Secuenciación Masiva y análisis de secuencias

Isabel Cuesta

isabel.cuesta@isciii.es

BU-ISCIII

Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

5 Noviembre 2024

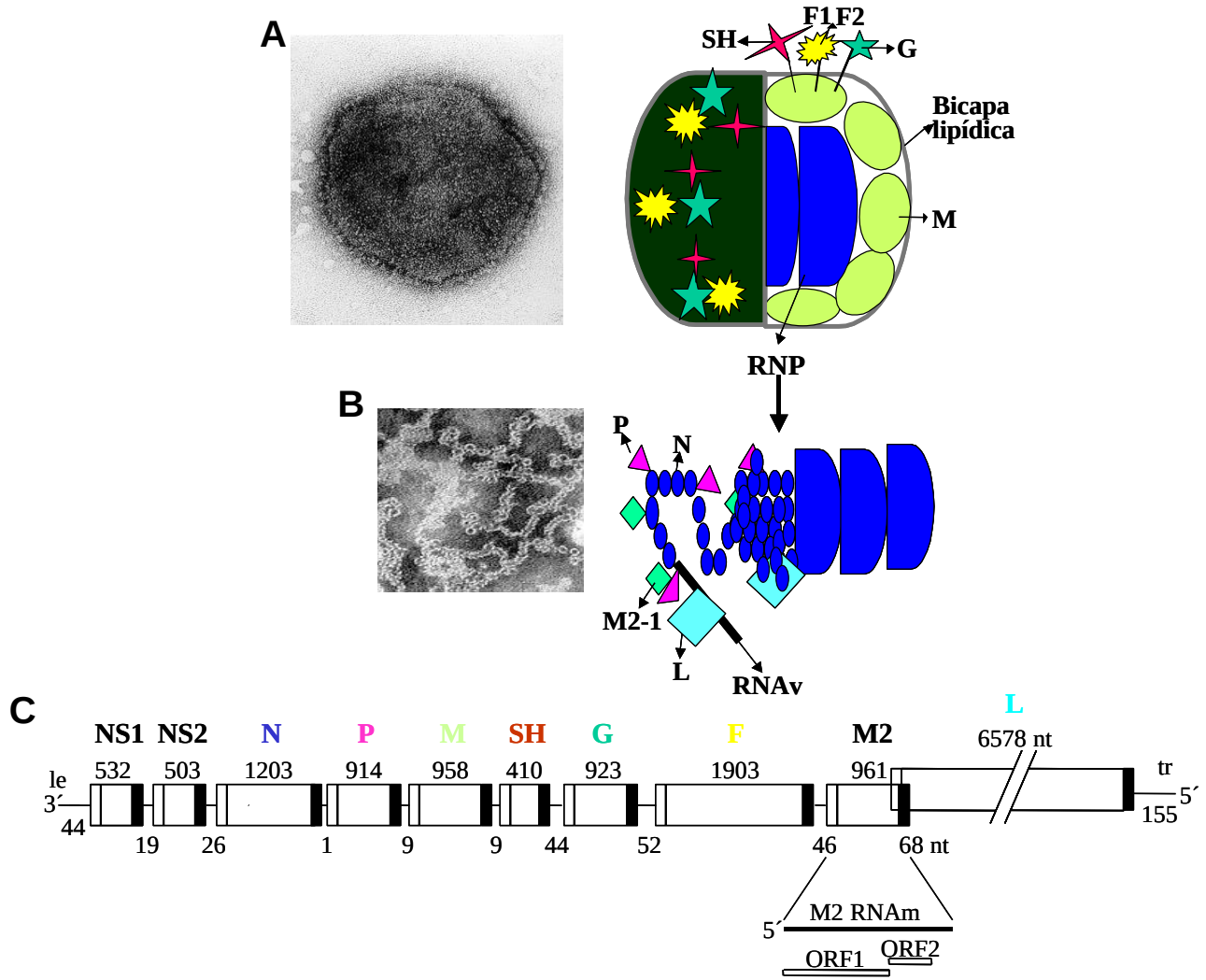
Master Virología

Index

- Secuenciación Masiva (Isabel Cuesta)
- Bioinformática (Isabel Cuesta)
- Conceptos análisis de datos: genoma de SARS-CoV-2 (Sara Monzón)
- Práctica: Viralrecon en Galaxy (Victor Manuel López)

¿Para que secuenciar genomas?

Virus Respiratorio Sincitial Humano




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

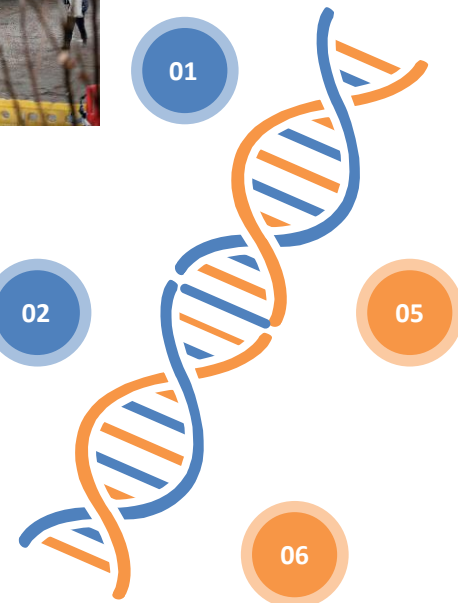
Pathogen discovery: new virus – SARS-CoV-2

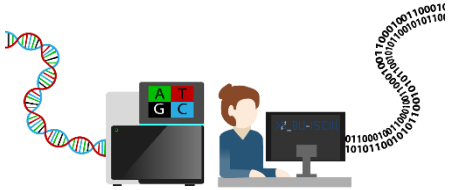
Deep Meta-Transcriptomic Sequencing

bronchoalveolar lavage fluid (BALF)

Meta-transcriptomic library

2x150 MiniSeq 56,565,928 sequences reads



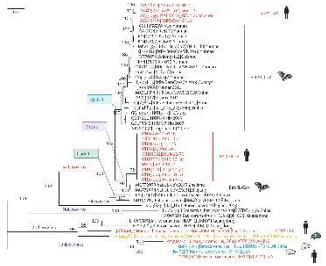


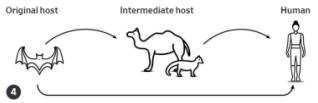
De novo-assembled - Megahit

384,096 Contigs

Screened for potential aetiological agents

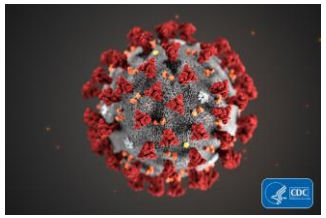
The longest 30,474 nt





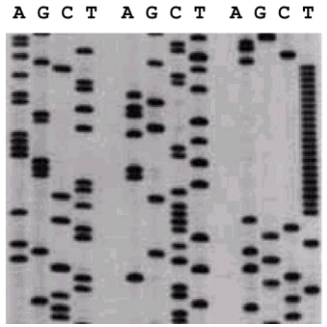
89.1% identity

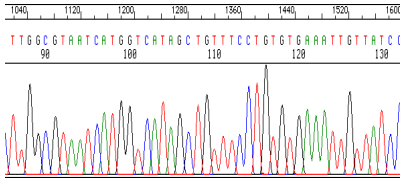
Closely related to a bat SARS-like coronavirus

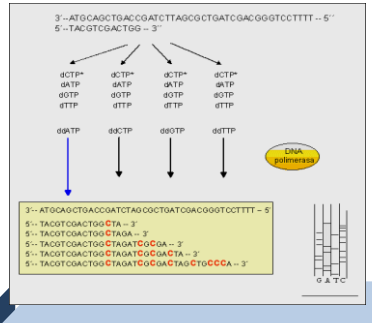


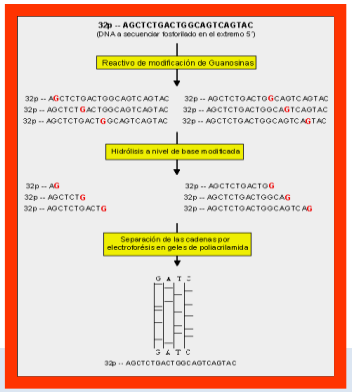
Wu et al., Nature 2020

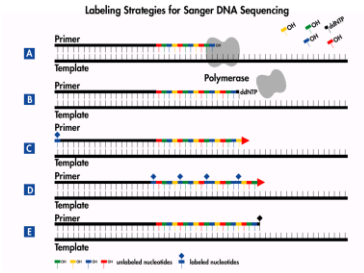
Métodos de secuenciación de DNA











MAXAM Y GILBERT

SANGER

SANGER
Secuenciadores Automáticos

1º GENERACIÓN

Evolution of DNA Revolution

A walk through the biological history: from Sanger to NGS

1953

1972

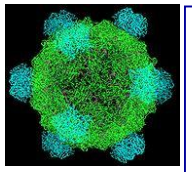
1977

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

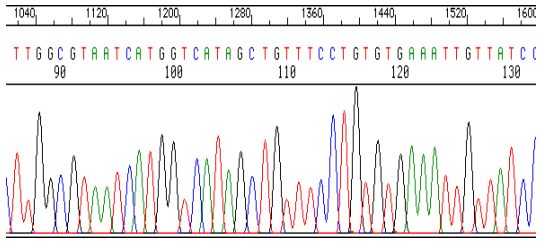
Evolution of DNA Revolution

A walk through the biological history: from Sanger to NGS

1986

1990

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



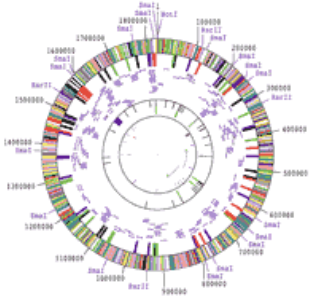
BLAST: Myers & Lipman publish the first algorithm to align DNA sequences (JMB, 215 1990).

Evolution of DNA Revolution

A walk through the biological history: from Sanger to NGS

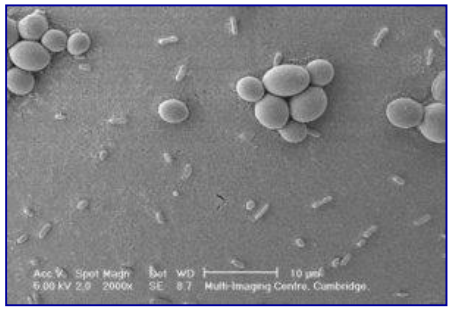
1995

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

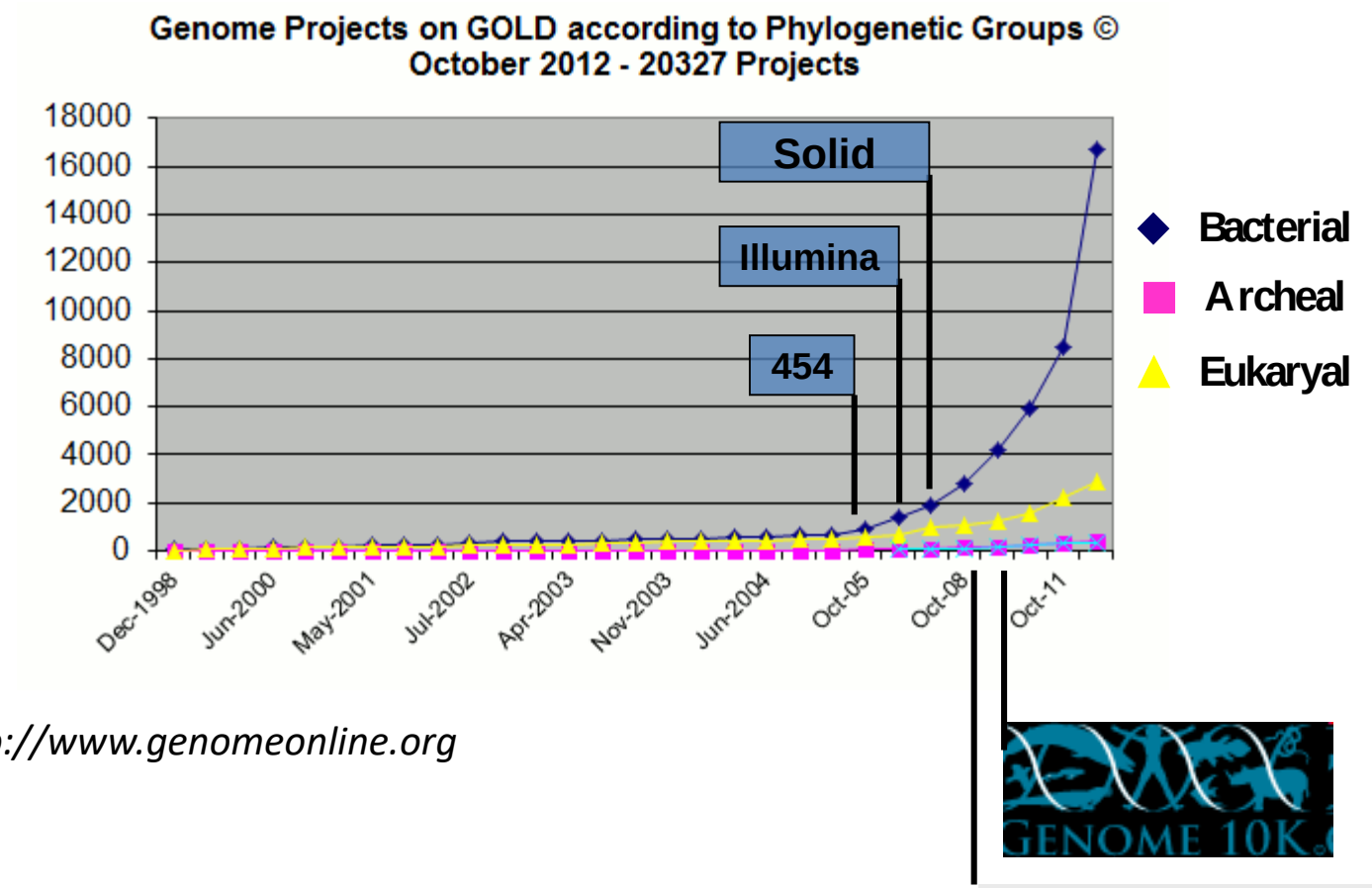
Bacterial Genomes

Bacillus subtilis
Escherichia coli

Eukaryotic Genomes

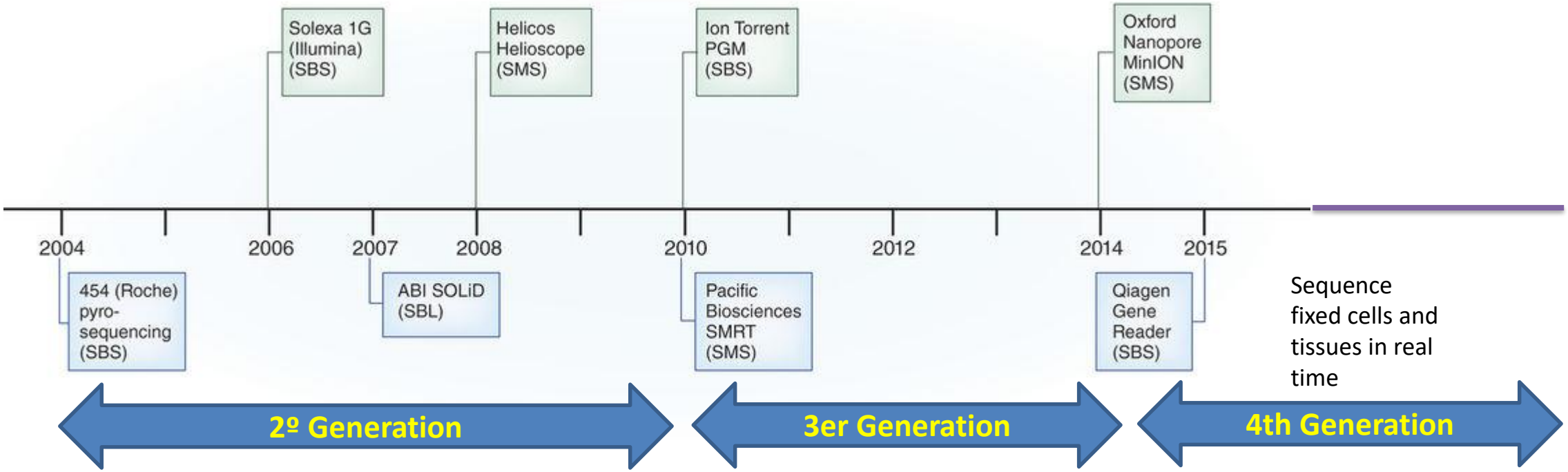
Caenorhabditis elegans
Arabidopsis thaliana
Drosophila melanogaster
Human genome **(1986-2003)**

Genomics Revolution Era



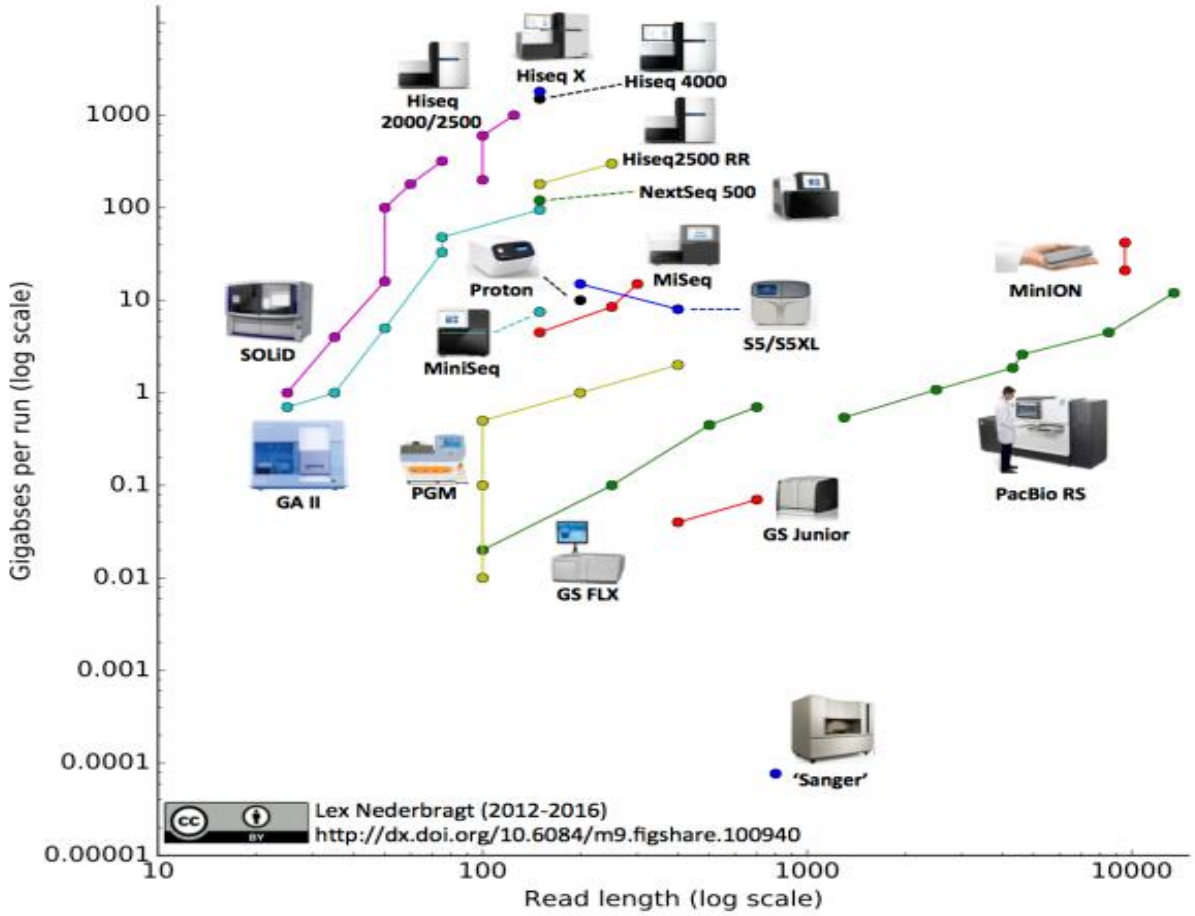
Source: http://www.genomeonline.org

NGS Platforms - Timelime



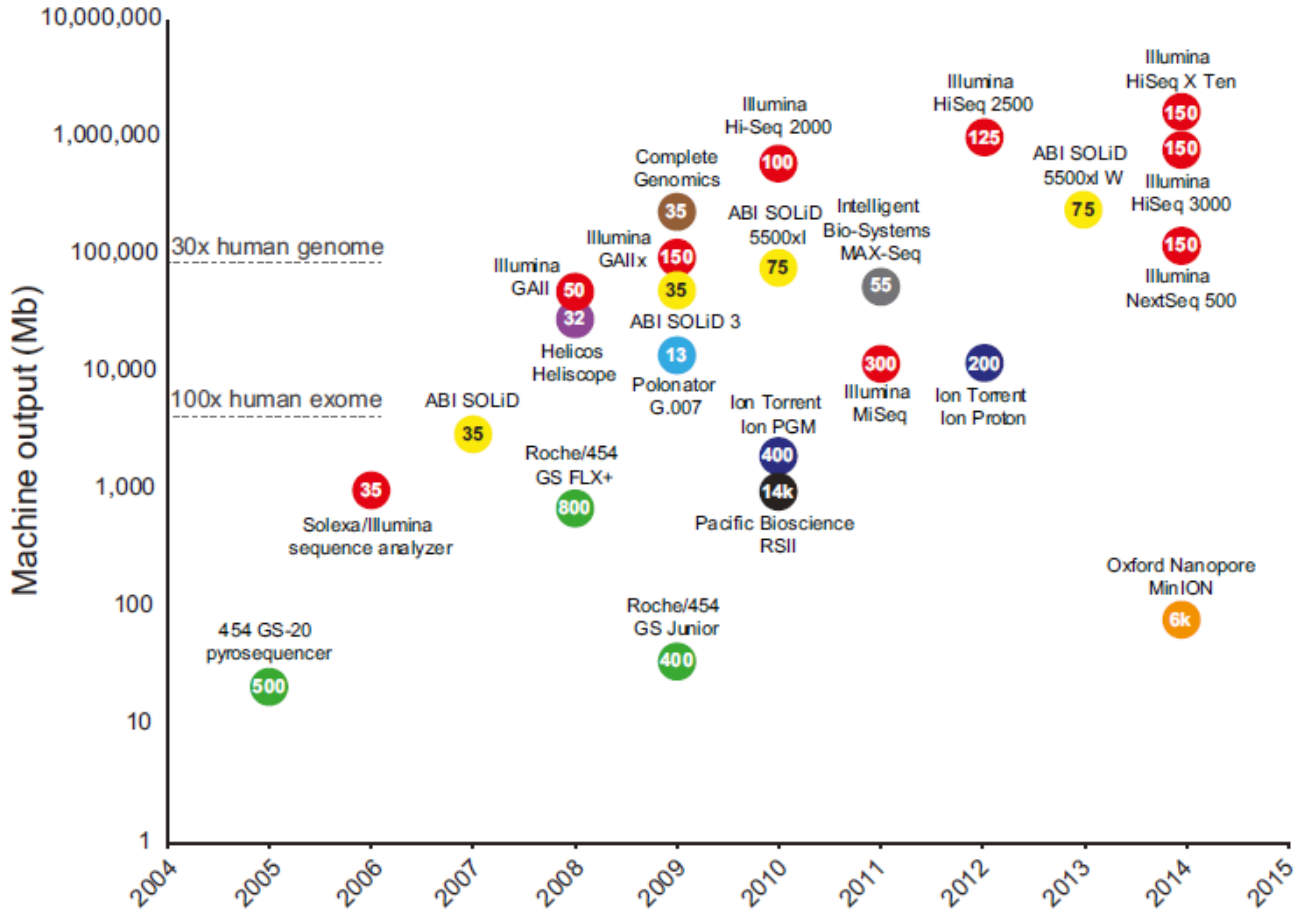
Mardis, Nature Protocols 2017

High-Throughput Sequencing Technologies



https://flxlexblog.wordpress.com/

High-Throughput Sequencing Technologies

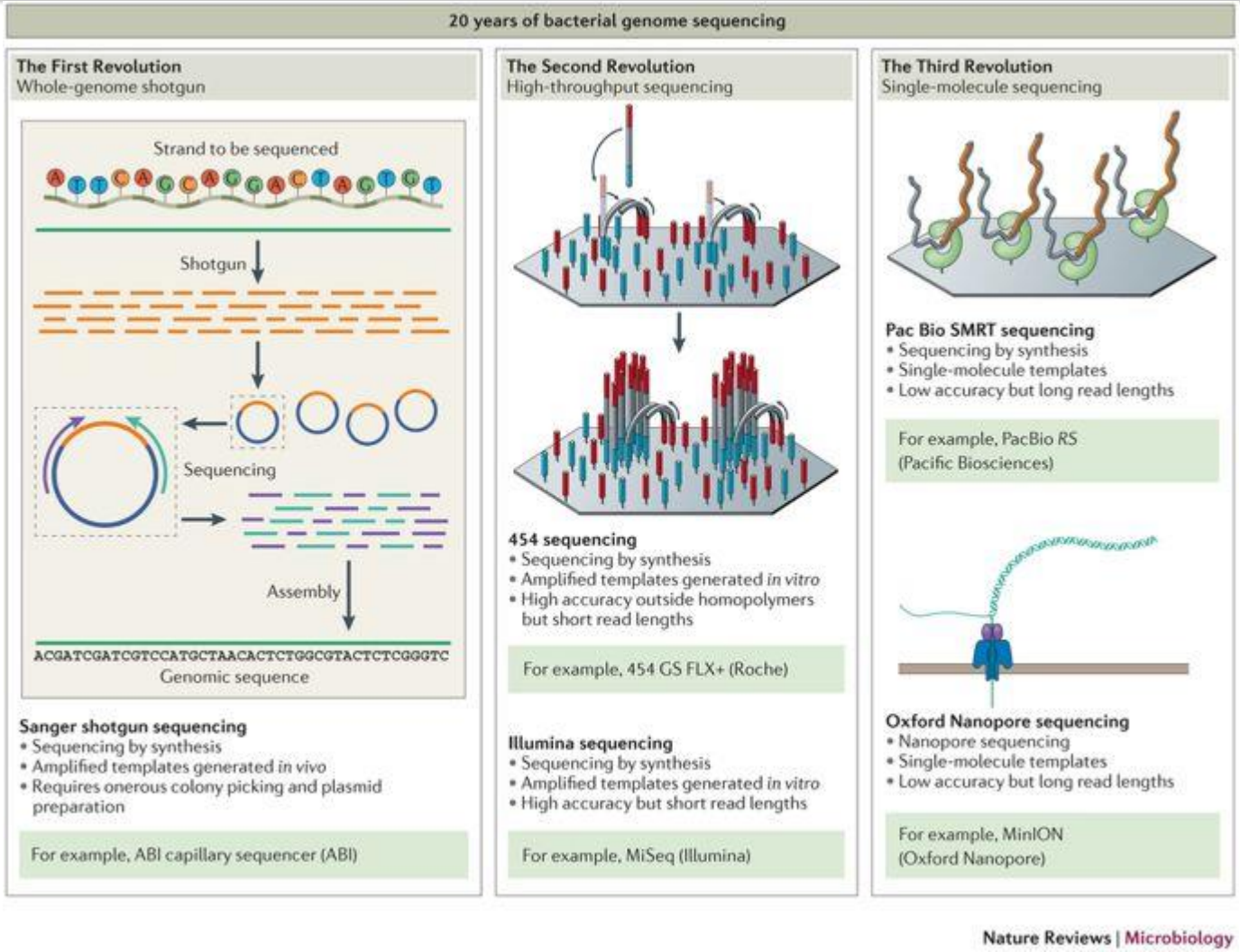


Numbers inside data points denote current read lengths. Sequencing platforms are color coded.

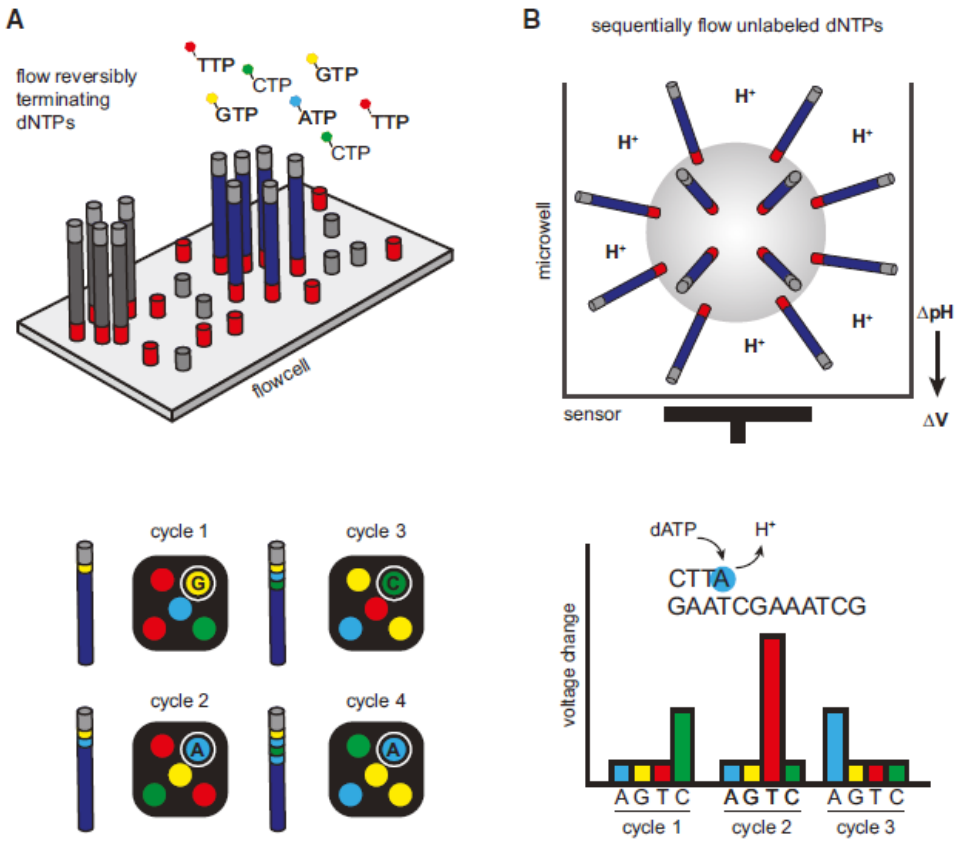
Reuter et al., Mol Cell 2015

High-Throughput Sequencing Technologies

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



The Second-generation Sequencing Technologies



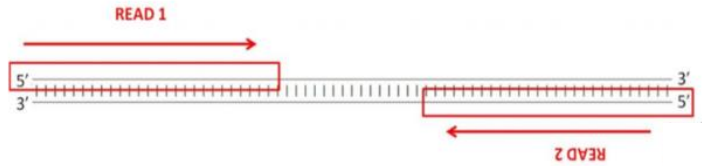
Clonal Amplification-Based Sequencing Platforms

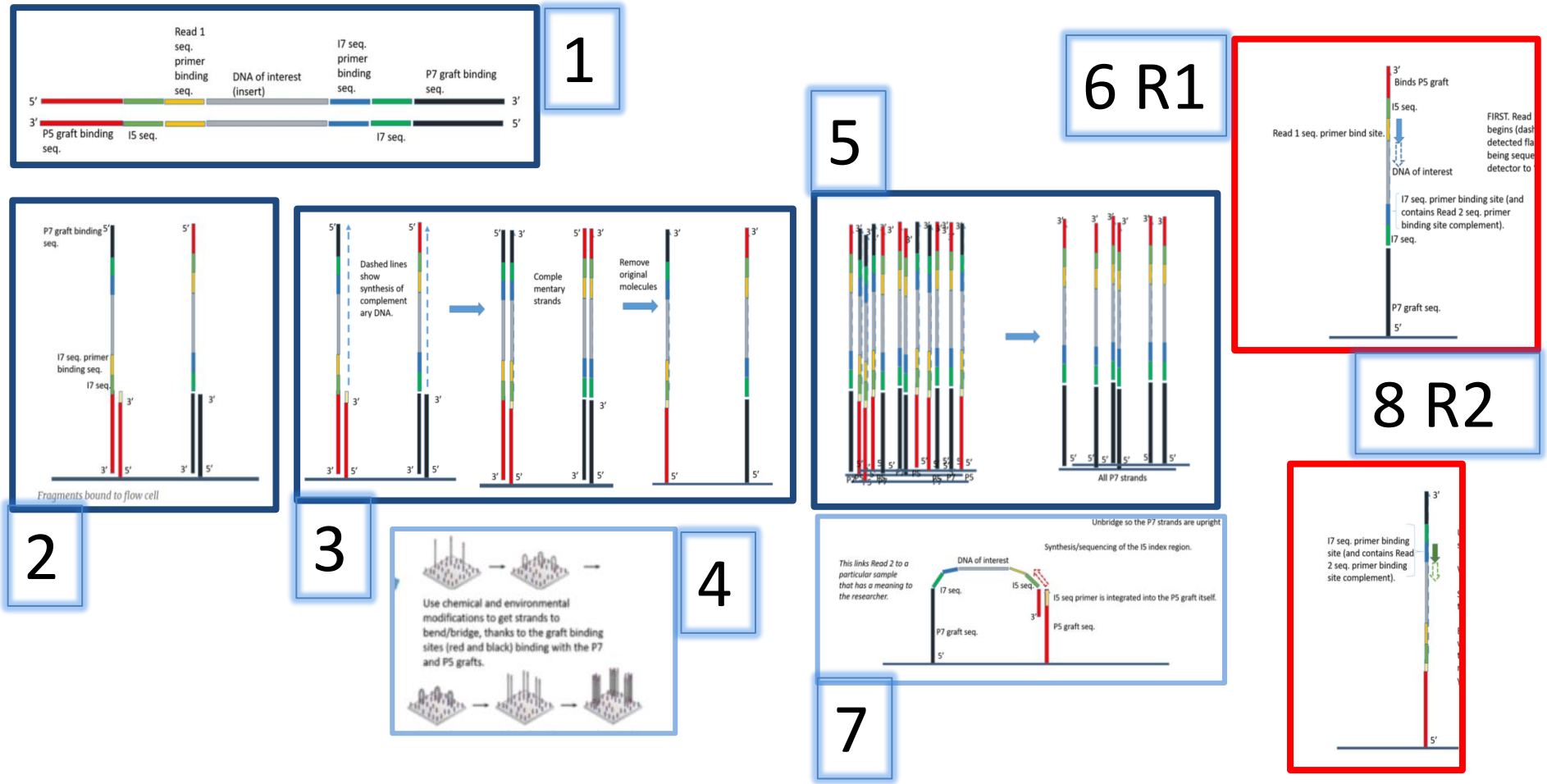
(A) Illumina's four-color reversible termination sequencing method.

(B) Ion Torrent's semiconductor sequencing method.

Reuter et al., Mol Cell 2015

Illumina sequencing





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

Illumina Benchtop Sequencers

Pervez et al., BioMed Research International 2022

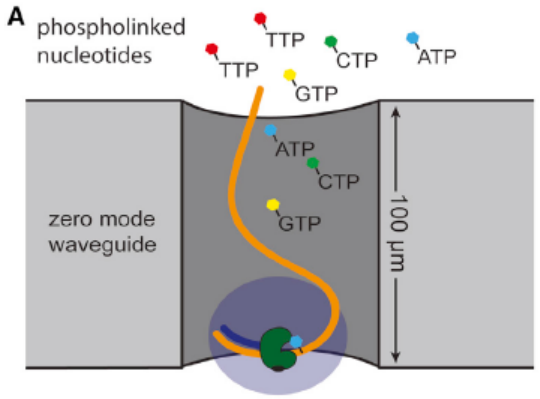
Methods/applications	iSeq 100	MiniSeq	MiSeq series	NextSeq 550 series	Next Seq 1000 & 2000
Ideal for	Every size lab	TG sequencing	Long read applications	Exome and transcriptome sequencing	miRNA and sRNA analysis
Major applications	sWGS (microbes) and TGS	iSeq 100+TG EP and 16S MS	iSeq 100+16S MGS	iSeq 100+TCS	sWGS (microbes), ES, SC profiling, TS, miRNA, and sRNA analysis
Max. data quality	>85% > Q30	>85% > Q30	>90% > Q30	>80% > Q30	>90% > Q30
Run time	9.5–19 h	4–24 hours	4–55 hours	12–30 hours	11-48 hours
Maximum output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	330 Gb*
Maximum reads per run	4 million	25 million	25 million	400 million	1.1 billion
Maximum read length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

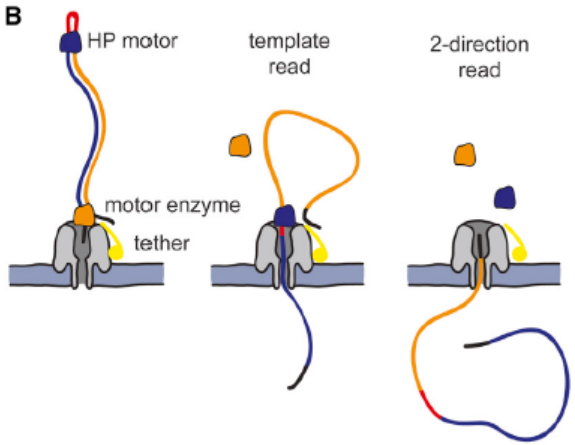
Illumina Production Scale Sequencers

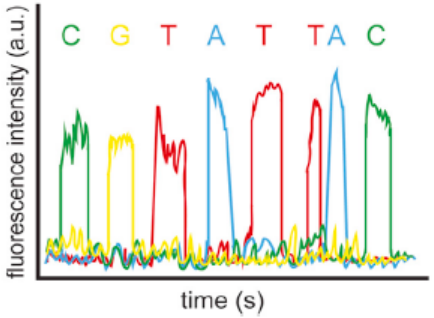
Pervez et al., BioMed Research International 2022

Methods/applications	NextSeq 550	NextSeq 550Dx	NextSeq 1000 & 2000	NovaSeq 6000
Ideal for	Research	Research+in vitro diagnostic	Targeted sequencing	Long read applications
Major applications	sWGS (microbes), TGS, and TCS	NextSeq 550+clinical NGS applications	NextSeq 550 series+SCP	NextSeq 550 series+NextSeq 1000 & 2000+lWGS
Max. data quality	>80% > Q30	>75% > Q30	>90% > Q30	>90% > Q30
Run time	12-30 hours	35 hours	11-48 hours	13-44 hours
Maximum output	120 Gb	90 Gb	360 Gb	6000 Gb
Maximum reads per run	400 million	300 million	1.2 billion	20 billion
Maximum read length	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 250 bp

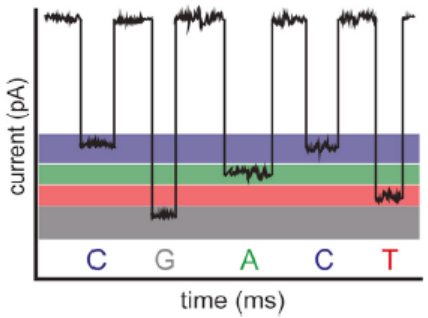
The Third-generation Sequencing Technologies







Pacific Bioscience's SMRT sequencing



Oxford Nanopore's sequencing strategy

Reuter et al., Mol Cell 2015

Single Molecule Sequencing Platforms

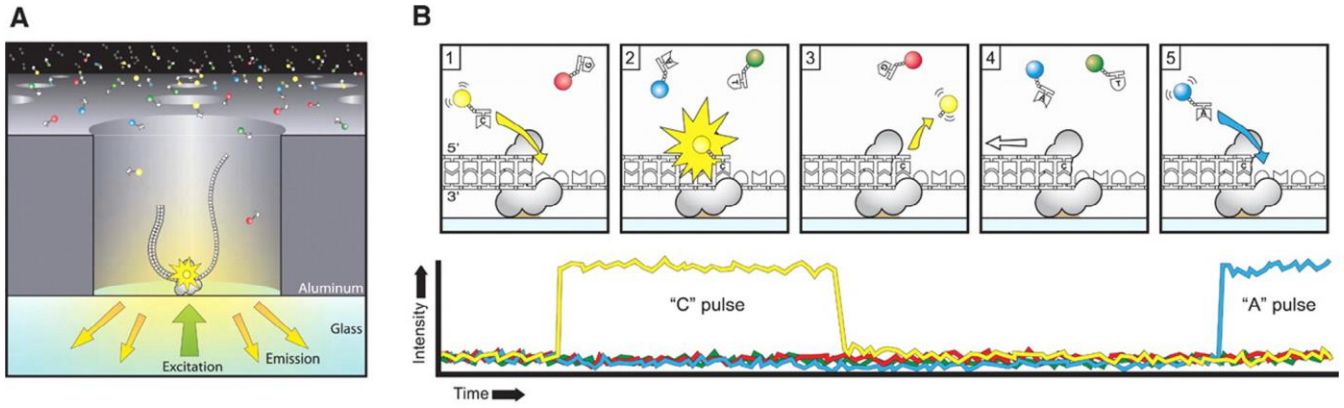
PacBio sequencing and its applications

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.

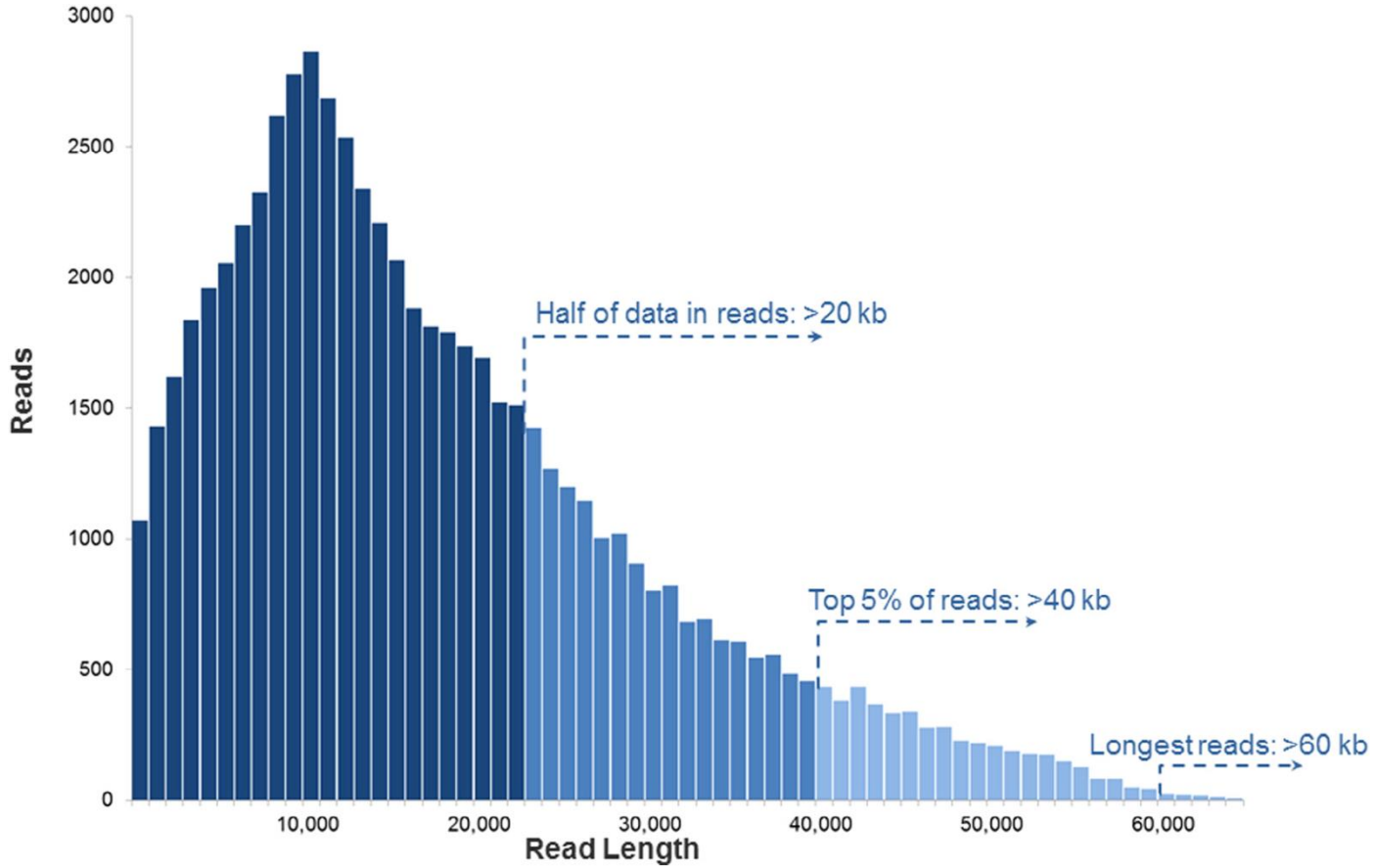
Sequencing by light pulses: The replication processes in all ZMWs of a SMRTcell are recorder by a movie of light pulses, and the pulses corresponding to each ZMW can be interpreted to be a sequence of bases (**continuous long read, CLR**).
Both strands can be sequenced multiple times (passes) in a single 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

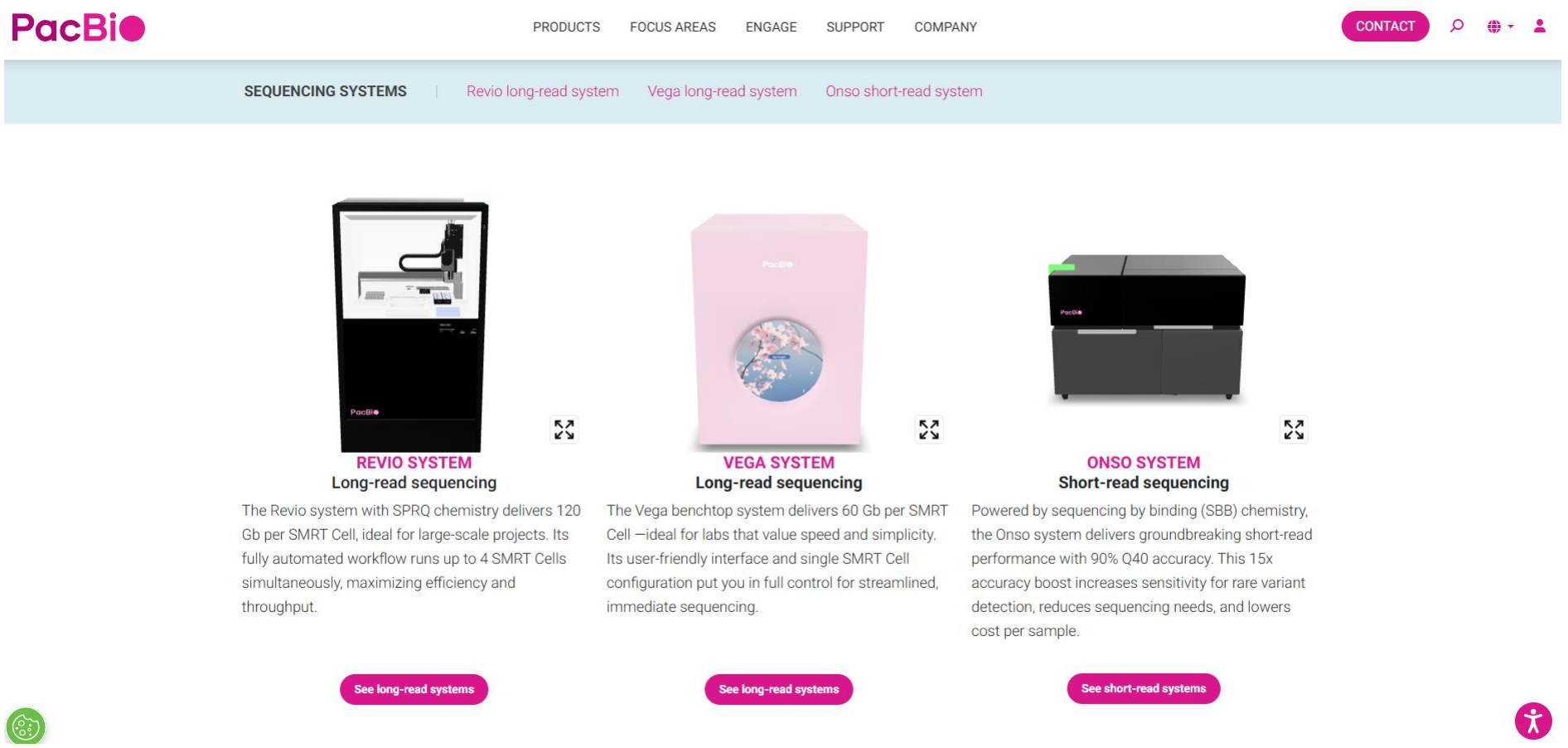
Discovered STRs (short tandem repeats)

Limitations

Both strands can be sequenced several times if the lifetime of the polymerase is long enough.

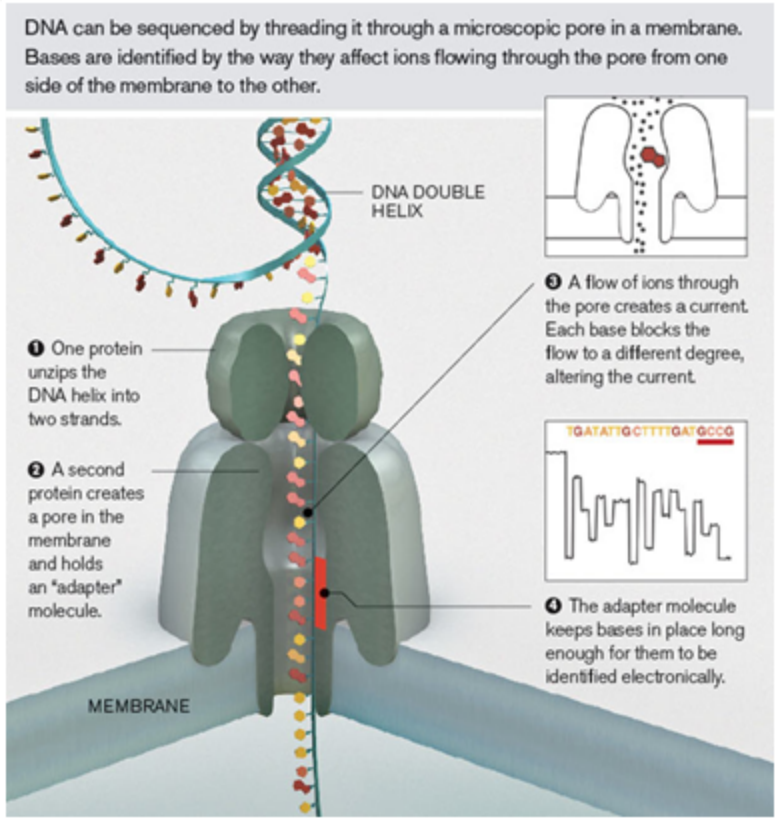
PacBio sequencing and its applications

https://www.pacb.com/sequencing-systems/



Nanopore-based fourth-generation DNA sequencing technology.

ONT, Oxford Nanopore Technologies



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

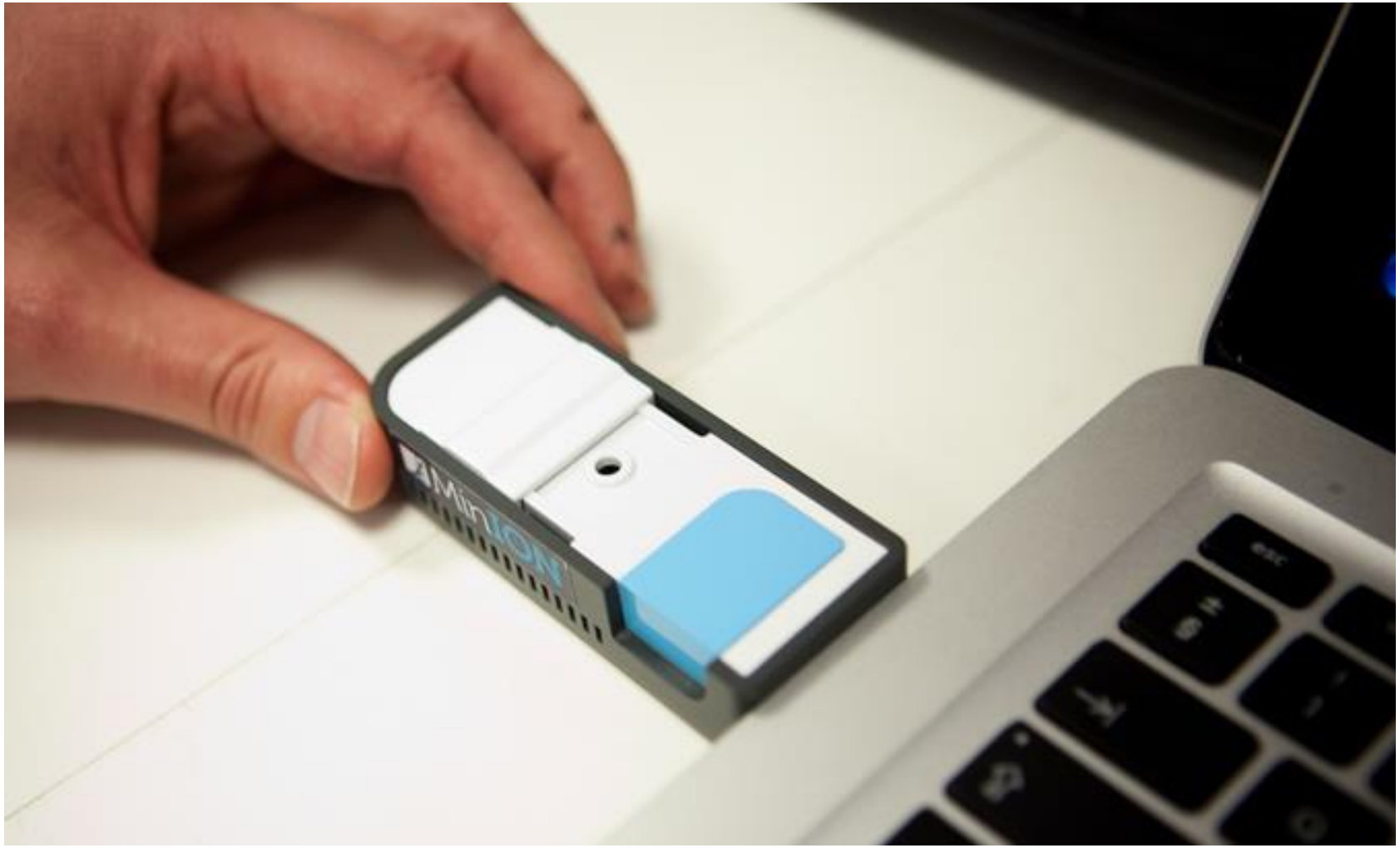
Nanopore sequencing also offers, for the first time, direct RNA sequencing, as well as PCR or PCR-free cDNA sequencing.

https://nanoporetech.com/applications/dna-nanopore-sequencing

Feng et al , Gen Prot Bioinf 2015

Nanopore-based fourth-generation DNA sequencing technology.

ONT, Oxford Nanopore Technologies



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

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

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

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

Comparison of various high-performing sequencing instruments

Pervez et al., BioMed Research International 2022

Manufacturer	Read length	Data output	Max. run time (hours)	Chemistry	Key applications**
Illumina (NovaSeq 6000)	300 PE	6 Tb (6000 Gb)	44	Sequencing by synthesis	SS-WGS and TGS, TGEP, 16sMGS, WES, SCP, LS-WGS, CA, MS, MGP, CFS, LBA
Thermo Fisher Scientific Ion Torrent (Ion GeneStudio S5 Prime)	600 SE	50 Gb	12	Sequencing by synthesis	WGS, WES, TGS
GenapSys (16 chips)	150 SE	2 Gb	24	Sequencing by synthesis	TS, SS-WGS, GEV, 16S rRNA sequencing, sRNA sequencing, TSCAS
QIAGEN (GeneReader)	100 SE	Not available	Not available	Sequencing by synthesis	Cancer research and identifying mutations
BGI/Complete Genomics	400 SE	6 Tb (6000 Gb)	40	DNA nanoball	Small and large WGS, WES and TGS
PacBio (HiFi Reads)	25 Kb	66.5 Gb	30	Real-time sequencing	DN sequencing, FT, identifying ASI, mutations, and EPM
Nanopore (PromethION)	4 Mb	14 Tb (14000 Gb)	72	Real-time sequencing	SV, GS, phasing, DNA and RNA base modifications, FT, and isoform detection

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

Advantages & disadvantages of sequencing platforms

Pervez et al., BioMed Research International 2022

Sequencing generation	Advantages	Disadvantages
First generation	High accuracy Helps in validating findings of NGS	High cost Low throughput
Second generation	High throughput Low cost Have clinical applications Short run time	Short read length Difficult sample preparation PCR amplification Long run time
Third generation	No PCR amplification Require less starting material Longer read lengths Very low cost Low error rate during library preparation Advantages of 3rdGS+	High sequencing error rate 10–15% in the PacBio and 5–20% in the ONT Fresh DNA requires for ensuring quality of ultralong reads Database systems and algorithms/tools are rare for analyzing 3rd and 4th GS data
Fourth generation	Ultrafast: scan of whole genome in 15 minutes Spatial distribution of the sequencing reads over the sample can be seen	

Advantages & disadvantages for short vs. Long read sequencing

	Advantages	Limitations
Short-read sequencing	· Higher sequence fidelity · Cheap · Can sequence fragmented DNA	· Not able to resolve structural variants, phasing alleles or distinguish highly homologous genomic regions · Unable to provide coverage of some repetitive regions
Long-read sequencing	· Able to sequence genetic regions that are difficult to characterize with short-read seq due to repeat sequences · Able to resolve structural rearrangements or homologous regions · Able to read through an entire RNA transcript to determine the specific isoform · Assists *de novo* genome assembly	· Lower per read accuracy · Bioinformatic challenges, caused by coverage biases, high error rates in base allocation, scalability and limited availability of appropriate pipelines

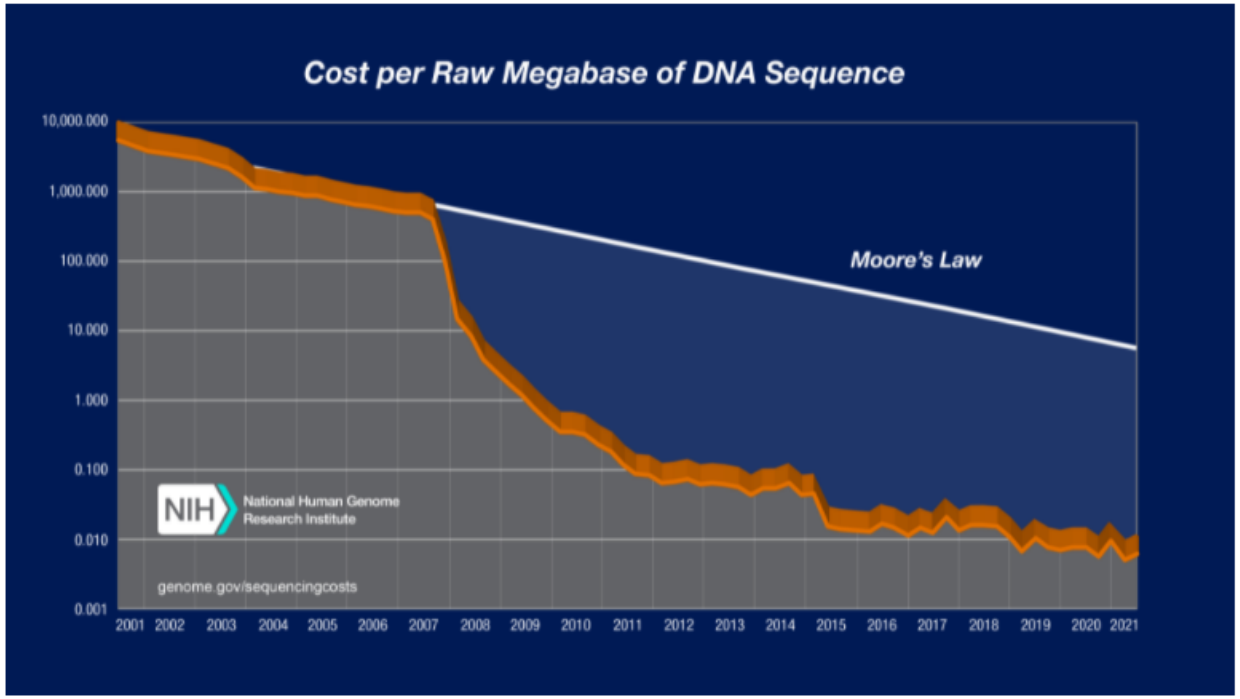
https://www.technologynetworks.com/genomics/articles/an-overview-of-next-generation-sequencing-346532

NGS PLATFORMS, main characteristics

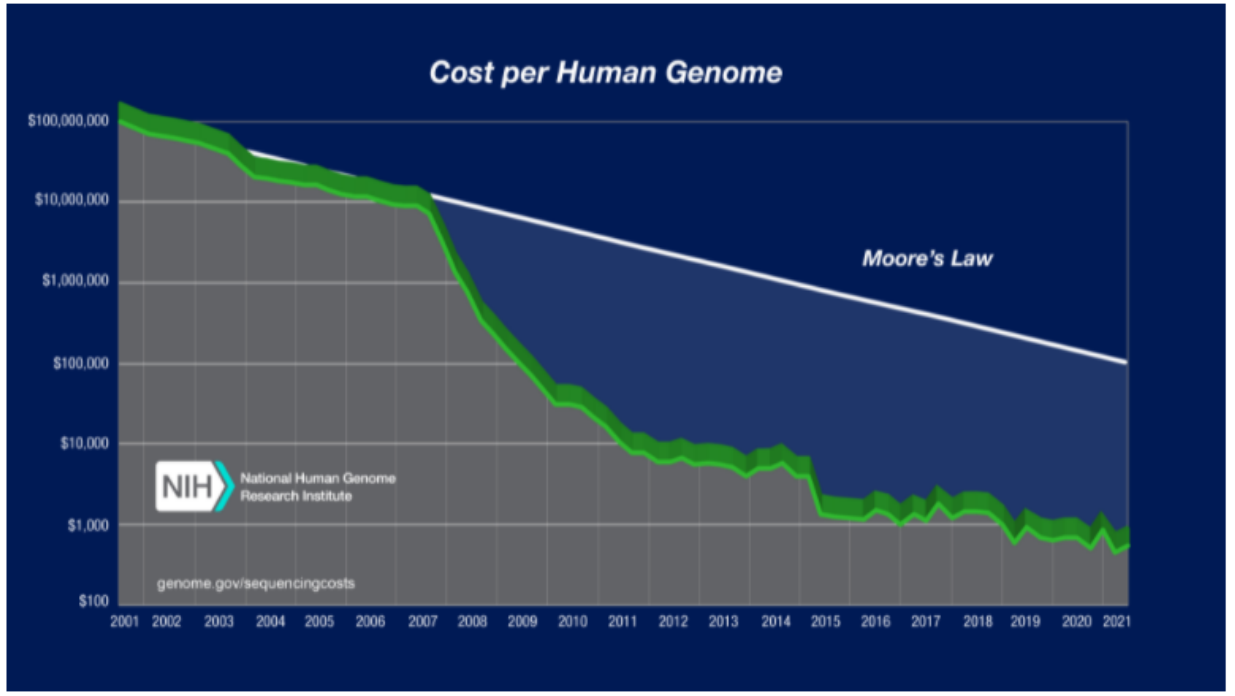
- Numero de bases que secuencia
- Numero lecturas → aplicaciones
- Longitud de las lecturas -→ importante para las aplicaciones ensamblado genomas, de illumina a PacBio

- Error de la base → Corrección con profundidad de lectura
- Formato fichero salida
- Software dedicado, universal fastq

Secuencing cost



Sequencing cost per megabase - 2021



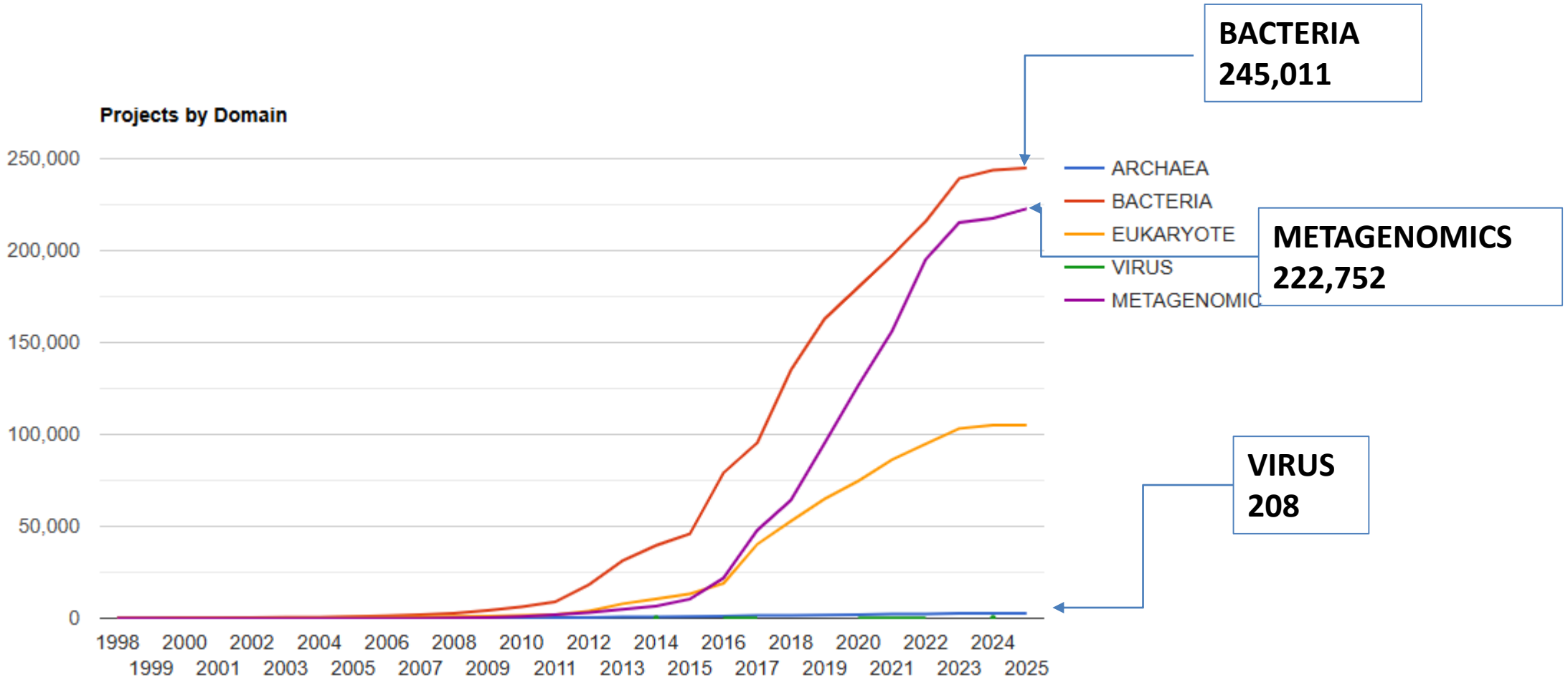
Cost per genome data - 2021

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

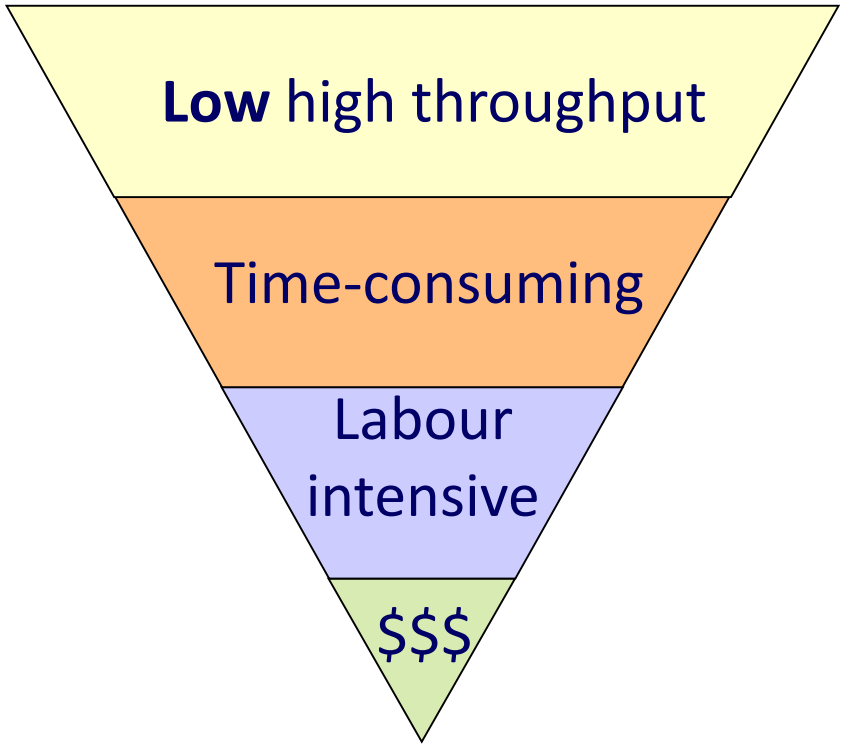
Sequencing projects

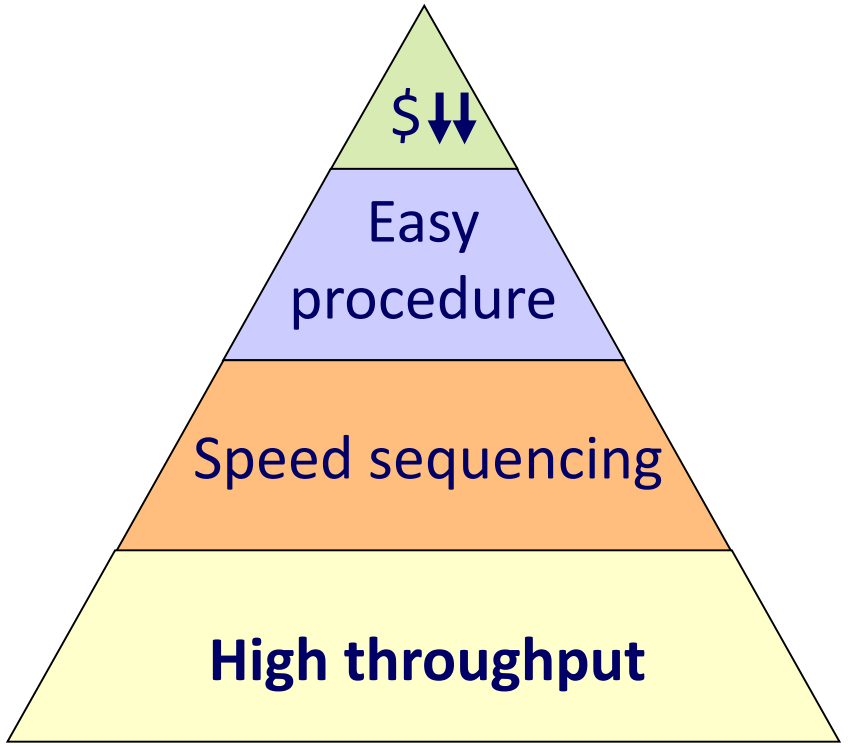
https://gold.jgi.doe.gov/

GOLD, Genome Online DataBase



Sanger vs SM, advantages of new technologies

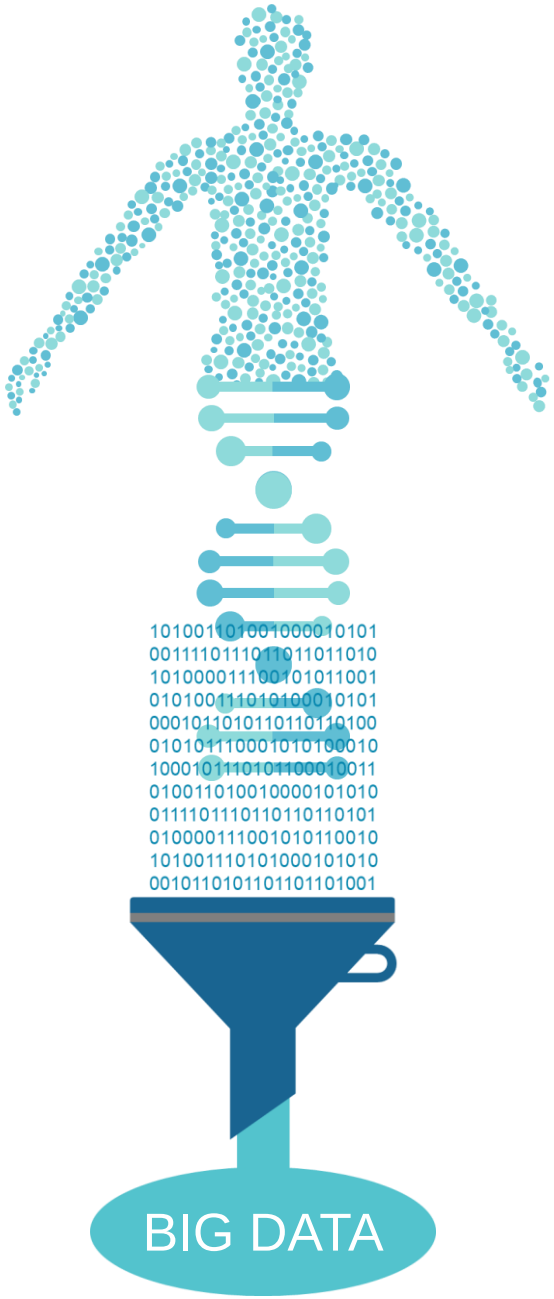


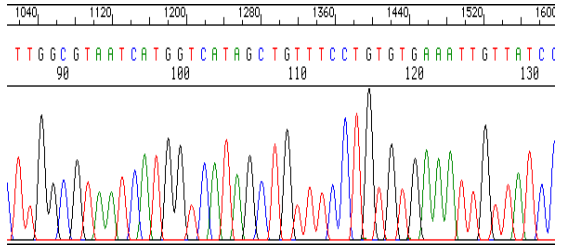


Semiautomatic **Sanger** capillary-based sequencing technology

NGS
Next Generation Sequencing =
Now Generation Sequencing

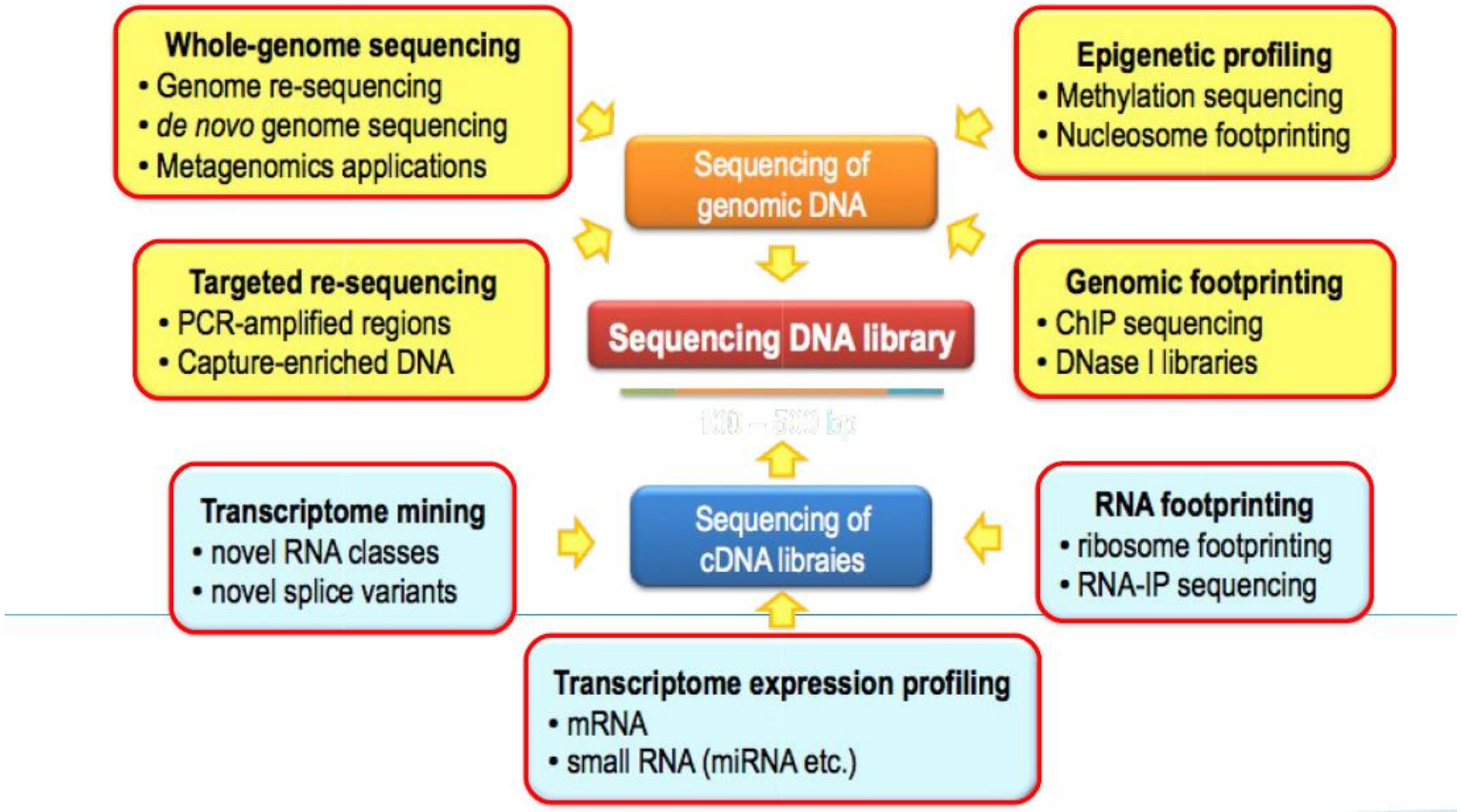
BIG DATA






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

Massive sequencing applications



APLICACIONES DE NGS EN EL DIAGNÓSTICO VIROLÓGICO

- 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
- Estudio de la Evolución viral
- Control de calidad de las vacunas virales vivas atenuadas

….

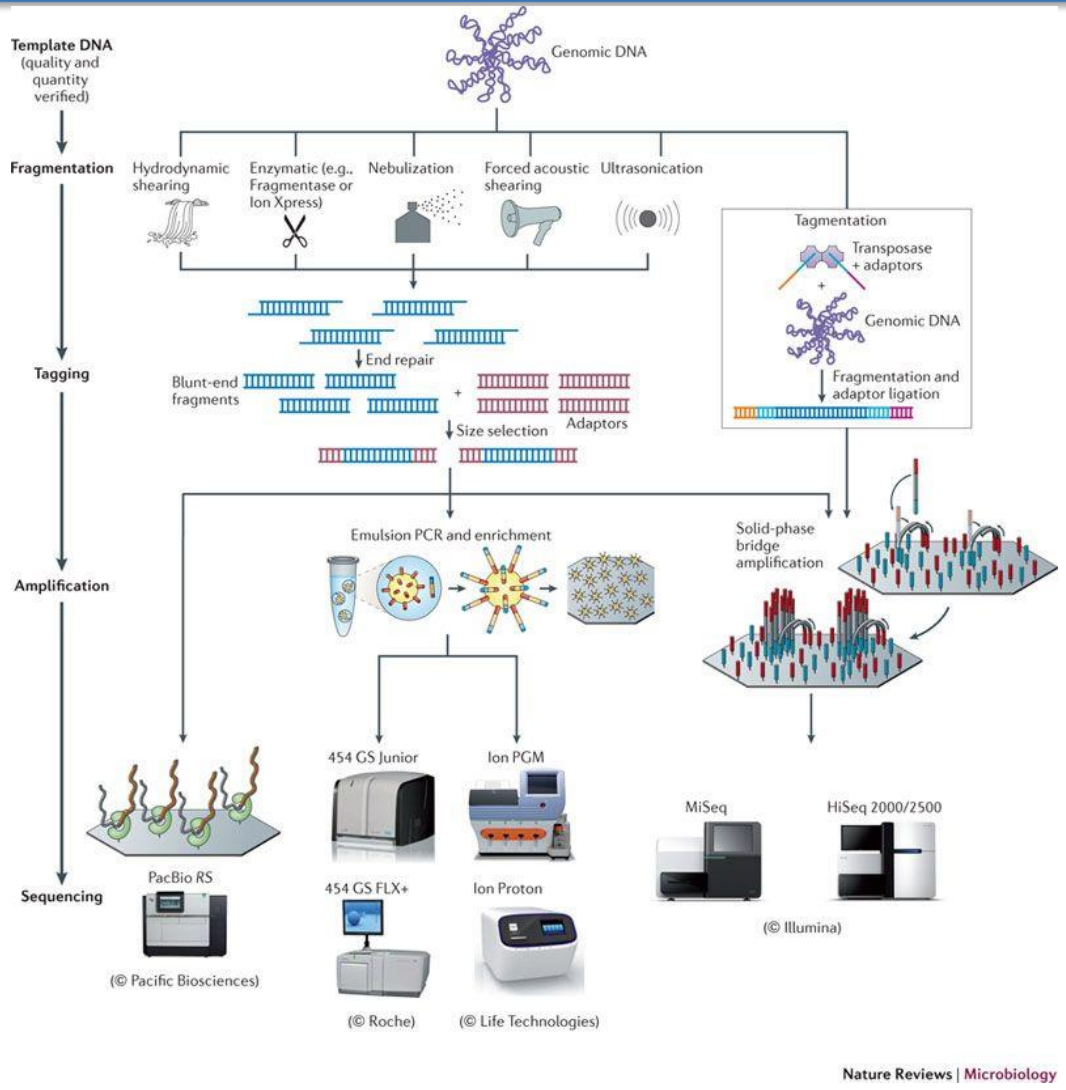
High throughput sequencing in clinical virology, Applications

- Virus discovery: Metagenomics, sequence-independent amplification
- Viral genome reconstruction
- Variants identification & quasispecies

Main Steps of Viral Genome Sequencing by NGS or HTS

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

High-throughput sequencing platforms



Loman et al, 2012

General features of viral genomes

S. No.	Class	Sequenced genomes	Size (Nt)	Proteins
1	DsDNA	414	4697–335,593	6–240
2	SsDNA	230	1360–10,958	6–11
3	DsRNA	61	3090–29,174	2–13
4	SsRNA (+)	421	2343–31,357	1–11
5	SsRNA (−)	81	8910–25,142	5–6

K. V. Chaitanya, Genome and Genomics,
https://doi.org/10.1007/978-981-15-0702-1_1

Genomes of Single Stranded DNA Viruses and their Mosaicism

Table 1.2 Morphological diversity of single stranded DNA viruses

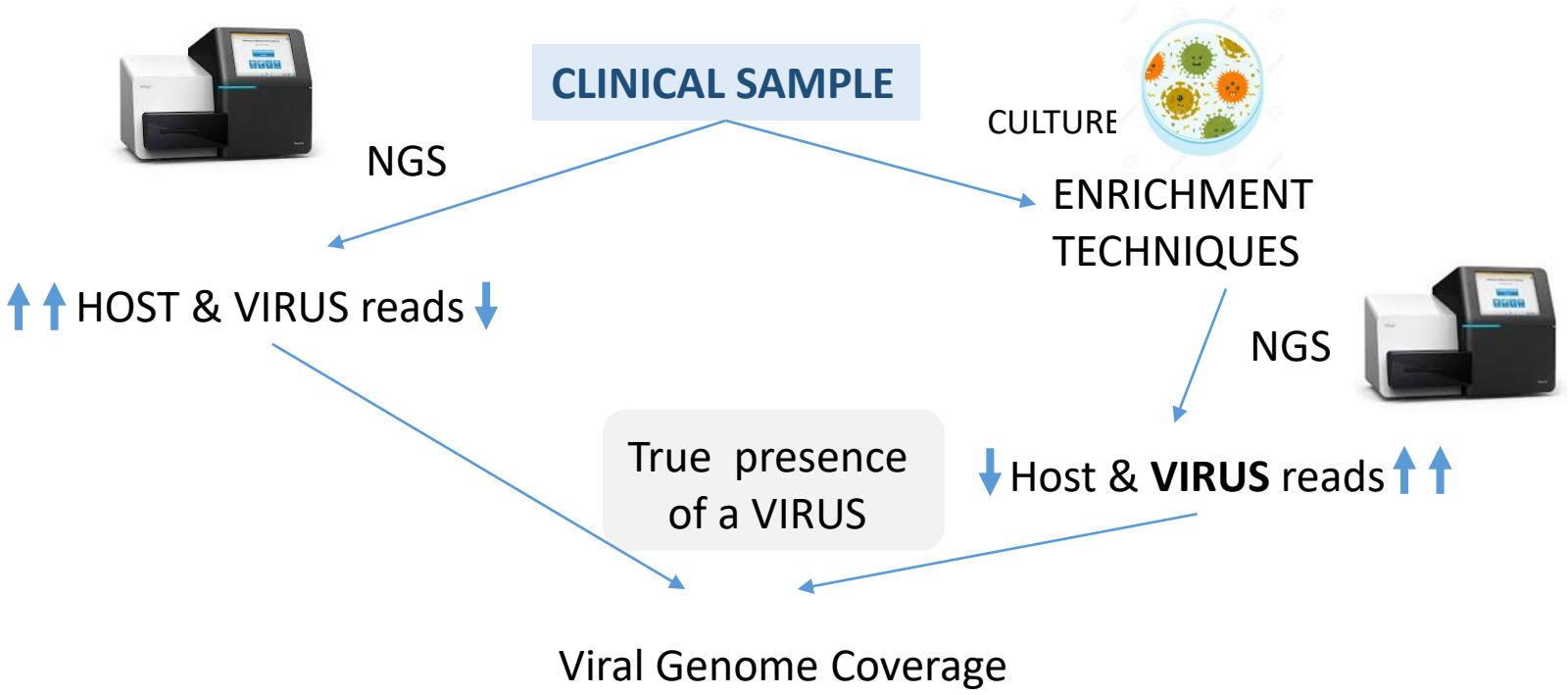
Host virus taxon	Virion morphology	Genome topology	Genome size
Microviridae	Icosahedral	Circular	4.4–6.1
Inoviridae			
Inovirus	Filamentous		5.8–12.4
Plectrovirus	Rod-shaped		4.5–8.2
Pleolipoviridae	Pleomorphic	Circular	7–10.6
Spiraviridae	Coil-shaped	Circular	24.9
Anelloviridae	Icosahedral	Circular	2–4
Bidnaviridae	Icosahedral	Linear, segmented, 6.5 per segment	
Circoviridae	Icosahedral	Circular	1.7–2.3
Geminiviridae	Icosahedral	Circular, segmented 3 per segment	
Nanoviridae	Icosahedral	Circular, segmented 0.98–1.1per segment	
Parvoviridae	Icosahedral	Linear	4–6.3

Table 1.3 Size of influenza virus segments and the proteins they encode

RNA segment	No. of nucleotides	Encoding protein	No. of amino acids
1	2341	Polymerase PB2	759
2	2341	Polymerase PB1	757
3	2233	Polymerase PA	716
4	1778	Haemagglutinin HA	566
5	1565	Nucleoprotein NP	498
6	1413	Neuraminidase NA	454
7	1027	Matrix protein M1	252
		Matrix protein M2	97
8	890	Non-structural protein NS1	230
		Non-structural protein NS2	121

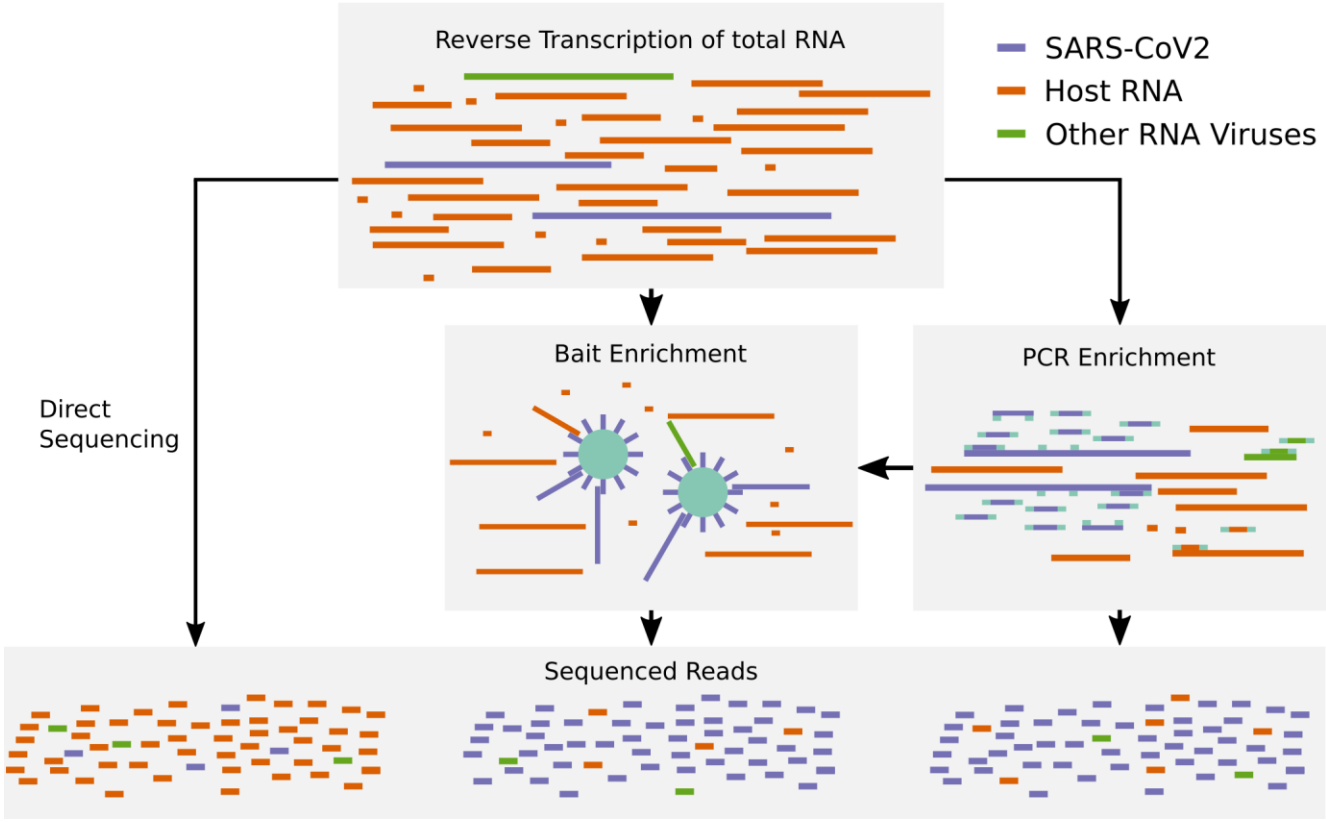
K. V. Chaitanya, Genome and Genomics,
https://doi.org/10.1007/978-981-15-0702-1_1

Viral Genome Sequencing

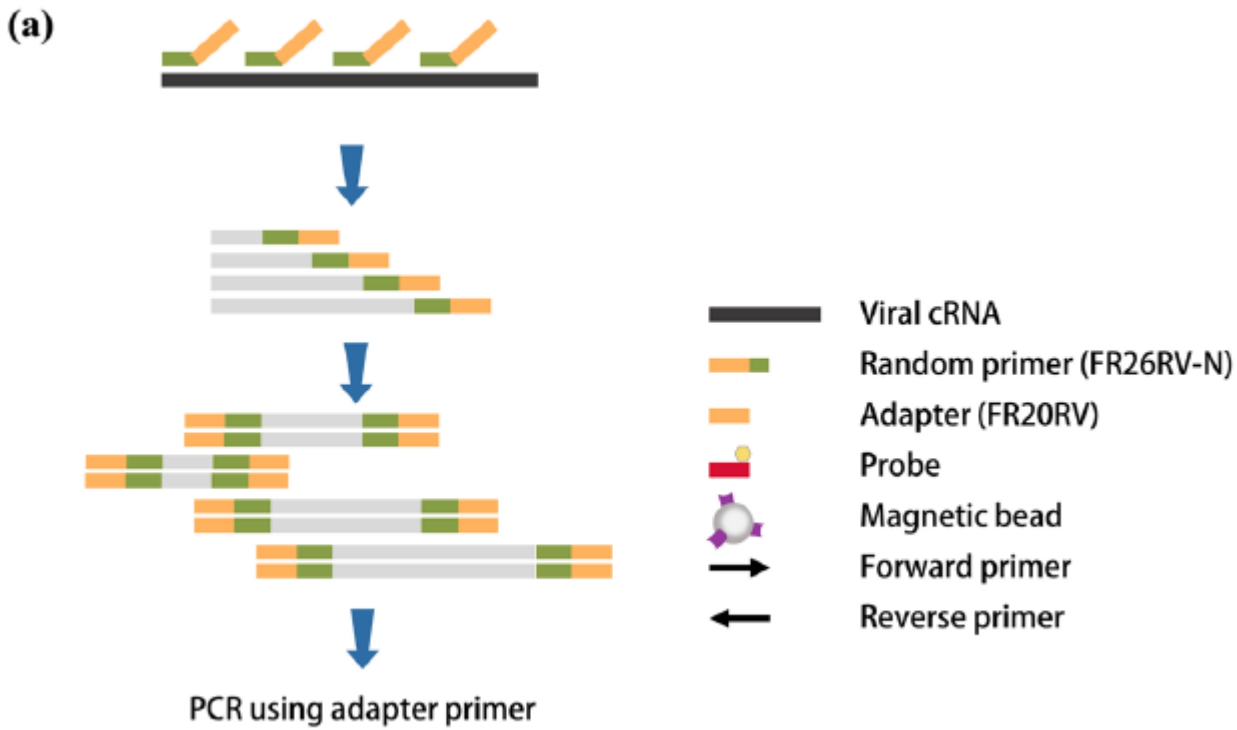


NGS needs a cutoff to determine the true presence of a pathogen versus carry-over or contamination between specimens or other non-specific reads.

Enrichment Techniques



Nucleic Acid Extraction: SISPA NGS method



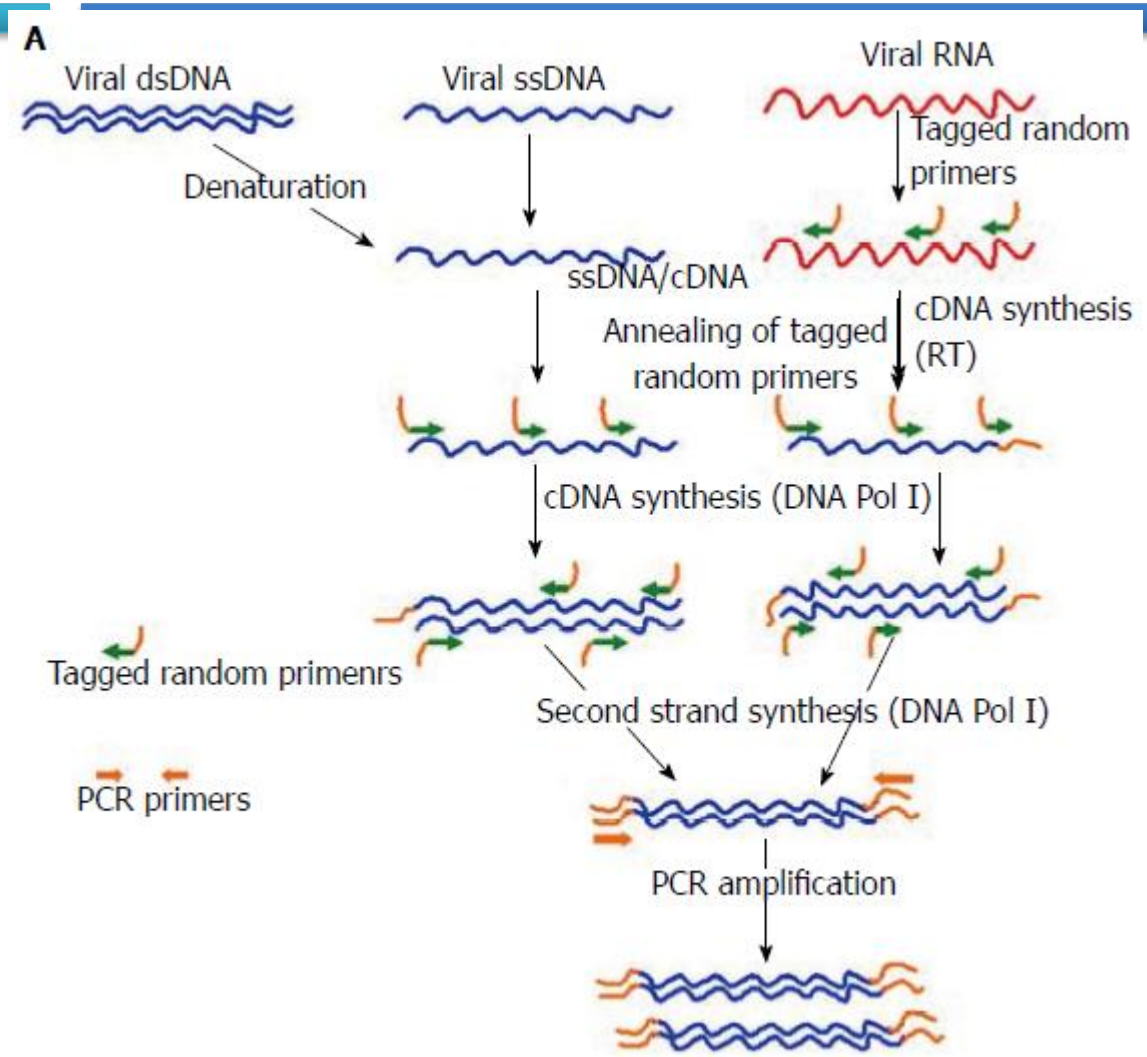
VIRUS DISCOVERY

- Nucleic acid **sequence-independent** amplification approaches
- Next-generation sequencers-based **metagenomic** approaches

RNA was reverse-transcribed using a random primer (FR26RV-N) and then cDNA was amplified using a single primer (FR20RV)

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

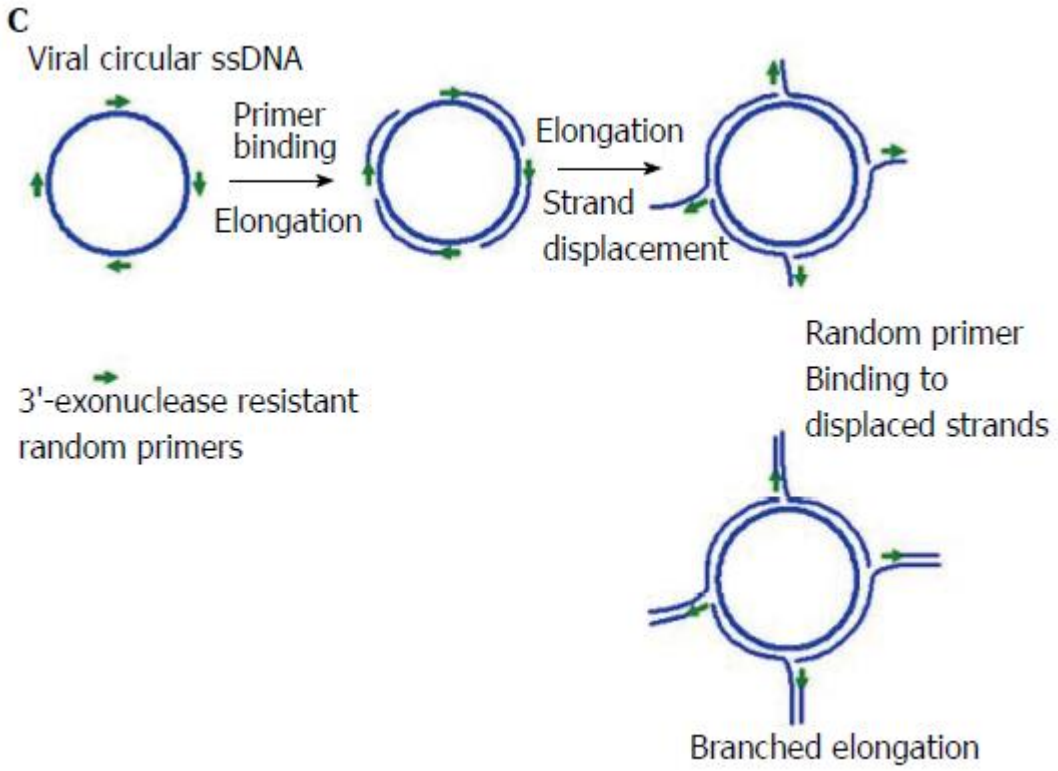
Sequence-independent single-primer amplification



Initially viral RNA and ssDNA is transcribed into complementary DNA (cDNA) using reverse transcriptase (RT) and DNA Pol I respectively, with the help of **tagged-primers having defined sequence at the 5' end while random nucleotides at the 3' end**. Subsequently, second strand synthesis is performed using DNA Pol I (Klenow) to make the cDNA double stranded (dsDNA). Now all the nuceic acids present in the reaction are dsDNA fragments have tagged sequence at their ends. Finally, **anchored dsDNA is amplified with primers annealing to the adapter specific sequences,** PCR product are checked and ready for analysis through cloning-sequencing or direct sequencing through next-generation sequencers (NGS);

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

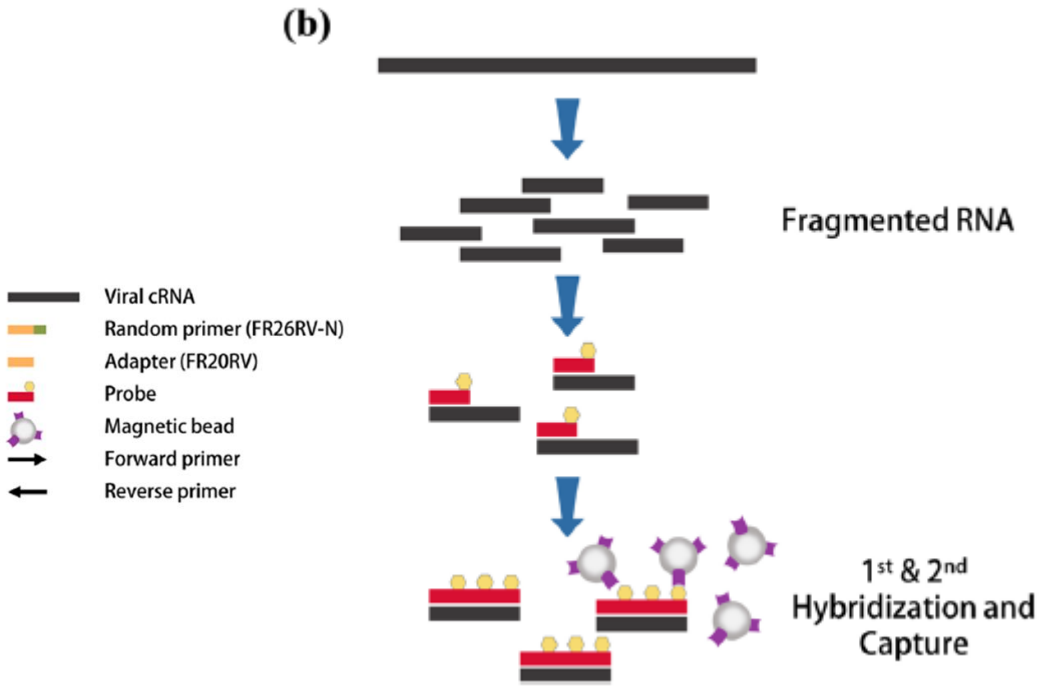
Rolling circle amplification



Amplification of multiply primed single stranded circular viral genomes. 3'-exonuclease resistant **primers randomly bind the genome and are elongated by the Phi29 polymerase**. The growing strand subsequently displaces the preceding strand of the DNA, making the strand available for binding of random primers and further elongation. This cyclic displacement and elongation leads to a highly branched structure of growing DNA, which is linear in topology.
Rolling circle amplification has the capability to specifically enrich the circular ssDNA genomes in an environment of other genetic materials, and could then be characterized by NGS.

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

Nucleic Acid Extraction: Target capture NGS

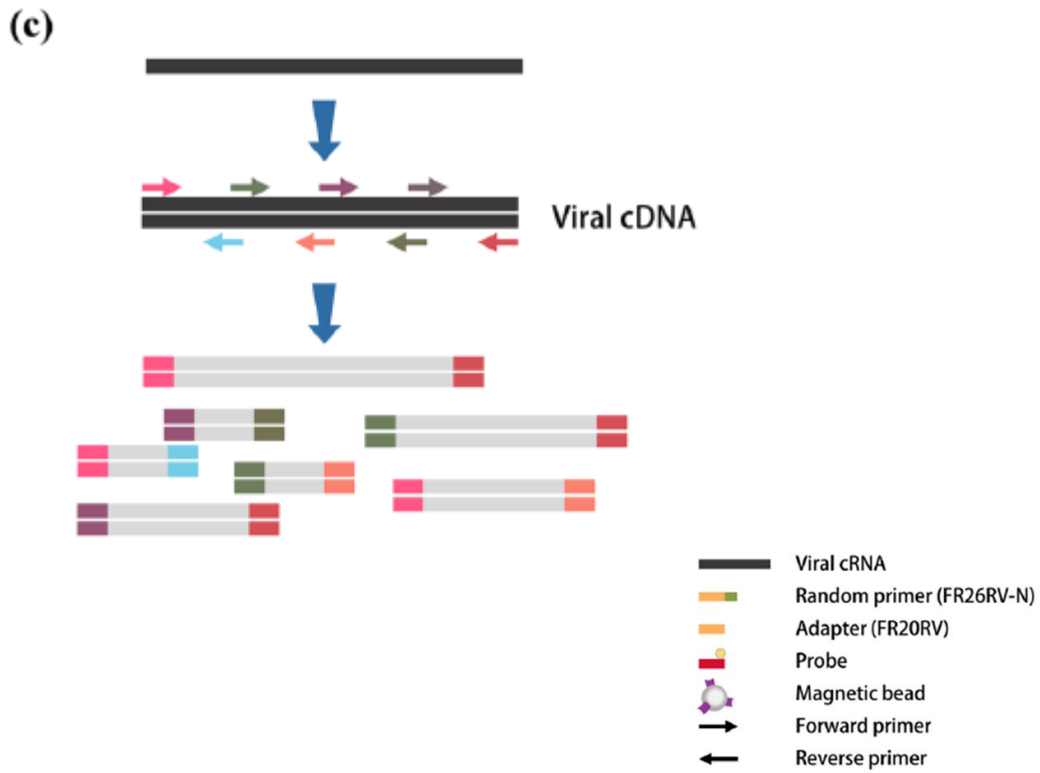


WHOLE VIRAL GENOME RECONSTRUCTION

Target capture NGS captures viral genomes by using **virus-specific probes**. RNA samples were fragmented and then cDNA was synthesized using random hexamer. To enrich the target interest, first & second hybridization and capture were performed using virus-specific probes.

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

Nucleic Acid Extraction: Amplicon NGS



WHOLE VIRAL GENOME RECONSTRUCTION

CHARACTERIZATION OF INTRA-HOST VARIABILITY

Amplicon NGS was applied to enrich viral genomes. Viral cDNA was amplified using **VIRUS-specific multiple primer mixture**

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

Nucleic Acid Extraction COMPARISON, Hantaan orthohantavirus genome sequencing

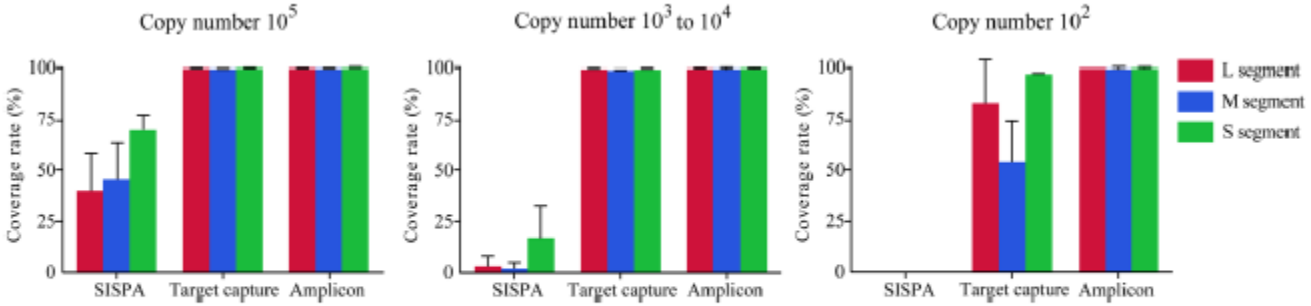


Figure 3. Coverage rate of consensus sequences for Hantaan orthohantavirus (HTNV) by viral copy number. The percentage coverage rate of consensus sequences by viral copy number among the three next-generation sequencing methods. Coverage rate was calculated by matching the consensus sequences with the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

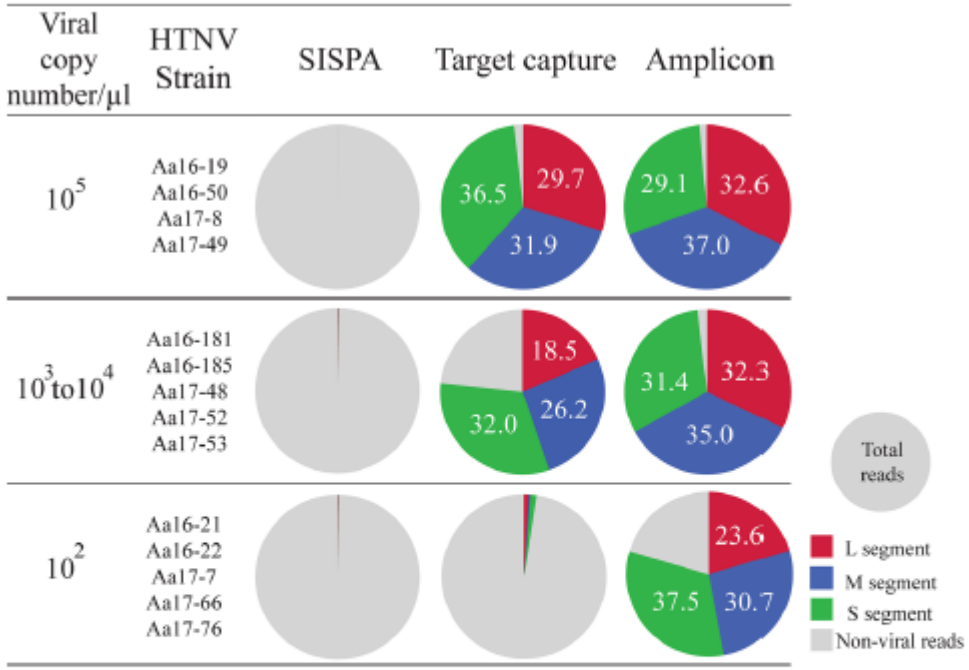
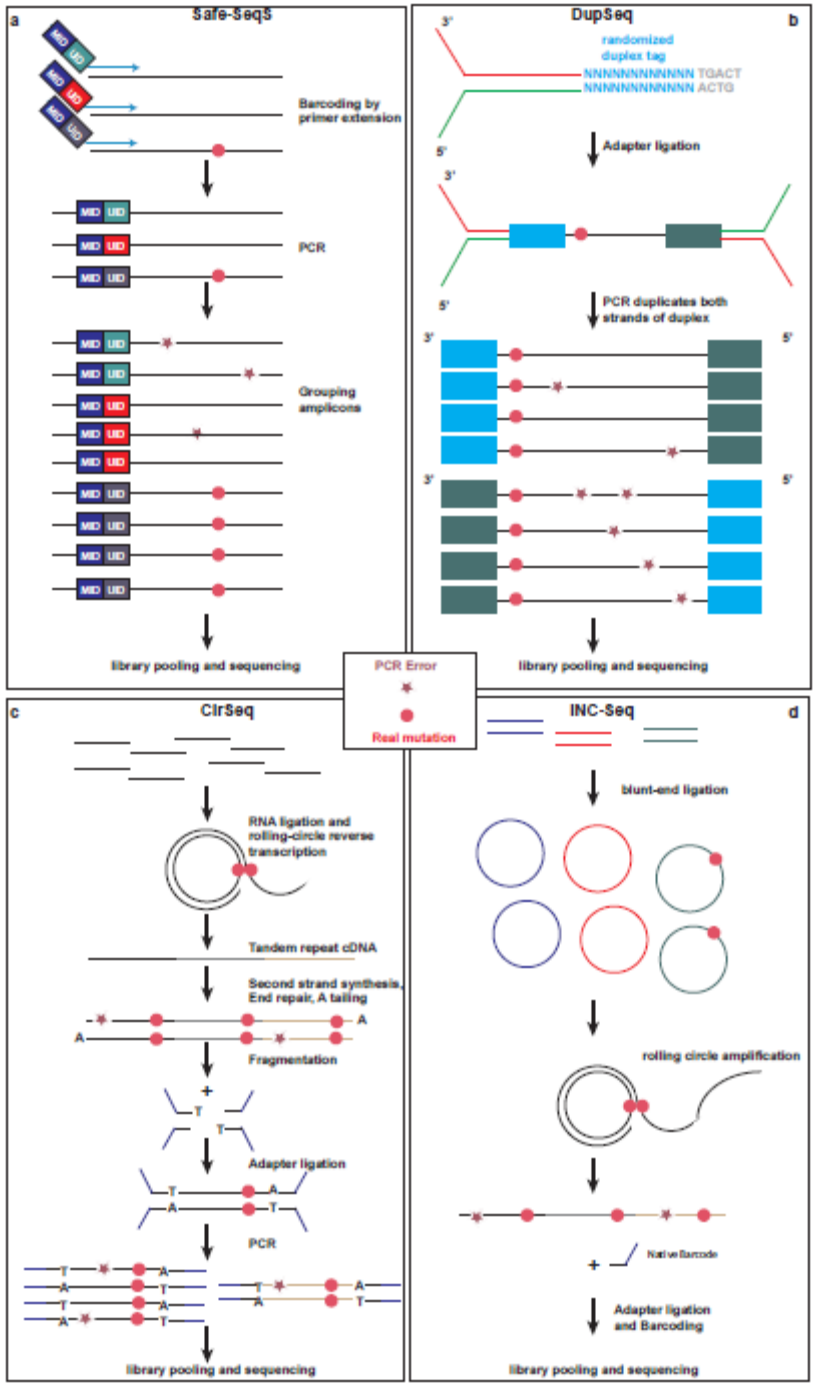


Figure 4. The composition of Hantaan orthohantavirus (HTNV) genomic reads in the total reads generated by three next-generation sequencing methods. The composition of viral reads mapped to HTNV was exhibited in the total number of reads. The total reads were produced by performing SISPA, target capture, and amplicon NGS. These reads were mapped to HTNV 76–118 tripartite genomes (GenBank accession number: L segment, NC_005222; M segment, NC_005219; S segment, NC_005218). A circle represents for total reads obtained by SISPA, target capture, and amplicon NGS methods. Red, Blue, and Green colors indicate the composition of viral reads for HTNV L, M and S segments, respectively, over the total reads. Gray color indicates non-viral reads in the total reads. The HTNV reads were shown as a percentage (%) evaluated by the ratio of viral reads over the total reads.

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

Nucleic Acid Extraction
low frequency variant identification, quasispecies



Lu et al., Virus Research 2020

Schema main steps of clinical virus Discovery by NGS

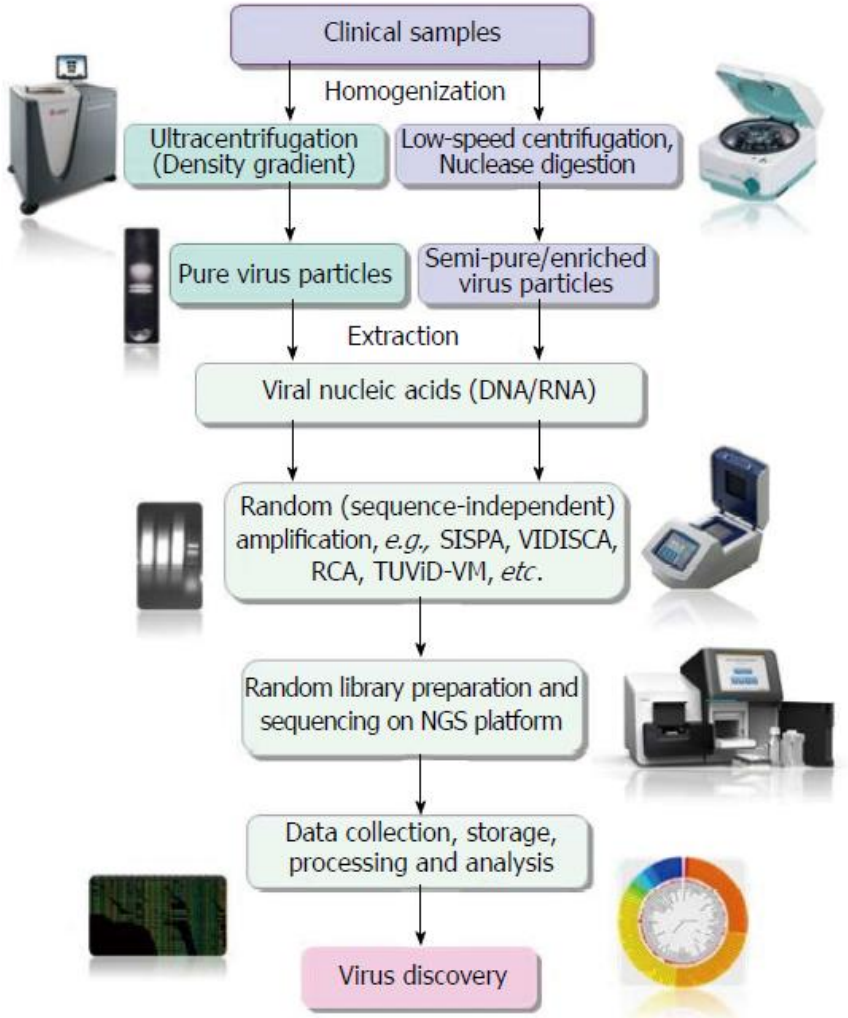


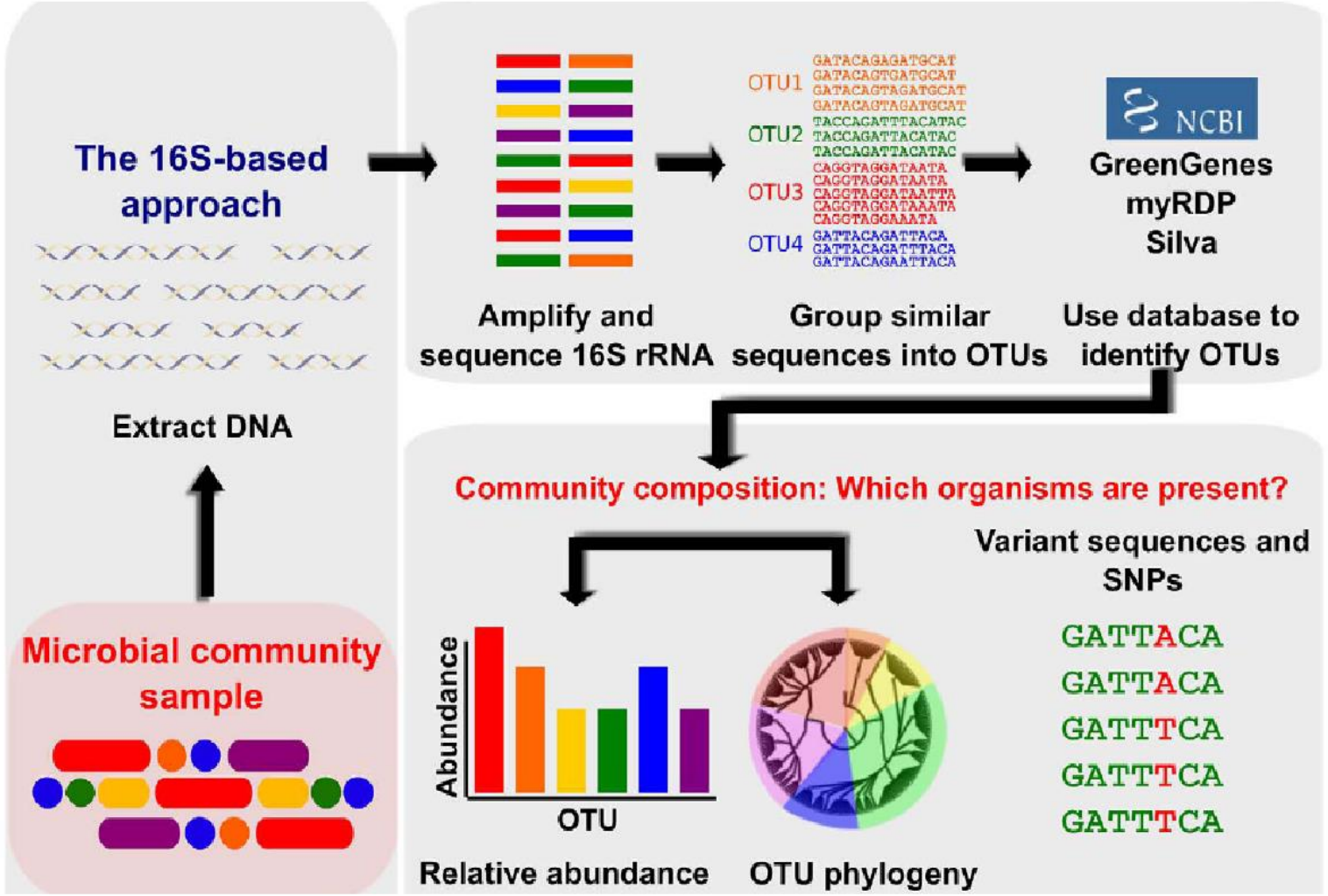
Figure 1 Diagrammatic representation of main steps of clinical virus discovery by next-generation sequencer based technologies.

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

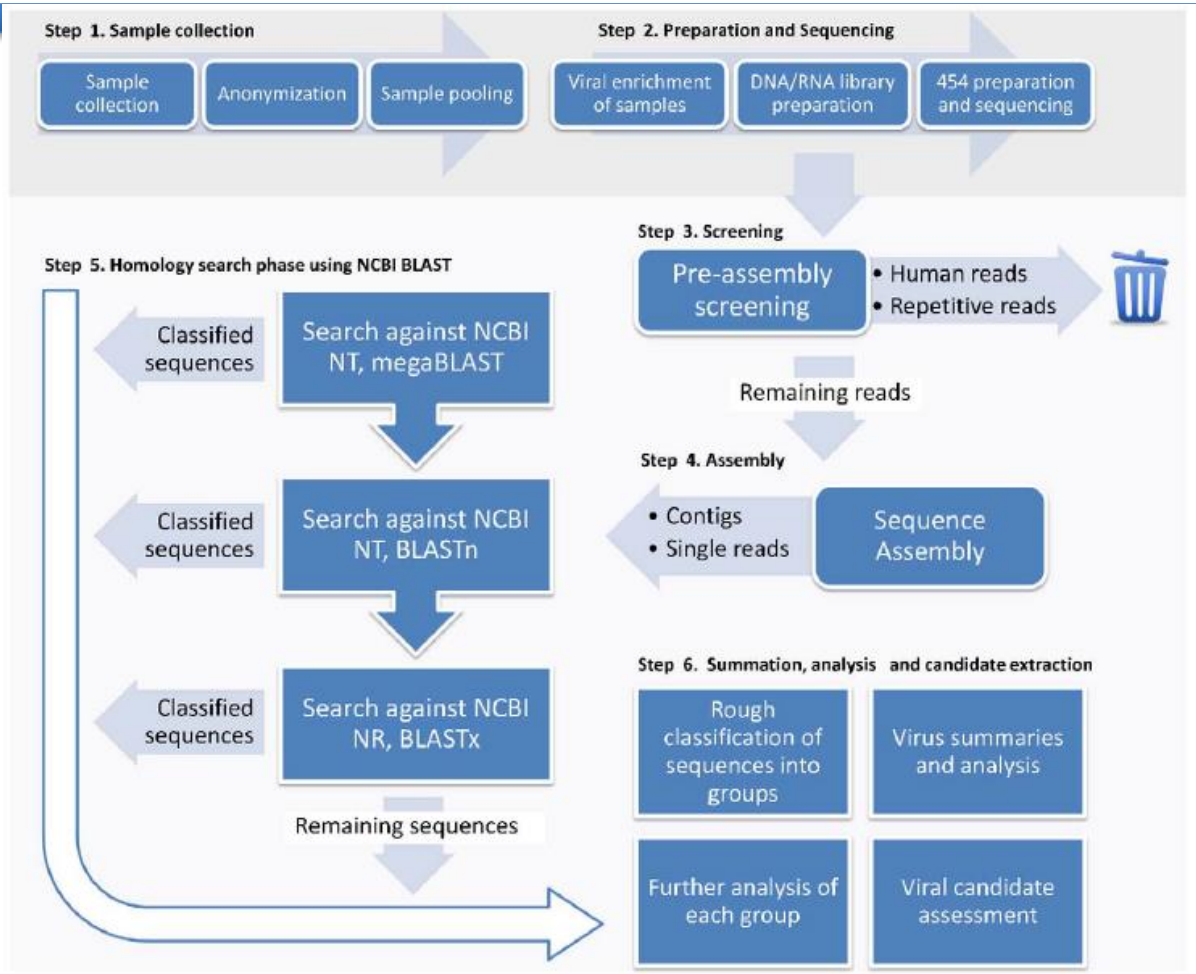
Metataxonomics vs Metagenomics (16S vs Shotgun)

	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

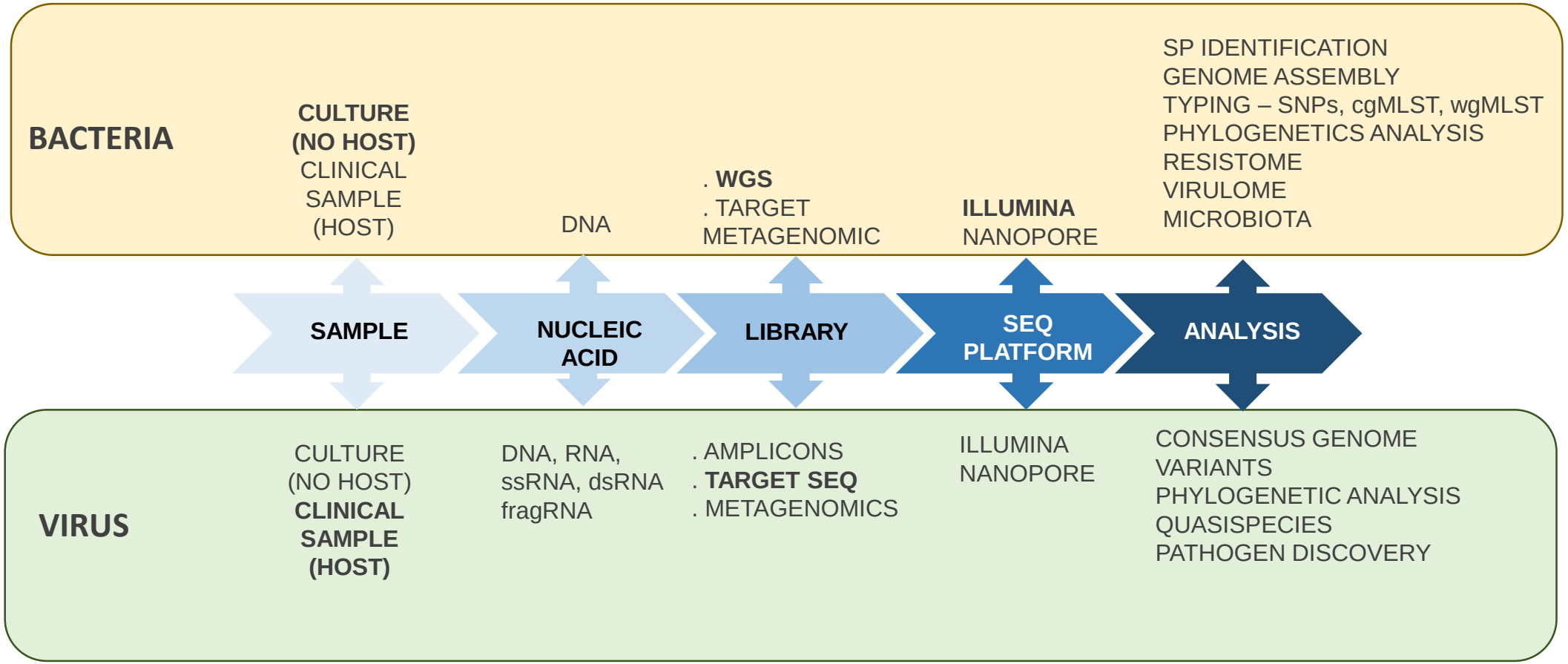


Metagenómica, pipeline de análisis

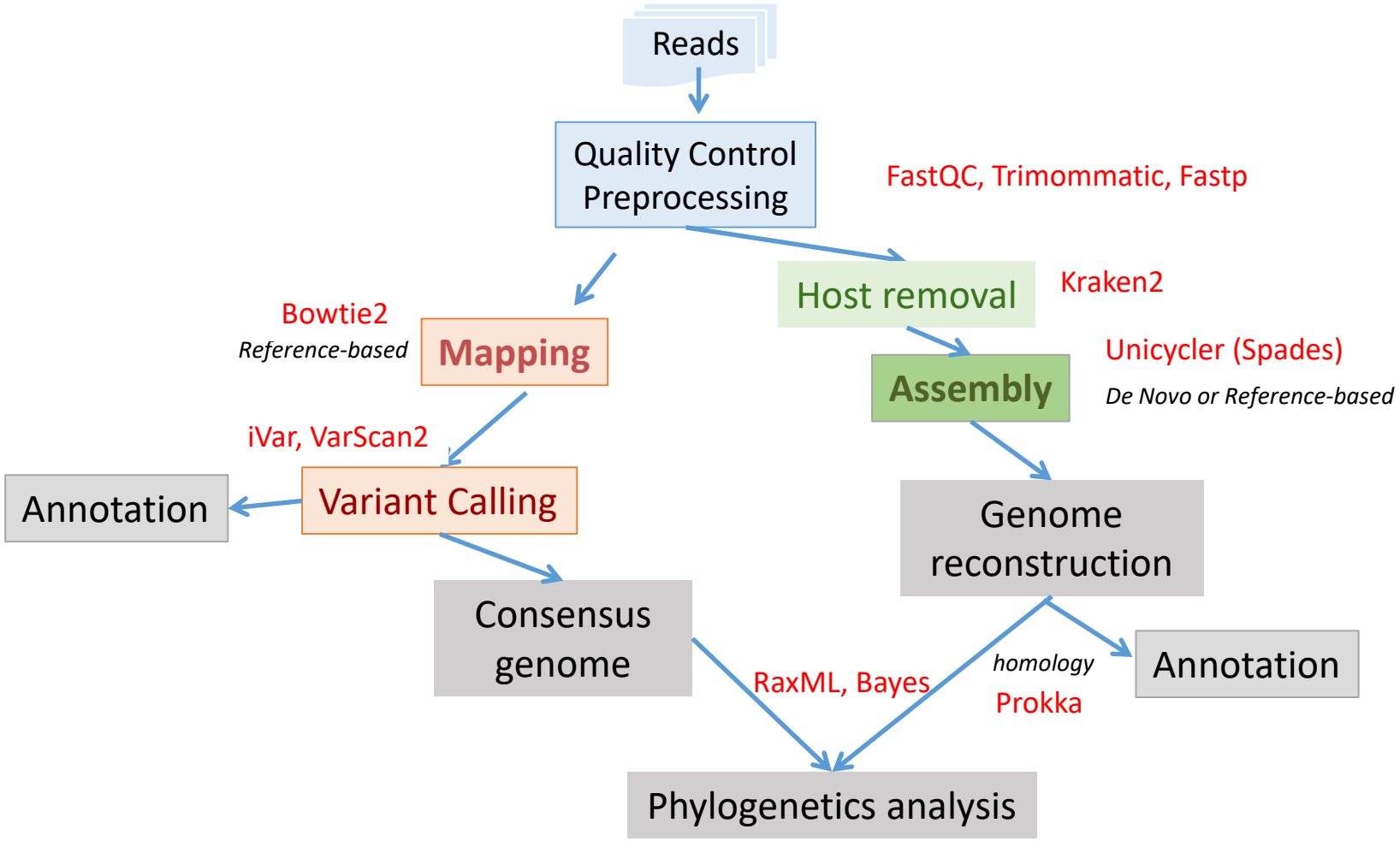


Lysholm et al., Plos One 2012:7,2, e30875

Bacterial and Viral Genome Sequencing



DATA ANALYSIS: Workflow example for virus genome analysis



Sequencing terms

Breadth of coverage

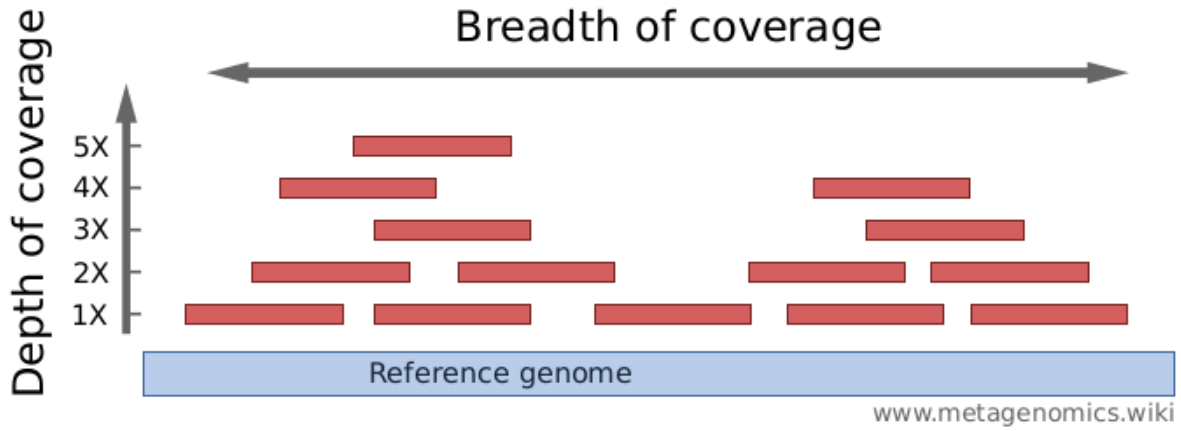
How much of a genome is "covered" by short reads? Are there regions that are not covered, even not by a single read?

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).



Depth of coverage and genome coverage

Depth of coverage

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

Genome coverage

% length sequence genome

Increase number of raw reads

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

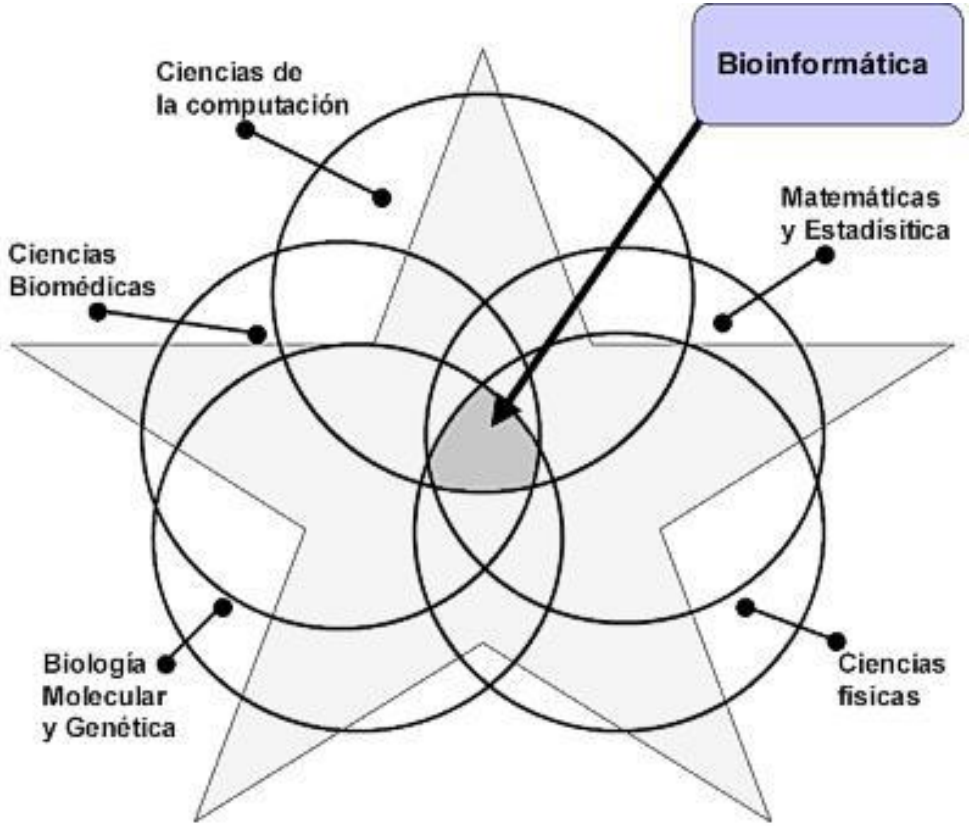
Qué es la Bioinformática?

PROBLEMAS BIOLÓGICOS

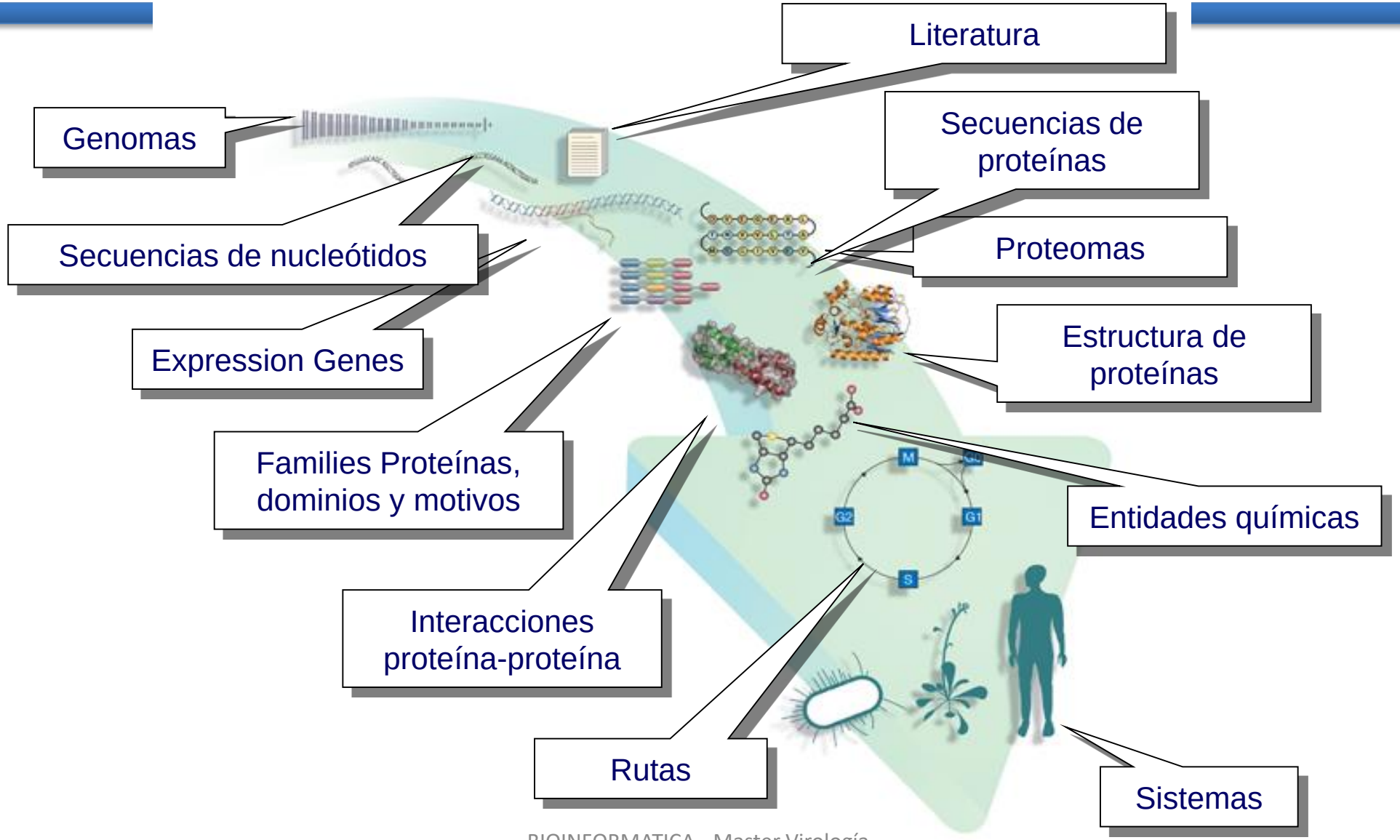
Procesamiento de datos

MÉTODOS COMPUTACIONALES

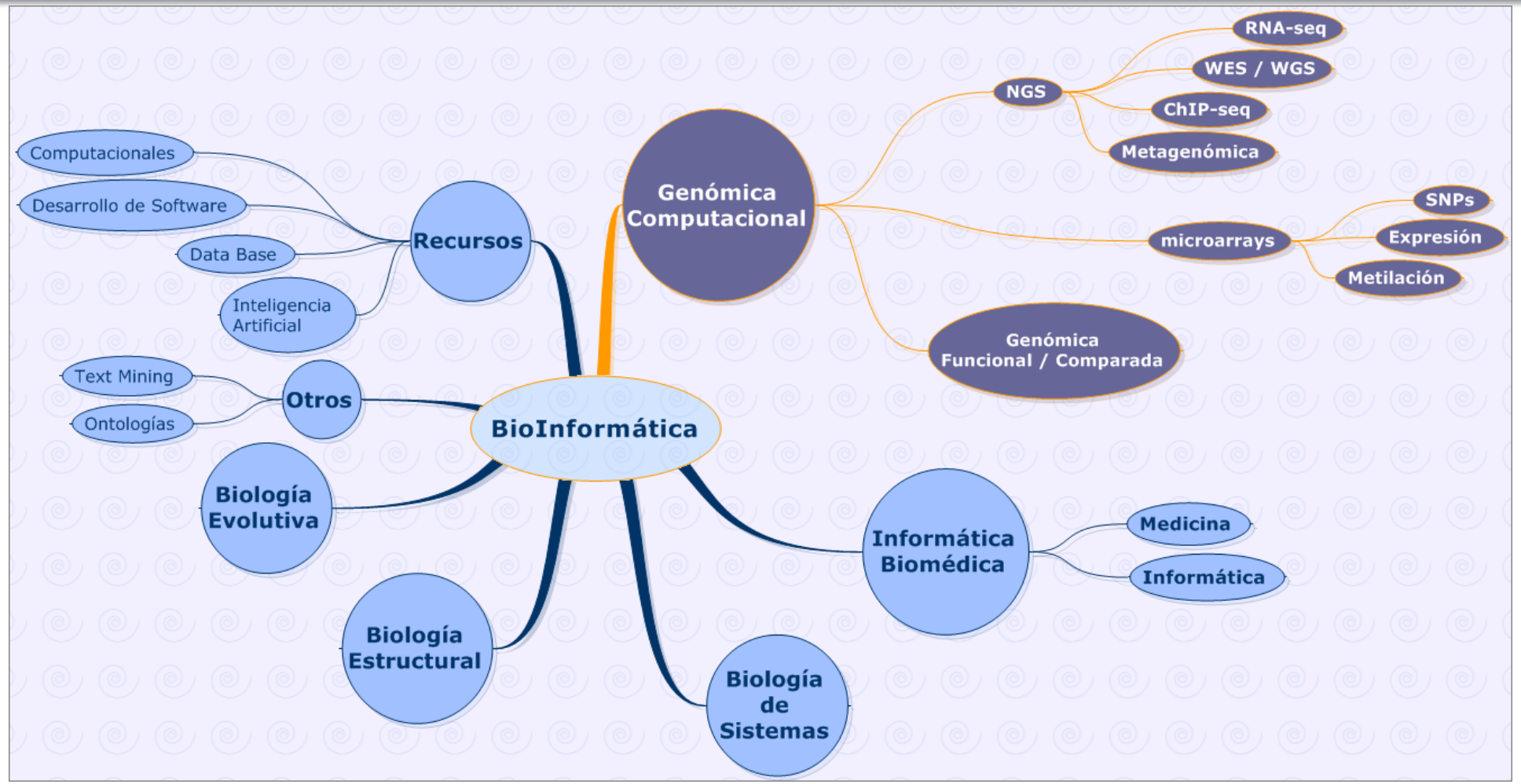
Bioinformática es multidisciplinar



Tipos de datos dan idea de la dimensión de la Bioinformática



AMBITO DE LA BIOINFORMÁTICA

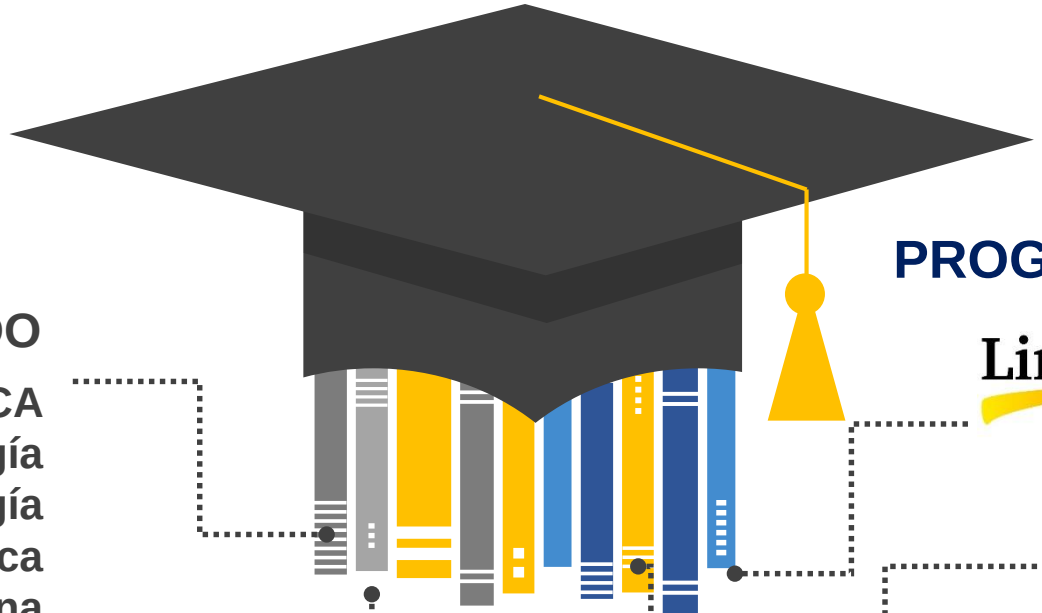


FORMACIÓN EN BIOINFORMÁTICA

Universidad Barcelona.

GRADO

BIOINFORMÁTICA
Biología
Biotecnología
Bioquímica
Medicina
Matemáticas
Química
Fisica
Informática
Telecomunicaciones



MASTER EN BIOINFORMÁTICA

PROGRAMACIÓN

Linux

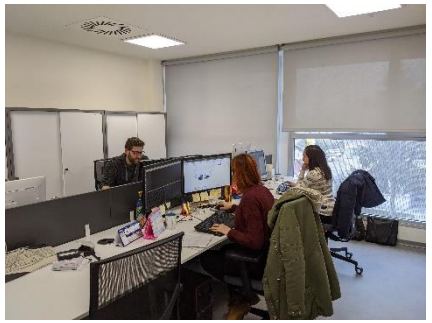
python

R

DOCTORADO

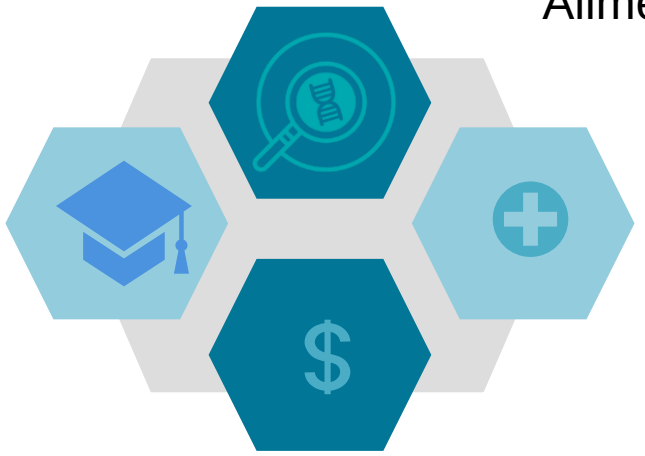
BIOLOGÍA – ANÁLISIS DE DATOS

¿Dónde trabaja un Bioinformático?



CENTRO DE INVESTIGACIÓN

Biomedicina
Agricultura
Alimentación



UNIVERSIDAD

Biociencias
Informática

HOSPITAL

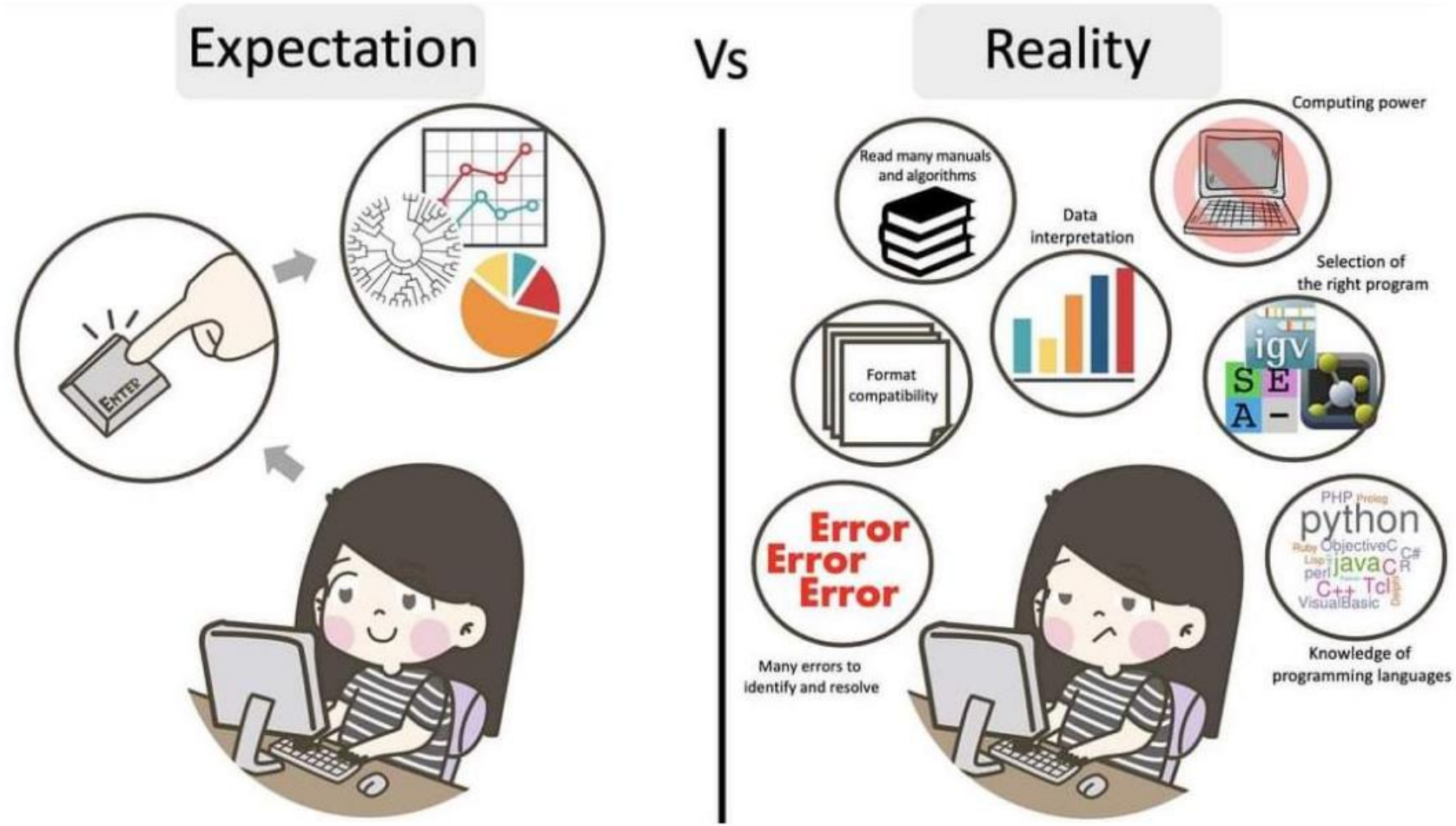
BIOINFORMÁTICO CLÍNICO
Genética
Oncología
Cardiología

EMPRESA

Bioinformática
Genética
Genómica

The truth about bioinformatics

.image-100[



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

Thanks for your attention!

Questions???