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

Unidad de Bioinformática (BU-ISCIII)
Unidades Comunes Científico Técnicas – SGAFI-ISCIII

17-21 Junio 2019, 7ª 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 de análisis de los datos procedentes de la SM

Sesión 1 - Secuenciación Masiva Plataformas de Secuenciación

Isabel Cuesta

Unidad de Bioinformática
Unidades Comunes Científico Técnicas – SGAFI-ISCIII

17-21 Junio 2019, 7ª Edición
Programa Formación Continua, ISCIII

INDICE

- Unidad de Bioinformática
  Servicios ofertados

- Evolución de la secuenciación

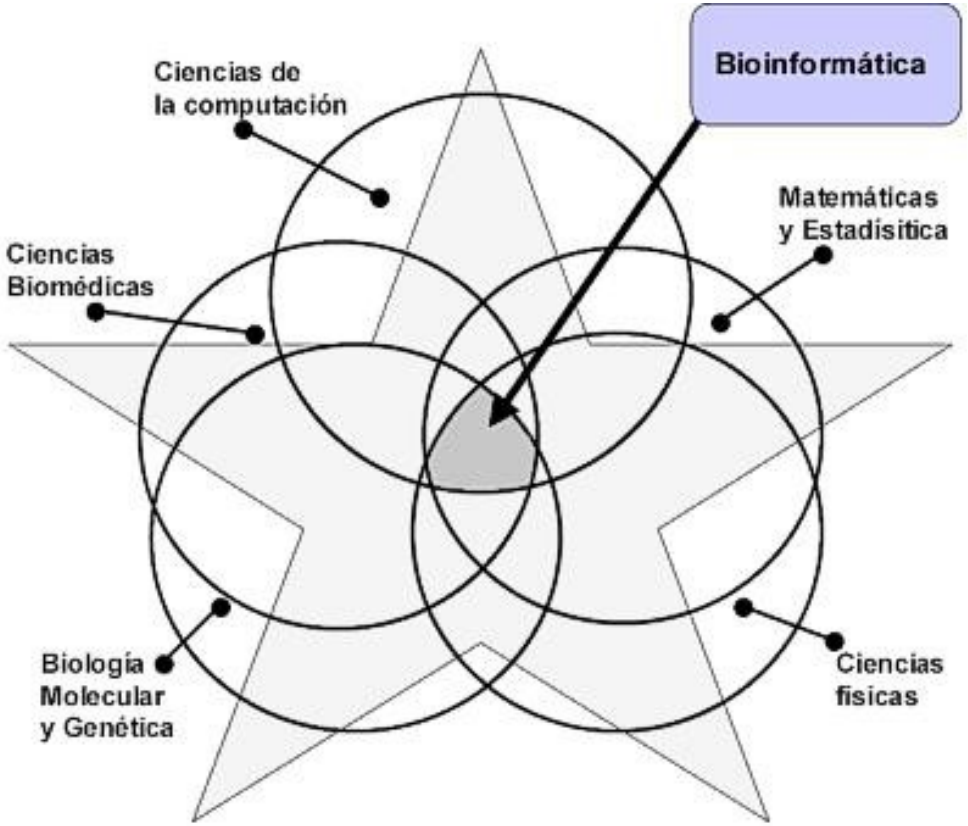
- Plataformas de secuenciación masiva (NGS)

Bioinformatics (i /ˌbaɪoʊˌɪnfərˈmætɪks/) is the application of statistics and computer science to the field of molecular biology.

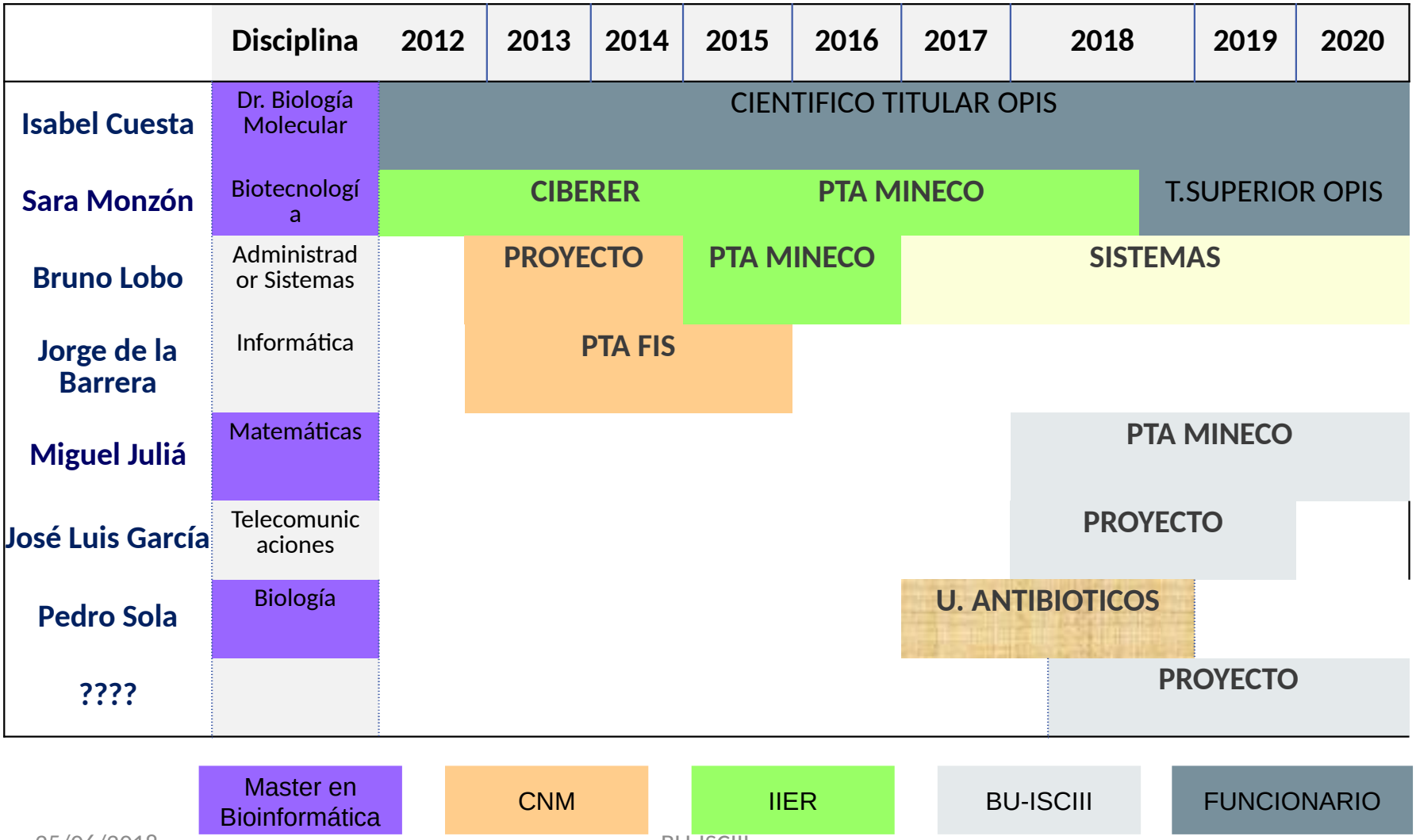
(Quantitative+Computable) Molecular Biology

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques and theory to solve formal and practical problems arising from the management and analysis of biological data.

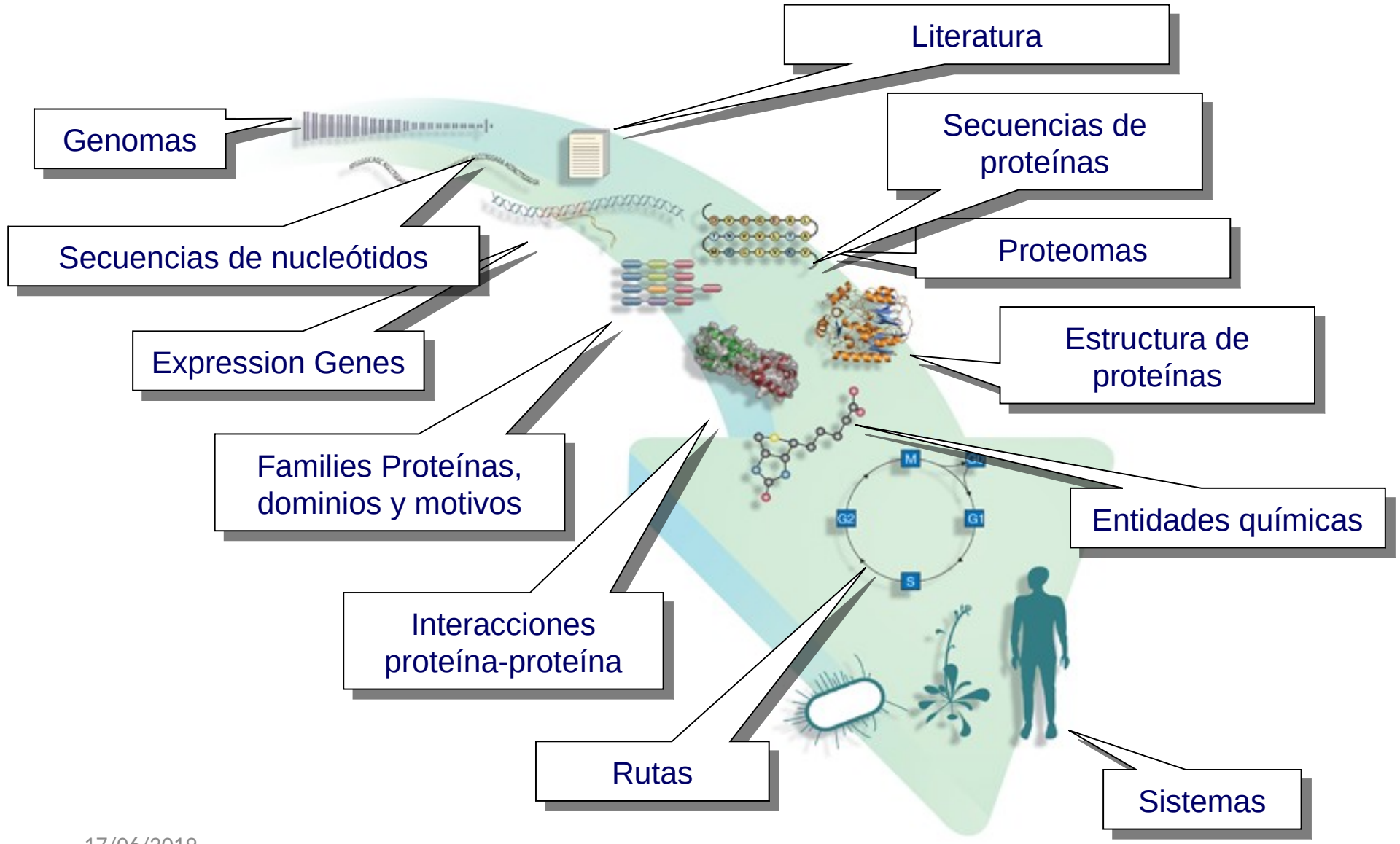
Bioinformática es multidisci



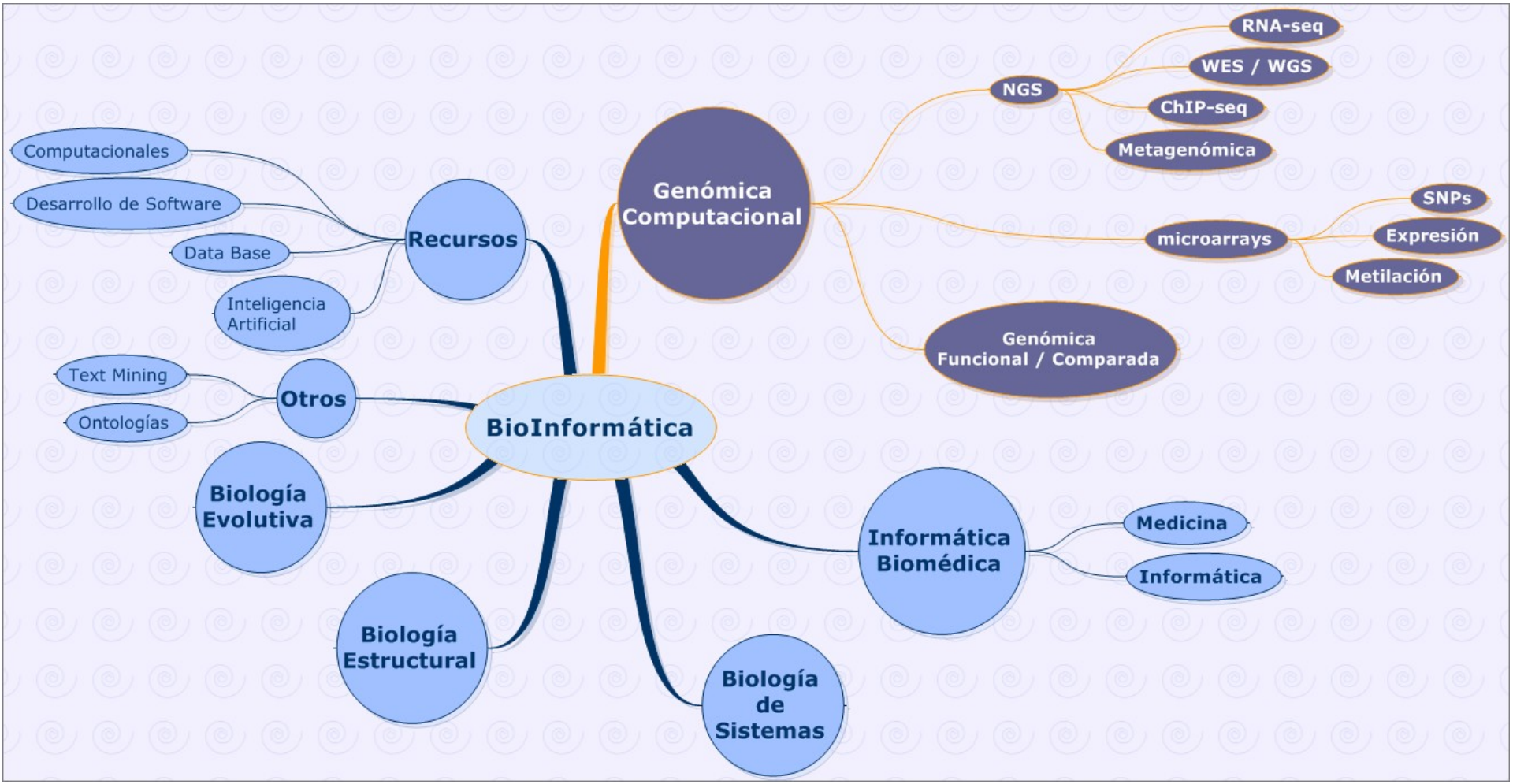
PERSONAL DE LA UNIDAD DE BIOINFORMATICA



os de datos dan idea de la dimensión de la Bioinformática



ESPECIALIZACIÓN



BU-ISCIII

RECURSOS INFORMÁTICOS

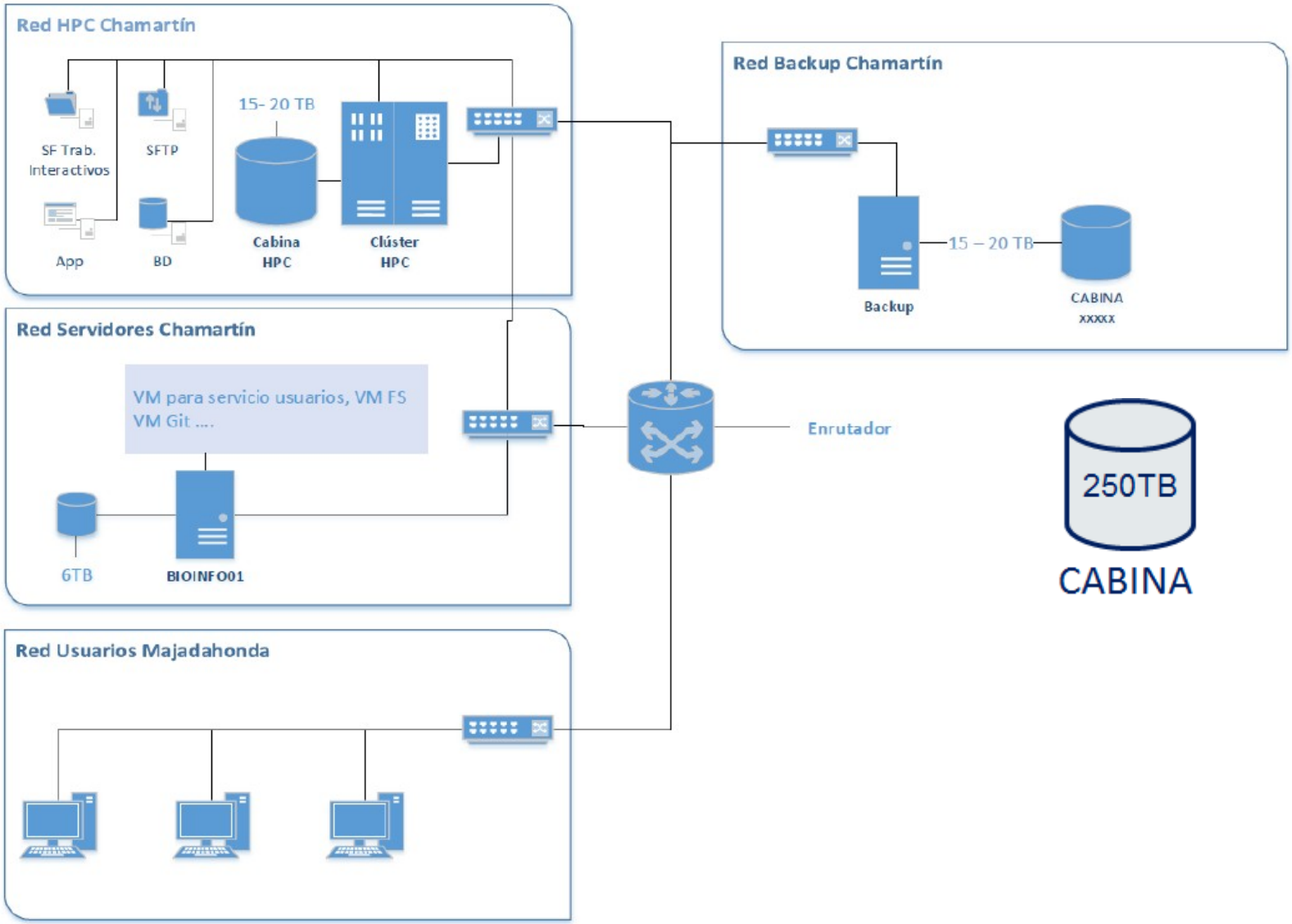
- Marco de relación con UTIC, establecido en 2013, administración compartida.
- Workstation (5), 4nucleos, 32Gb Ram, 4TB almacenamiento
- Servidor de la Unidad, 4-quad, 120Gb Ram,
- HPC 320 cores, 8TB RAM (16 nodos: Por nodo: 2 procesadores de 10 cores (20 threads) a 2,5 GHZ. 256 GB de RAM. Almacenamiento interno de 500 GB. 2 Interfaces de red a 10 Gbps.)
- 2 cabinas de almacenamiento, NetApp, 70 TB y 250TB escalable, ubicadas en el nuevo CPD del ISCIII.

Nuevo CPD

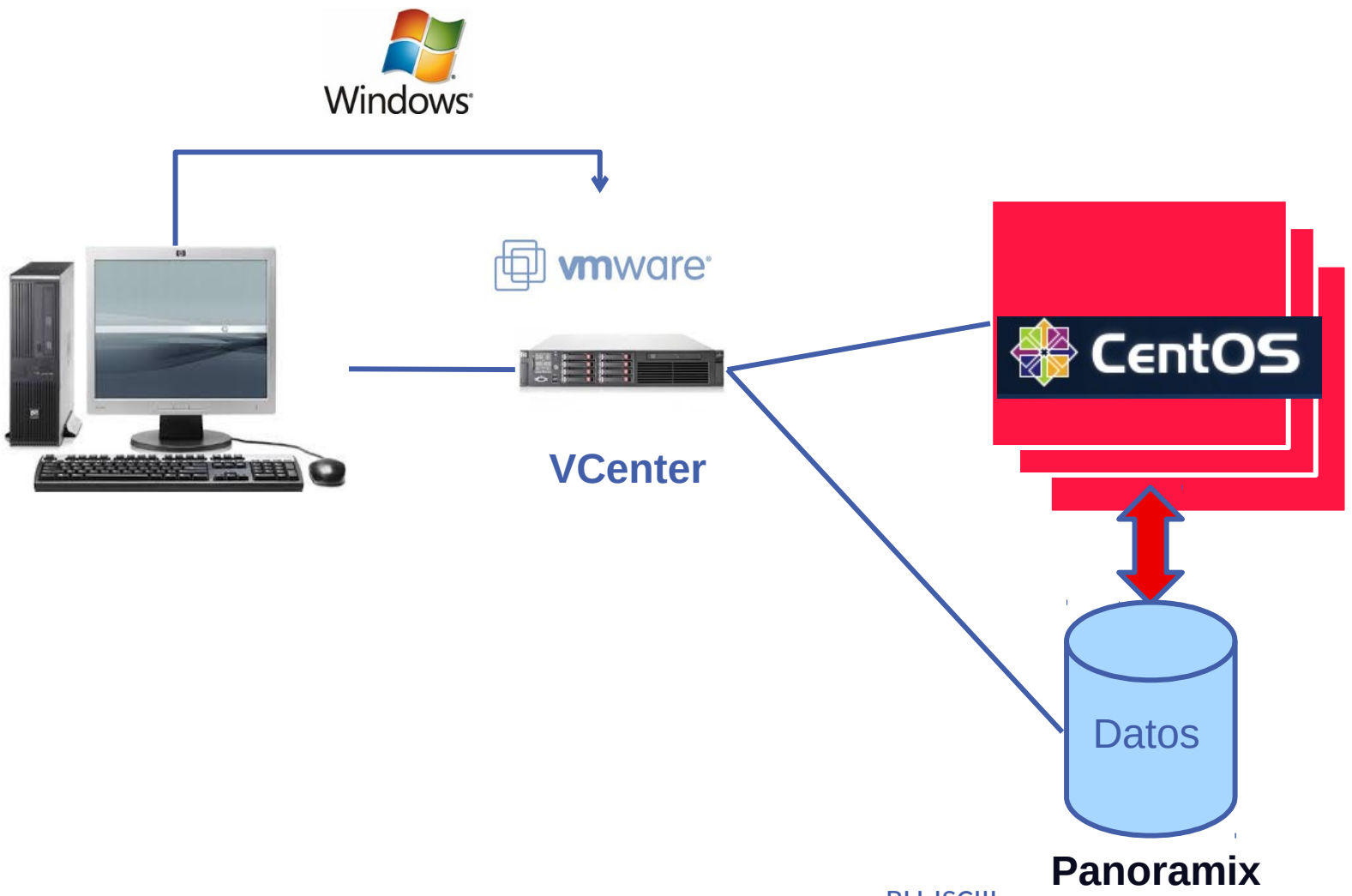
Pabellón 4 – planta semisótano

CPD respaldo Majadahonda

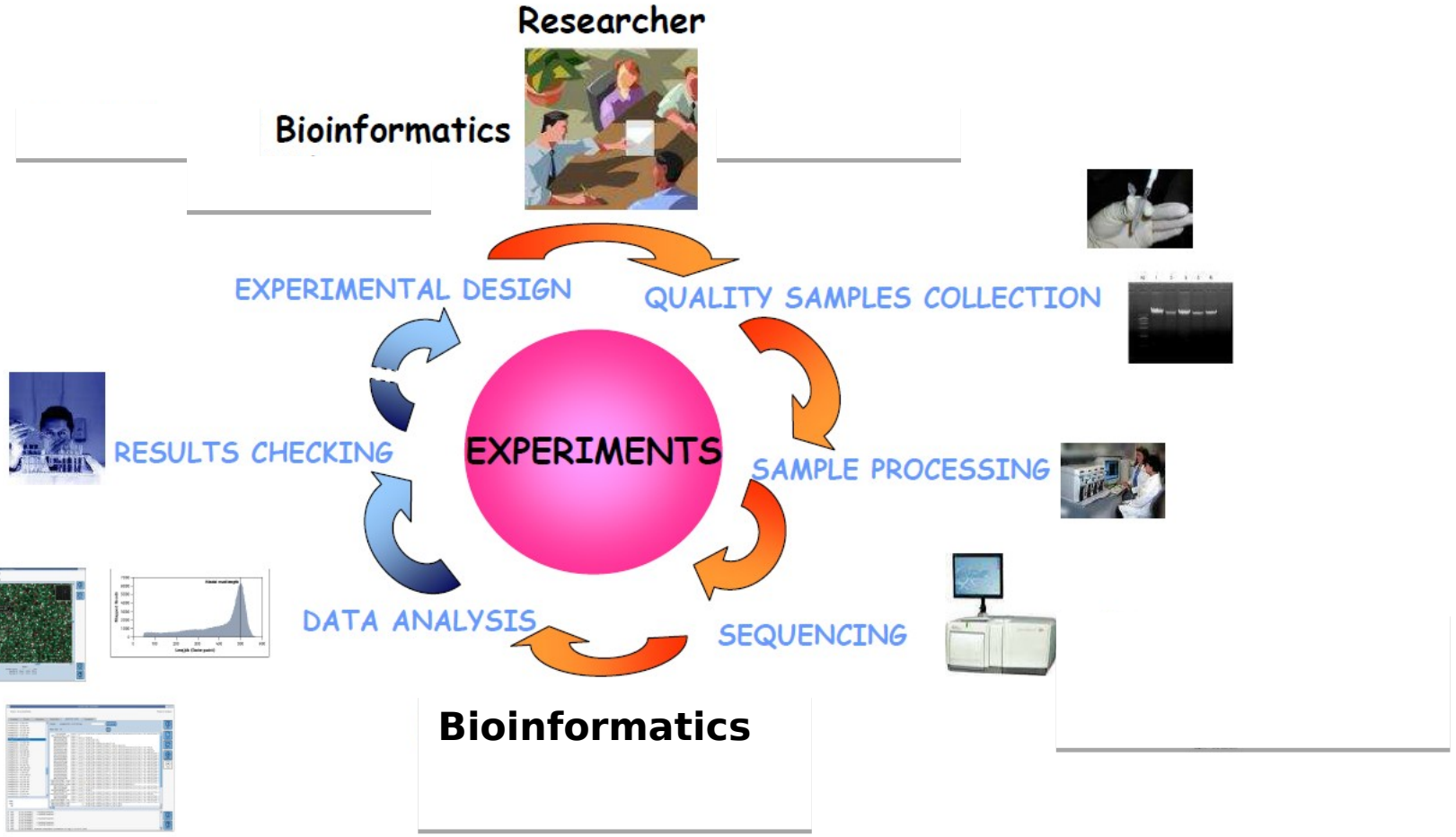
INFRAESTRUCTURA



Recursos Informáticos para el curso



Workflow en NGS



Cartera de Servicios

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

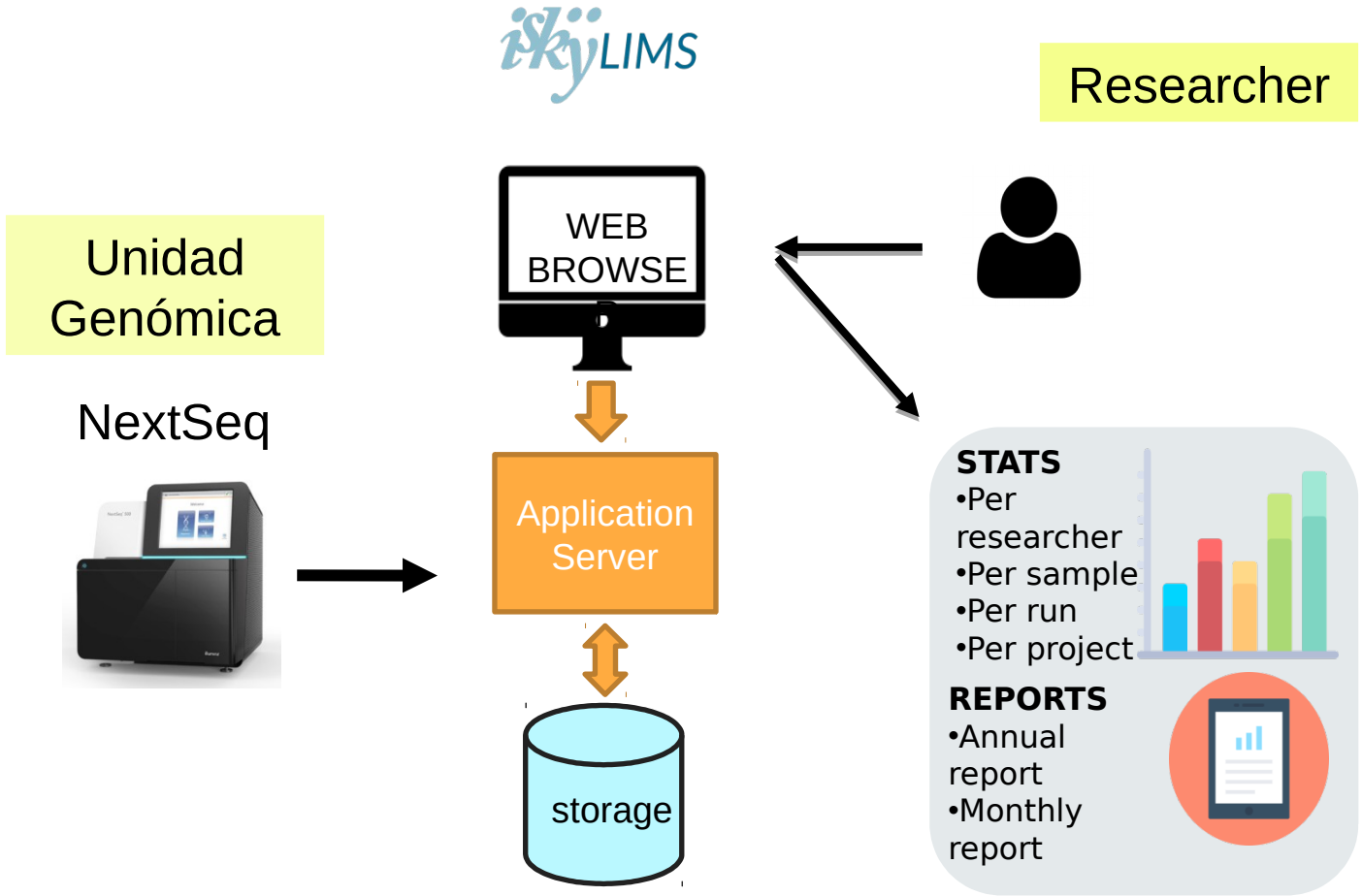
- **ASESORIA Y FORMACIÓN EN BIOINFORMÁTICA**
  Orientación en el análisis bioinformático
  Organización de cursos internos y externos

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

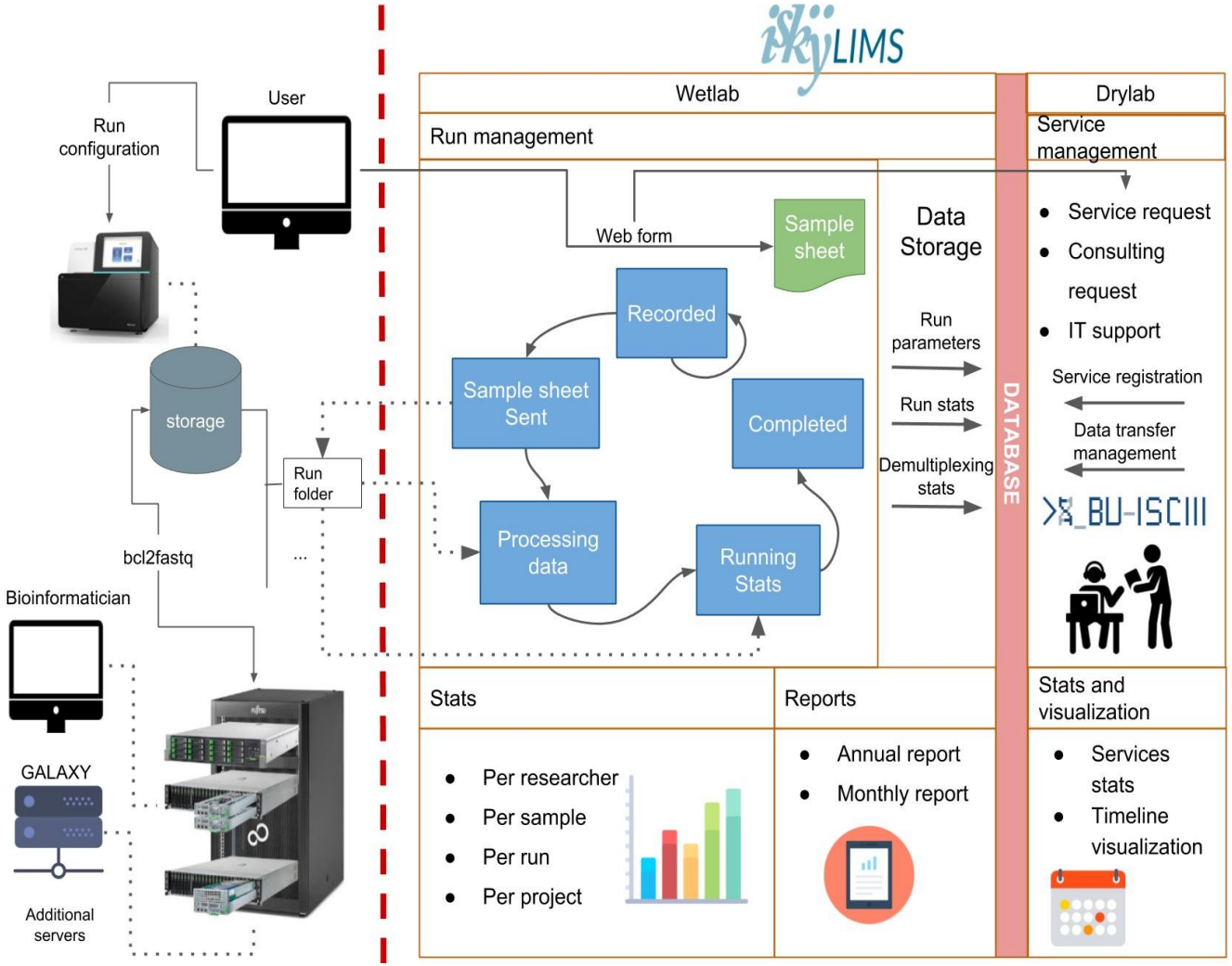
Cartera de Servicios de la Unidad – Análisis de datos Secuenciación Masiva

		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 MeTRS Kaiju
	Shotgun			Genome Ref Seq		Pathogen / Genome coverage	PikaVirus

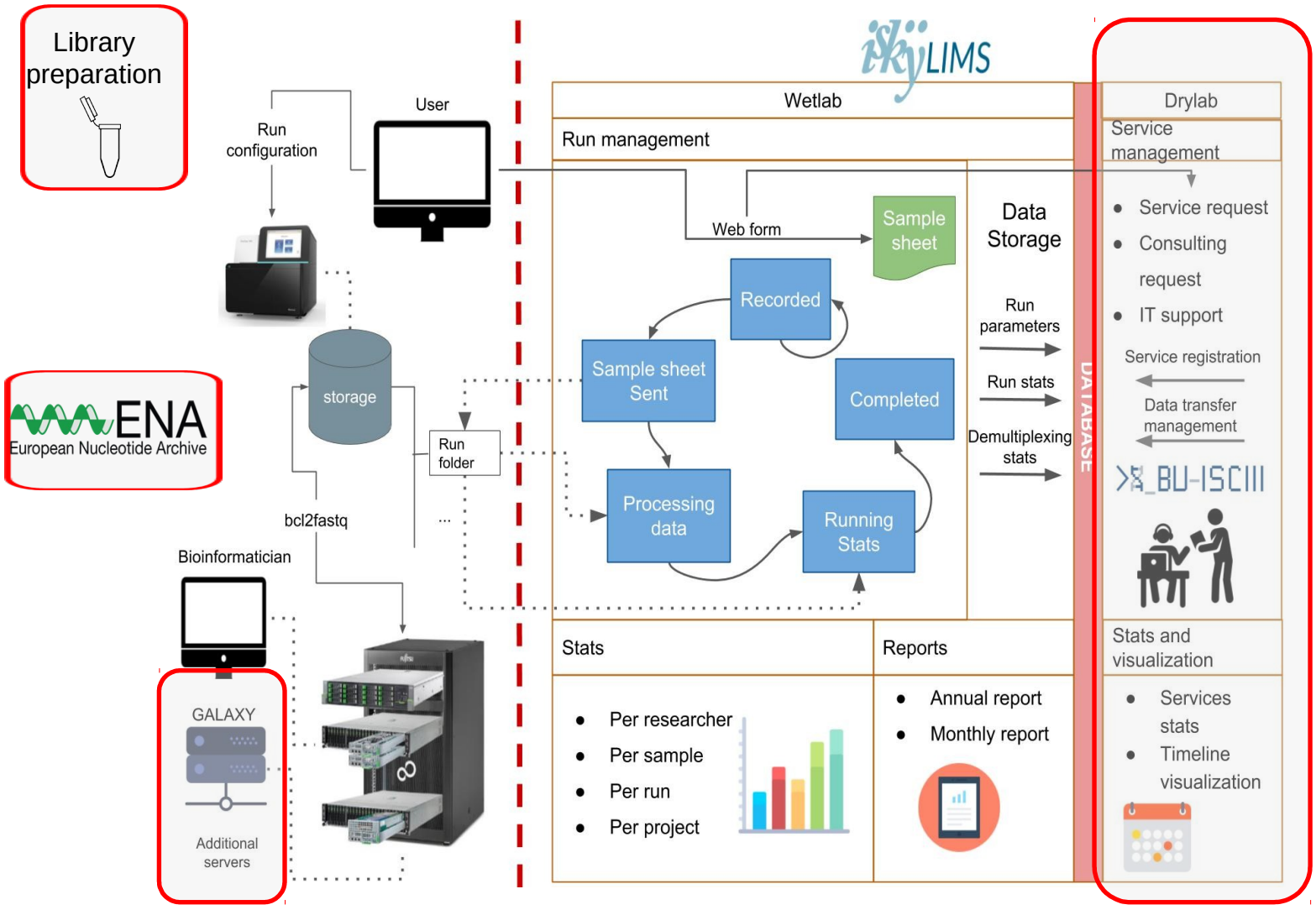
iSkyLIMS


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

iSkyLIMS


iSkyLIMS
Wetlab
Drylab
Run management
Service management
User
Run configuration
Web form
Sample sheet
Data Storage
Service request
Consulting request
IT support
Recorded
Sample sheet Sent
Completed
Processing data
Running Stats
Run parameters
Run stats
Demultiplexing stats
DATABASE
Service registration
Data transfer management
storage
Run folder
bcl2fastq
...
BU-ISCIII
Bioinformatician
GALAXY
Additional servers
Stats
Reports
Stats and visualization
Per researcher
Per sample
Per run
Per project
Annual report
Monthly report
Services stats
Timeline visualization

iSkyLIMS


Library preparation
Run configuration
User
storage
Run folder
bcl2fastq
...
Bioinformatician
GALAXY
Additional servers
ENA
European Nucleotide Archive
iSkyLIMS
Wetlab
Run management
Web form
Sample sheet
Recorded
Sample sheet Sent
Completed
Processing data
Running Stats
Data Storage
Run parameters
Run stats
Demultiplexing stats
DATABASE
Stats
Per researcher
Per sample
Per run
Per project
Reports
Annual report
Monthly report
Drylab
Service management
Service request
Consulting request
IT support
Service registration
Data transfer management
>_BU-ISCIII
Stats and visualization
Services stats
Timeline visualization

SERVICIOS DE LA UNIDAD

smonzon | Logout | My account

BioInformatics

iSkyLIMS: DryLab

Welcome

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

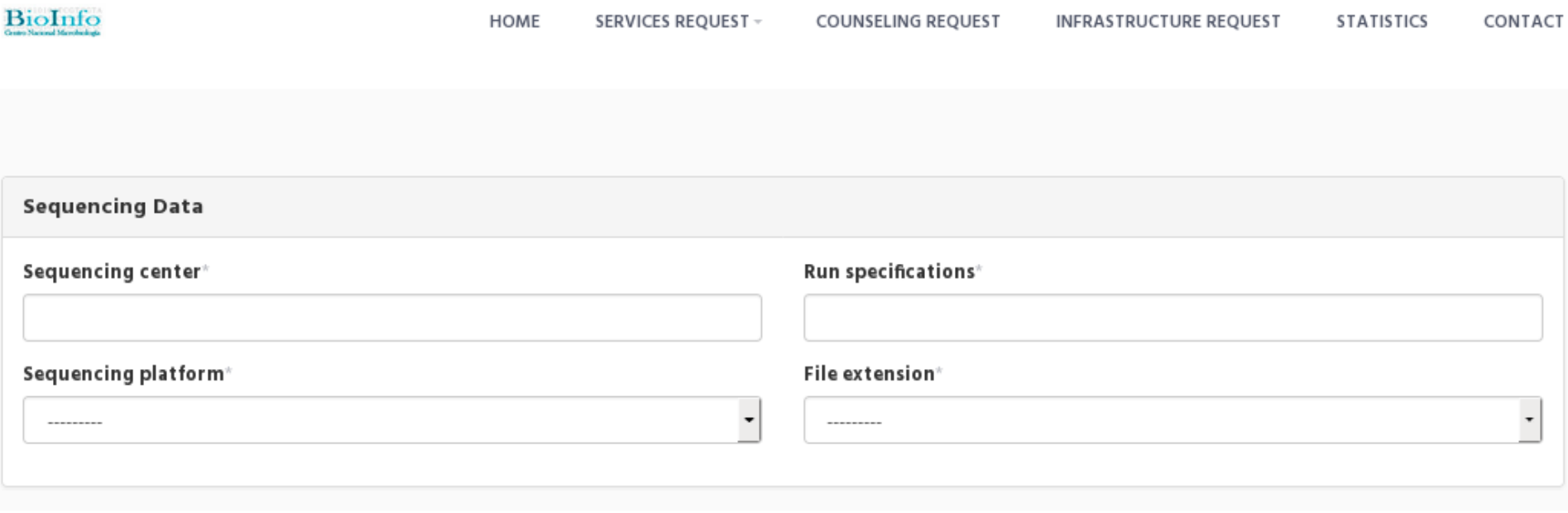
Services ongoing and queued

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

Timeline of services

Under construction. Kind of diagram with services dates.

SERVICES REQUEST



SERVICES REQUEST

Service selection

AvailableServices *

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

SERVICES REQUEST

Service Description

Service description file*

Browse... No file selected.

Service Notes*

Save

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

Service selection

AvailableServices *

- [ ] **User support**
  - [ ] **Installation and support of bioinformatic software on Linux OS**
  - [ ] **Installation and access to Virtual machines in the Unit server containing bioinformatic software**
  - [ ] **Code snippets development**

Service Description

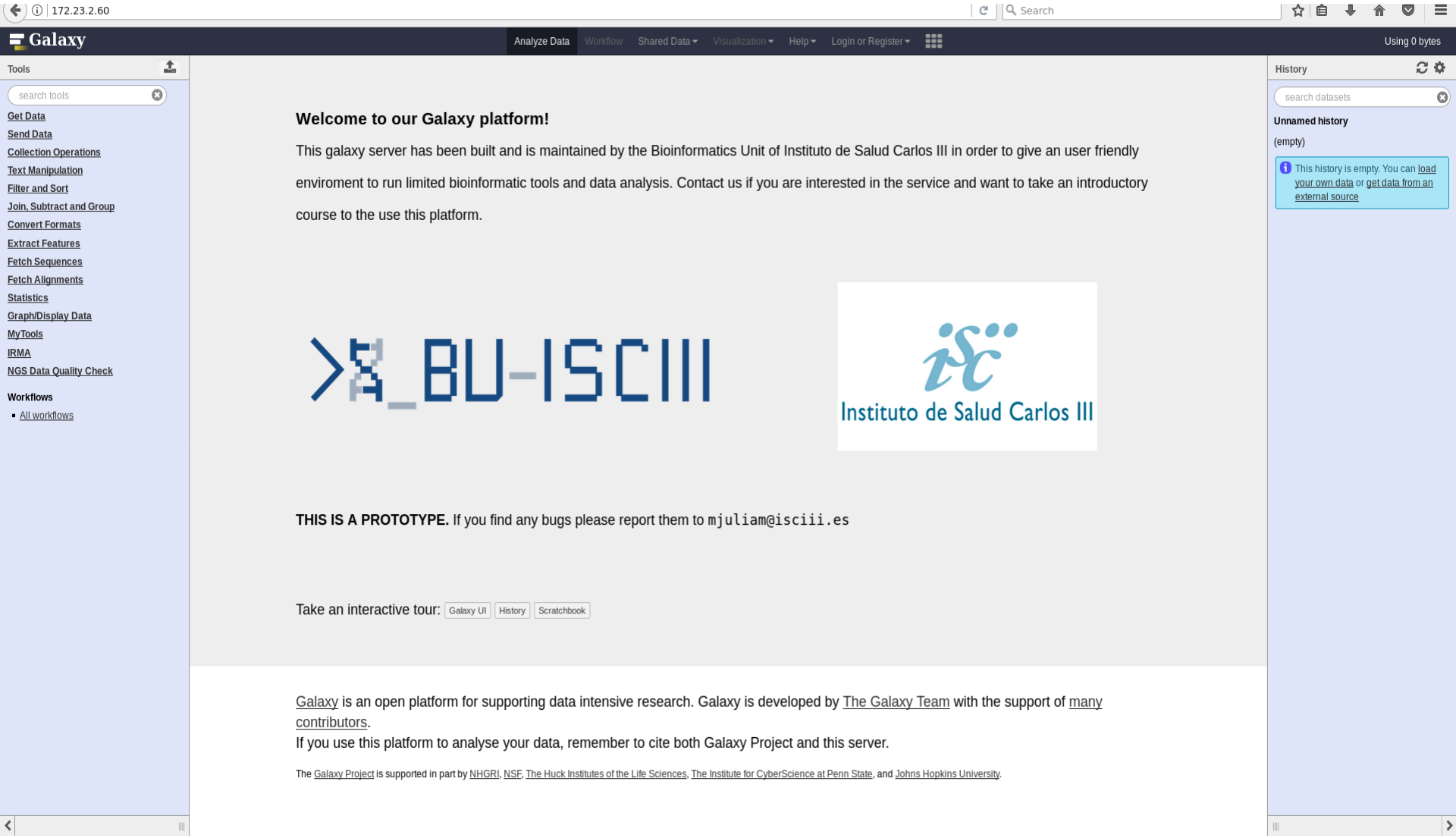
Service description file*

Browse... No file selected.

Service Notes*

Save

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

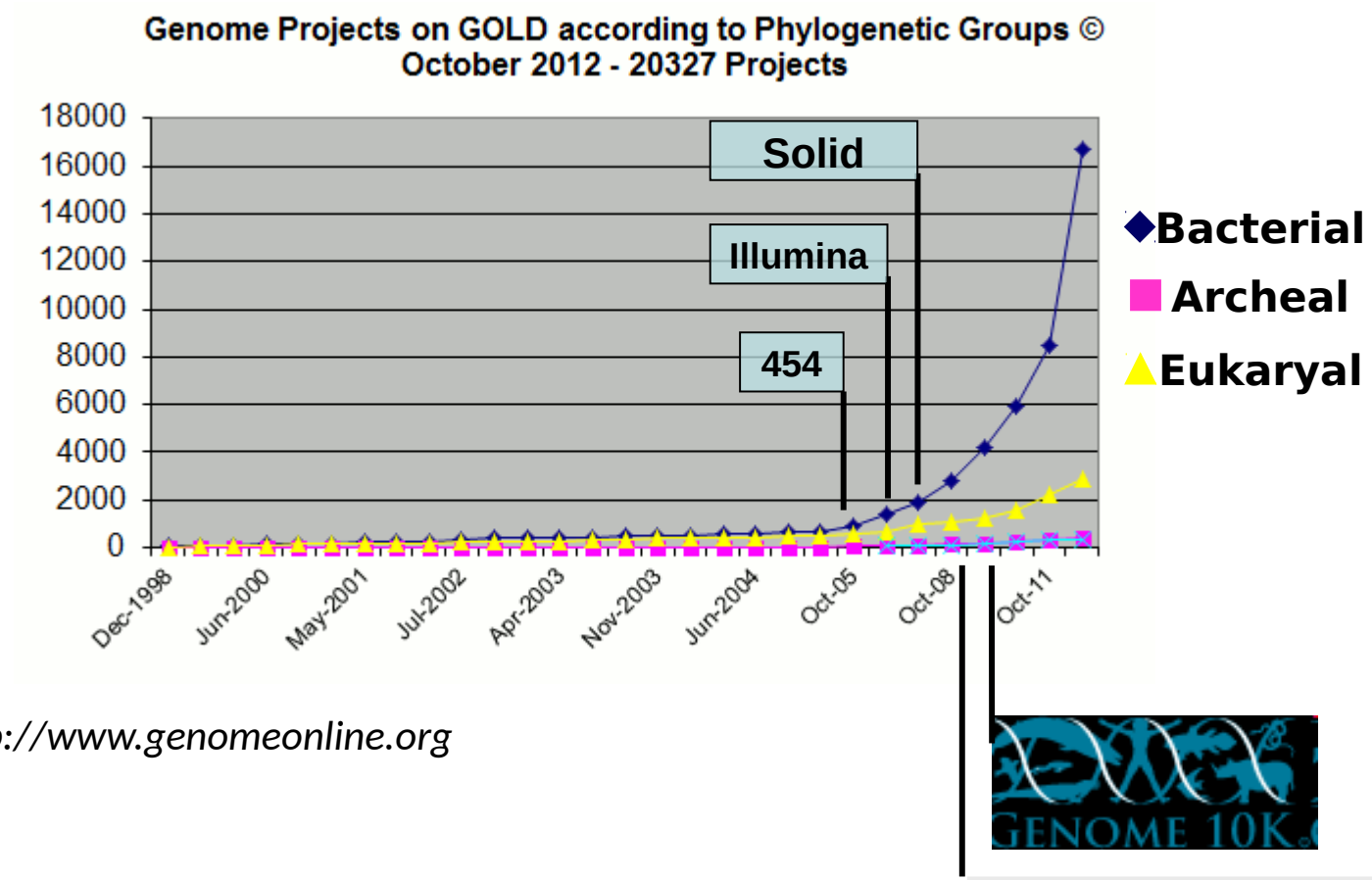
INDICE

- Unidad de Bioinformática
  Servicios ofertados

- Evolución de la secuenciación

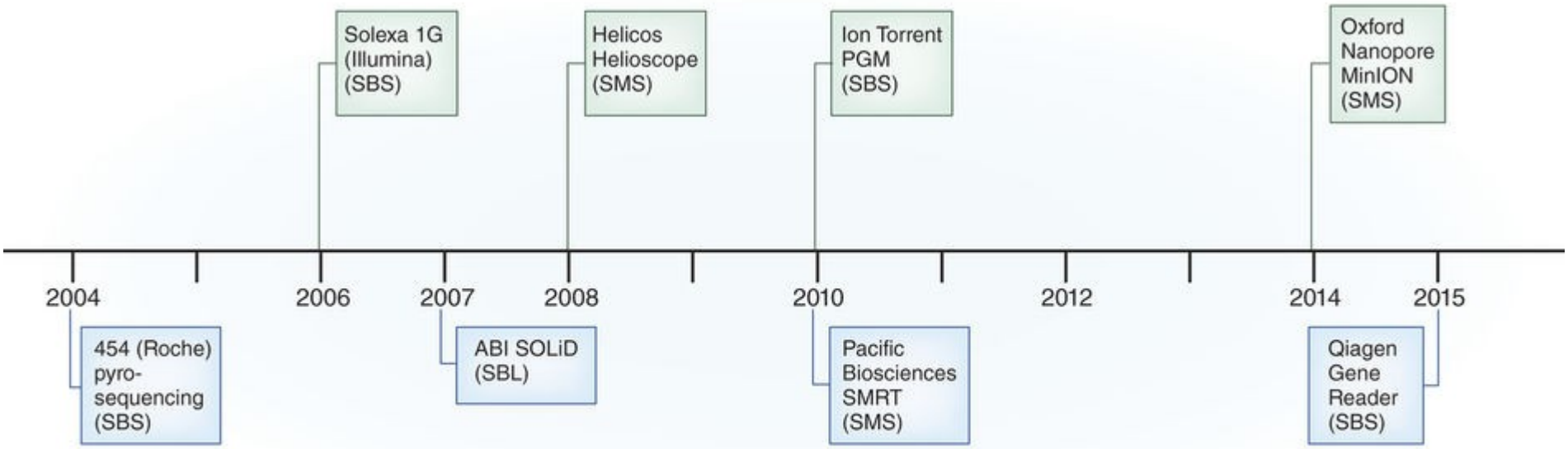
- Plataformas de secuenciación masiva (NGS)

Genomics Revolution Era

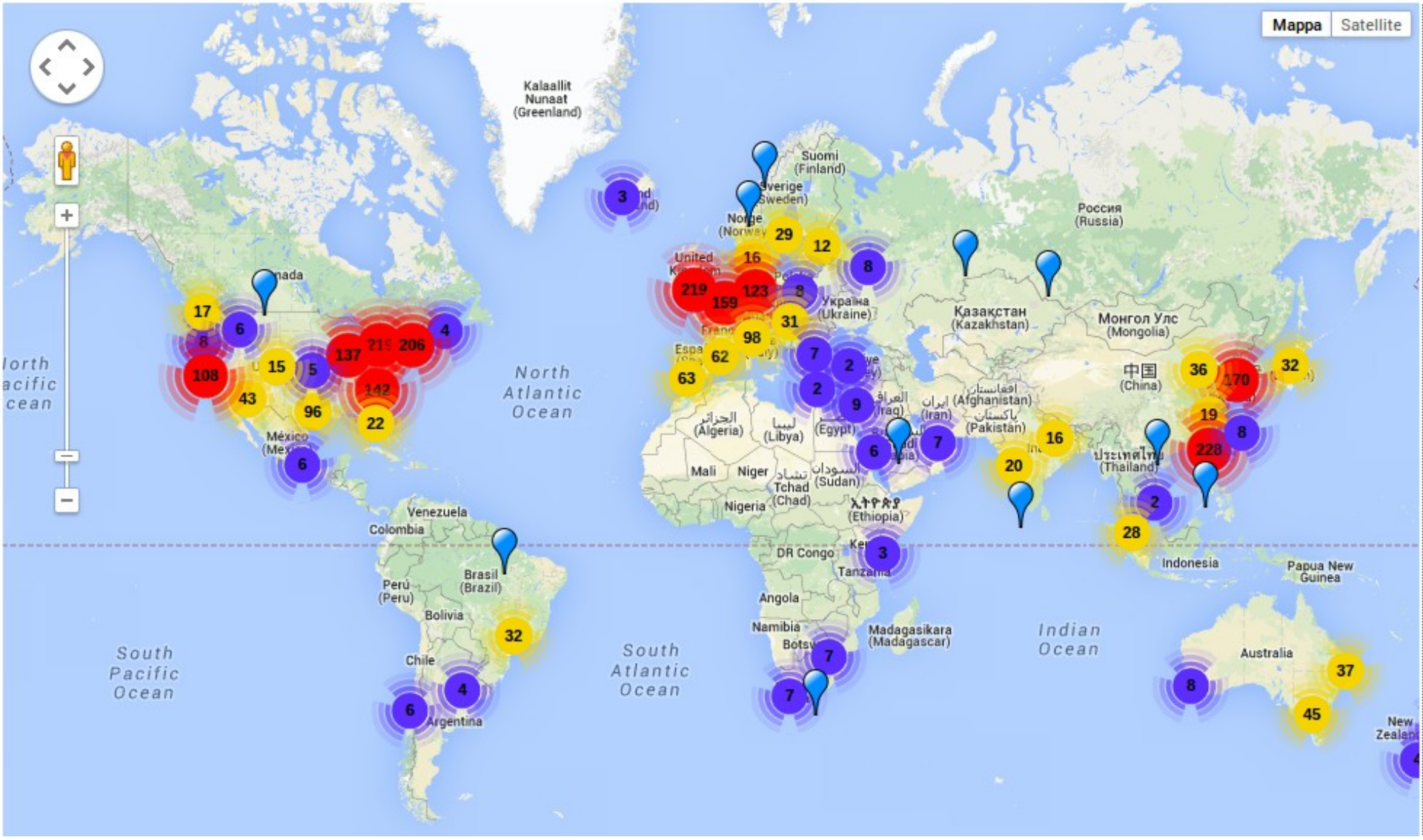


Source: http://www.genomeonline.org

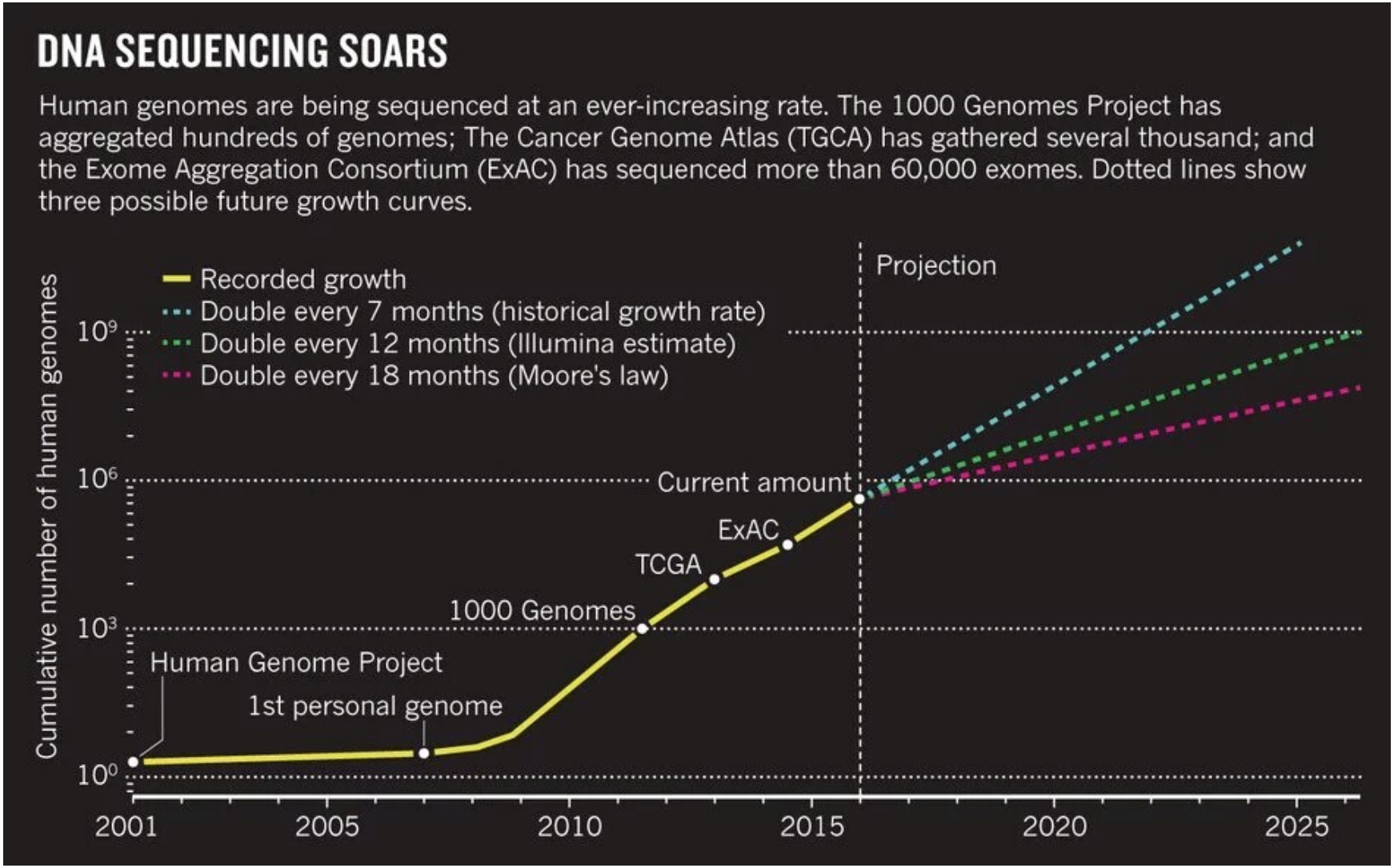
NGS PLATFORMS: TIMELINE



Democratización de la secuenciación. Mapa de lo



SEQUENCING PROJECTS



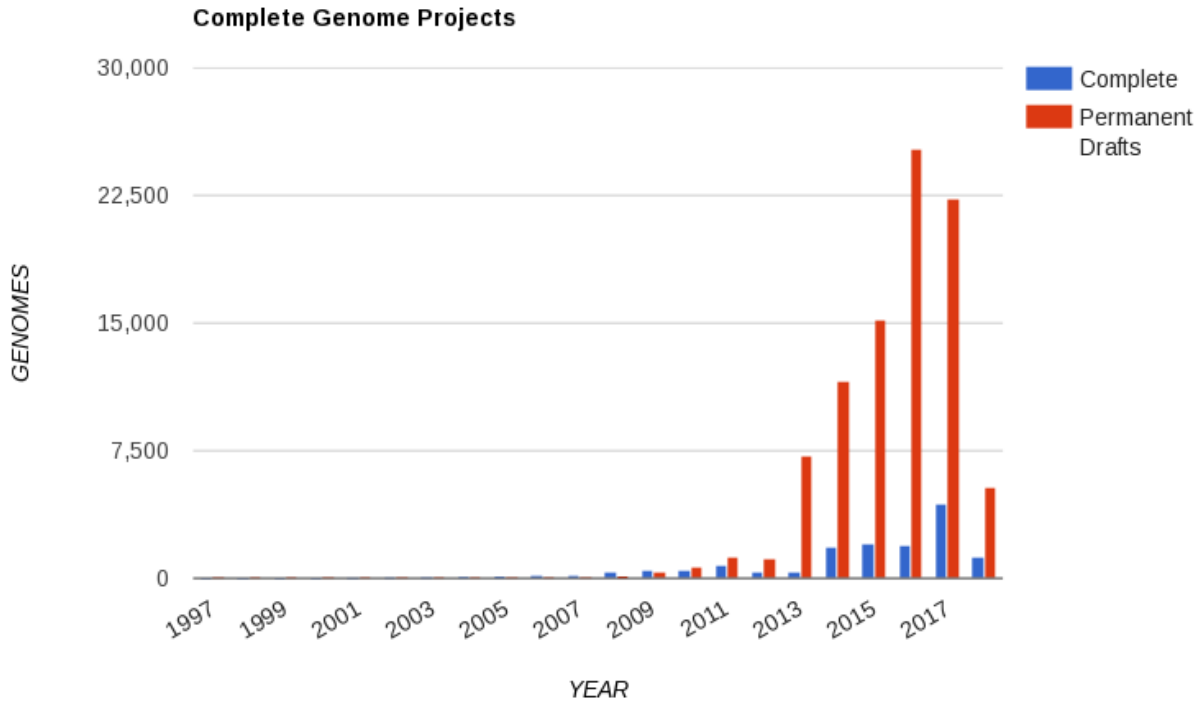
Genomics Revolution Era

JGI GOLD GENOMES ONLINE DATABASE

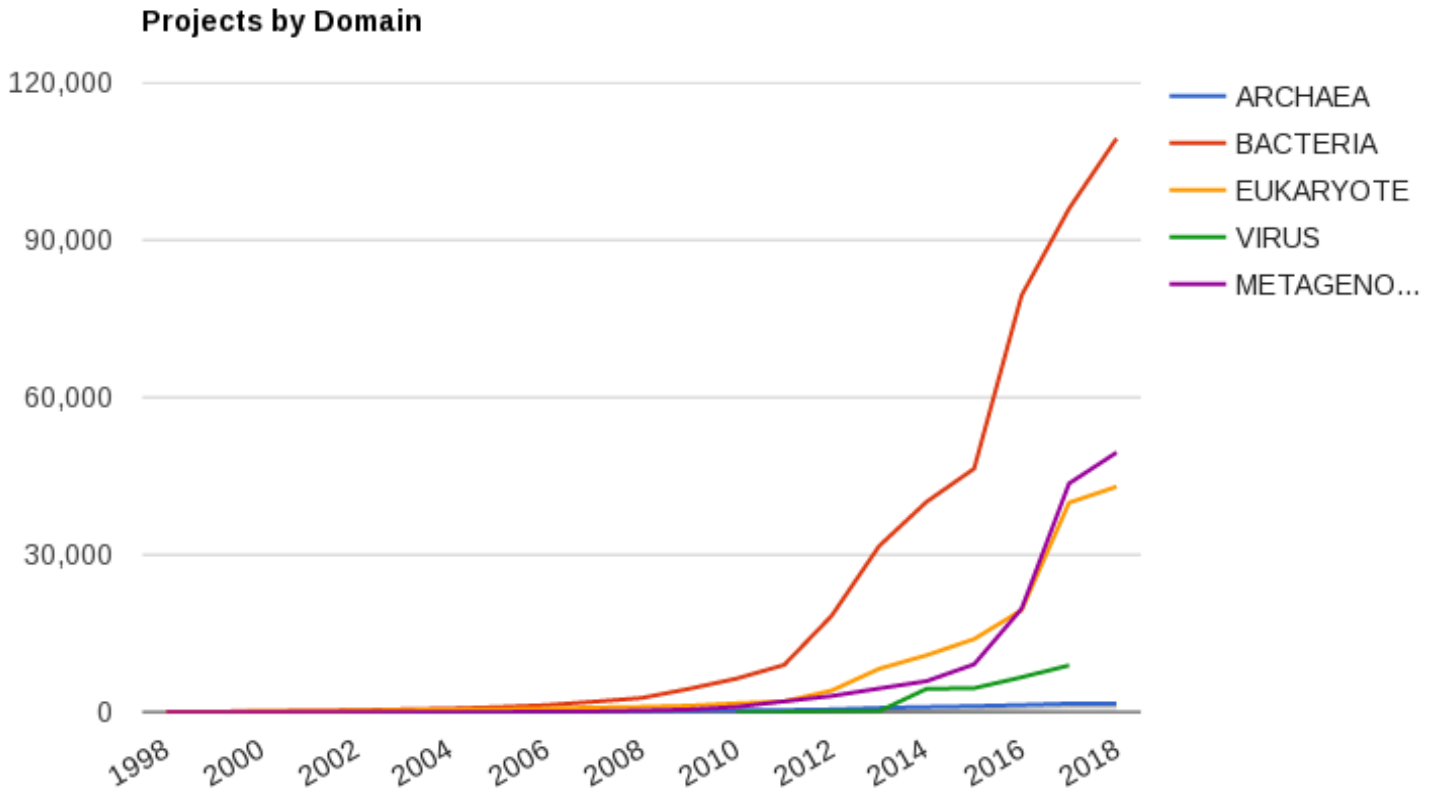
JGI HOME LOG IN

Home Search Distribution Graphs Biogeographical Metadata Statistics References Team Help News

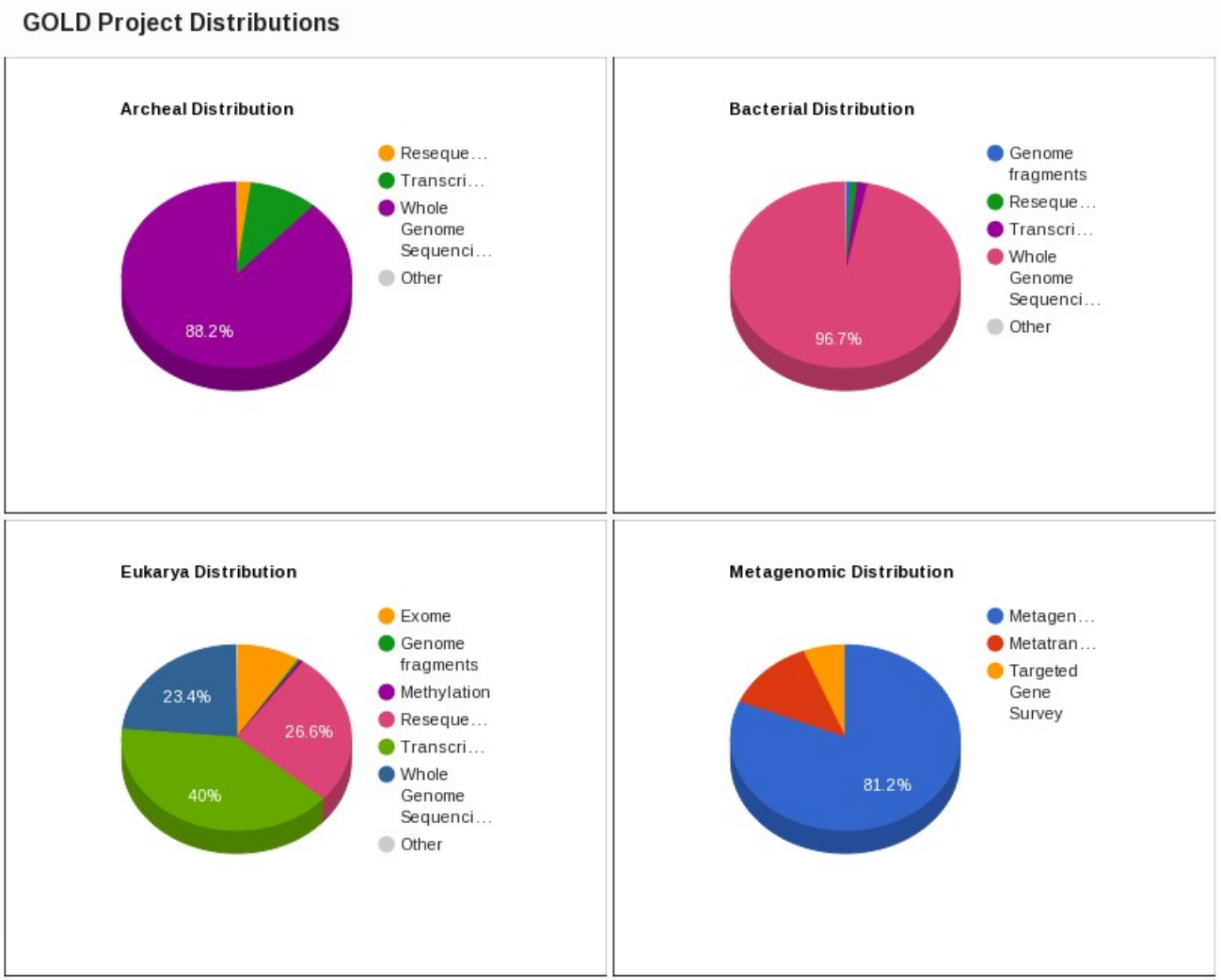
Source: https://gold.jgi.doe.gov/statistics



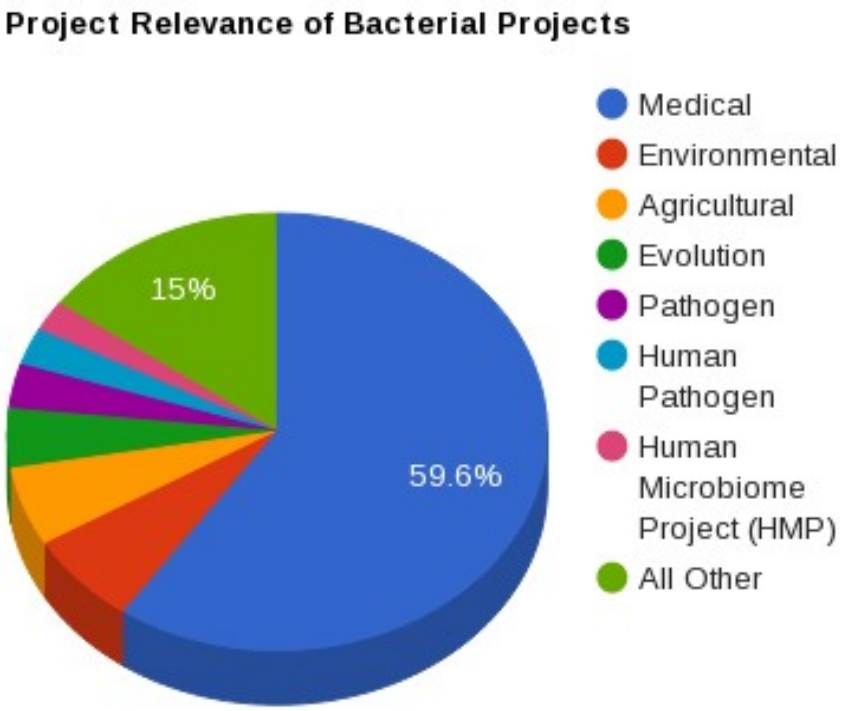
Genomics Revolution Era



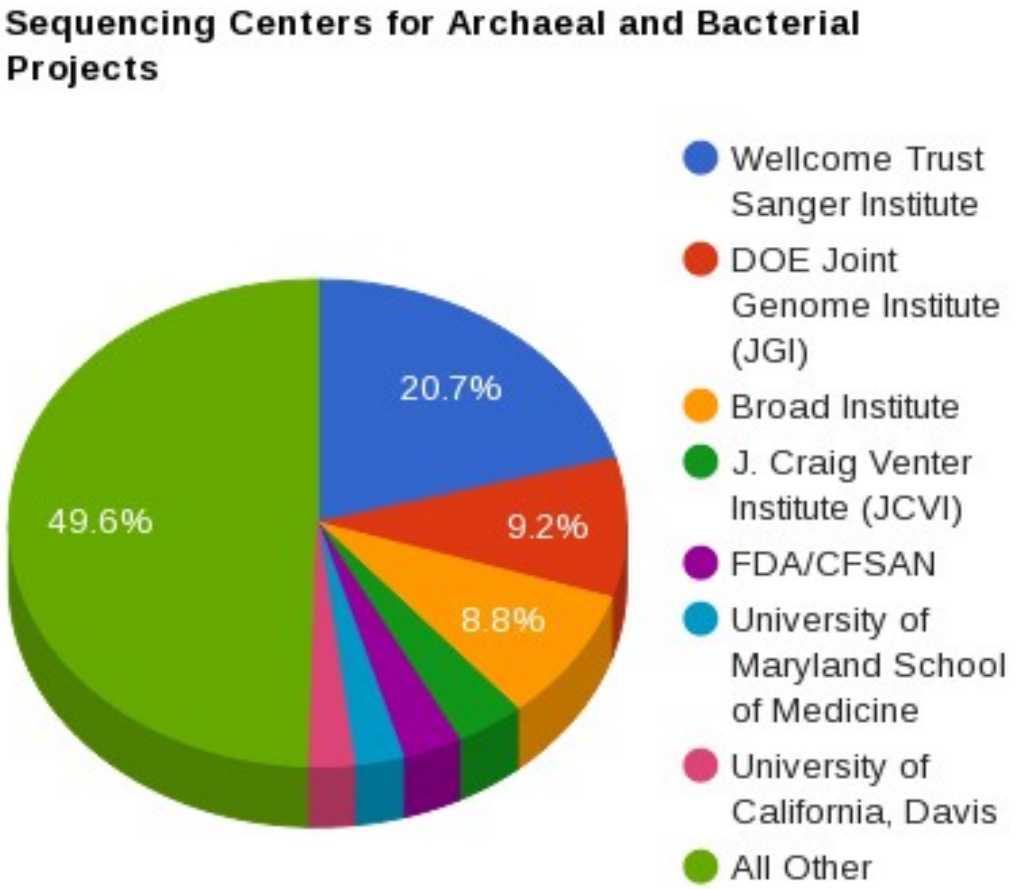
Genomics Revolution Era



Genomics Revolution Era



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)

High-Throughput Sequencing Platforms

Roche 454 SEQUENCING

GS-FLX System

illumina

Genome Analyzer IIx

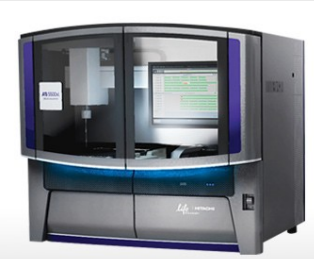
Hiseq 2000/1000

Hiseq 2500/1500

NextSeq

life technologies™

SOLID 3 Plus/4



Genetic Analyzers 5500W Series

Benchtop High-Throughput Sequencing Platform

Roche 454 SEQUENCING

Roche 454 GS Junior

illumina

MiSeq

iSeq 100

MiniSeq

life technologies™

Ion Proton™ System

Ion PGM™ System

Illumina Benchtop Sequencers

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




Popular Applications & Methods	iSeq 100 System Key Application	MiniSeq System Key Application	MiSeq Series ⊕ Key Application	NextSeq Series ⊕ Key Application
Large Whole-Genome Sequencing (human, plant, animal)				●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●	●
Exome Sequencing				●
Targeted Gene Sequencing (amplicon, gene panel)	●	●	●	●
Whole-Transcriptome Sequencing				●
Gene Expression Profiling with mRNA-Seq				●
Targeted Gene Expression Profiling	●	●	●	
Long-Range Amplicon Sequencing*	●	●	●	
miRNA & Small RNA Analysis	●	●	●	●
DNA-Protein Interaction Analysis			●	●
Methylation Sequencing				●
16S Metagenomic Sequencing		●	●	●

Run Time	9–17.5 hours	4–24 hours	4–55 hours	12–30 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb
Maximum Reads Per Run	4 million	25 million	25 million †	400 million
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp

Illumina Production-Scale Sequencers

Benchtop Sequencers	Production-Scale Sequencers		
			
NextSeq Series ✪	HiSeq Series ✪	HiSeq X Series†	NovaSeq 6000 System ✪
Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)	●	●	●	●
Small Whole-Genome Sequencing (microbe, virus)	●	●		●
Exome Sequencing	●	●		●
Targeted Gene Sequencing (amplicon, gene panel)	●	●		●
Whole-Transcriptome Sequencing	●	●		●
Gene Expression Profiling with mRNA-Seq	●	●		●
miRNA & Small RNA Analysis	●	●		●
DNA-Protein Interaction Analysis	●	●		●
Methylation Sequencing	●	●		●
Shotgun Metagenomics	●	●		●
Run Time	12–30 hours	< 1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	< 3 days	16–36 hours (Dual S2 flow cells) 44 hours (Dual S2 flow cells)
Maximum Output	120 Gb	1500 Gb	1800 Gb	6000 Gb§
Maximum Reads Per Run	400 million	5 billion	6 billion	20 billion‖
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp

High-Throughput Single Molecule Sequencing Pl

3ª GENERACIÓN

Helicos BioSciences Corporation

Helicos Genetic Analysis System

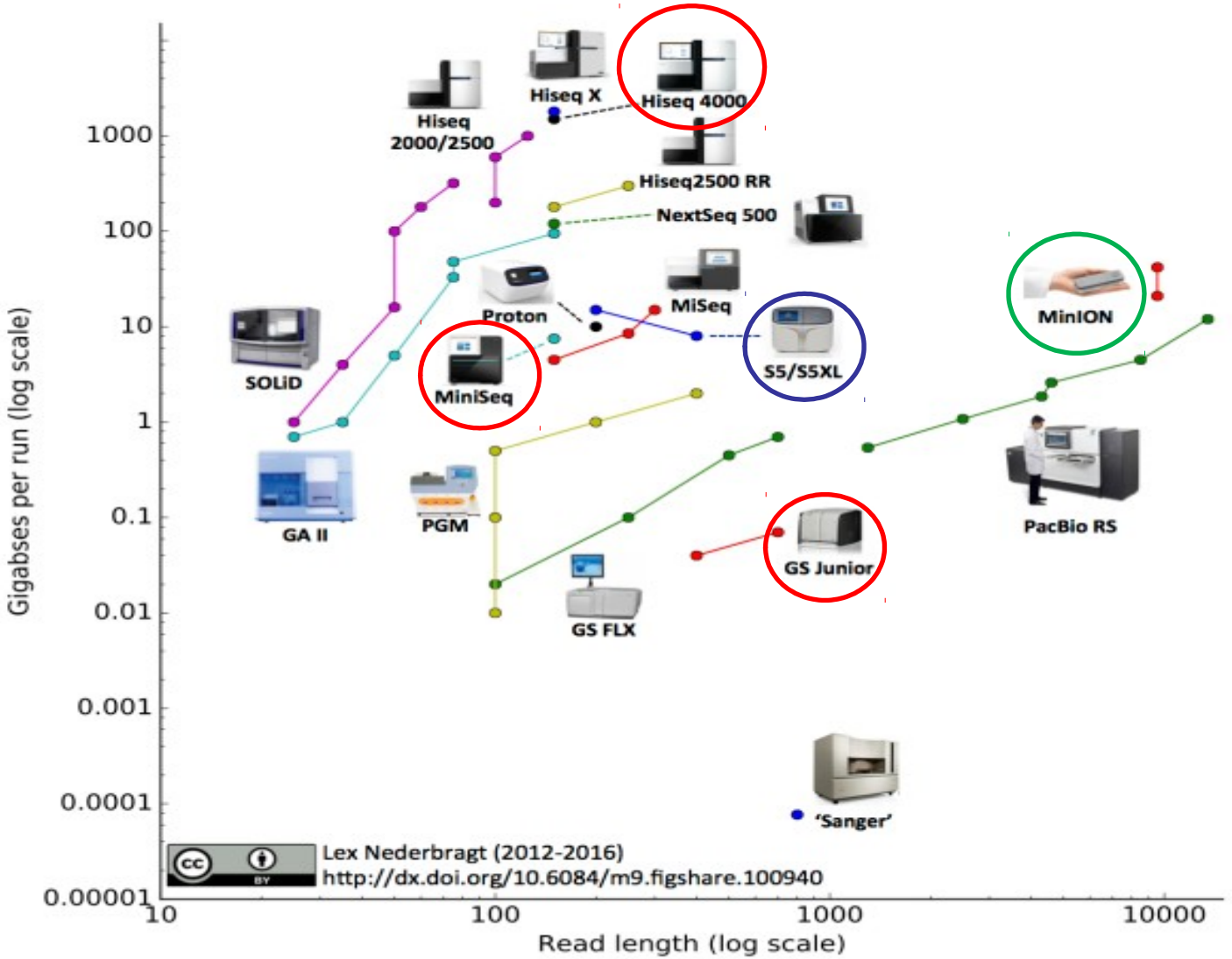
PACIFIC BIOSCIENCES®

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

Oxford Nanopore

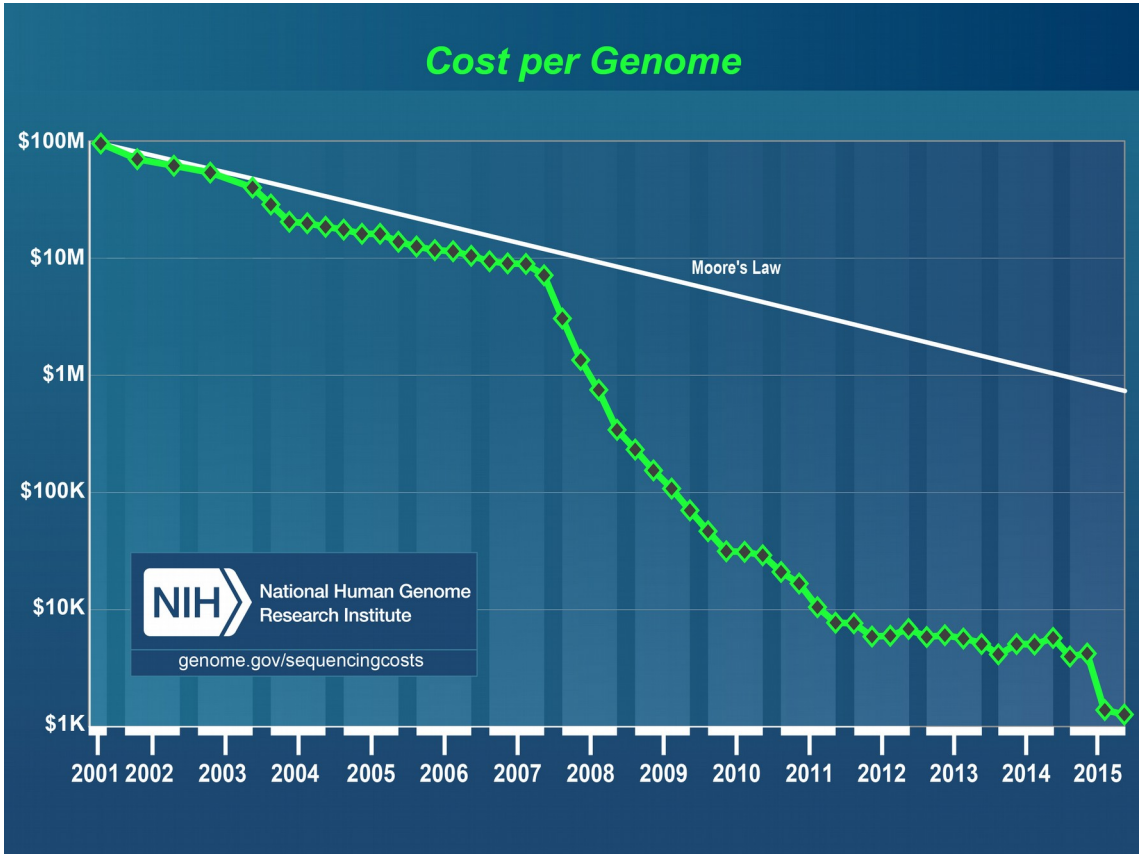
MinIon™ Device

PLATAFORMAS DE SECUENCIACIÓN, 2016 Editio

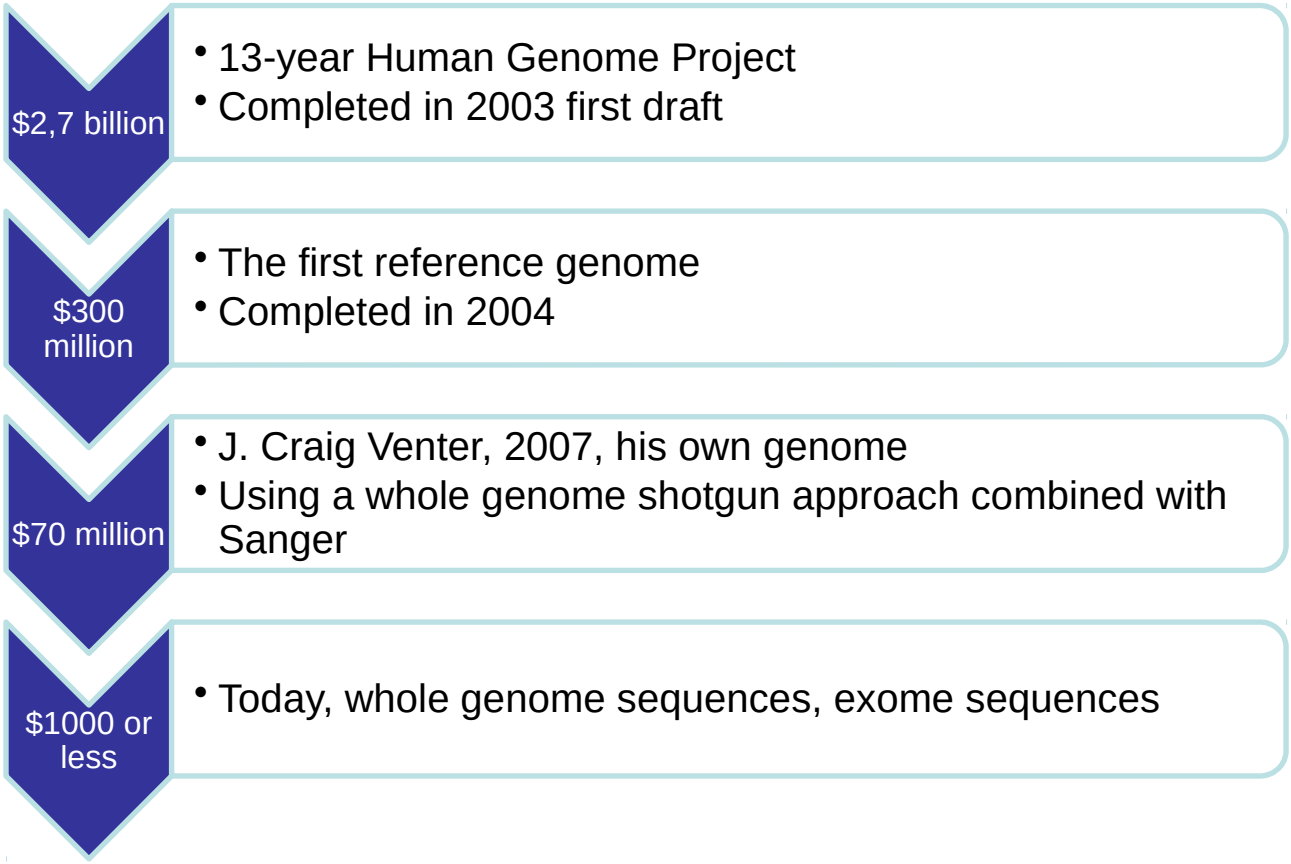


https://flxlexblog.wordpress.com/

Coste actual de la secuenciación



Evolución del coste de la secuenciación de u



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

Differences Between Sanger and Next Gen
samples, tracking, and data relationships

	Sanger	Next Gen
Sequencing Samples	Clones, PCR	DNA libraries
Sample Tracking	Many samples in 96, 384 well plates	Few
Preparation Steps	Few, sequencing reactions clean up	Many, complex procedures
Data collection	Samples in plates 96, 384	Samples on slides 1 - 16+
Data	One read / sample	Thousands and millions reads / sample

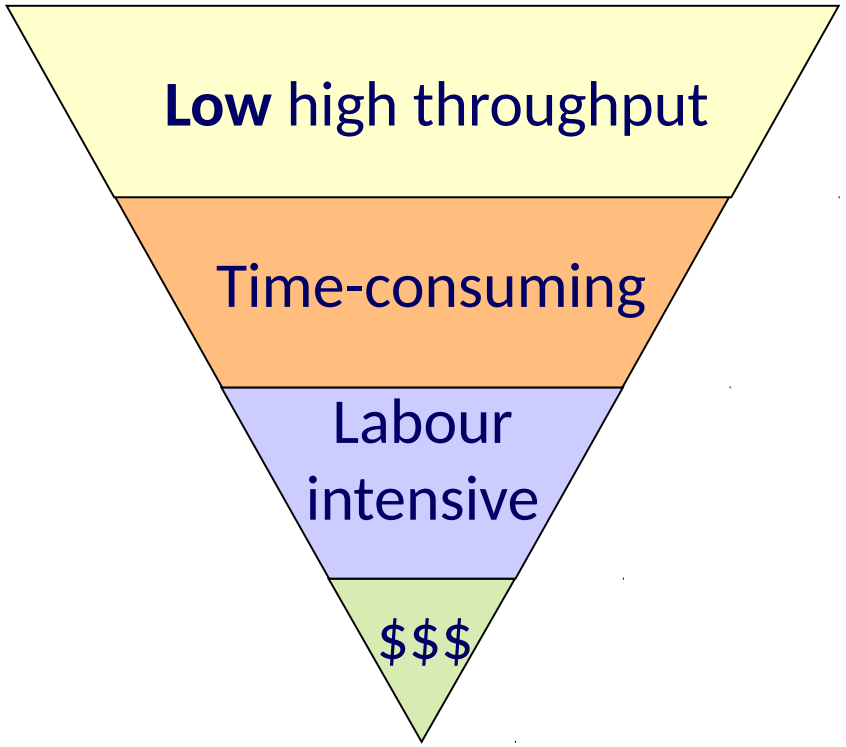
geospiza™ FROM SAMPLE TO RESULTS™

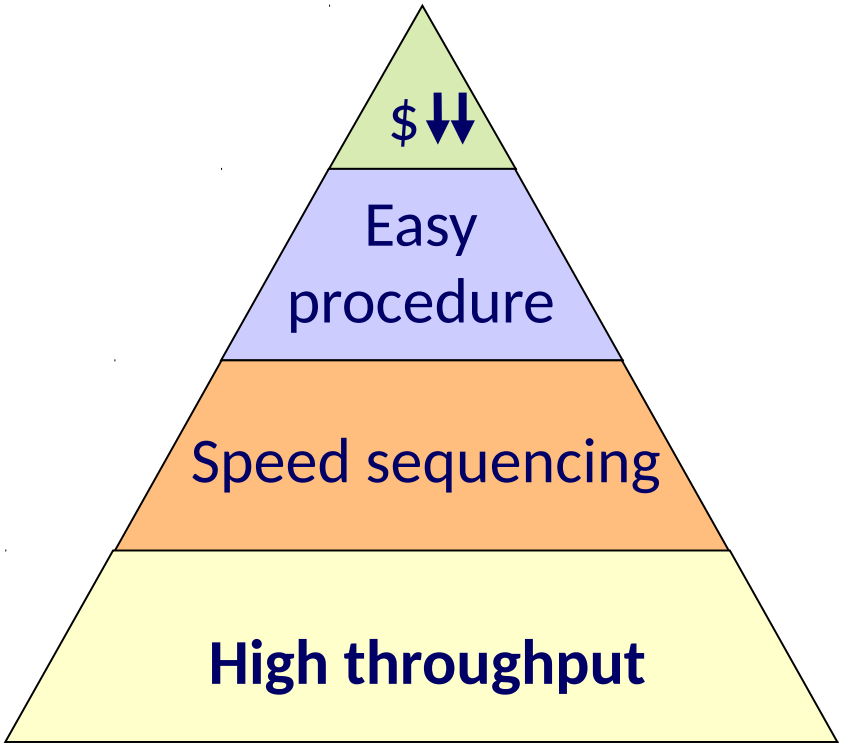
En 2008, hoy 96+

Kits comerciales

Sanger vs NGS

advantages of new technologies

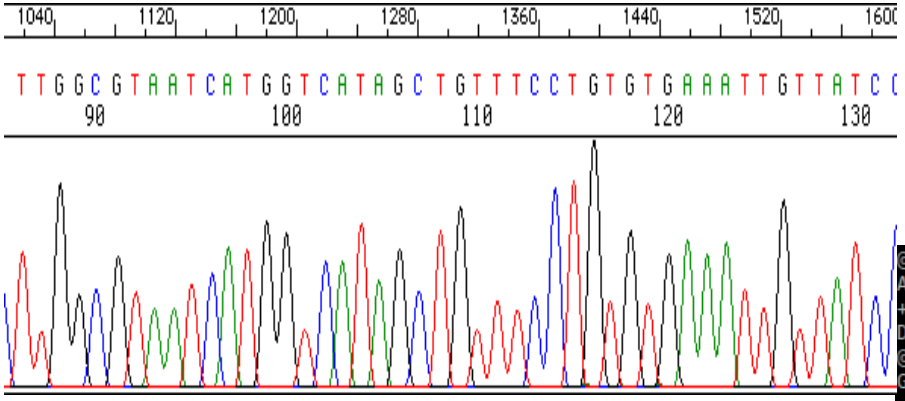




Semiautomatic **Sanger** capillary-based sequencing technology

NGS
Next Generation Sequencing =
Now Generation Sequencing

Nuevo escenario en el análisis de datos




```
@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

Primera generación

- Sanger

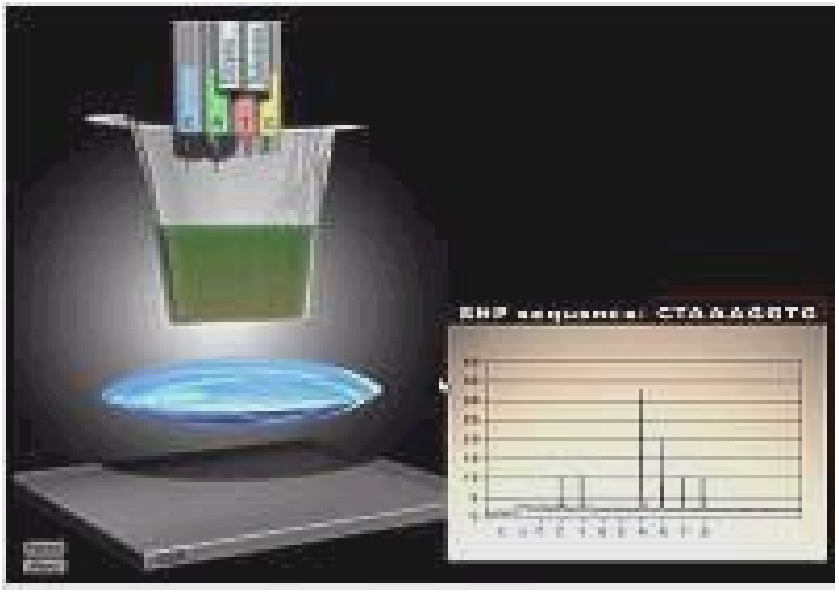
Segunda Generación

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

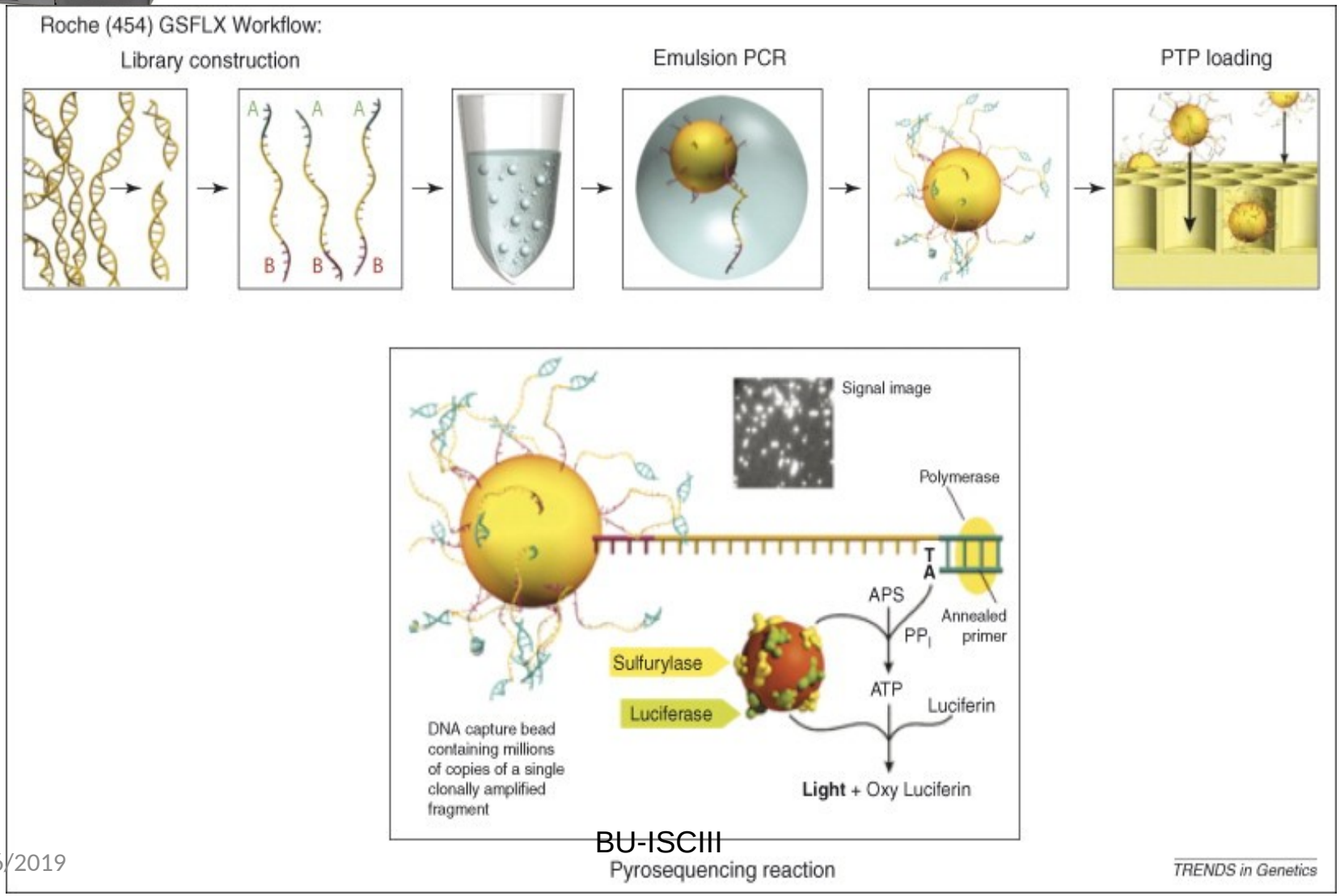
Tercera Generación

- Pacific Biosciences
- Nanopore

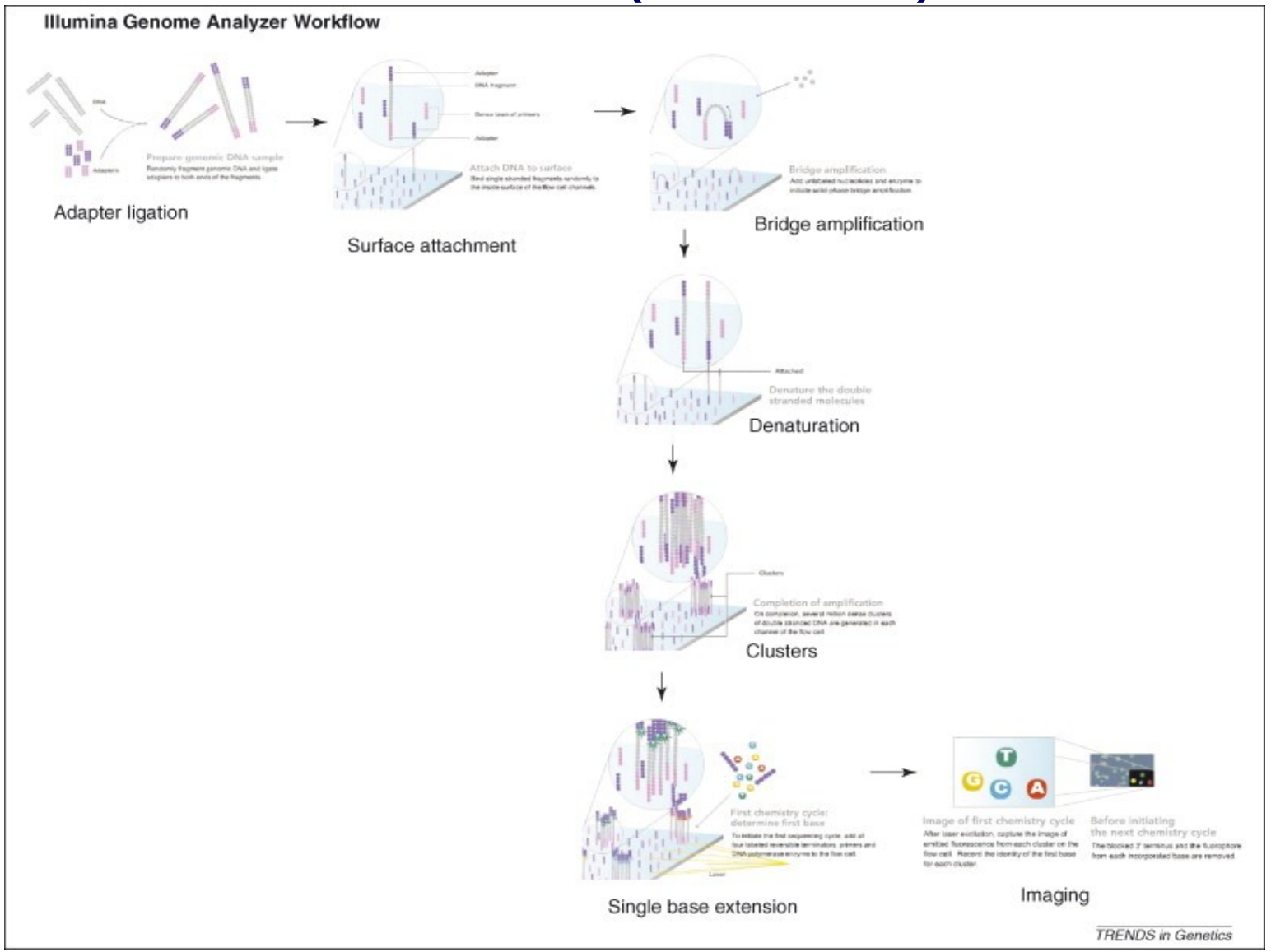
PIROSECUENCIACIÓN

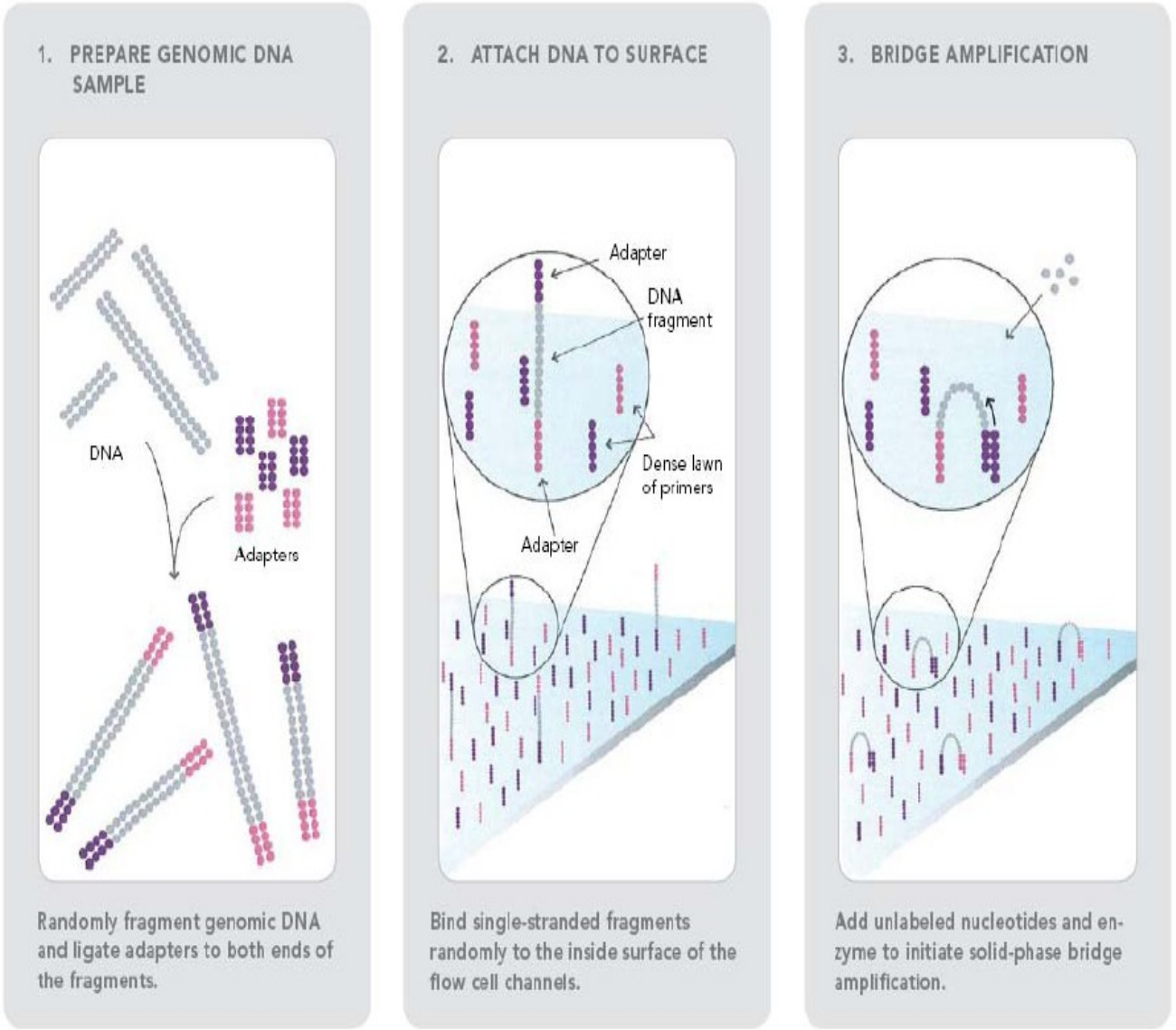


Roche (454) Workflow

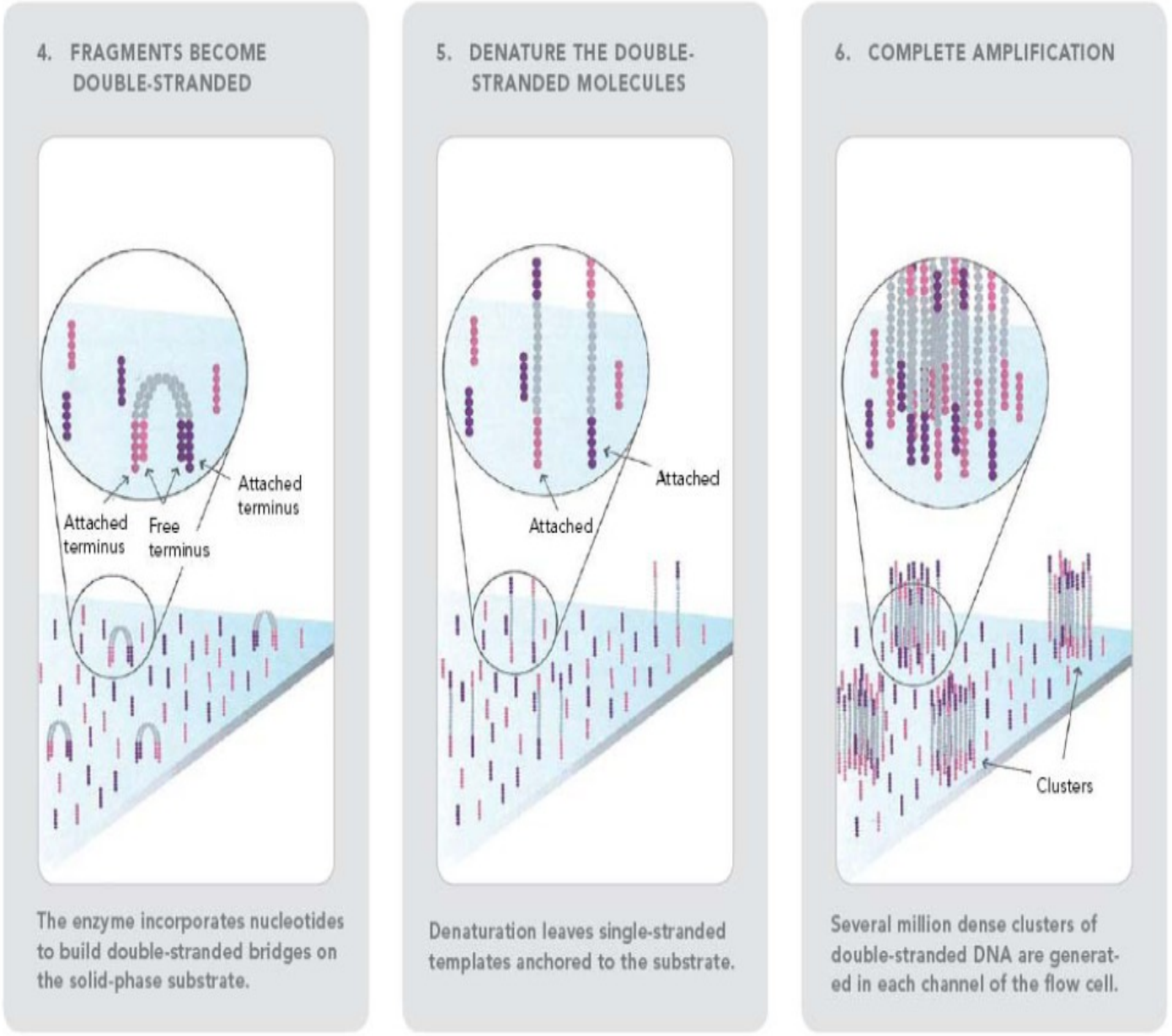


Illumina (Solexa) Workflow

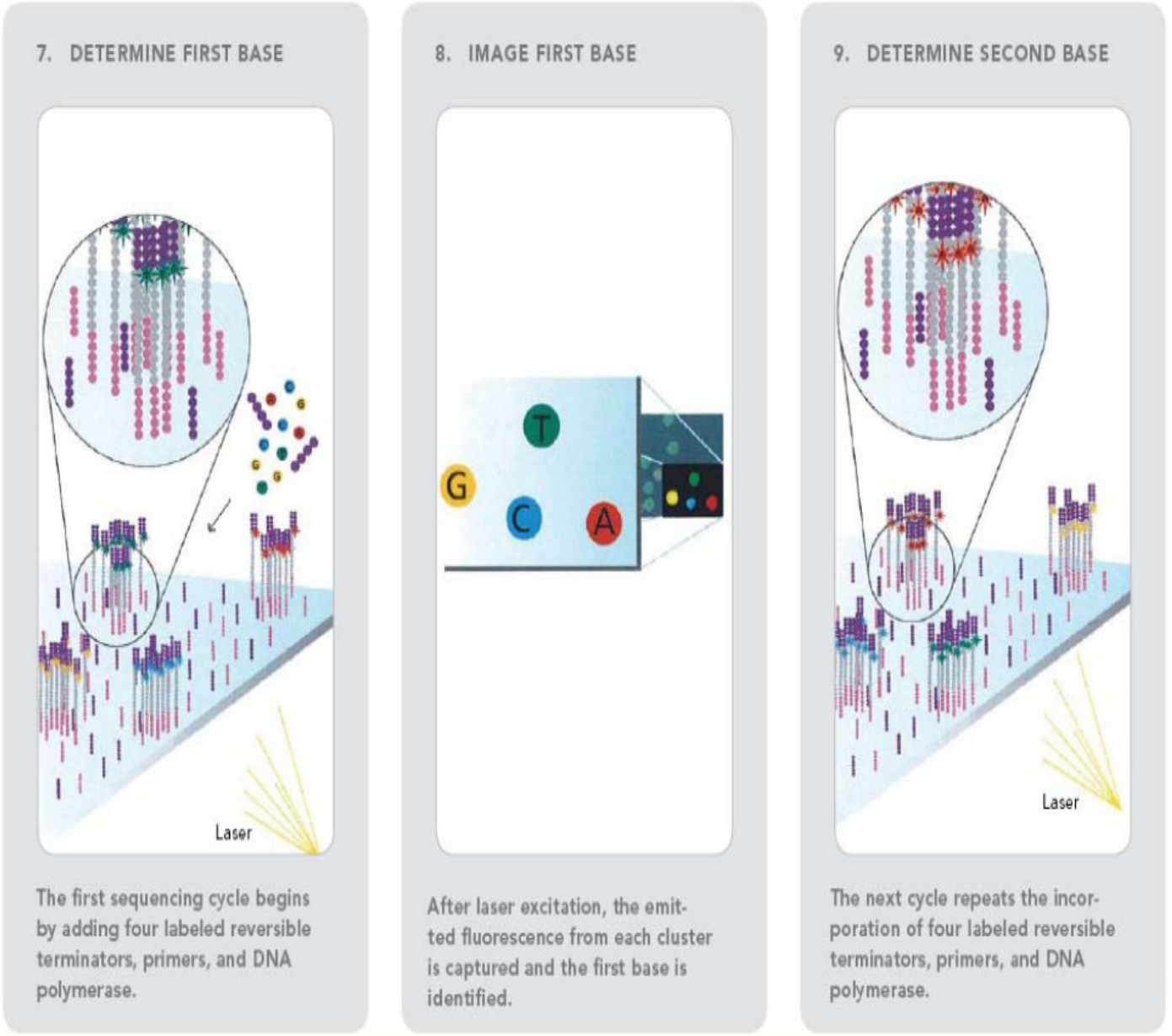




amplificación en fase sólida

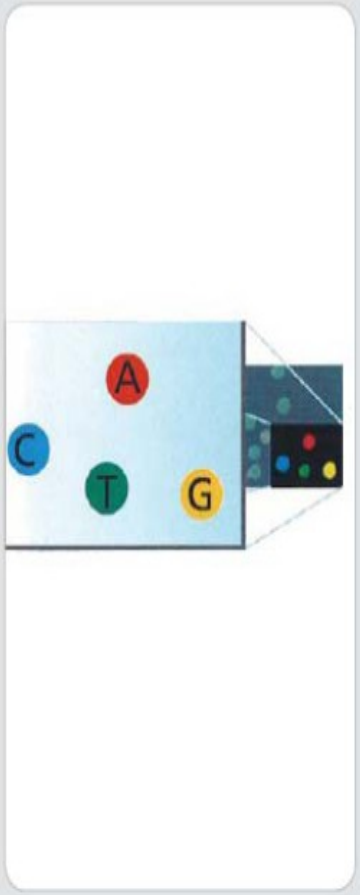


> 1000 copies in ≤ 1 µm; 10^7 clusters per cm^2



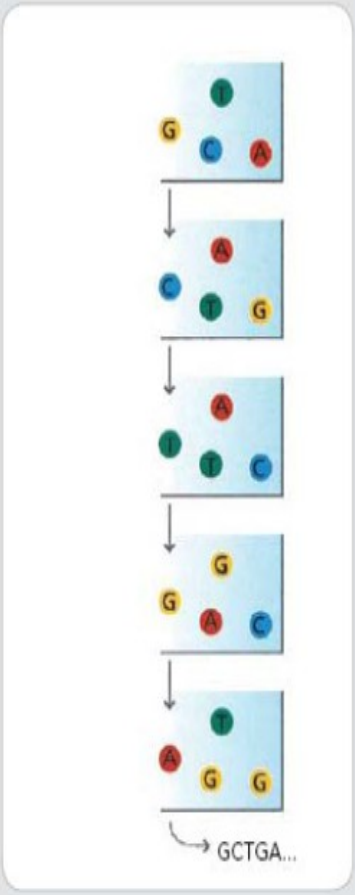
Secuenciación con nucleótidos terminadores reversibles

10. IMAGE SECOND CHEMISTRY CYCLE



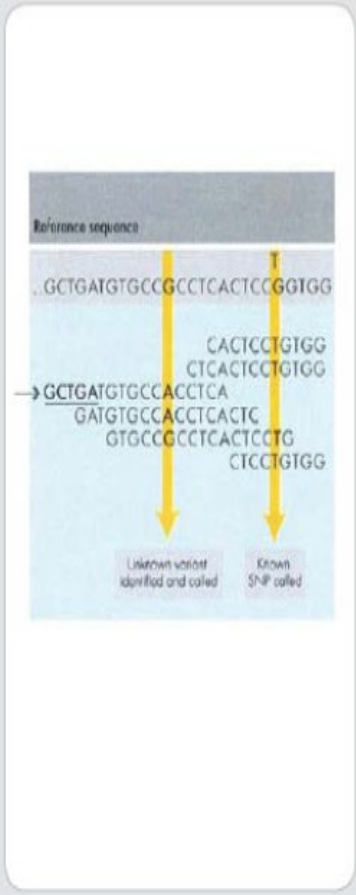
After laser excitation, the image is captured as before, and the identity of the second base is recorded.

11. SEQUENCING OVER MULTIPLE CHEMISTRY CYCLES



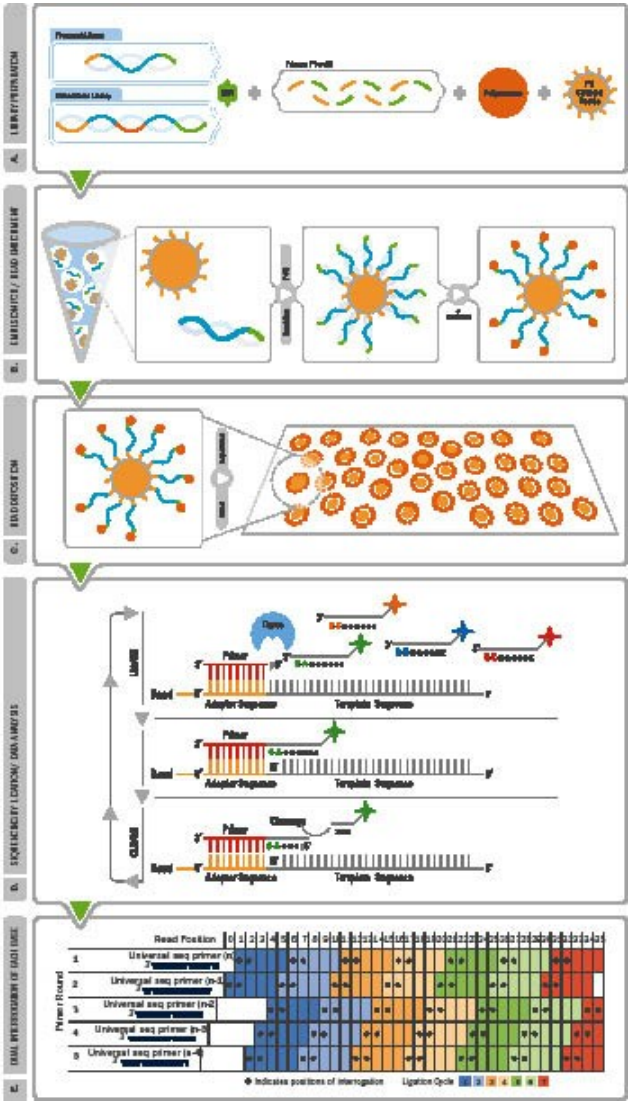
The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

12. ALIGN DATA



The data are aligned and compared to a reference, and sequencing differences are identified.

Life Technologies / Solid



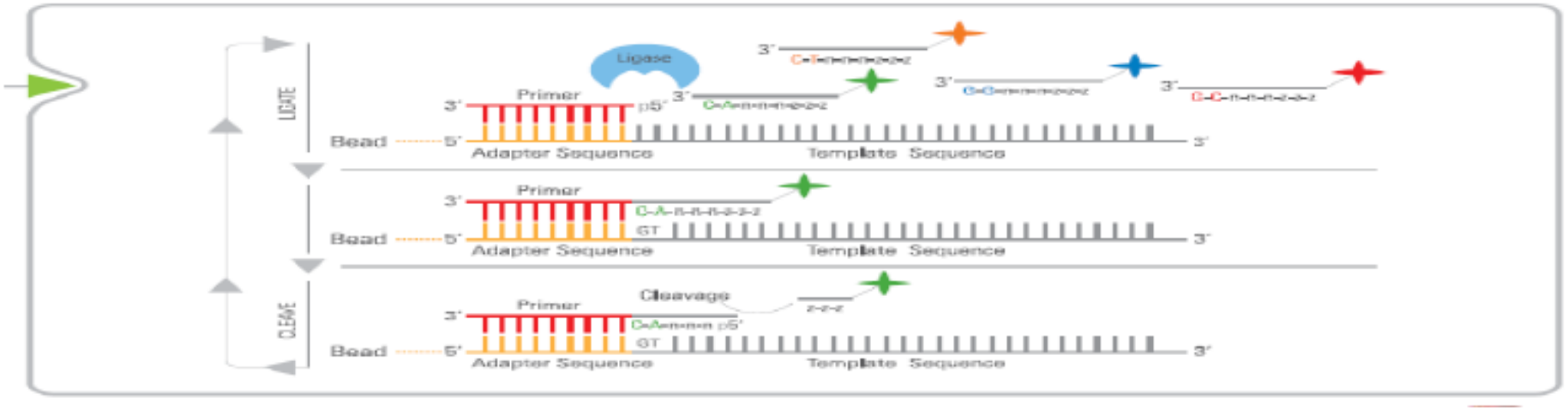
Library Construction

Emulsion PCR

The beads are deposited onto a slide

Life Technologies / Solid

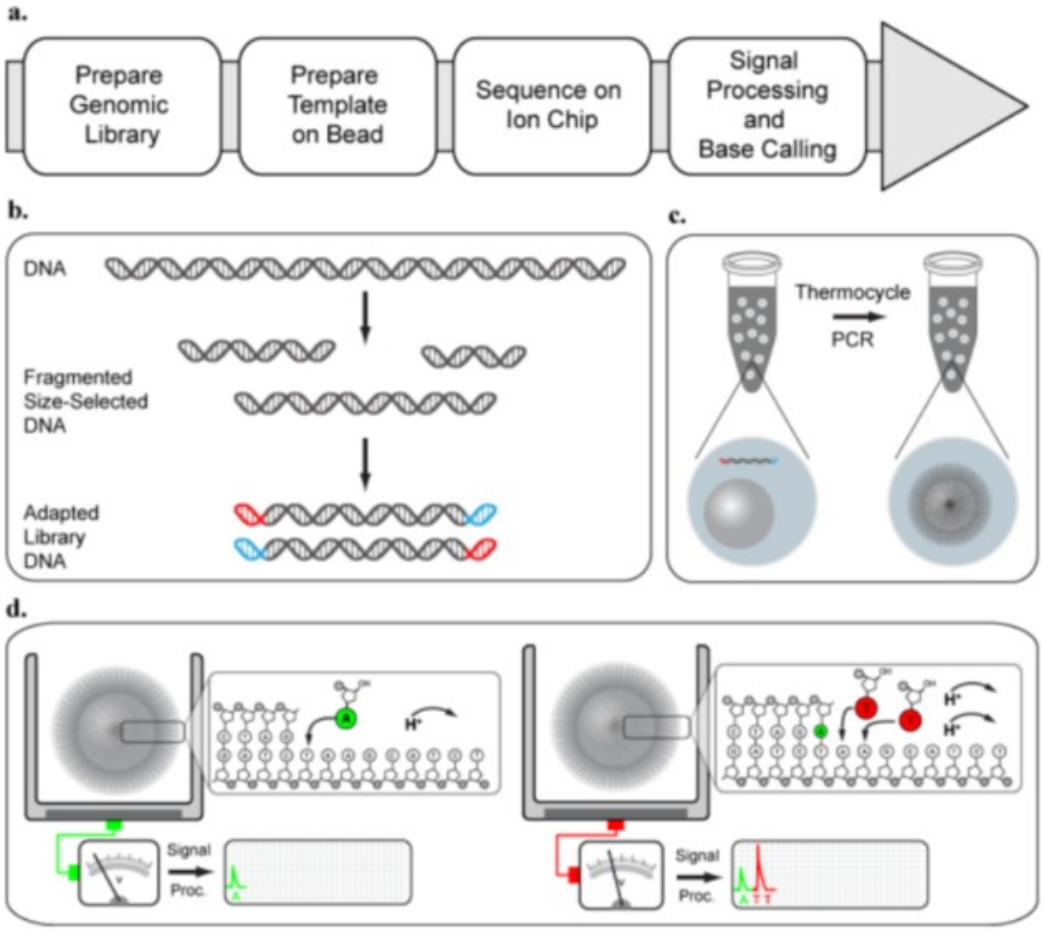
Sequencing by Ligation



Primers hybridize to the P1 adapter sequence within the library template.
A set of four fluorescently labeled di-base probes compete for ligation to the sequencing primer. Specificity of the di-base probe is achieved by interrogation every 1st and 2nd base in each ligation reaction.
Multiple cycles of ligation, detection and cleavage are performed.

Ion Torrent PGM

Personal Genome Machine



Secuenciadores

Primera generación

- Sanger

Segunda Generación

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

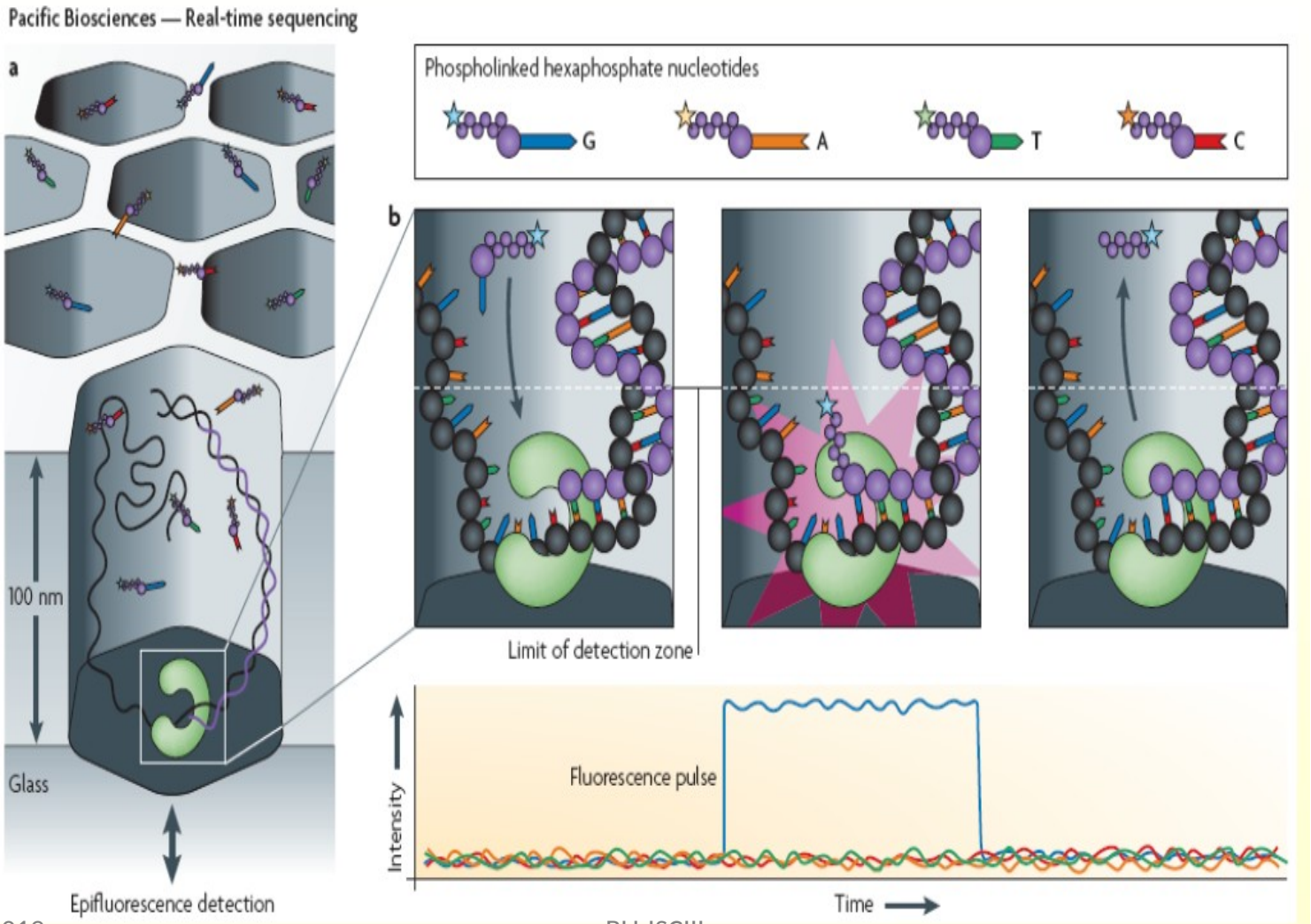
Tercera Generación

- Pacific Biosciences
- Nanopore

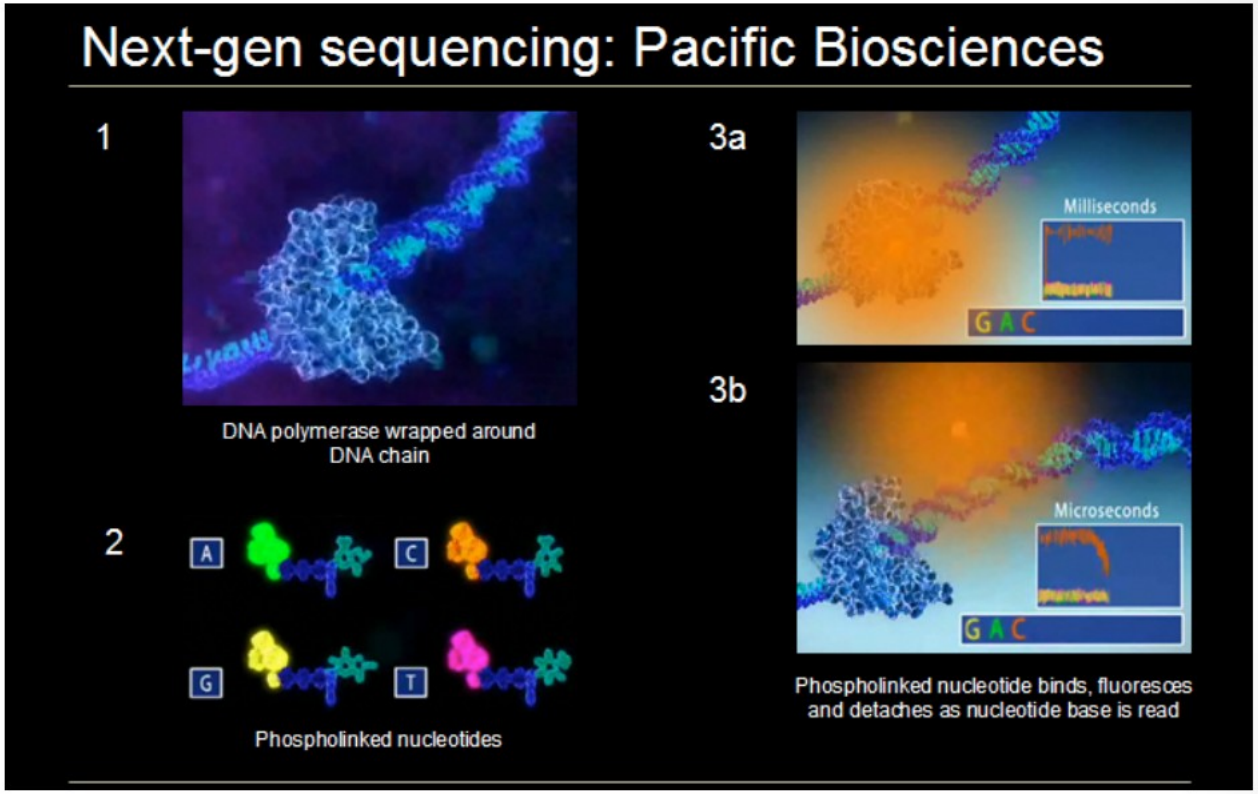
3ª GENERACIÓN: LECTURAS MAS LA

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

ULTRASECUENCIACIÓN A TIEMP



ULTRASECUENCIACIÓN A TIEMP

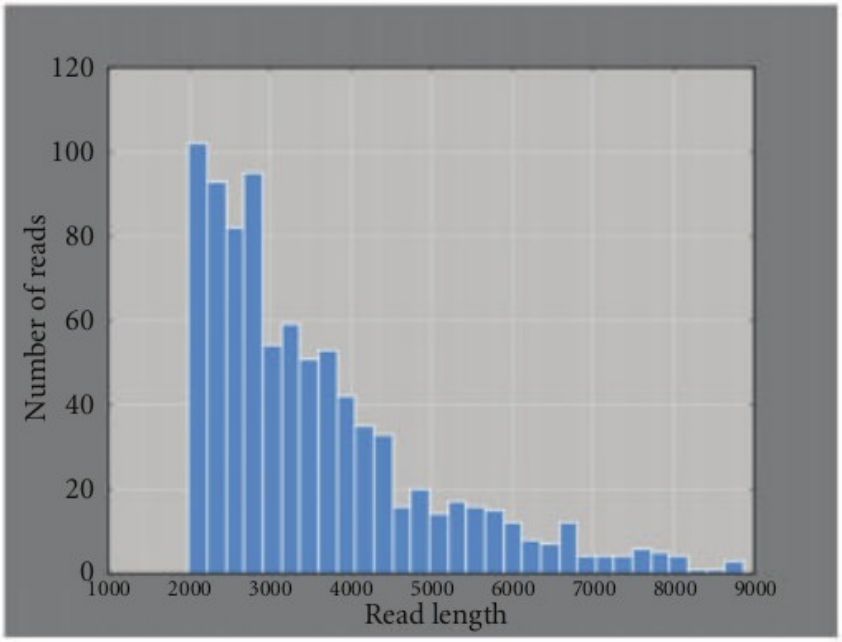


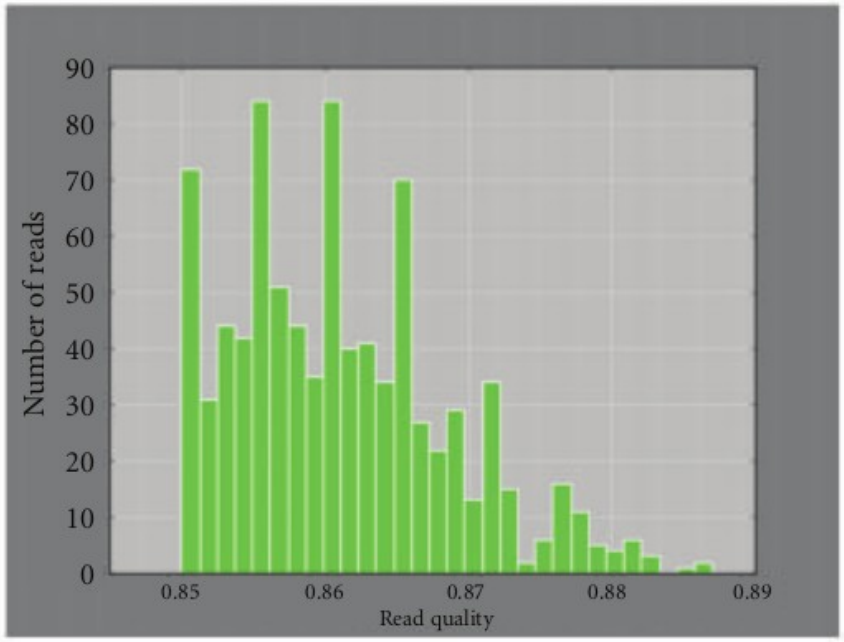
http://www.melanieswan.com/FOLS.html

Secuenciación de un fosmido DNA usando el secuenciador PacBioscience

	Prefilter	Post-QC filter*
Number of bases	84, 110, 272 bp	22, 373, 400 bp
Number of reads	46, 861	6, 754
Mean read length	513 bp	2, 566 bp
Mean read score	0.144	0.819

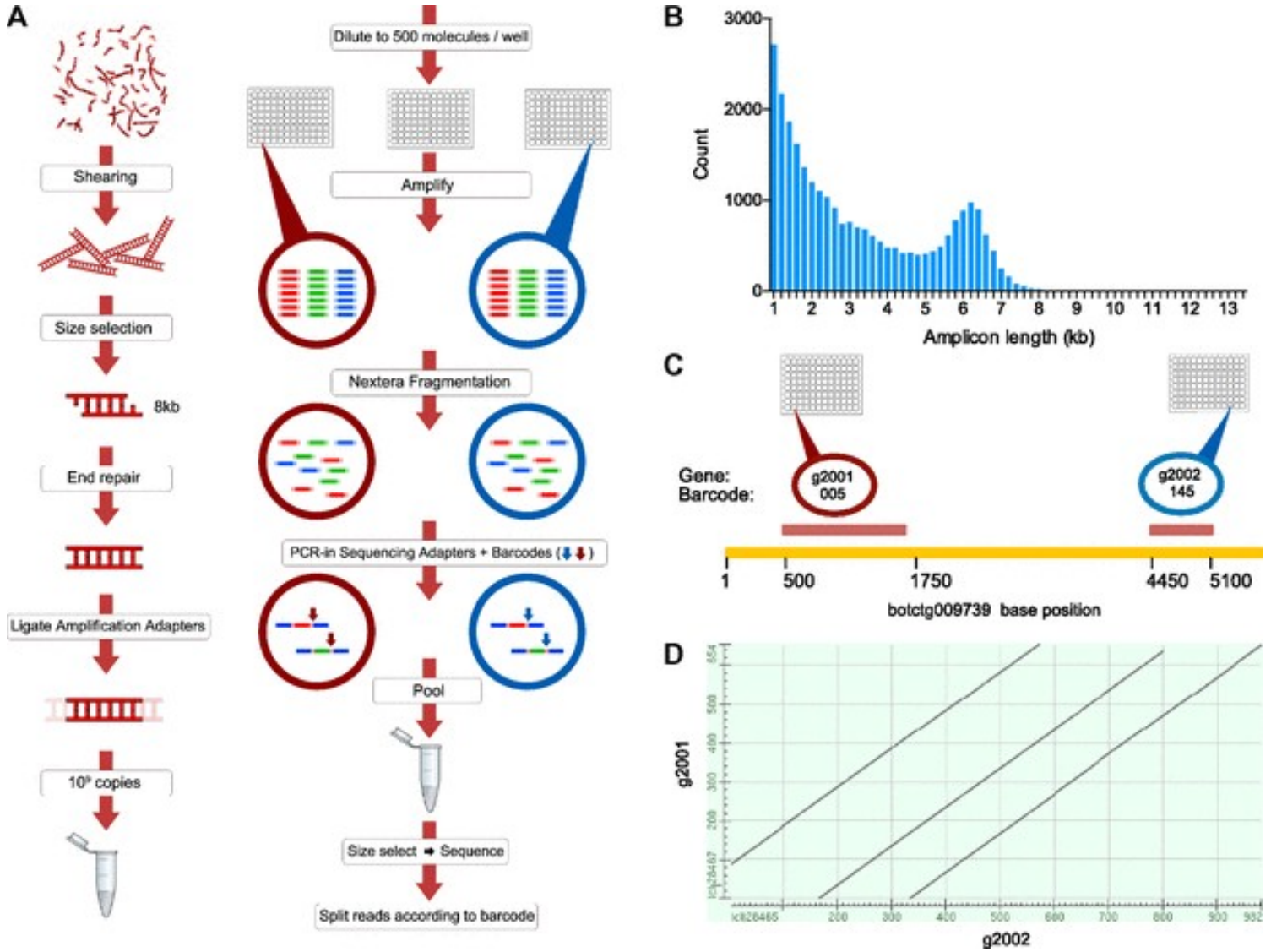
* MinRL = 50, MinRS = 0.75





Liu et al., J. Biomedicine and Biotechnology, 2012

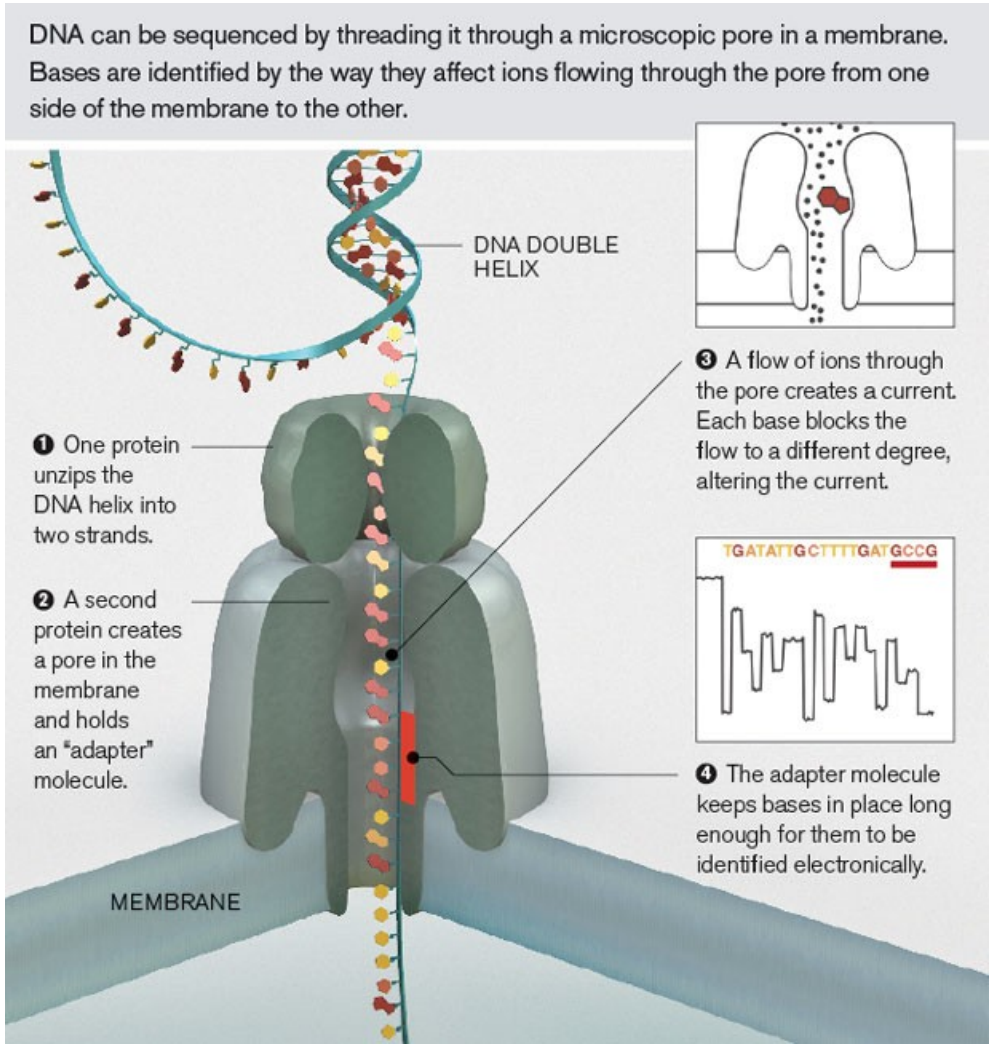
MOLECULO, ILLUMINA



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

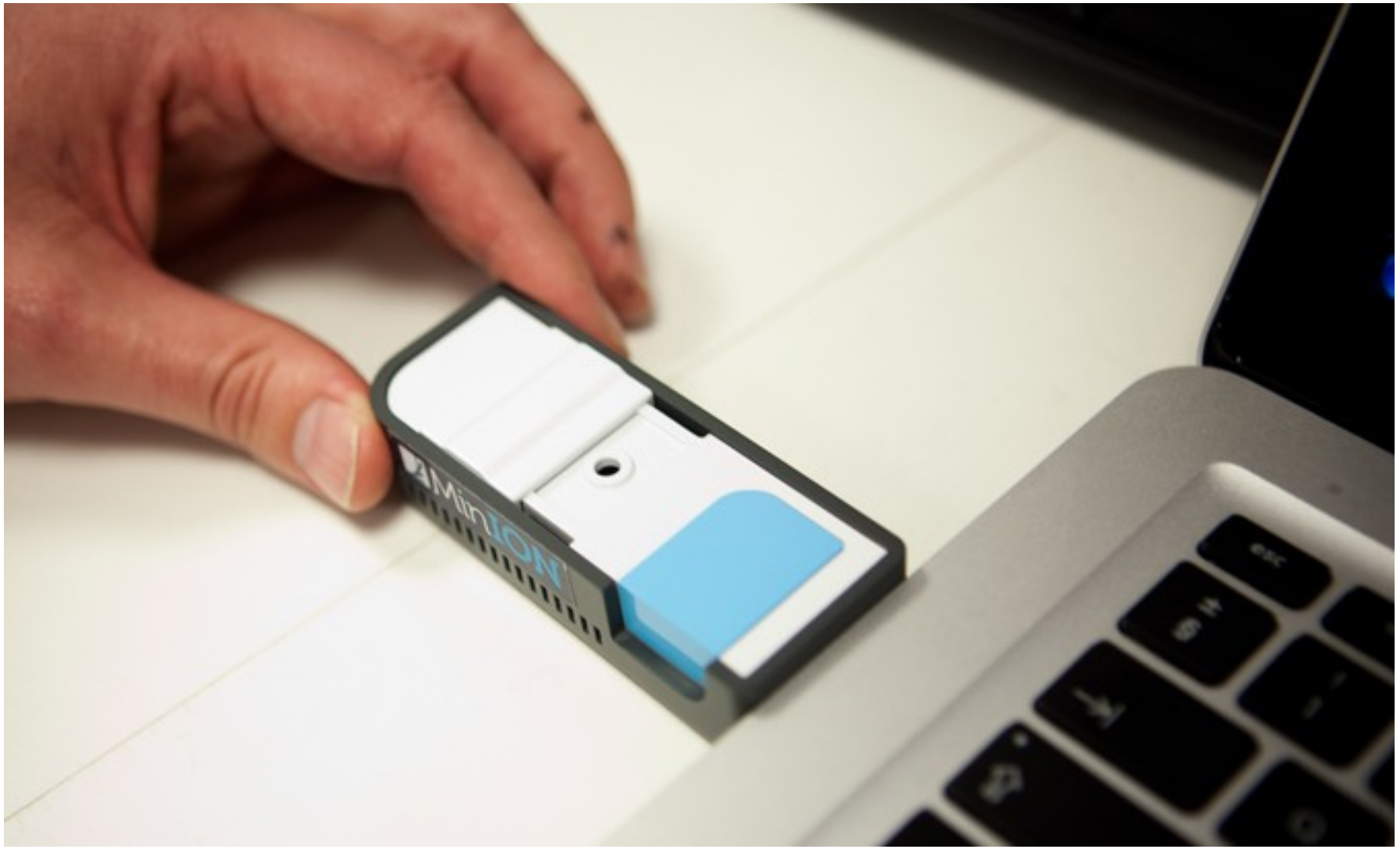
Voskoboynik et al., elife 2013,2:e00569

MinIon, OXFORD NANOPORE



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

MinIon, OXFORD NANOPORE



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

OXFORD NANOPORE

MinION

- Pocket-sized
- Up to 512 nanopore channels
- Simple 10-minute sample prep available
- Many publications illustrate broad usage
- Commercially available

Learn more

Buy MinION

GridION

- Multiple sequencing devices, one compute module
- Use up to five MinION Flow Cells at a time
- Benchtop processor capable of handling high data volumes in real time
- Rapid, real-time applications such as Read Until ...

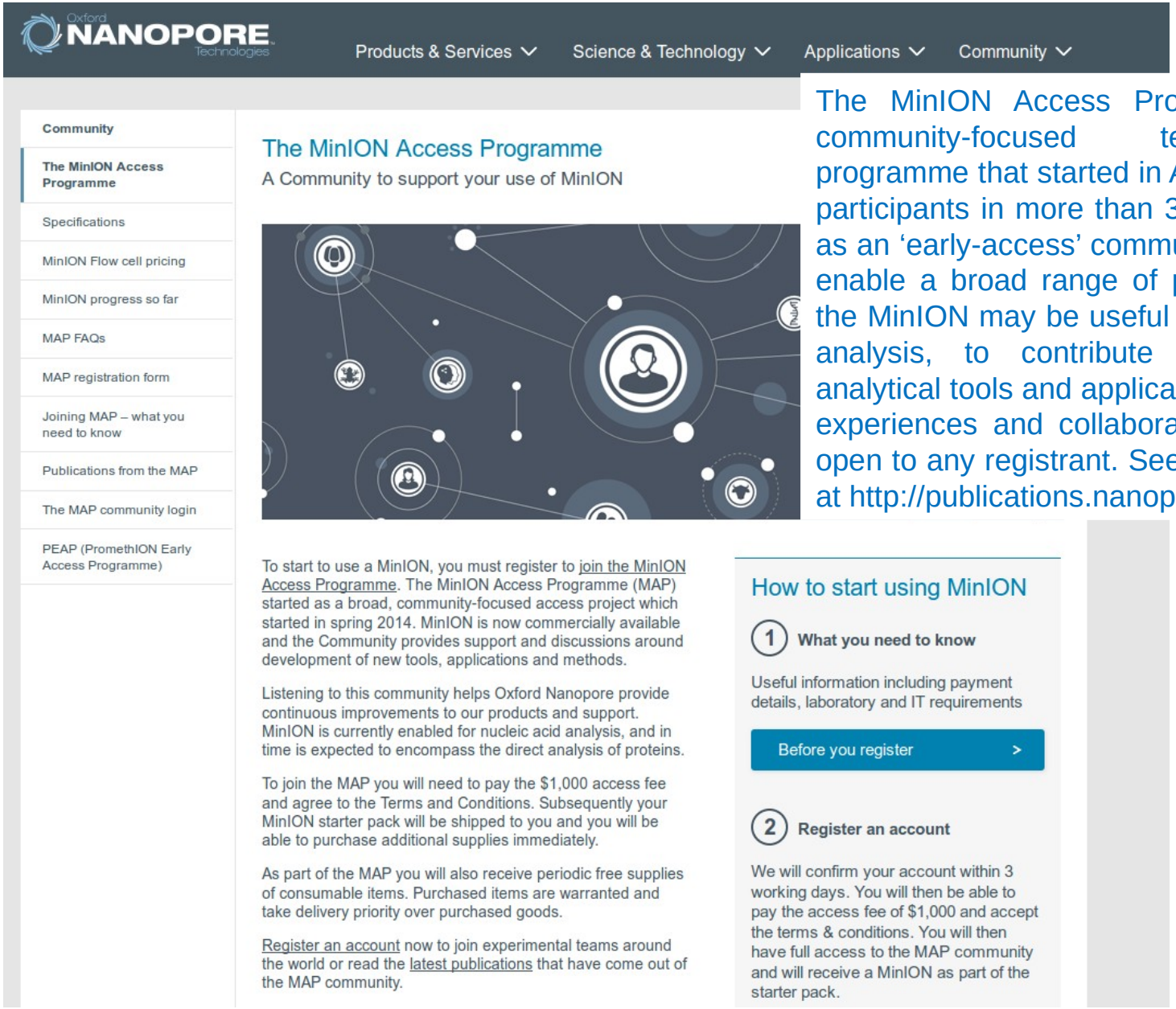
Learn more about GridION

PromethION

- Benchtop system
- Up to 48 flow cells of 3,000 nanopore channels each
- Simple 10-minute sample prep available
- Flow cells may be run individually or concurrently

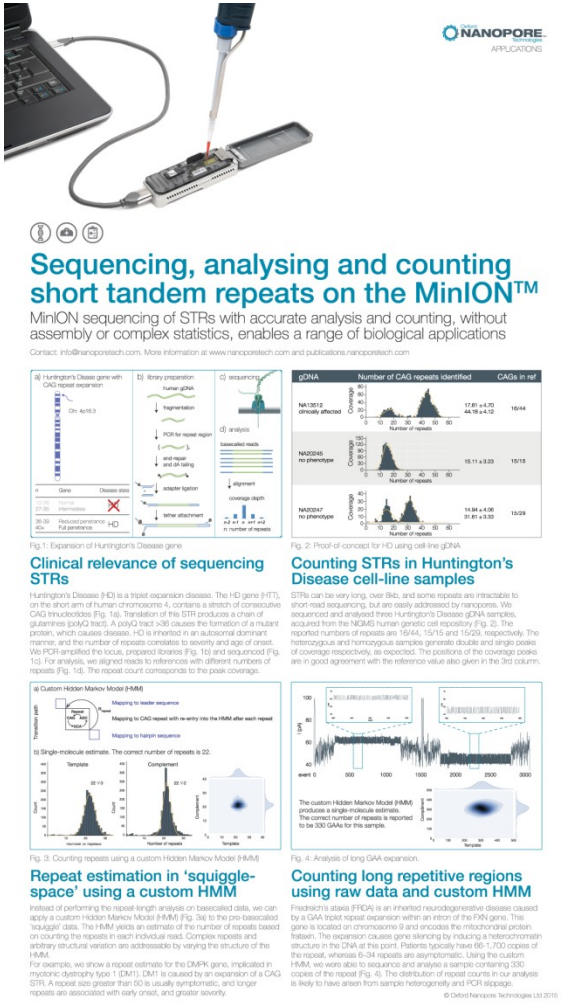
Learn more

Register for early access



The MinION Access Programme (MAP) is a community-focused technology access programme that started in April 2014 and includes participants in more than 30 countries. Launched as an 'early-access' community the MAP seeks to enable a broad range of people to explore how the MinION may be useful to them in nucleic acid analysis, to contribute to developments in analytical tools and applications and to share their experiences and collaborate. The MAP is now open to any registrant. See tools and publications at http://publications.nanoporetech.com

The Nanopore analysis channel on F1000Research will be a central, open platform on which scientists can publish and discuss new applications and analyse workflows for nanopore sequencing data. People can access and contribute data easily, so that the wider life-science community can realise the full potential of this new technology quickly.”



EMBL-EBI Services Research Train

News Press releases

Overview | Leadership | Funding | Background | Collaborations | Jobs | People & groups | News | Events | Visit us | Contact

About us > News > Press releases > Mini DNA sequencer tests true

Mini DNA sequencer tests true

- Public access to Oxford Nanopore's MinION™ miniature sensing device enabled an international consortium to evaluate the technology and provide a standard protocol for its use;
- Preliminary analysis of data generated in five very different laboratories indicates the performance and accuracy of the device is consistently good;
- Data are freely available for re-analysis and innovation in the Nanopore analysis channel on *F1000Research*.

---

15 October 2015 – The MinION, a handheld DNA-sequencing device developed by Oxford Nanopore, has been tested and evaluated by an independent, international consortium coordinated by EMBL's European Bioinformatics Institute (EMBL-EBI). The innovative device opens up new possibilities for using sequencing technology in the field, for example in tracking disease outbreaks, testing packaged food or the trafficking of protected species.

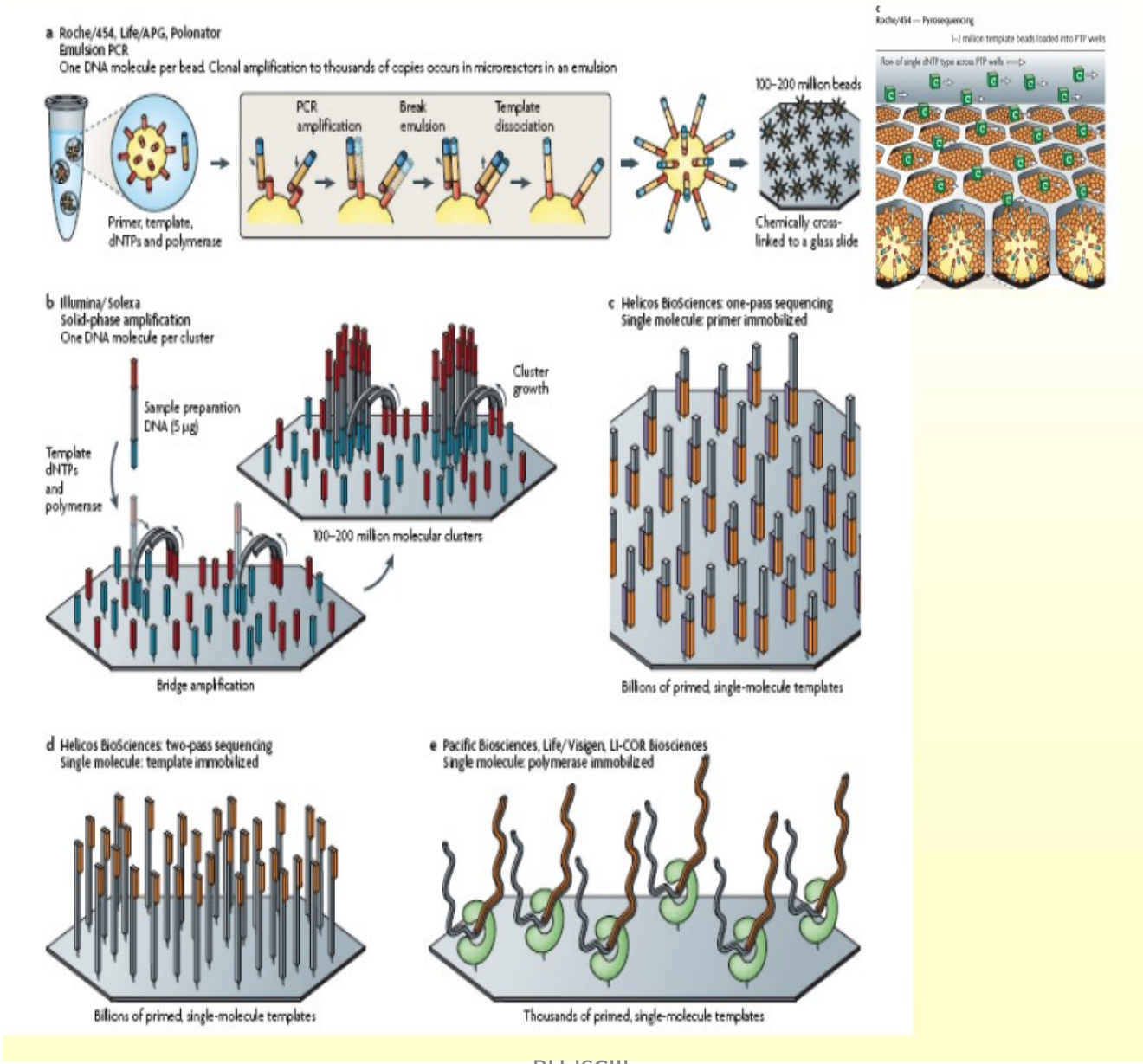
http://www.ebi.ac.uk/about/news/press-releases/mini-dna-sequencer-tests-true

SmidgION, OXFORD NANOPORE

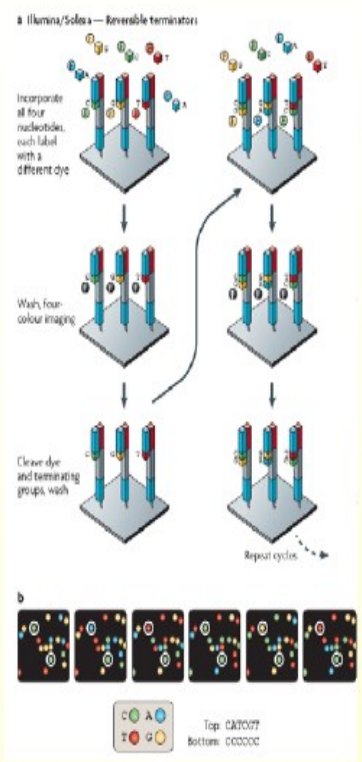
One of the biggest highlights from the talk is SmidgION, the smallest sequencing device that can be plugged into a smartphone for sequencing and analysis. SmidgION will have 256 pores, can sequence for four hours with a smartphone,and produce 230 Mb/hour. SmidgION is expected to be available in late 2017.

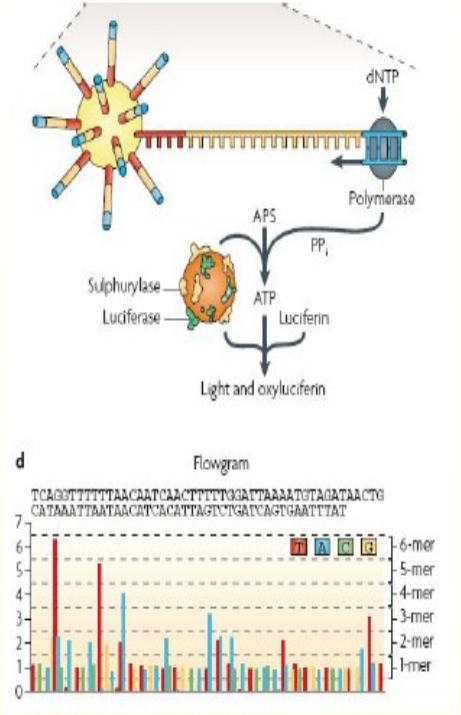
http://nextgenseek.com/2016/05/oxford-nanopore-introduces-smidgion-at-london-calling-2016/

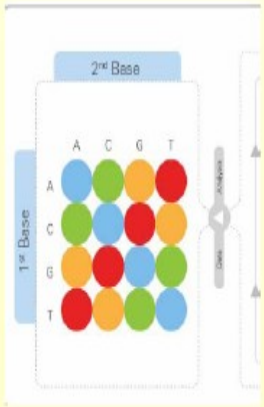
ESTRATEGIAS DE INMOVILIZACIÓN

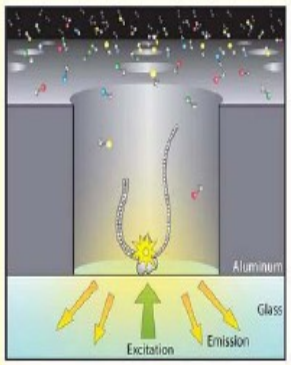


ESTRATEGIAS DE LECTURA

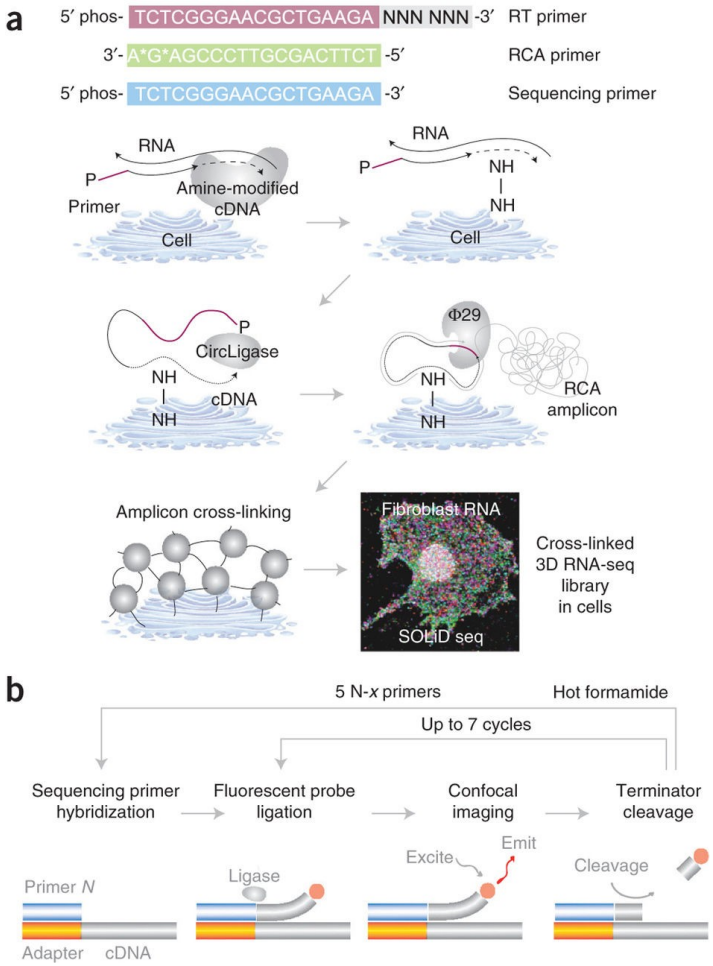








Secuenciación in situ!



Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues

Lee et al., Nat protocols 10 442-458 (2015)

Comparative of NGS platforms

Next-Gen Sequence	Machine Cost	Cost per run	Minimum Throughput	Sequencing Run Time	Cost Per Mb
Illumina MiSeq	$125,000	$750	1500 Mb (2 x 150 Bases)	27 Hours	$0.5
454 GS Junior	$108,000	$1,100	35 Mb (400 Bases)	8 Hours	$31
Ion Torrent PGM - 314 Chip	$80,490	$225	10Mb (100 Bases)	3 Hours	$22.5
Ion Torrent PGM - 316 Chip	$80,490	$425	100Mb	3 Hours	$4.25
Ion Torrent PGM - 318 Chip	$80,490	$625	1000Mb	3 Hours	$0.63

http://nextgenseek.com/2012/08/comparing-price-and-tech-specs-of-illumina-miseq-ion-torrent-pgm-454-gs-junior-and-pacbio-rs/

Comparative of NGS platforms

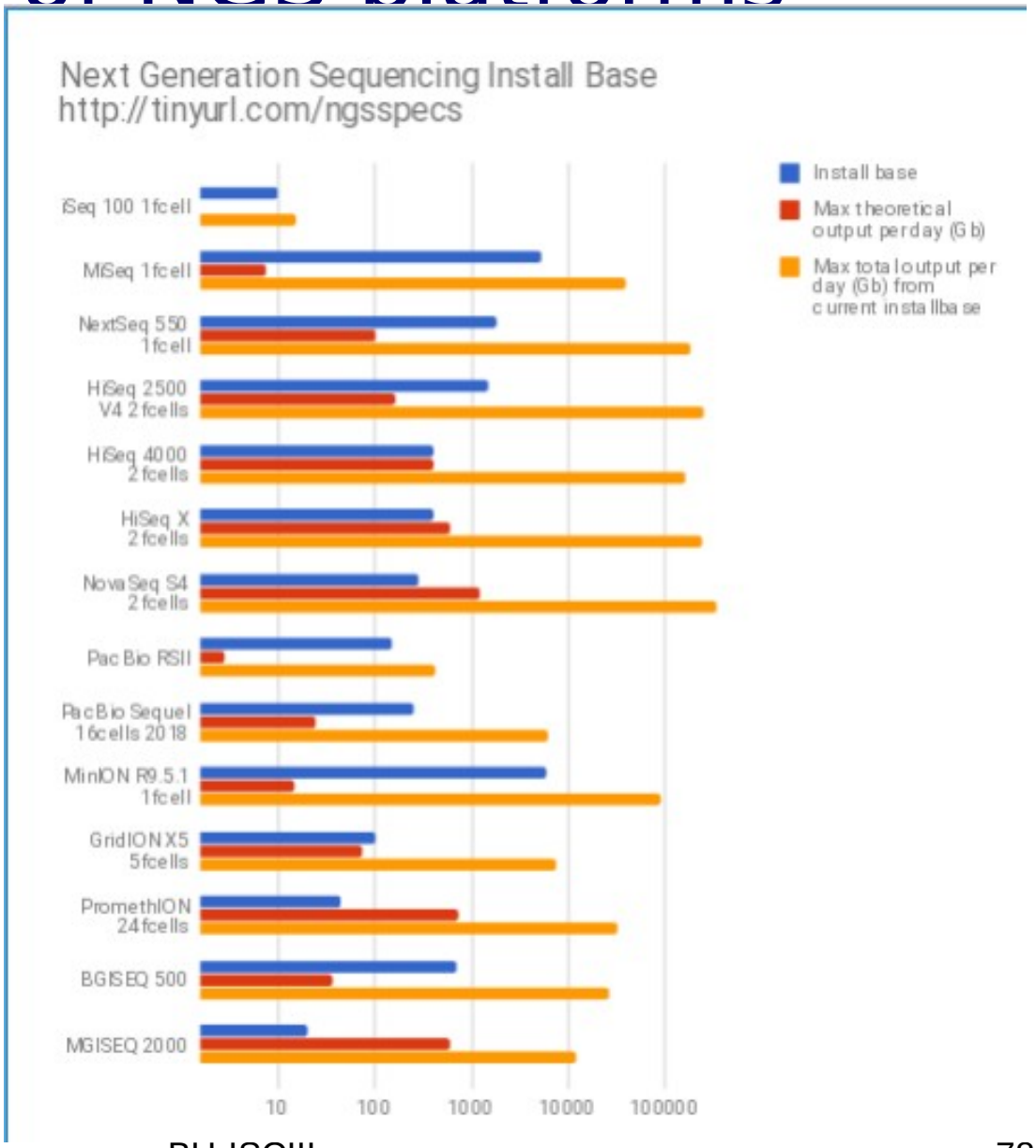
	Machine Cost	Throughput Per Run	Sequencing Cost per Gb	Sequencing Run Time	Observed Error rate
Illumina MiSeq	$128,000	1.5-2 Gb	$502	27 Hours	0.8%
PacBio RS	$695,000	100 Mb	$2000	2 Hours	12.86%
Ion Torrent -318 Chip	$80,000	1Gb	$1000	2 Hours	1.71%
Illumina GAIIx	$256,000	30 Gb	$148	10 Days	0.76%
Illumina HiSeq 2000	$654,000	600 Gb	$41	11 Days	0.26%

http://nextgenseek.com/2012/08/comparing-price-and-tech-specs-of-illumina-miseq-ion-torrent-pgm-454-gs-junior-and-pacbio-rs/

Comparative of NGS platforms

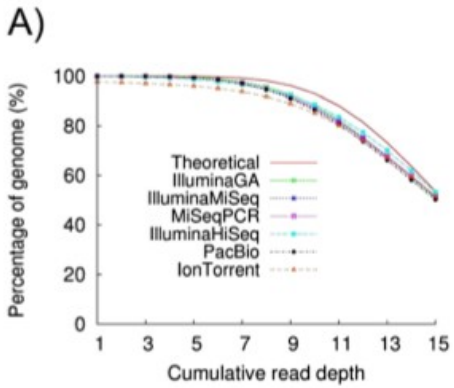
https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/edit?hl=en_GB&hl=en_GB#gid=1569422585

https://twitter.com/albertvilella/status/655814404106244096



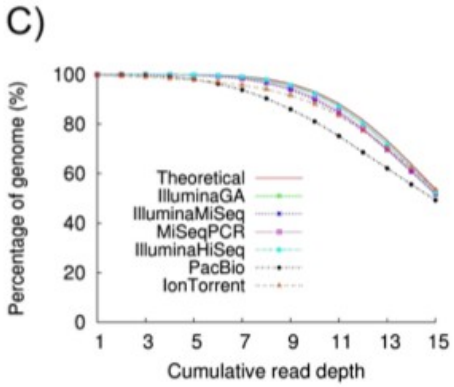
Uniformidad de cobertura a lo largo del genoma

%GC



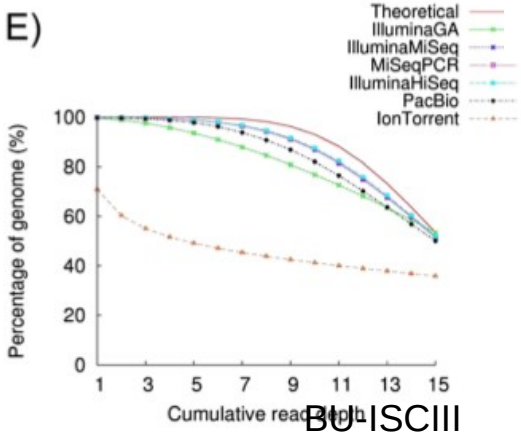
Bordetella pertussis

%GC



Staphylococcus aureus

%AT



Plasmodium falciparum

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

Variación en la cobertura dependiendo del Sistema de Secuenciación

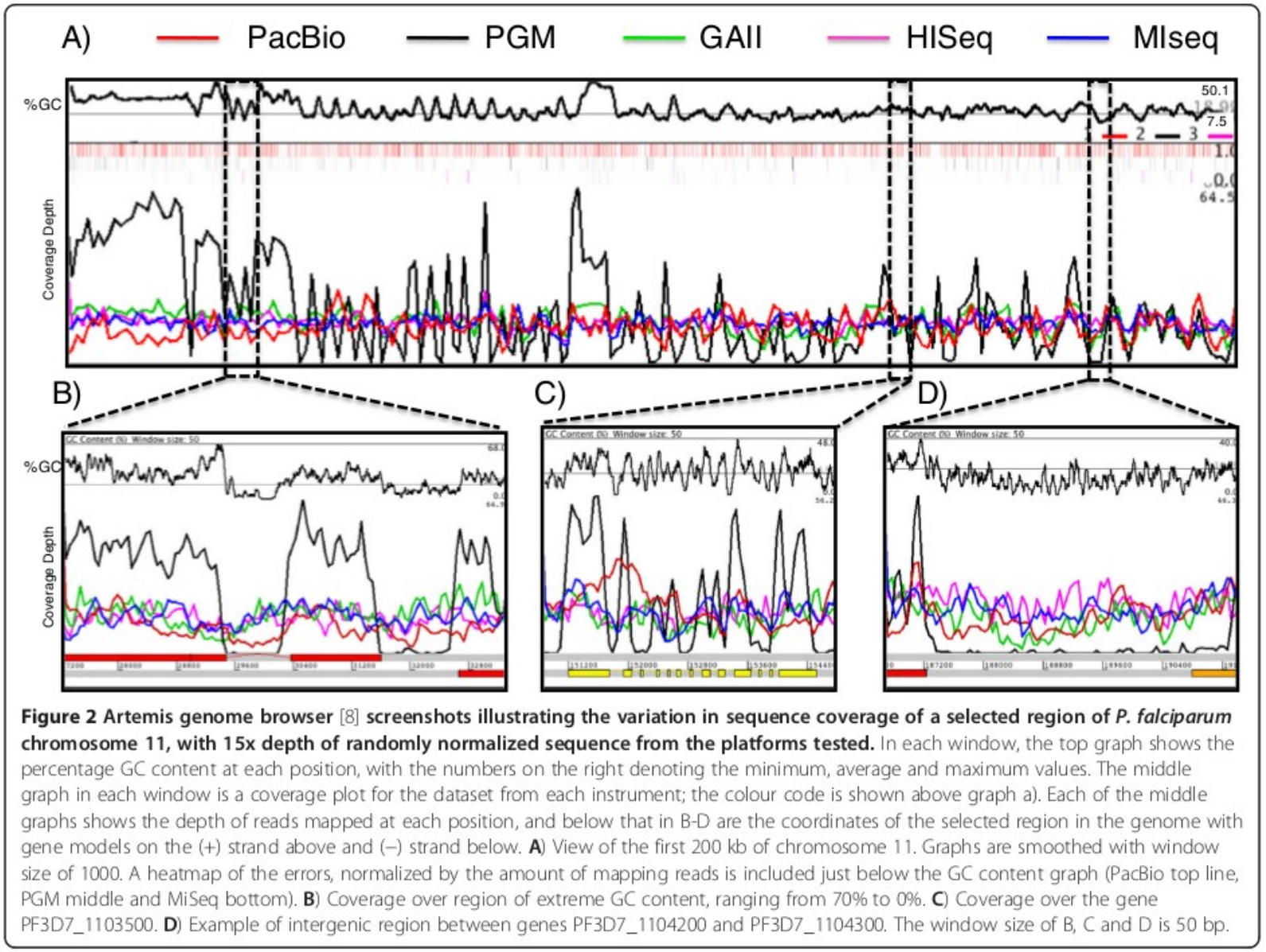


Figure 2 Artemis genome browser [8] **screenshots illustrating the variation in sequence coverage of a selected region of *P. falciparum* chromosome 11, with 15x depth of randomly normalized sequence from the platforms tested.** In each window, the top graph shows the percentage GC content at each position, 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

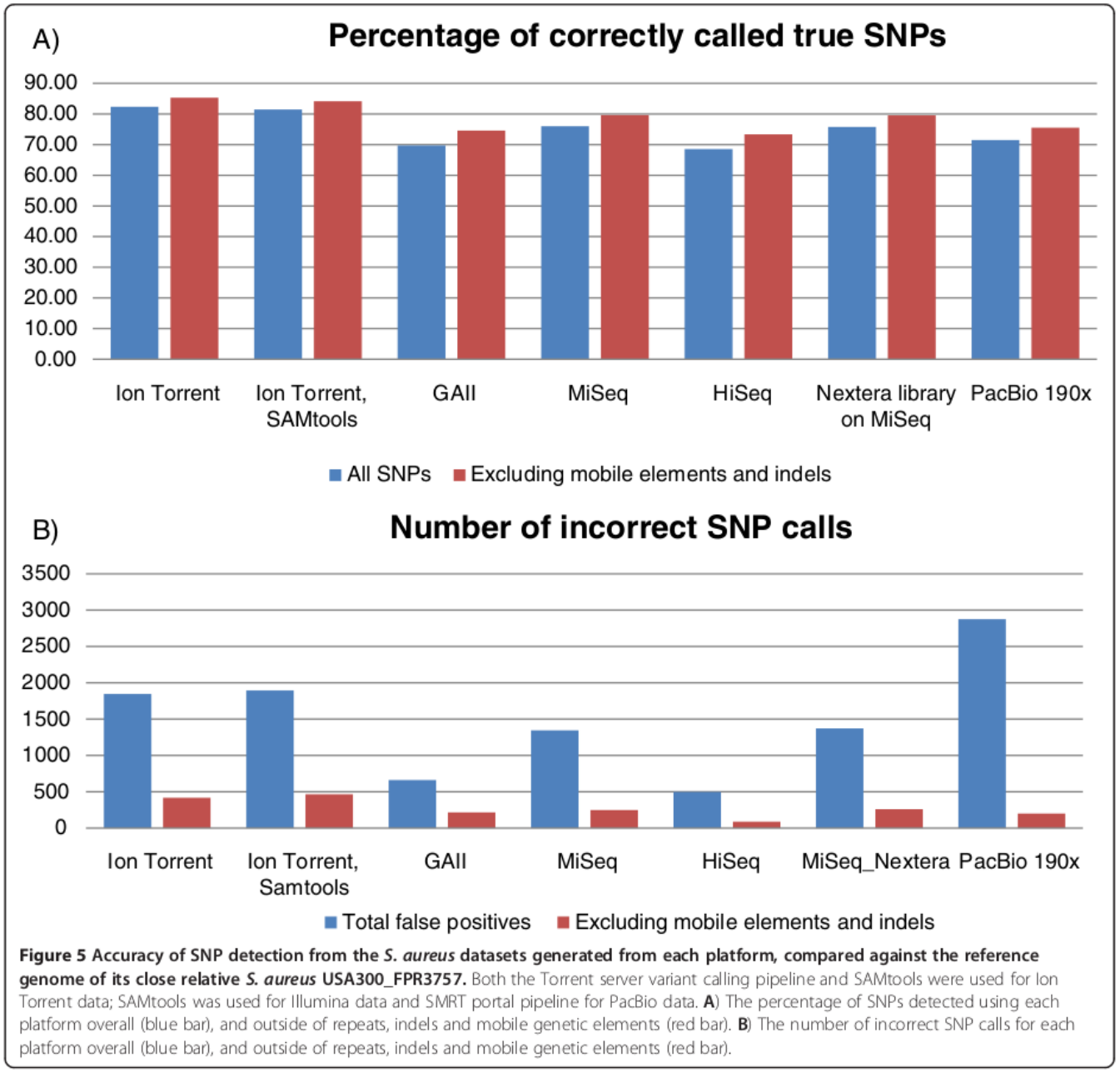


Figure 5 Accuracy of SNP detection from the *S. aureus* datasets generated from each platform, compared against the reference genome of its close relative *S. aureus* USA300_FPR3757. Both the Torrent server variant calling pipeline and SAMtools were used for Ion Torrent data; SAMtools was used for Illumina 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

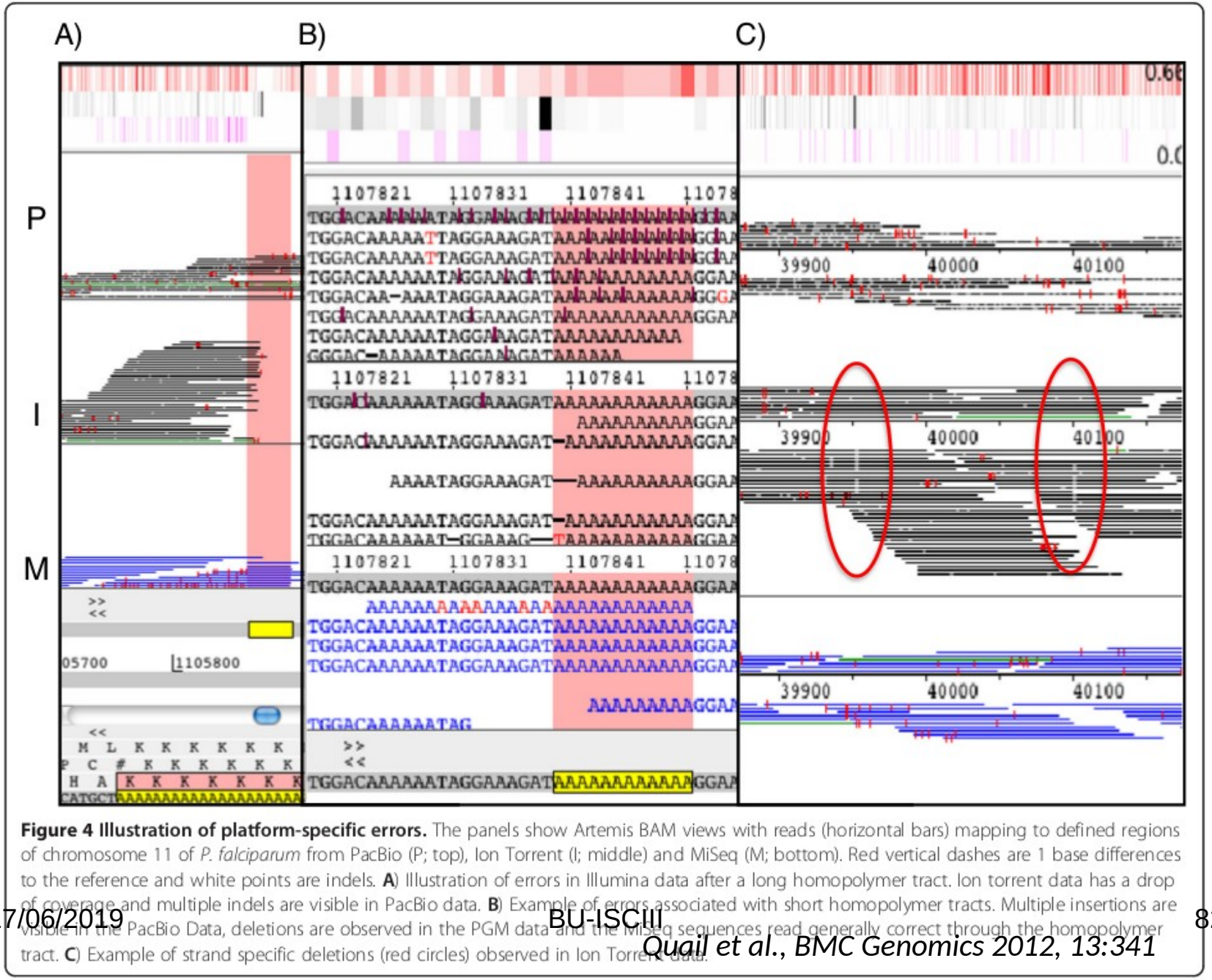


Figure 4 Illustration of platform-specific errors. The panels show Artemis BAM views with reads (horizontal bars) mapping to defined regions of chromosome 11 of *P. falciparum* from PacBio (P; top), Ion Torrent (I; middle) and MiSeq (M; bottom). Red vertical dashes are 1 base differences to the reference 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.
- Illumina, preparación de librería adecuada a amplicones para generar correctamente los clusters
- Aplicaciones dependientes de plataforma

Conclusiones

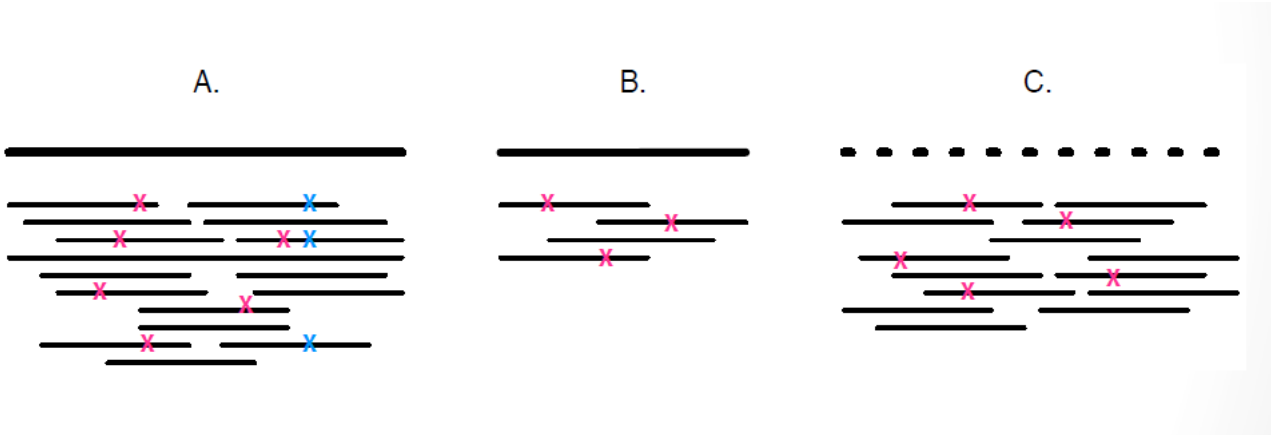
A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers
Quail et al., BMC Genomics 2012, 13:341

- 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

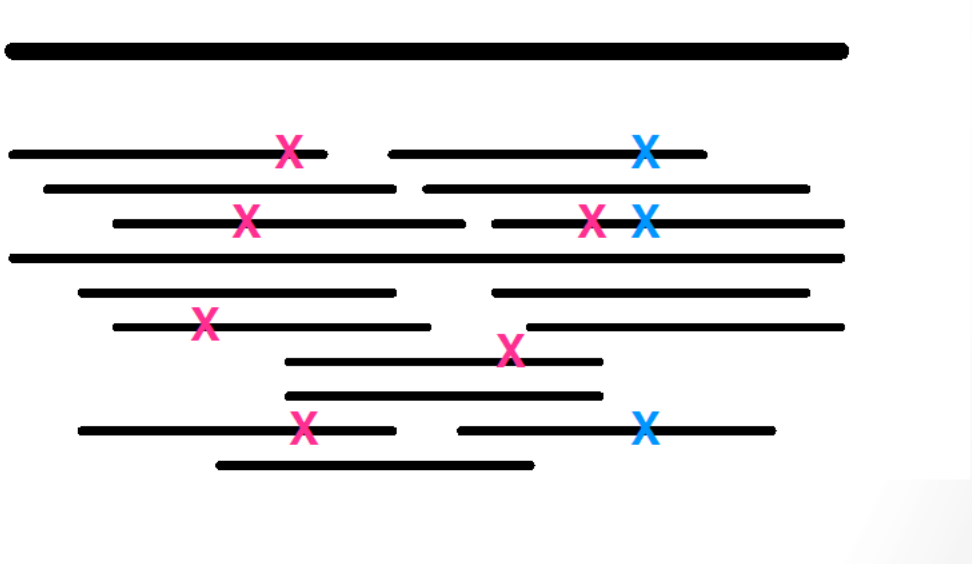
Básicamente tres problemas

Resecuenciación, Conteo y ensamblado



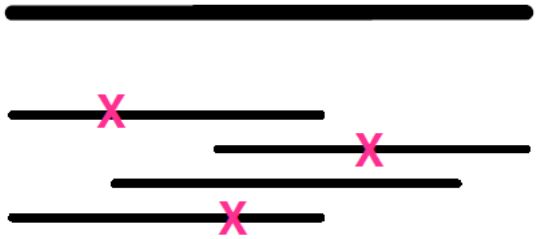
Resecuenciación

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



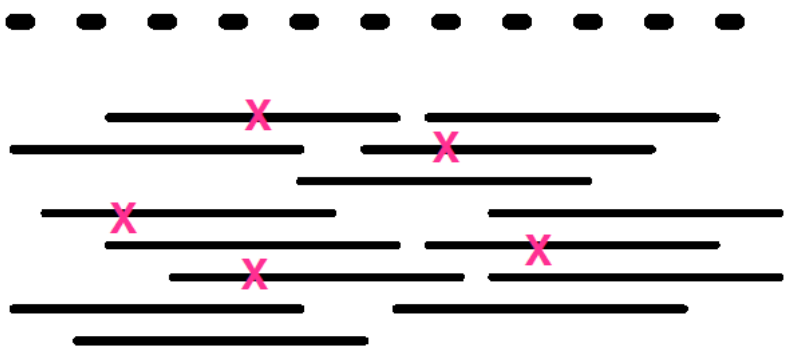
Conteo

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



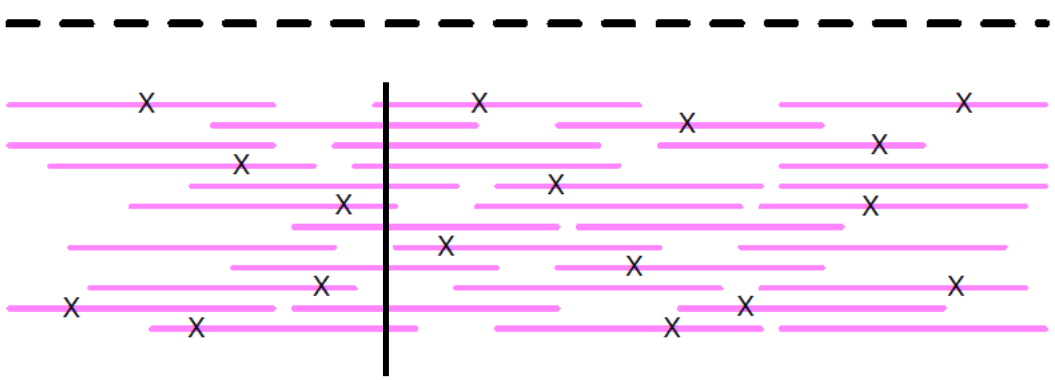
Ensamblado

No hay genoma de referencia y lo construimos de novo



Cobertura (depth of covarage)

Número medio de lecturas por base. i.e. 10x



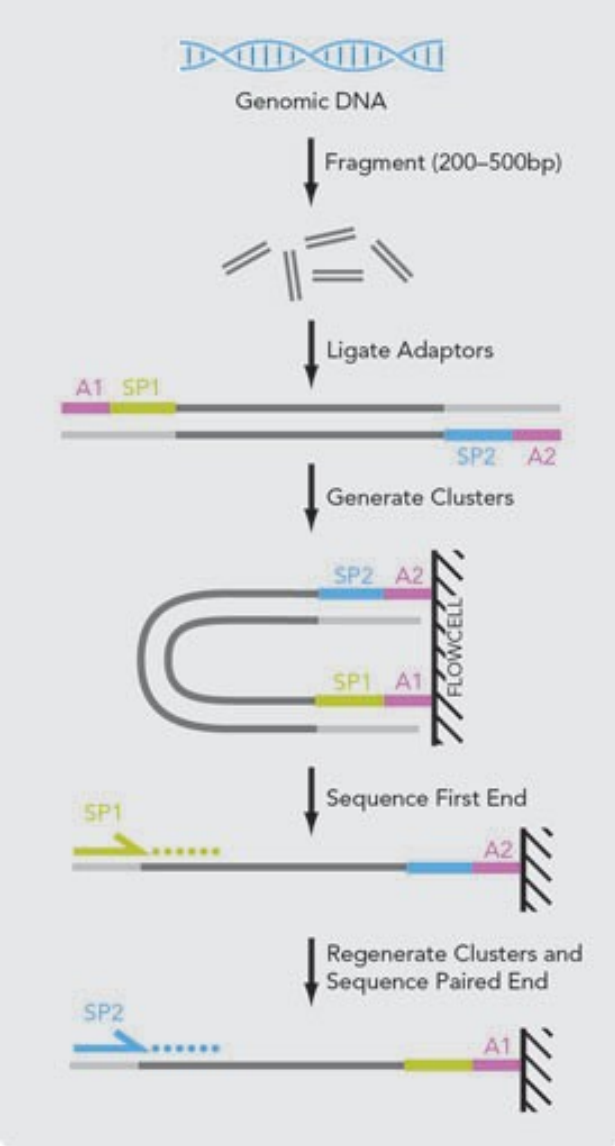
Sequencing coverage

as the average number of reads that align known reference bases

Number of reads x read length / target size

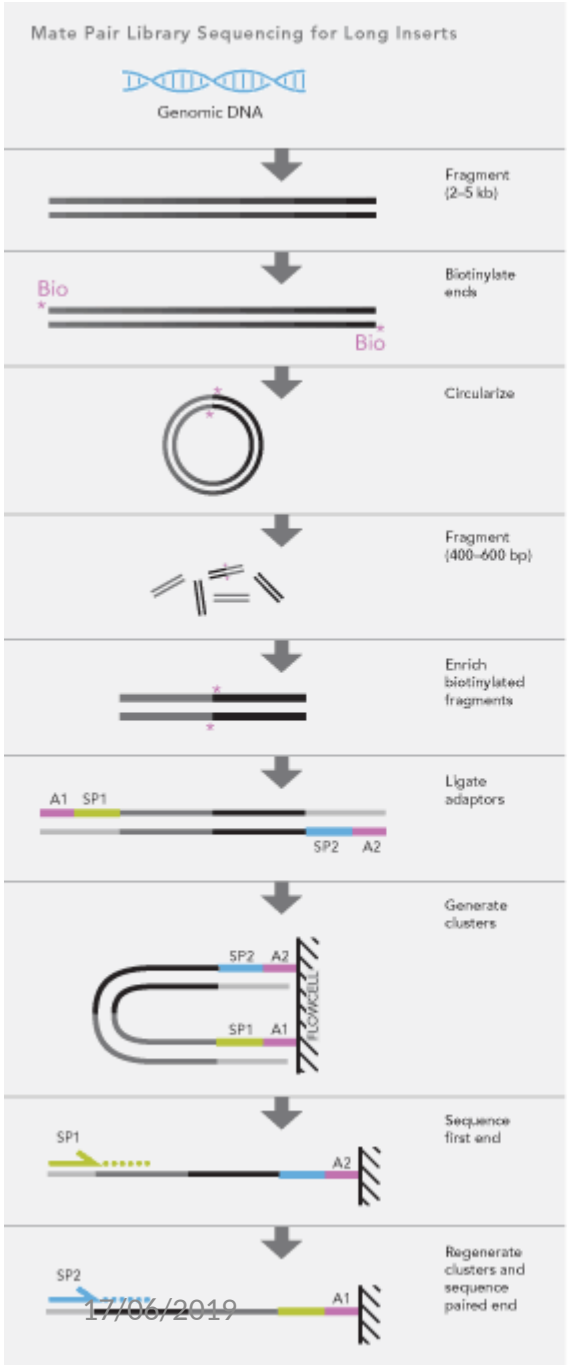
assuming that reads are randomly distributed across the genome.

Que es Pair-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

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/

Resumen

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 Moleculo mayor longitud de lecturas (corrección de errores con Illumina)

Gracias por la atención
Preguntas ???

Isabel Cuesta

Unidad de Bioinformática – Unidades Científico Técnicas - ISCIII

isabel.cuesta@isciii.es