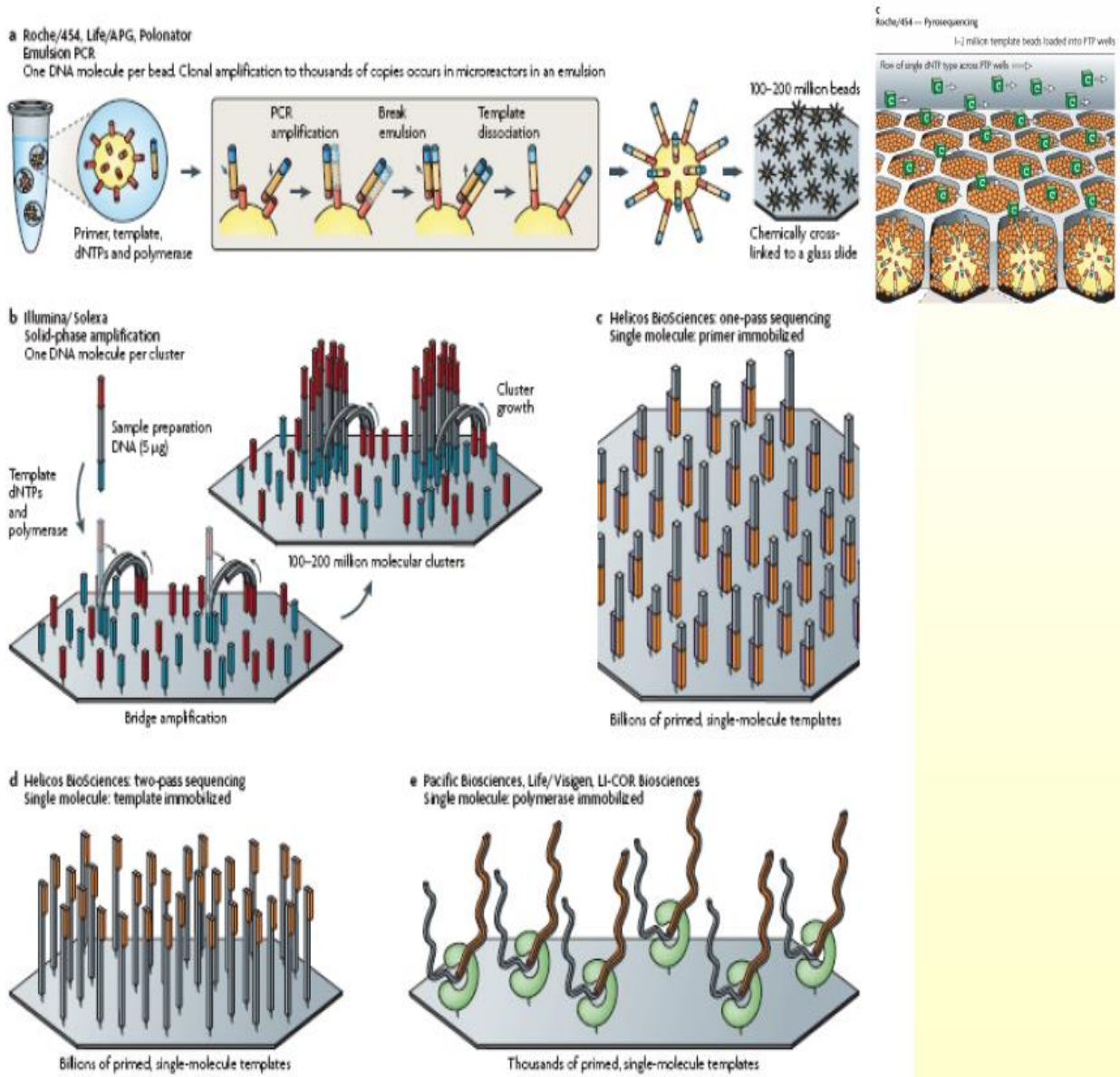
Oxford Nanopore has now started developing an even smaller device, SmidgION.

potential applications may include remote monitoring of pathogens in a breakout or infectious disease; the on-site analysis of environmental samples such as water/metagenomics samples, real time species ID for analysis of food, timber, wildlife or even unknown samples; field-based analysis of agricultural environments, and much more.

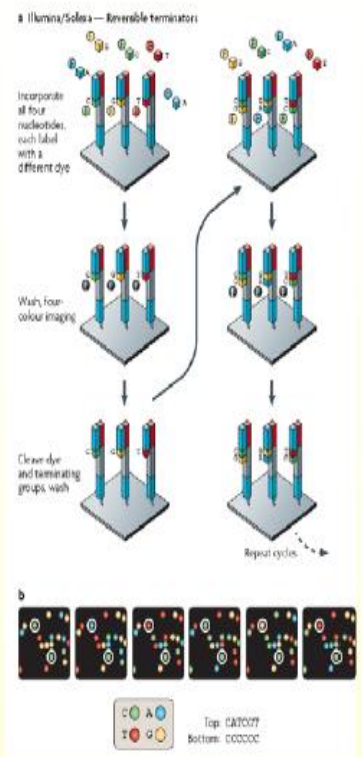
Long-read Sequencing Platforms characteristics

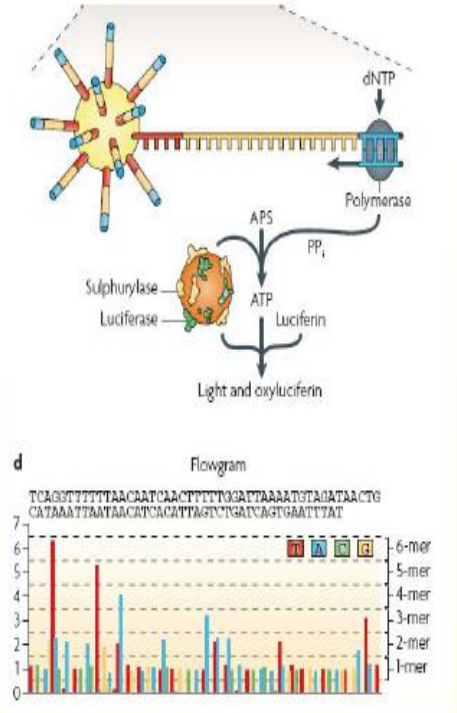
Company	Pacific Biosciences			Oxford Nanopore			
System Platform	Sequel	Sequel II	Sequel IIe	Flongle	MinION	GridION	PromethION
Sequencing Principle	PacBio Single Molecule Sequencing			Nanopore Single molecule Sequencing			
Detection	Fluorescent			Electrical Conductivity			
Applications	Whole genome *de novo* assembly, variant detection, structural variation detection, full length transcript sequencing, targeted/amplicon sequencing, metagenomics sequencing			DNA, amplicons, cDNA, Direct RNA sequencing			
Maximum Read length (bases)	300 kb			Longest read so far: > 4 Mb			
Flow cells/device	12 SMRT Cells 1M can be used at a time, and 8 SMRT Cell 8M can be used serially			1 (126 channels per flow cell)	1 (512 channels per flow cell)	5 (512 channels per flow cell)	24 or 48 (3000 channels per flow cell)
Output (per flow cell)	75 Gb	600 Gb		1 - 2 Gb[a]	10 - 30 - 50 Gb[a]		100 - 200 - 300 Gb[a]
Sequencing Run time	Up to 20 hr	Up to 30 hr		1 min - 16 hr	1 min - 72 hr		
Accuracy/Quality Score	Number of HiFi Reads >99% Accuracy: Up to 5,000,000 reads	Number of HiFi Reads >99% Accuracy: Up to 4,000,000 reads		**Single Molecule:** R9 modal Accuracy >98.3%, R10 modal Accuracy >97.5%. New chemistry Accuracy >99% (coming soon) **Consensus:** R9.4.1: Current best Q45 (>99.99%) R10: Current Best Q50 (99.999%)			
Equipment Cost (USD)	approximately $525,000			$1,460 (12 flow cells included)	$9,300	$69,955	24 flow cells: $335,455 48 flow cells: $530,000
Dimensions	92.7 x 91.4 x 167.6 cm			105 x 23 x 8 mm	Mk1b: 105 x 23 x 33 mm; Mk1c: 140 x 30 x 116 mm	365 x 220 x 360 mm	Sequencer: 590 x 190 x 430 mm; Data Acquisition unit: 178 x 440 x 470 mm
Weight	362 kg			20 g	Mk1b: 87 g Mk1c: 450 g	11 kg	Sequencer: 28 kg; Data Acquisition unit: 25 kg
Advantages	Very long reads can help resolve ambiguities; no DNA amplification required, comparatively faster turnaround time			Fast-sequencing; Small instrument footprint; Portability; Real-time data analysis			
Disadvantages	Sequencing equipment is expensive, which could be cost-prohibitive for smaller clinical laboratories, large footprint of the equipment, historically higher error rate (continues to improve)			Historically higher error rate (continues to improve)			

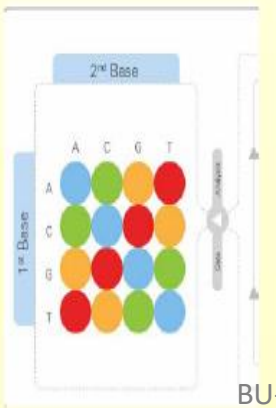
ESTRATEGIAS DE INMOVILIZACIÓN

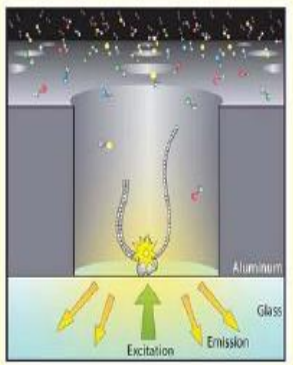


ESTRATEGIAS DE LECTURA









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	$1.0–1.5 \times 10^4$ on average	13	$3.5–7.5 \times 10^4$	0.5–4 h	0.40–0.80	[5,12,15]
Oxford Nanopore MinION	3rd	$2–5 \times 10^3$ on average	38	$1.1–4.7 \times 10^4$	50 h	6.44–17.90	[22,23]

Characteristics, strengths and weaknesses of commonly used sequencing platforms

Table 2
Characteristics, strengths and weaknesses of commonly used sequencing platforms

Platform \ Instrument	Throughput range (Gb)[a]	Read length (bp)	Strength	Weakness
Sanger sequencing				
ABI 3500/3730	0.0003	Up to 1 kb	Read accuracy and length	Cost and throughput
Illumina				
MiniSeq	1.7–7.5	1×75 to ×150	Low initial investment	Run and read length
MiSeq	0.3–15	1×36 to 2×300	Read length, scalability	Run length
NextSeq	10–120	1×75 to 2×150	Throughput	Run and read length
HiSeq (2500)	10–1000	×50 to ×250	Read accuracy, throughput,	High initial investment, run
NovaSeq 5000/6000	2000–6000	2×50 to ×150	Read accuracy, throughput	High initial investment, run
IonTorrent				
PGM	0.08–2	Up to 400	Read length, speed	Throughput, homopolymers[c]
S5	0.6–15	Up to 400	Read length, speed,	Homopolymers[c]
Proton	10–15	Up to 200	Speed, throughput	Homopolymers[c]
Pacific BioSciences				
PacBio RSII	0.5–1[b]	Up to 60 kb	Read length, speed (Average 10 kb, N50 20 kb)	High error rate and initial
Sequel	5–10[b]	Up to 60 kb	Read length, speed (Average 10 kb, N50 20 kb)	High error rate
Oxford Nanopore				
MInION	0.1–1	Up to 100 kb	Read length, portability	High error rate, run length,

[a] The throughput ranges are determined by available kits and run modes on a per run basis. As an example of a 15-GB throughput, thirty-five 5-MB genomes can be sequenced to a minimum coverage of 40× on the Illumina MiSeq using the v3 600 cycle chemistry.
[b] Per one single-molecule real-time cell.
[c] Results in increased error rate (increased proportion of reads containing errors among all reads) which in turn results in false-positive variant calling.

Besser et al., Clin Micr Infect, 2018

Comparison of various high-performing sequencing instruments

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

Advantages and disadvantages of sequencing generations

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	High sequencing error rate 10–15% in the PacBio and 5–20% in the ONT Fresh DNA requires for ensuring quality of ultralong reads Database systems and algorithms/tools are rare for analyzing 3rd and 4th GS data
Fourth generation	Advantages of 3rdGS+ 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	· Higher sequence fidelity · Cheap · Can sequence fragmented DNA	· Not able to resolve structural variants, phasing alleles or distinguish highly homologous genomic regions · Unable to provide coverage of some repetitive regions
Long-read sequencing	· Able to sequence genetic regions that are difficult to characterize with short-read seq due to repeat sequences · Able to resolve structural rearrangements or homologous regions · Able to read through an entire RNA transcript to determine the specific isoform · Assists *de novo* genome assembly	· Lower per read accuracy · Bioinformatic challenges, caused by coverage biases, high error rates in base allocation, scalability and limited availability of appropriate pipelines

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

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

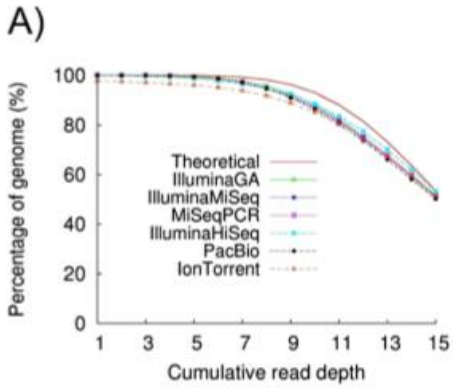
NGS PLATFORMS, main characteristics

- 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 y evolución de la tecnología
- Formato fichero salida
- Software dedicado, universal fastq

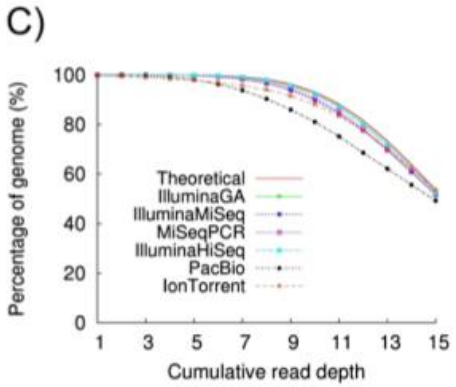
Uniformidad de cobertura a lo largo del genoma

%GC



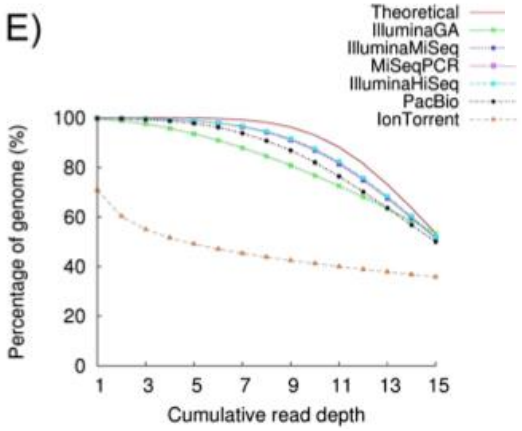
Bordetella pertussis

%GC



Staphylococcus aureus

%AT



Plasmodium falciparum

Quail et al., BMC Genomics 2012, 13:341

nature biotechnology
FOCUS | ARTICLES
https://doi.org/10.1038/s41587-021-01049-5
Check for updates

Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study

Foox et al., Nature Biotecnology 2021

Interlaboratory and intralaboratory DNA-seq replicates of

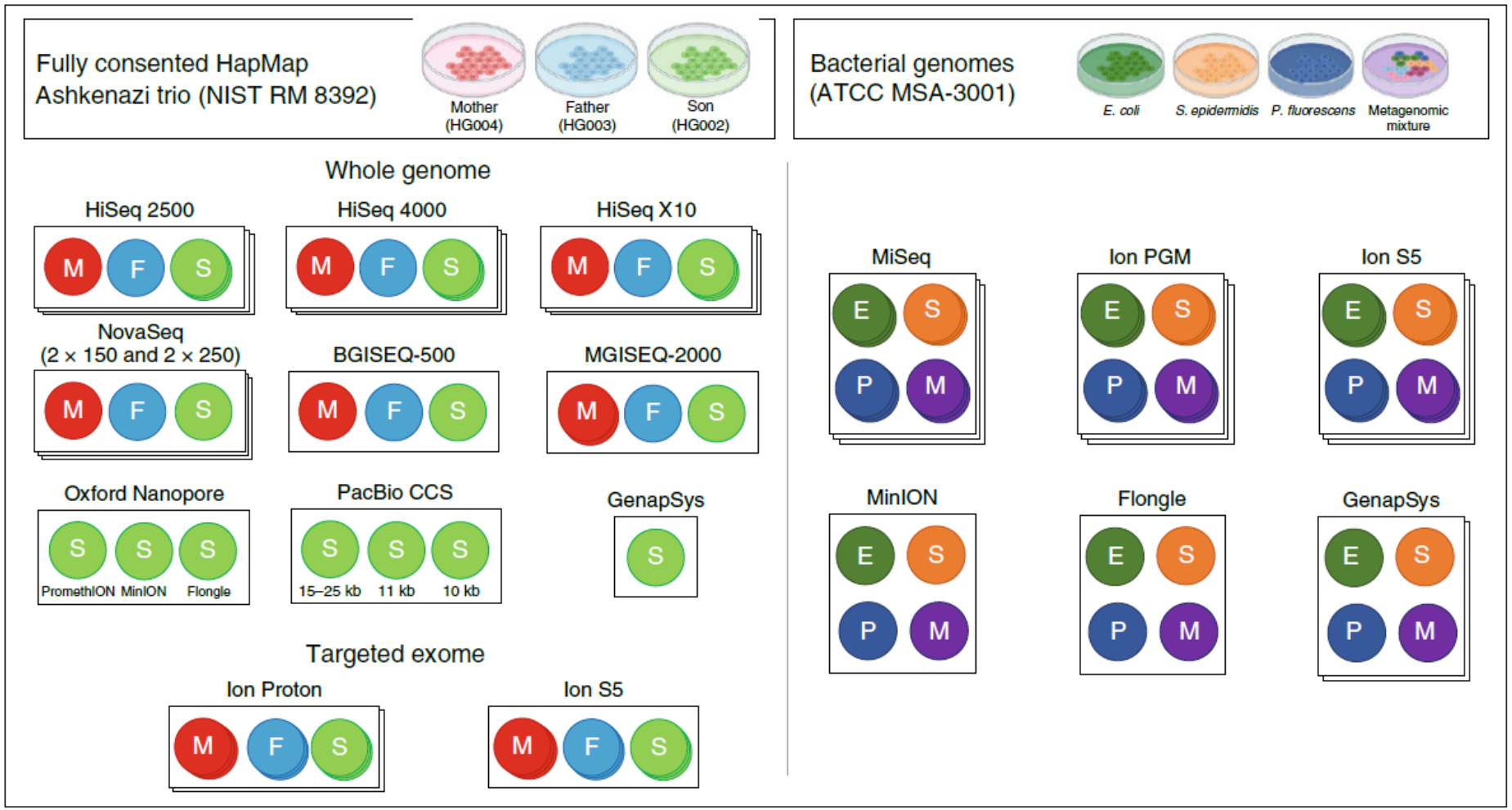
- The Ashkenazi trio (NIST reference material)
- Three individual bacterial strains
- Metagenomic mixture of ten bacterial species

TO STUDY THE EFFECTS OF GC CONTENT AND LIBRARY COMPLEXITY.

These replicates were generated across

- five Illumina platforms
- Three ThermoFisher Ion Torrent platforms, the BGISEQ-500 and MGISEQ-2000 platforms, the GenapSys GS111 Platform
- Oxford Nanopore Technologies (ONT) Flongle, MinION and PromethION flow cells
- Publicly available PacBio CCS data for HG002.

Foox et al., Nature Biotecnology 2021



Box 1 | Best-practice recommendations

We summarize below ten best-practice recommendations for the community based on our analysis.

1. Mapping efficiency rates are both platform specific and species specific. Illumina instruments are most comparable to one another. BGISEQ-500 and GenapSys GS111 instruments, providing the shortest read lengths of platforms in this study, return the lowest uniquely mapping rate and highest multi-mapping rate. BGI/MGISEQ libraries have the lowest duplicate read rate. PacBio CCS datasets have the highest rates of unique mapping and lowest non-mapping rate. Short fragments in ONT data bring down overall mapping efficiency.
2. Alignments (BAM files) can be normalized by calculating mean autosomal coverage using mosdepth and then downsampling using Picard DownsampleSam. However, even within normalized data, coverage dramatically varies within repetitive and low-complexity regions, even among replicates sequenced on the same instrument. Long-read technologies provide the highest coverage within these genomic regions. For short-read platforms, HiSeq 4000 and X10 provide the most consistent, highest coverage.
3. Sequencing error can be calculated with BBMap reformat.sh and comparing mismatch histogram tables. All instruments have some level of sequencing error, ranging from 0.1% up to 20% in poorly defined satellite repeat regions. BGI/MGISEQ provide the lowest sequencing error rates among short-read technologies. PacBio CCS provides the lowest error rate out of all technologies. Although the error rate is highest of all platforms, ONT instruments perform highly consistently from the smallest throughput (Flongle) flow cell to the largest (PromethION R9.4).
4. Mismatch rates are elevated in areas of high and low GC content and by base to a lesser extent. Errors are more frequent in regions with larger repeat sizes of homopolymers and lower entropy of STRs, except for ONT instruments, which show flat (although currently still high) error rates, irrespective of sequence content. PacBio CCS has the lowest error rate in these contexts, whereas GenapSys has elevated STR error rates compared with other short-read platforms.
5. For calling known variants, DeepVariant is the most sensitive and precise software, although this software is trained on immortalized B lymphocyte cell line data and may be overfitted. Strelka2 is as precise as DeepVariant, whereas GATK HaplotypeCaller is as sensitive. Sentieon Haplotyper is very nearly as sensitive as GATK HaplotypeCaller, but is by far the most computationally efficient. Default parameters may be used for each caller.
6. Sensitivity and specificity of variant detection can be assessed with RTG vcfeval. Among known variants, true positives in L1/L2/STR regions are recalled the most easily, whereas variants in simple repeats and low-complexity regions are the hardest to capture. Read length makes an impact on the ability to call true positives because data with shorter read lengths (HiSeq 2500 2×125 bp and BGISEQ-500 2×100 bp) capture the lowest proportion of true positives across RepeatMasker regions examined.
7. The length of INDELs captured by each platform can be evaluated using RTG vcf-stats with the –allele-lengths flag. INDEL detection is highly platform specific, in particular for insertions (deletions are more comparable among platforms). ONT instruments capture the lowest proportion, followed by BGISEQ-500, Illumina HiSeq platforms and then PacBio CCS. The NovaSeq 6000 using 2×250-bp read chemistry is the most robust instrument for capturing known INDELs.
8. SV calling consistency is most impacted by the variant caller used. This can be evaluated by calling SVs with Delly, Manta and Lumpy, and then consolidating calls with SURVIVOR. Sequencing instrument is the second highest source of variability, followed by within-instrument replicates. The majority of unique SVs are likely due to sequencing artifacts and can be considered false negatives.
9. A genome-wide distribution of roughly 20,000 SVs is common with a given genome, which is slightly higher than previous estimates and benefits from longer reads. Within those, the majority (70%) will be called as deletions, followed by translocations (14%), insertions (6%), duplications (5%) and inversions (4%). No significant clustering of SVs is seen within the genomes examined in this study, indicating that overlapping SVs between replicates or instruments can be considered true positives, rather than mapping artifacts.
10. In mixed metagenomic samples, the rate of mapping is linked to the GC content of the reference genome for each taxon. High- and low-GC content taxa tend to be underrepresented in reference-based alignment. This can be determined using mosdepth with the -F 3844 flag to assess the number of reads uniquely mapping to each genome within the mixed reference set.

Foox et al., Nature Biotecnology 2021

CONCLUSIONS **Foox et al., Nature Biotecnology 2021**

- Among short-read instruments, **HiSeq 4000 and X10** provided the **most consistent, highest genome coverage**, while **BGI/MGISEQ** provided the **lowest sequencing error rates**

- The long-read instrument **PacBio CCS** had the **highest reference-based mapping rate and lowest non-mapping rate.**

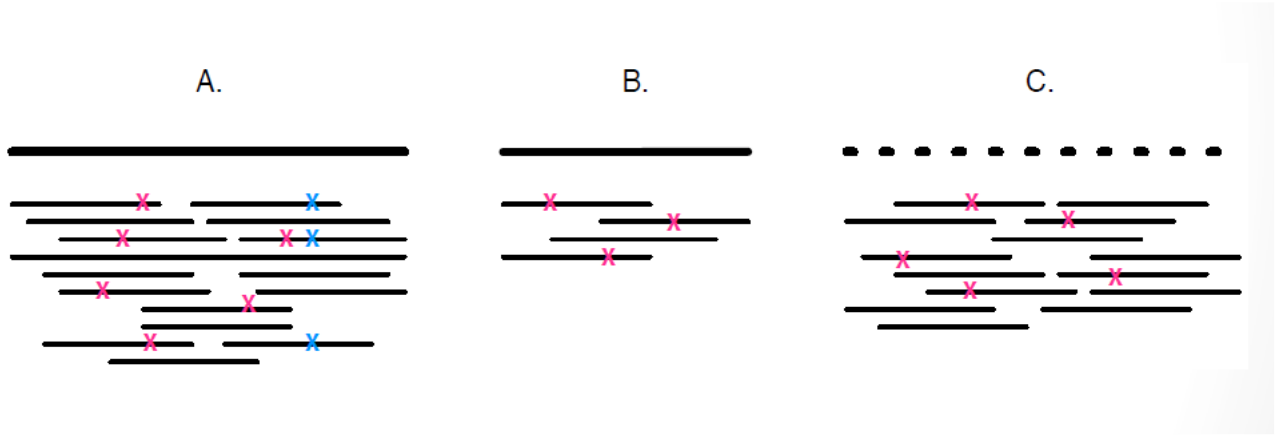
- The two long-read platforms **PacBio CCS and PromethION/MinION** showed the **best sequence mapping in repeat-rich areas and across homopolymers**.

- **NovaSeq 6000 using 2 × 250-bp** read chemistry was the most robust instrument for **capturing known insertion/deletion events.**

Algunos conceptos en secuenciación

Básicamente tres problemas

Resecuenciación, Conteo y ensamblado

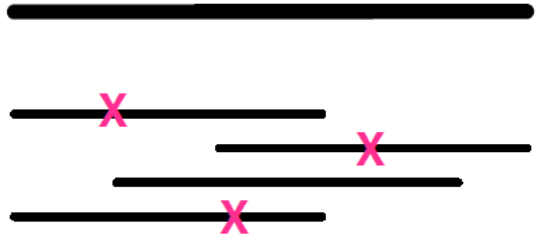


Resecuenciación

Conocemos el genoma, genoma de referencia, y queremos identificar variaciones (azul), en un background de errores (rosa)

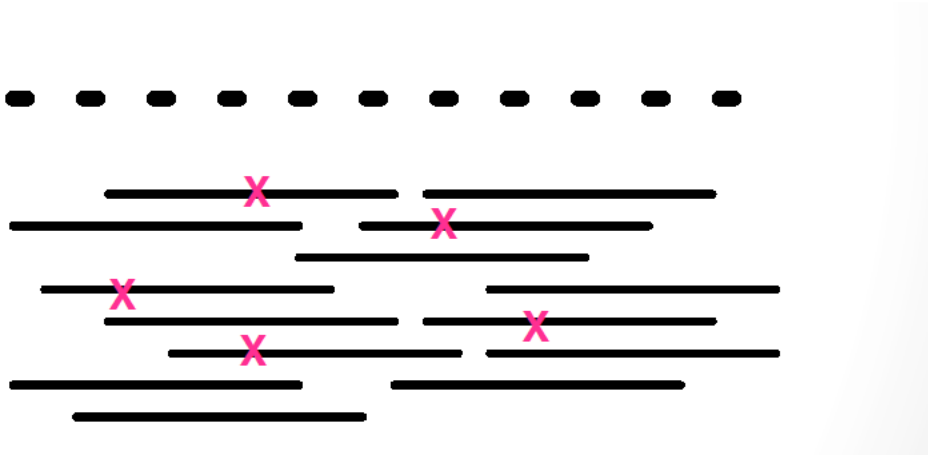
Conteo

Número de lecturas de un gen (amplicón) o mRNA (RNAseq). Equivalente a expresión en Microarrays.



Ensamblado

No hay genoma de referencia y lo construimos de novo



Sequencing terms

Breadth of coverage

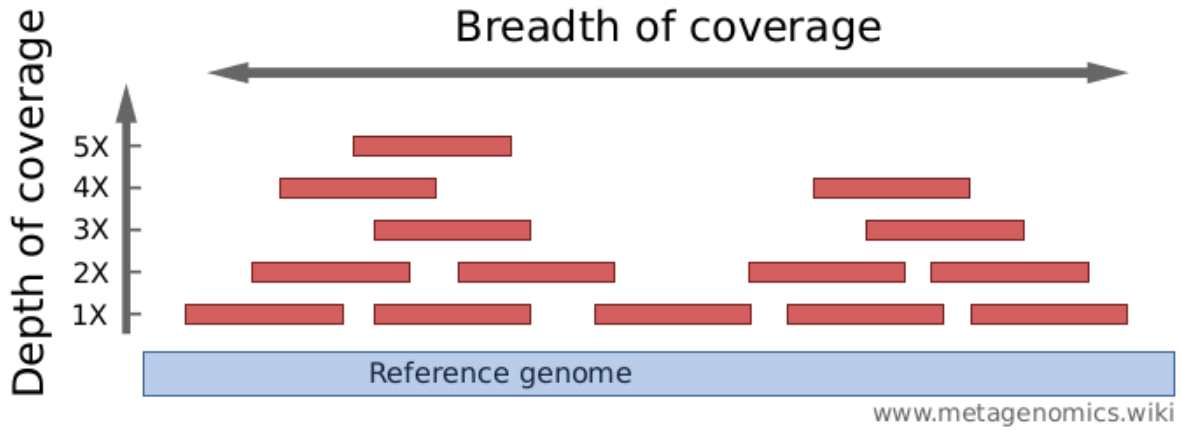
How much of a genome is "covered" by short reads? Are there regions that are not covered, even not by a single read?

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

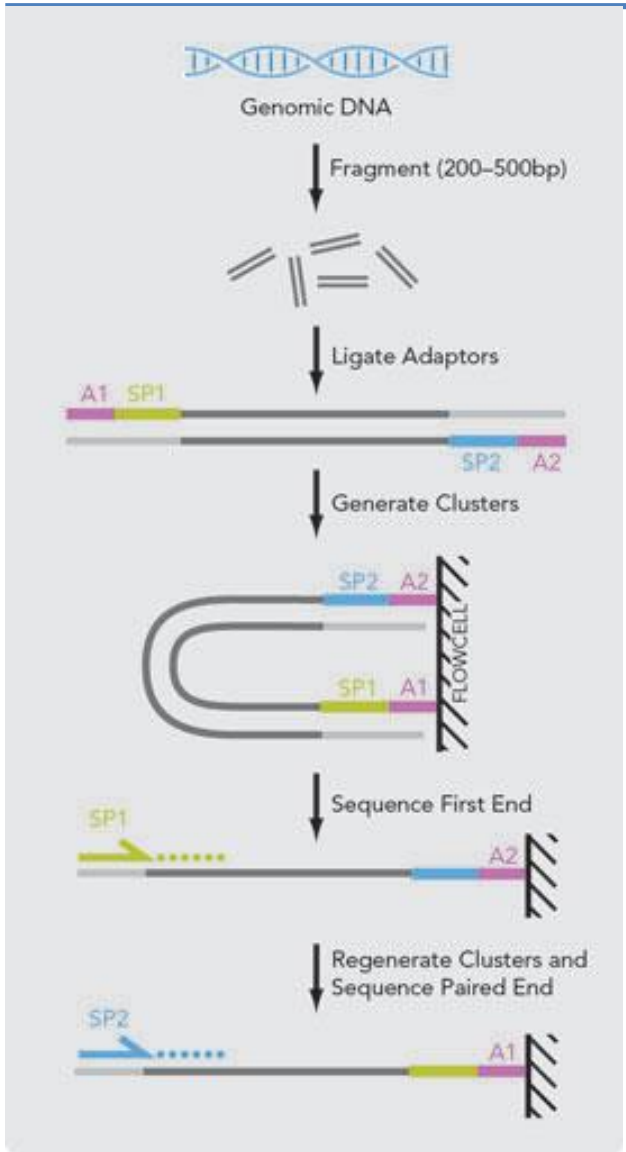
Depth of coverage

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

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

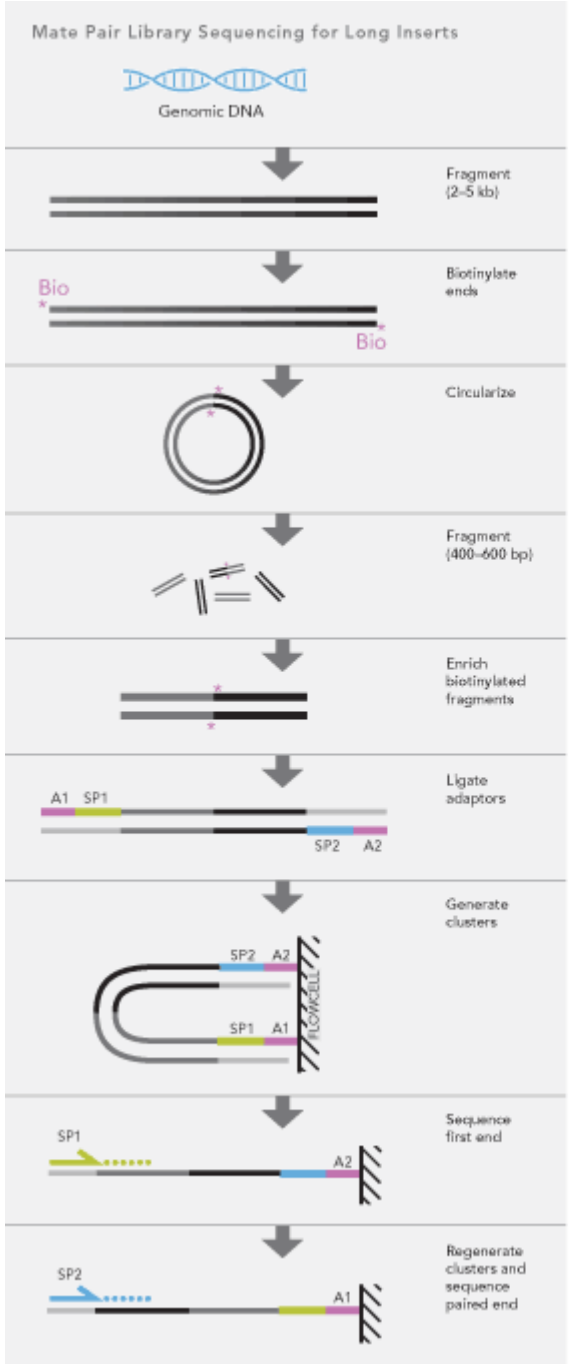


Que es Paired-end?



Secuenciación de un fragmento (bp)

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



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.

Que es Mate-pair?

Secuenciación de dos fragmentos separados kb.

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

Resumen

Número de lecturas por muestra (cobertura y profundidad de cobertura)
Número de muestras que secuencio simultáneamente

Elegir secuenciador: Número de bases que lee un secuenciador y longitud de lectura, single o paired-end (kit de secuenciación)

Cobertura es importante para la llamada a variantes, RNAseq. Plataforma con mayor rendimiento Illumina

La longitud de las lecturas es importante para el ensamblado
PacBio y MinIon mayor longitud de lecturas (corrección de errores con Illumina)

Coverage and Read Depth Recommendations by Sequencing Application

Table 1: Coverage and Read Recommendations by Application

Category	Detection or Application	Recommended Coverage (x) or Reads (millions)	References
Whole genome sequencing	Homozygous SNVs	15x	Bentley et al., 2008
	Heterozygous SNVs	33x	Bentley et al., 2008
	INDELs	60x	Feng et al., 2014
	Genotype calls	35x	Ajay et al., 2011
	CNV	1-8x	Xie et al., 2009; Medvedev at al., 2010
Whole exome sequencing	Homozygous SNVs	100x (3x local depth)	Clark et al., 2011; Meynert et al., 2013
	Heterozygous SNVs	100x (13x local depth)	Clark et al., 2011; Meynert et al., 2013
	INDELs	not recommended	Feng et al., 2014
Transcriptome Sequencing	Differential expression profiling	10-25M	Liu Y. et al., 2014; ENCODE 2011 RNA-Seq
	Alternative splicing	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	Allele specific expression	50-100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq
	De novo assembly	>100M	Liu Y. et al., 2013; ENCODE 2011 RNA-Seq

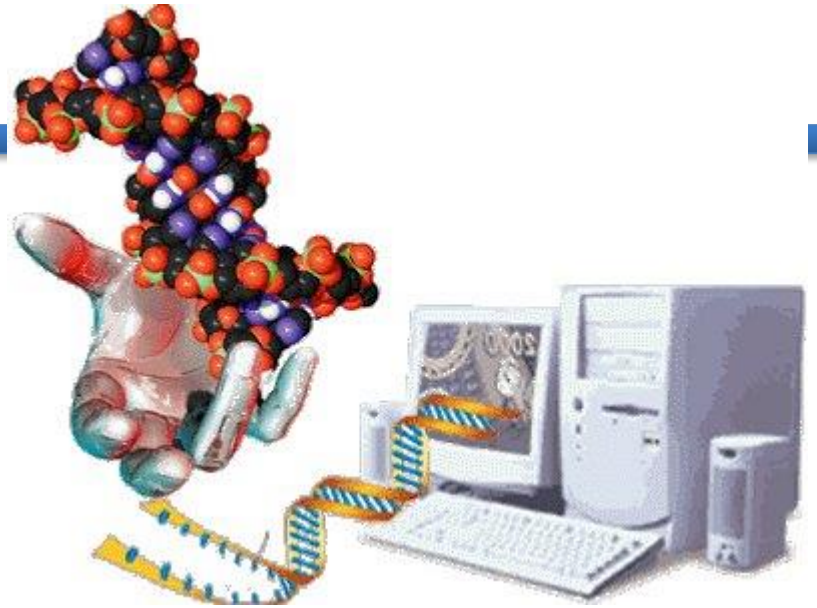
https://genohub.com/recommended-sequencing-coverage-by-application/

Coverage and Read Depth Recommendations by Sequencing Application

DNA Methylation Sequencing	CAP-Seq	>20M	Long, H.K. et al., 2013
	MeDIP-Seq	60M	Taiwo, O. et al., 2012
	RRBS (Reduced Representation Bisulfite Sequencing)	10X	ENCODE 2011 Genome
	Bisulfite-Seq	5-15X; 30X	Ziller, M.J et al., 2015; Epigenomics Road Map
RNA-Target-Based Sequencing	CLIP-Seq	10-40M	Cho J. et al., 2012; Eom T. et al., 2013; Sugimoto Y. et al., 2012
	iCLIP	5-15M	Sugimoto Y. et al., 2012; Rogelj B. et al., 2012
	PAR-CLIP	5-15M	Rogelj B. et al., 2012
	RIP-Seq	5-20M	Lu Z. et al., 2014
Small RNA (microRNA) Sequencing	Differential Expression	~1-2M	Metpally RPR et al., 2013; Campbell et al., 2015
	Discovery	~5-8M	Metpally RPR et al., 2013; Campbell et al., 2015

https://genohub.com/recommended-sequencing-coverage-by-application/

Gracias por la atención
Preguntas ???



Isabel Cuesta

Unidad de Bioinformática – Unidades Científico Técnicas - ISCIII

isabel.cuesta@isciii.es