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

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

17-28 Mayo 2021, 8ª 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

17-28 Mayo 2021, 8ª Edición
Programa Formación Continua, ISCIII

INDICE

- Unidad de Bioinformática
  Servicios ofertados

- Evolución de la secuenciación

- Plataformas de secuenciación masiva (NGS o HTS)

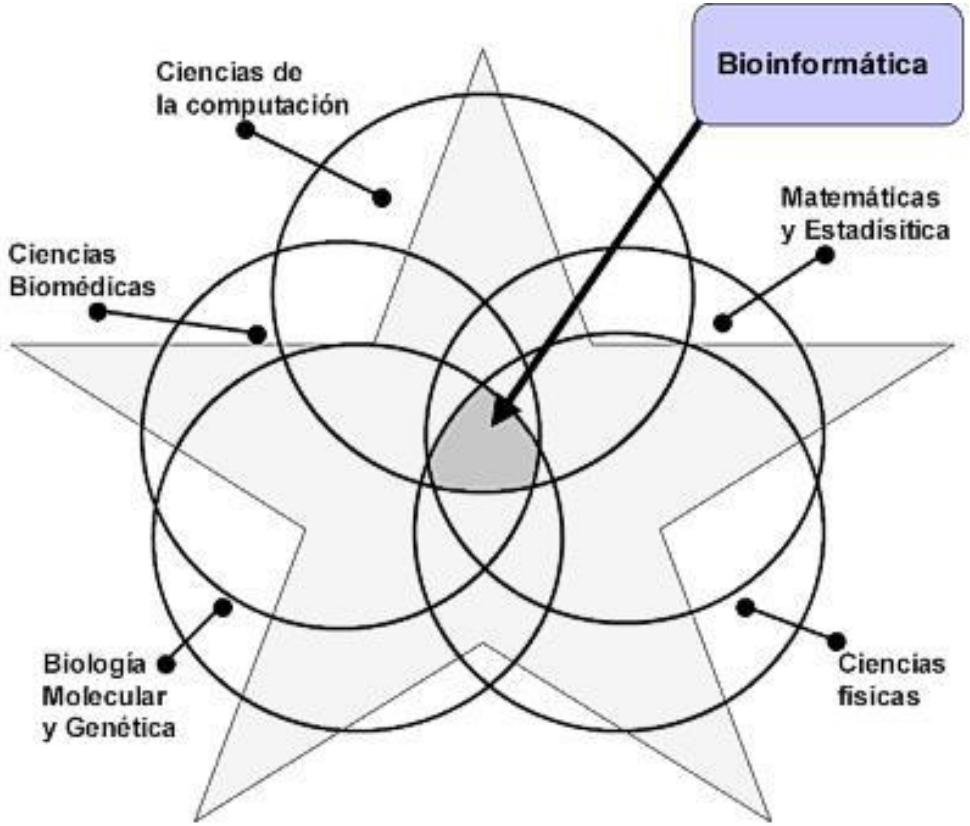
Qué es la Bioinformática?

PROBLEMAS BIOLÓGICOS

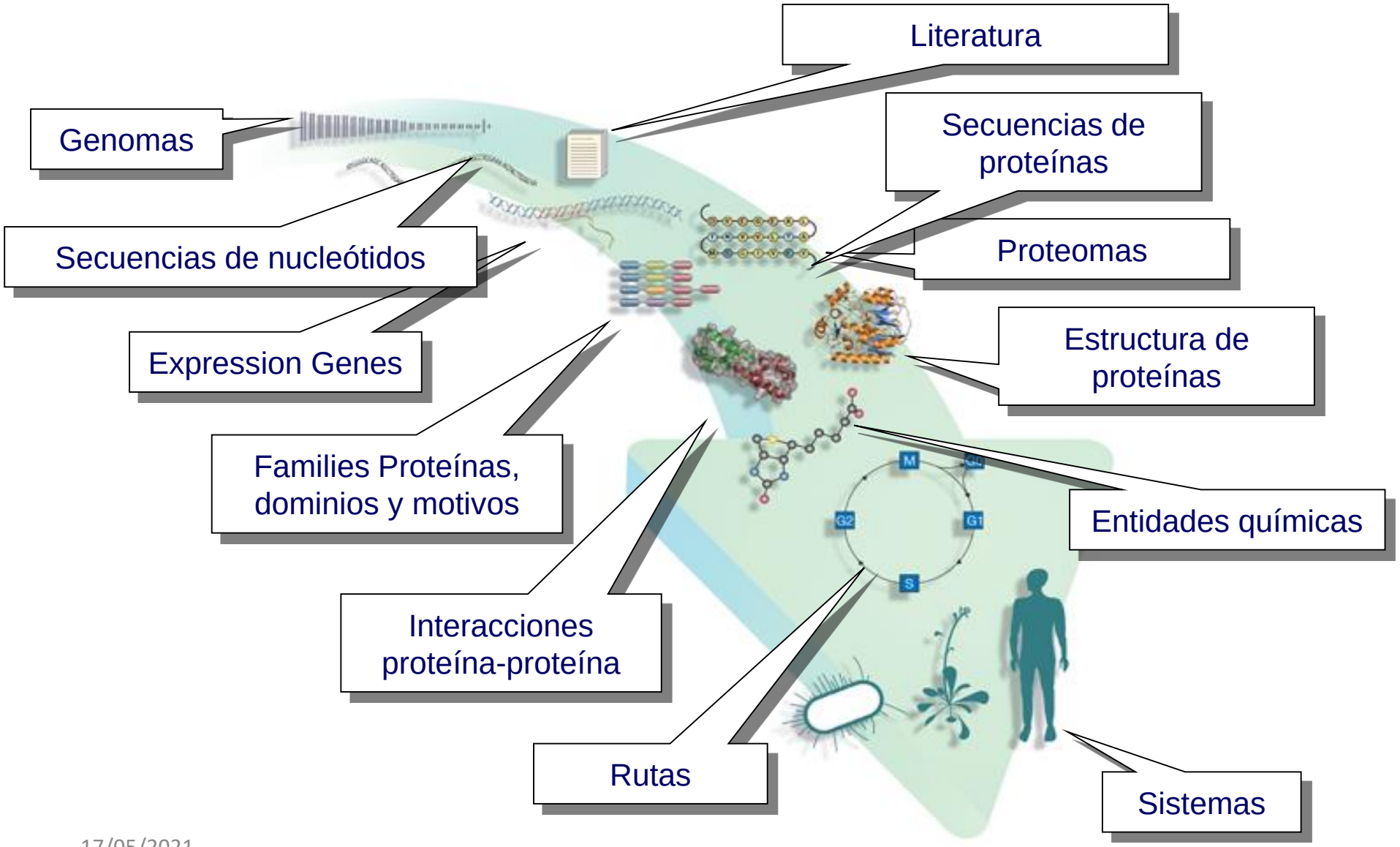
Procesamiento de datos

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

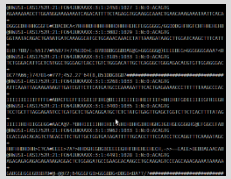
MinIon - Nanopore

2021

NovaSeq 6000

Bioinformatics Unit

2012



Service & Support to Researchers on HTS Data Analysis

National Microbiology Centre (CNM)

Research Institute for Rare Diseases (IIER)

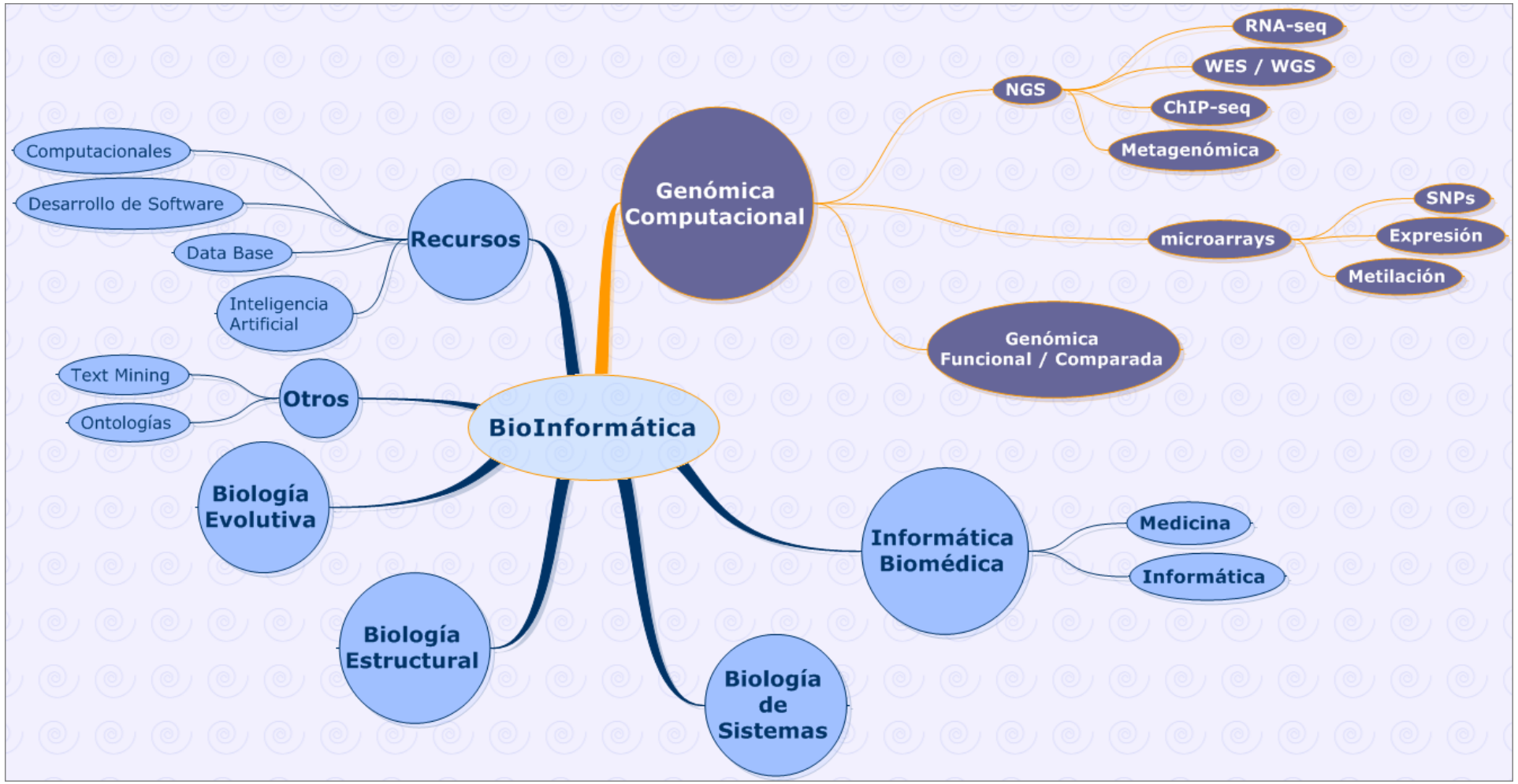
Functional Unit for Research in Chronic Disease

Network of Biological Alerts

National Centre of Tropical Medicine

National Environment Health Centre

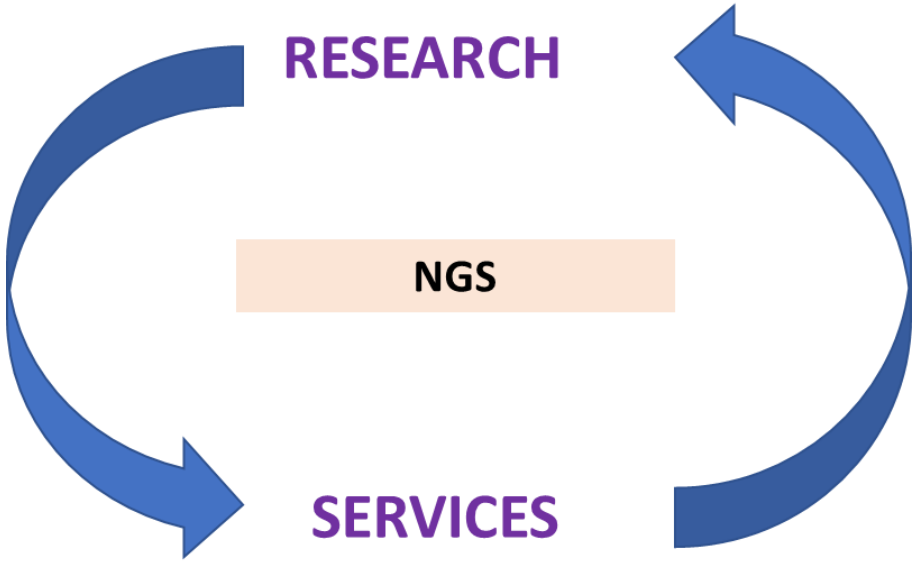
BU-ISCIII Mission - Activities



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

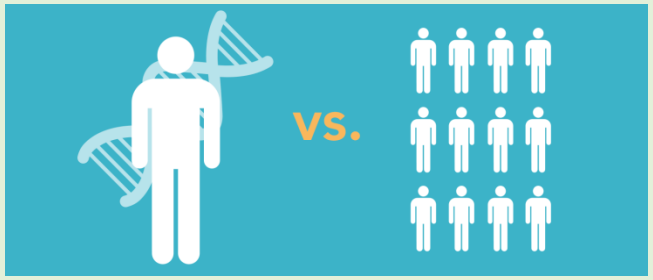
Provide scientific and technical solutions for using NGS in the diagnostic routine or research activity from different ISCIII labs

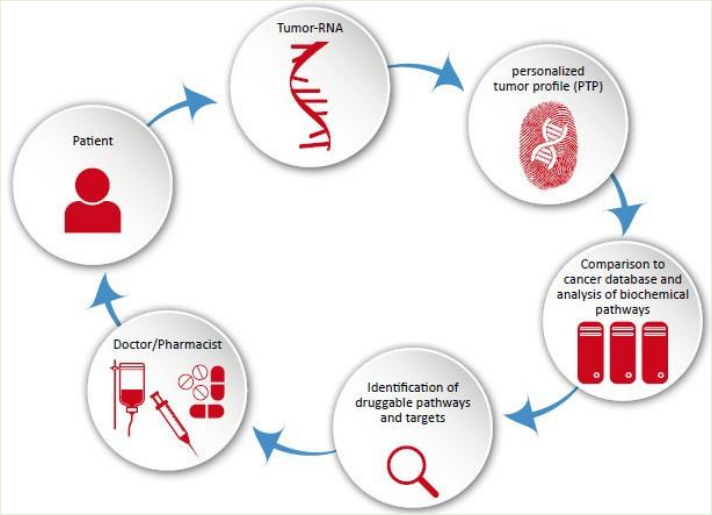


Clinical Bioinformatics - Precision Medicine

IIER

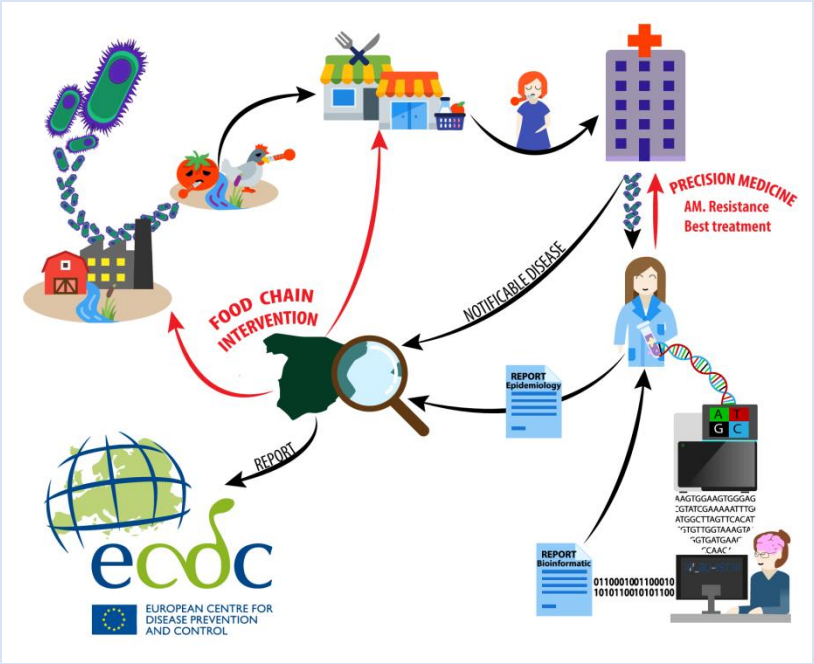
Tumour or Rare Diseases



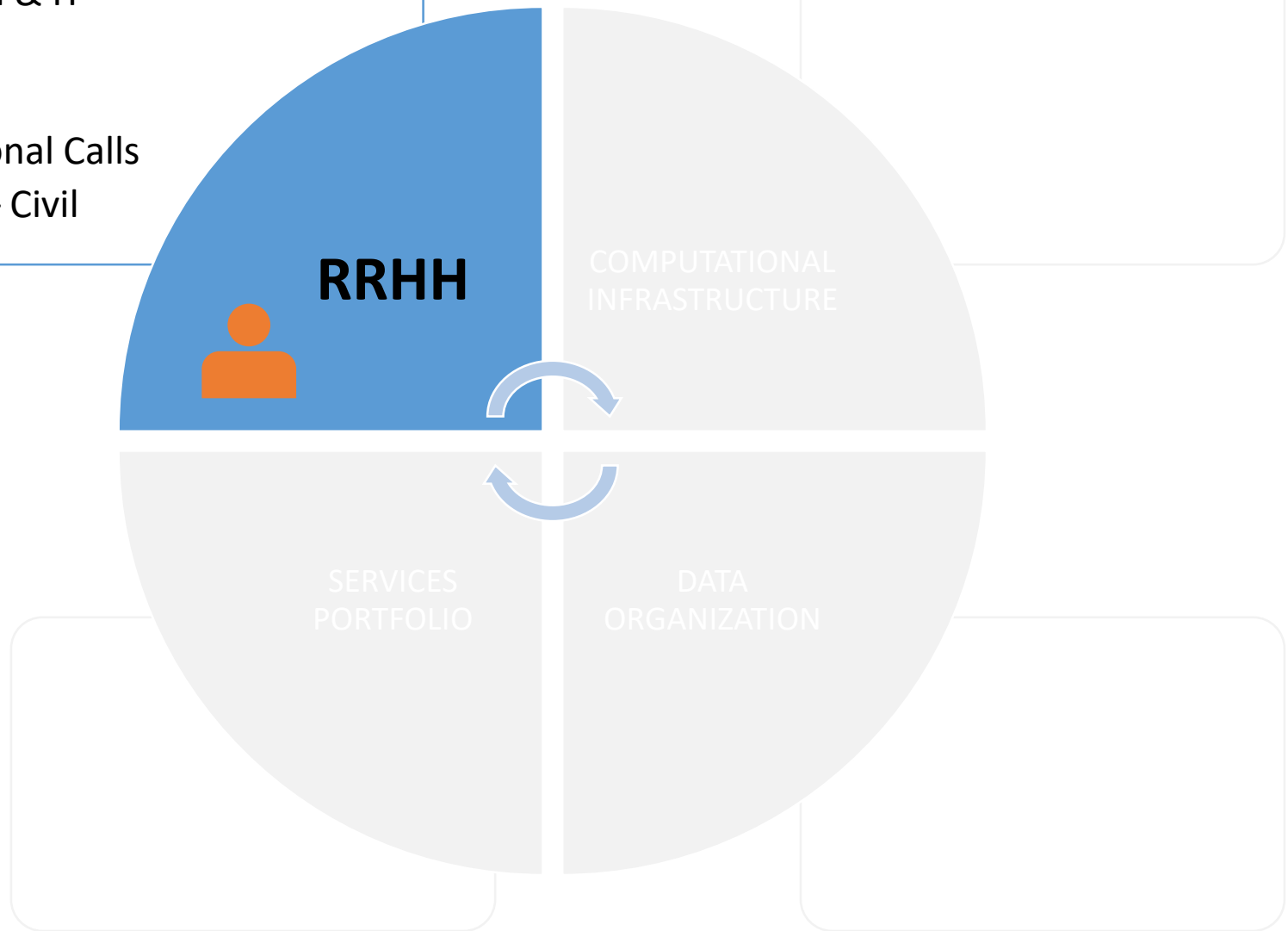


CNM

Pathogen associated



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

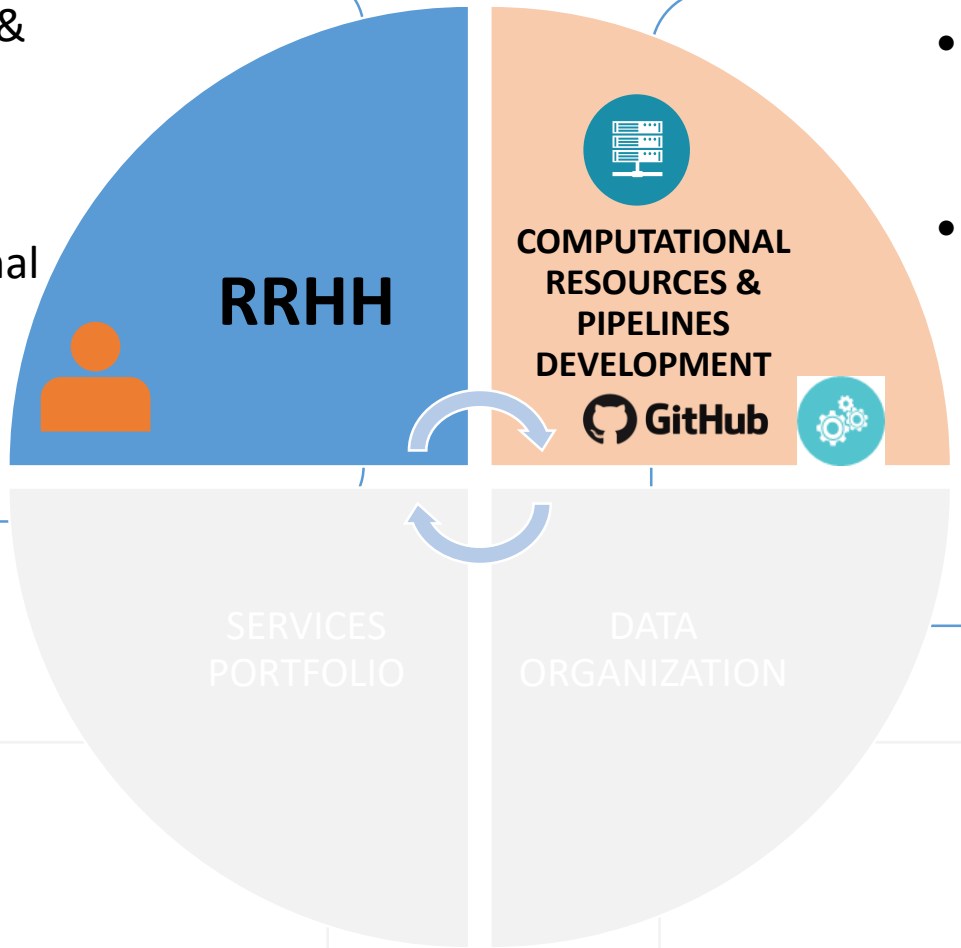


Human resources

	Disciplina	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
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 (2022)		
Luis Chapado									COLABORADOR		

Master en Bioinformática · CNM · IIER · BU-ISCIII · FUNCIONARIO

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



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

Computational Resources

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

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

Data Centre (CPD-ISCIII)

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

- 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

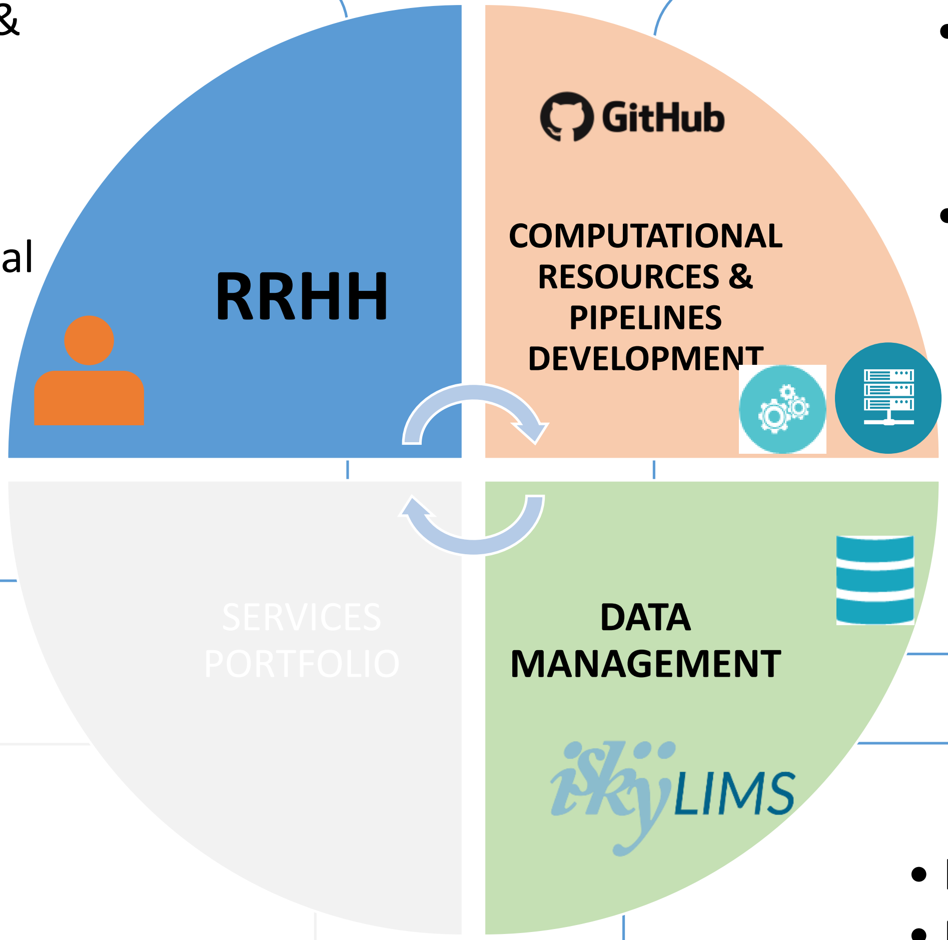
Lab ID	Allele based analysis					
	Approach	Allelic calling method	Assembler	Scheme	Difference within cluster	Difference outside cluster
EQA provider	BioNumerics	Assembly- and mapping-based	SPAdes	Applied Math (cgMLST/Pasteur)	0-3	24-1112
19	BioNumerics	Assembly- and mapping-based	SPAdes	Applied Math (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			

Table 4. Results of SNP-based cluster analysis

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

Fifth external quality assessment scheme for Listeria monocytogenes typing

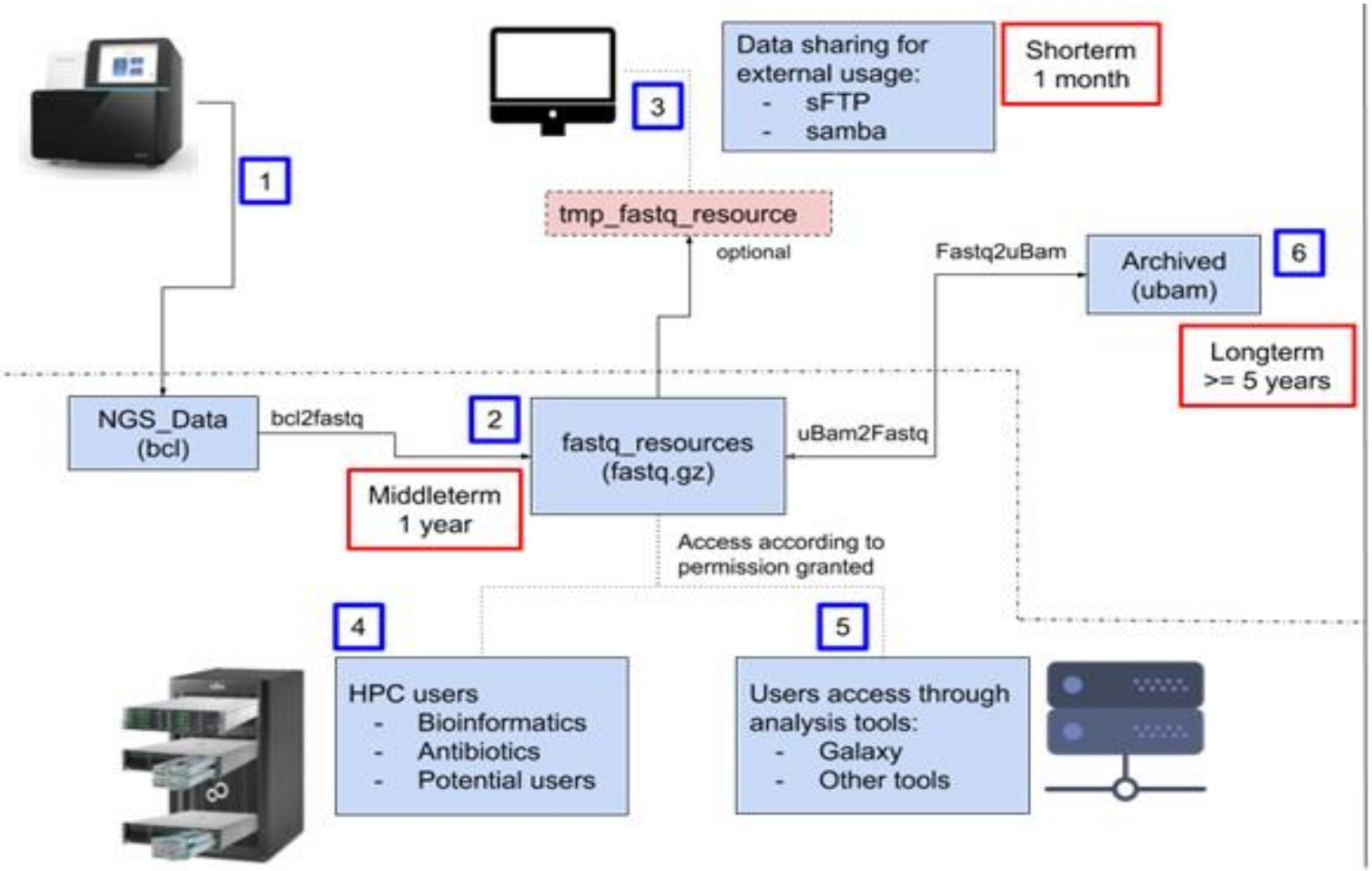
- TEAM: Bioinformatician & IT
- SOURCE OF FUNDING:
  - Research Project
  - National or International Calls
  - Permanent position – Civil Servant



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

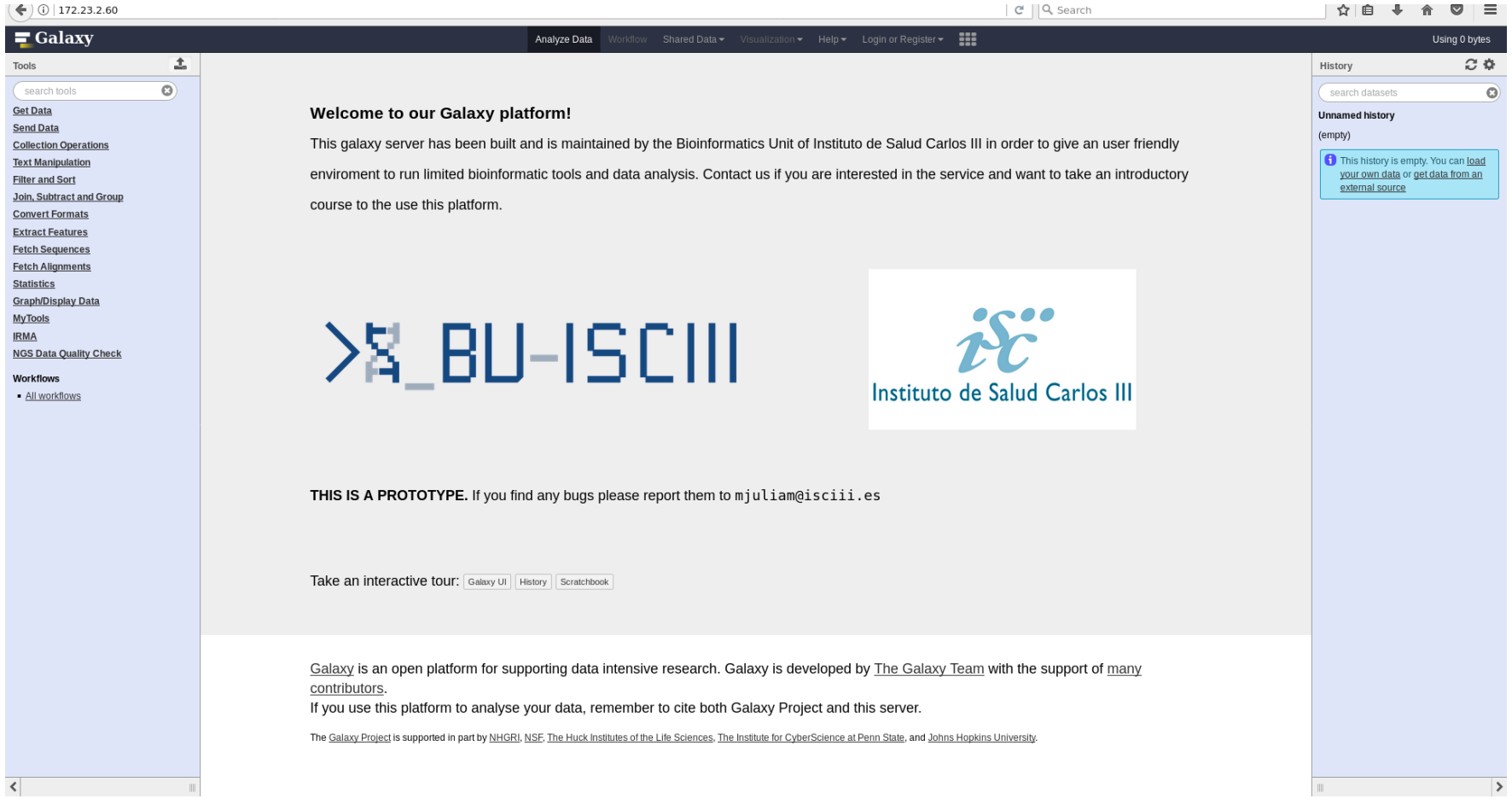
- INFRASTRUCTURE
- DATA WORKFLOW
- LIMS

Infrastructure and data management

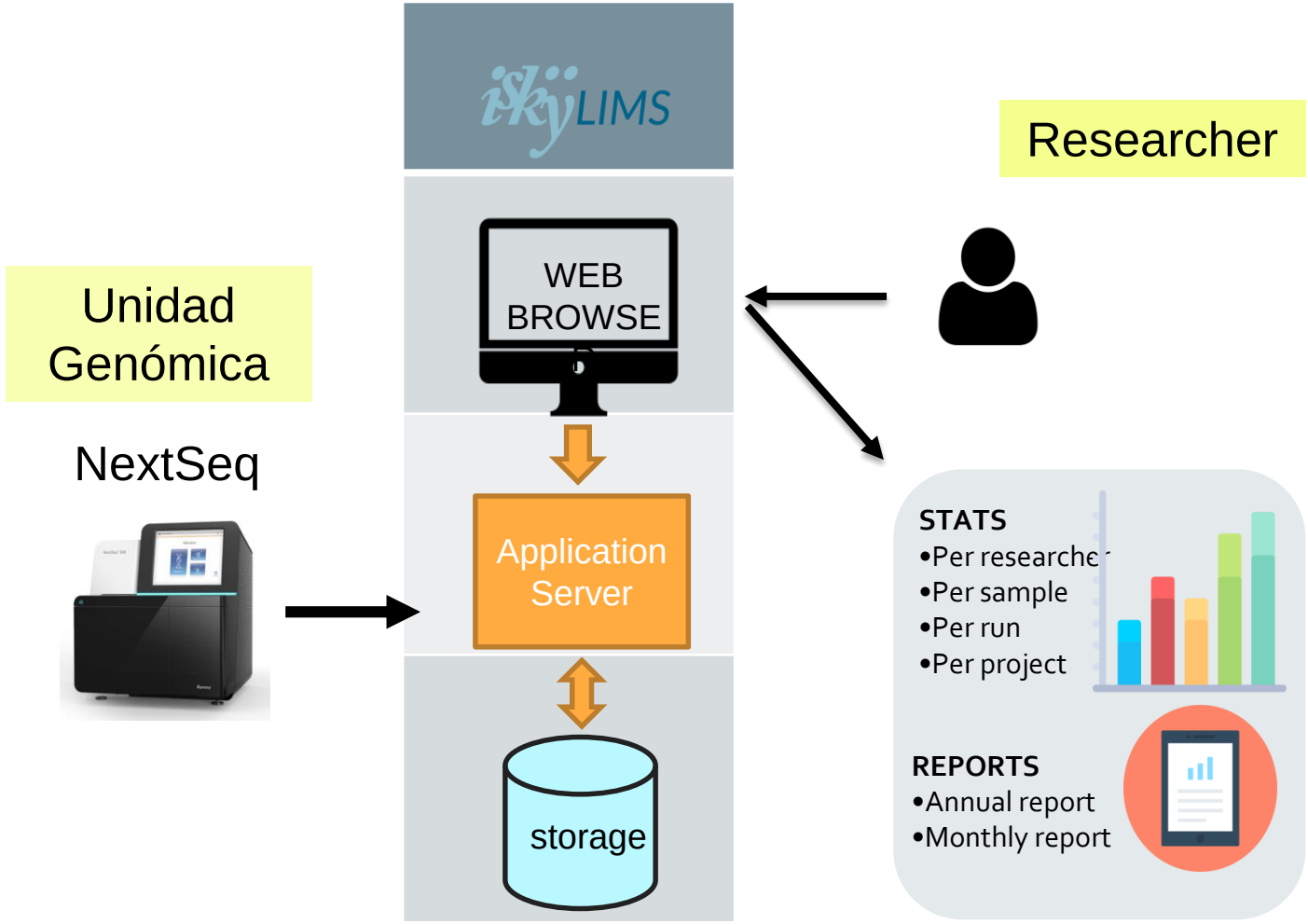


- Minimize I/O issues
- Maximize storage uses

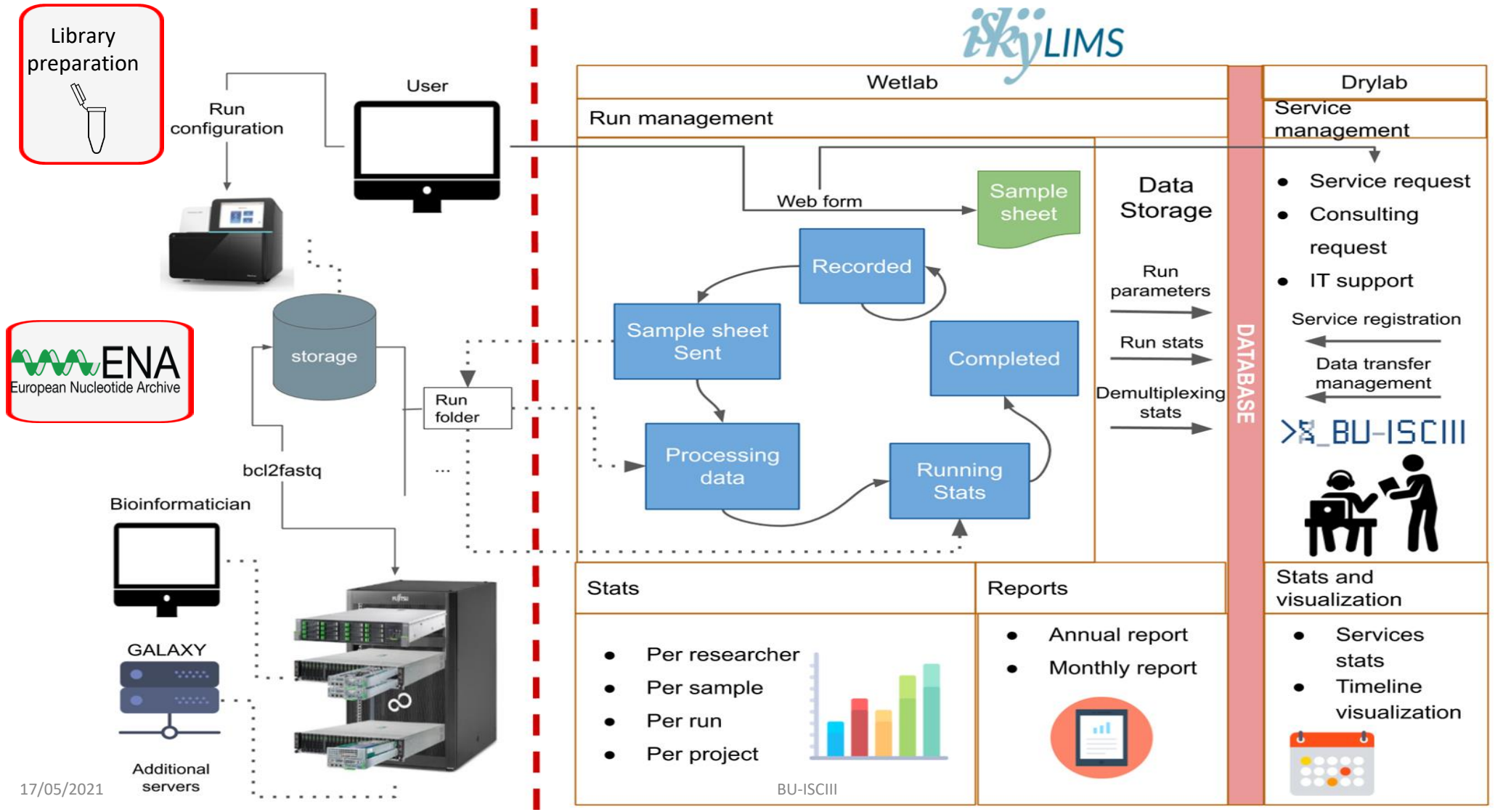
Galaxy


172.23.2.60
Search
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.
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

iSkyLIMS

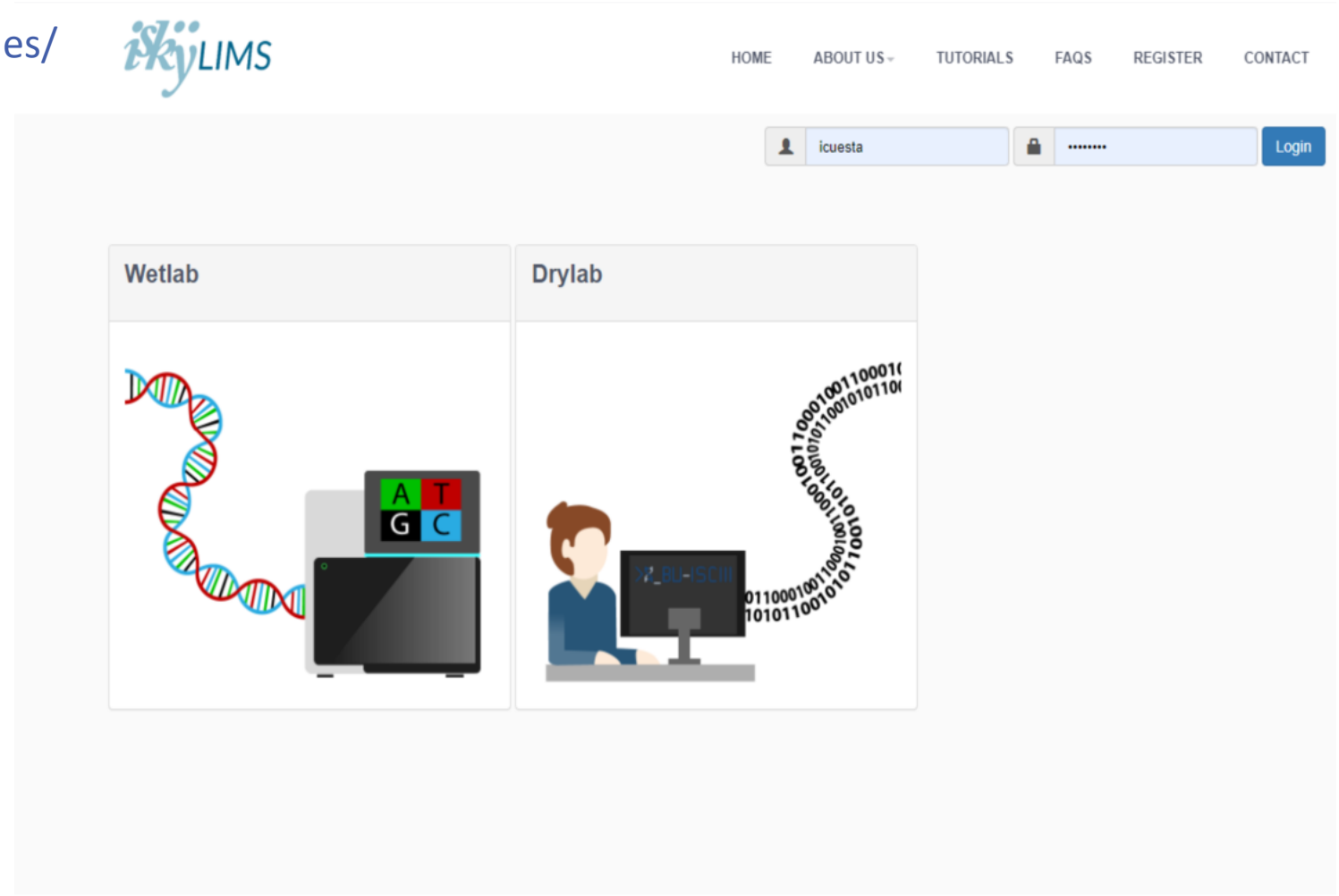

iSkyLIMS
Researcher
WEB BROWSE
Unidad Genómica
NextSeq
Application Server
storage
STATS
•Per researche
•Per sample
•Per run
•Per project
REPORTS
•Annual report
•Monthly report

Infrastructure and data management: LIMS



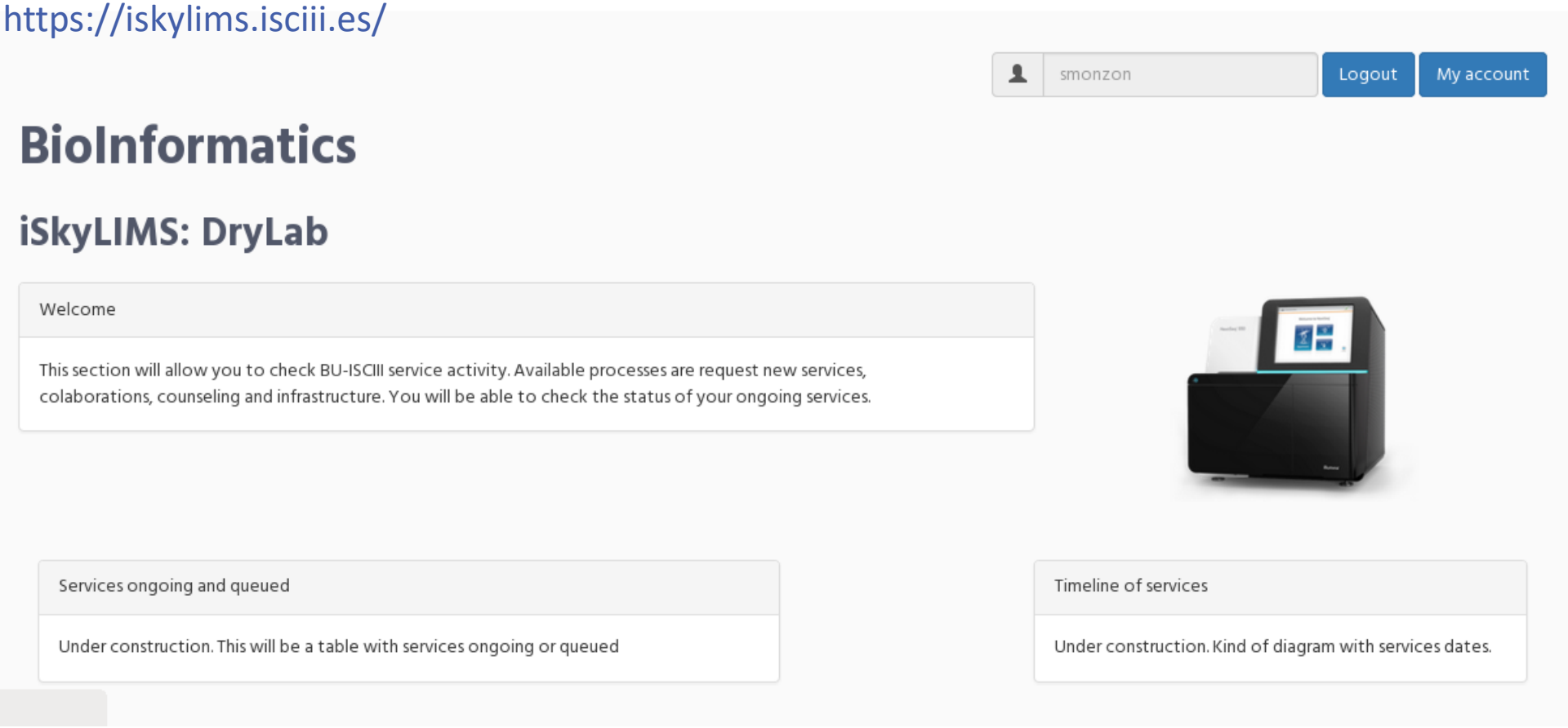
SERVICIOS DE LA UNIDAD

https://iskylims.isciii.es/

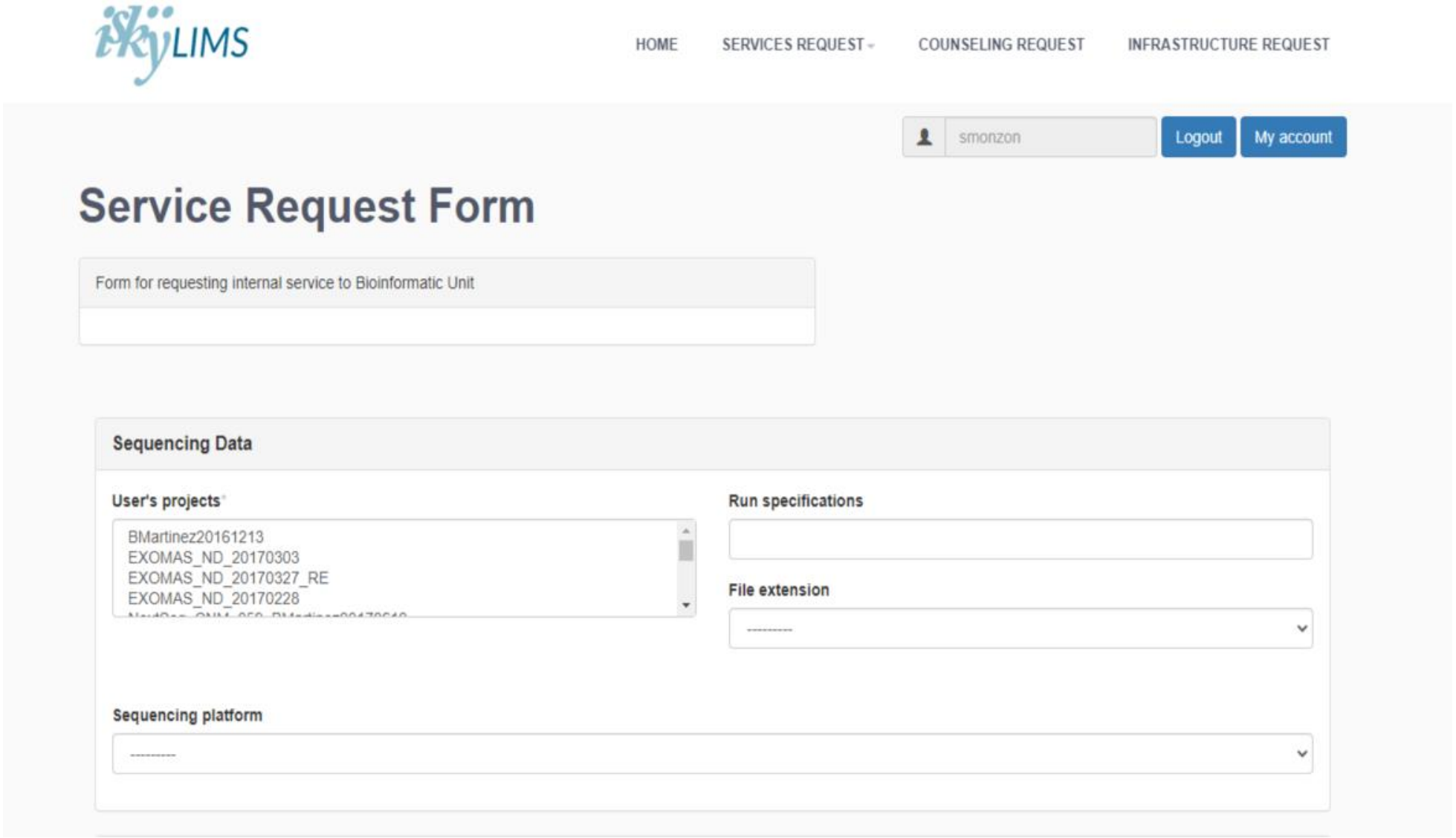


SERVICIOS DE LA UNIDAD

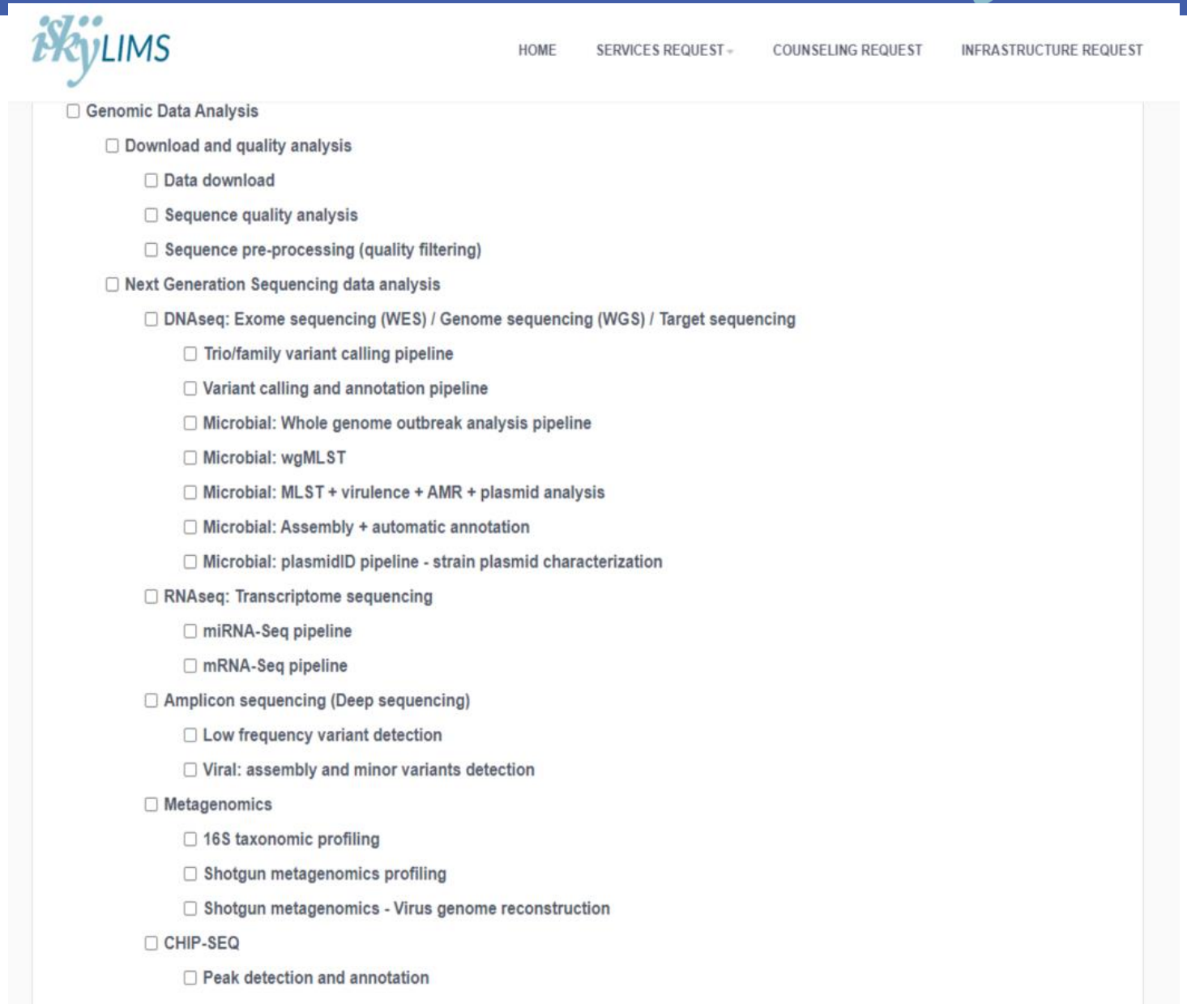
https://iskylims.isciii.es/



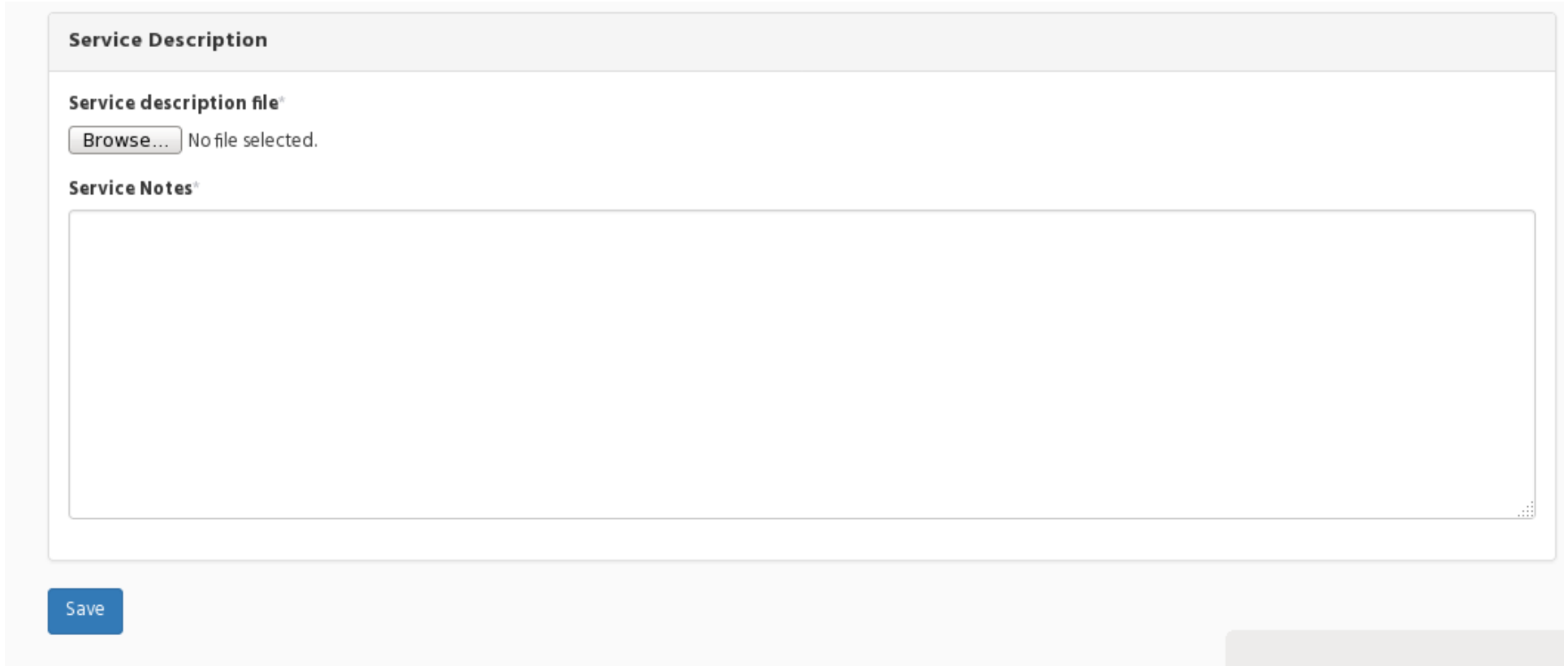
SERVICES REQUEST



SERVICES REQUEST



SERVICES REQUEST



COUNSELING REQUEST

Service selection

AvailableServices *

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

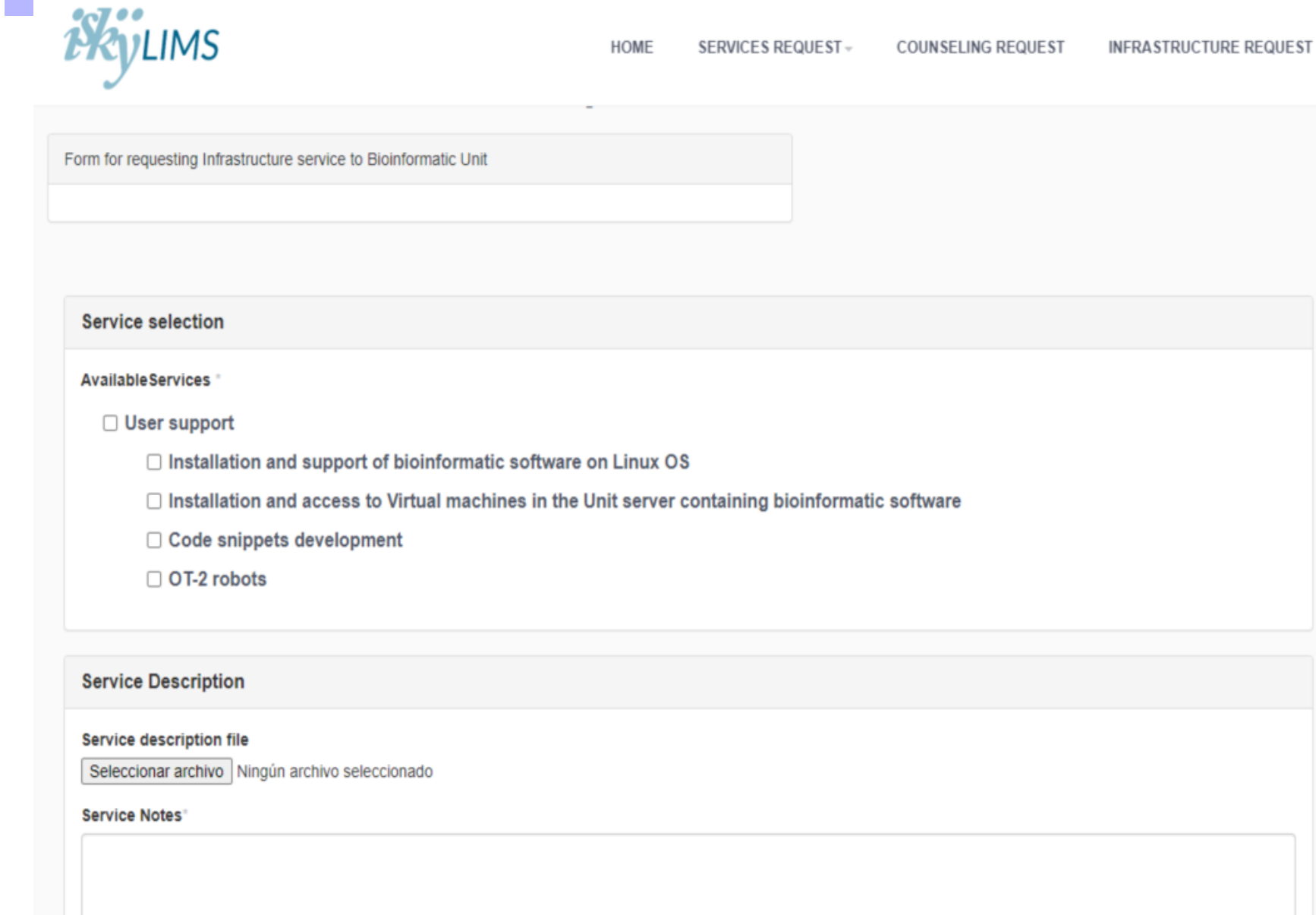
Service Description

Service description file*

Browse... No file selected.

Service Notes*

INFRASTRUCTURE REQUEST



Infrastructure and data management

iSkyLIMS

HOME · RUN PREPARATION · SEARCH · STATISTICS · REPORTS

bioinfoadm · Logout · My account

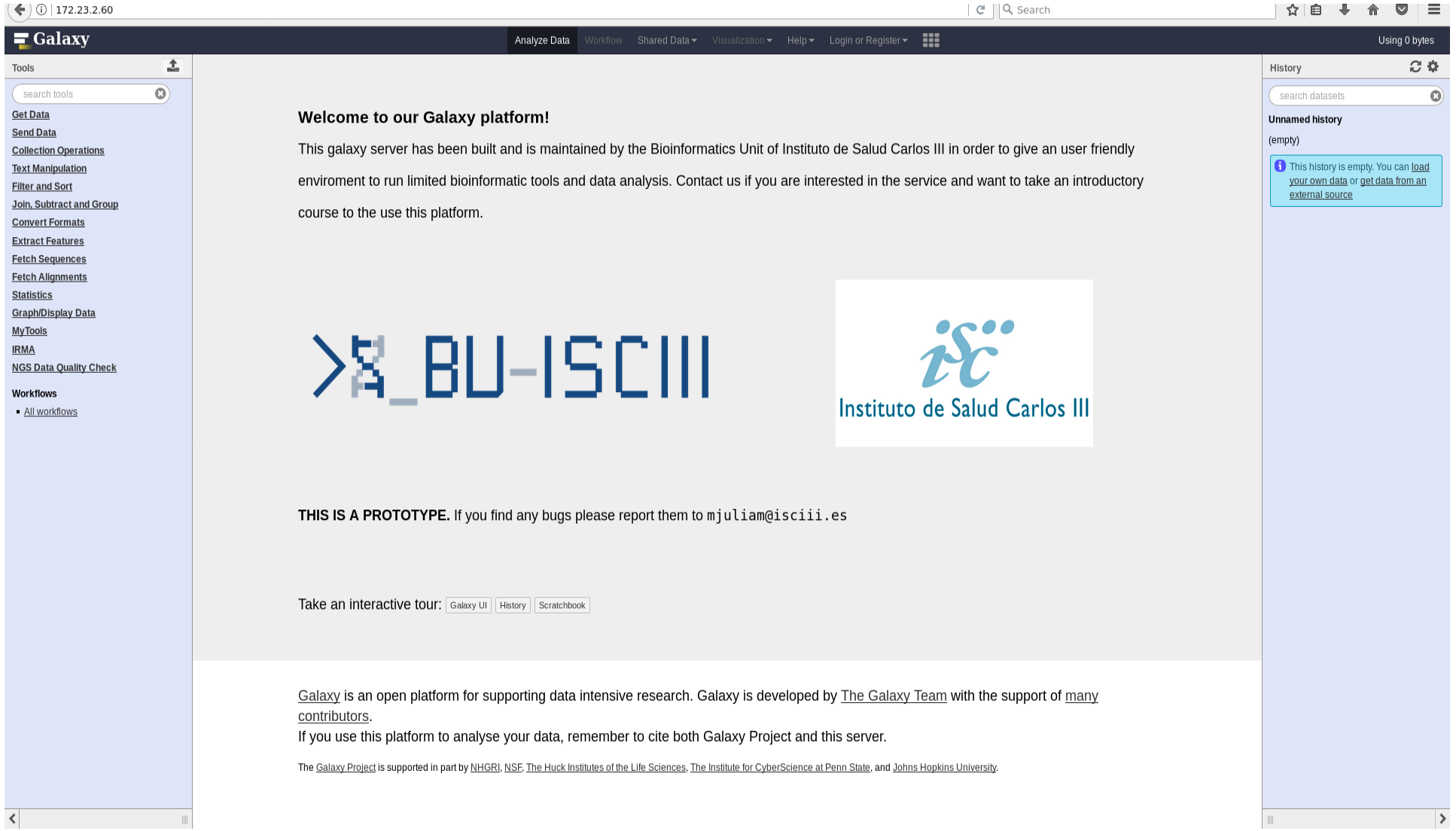
Statistics results for Investigator rabad

Projects using the sequencer NS500454 :

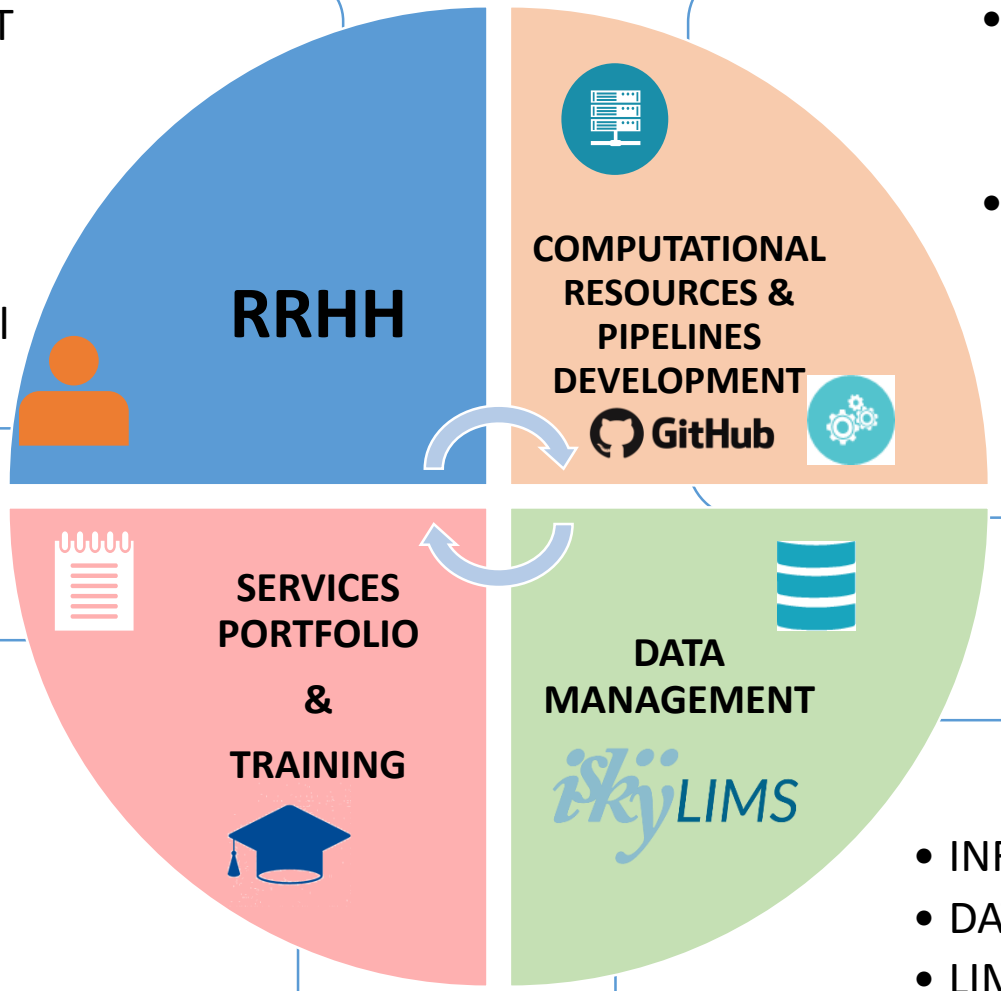
Export Table To Excel

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

Galaxy


172.23.2.60
Search
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.
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

- 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

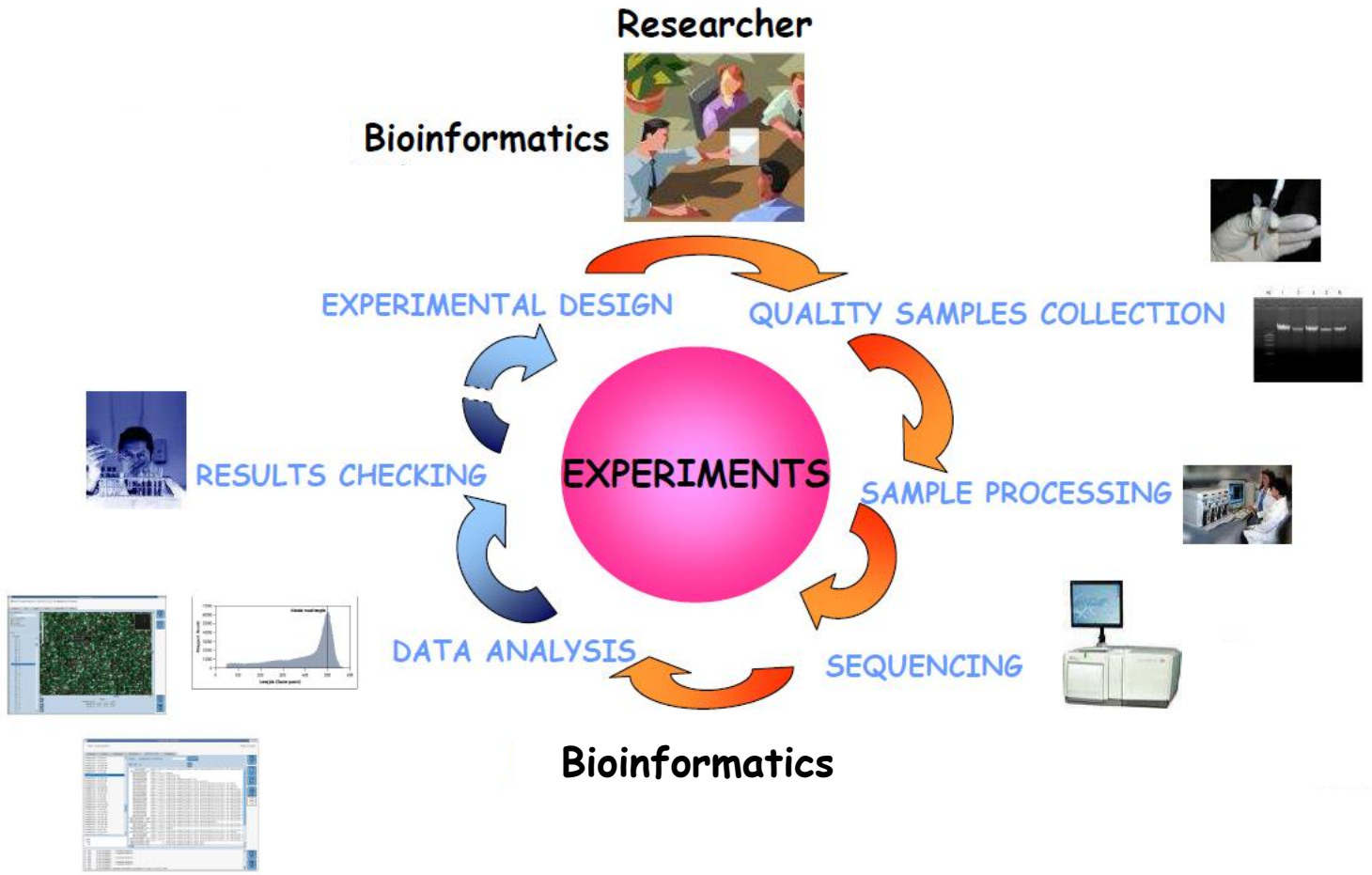
Cartera de Servicios

- **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

Workflow en NGS

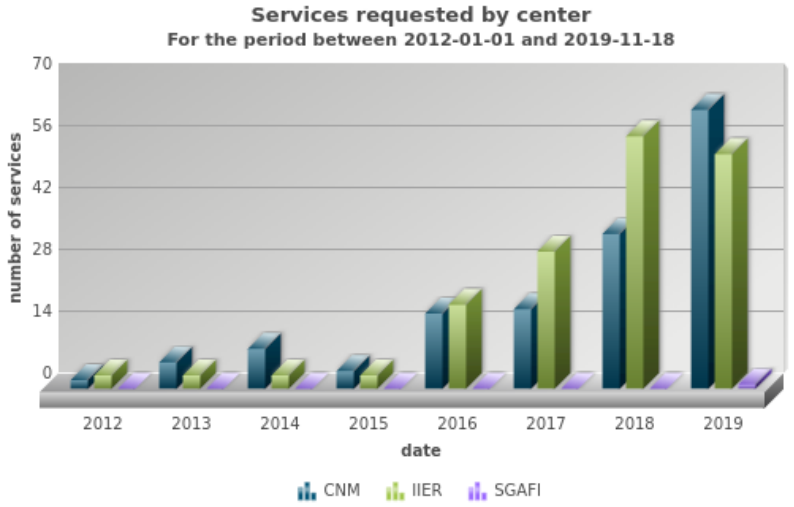


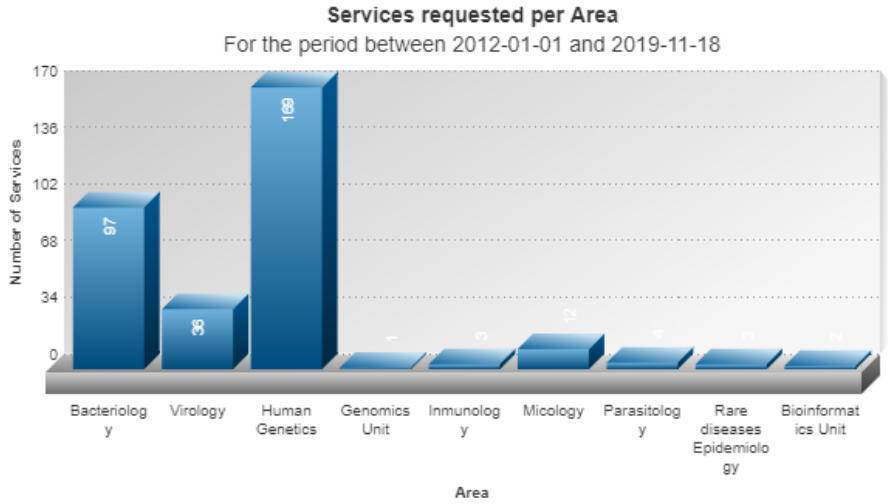
Services Portfolio

		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 – 2019

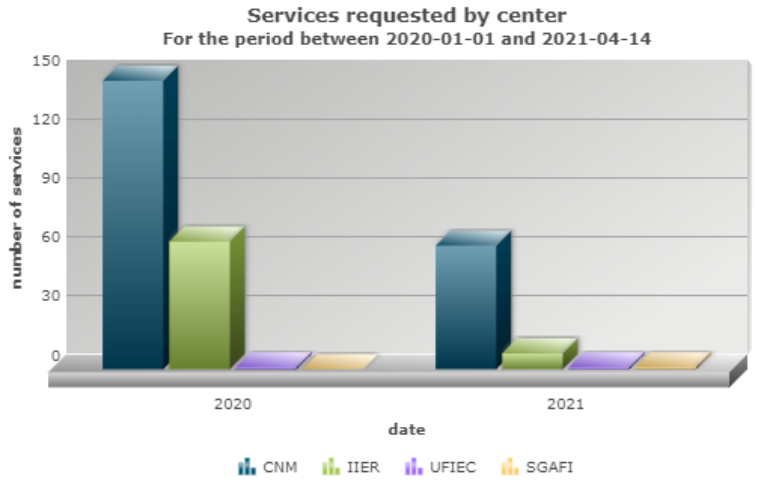
- 327 Services per center / CNM – IIER / 22 Researchers

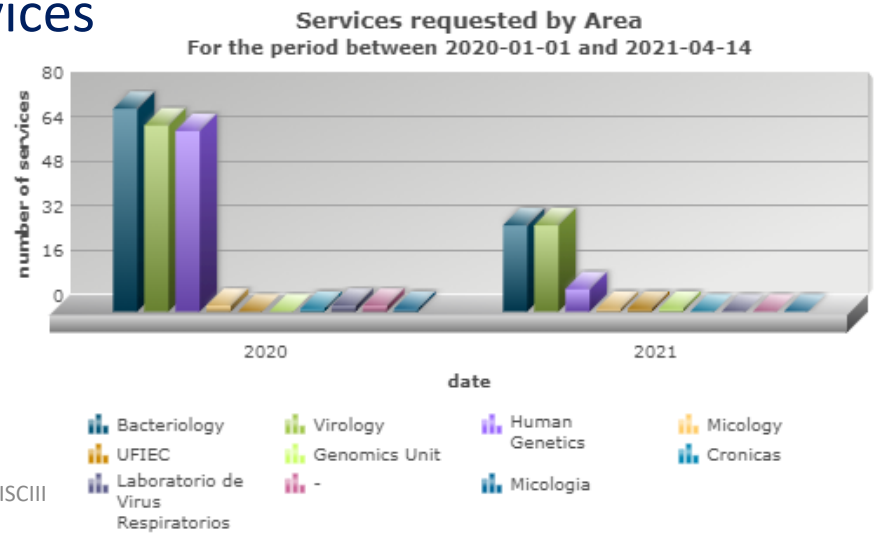




2020-2021

200 Services





Training

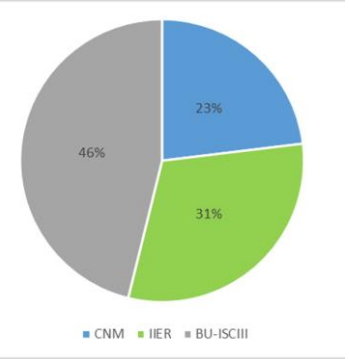
Courses

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)

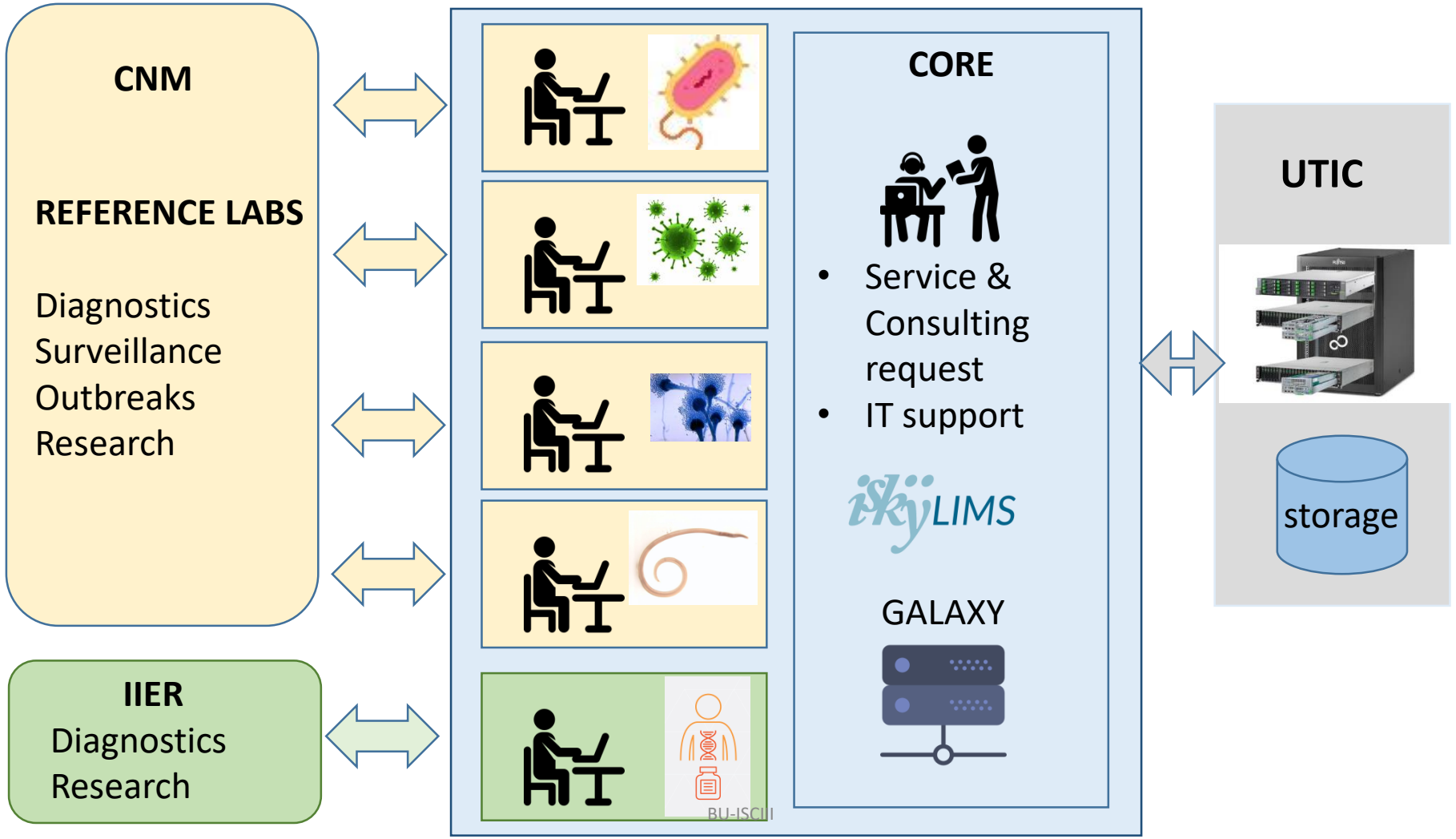


Master & Grade Students

- 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

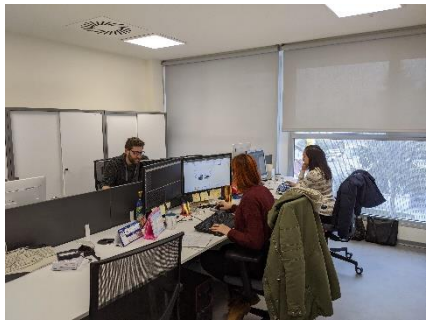
python

R

DOCTORADO

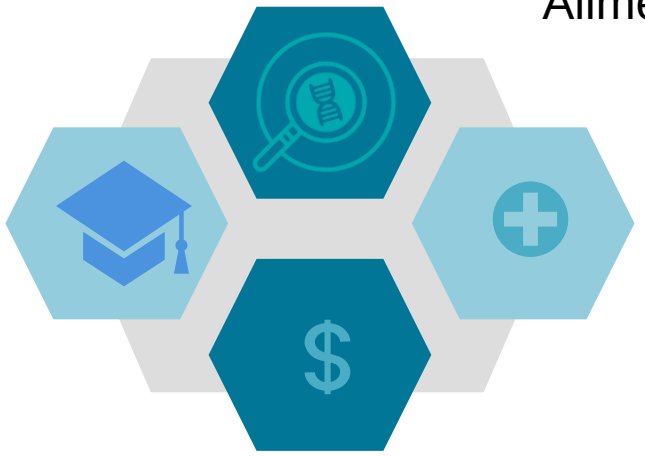
BIOLOGÍA – ANÁLISIS DE DATOS

¿Dónde trabaja un Bioinformático?



CENTRO DE INVESTIGACIÓN

Biomedicina
Agricultura
Alimentación



UNIVERSIDAD

Biociencias
Informática

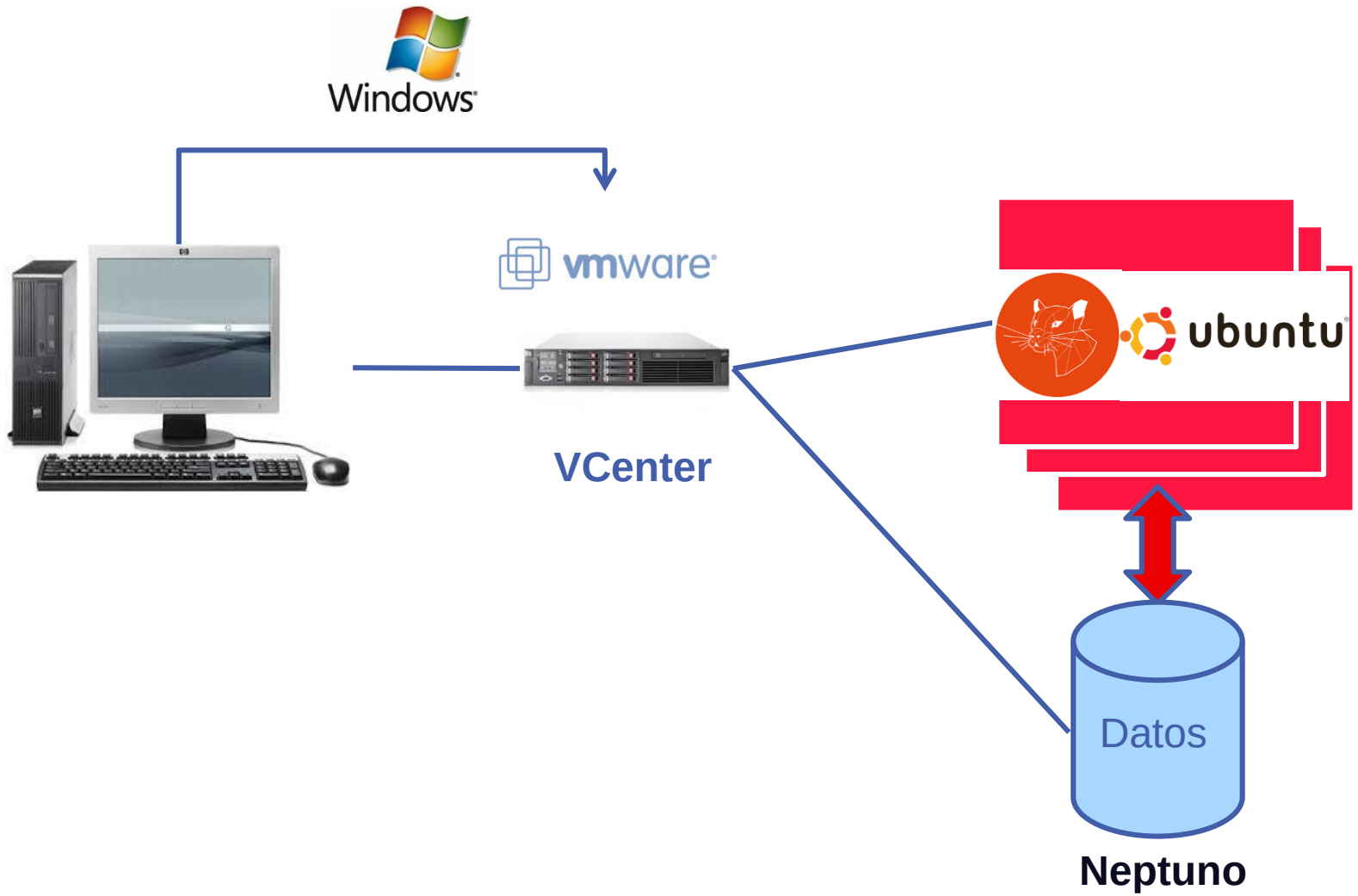
HOSPITAL

BIOINFORMÁTICO CLÍNICO
Genética
Oncología
Cardiología

EMPRESA

Bioinformática
Genética
Genómica

Recursos Informáticos para el curso

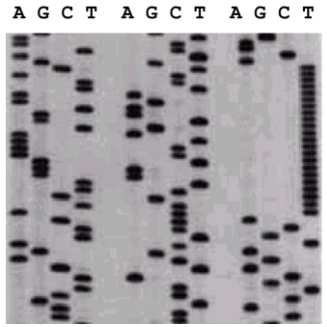


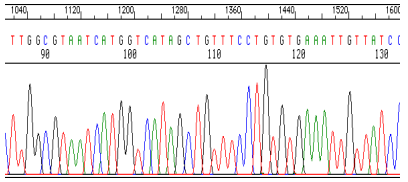
INDICE

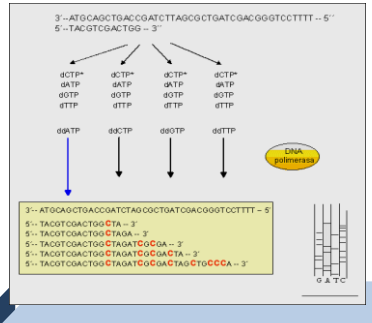
- Unidad de Bioinformática
- Servicios ofertados

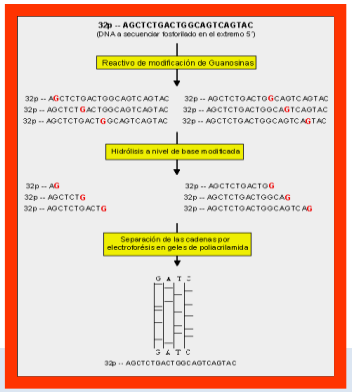
- Evolución de la secuenciación
- Plataformas de secuenciación masiva (NGS)

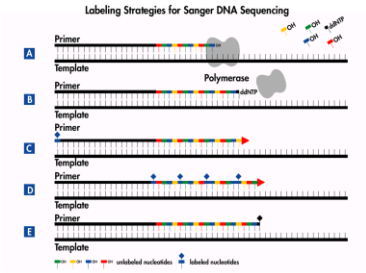
Métodos de secuenciación de DNA











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).*

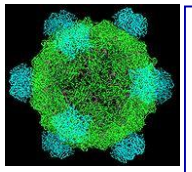
1972

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

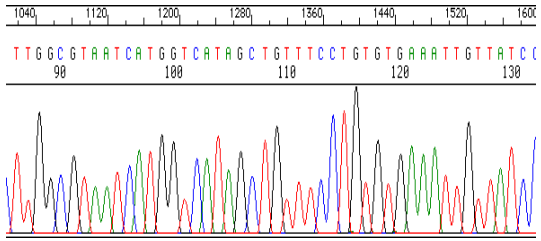
Evolution of DNA Revolution

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)



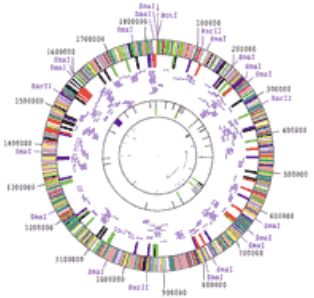
BLAST: Myers & Lipman publish the first algorithm to align DNA sequences (JMB, 215 1990).

Evolution of DNA Revolution

A walk through the biological history: from Sanger to NGS

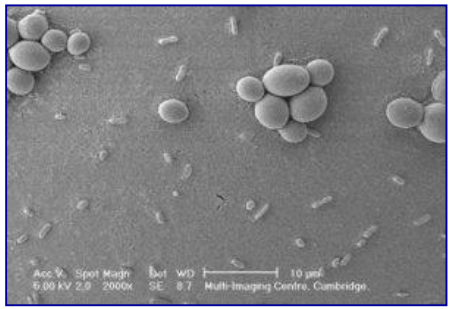
1995

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).



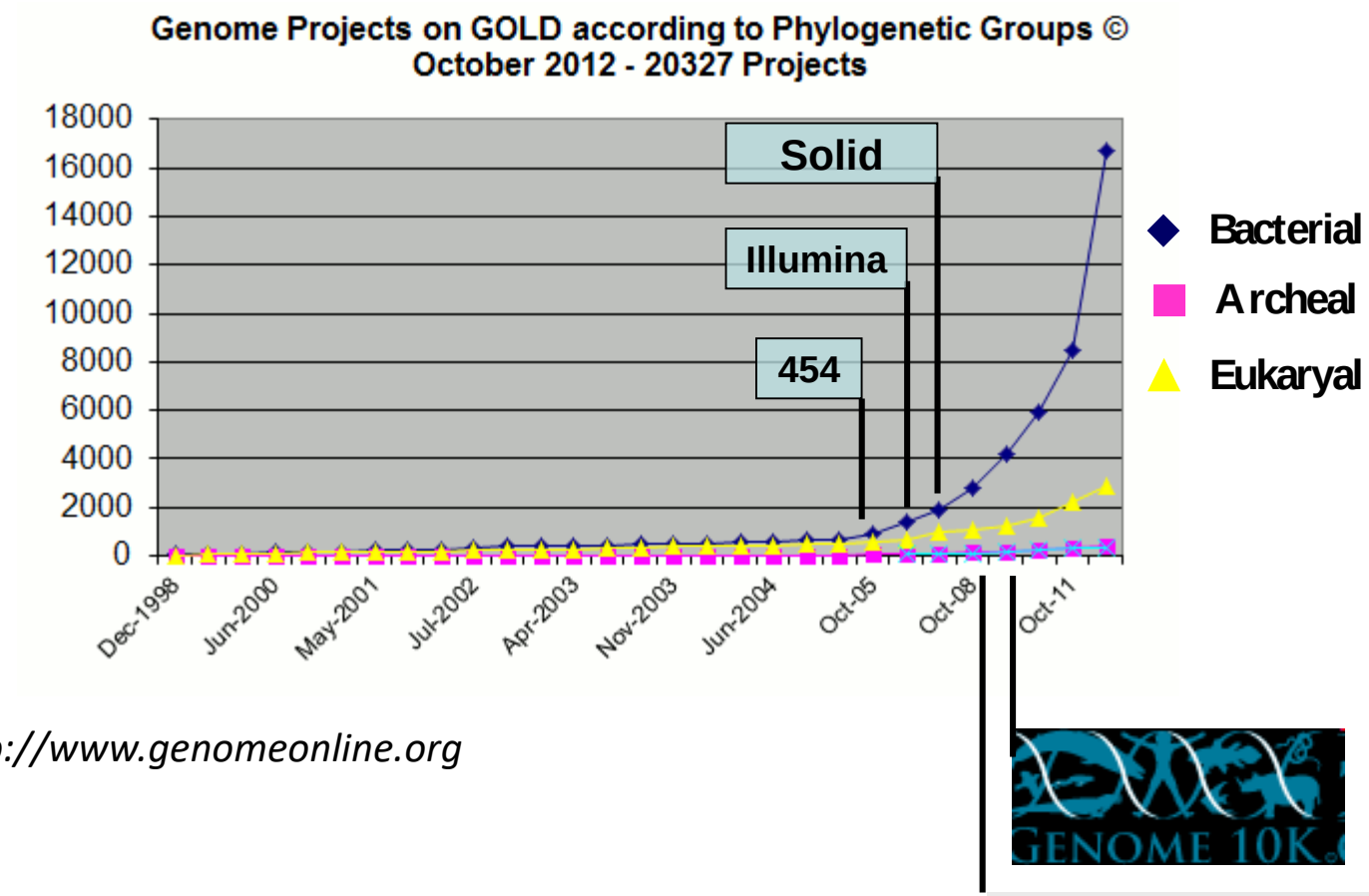
2003

Bacterial Genomes

Bacillus subtilis
Escherichia coli

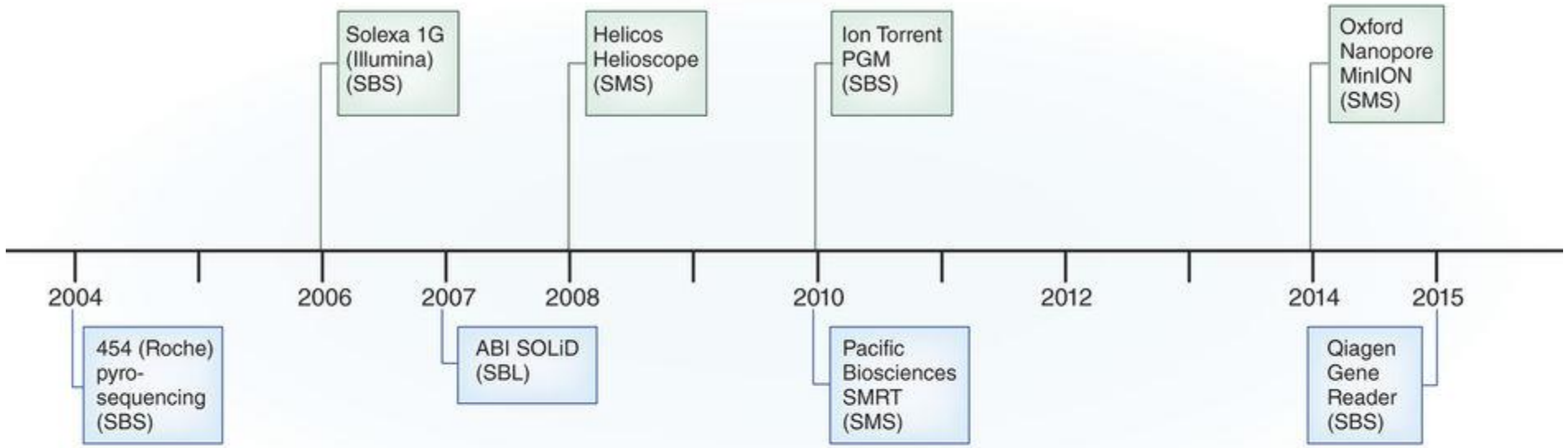
Eukaryotic Genomes
Caenorhabditis elegans
Arabidopsis thaliana
Drosophila melanogaster
Human genome **(1986-2003)**

Genomics Revolution Era

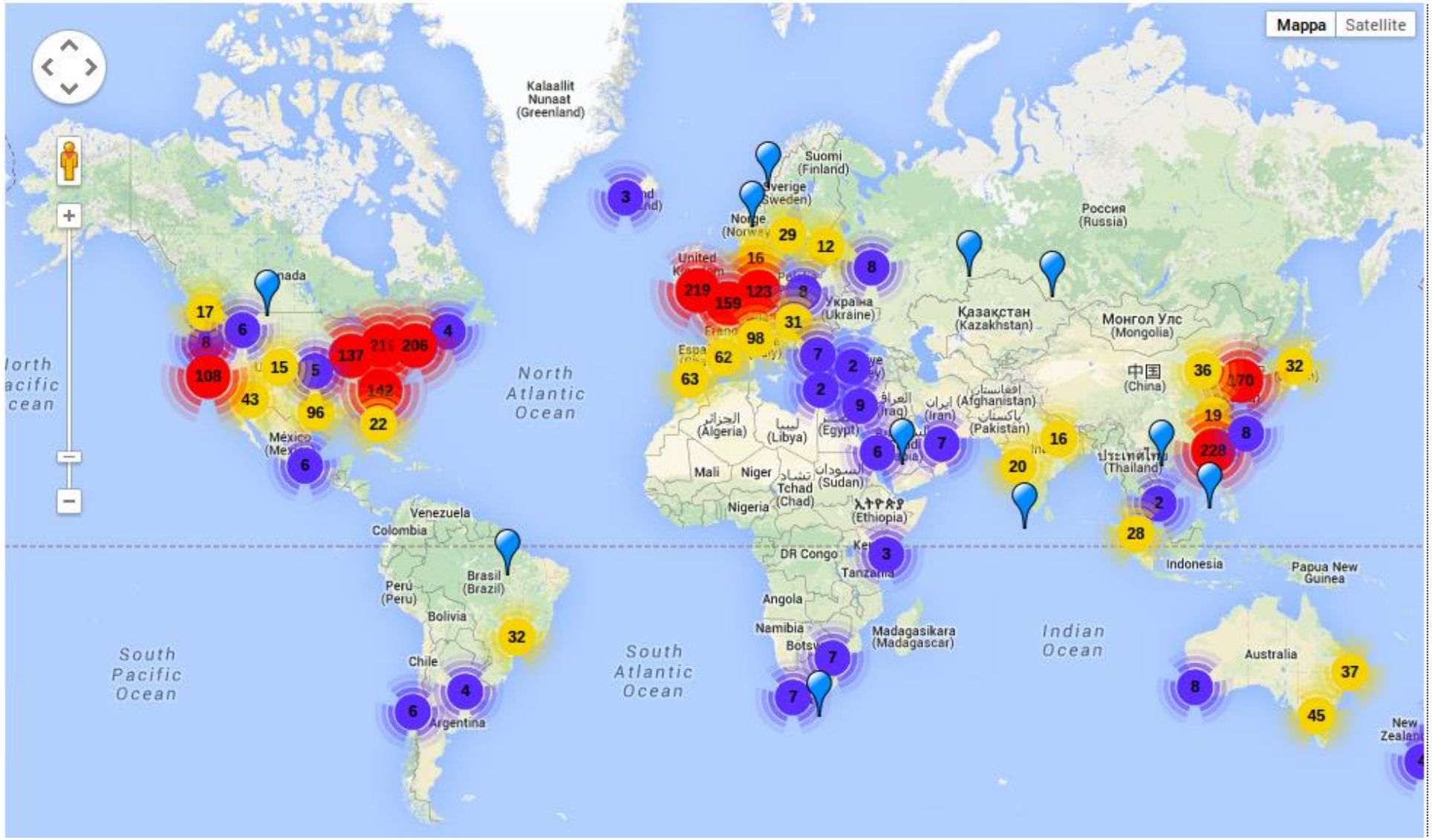


Source: http://www.genomeonline.org

NGS PLATFORMS: TIMELINE

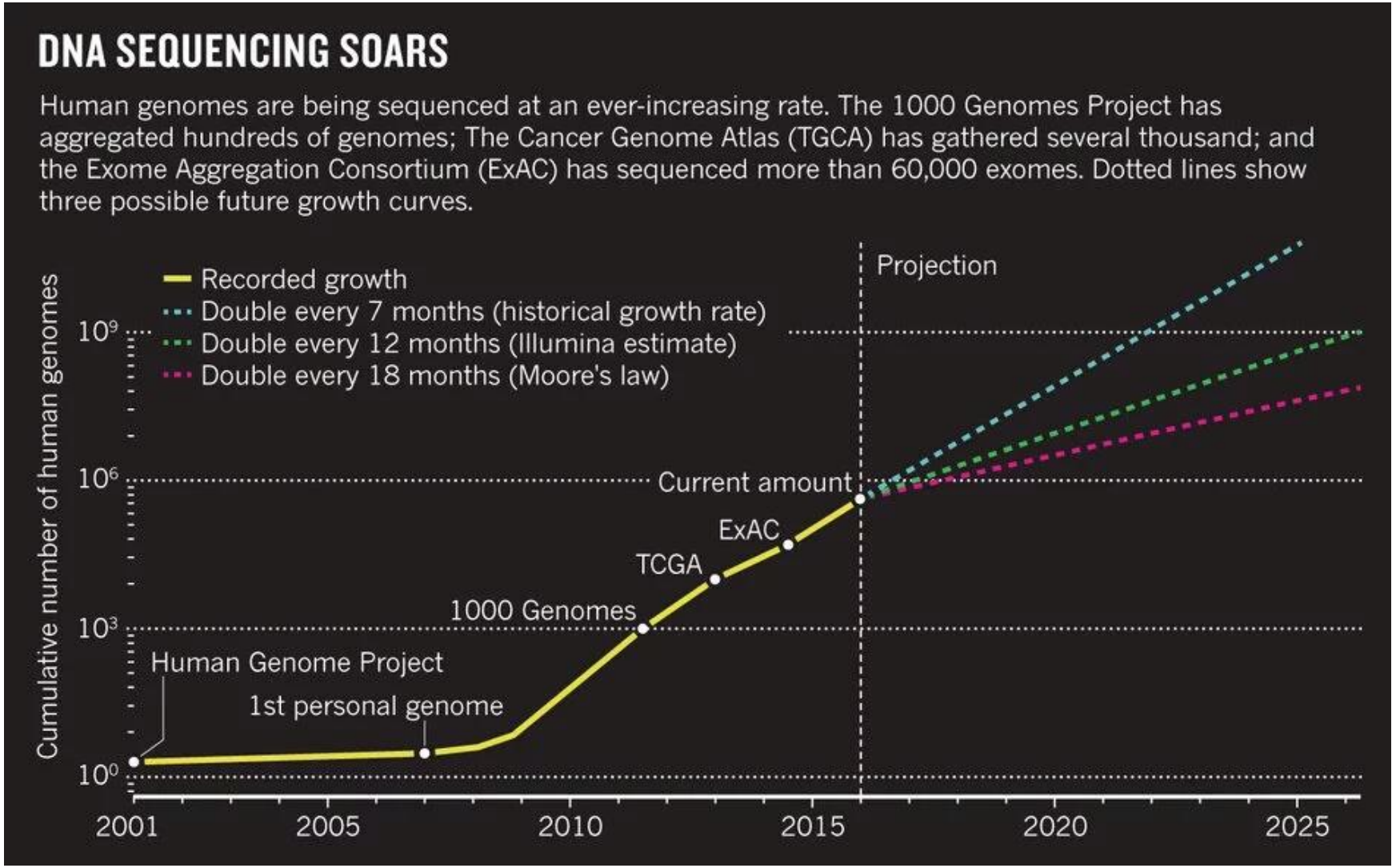


Democratización de la secuenciación. Mapa de los secuenciadores de alto rendimiento



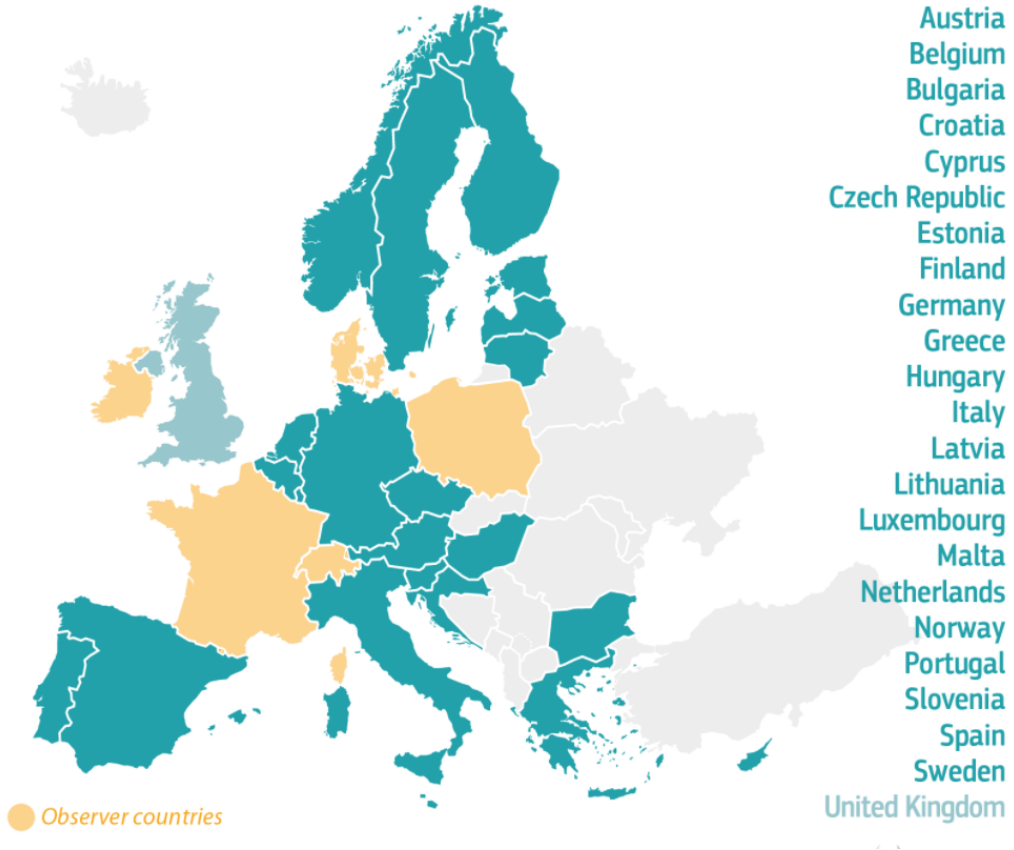
http://omicsmap.com

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

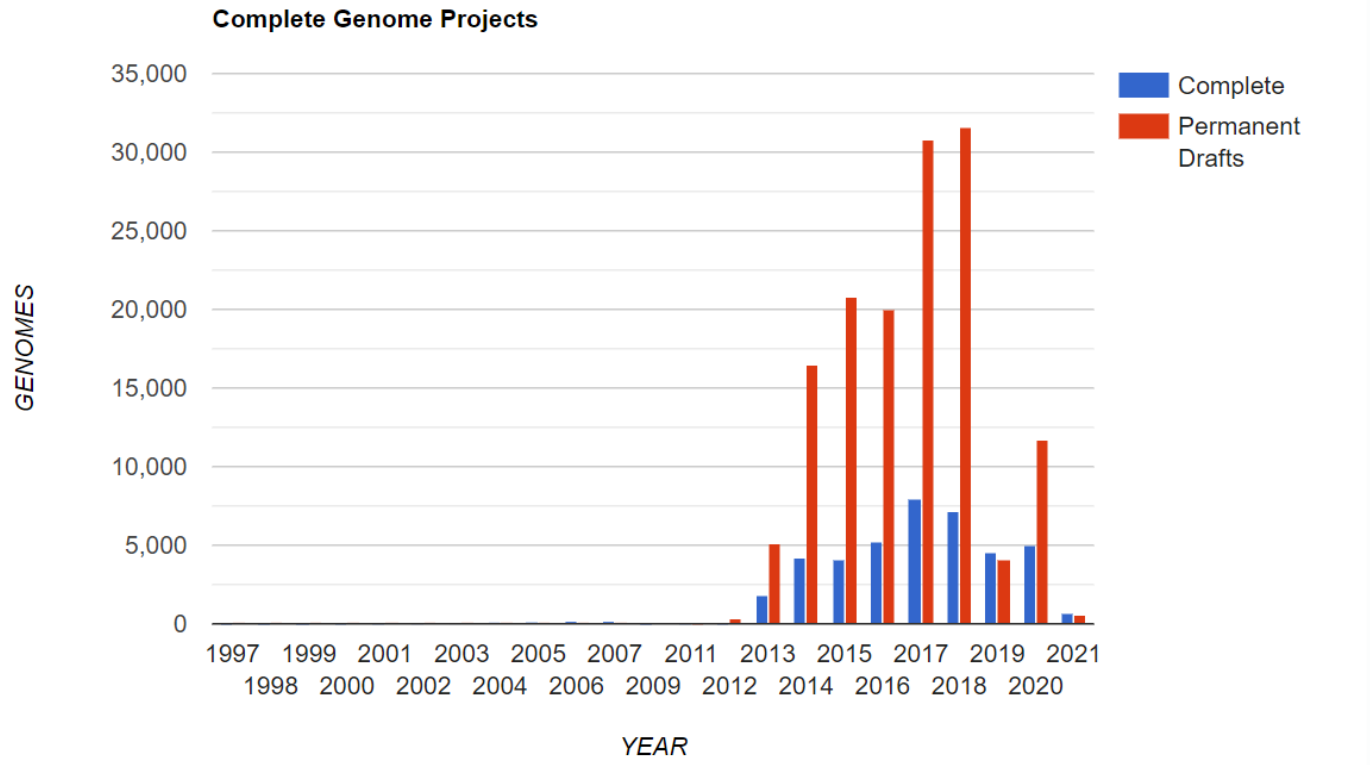
Genomics Revolution Era

JGI GOLD GENOMES ONLINE DATABASE

JGI HOME LOG IN

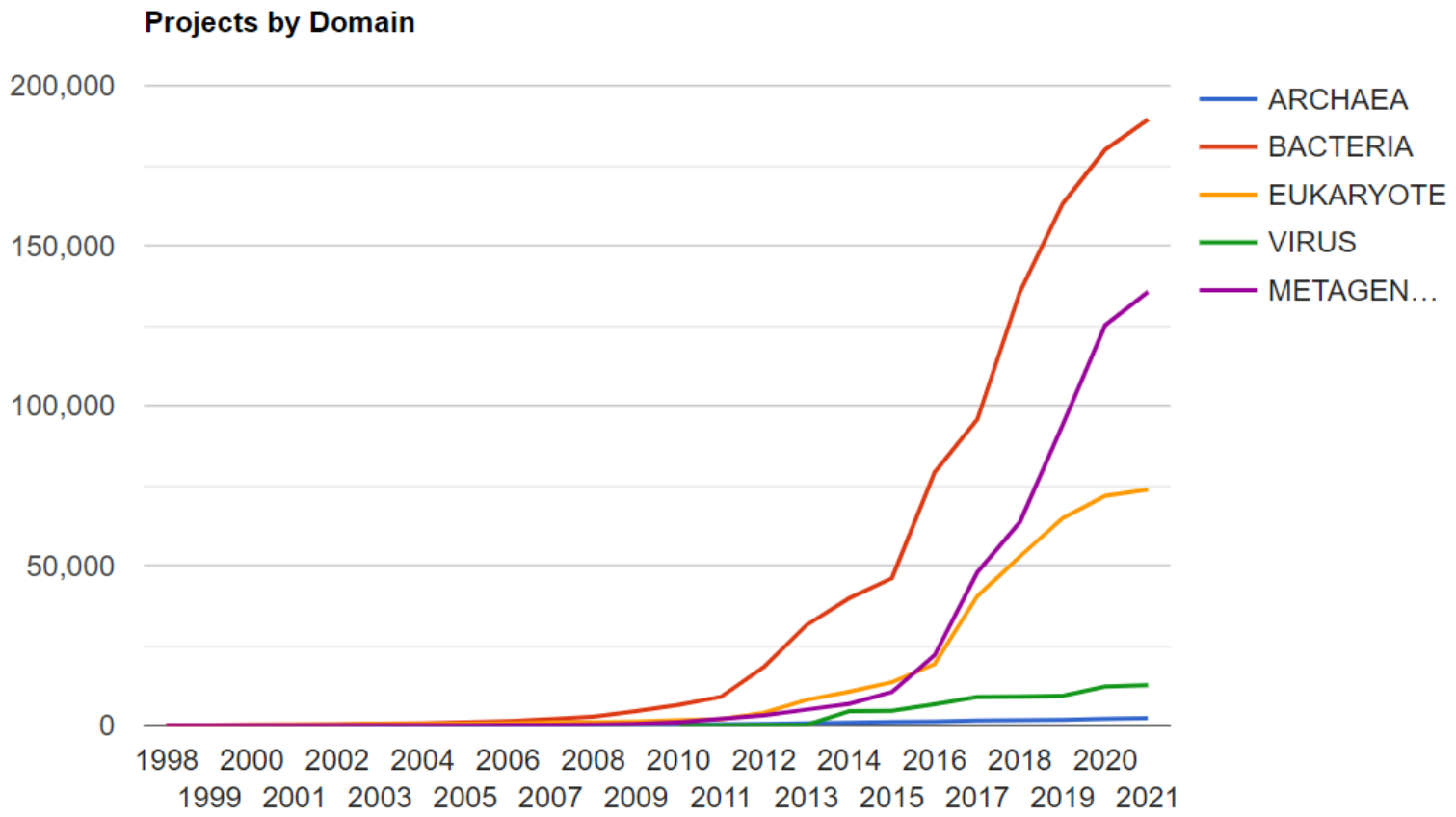
Home Search Distribution Graphs Biogeographical Metadata Statistics References Team Help News

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



Genome Totals in GOLD (by year and status)

Genomics Revolution Era



Secuenciadores

Primera generación

- Sanger

Segunda Generación

- 454/Roche
- Solexa/Illumina
- Solid
- Ion Torrent

Tercera Generación

- Pacific Biosciences
- Nanopore

High-Throughput Sequencing Platforms

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

High-Throughput Sequencing Platforms

Roche 454 SEQUENCING

GS-FLX System

illumina

Genome Analyzer IIx

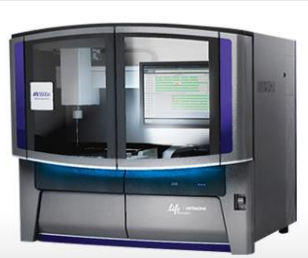
Hiseq 2000/1000

Hiseq 2500/1500

NextSeq

life technologies™

SOLID 3 Plus/4



Genetic Analyzers 5500W Series

Benchtop High-Throughput Sequencing Platforms

Roche 454 SEQUENCING

Roche 454 GS Junior

illumina

MiSeq

iSeq 100

MiniSeq

life technologies™

Ion Proton™ System

Ion PGM™ System

High-Throughput Single Molecule Sequencing Platforms

3ª GENERACIÓN

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

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	●	●	●

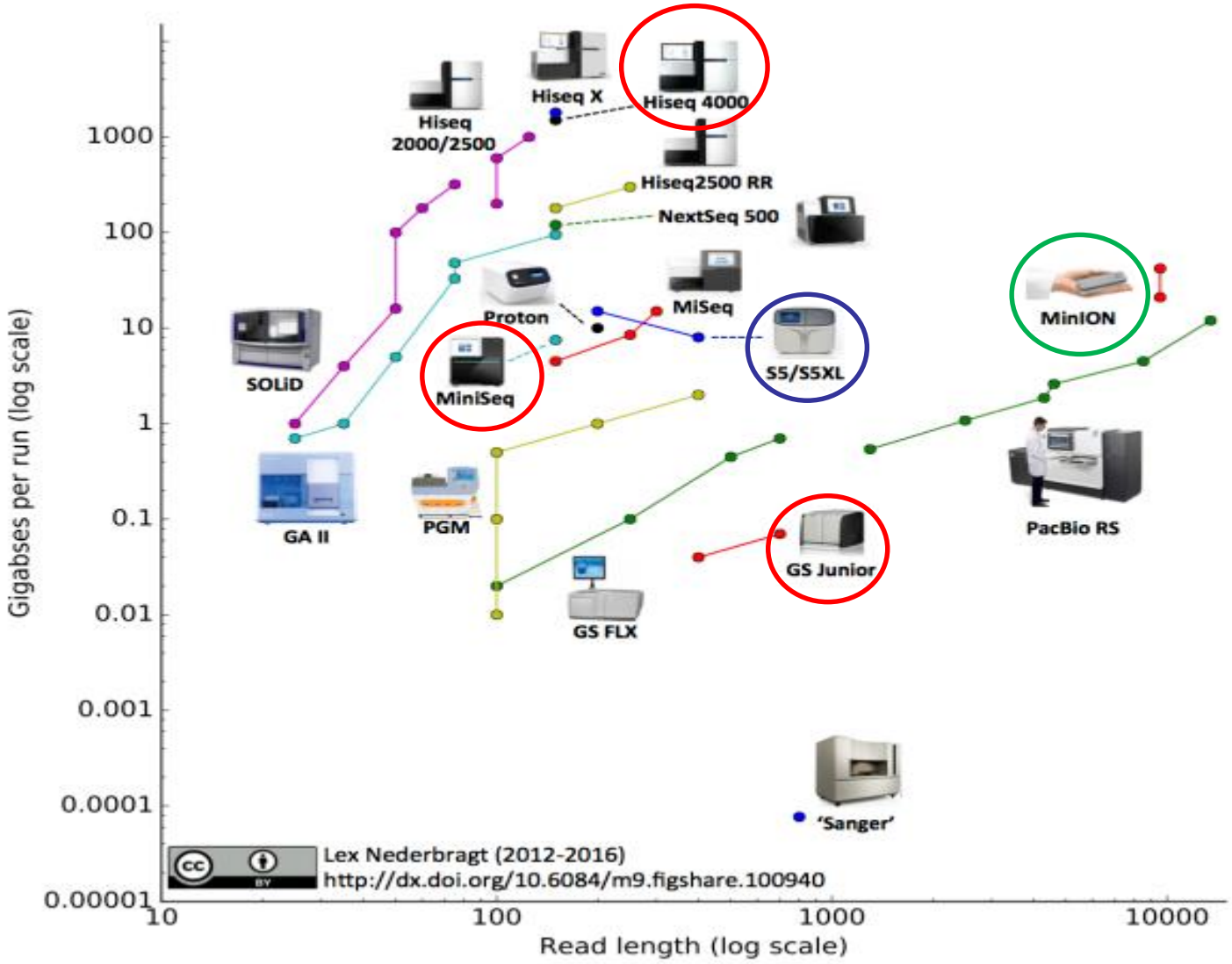
Optimized NGS Sample Tracking and Workflows

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 ›

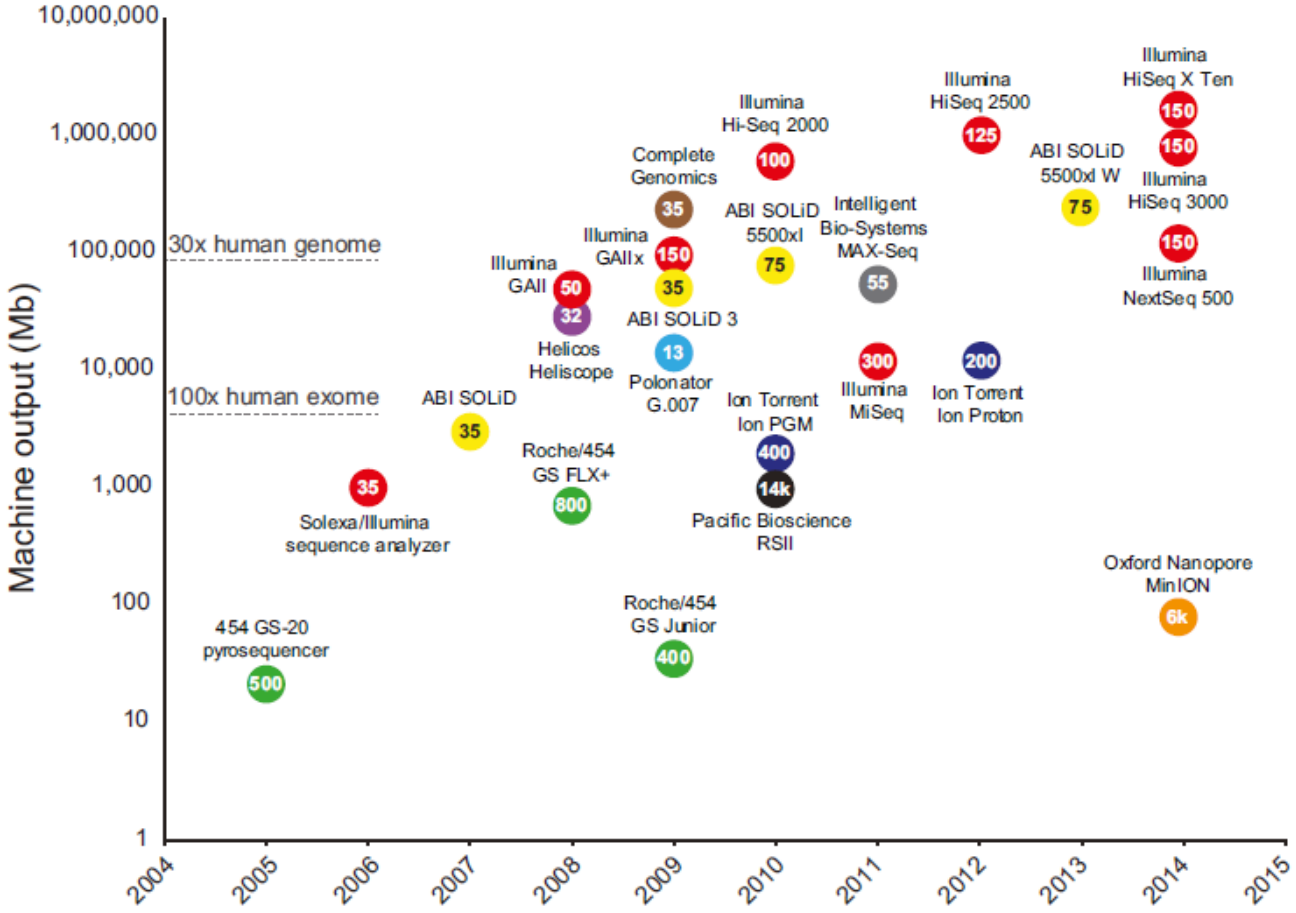
Run Time	12–30 hours	11–48 hours	~13 - 38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)
Maximum Output	120 Gb	330 Gb*	6000 Gb
Maximum Reads Per Run	400 million	1.1 billion*	20 billion
Maximum Read Length	2 × 150 bp	2 × 150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

PLATAFORMAS DE SECUENCIACIÓN. 2016 Edition



https://flxlexblog.wordpress.com/

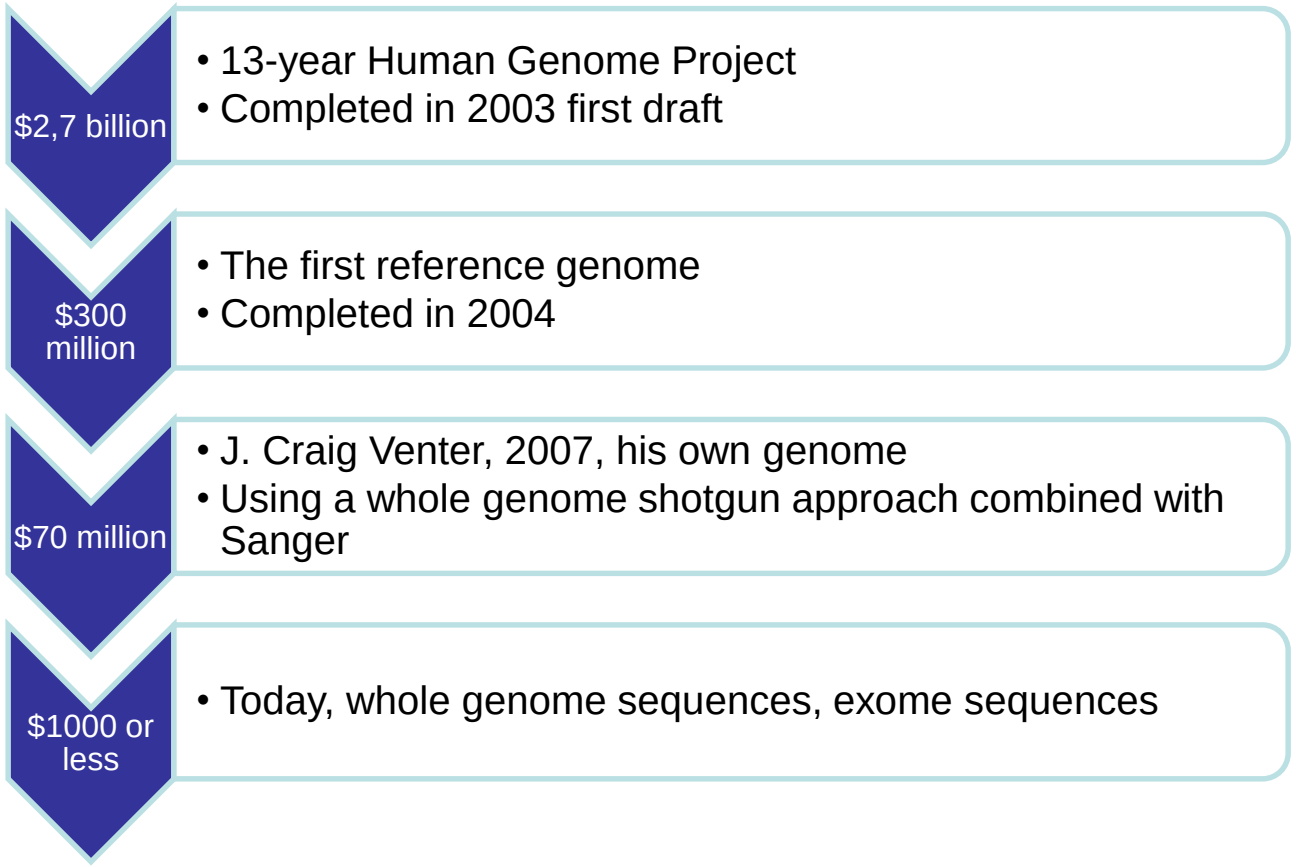
High-Throughput Sequencing Technologies



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

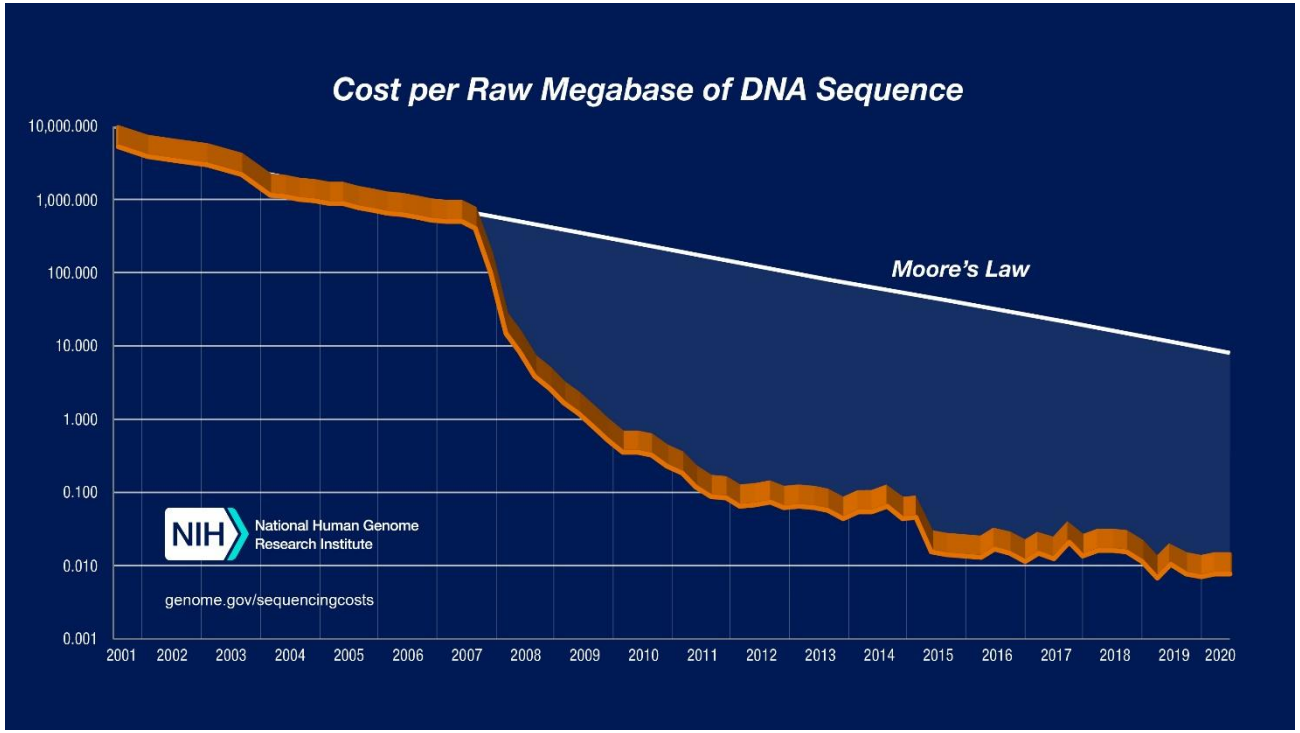
Reuter et al., Mol Cell 2015

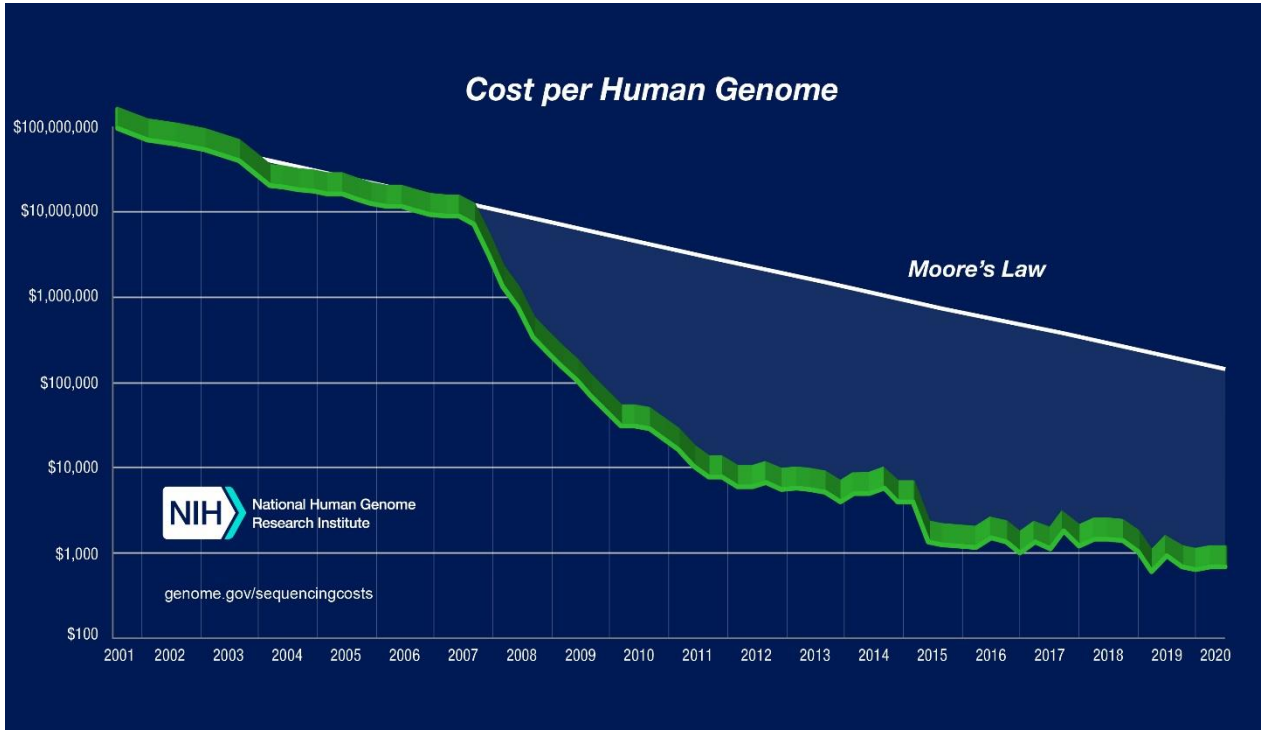
Evolución del coste de la secuenciación de un genoma humano



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

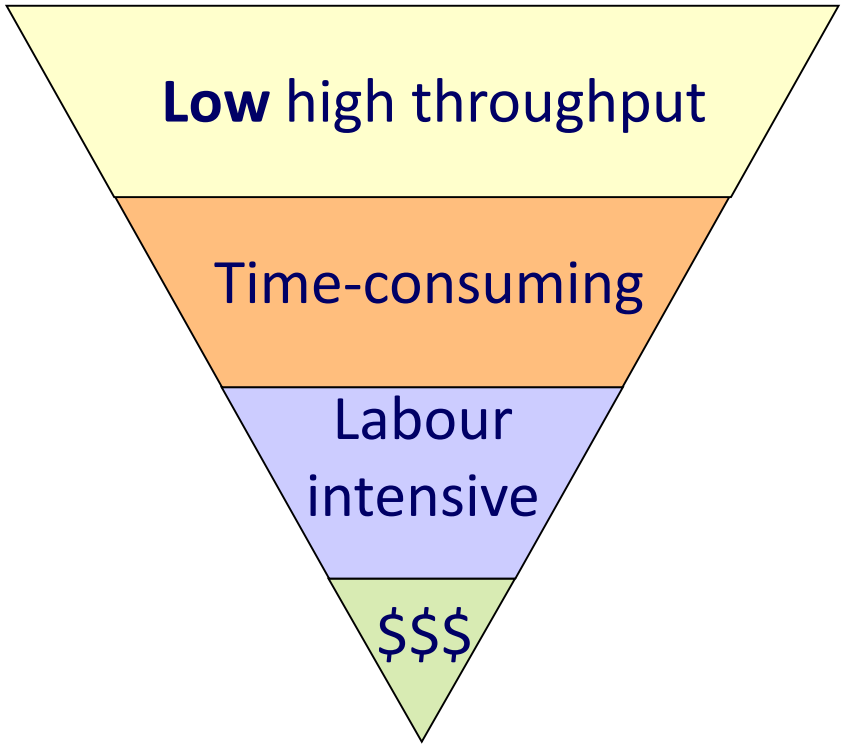
Coste actual de la secuenciación

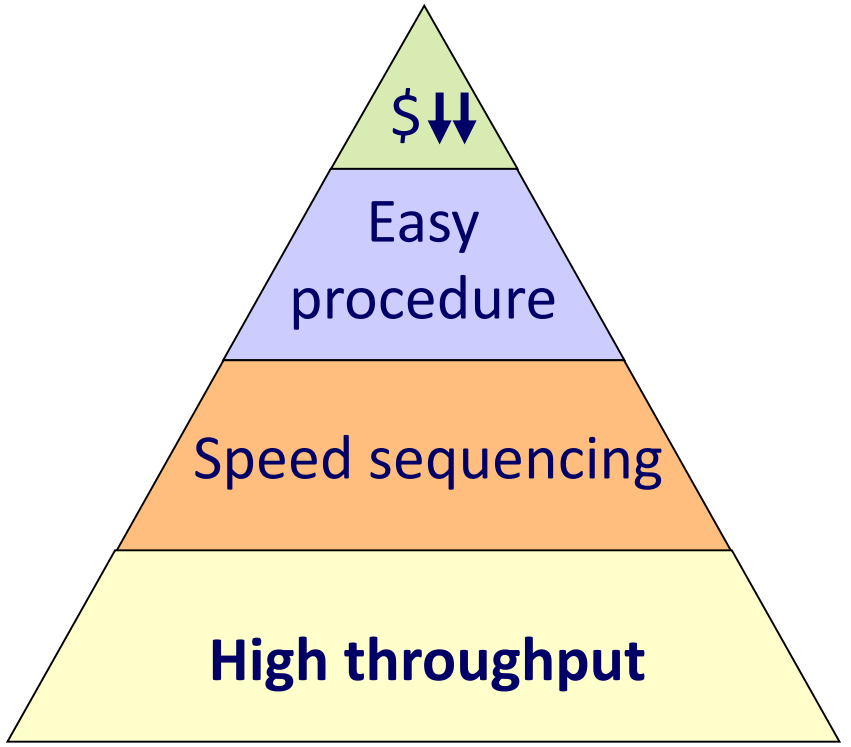




Sanger vs NGS

advantages of new technologies

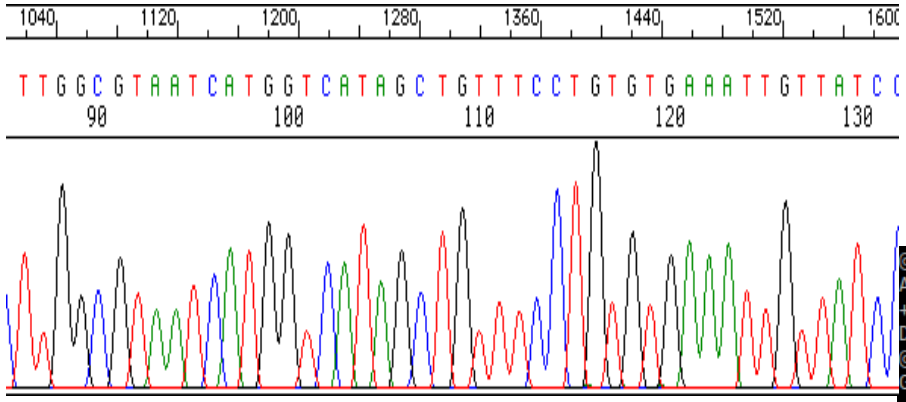




Semiautomatic **Sanger** capillary-based sequencing technology

NGS
Next Generation Sequencing =
Now Generation Sequencing

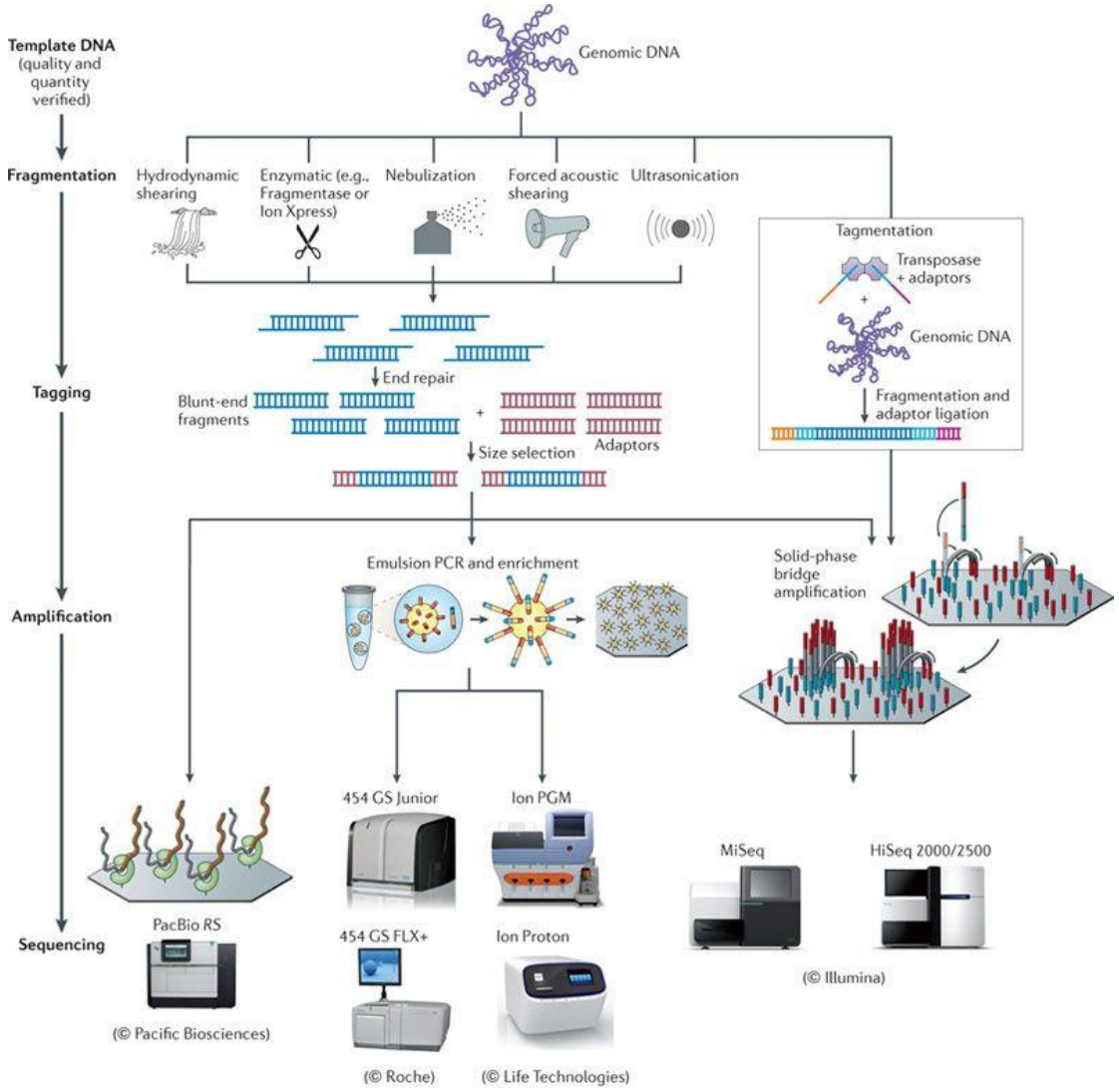
Nuevo escenario en el análisis de datos, BIG DATA




```
@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>HHIHIFGDIIIIIIGFHIEGH
@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;64GGGFGB>GGGBDG<DBGB<DA??/?########################
```

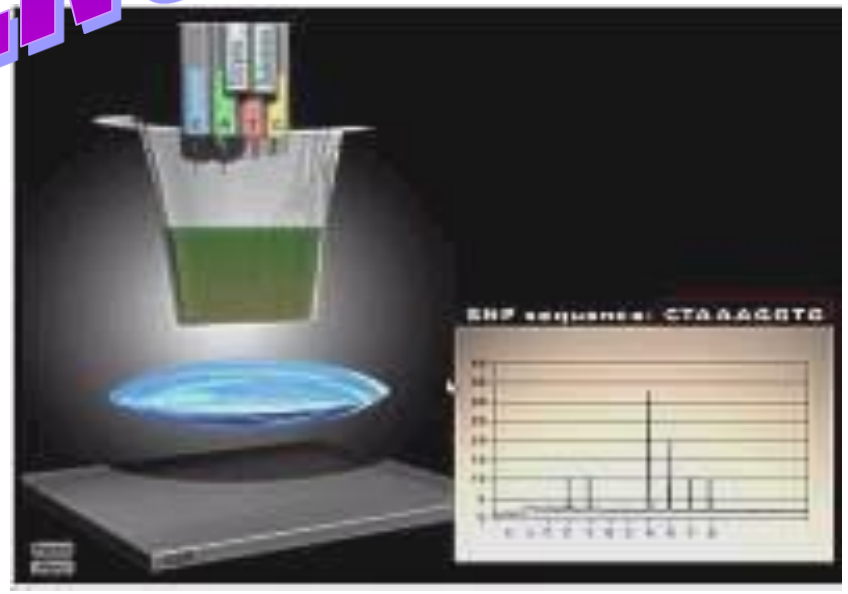
Secuenciadores

High-throughput sequencing platforms

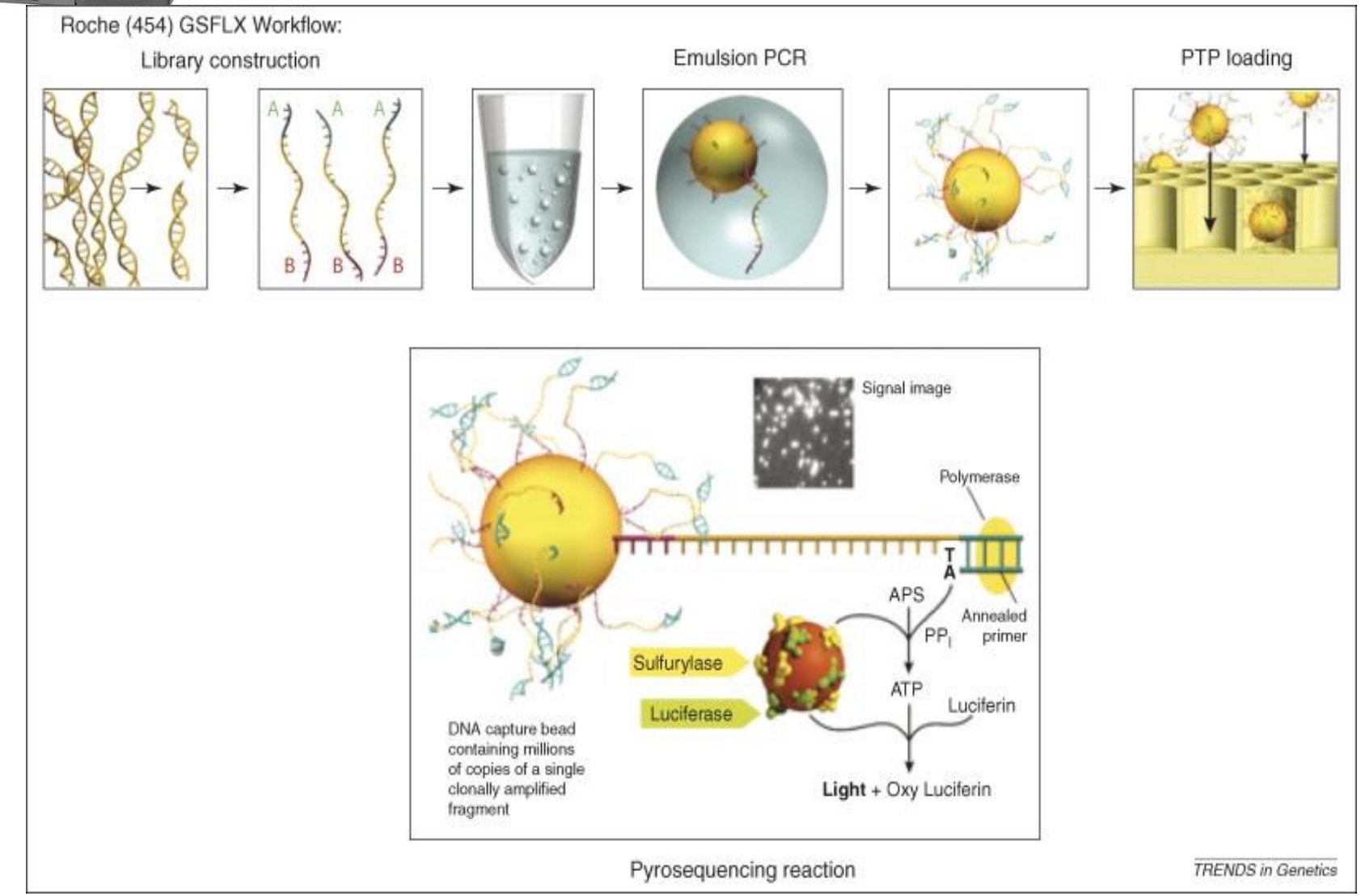


Nature Reviews | Microbiology Loman et al, 2012

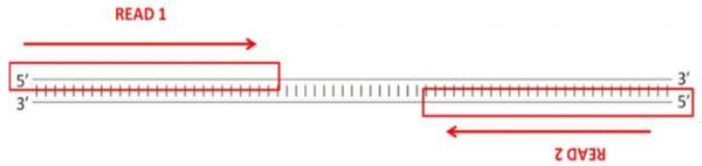
PIROSECUENCIACIÓN

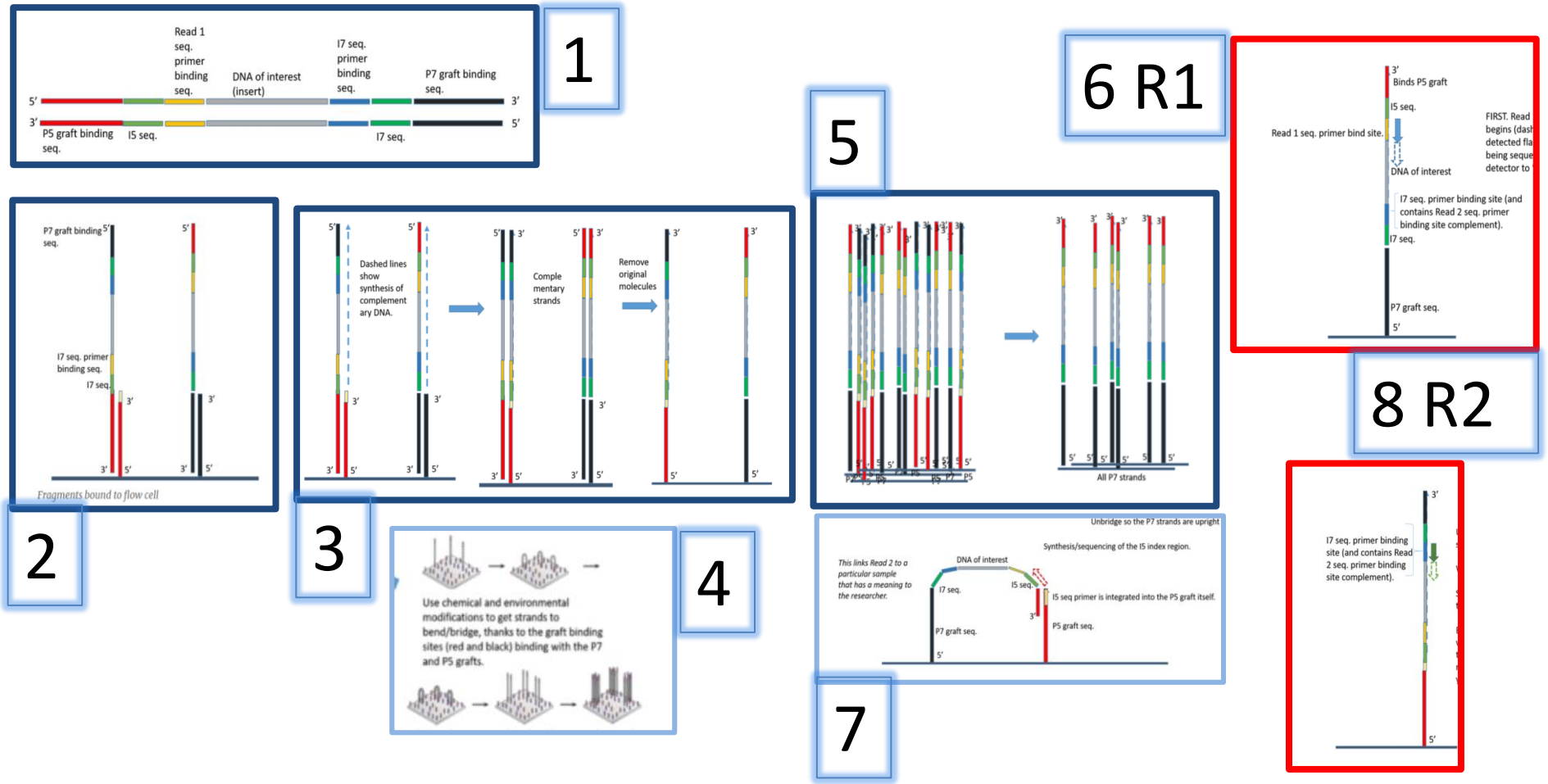


Roche (454) Workflow



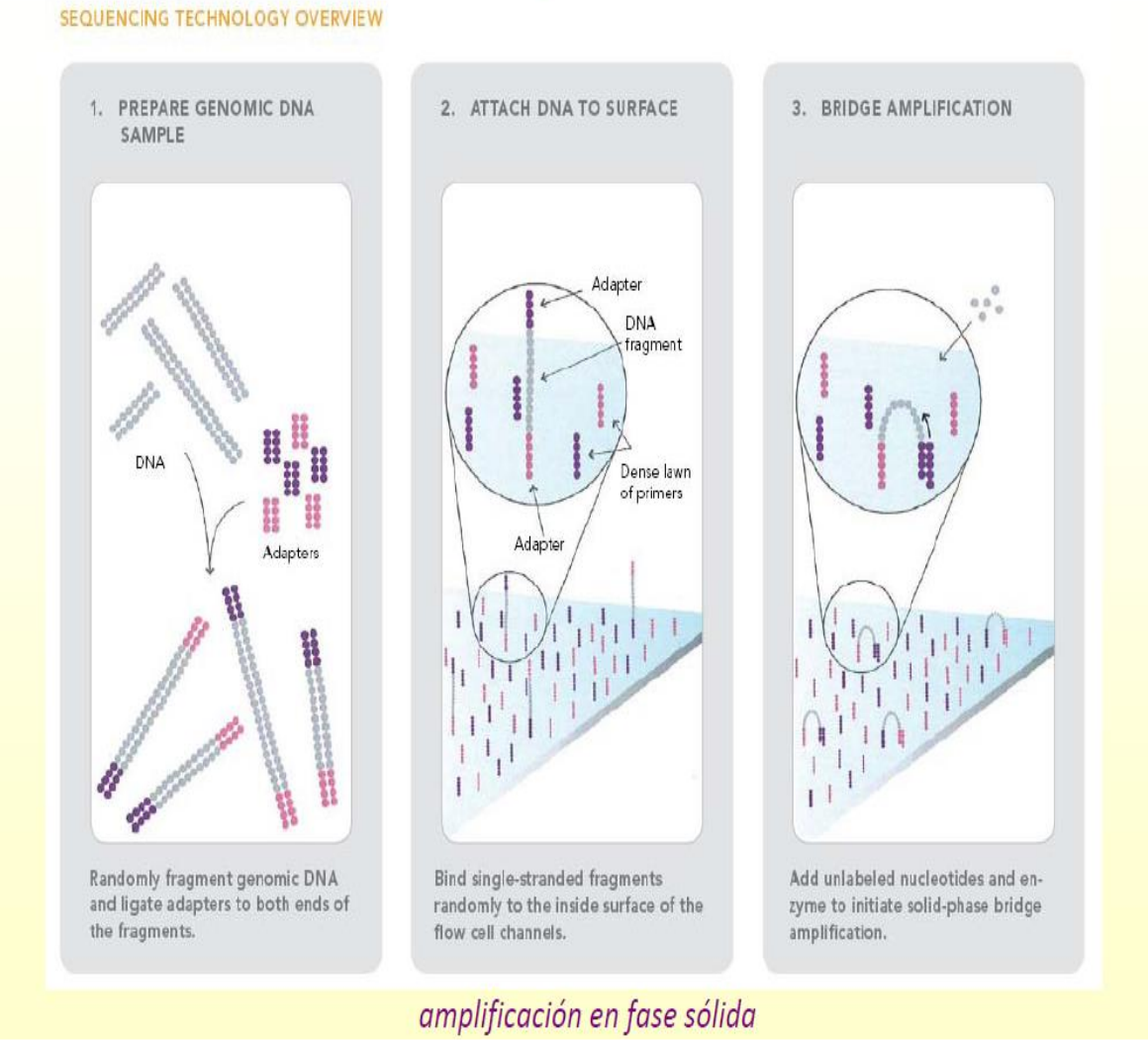
Illumina sequencing





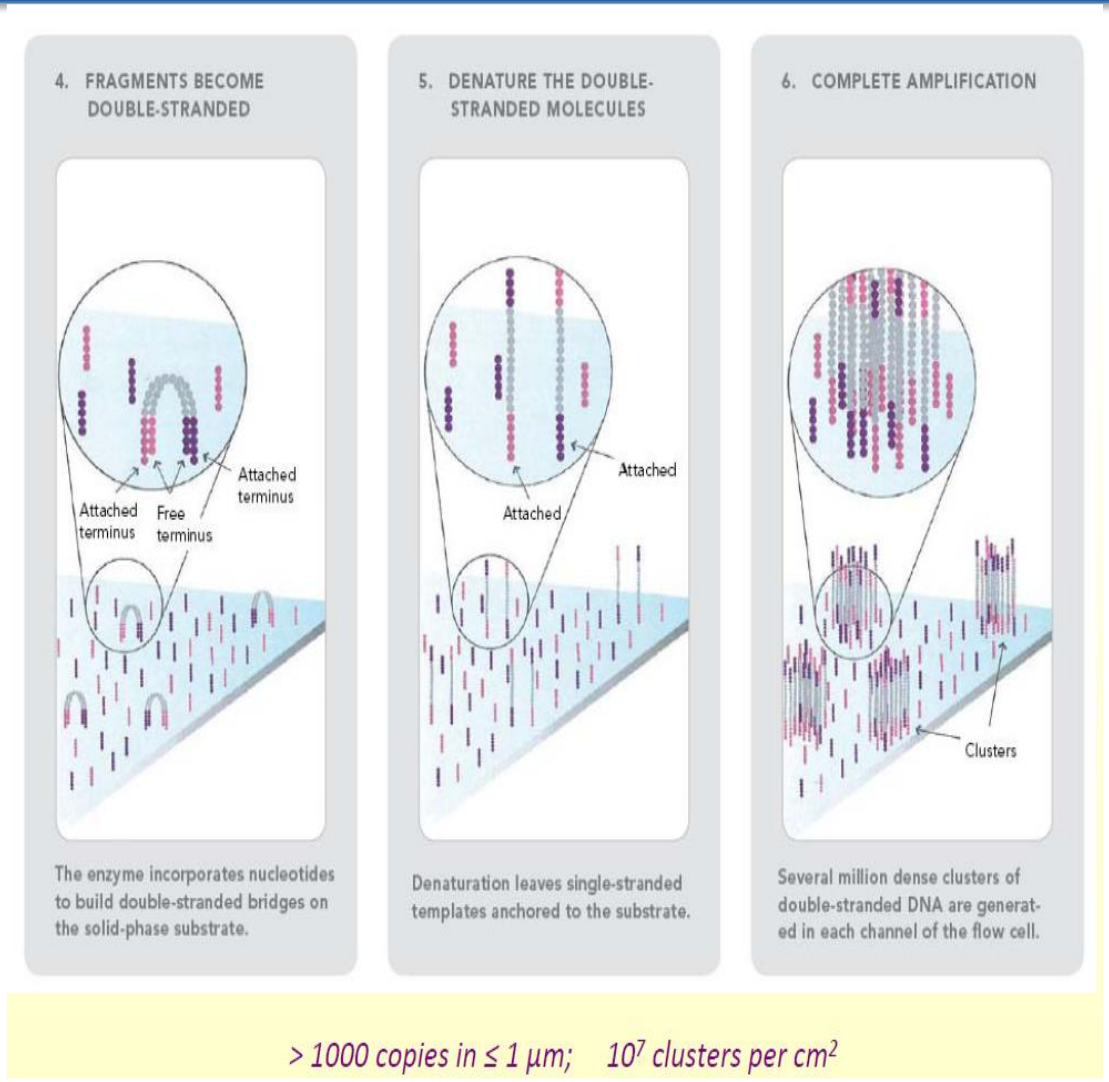
https://kscbioinformatics.wordpress.com/2017/02/13/illumina-sequencing-for-dummies-samples-are-sequenced/

Illumina sequencing



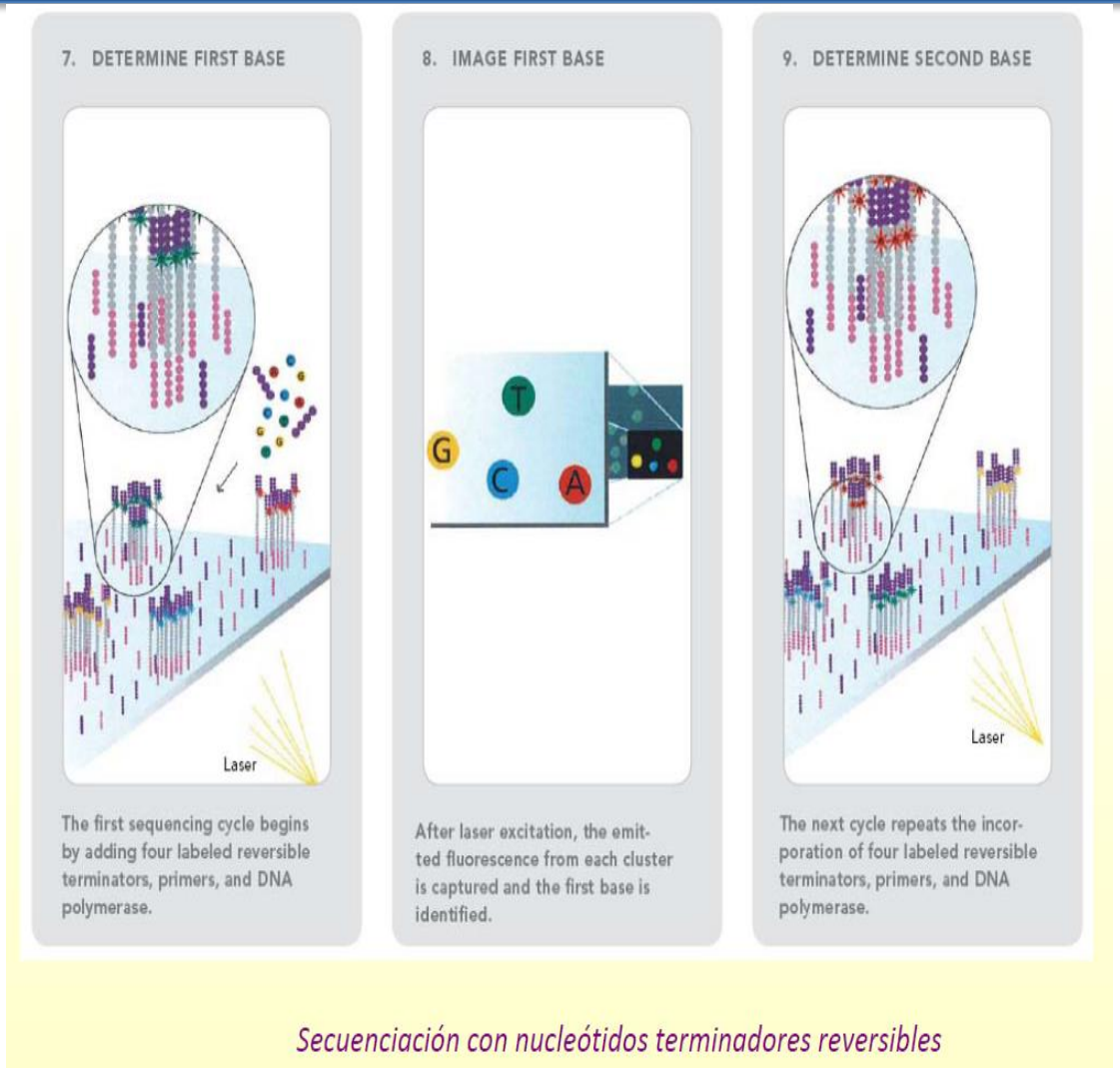
amplificación en fase sólida

Illumina sequencing



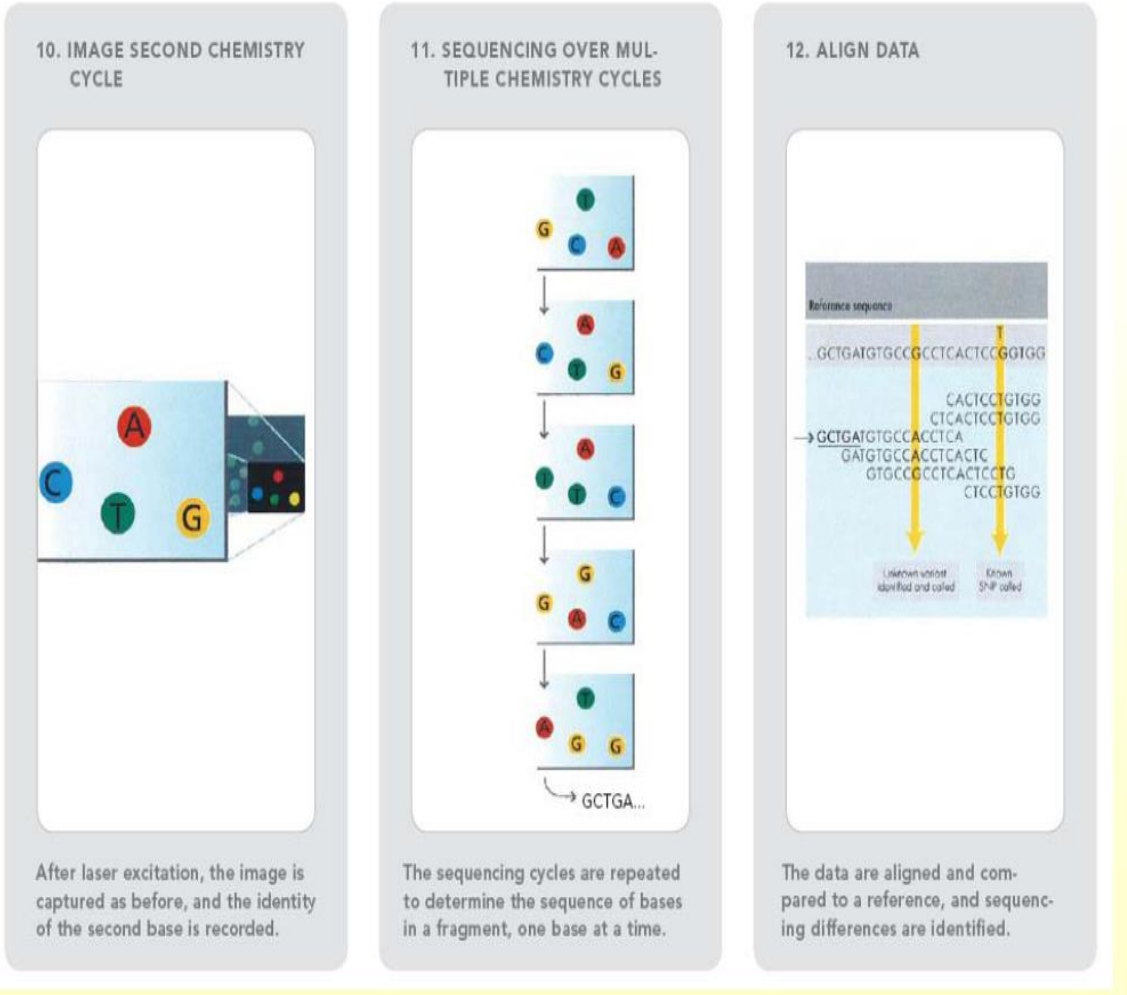
> 1000 copies in ≤ 1 µm; 10^7 clusters per cm^2

Illumina sequencing



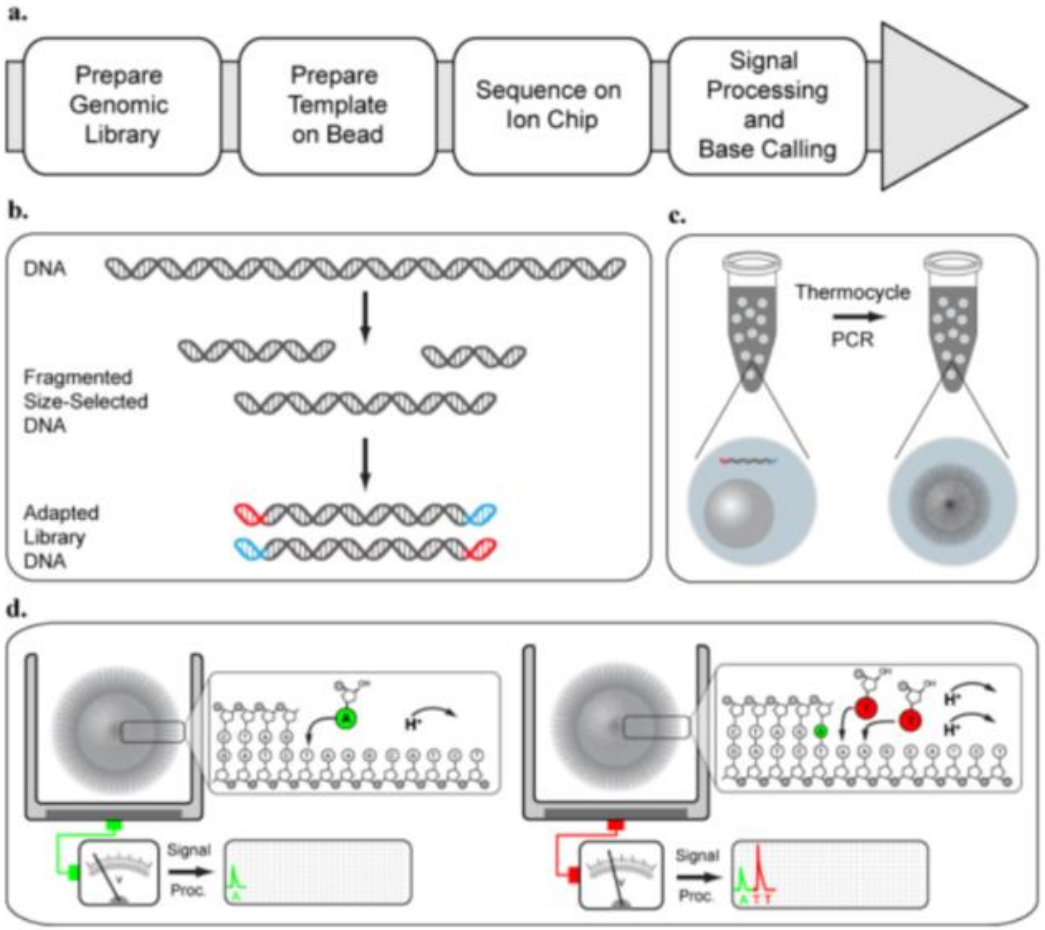
Secuenciación con nucleótidos terminadores reversibles

Illumina sequencing

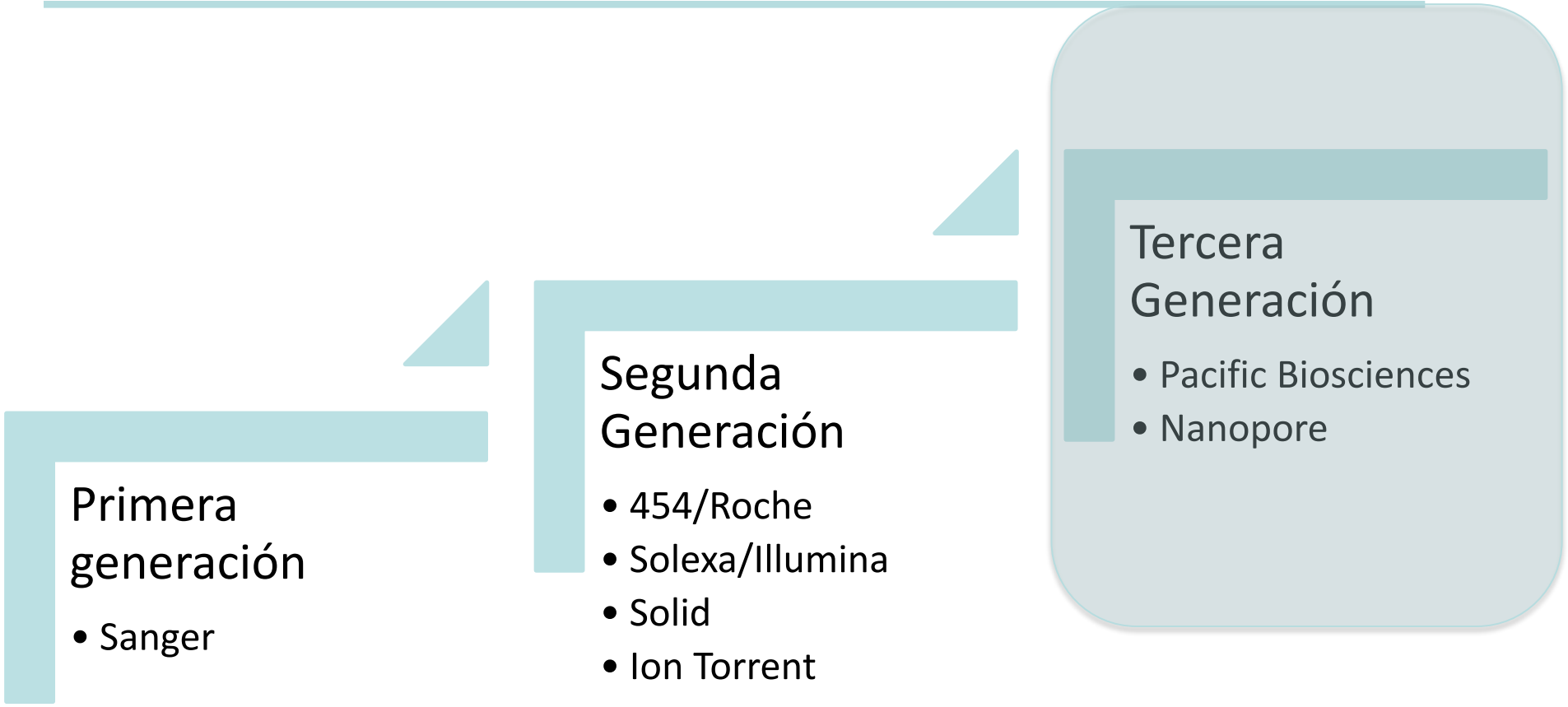


Ion Torrent PGM

Personal Genome Machine



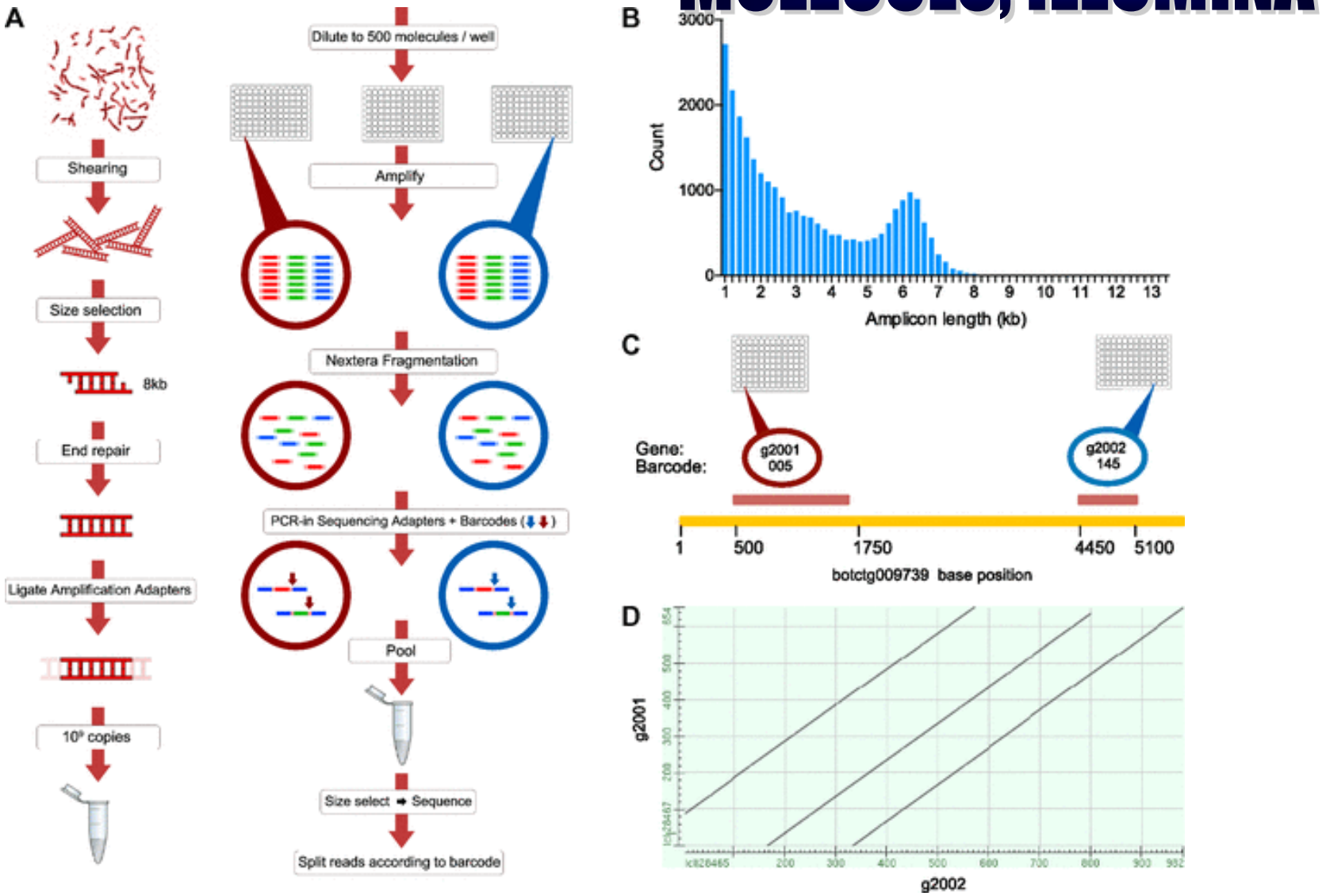
Secuenciadores



3ª GENERACIÓN: LECTURAS MAS LARGAS Y MOLECULA ÚNICA

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

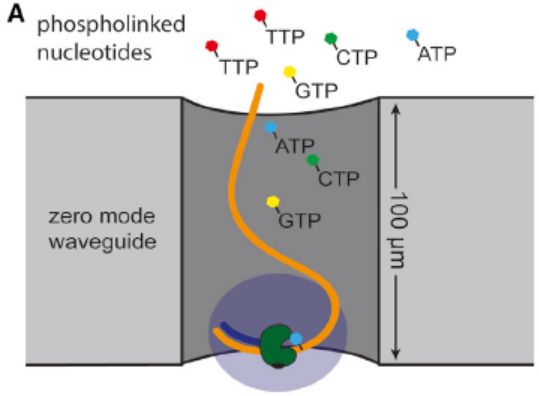
MOLECULO, ILLUMINA

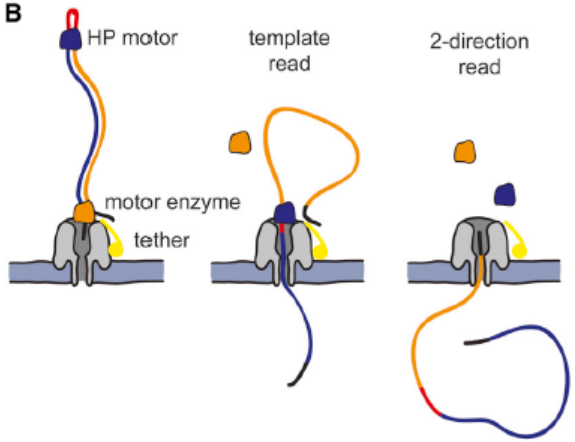


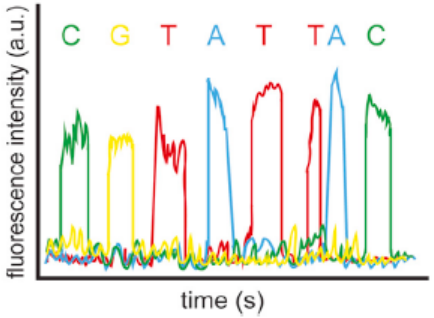
Moleculo, acquired by Illumina in late 2012, developed an innovative technology for generating long reads that combines a new library prep method and genome analysis tools

Voskoboynik et al., elife 2013,2:e00569

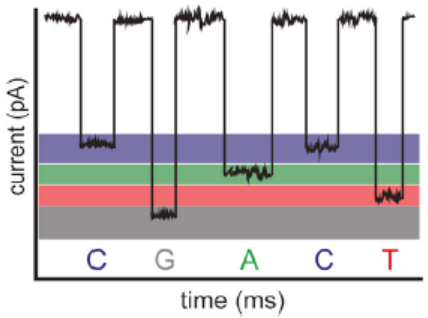
The Third-generation Sequencing Technologies







Pacific Bioscience's SMRT sequencing



Oxford Nanopore's sequencing strategy

Reuter et al., Mol Cell 2015

Single Molecule Sequencing Platforms

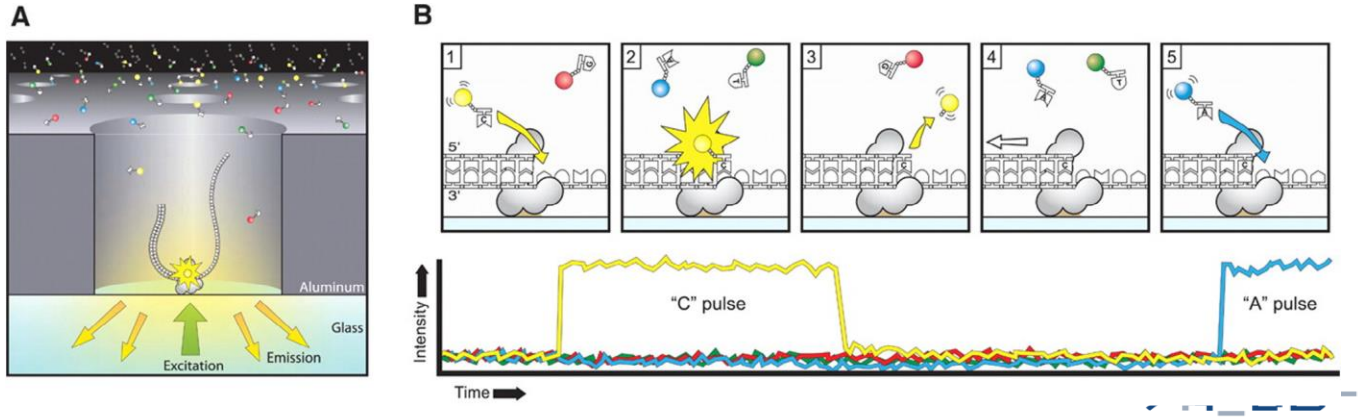
PacBio sequencing and its applications

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

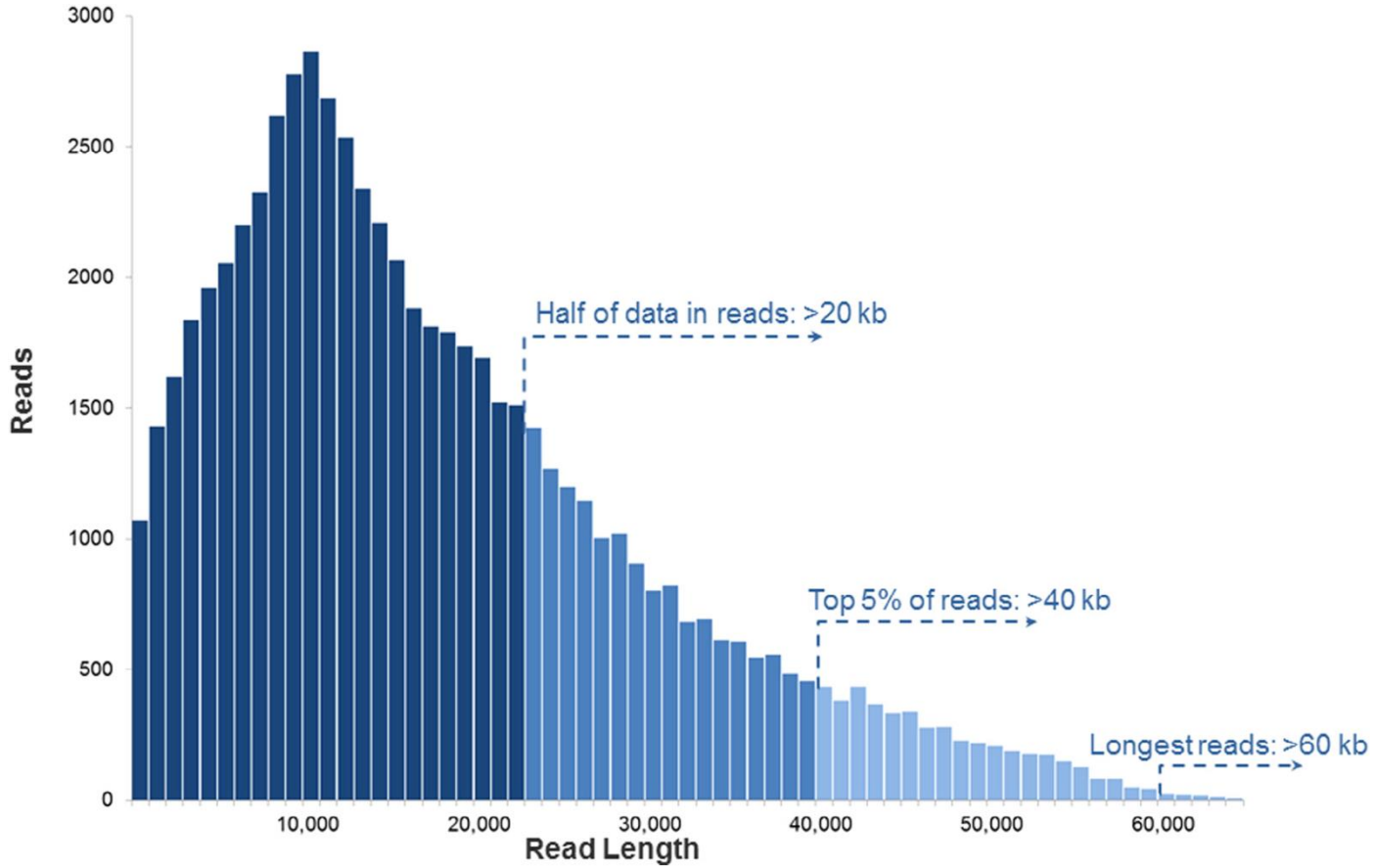
A single SMRT cell: this contains 150000 ZMWs (zero-mode waveguide). A SMRTbell diffuses into a ZMW.
Approx 35000 -75000 ZMWs produce a read in a run lasting 0,5-4h resulting in 0,5-1Gb.



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

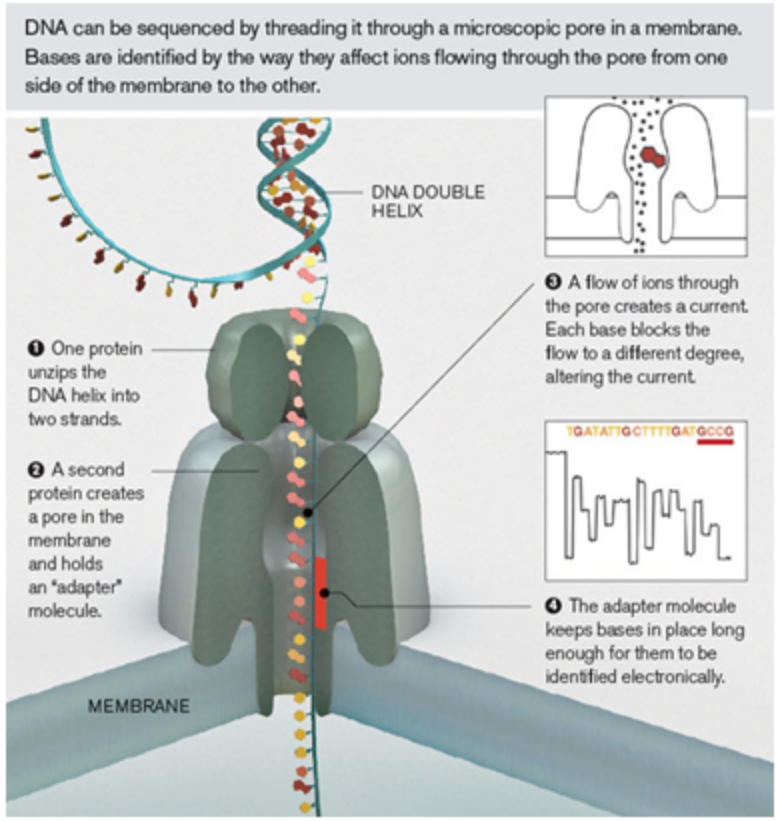
Produced highly-contiguous assemblies of bacterial and eukaryotic genomes

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



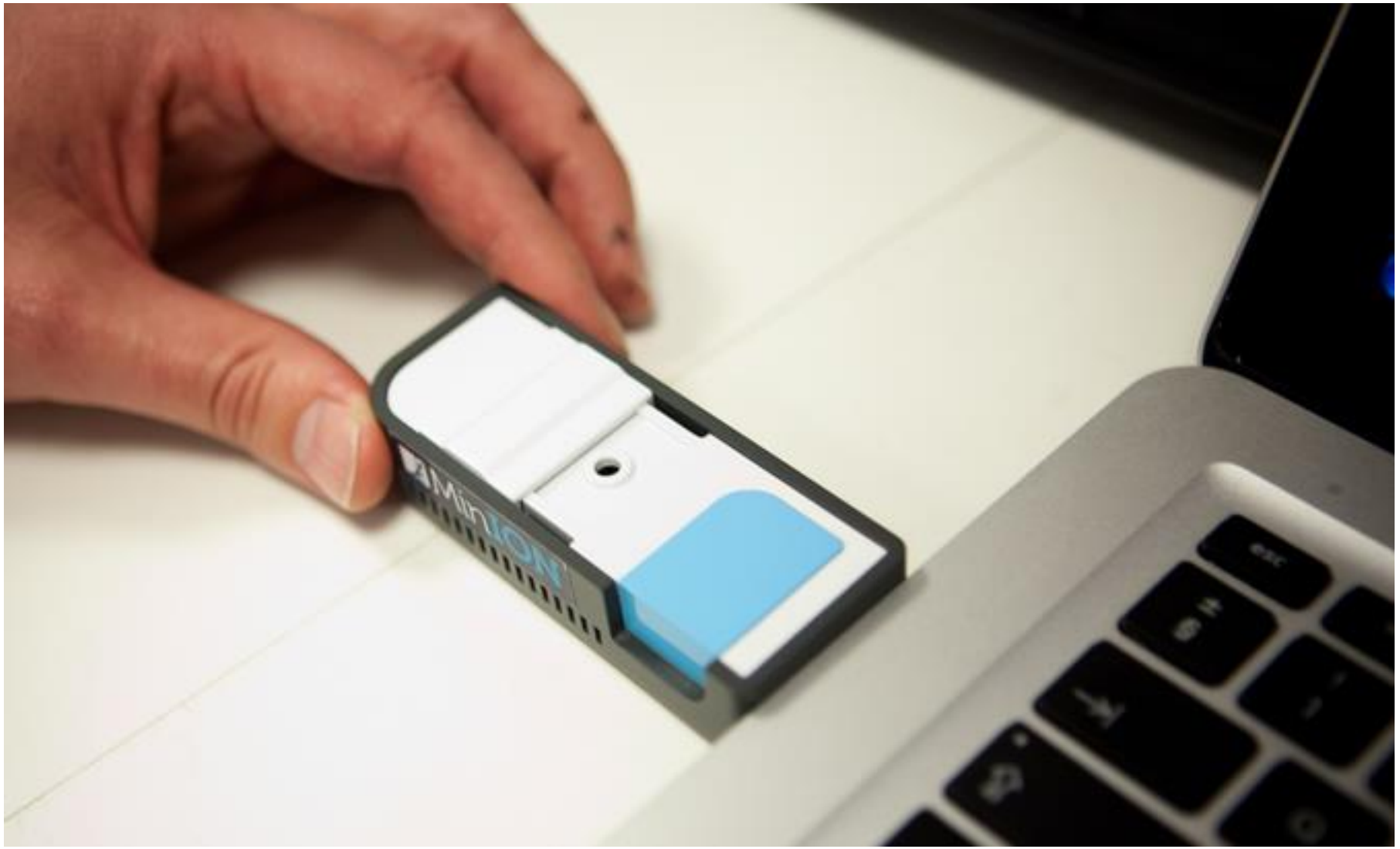
Strand sequencing' is a technique that passes intact DNA polymers through a protein nanopore, sequencing in real time as the DNA translocates the pore.

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

MinIon, OXFORD NANOPORE



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.

Oxford Nanopore Technologies

https://nanoporetech.com/products/comparison

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:

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)

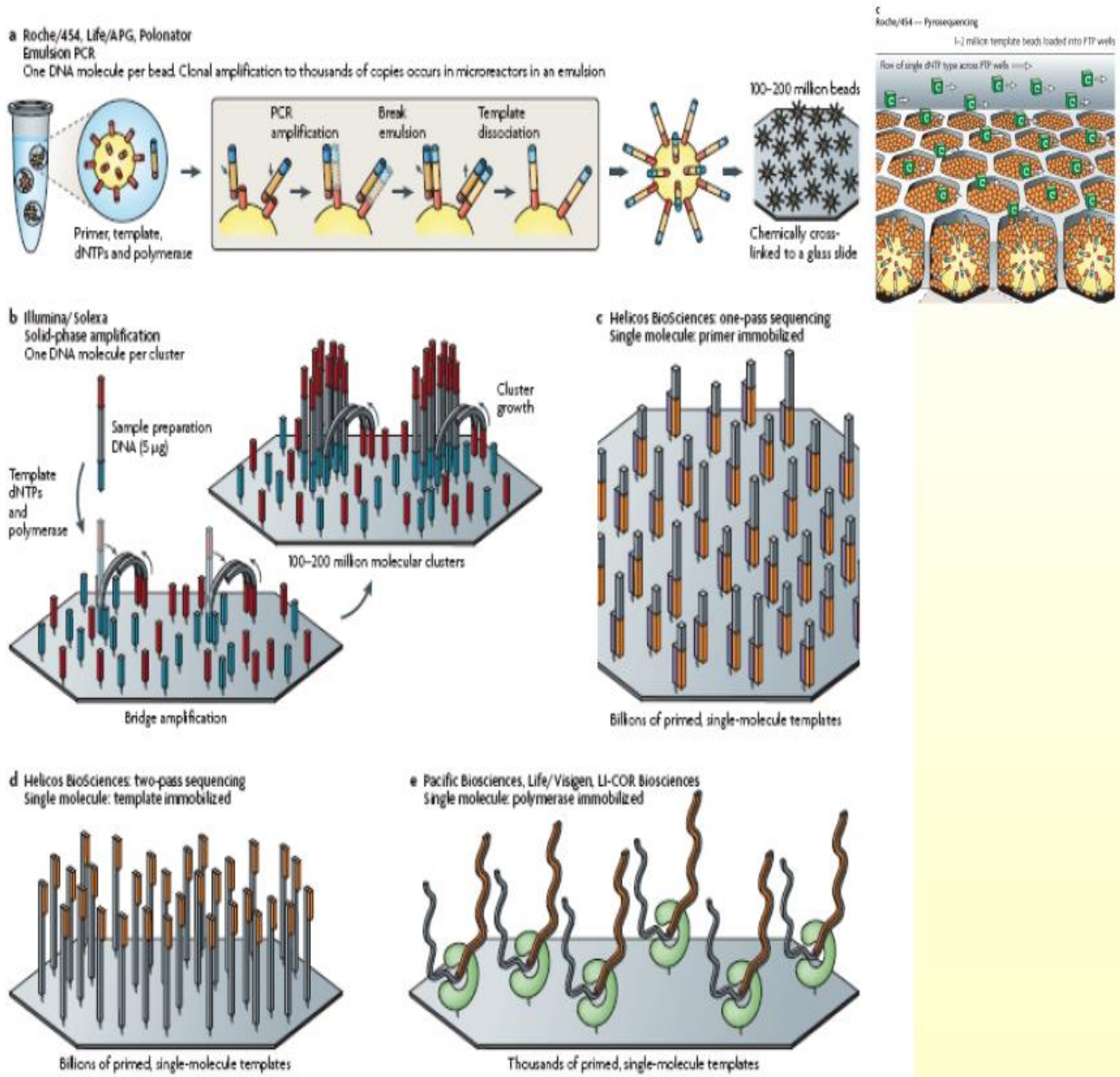
MinIT has now been replaced by the MinION Mk1C, which combines the real-time, portable sequencing of MinION, with powerful integrated compute, a high-resolution touchscreen, and full connectivity.

SmidgION, Mobile analysis

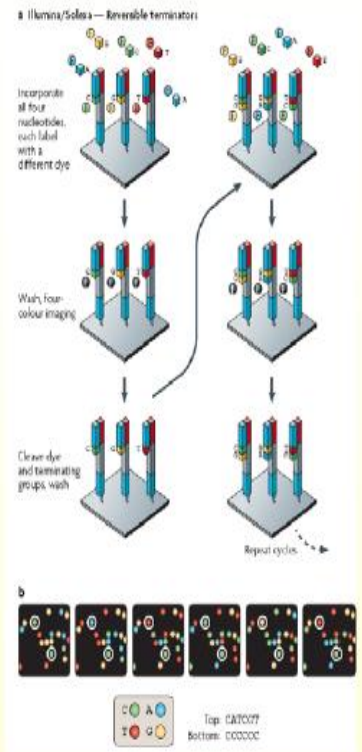
Oxford Nanopore has now started developing an even smaller device, SmidgION.

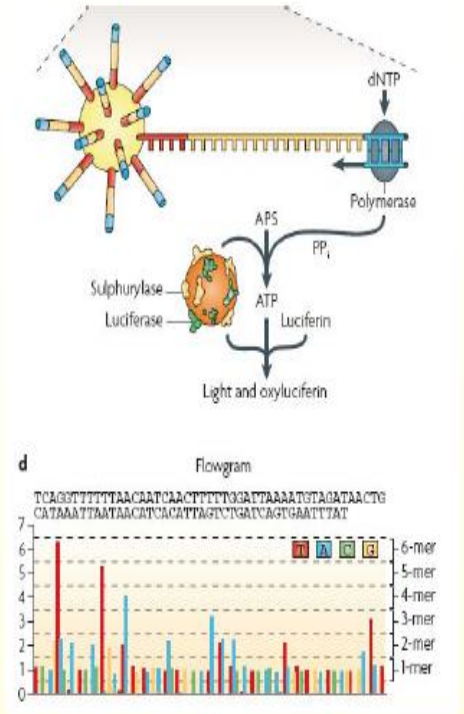
potential applications may include remote monitoring of pathogens in a breakout or infectious disease; the on-site analysis of environmental samples such as water/metagenomics samples, real time species ID for analysis of food, timber, wildlife or even unknown samples; field-based analysis of agricultural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

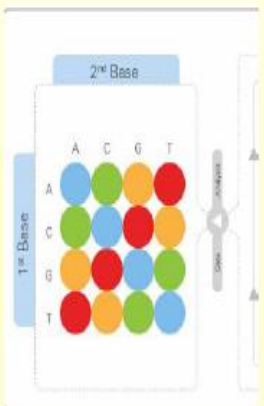
ESTRATEGIAS DE INMOVILIZACIÓN

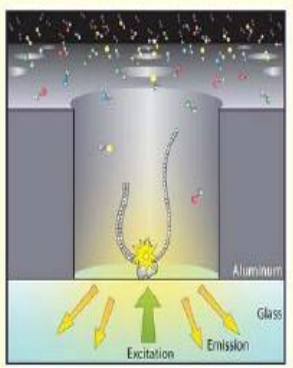


ESTRATEGIAS DE LECTURA









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,

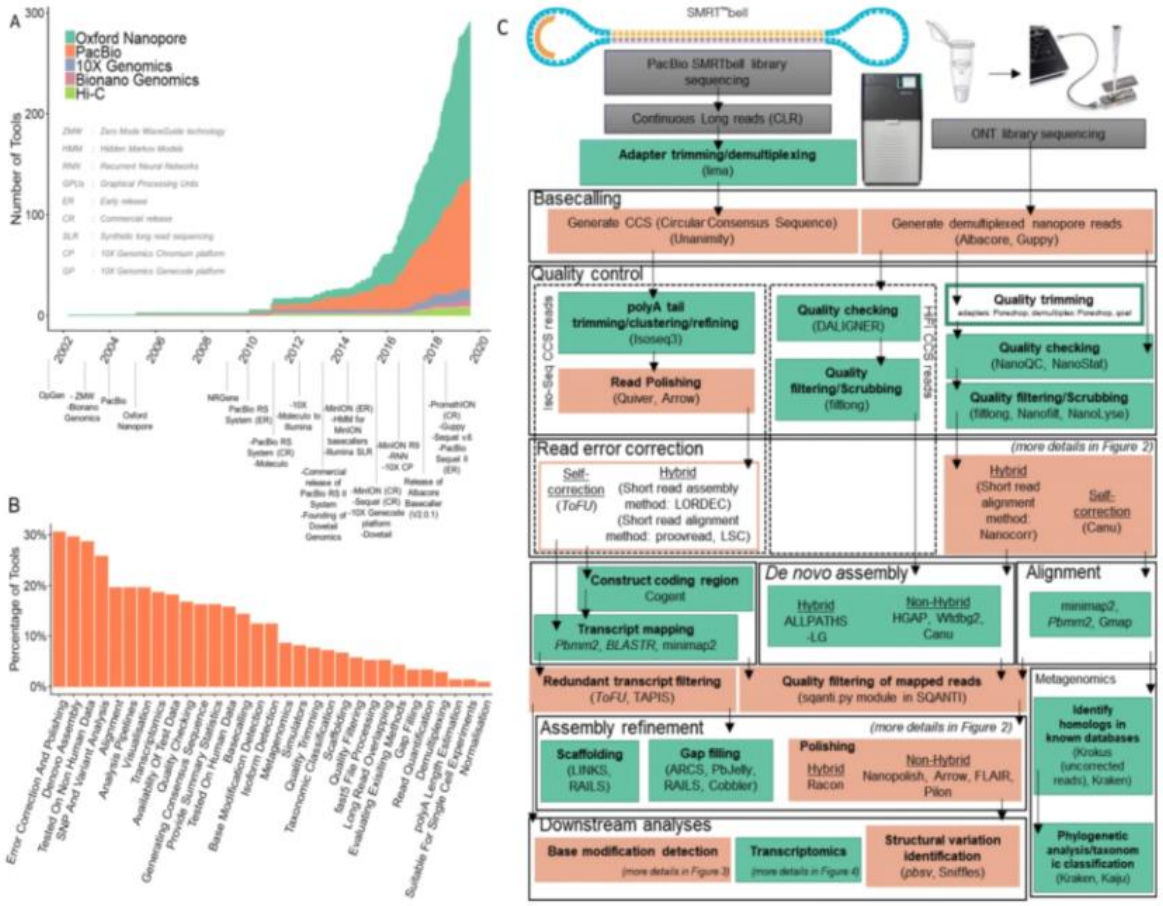
[a] The throughput ranges are determined by available kits and run modes on a per run basis. As an example of a 15-GB throughput, thirty-five 5-MB genomes can be sequenced to a minimum coverage of 40× on the Illumina MiSeq using the v3 600 cycle chemistry.
[b] Per one single-molecule real-time cell.
[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

Long-read sequencing data analysis

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

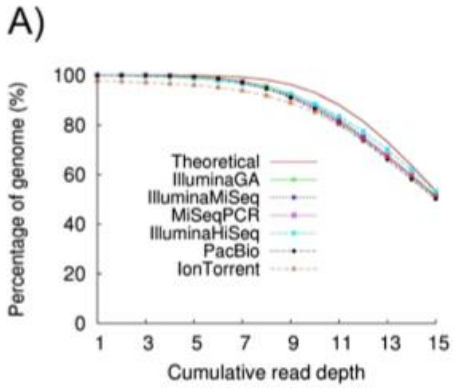
From: Opportunities and challenges in long-read sequencing data analysis



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

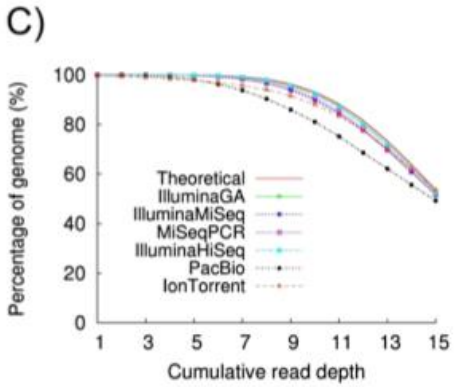
Uniformidad de cobertura a lo largo del genoma

%GC



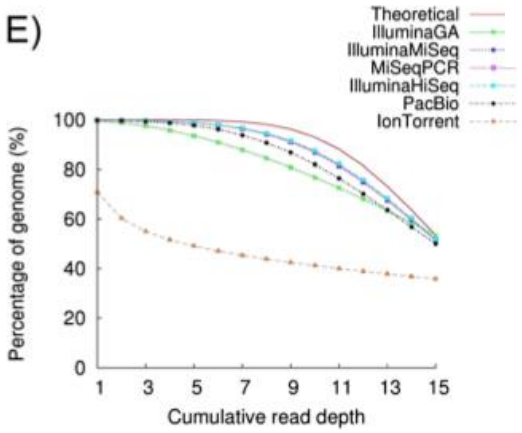
Bordetella pertussis

%GC



Staphylococcus aureus

%AT



Plasmodium falciparum

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

Variación en la cobertura dependiendo del Sistema de Secuenciación

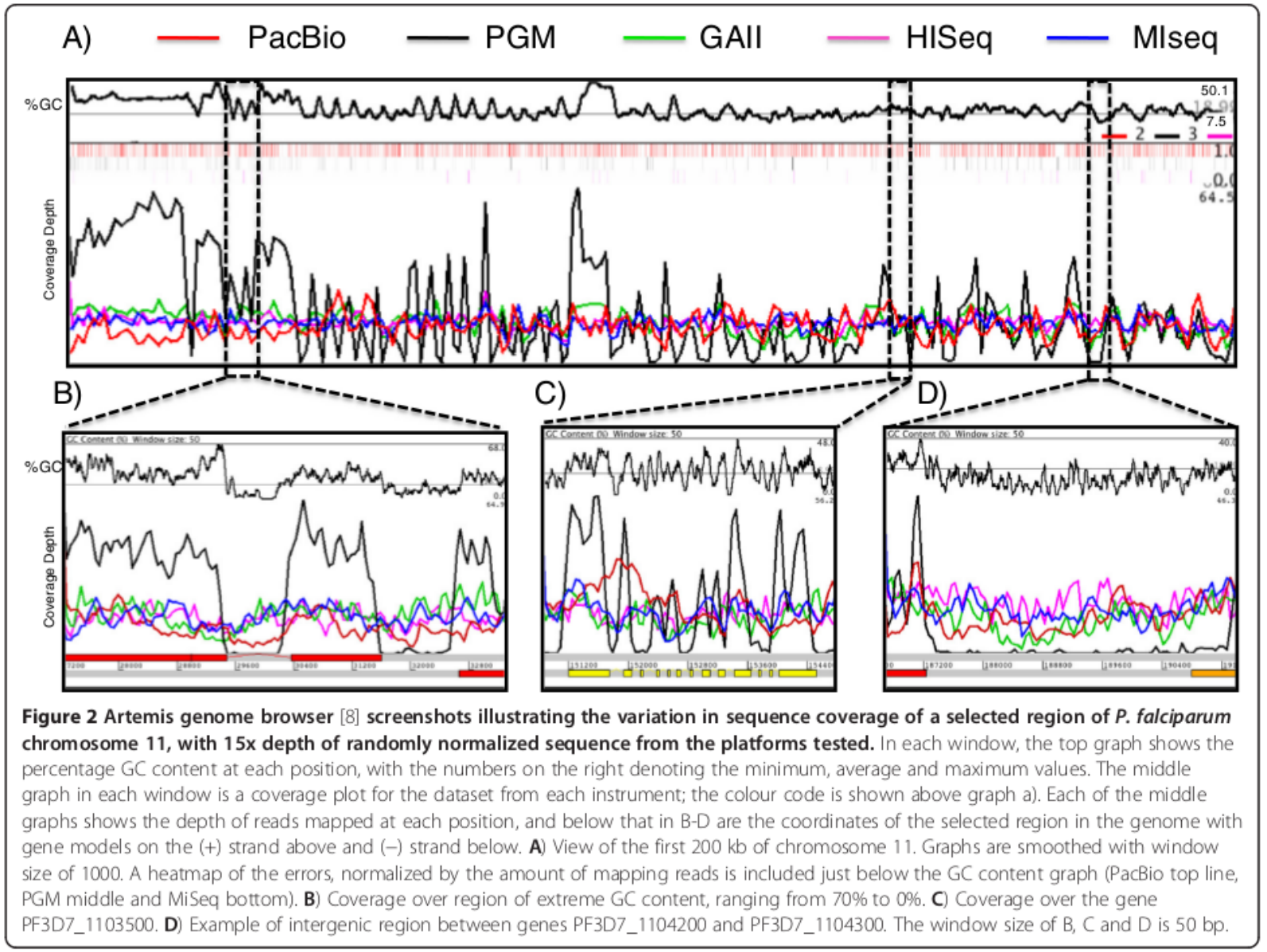


Figure 2 Artemis genome browser [8] **screenshots illustrating the variation in sequence coverage of a selected region of *P. falciparum* chromosome 11, with 15x depth of randomly normalized sequence from the platforms tested.** In each window, the top graph shows the percentage GC content at each position, with the numbers on the right denoting the minimum, average and maximum values. The middle graph in each window is a coverage plot for the dataset from each instrument; the colour code is shown above graph a). Each of the middle graphs shows the depth of reads mapped at each position, and below that in B-D are the coordinates of the selected region in the genome with gene models on the (+) strand above and (–) strand below. **A**) View of the first 200 kb of chromosome 11. Graphs are smoothed with window size of 1000. A heatmap of the errors, normalized by the amount of mapping reads is included just below the GC content graph (PacBio top line, PGM middle and MiSeq bottom). **B**) Coverage over region of extreme GC content, ranging from 70% to 0%. **C**) Coverage over the gene PF3D7_1103500. **D**) Example of intergenic region between genes PF3D7_1104200 and PF3D7_1104300. The window size of B, C and D is 50 bp.

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

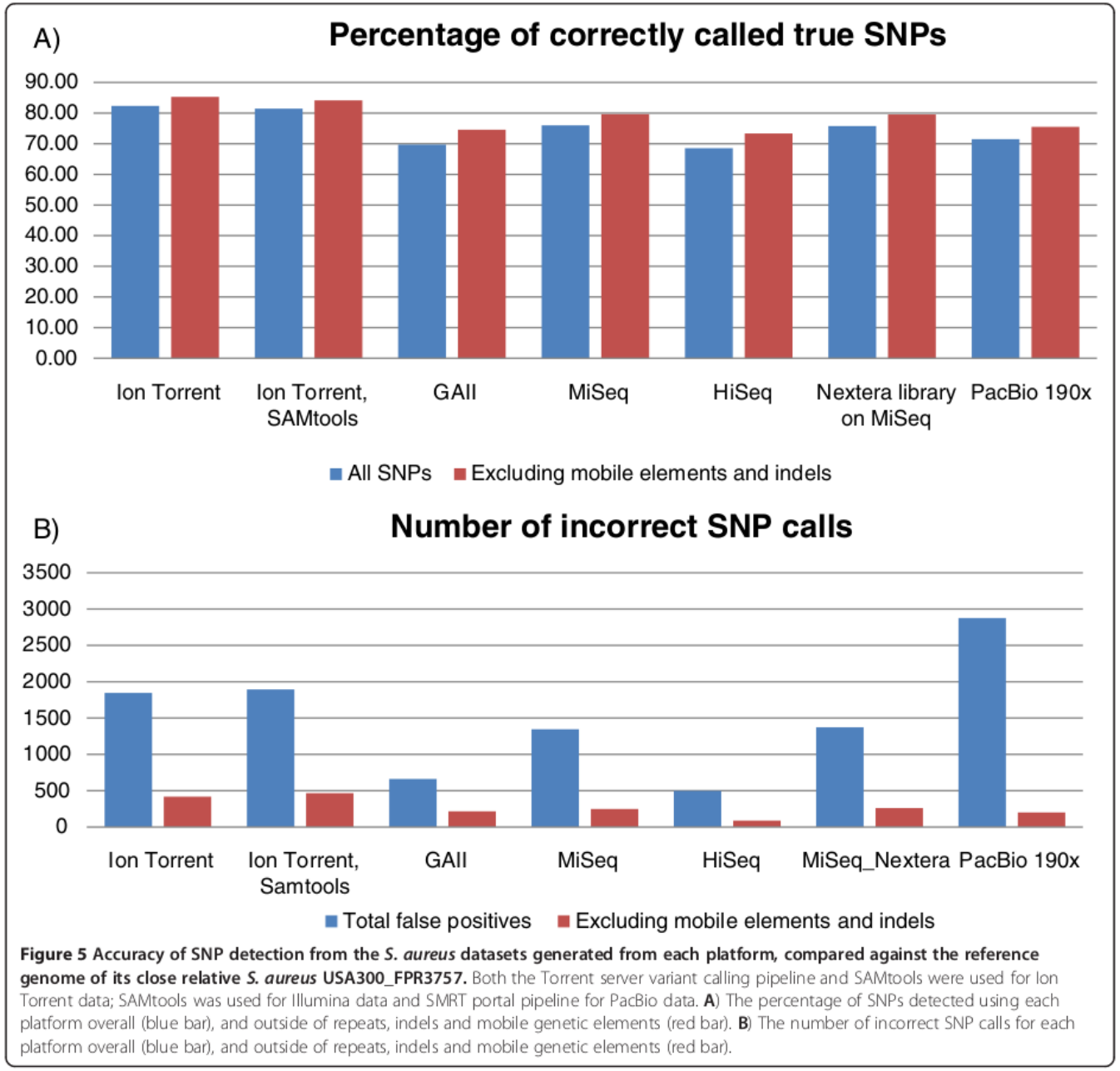


Figure 5 Accuracy of SNP detection from the *S. aureus* datasets generated from each platform, compared against the reference genome of its close relative *S. aureus* USA300_FPR3757. Both the Torrent server variant calling pipeline and SAMtools were used for Ion Torrent data; SAMtools was used for Illumina data and SMRT portal pipeline for PacBio data. **A**) The percentage of SNPs detected using each platform overall (blue bar), and outside of repeats, indels and mobile genetic elements (red bar). **B**) The number of incorrect SNP calls for each platform overall (blue bar), and outside of repeats, indels and mobile genetic elements (red bar).

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

Errores específicos de plataforma

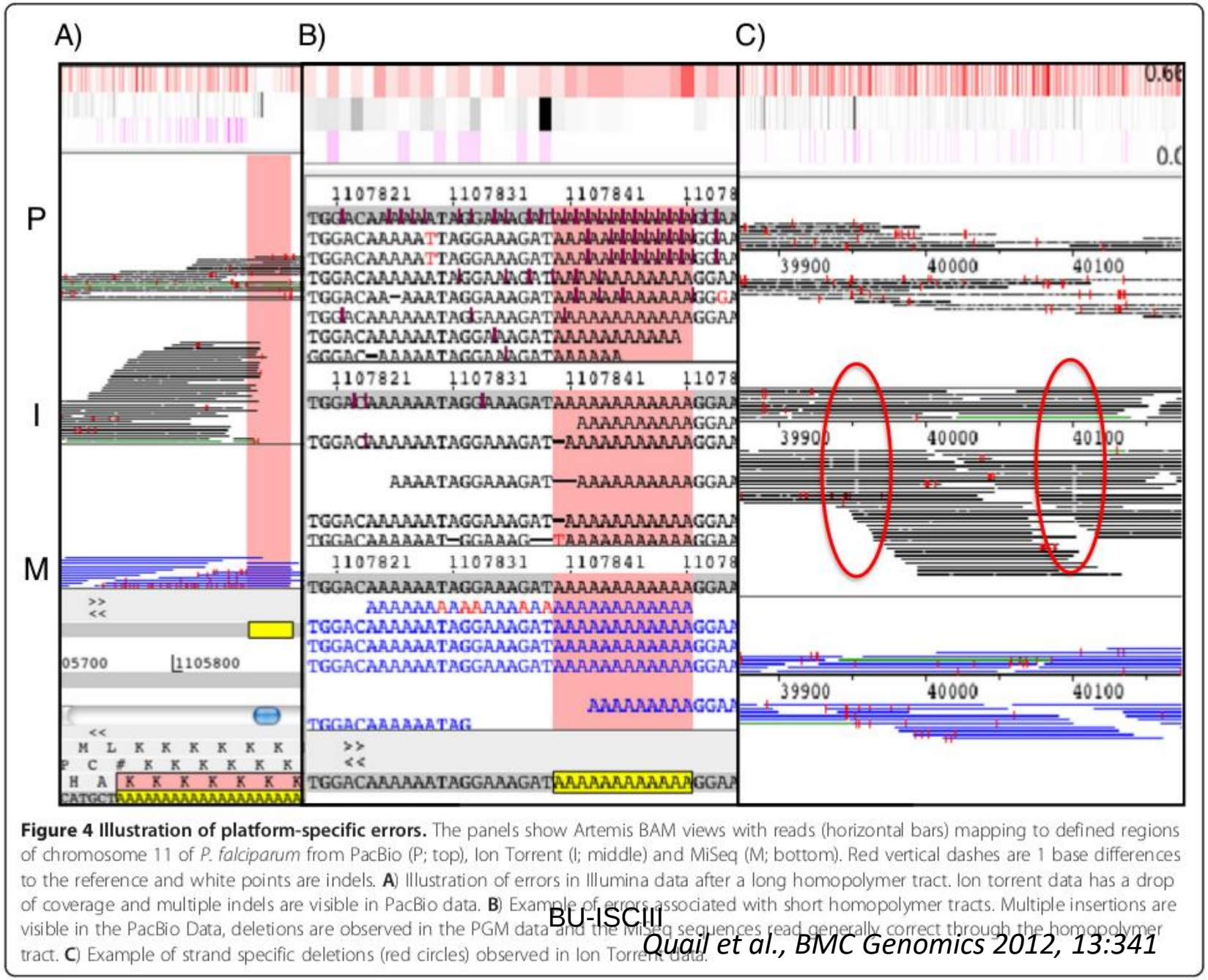


Figure 4 Illustration of platform-specific errors. The panels show Artemis BAM views with reads (horizontal bars) mapping to defined regions of chromosome 11 of *P. falciparum* from PacBio (P; top), Ion Torrent (I; middle) and MiSeq (M; bottom). Red vertical dashes are 1 base differences to the reference and white points are indels. **A**) Illustration of errors in Illumina data after a long homopolymer tract. Ion torrent data has a drop of coverage and multiple indels are visible in PacBio data. **B**) Example of errors associated with short homopolymer tracts. Multiple insertions are visible in the PacBio Data, deletions are observed in the PGM data and the MiSeq sequences read generally correct through the homopolymer tract. **C**) Example of strand specific deletions (red circles) observed in Ion Torrent data.

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

Conclusiones

A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers
Quail et al., BMC Genomics 2012, 13:341

- Ion Torrent no es recomendable para secuenciar genomas con bajo contenido en GC
- Pac Bio:
  - Requiere elevada cantidad de DNA (no amplificación)
  - No recomendable para aplicaciones de counting (RNAseq, Chipseq, exoma)
  - No valido para identificar SNPs
  - Necesario adaptar software para aprovechar la longitud de lectura en el assembly.
- Aplicaciones dependientes de plataforma

Conclusiones

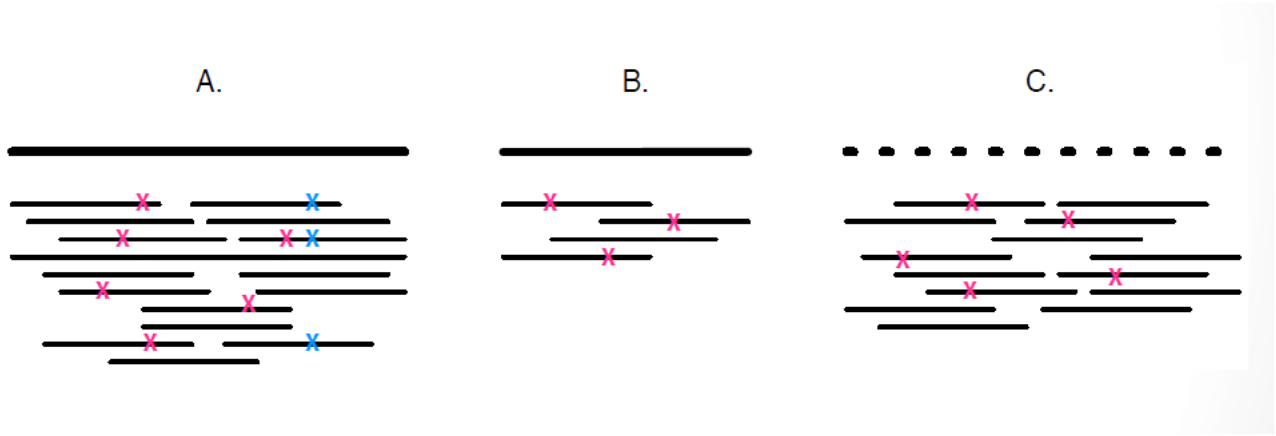
A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers
Quail et al., BMC Genomics 2012, 13:341

- Conclusiones obtenidas con reactivos y plataformas a 2011
- Hay errores intrínsecos a la plataforma
- Otros errores se solucionaran con la actualización de las plataformas

Algunos conceptos en secuenciación

Básicamente tres problemas

Resecuenciación, Conteo y ensamblado

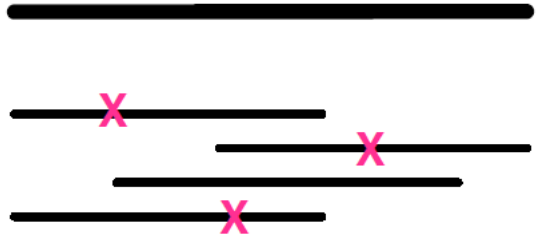


Resecuenciación

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

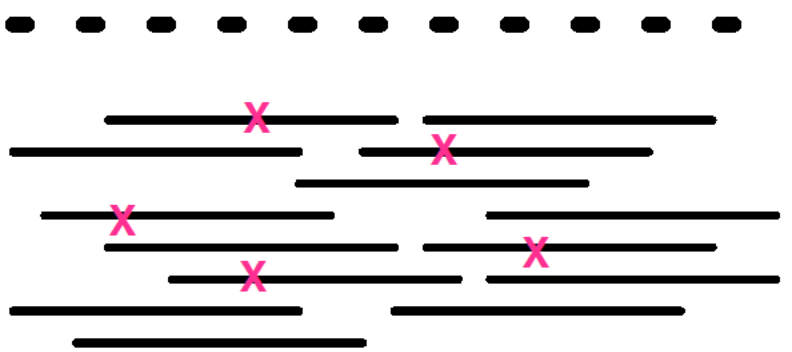
Conteo

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

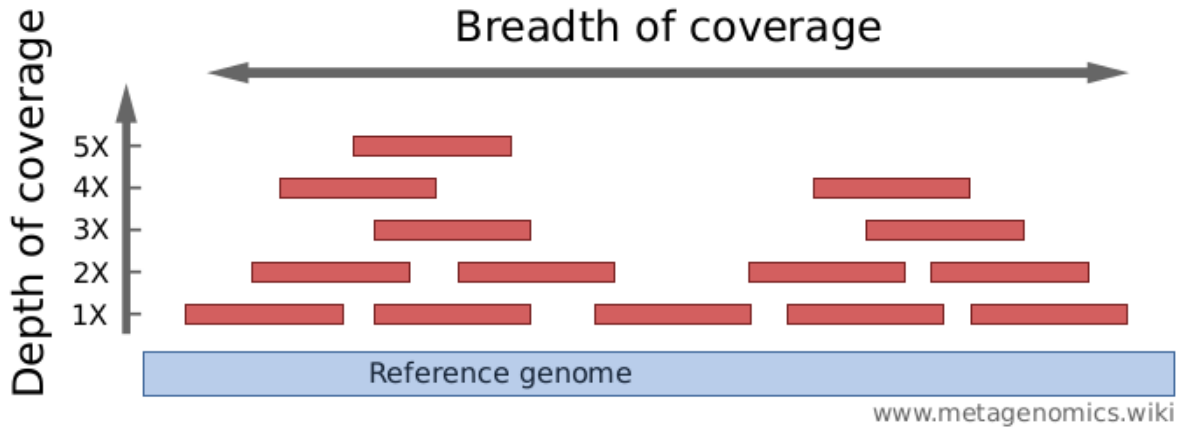
How much of a genome is "covered" by short reads? Are there regions that are not covered, even not by a single read?

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.

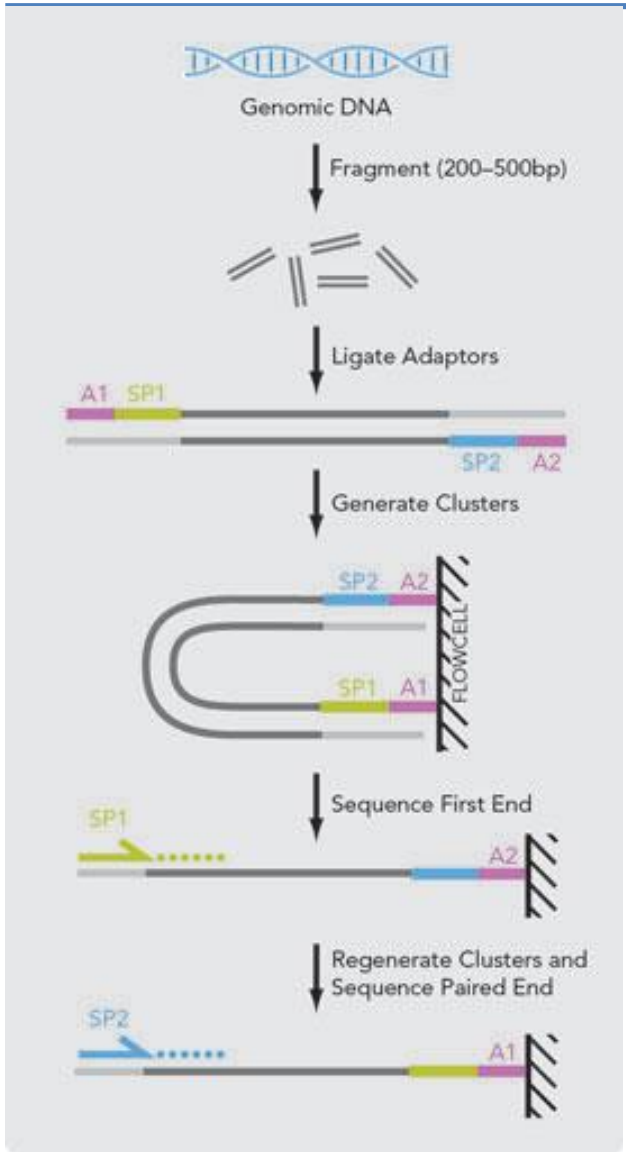
Depth of coverage

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).



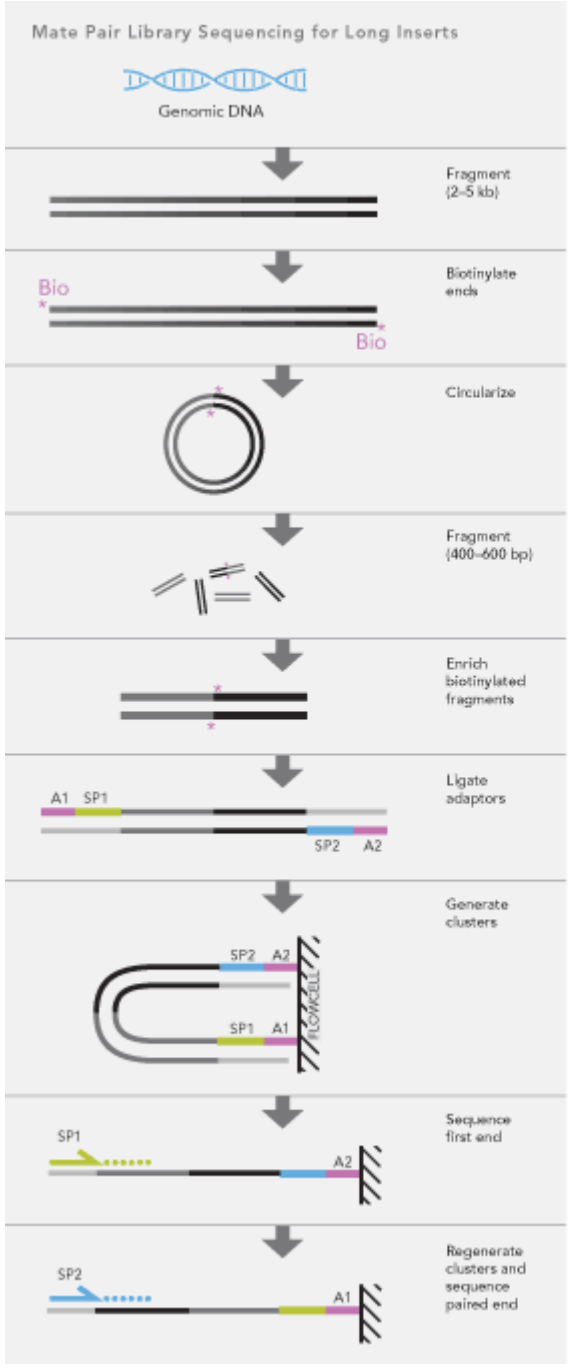
Que es Paired-end?



Secuenciación de un fragmento (bp)

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

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

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

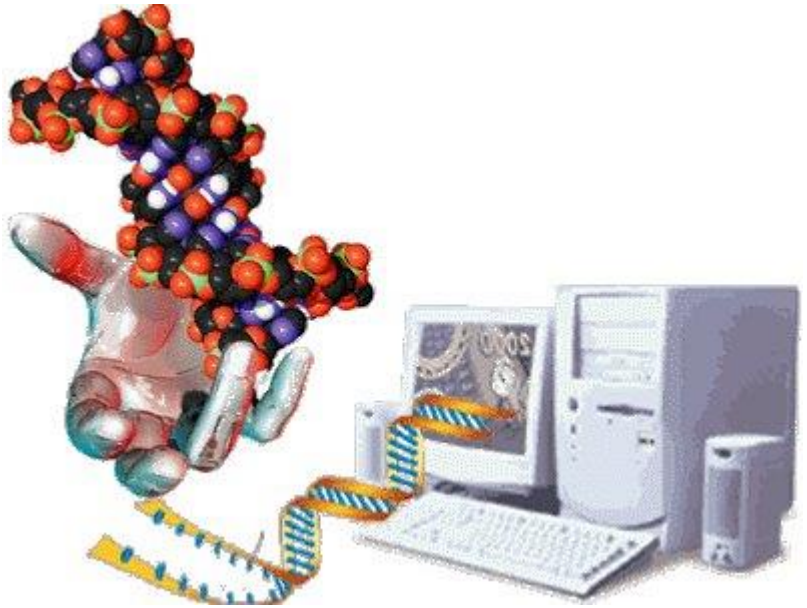
https://genohub.com/recommended-sequencing-coverage-by-application/

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