Secuenciación Masiva y análisis de secuencias

Isabel Cuesta

isabel.cuesta@isciii.es

BU-ISCIII

Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

25 Noviembre 2020

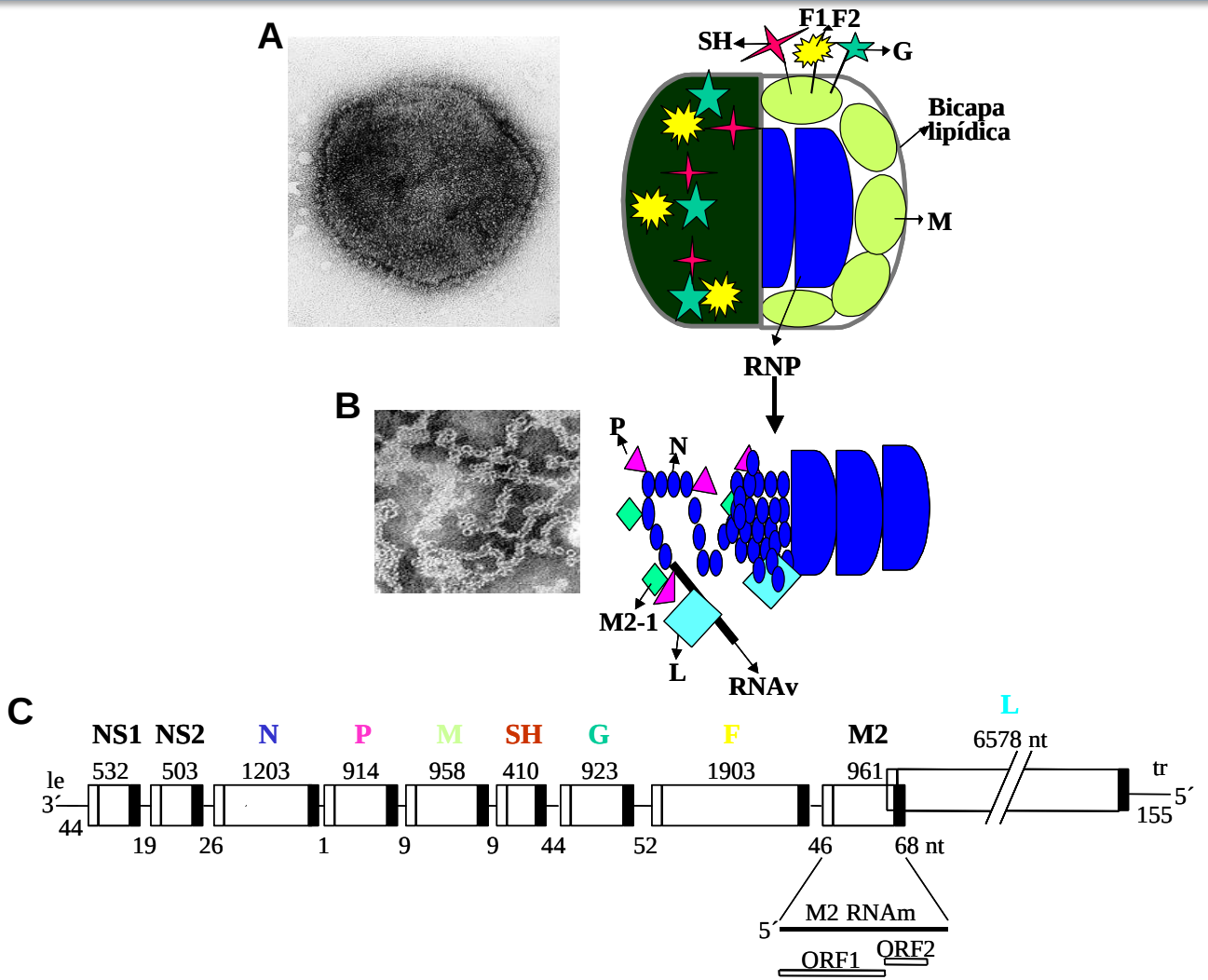
Master Virología

Index

- Secuenciación Masiva (Isabel Cuesta)
- Bioinformática (Isabel Cuesta)
- Conceptos análisis de datos: genoma de SARS-CoV-2 (Sara Monzón)
- Práctica: ViralRecon y Metagenómica (Sarai Varona y Miguel Juliá)

¿Para que secuenciar genomas?

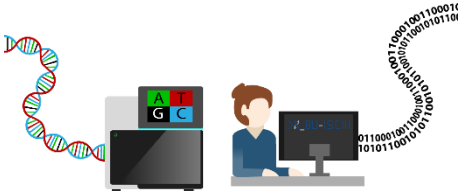
Virus Respiratorio Sincitial Humano

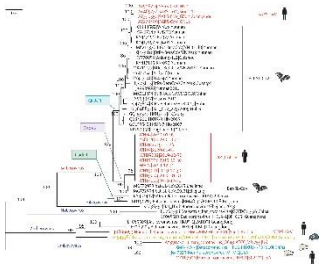


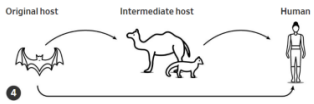

CORONAVIRUS
NOVEL 2019-NCOV
DANGEROUS
DIAGNOSTIC
CHINA
DEATH
ALERT
DIAGNOSIS
COUGH
PATIENT
INFECTION
HEALTH
EPIDEMIC
TREATMENT
PNEUMONIA
FEVER
VIRUS
CORONAVIRIDAE
2019-NCOV
ILLNESS
MEDICAL
DISEASE
HEALTHCARE
WUHAN
ALERT
HYGIENE
CHINESE
NEW
RESPIRATORY
PNEUMONIA
CORONAVIRIDAE
NOVEL CORONAVIRUS

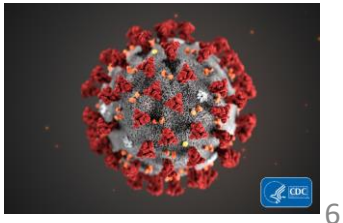
Descubrimiento de un Nuevo virus

Secuenciando su genoma



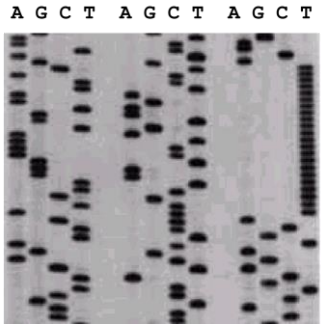


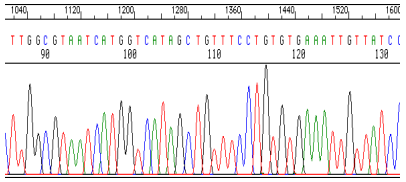


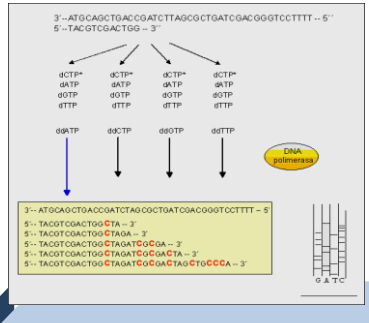


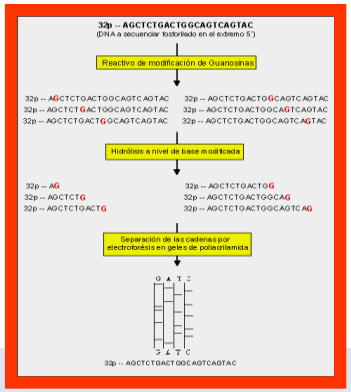
01 02 03 04 05 06

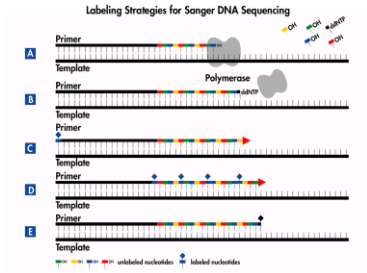
Métodos de secuenciación de DNA











SANGER

SANGER
Secuenciadores Automáticos

MAXAM Y GILBERT

1ª GENERACIÓN

Evolution of DNA Revolution

A walk through the biological history: from Sanger to NGS

1953

1972

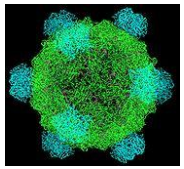
1977

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

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

Development of recombinant genetic engineering

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



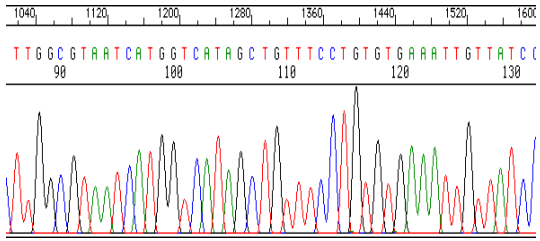
Bacteriofago ΦX174 5386nt plus and minus method

Evolution of DNA Revolution

A walk through the biological history: from Sanger to NGS

1986

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



1990

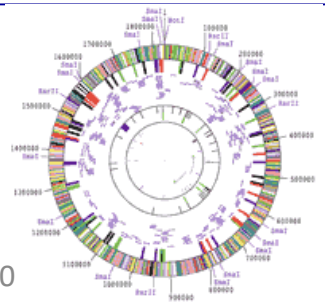
BLAST: Myers & Lipman publish the first algorithm to align DNA sequences (JMB, 215 1990).

Evolution of DNA Revolution

A walk through the biological history: from Sanger to NGS

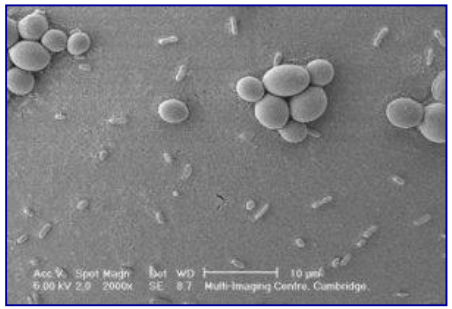
1995

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

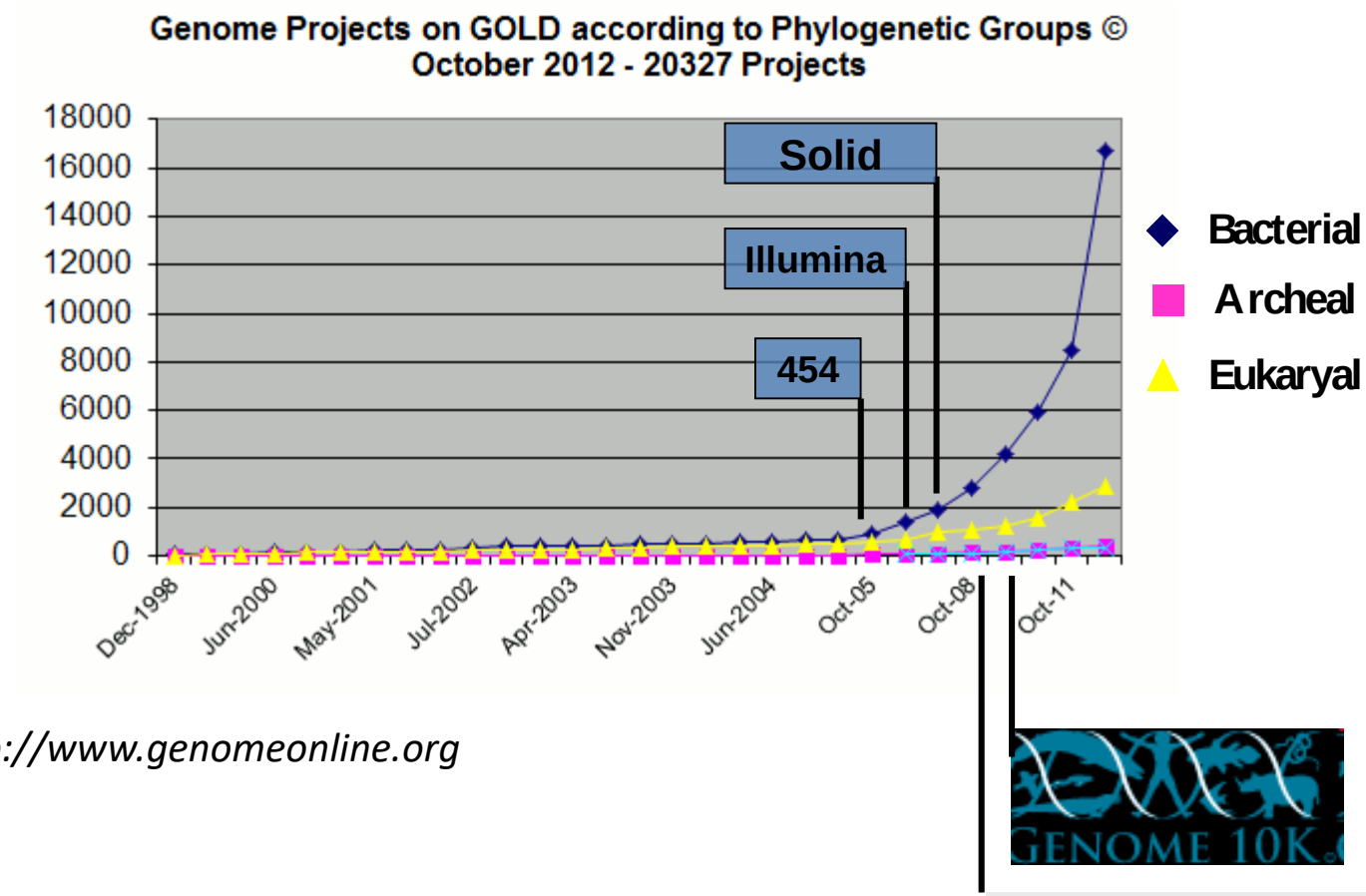
Bacterial Genomes

Bacillus subtilis
Escherichia coli

Eukaryotic Genomes

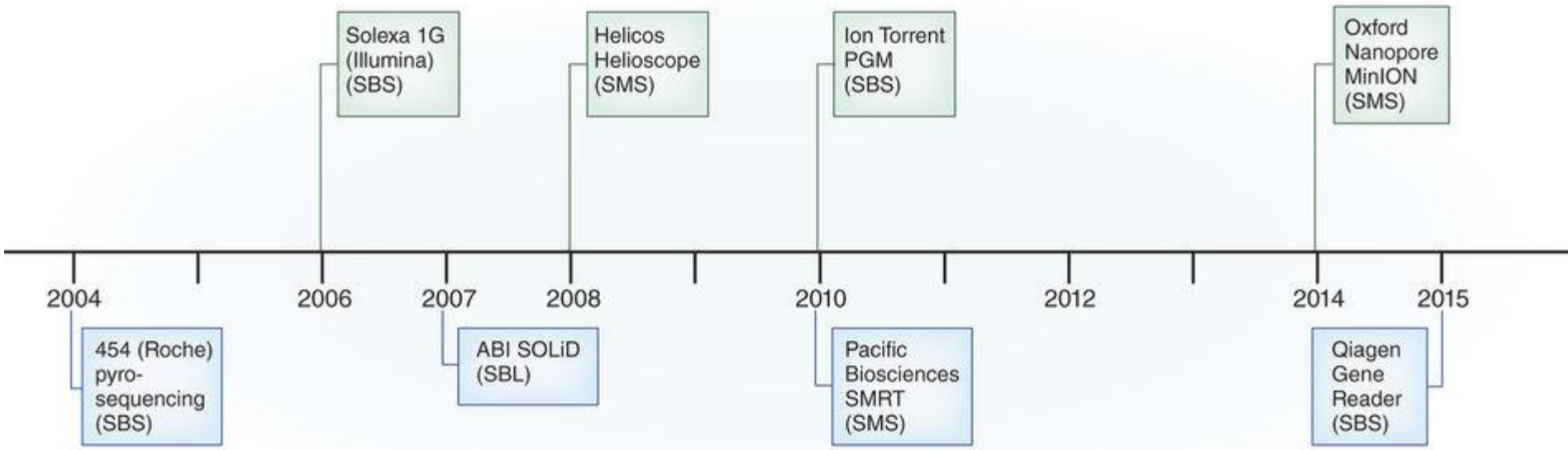
Caenorhabditis elegans
Arabidopsis thaliana
Drosophila melanogaster
Human genome **(1986-2003)**

Genomics Revolution Era



Source: http://www.genomeonline.org

DNA sequencing technologies 2006-2016



Mardis, Nature Protocols 2017

DNA sequencing technologies 2006-2016

Primera generación

- Sanger

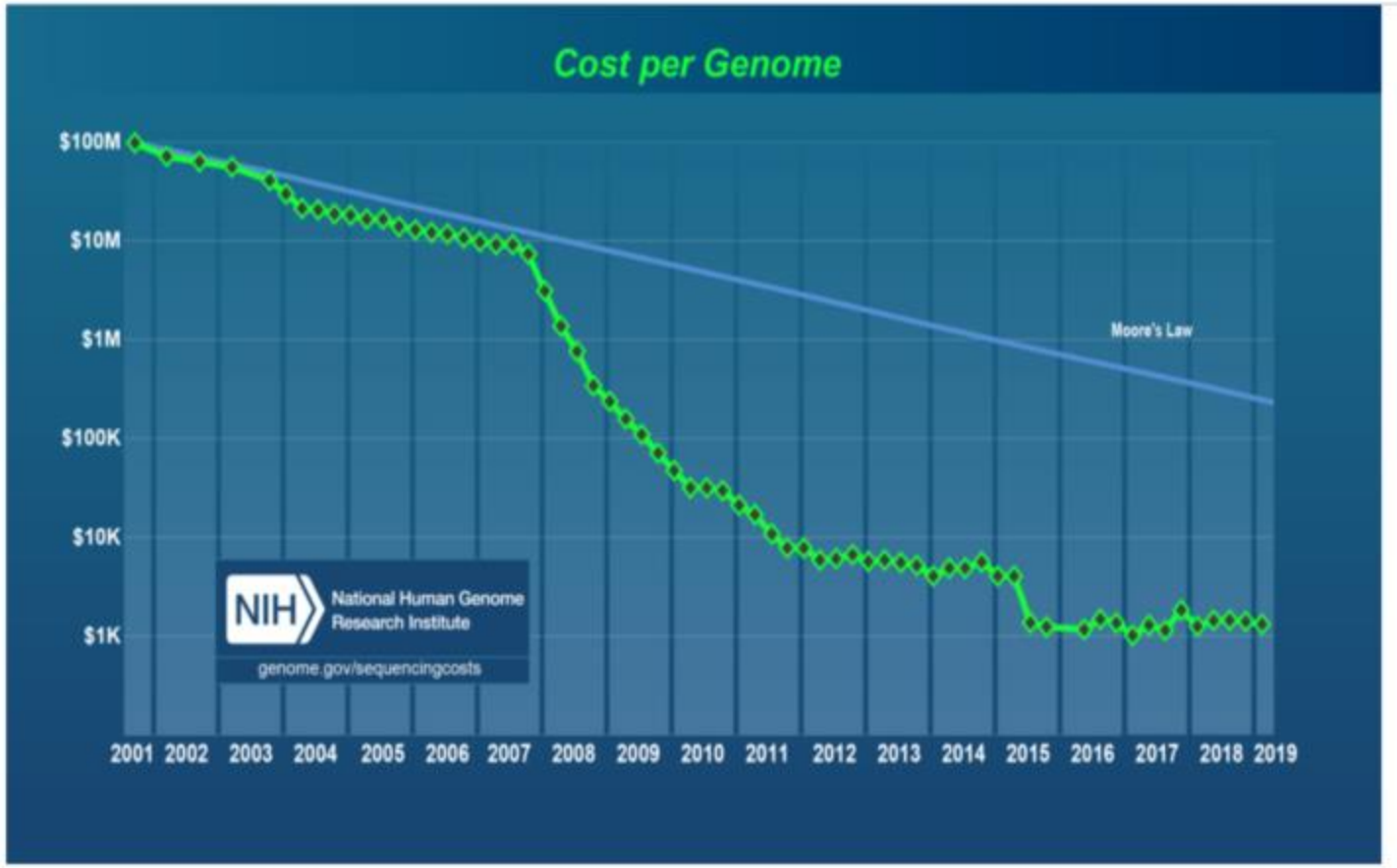
Segunda Generación

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

Tercera Generación

- Pacific Biosciences
- Nanopore

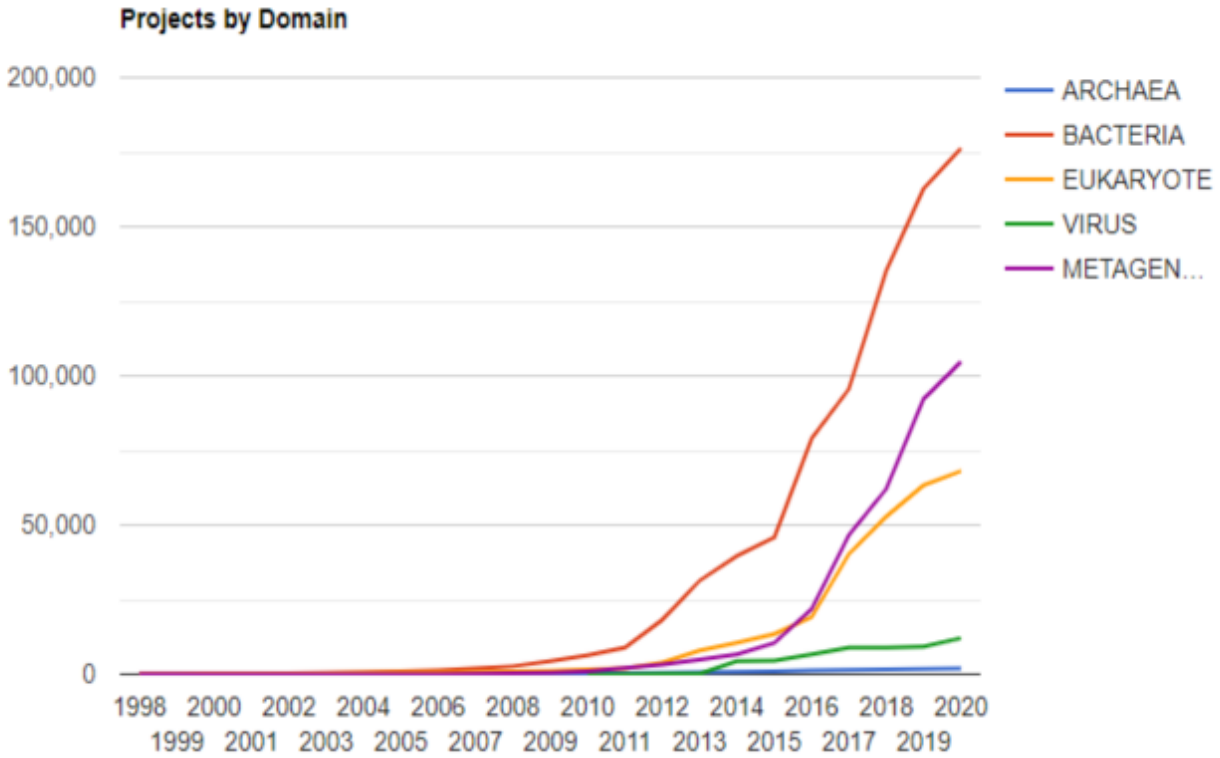
Sequencing cost



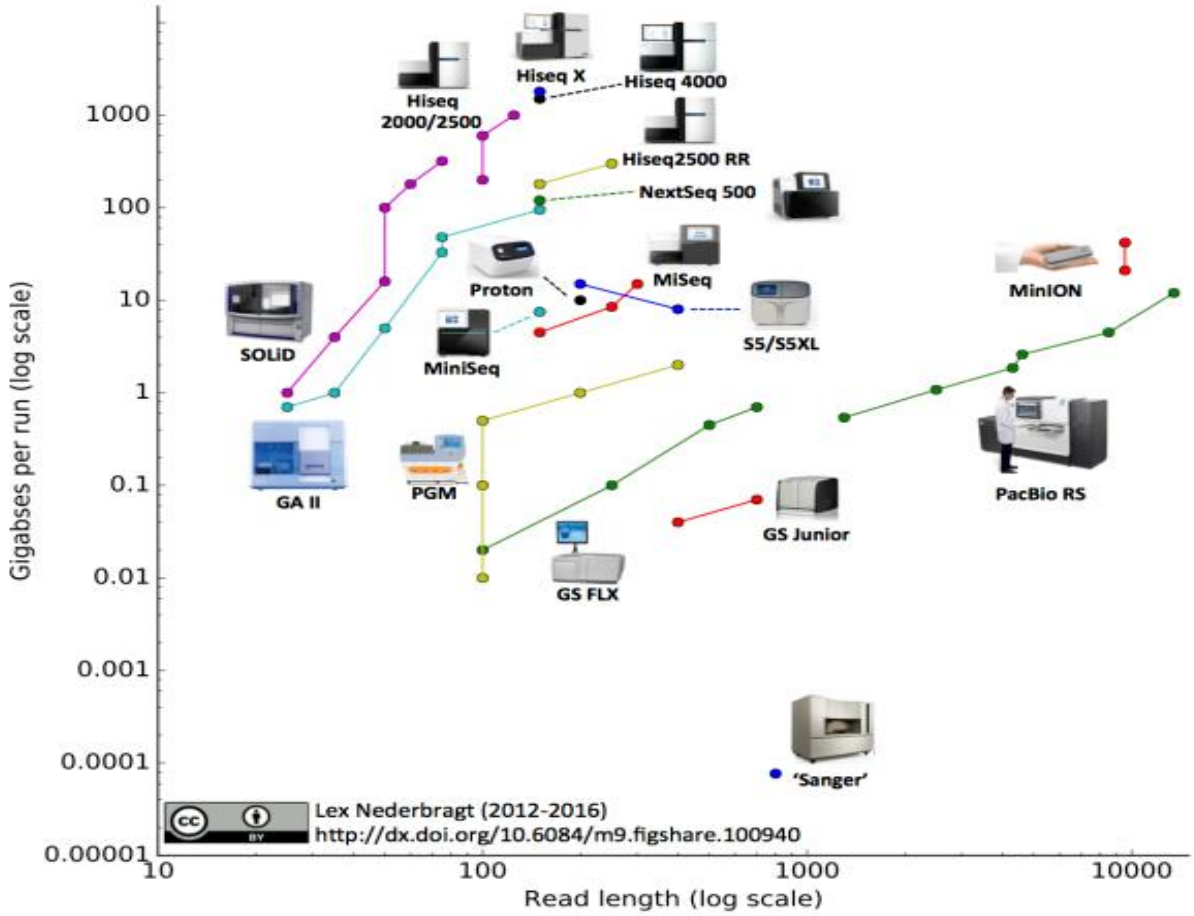
Sequencing projects

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

GOLD, Genome Online DataBase

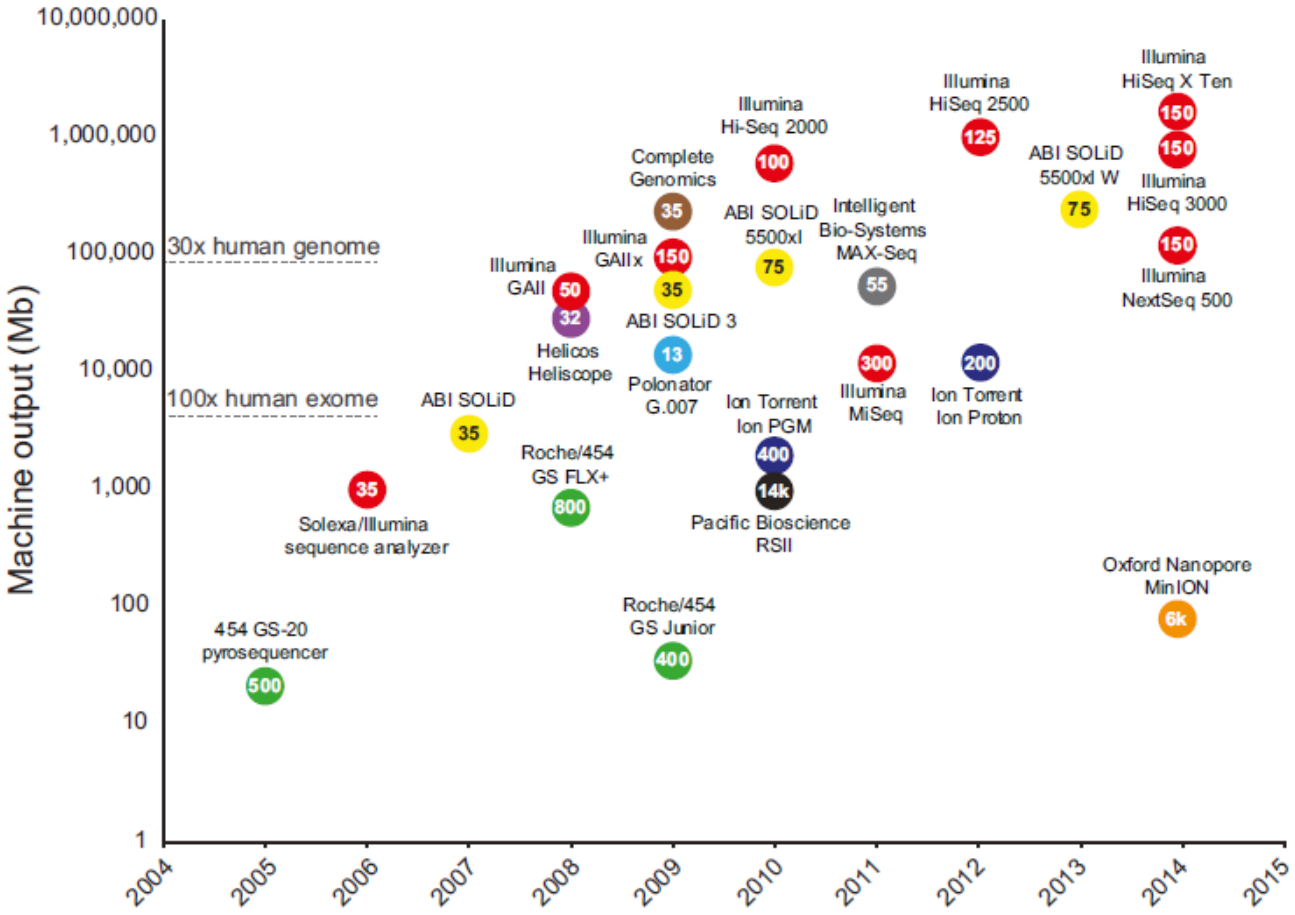


High-Throughput Sequencing Technologies



https://flxlexblog.wordpress.com/

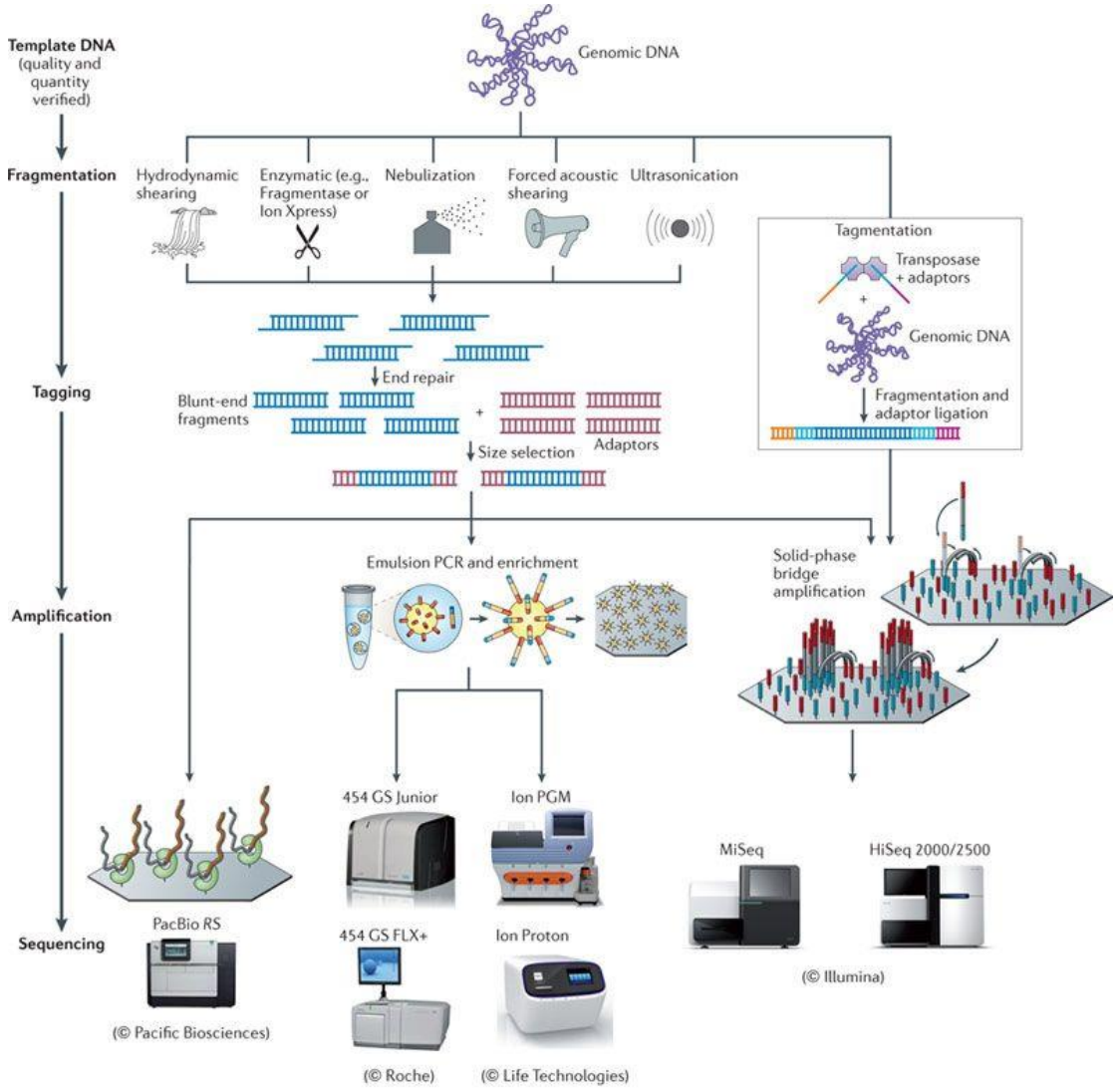
High-Throughput Sequencing Technologies



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

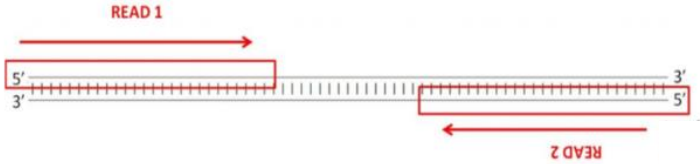
Reuter et al., Mol Cell 2015

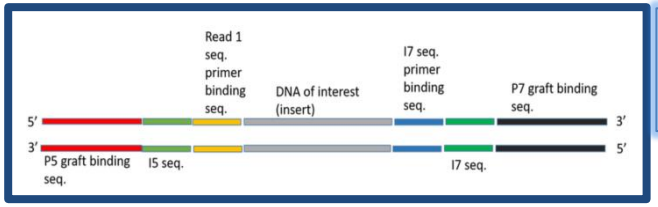
High-throughput sequencing platforms



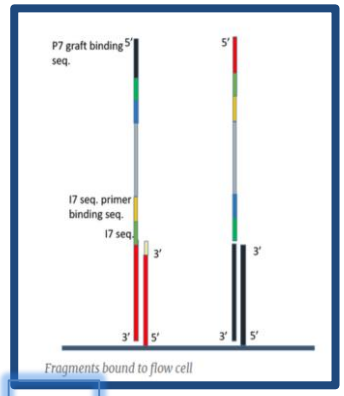
Nature Reviews | Microbiology Loman et al, 2012

Illumina sequencing

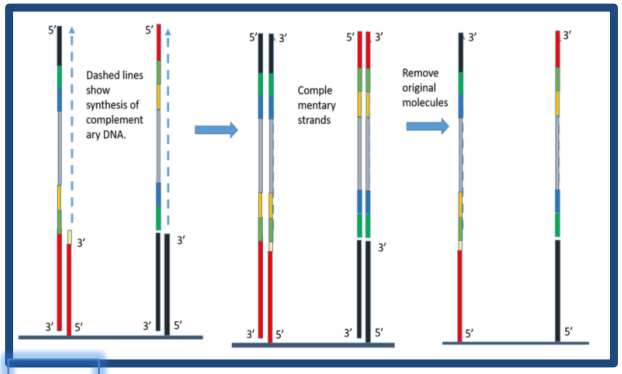




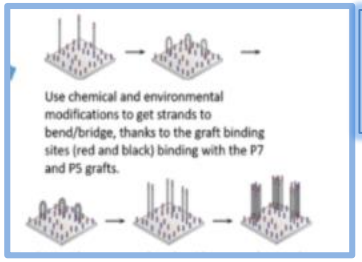
1



2

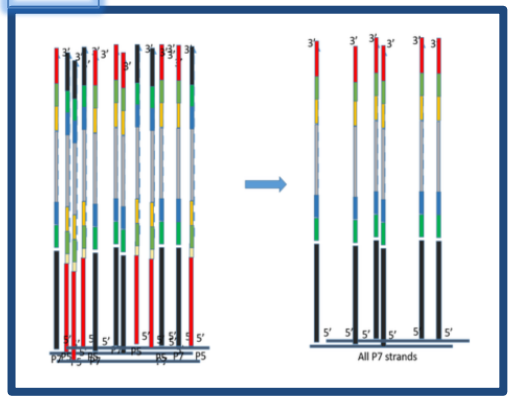


3

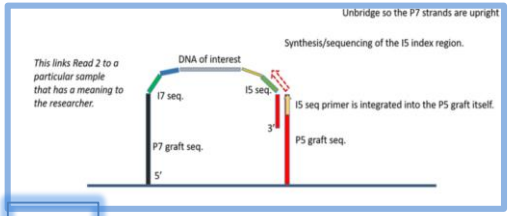


4

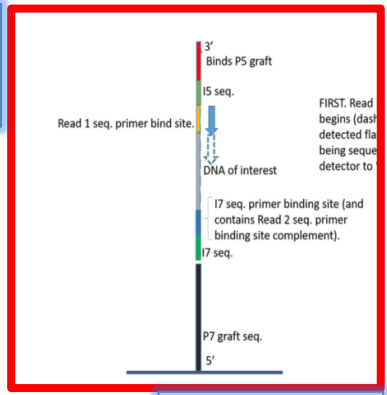
5



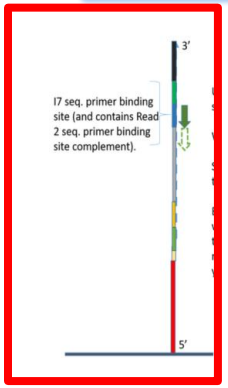
6 R1



7

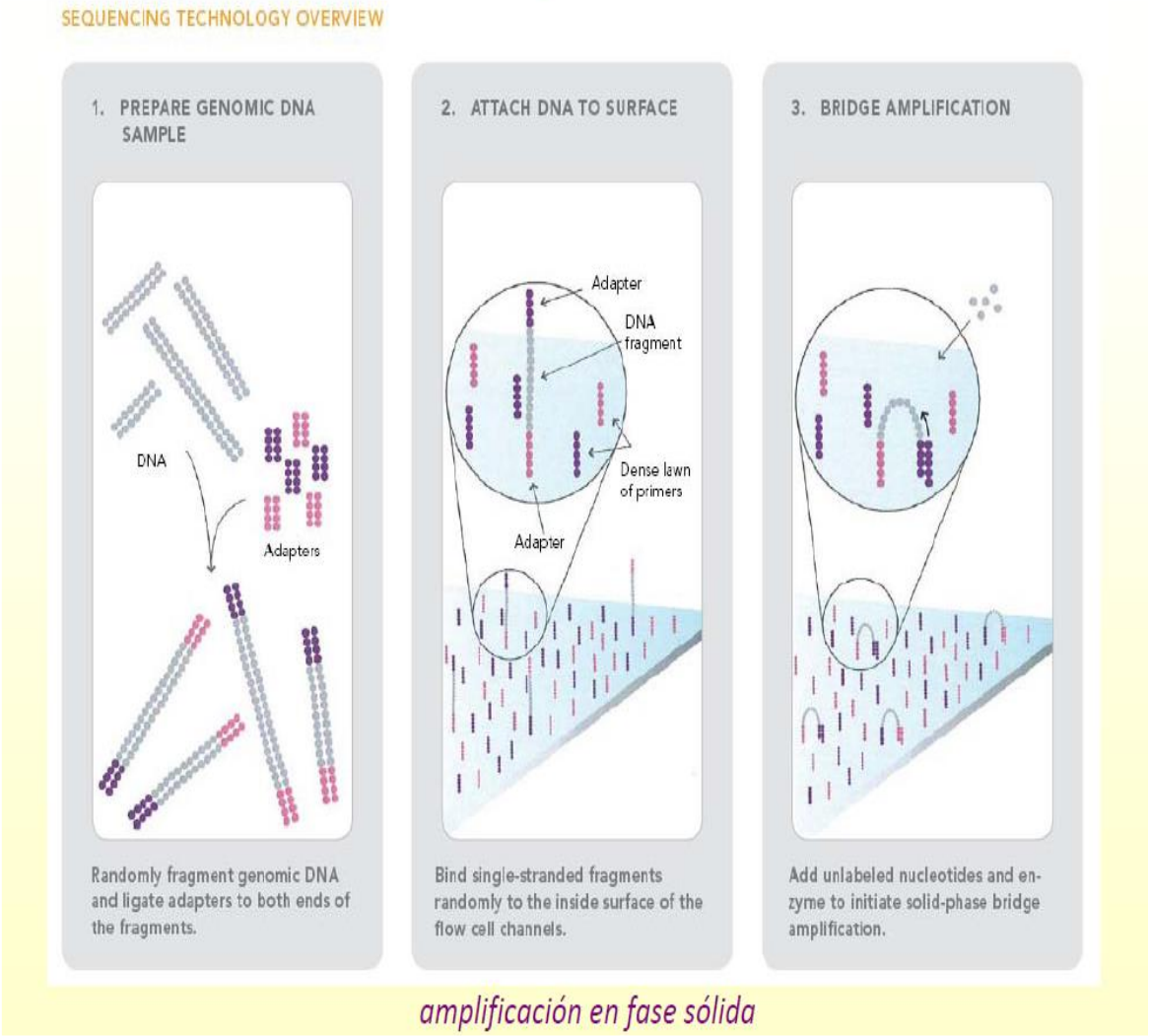


8 R2



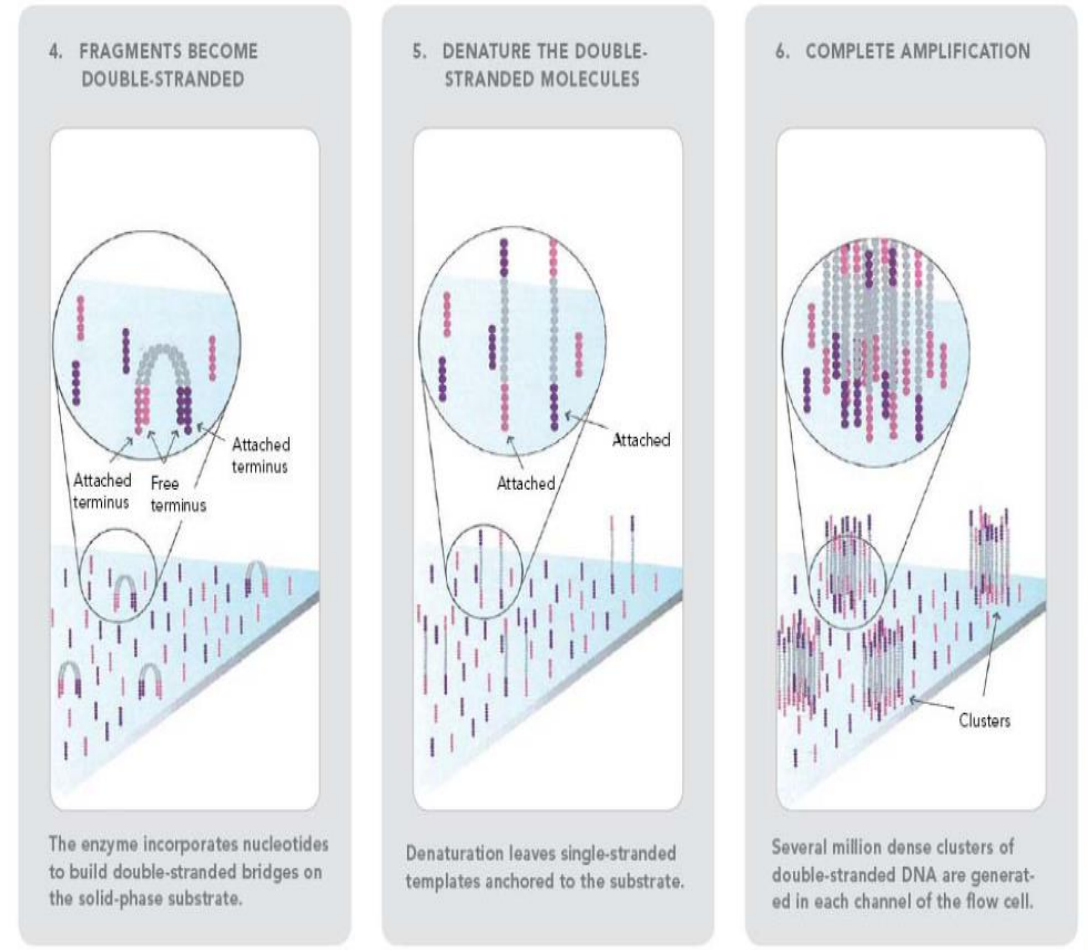
https://kscbioinformatics.wordpress.com/2017/02/13/illumina-sequencing-for-dummies-samples-are-sequenced/

Illumina sequencing



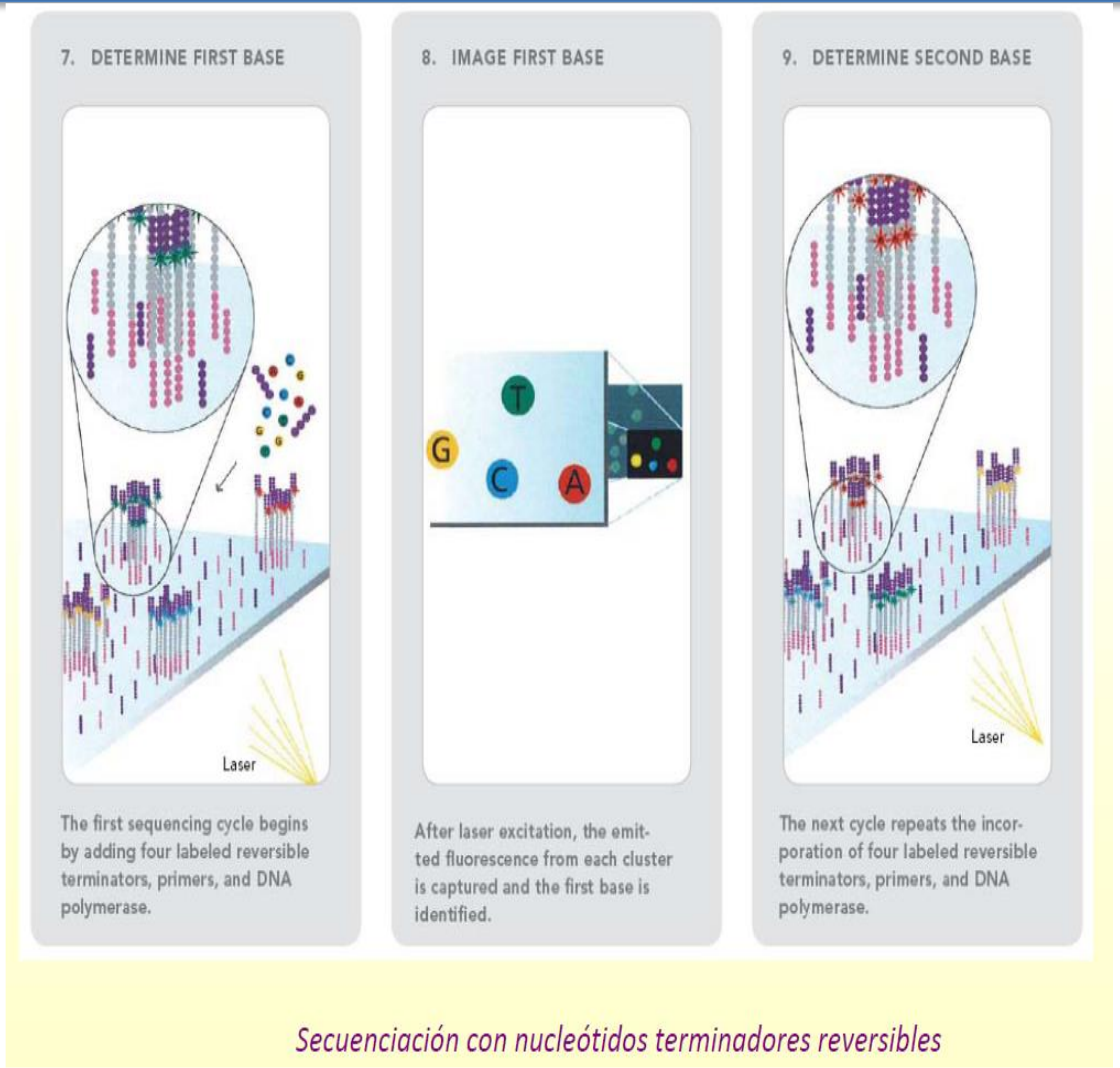
amplificación en fase sólida

Illumina sequencing



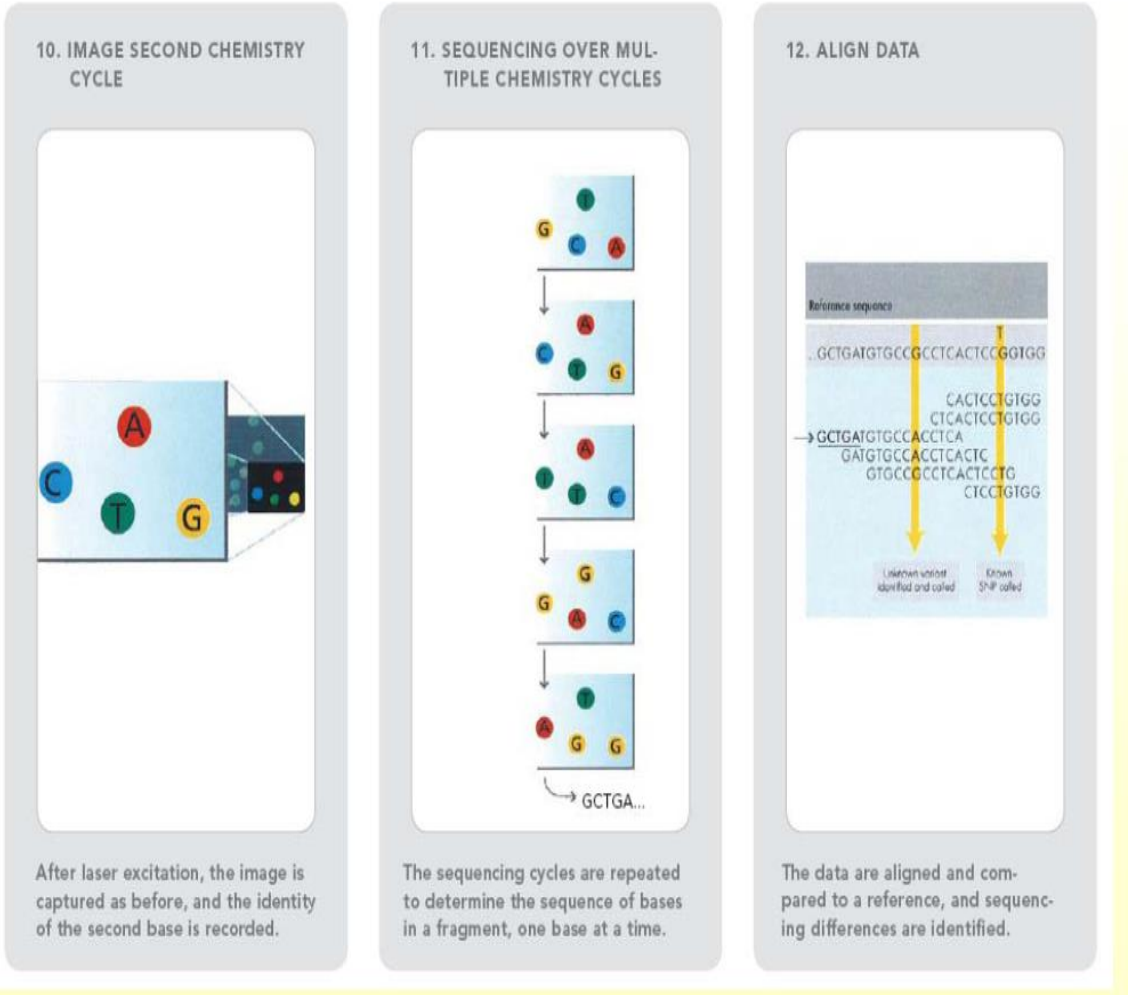
> 1000 copies in ≤ 1 µm; 10^7 clusters per cm^2

Illumina sequencing

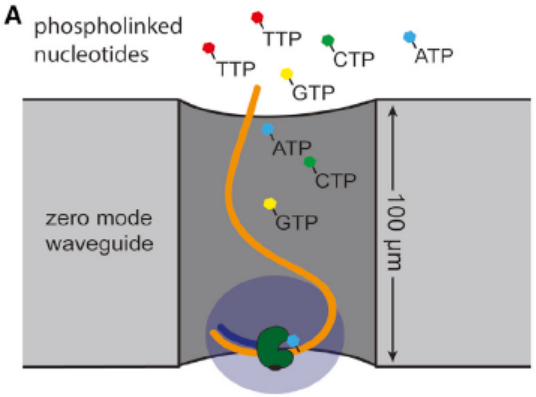


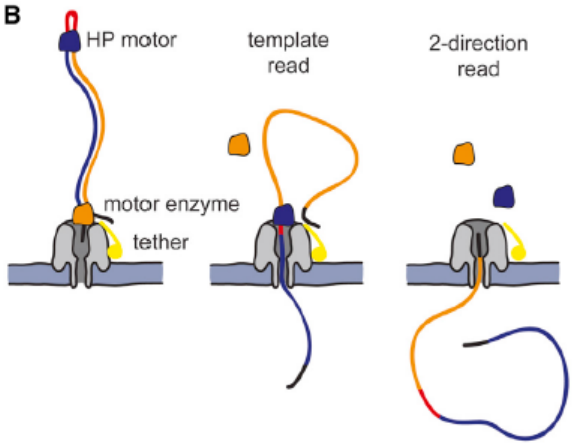
Secuenciación con nucleótidos terminadores reversibles

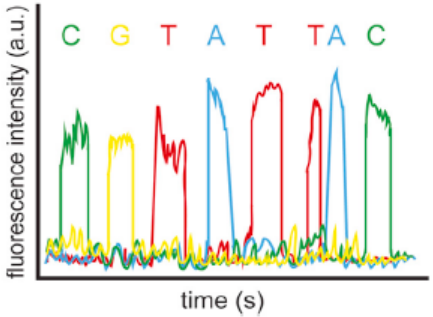
Illumina sequencing



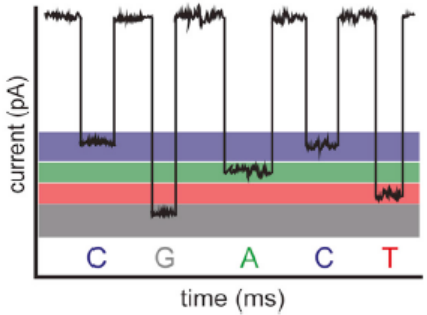
The Third-generation Sequencing Technologies







Pacific Bioscience's SMRT sequencing



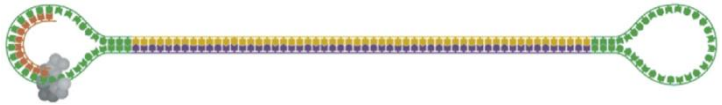
Oxford Nanopore's sequencing strategy

Reuter et al., Mol Cell 2015

Single Molecule Sequencing Platforms

PacBio sequencing and its applications

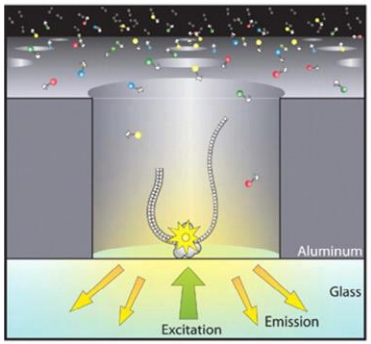
Rhoads & Au, Gen Prot Bioinf 2015

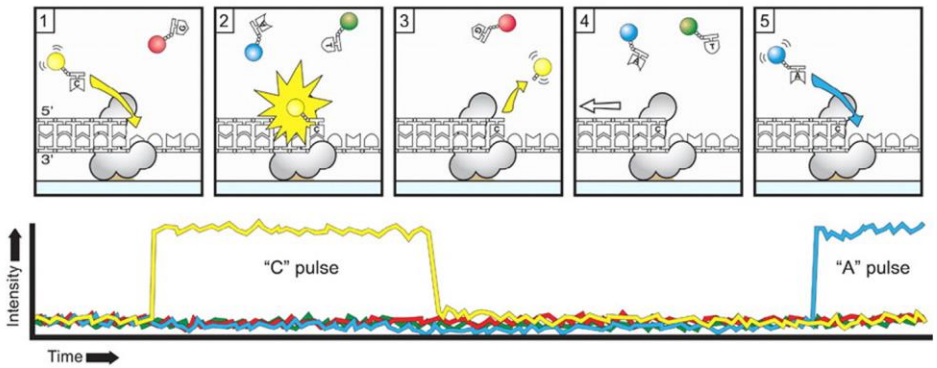


SMRTbell template: is a closed, single-stranded circular DNA that is created by ligating hairpin adaptors to both ends of a target dsDNA

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.

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

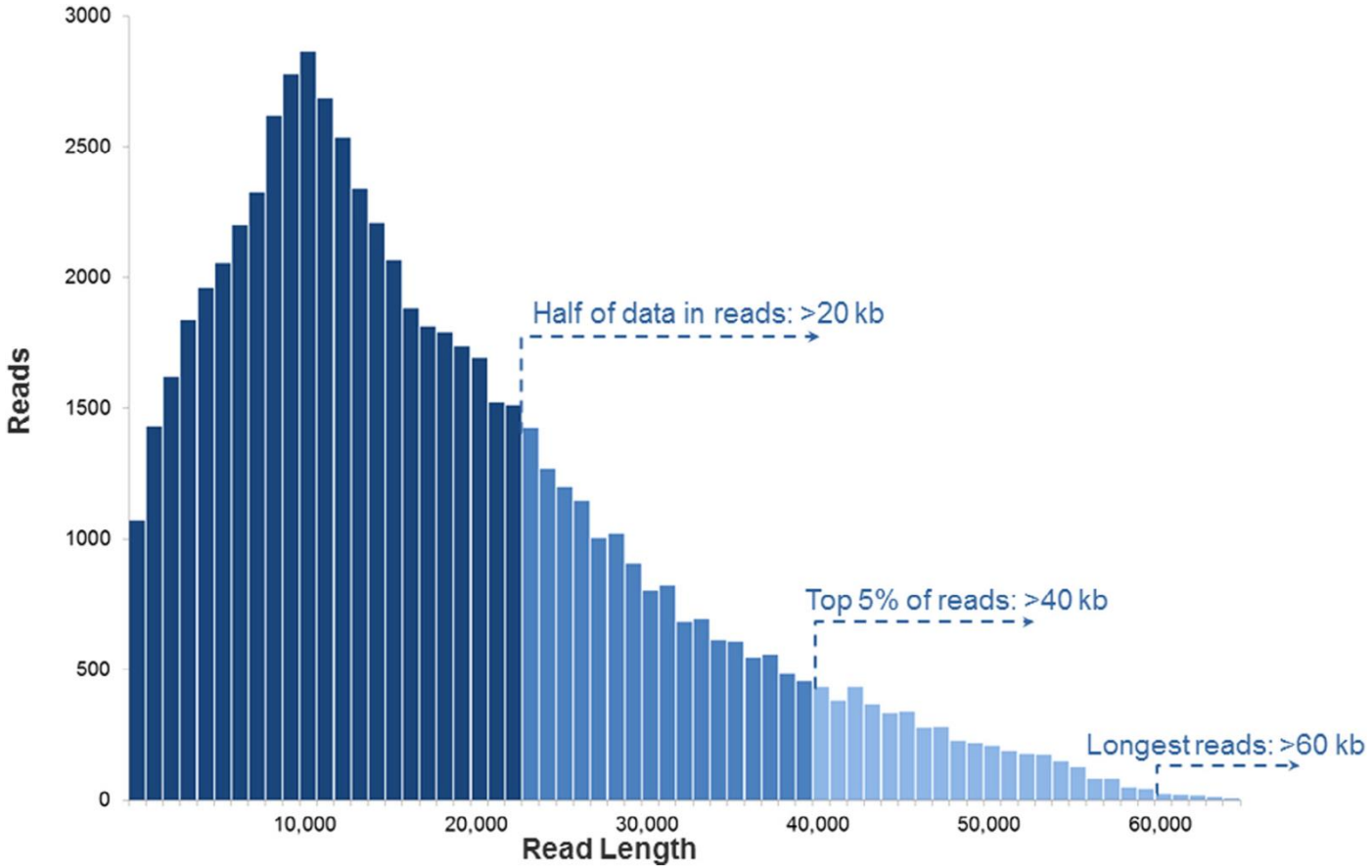




PacBio sequencing and its applications

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

Advantage

Closes gaps and completes genomes due to longer reads

Identifies non-SNP SVs

Achievements

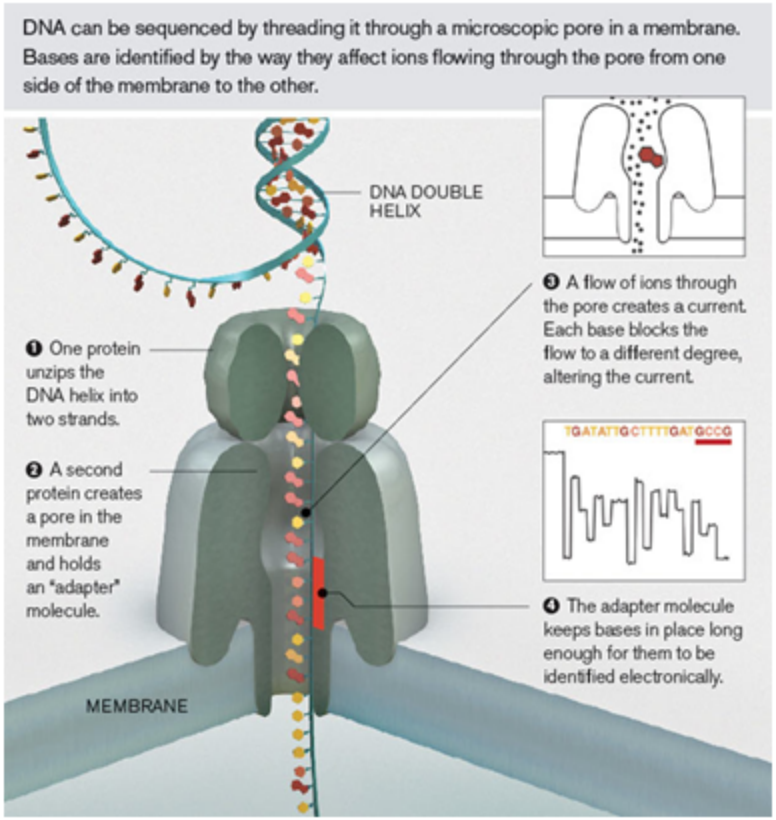
Produced highly-contiguous assemblies of bacterial and eukaryotic genomes

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



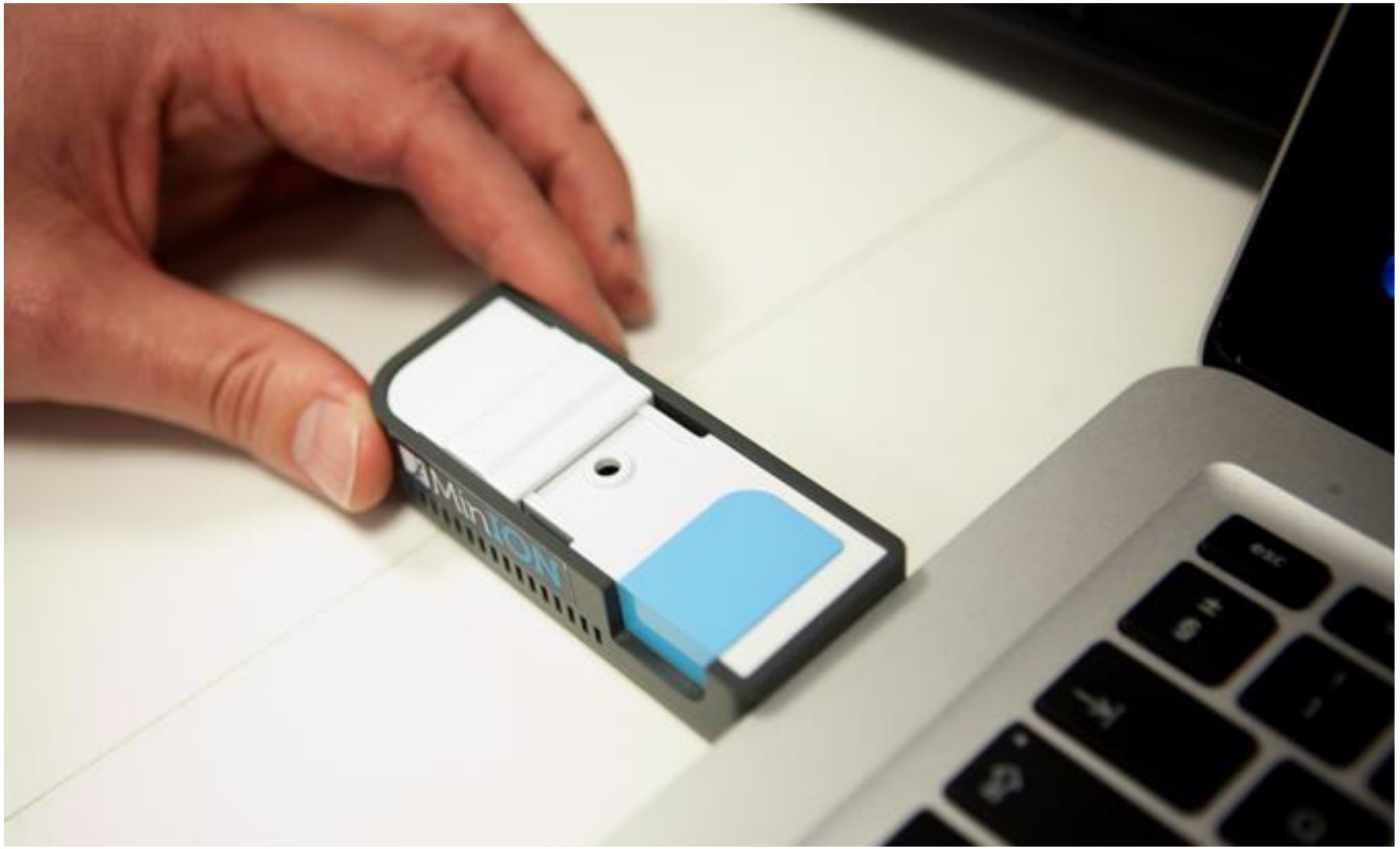
Strand sequencing' is a technique that passes intact DNA polymers through a protein nanopore, sequencing in real time as the DNA translocates the pore.

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

Flongle	MinION	GridION	PromethION
Long read, direct DNA/RNA/epigenetic sequencing, scalable, real time/rapid, on-demand sequencing that is easy to use and install.			
✔ Your portable device for smaller, individual, rapid tests.	✔ Your personal sequencer, putting you in control.	✔ High throughput sequencing, in modular form (up to 5 flow cells) to be on-demand.	✔ Very high throughput sequencing, in modular form (up to 48 flow cells) to be on-demand.
✔ When you don't want to multiplex samples or start a larger run.	✔ Whether in your lab or out in the field.	✔ For your lab or to offer as a service.	✔ For your lab or as a service.
✔ Amplicons, panels/targeted sequencing, quality testing and more.	✔ Whole genomes/exomes, metagenomics, targeted sequencing, whole transcriptome (cDNA), smaller transcriptomes (direct RNA), multiplexing for smaller samples and more.	✔ Larger genomes or projects, whole transcriptomes (direct RNA or cDNA) or where you have larger numbers of samples and more.	✔ Larger genomes or projects, whole transcriptomes (direct RNA or cDNA), very large numbers of samples and more.
✔ For use with MinIT or a laptop.	✔ For use with MinIT or a laptop.	✔ Compute included for real time data analysis and easy installation.	✔ Compute included for real time data analysis and easy installation.

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,

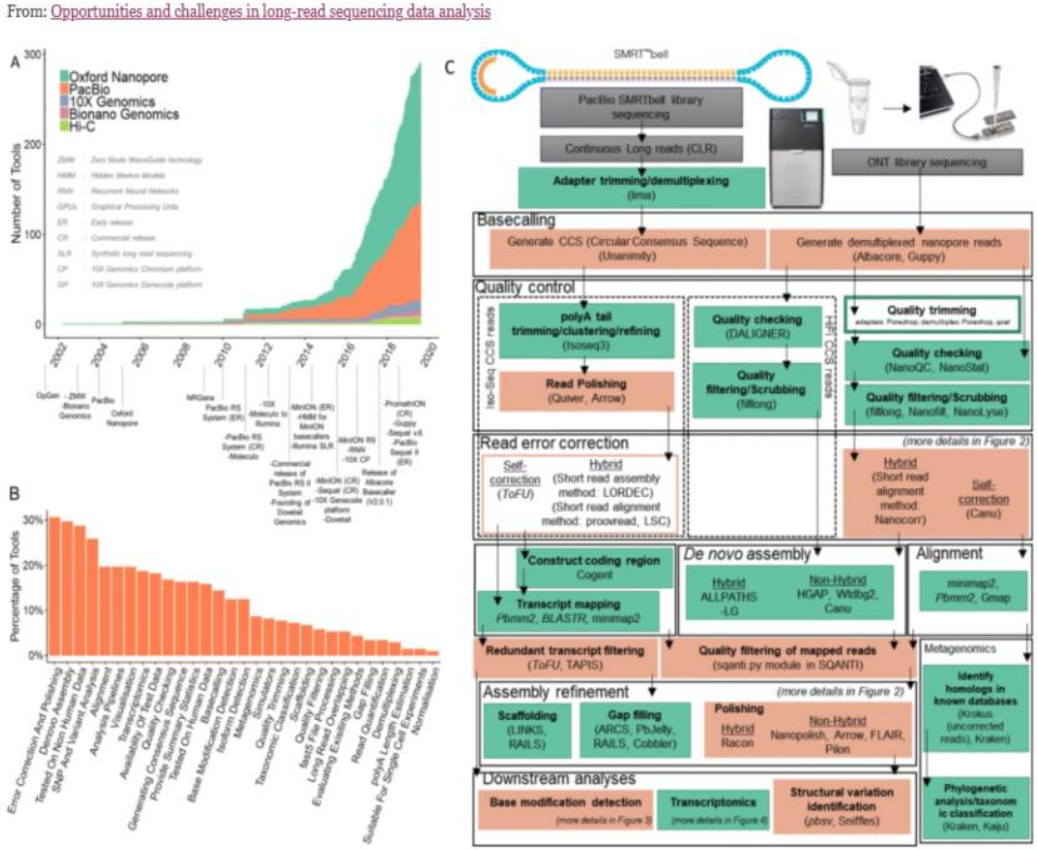
[a] The throughput ranges are determined by available kits and run modes on a per run basis. As an example of a 15-GB throughput, thirty-five 5-MB genomes can be sequenced to a minimum coverage of 40× on the Illumina MiSeq using the v3 600 cycle chemistry.
[b] Per one single-molecule real-time cell.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

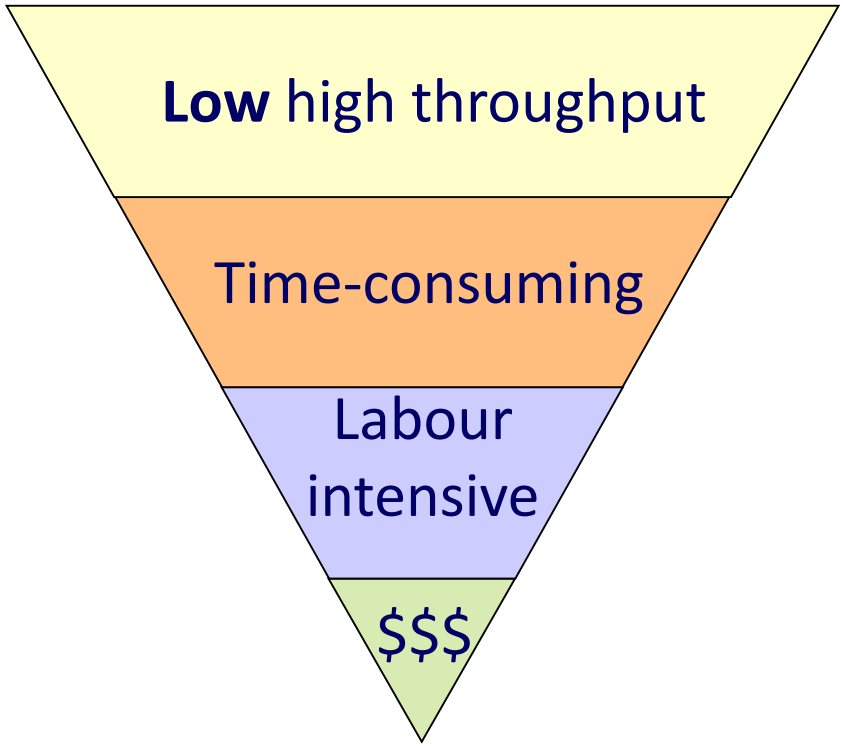
Amarasinghe et al., Genome Bioiogy 2020, 21:30

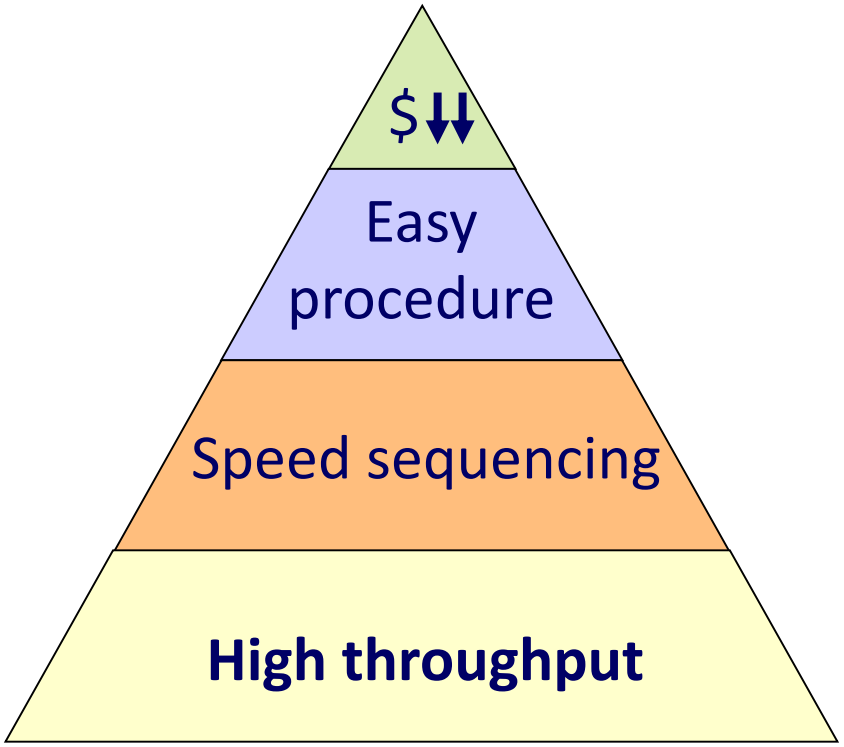
From: Opportunities and challenges in long-read sequencing data analysis



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

Sanger vs SM, advantages of new technologies

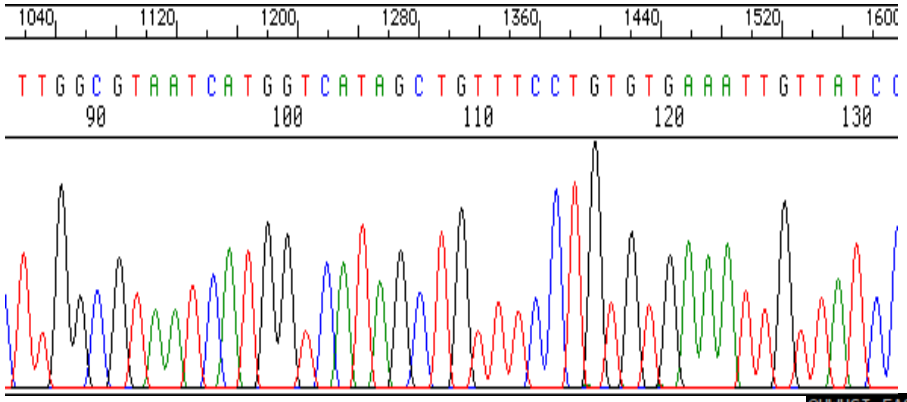




Semiautomatic **Sanger** capillary-based sequencing technology

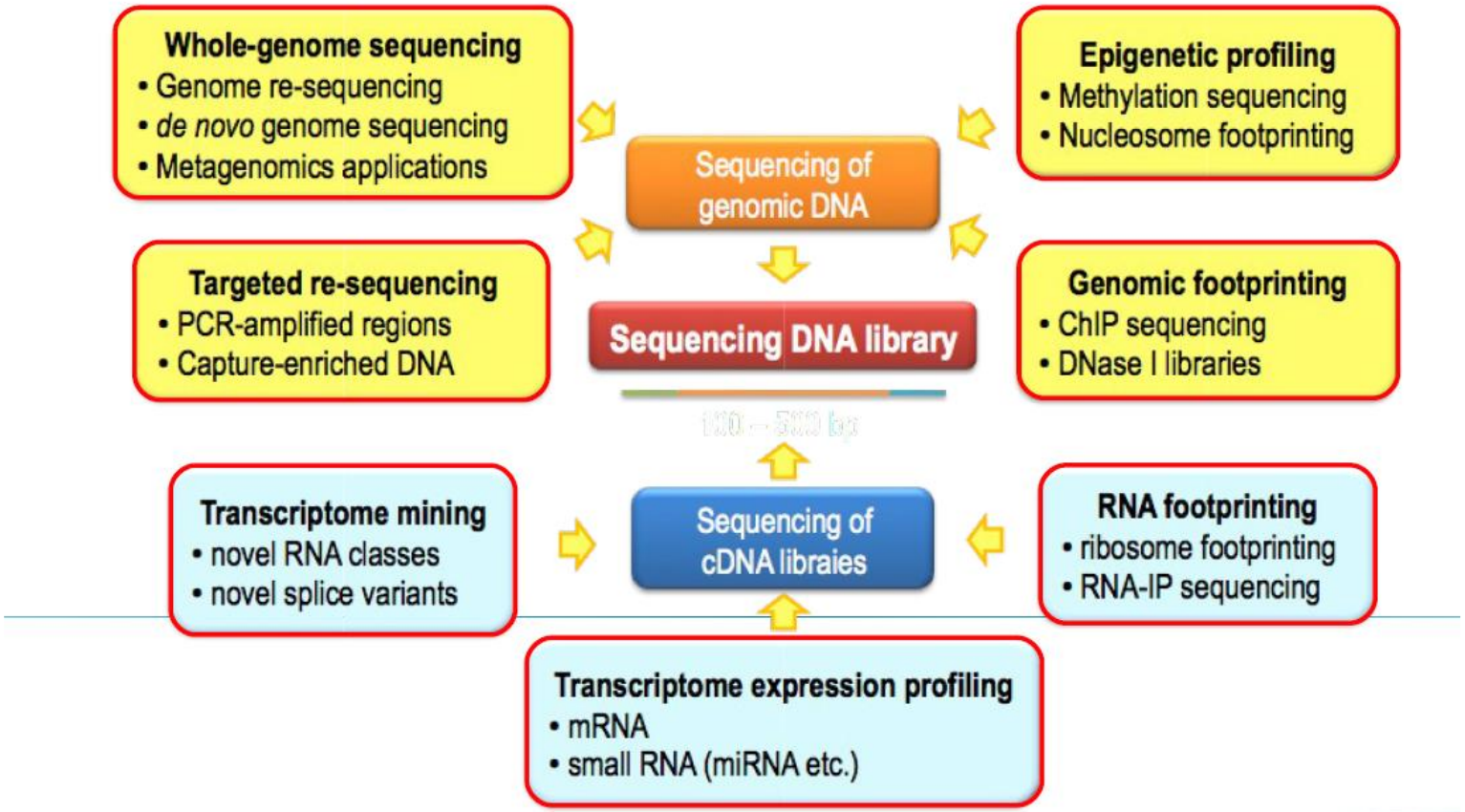
NGS
Next Generation Sequencing =
Now Generation Sequencing

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>HHIHIFGDIIIIIGFHIEGH
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3400:1035 1:N:0:ACAGTG
TCCTGCTTTAGGAGANTCCTCATGCTCTGACAGGATGCTCTCTATGTGAGTTGAGCTGGTCTTCTCACTTTTATAG
+
IIIIIHIHIIGGEGG#AACA@?=?BHHIIIIIHHIHIIIIHIHIHHGIHIHGHGIGIHGEGGGGHG@EFGGCEFAB
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3962:1033 1:N:0:ACAGTG
CCACCAACACAGTCTNCACCTTCTGTTGCTGGTGATAGATTTTTGCACCTTTCCATCCTCCAGGTTTCAAAATAGC
+
HHFHHDHDHH>C?CA#EEEE>?A?>HHDGHEGBGBCEEEEGHHF8HEHEEHECH,=>>==EAEE>BEBBAEAACAB
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:4491:1028 1:N:0:ACAGTG
AGAGAGAGAGAGAGANAGAGGACTCTGGAGATGCCGAAGCACAAGCCTGCAAGAGTCCCAGCAAAGAAAATAAAAA
+
GADGGEGGEGBBB?B#@=@@72;64GGGFGB>GGGBDG<DBGB<DA??/?#########################
```

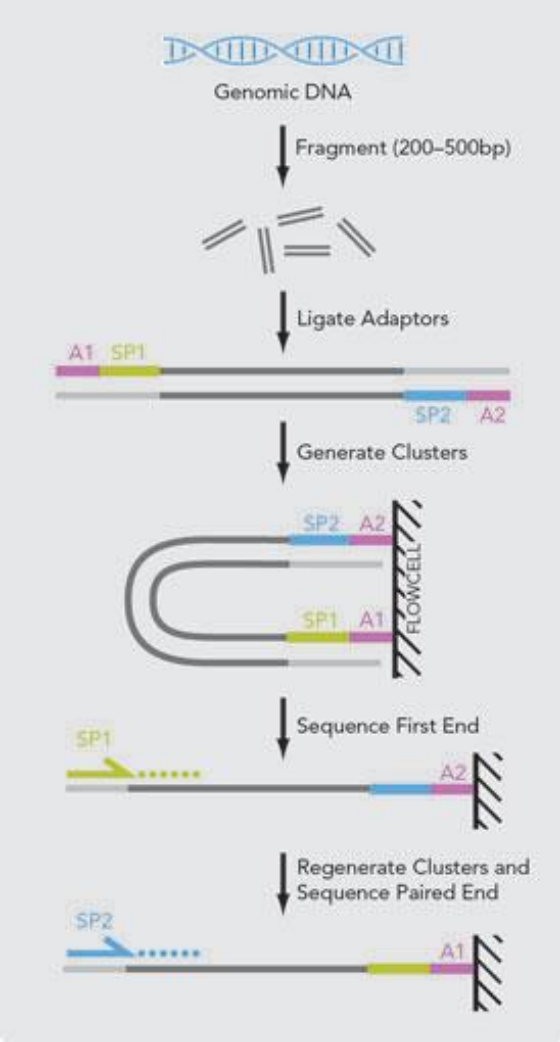
Massive sequencing applications



APLICACIONES DE NGS EN EL DIAGNÓSTICO VIROLÓGICO

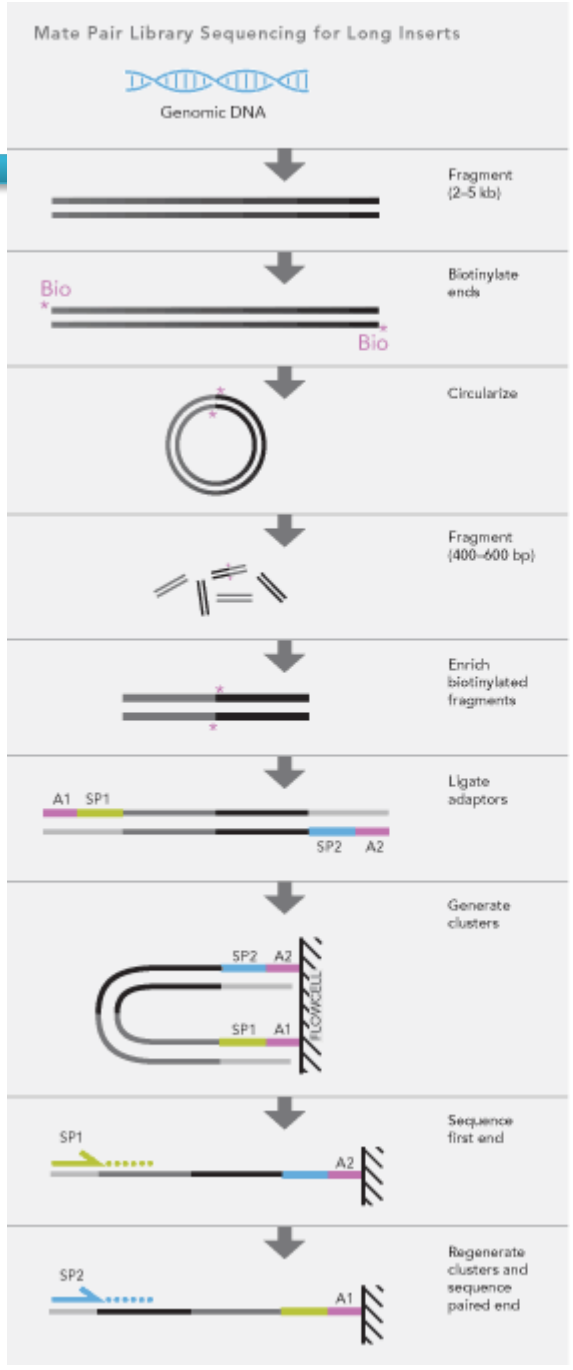
- Detección de patógenos virales no identificados en la muestra
- Identificación de virus nuevos
- Identificación de virus en muestras tumorales
- Caracterización del Viroma de un organismo
- Secuenciación del genoma viral completo
- Estudio de la Variabilidad genómica (Quasispecies)
- Monitorización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

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

Sequencing terms

Breadth of coverage

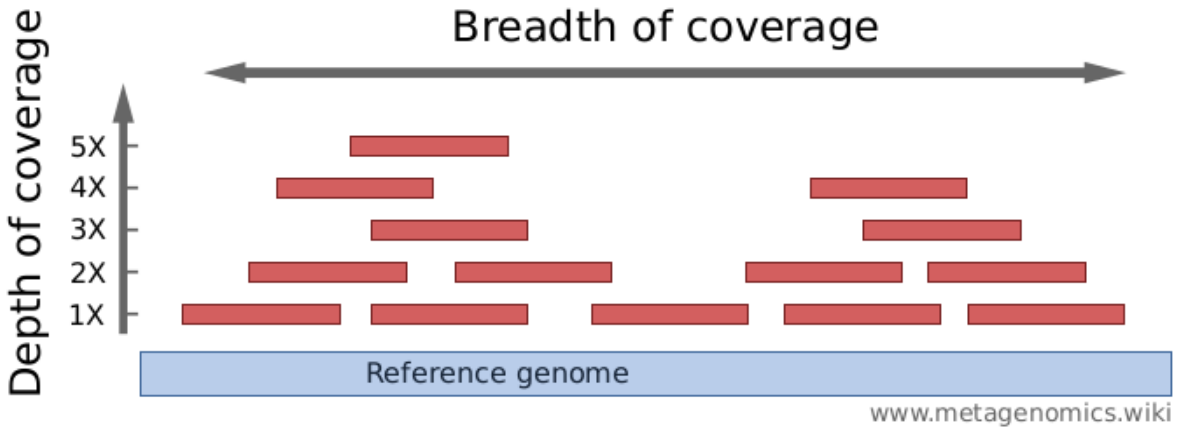
How much of a genome is "covered" by short reads? Are there regions that are not covered, even not by a single read?

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

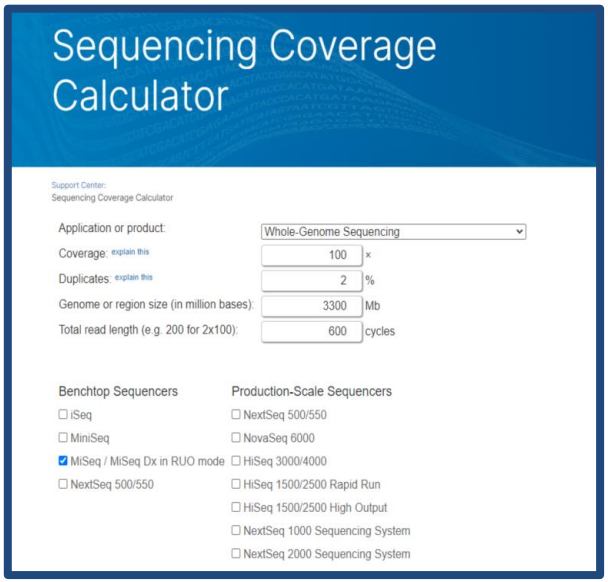
How strong is a genome "covered" by sequenced fragments (short reads)?

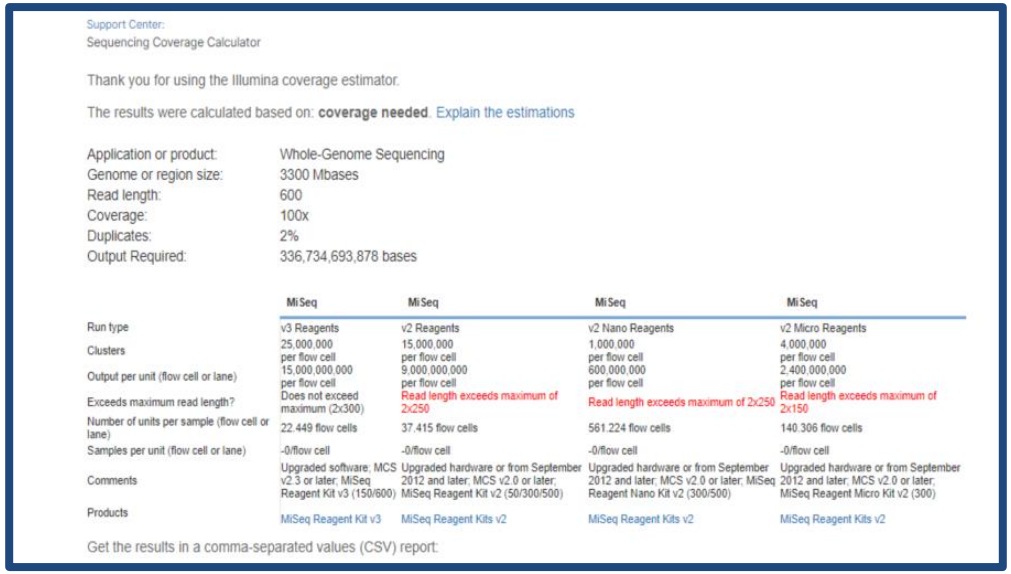
Per-base coverage is the average number of times a base of a genome is sequenced. The coverage depth of a genome is calculated as the number of bases of all short reads that match a genome divided by the length of this genome. It is often expressed as 1X, 2X, 3X,... (1, 2, or, 3 times coverage).



Calculo de cobertura: número de lecturas

Total output required = region size * coverage / ((1-duplicates/100) * on target/100)





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

PREPARACIÓN LIBRERÍA, estrategias

SECUENCIACIÓN GENOMA, EXOMA, TRANSCRIPTOMA

1. Sin amplificación
2. Amplificación con PCR
3. Sondas captura

- Tamaño de fragmento
- Longitud de la lectura
- Single o Pair-end
- Número de bases por muestra
- Profundidad de cobertura x

SECUENCIACIÓN GENOMAS

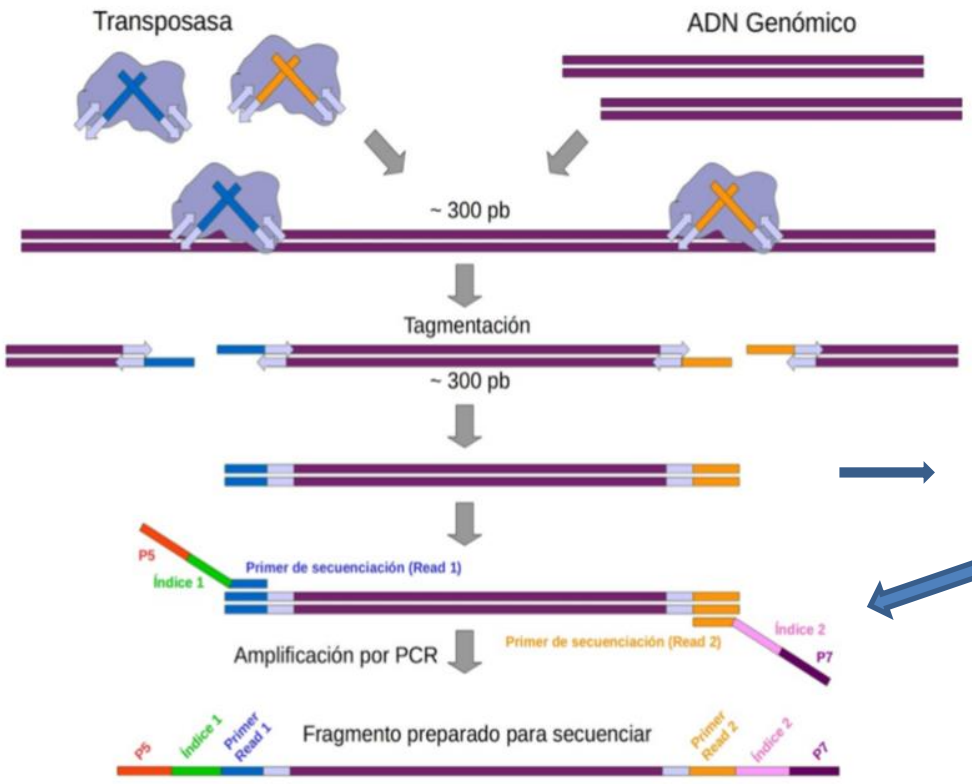
1. Metagenómica

IDENTIFICACIÓN MICROORGANISMOS

1. Metataxonomía

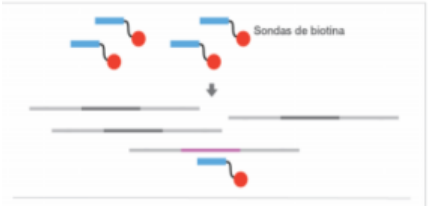
PREPARACIÓN LIBRERÍA

ENZIMÁTICA FÍSICA



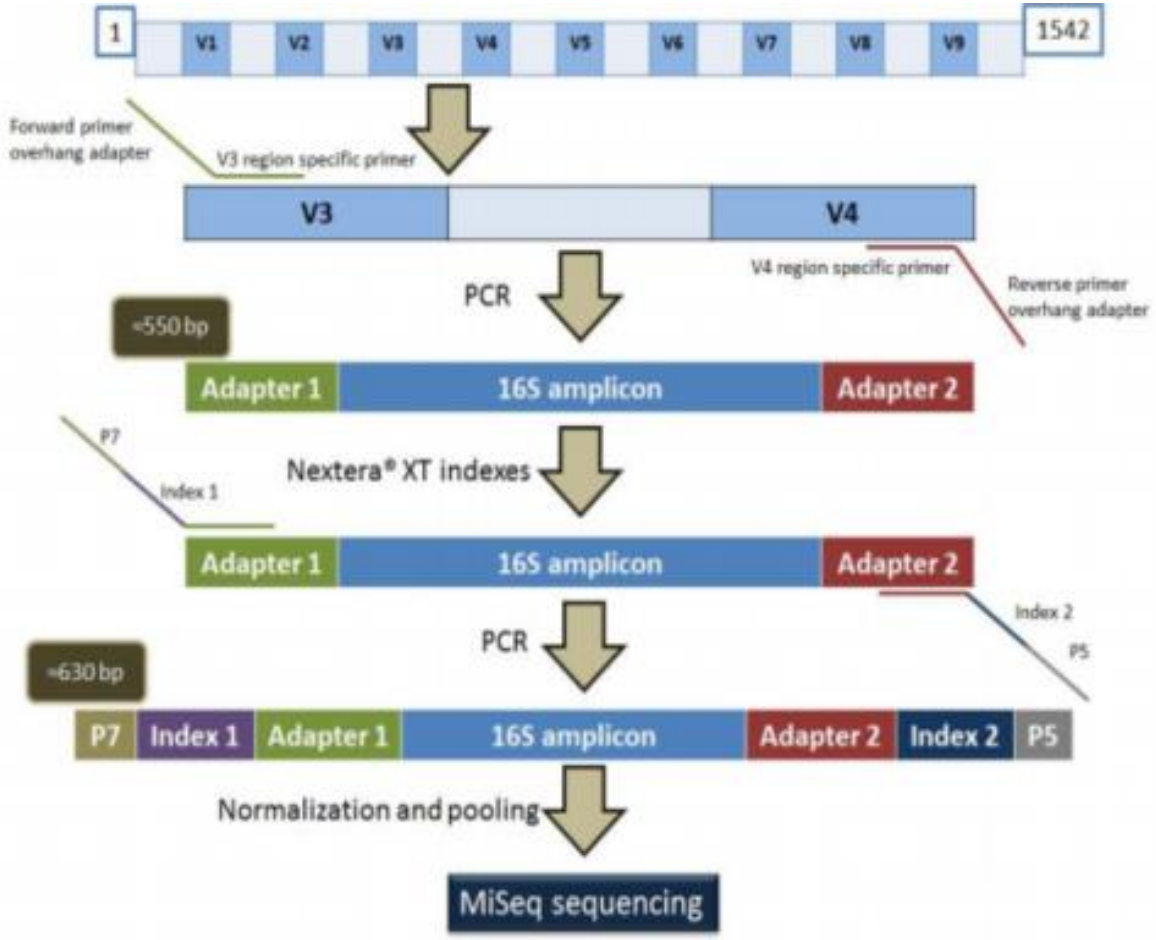
RNA -> cDNA

ENRIQUECIMIENTO:
PCR
CAPTURA SONDAS



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

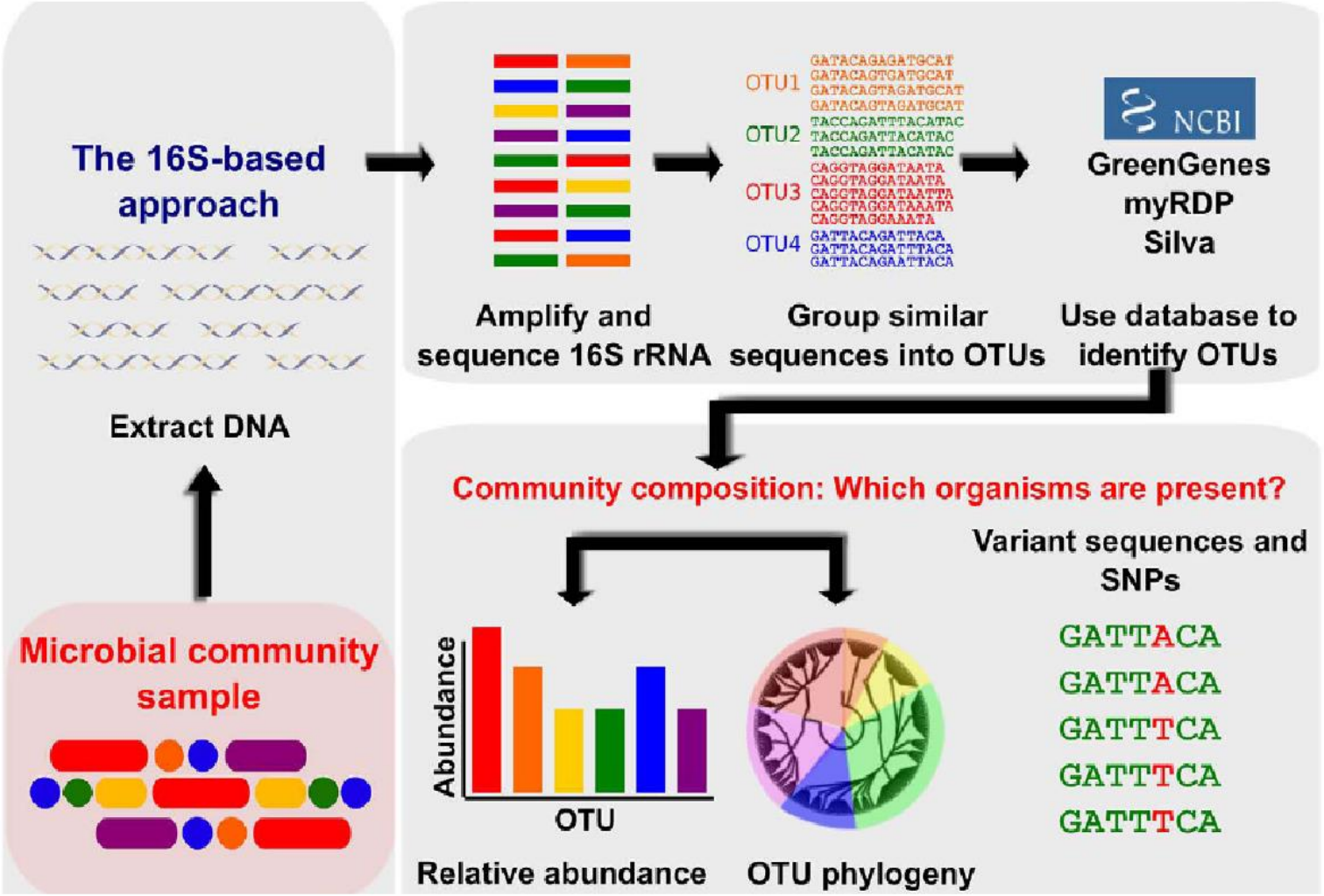
PREPARACIÓN LIBRERÍA, rRNA 16S, caracterización microbiota



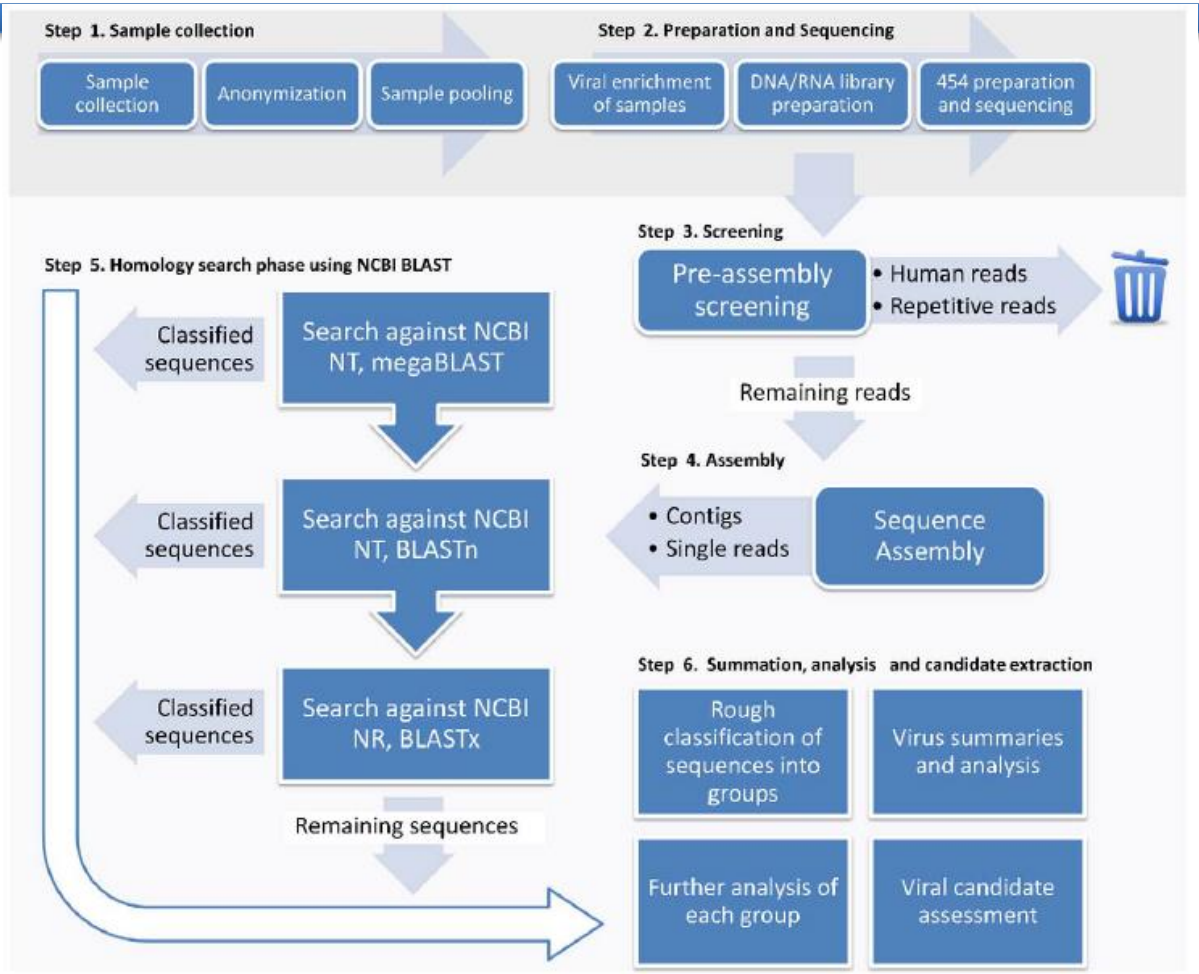
Metataxonomics vs Metagenomics (16S vs Shotgun)

	Metagenetics	**Metagenomics**
Amplified sequence	Marker regions	Whole genome
Computing time	Usually short	Usually long
Taxonomic composition	Yes	Yes
New pathogen detection	No	Yes
Genome coverage information	No	Yes

Metataxonomics



Metagenómica, pipeline de análisis



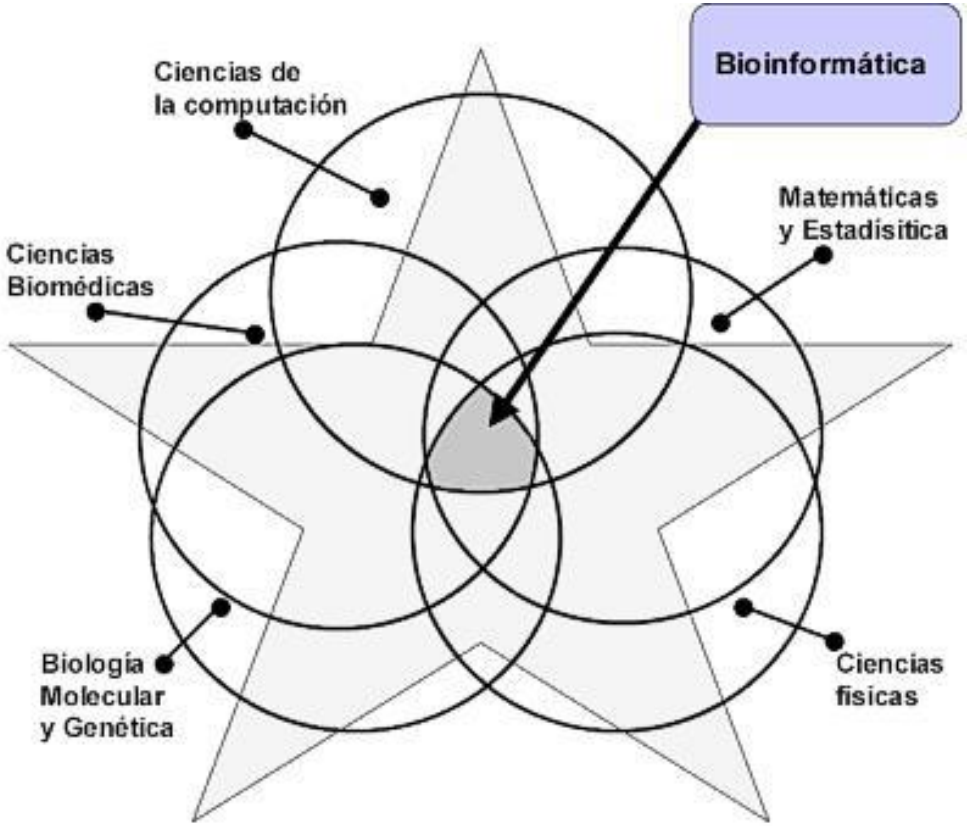
Lysholm et al., Plos One 2012:7,2, e30875

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

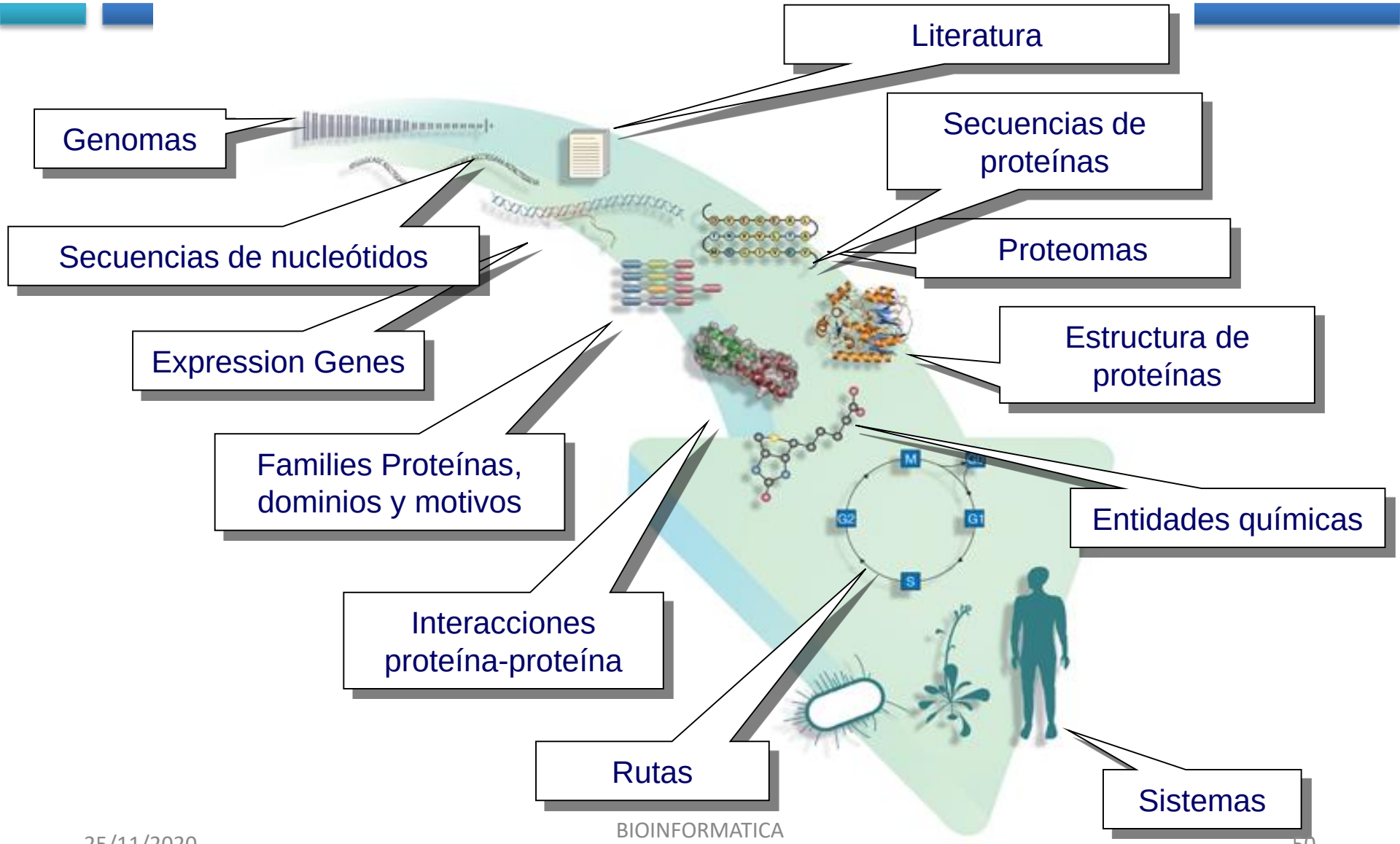
(Quantitative+Computable) Molecular Biology

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.

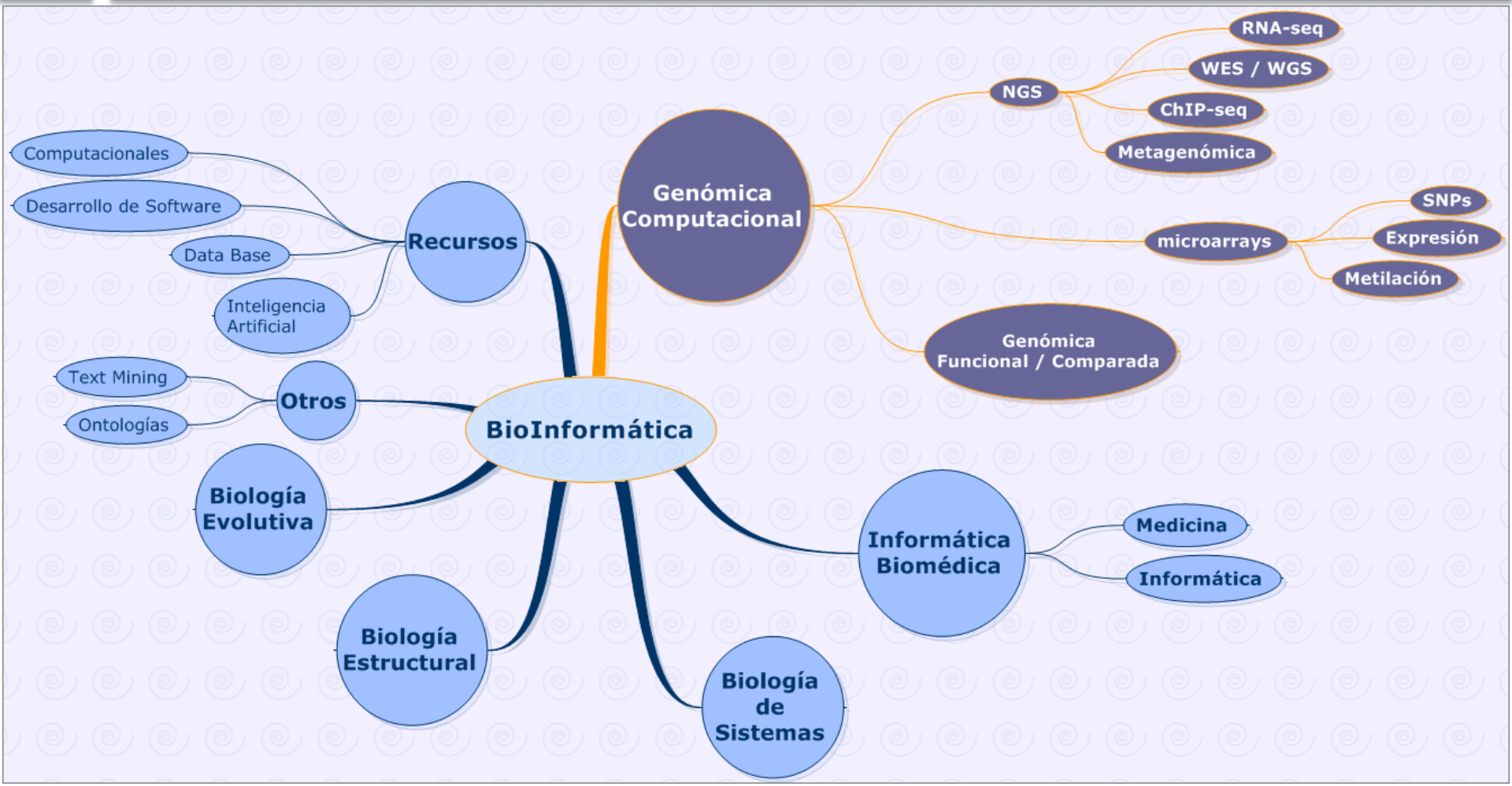
Bioinformática es multidisciplinar



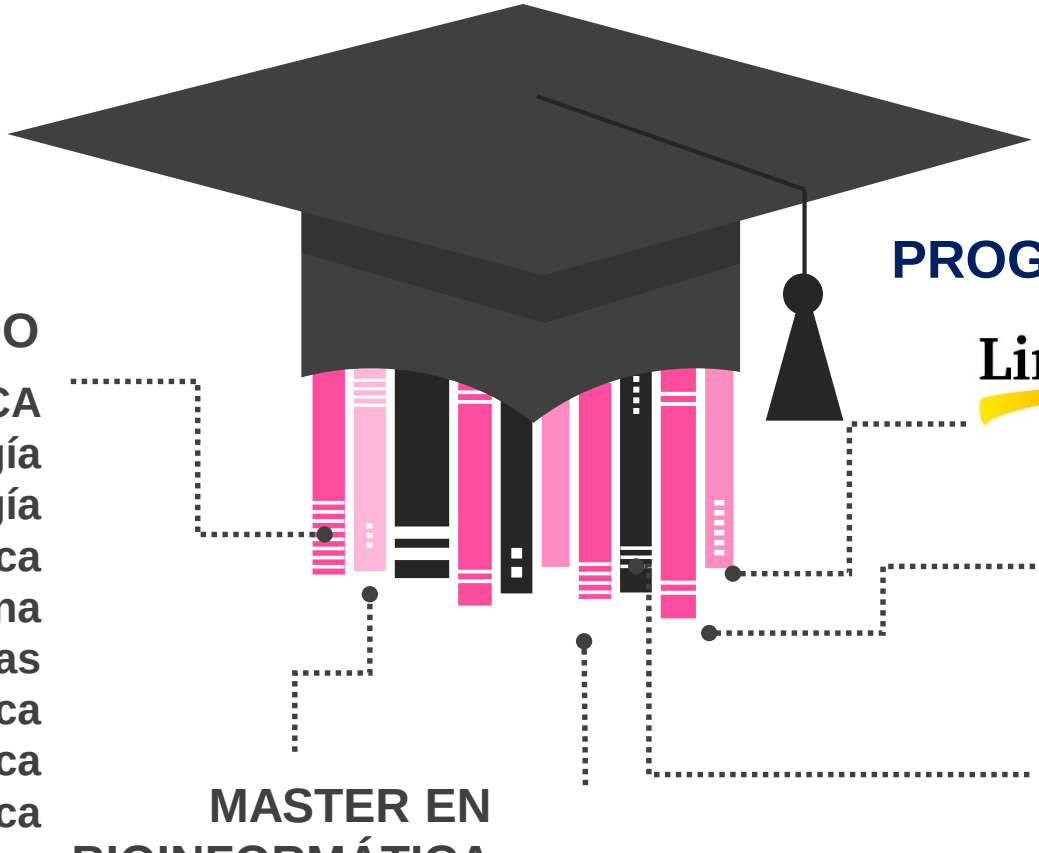
Tipos de datos dan idea de la dimensión de la Bioinformática



AMBITO DE LA BIOINFORMÁTICA



FORMACIÓN EN BIOINFORMÁTICA



Universidad Barcelona.

GRADO

BIOINFORMÁTICA
Biología
Biotecnología
Bioquímica
Medicina
Matemáticas
Química
Fisica
Informática
Telecomunicaciones

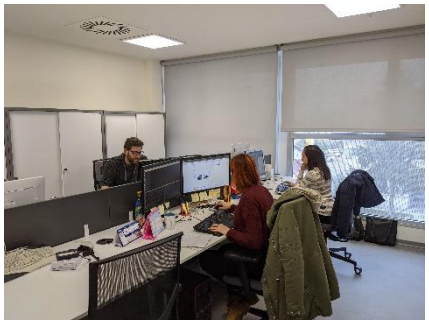
MASTER EN BIOINFORMÁTICA

DOCTORADO

BIOLOGÍA – ANÁLISIS DE DATOS

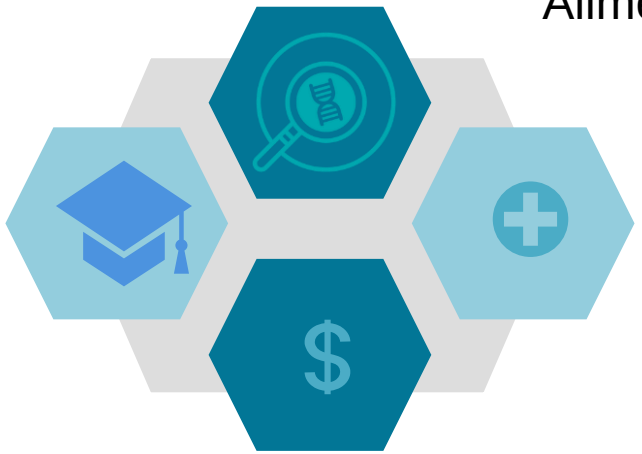
PROGRAMACIÓN

¿Dónde trabaja un Bioinformático?



CENTRO DE INVESTIGACIÓN

Biomedicina
Agricultura
Alimentación



UNIVERSIDAD

Biociencias
Informática

HOSPITAL

BIOINFORMÁTICO CLÍNICO
Genética
Oncología
Cardiología

EMPRESA

Bioinformática
Genética
Genómica

Thanks for your attention!

Questions???