Sesión 3 - Secuenciación Masiva Aplicaciones

Isabel Cuesta

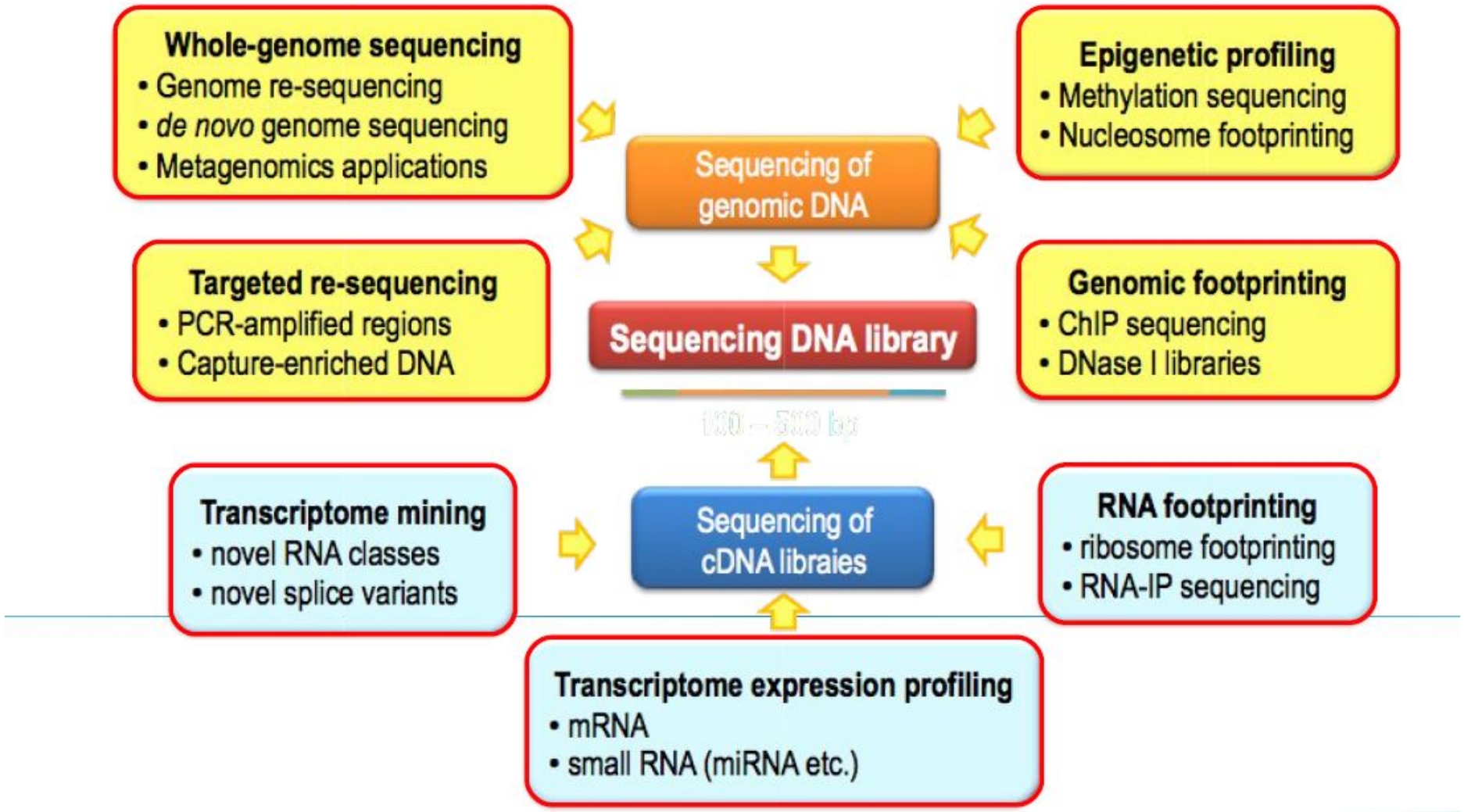
Unidad de Bioinformática (BU-ISCIII)
Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

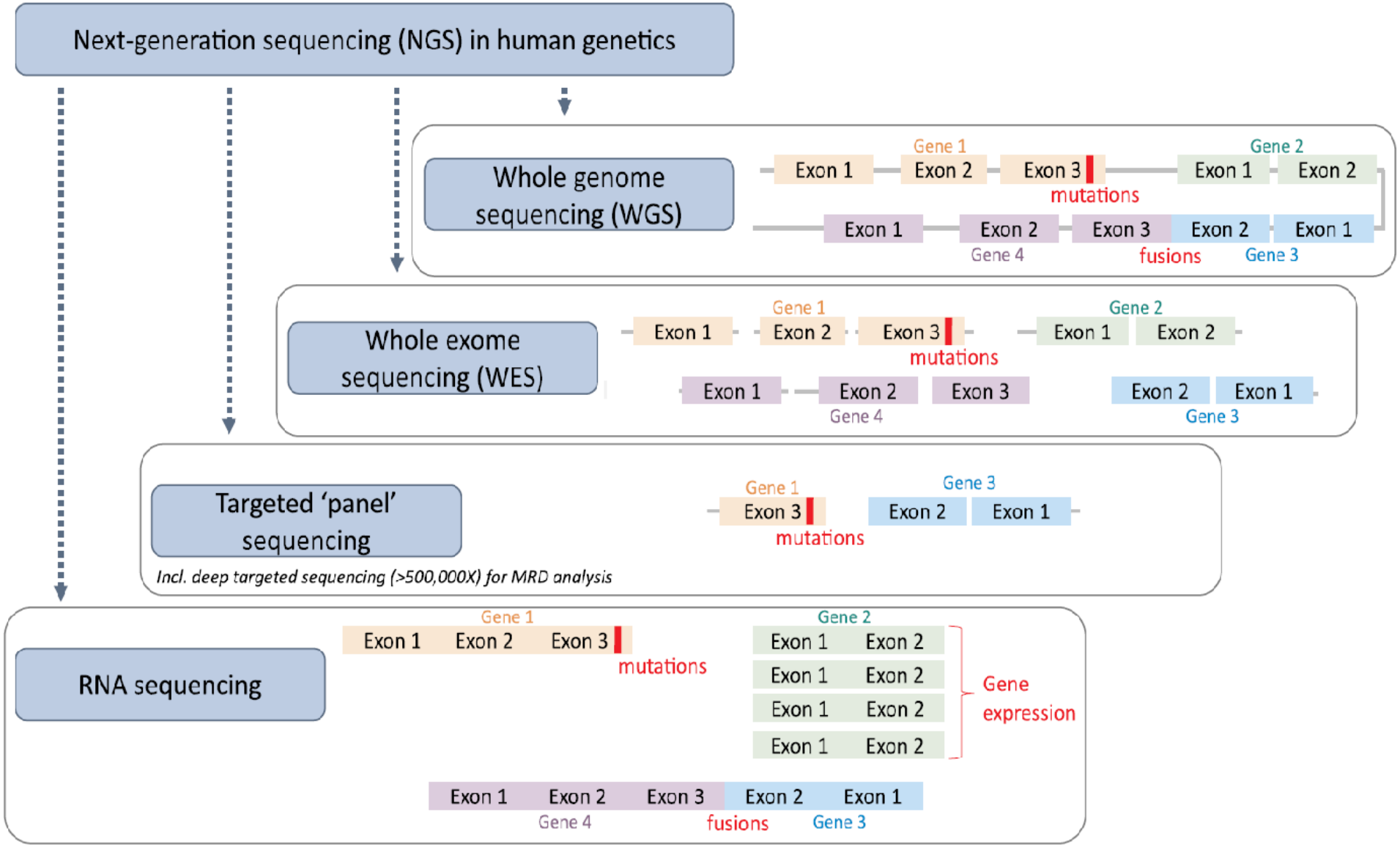
17-28 Mayo 2021, 8ª Edición
Programa Formación Continua, ISCIII

What has NGS changed?

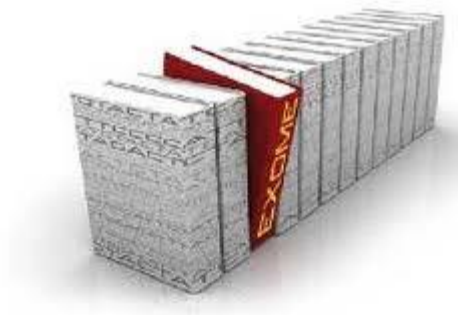
- ✓Functional genomics. Genome-Seq. Epigenetics
- ✓Molecular diagnostics. Complex diseases
- ✓Microbial Ecology. Metagenomics
- ✓Molecular Ecology. Population Genetics
- ✓Evolutionary Genomics
- ✓DNA-Protein Interactions. ChIPSeq
- ✓Pharmacogenomics
- ✓Transcriptomics. RNAseq
- ✓Systems Biology

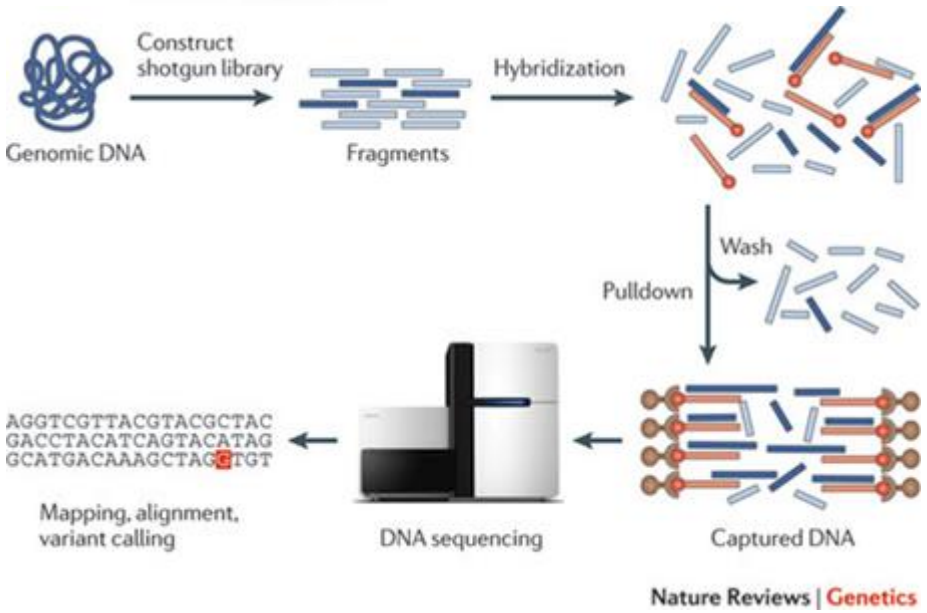
Aplicaciones de la secuenciación masiva

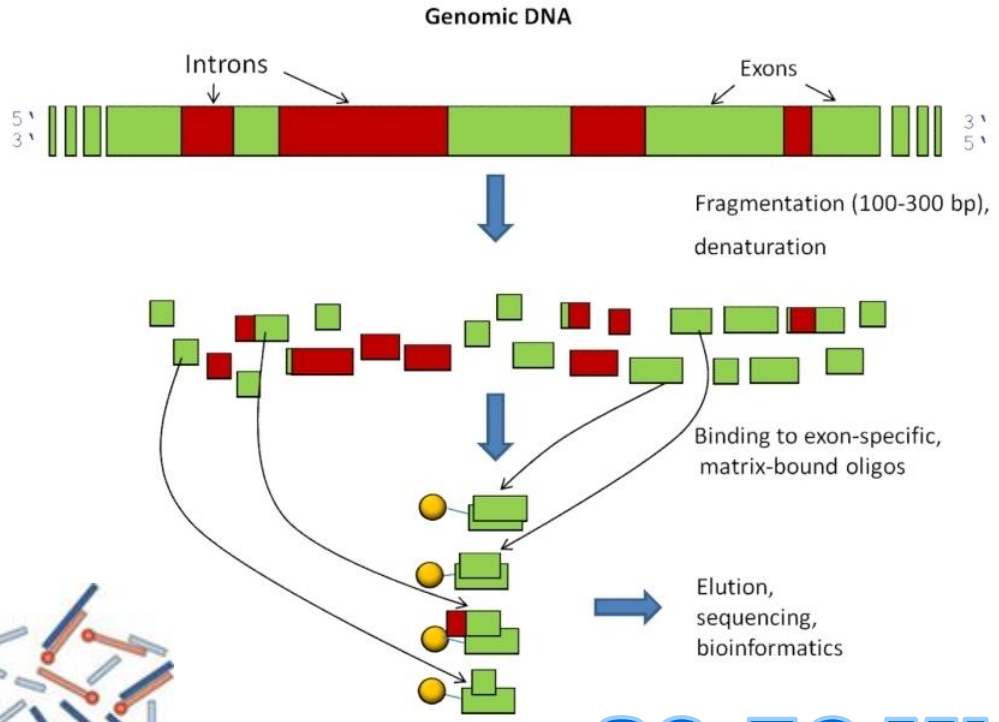



Next-generation sequencing (NGS) in human genetics
Whole genome sequencing (WGS)
Gene 1
Exon 1
Exon 2
Exon 3
mutations
Gene 2
Exon 1
Exon 2
Exon 1
Exon 2
Exon 3
Exon 2
Exon 1
Gene 4
fusions
Gene 3
Whole exome sequencing (WES)
Gene 1
Exon 1
Exon 2
Exon 3
mutations
Gene 2
Exon 1
Exon 2
Exon 1
Exon 2
Exon 3
Exon 2
Exon 1
Gene 4
Gene 3
Targeted 'panel' sequencing
Gene 1
Exon 3
mutations
Gene 3
Exon 2
Exon 1
Incl. deep targeted sequencing (>500,000X) for MRD analysis
RNA sequencing
Gene 1
Exon 1
Exon 2
Exon 3
mutations
Gene 2
Exon 1
Exon 2
Exon 1
Exon 2
Exon 1
Exon 2
Exon 1
Exon 2
Gene expression
Exon 1
Exon 2
Exon 3
Exon 2
Exon 1
Gene 4
fusions
Gene 3

EXOME







30-50 Mb

Human Exome

$500

Genoma, Exoma, Panel? desde un punto de vista clínico

PANEL

- Barato y rápido
- Util en enfermedades monogénicas
- Datos mas manejables, análisis y almacenamiento

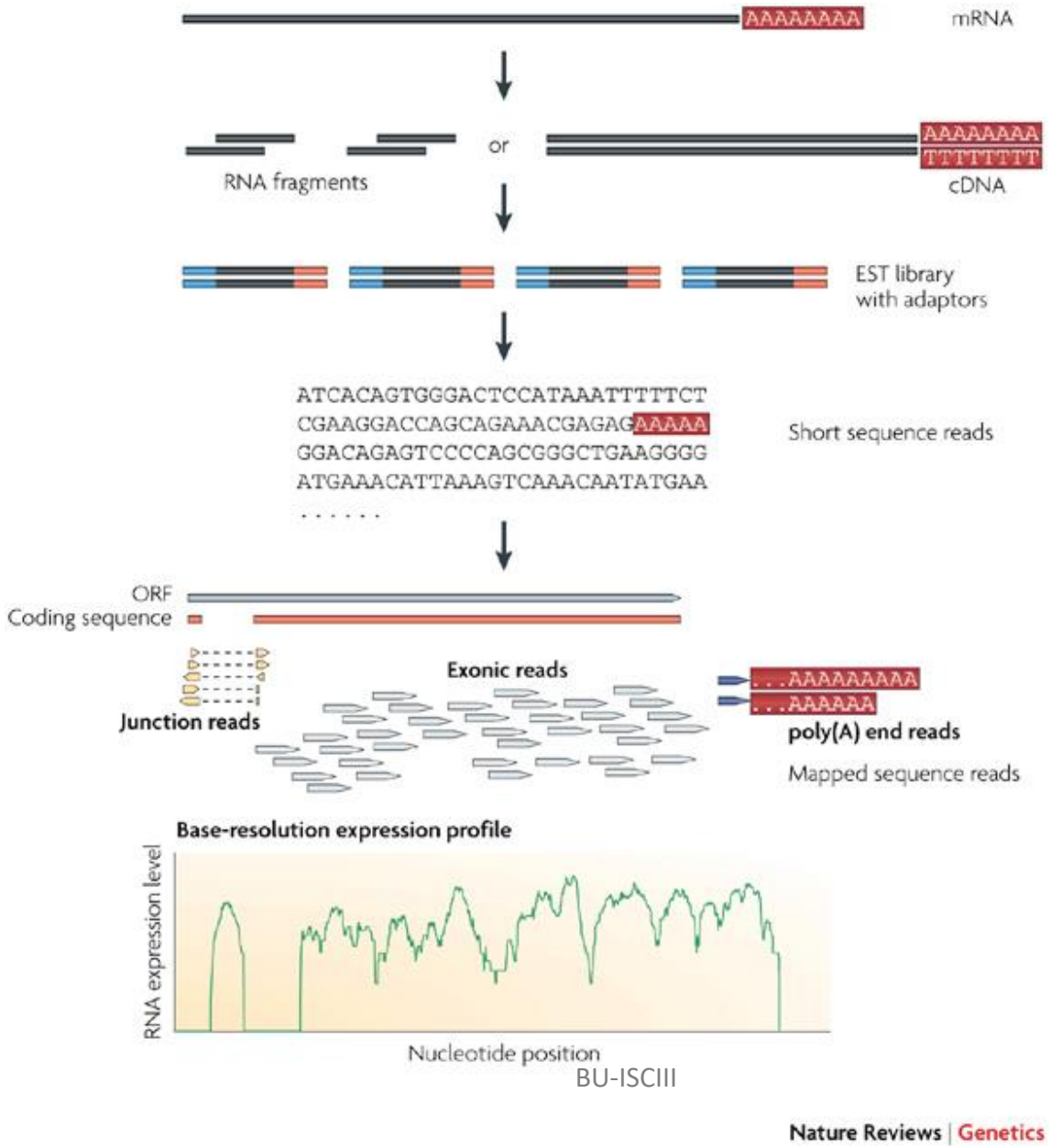
EXOMA

- Mas complejo y lento
- Necesario en enfermedades complejas
- Análisis mas complejo
- Mayor volumen de datos

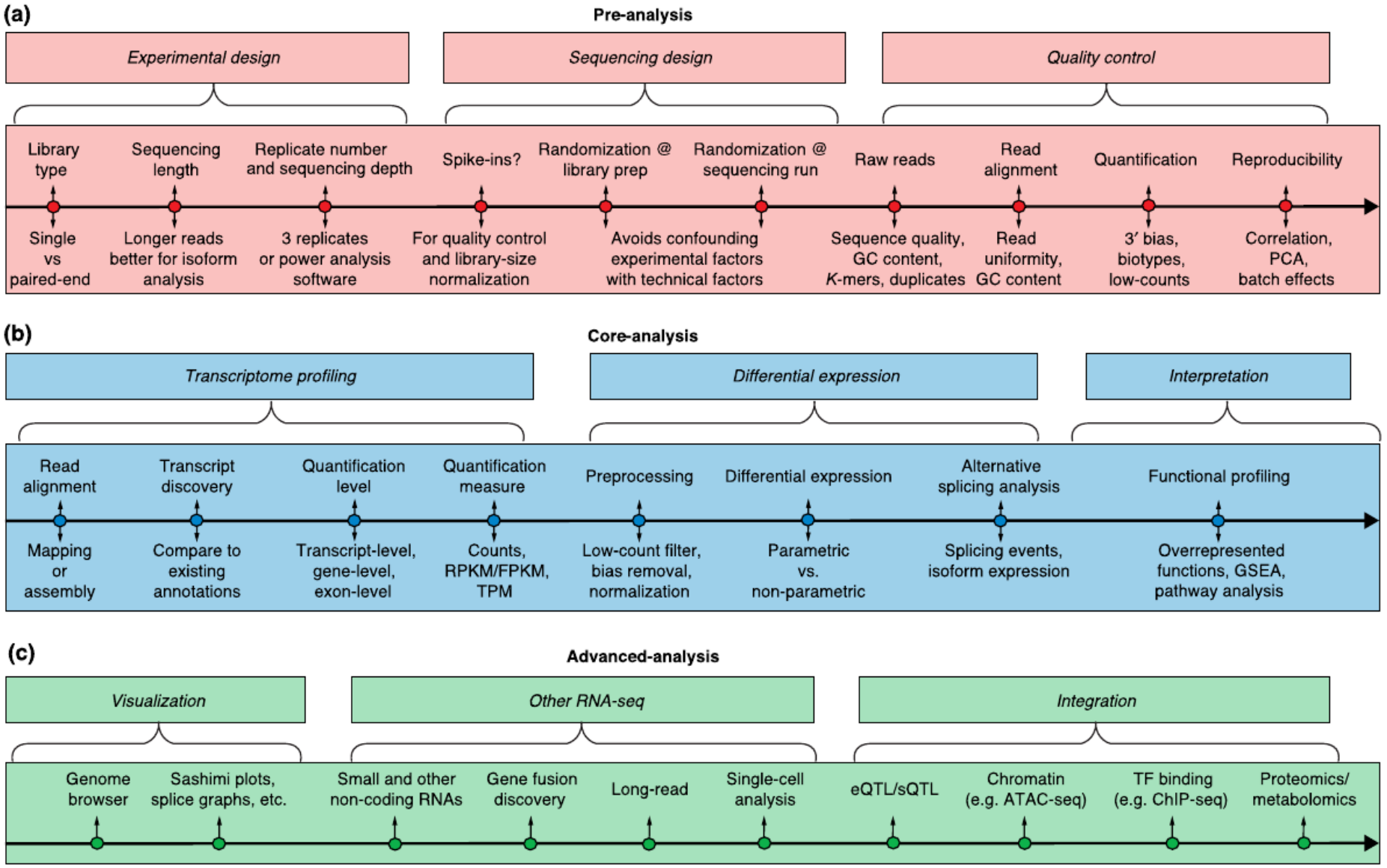
GENOMA

- Maxima complejidad en secuenciación y coste
- Información de regiones no codificantes
- Análisis de variaciones estructurales
- Elevado volumen de datos

RNA seq

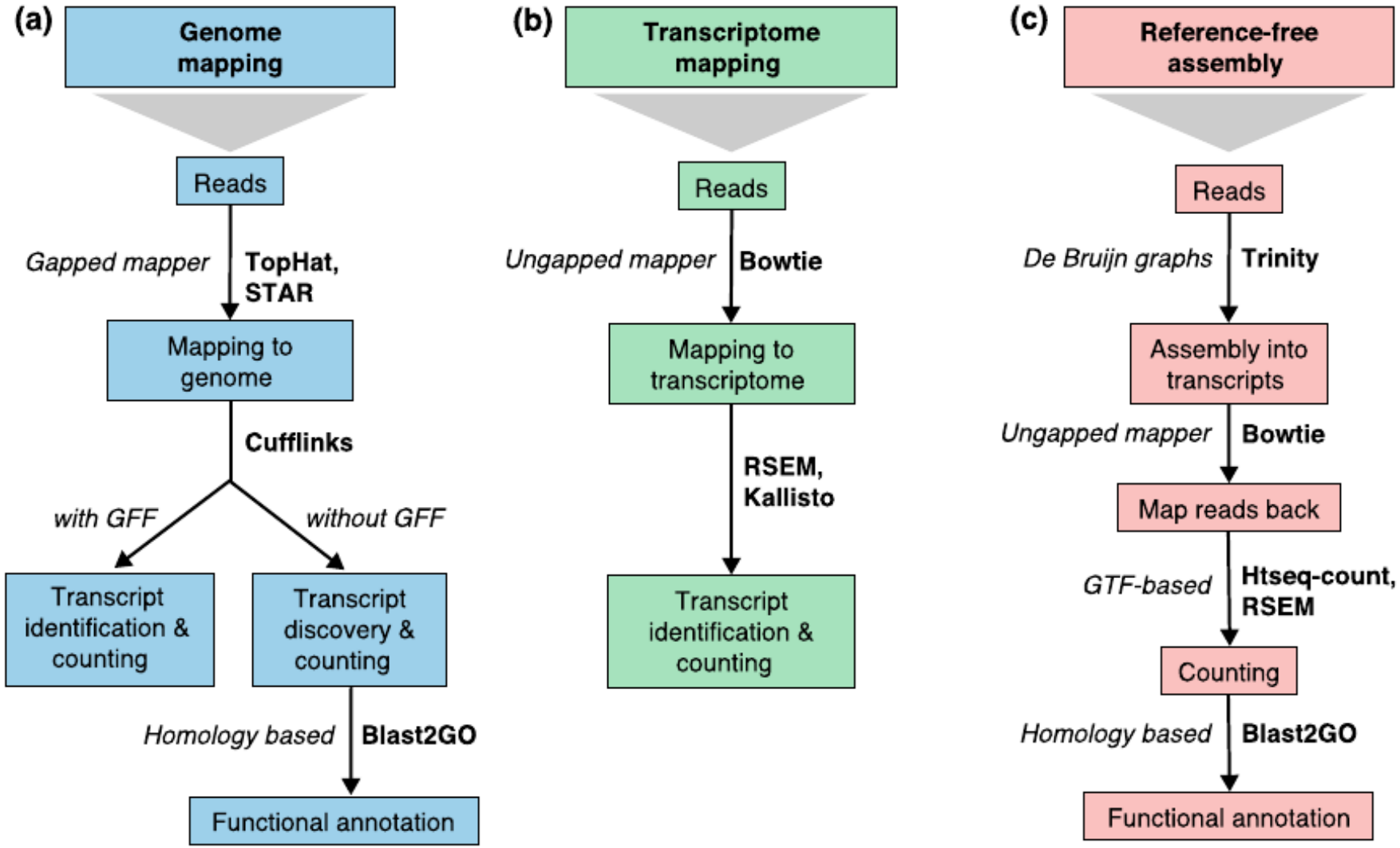

AAAAAAAA
mRNA
or
RNA fragments
AAAAAAAA
TTTTTTTT
cDNA
EST library
with adaptors
ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA
. . . . . .
Short sequence reads
ORF
Coding sequence
Exonic reads
Junction reads
. . .AAAAAAAAA
. . .AAAAAA
poly(A) end reads
Mapped sequence reads
Base-resolution expression profile
RNA expression level
Nucleotide position
Nature Reviews | Genetics

RNA seq



Conesa et al., Genome Biology (2016) 17:13

RNA seq: transcript identification strategies

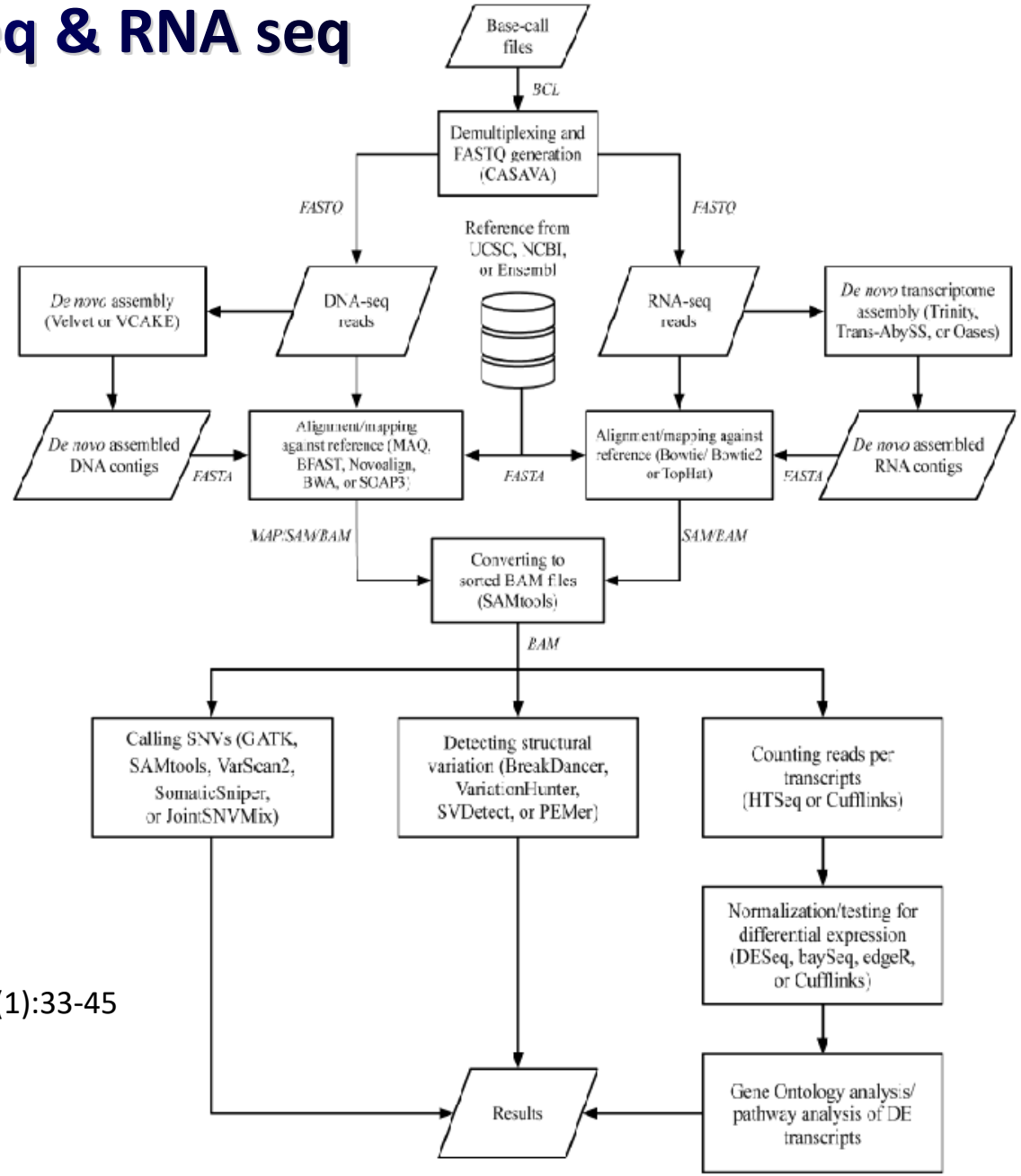


Conesa et al., Genome Biology (2016) 17:13

RNA seq, ventajas

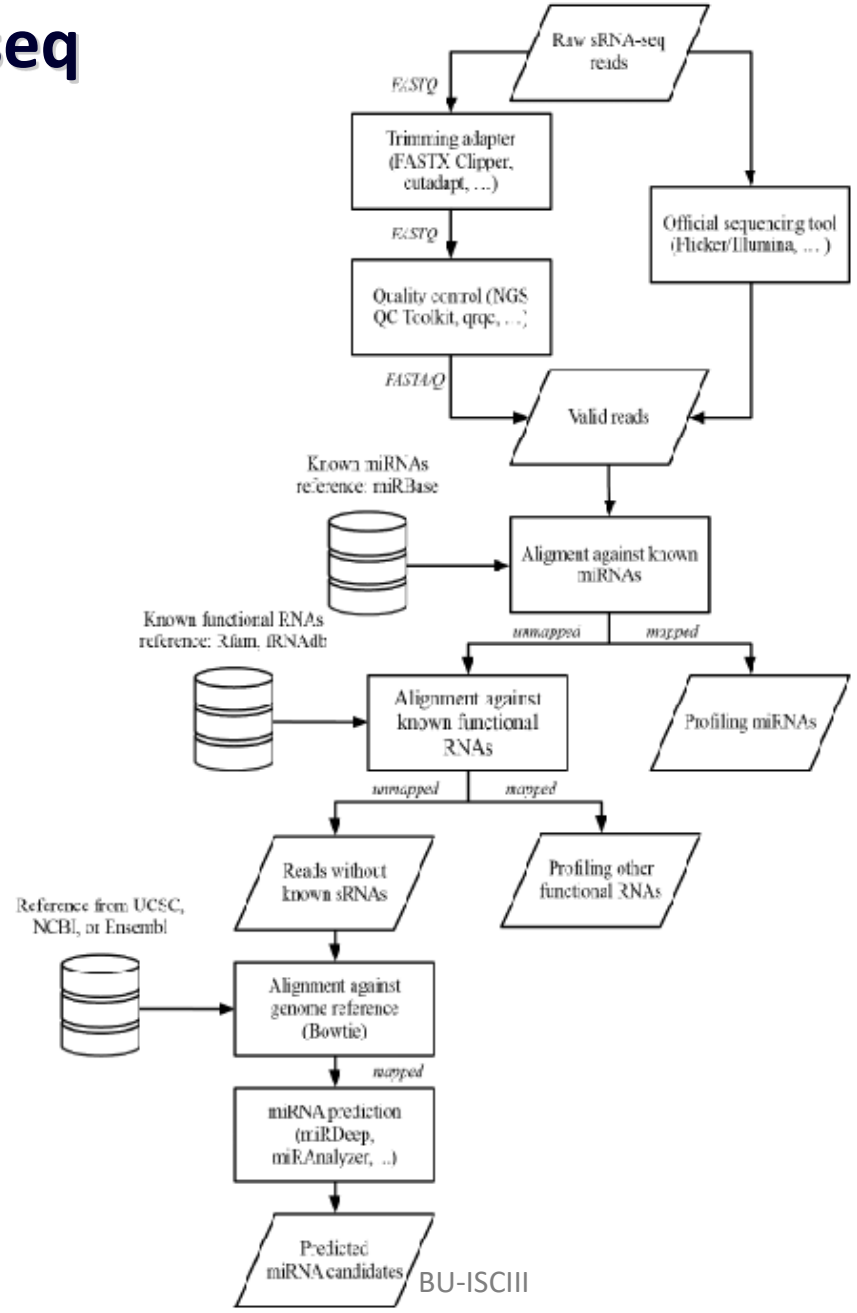
- A diferencia de microarrays no requiere conocimiento de la secuencia del genoma
- Bajo nivel de ruido (las lecturas se mapan correctamente)
- Cuantificación de los transcritos
- Identificación de nuevos transcritos
- Identificación de variaciones (SNPs)
- Disminución progresiva del coste
- Pipelines de análisis disponibles para organismos eucariotas superiores

Strategies of DNA-seq & RNA seq



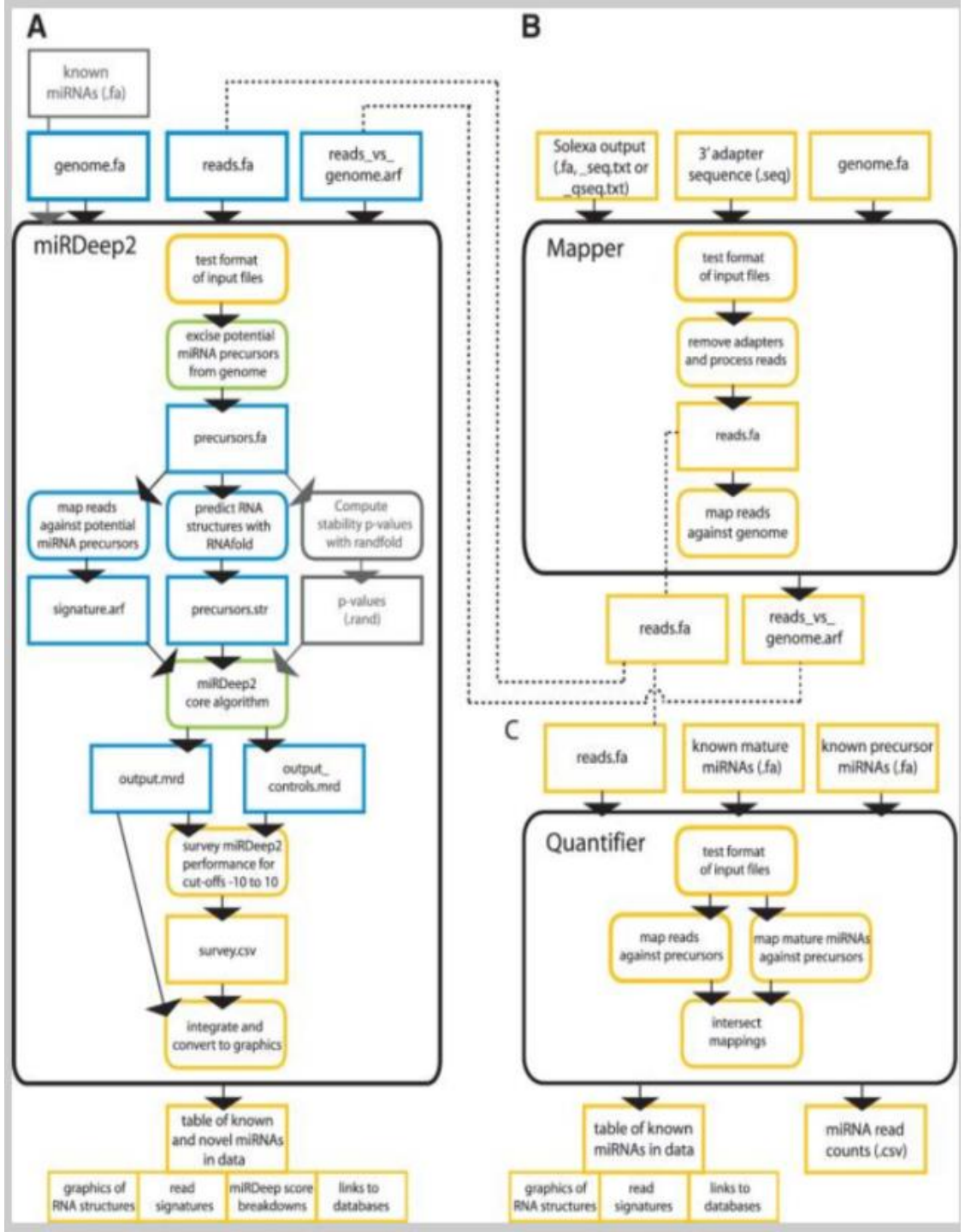
Lee et al., Transl Cancer Res 2013;2(1):33-45

Strategies of small RNA seq



Lee et al., Transl Cancer Res 2013;2(1):33-45

miRNAseq analysis



Applications in Human Genetics

- Identification of polymorphisms –SNVs (SNPs, indels)
- Identification of structural variants –SVs (1kb-3Mb) (indels, CNVs, inversions, translocations)
- Gene expression profiling analysis, differential expression – mRNA, SNVs
- Identification and quantification of : smallRNA, miRNA, lnRNA
- De novo transcriptome assembly
- Chip-Seq
- Metil-Seq
- Variant effect annotation

……

Clinical Bioinformatics - Precision Medicine

SpainUDP: "Spanish Undiagnosed Rare Diseases Program"

International Journal of *Environmental Research and Public Health*

MDPI

Article

SpainUDP: The Spanish Undiagnosed Rare Diseases Program

Estrella López-Martín [1,*], Beatriz Martínez-Delgado [2], Eva Bermejo-Sánchez [1], Javier Alonso [2], the SpainUDP Network [†] and Manuel Posada [1]

1 Institute of Rare Diseases Research (IIER) & Centre for Biomedical Network Research on Rare Diseases (CIBERER), Instituto de Salud Carlos III, 28029 Madrid, Spain; eva.bermejo@isciii.es (E.B.-S.); mposada@isciii.es (M.P.)

2 Institute of Rare Diseases Research (IIER), Instituto de Salud Carlos III, 28029 Madrid, Spain; bmartinezd@isciii.es (B.M.-D.); fjalonso@isciii.es (J.A.)

* Correspondence: elopez@isciii.es; Tel.: +34-91-822-2911

† SpainUDP (Spain) is a member of UDNI (Undiagnosed Diseases Network International) (network's members who are co-authors are listed in Appendix A), 28029 Madrid, Spain.

Received: 12 June 2018; Accepted: 7 August 2018; Published: 14 August 2018

check for updates

SpainUDP

RD-Connect

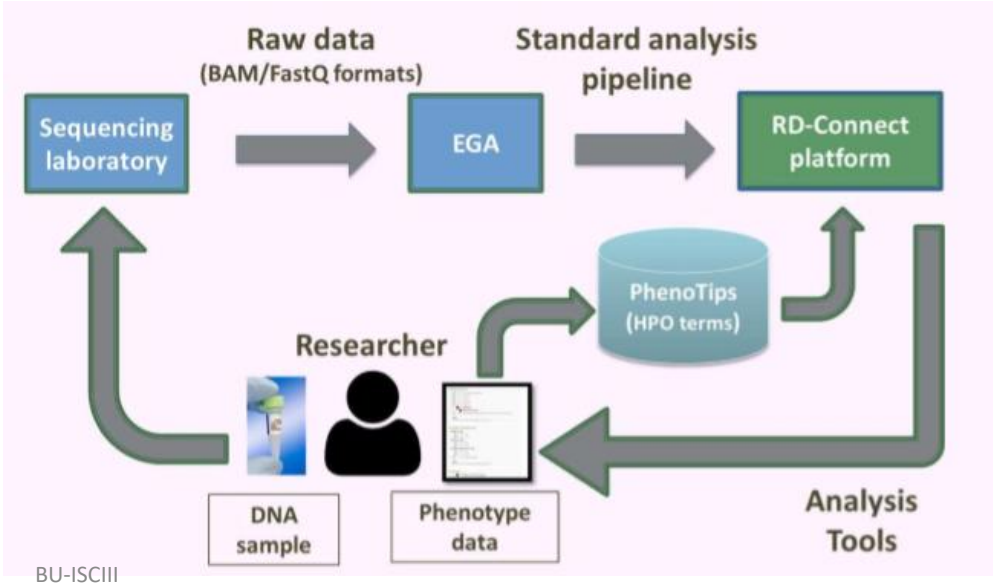
EUROPEAN JOINT PROGRAMME RARE DISEASES

SolveRD

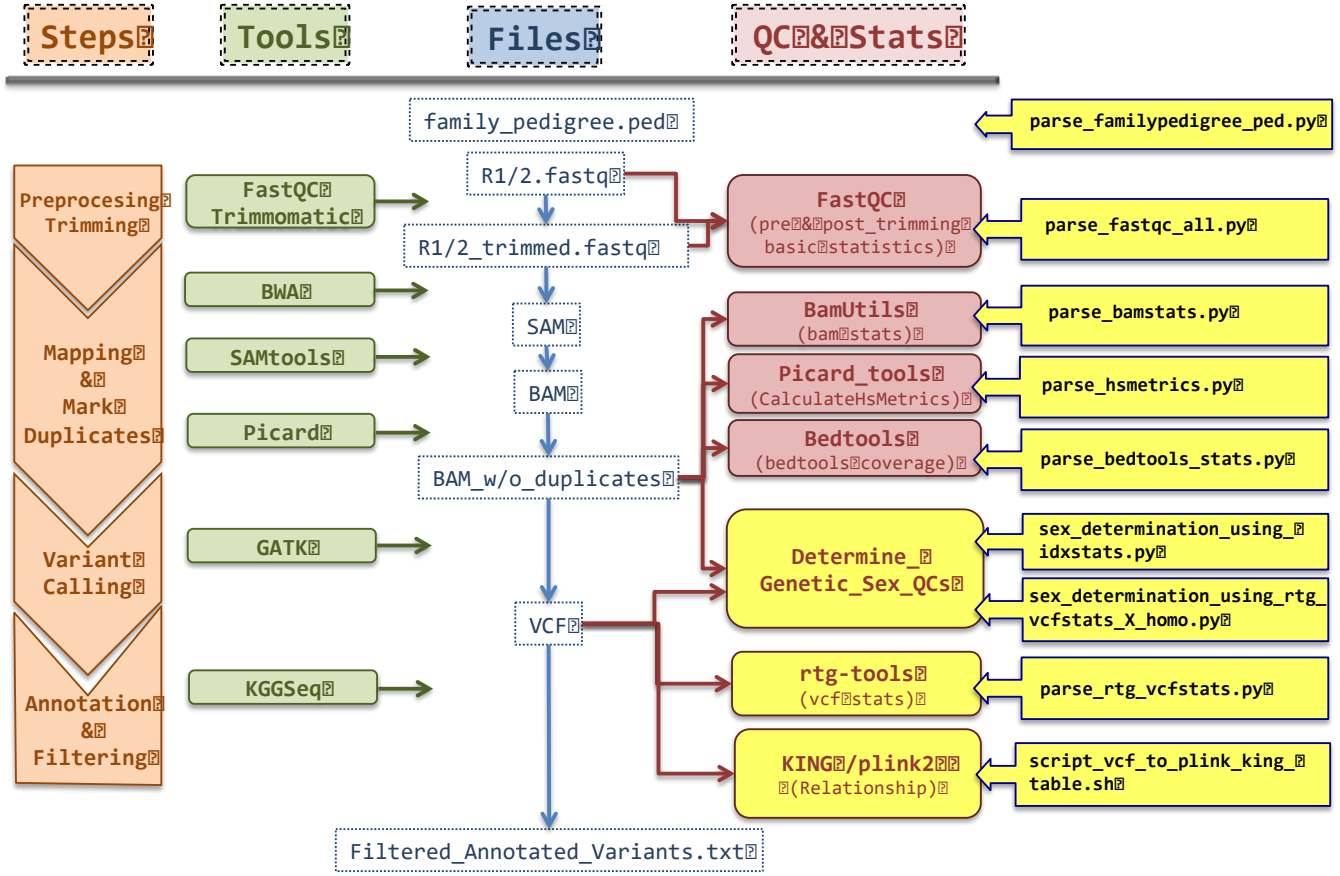
SpainUDP, Manuel Posada

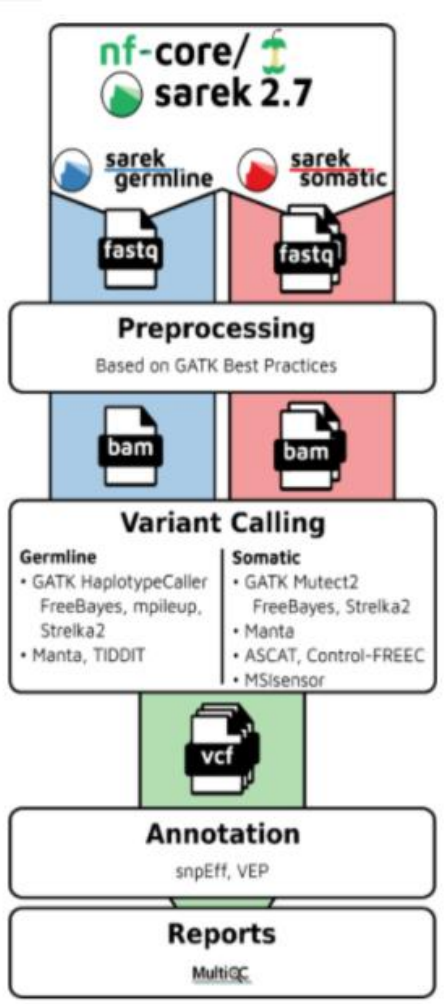
- El programa SpainUDP - IIER se creo con el objetivo de intentar obtener un diagnóstico clínico en pacientes con enfermedades raras sin diagnosticar. Se basa en tres pilares: Estudio individualizado y exhaustivo de cada caso; Uso de técnicas de NGS; Compartir datos a nivel internacional (RD-Connect)

- Se han analizado 120 WES, llegando a un 50% de resolución de casos. Se incorporo en 2020 WGS y RNAseq.

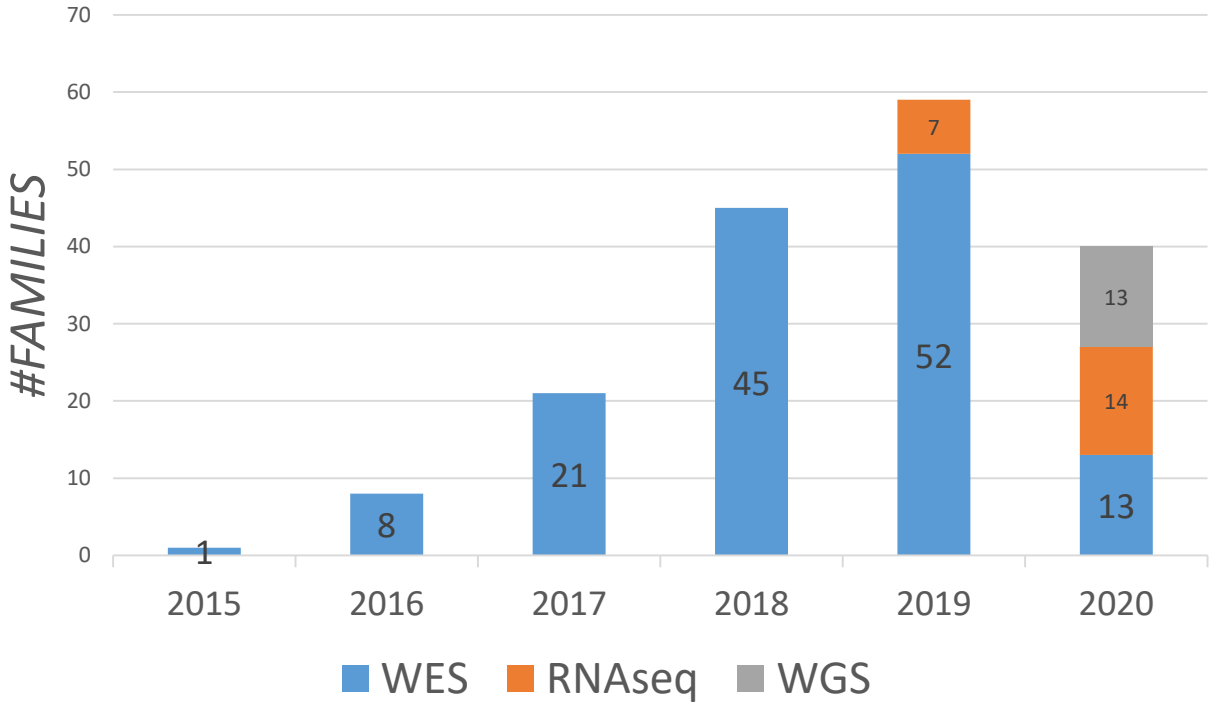


Data analysis: Pipeline





SpainUDP: number of families analyzed

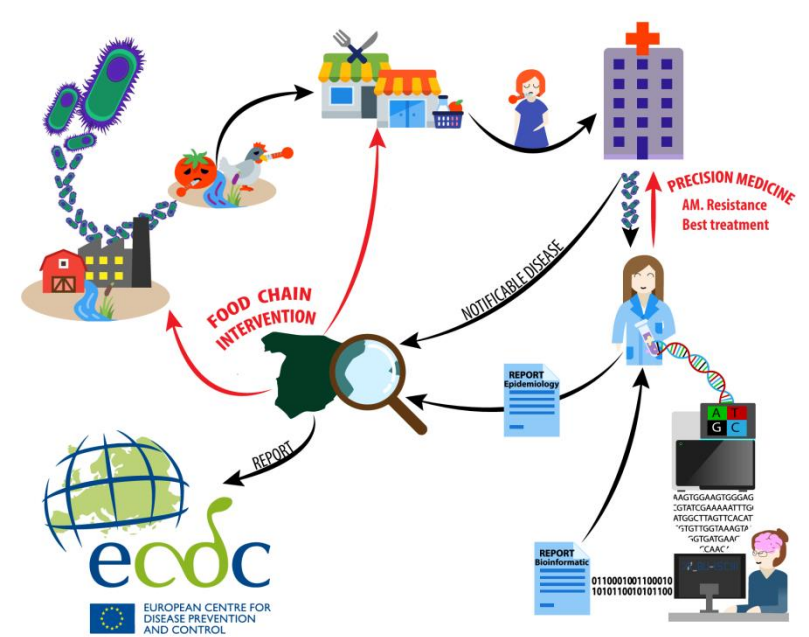


Applications Microbial Genomics

Research - Clinical Bioinformatics - Precision Medicine

CNM

Pathogen associated



AESI 2017-2019 BU-ISCIII – Genómica

AESI 2019-2021 BU-ISCIII - Genómica

AESI 2018 – 2021 **PLATAFORMA DE BIOINFORMATICA ISCIII-TransBioNet**

METAGENOMICS EQAE
Special Pathogens Unit,
P. Anda, R. Escudero, I. Jado

EMERGE
Efficient response to highly dangerous and emerging pathogens at EU level

GMI – HTS Standards, Databases Sharing and Guidelines

Global Microbial Identifier

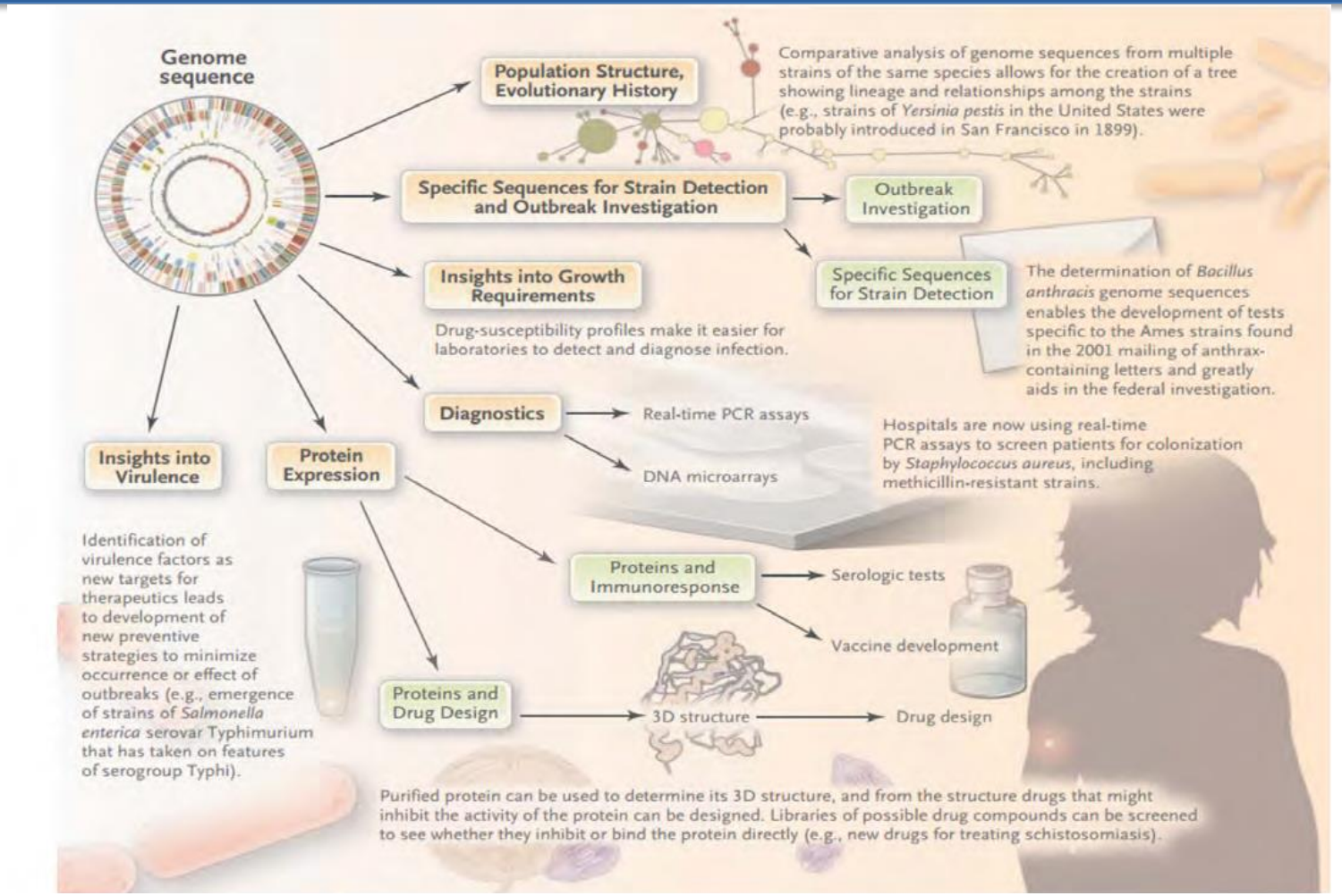
GMI – UNSGM PT for detection of biological threats by genomic analysis – **AESI 2019**

COMPARE Food Metage

compare

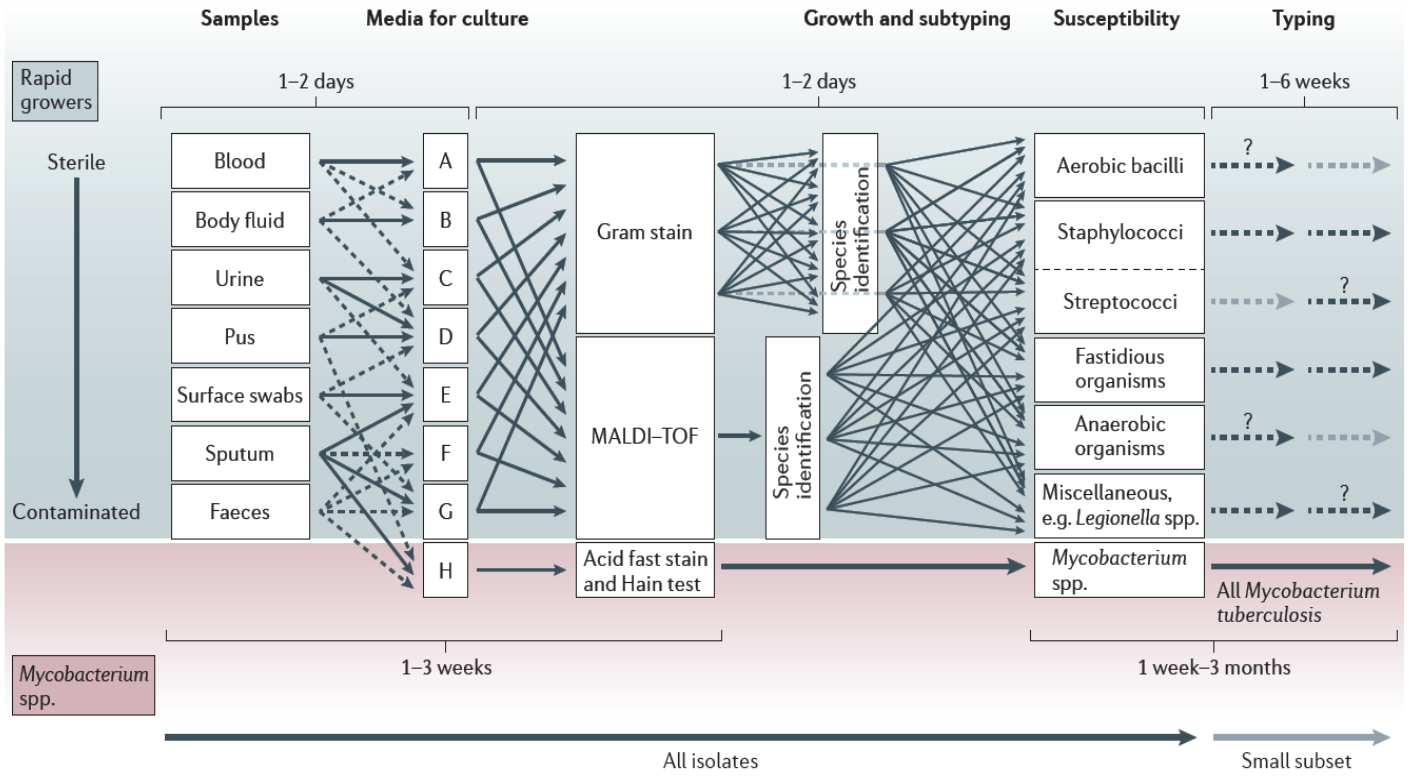
Use of microbial genomics for tool development

Report from The American Academy of Microbiology, 2015



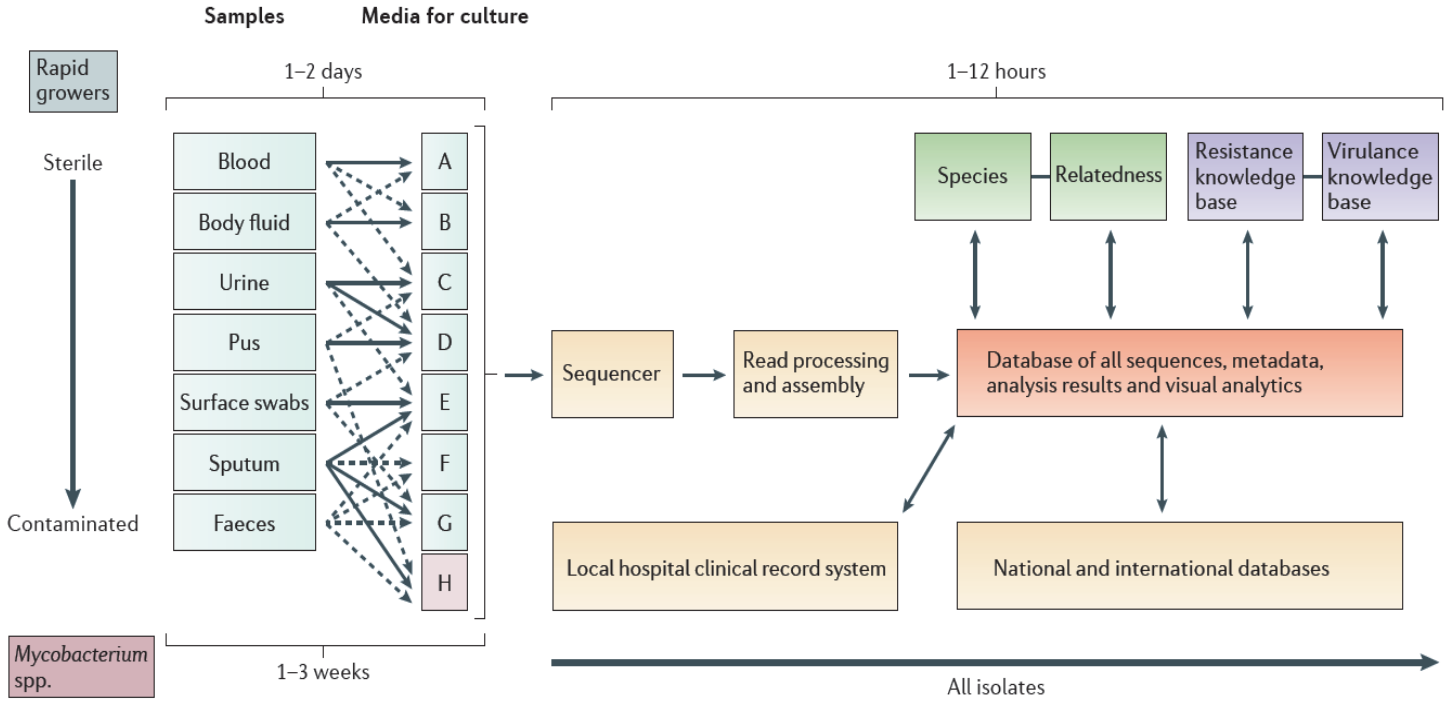
Workflow for processing samples for bacterial pathogens

Didelot et al., Nature Genet Review 2012, 13:601-612



Ongoing developments in DNA-sequencing technologies are likely to affect the diagnosis and monitoring of all pathogens, including viruses, bacteria, fungi and parasites.

The diagnostic and clinical applications of bacterial WGS



Didelot et al., Nature Genet Review 2012, 13:601-612

Foodborne outbreak identification "Crisis del pepino"

2011

Mayo

- 24 Primera muerte en Alemania
- 26 Alemania acusa a los pepinos españoles
- 30 Prohibición de importaciones de verduras de España y Alemania
- 31 Laboratorios alemanes desmienten oficialmente que los pepinos españoles sean el foco de infección

Junio

- 10 Resolución de la crisis

Causado por la toxi-infección de Escherichia coli enterohemorrágica (EHEC) (*Escherichia coli* O104:H4)

Muerte: 32 personas en Alemania, 1 Suecia y 1 Francia y 2263 infectados en 12 países de Europa.

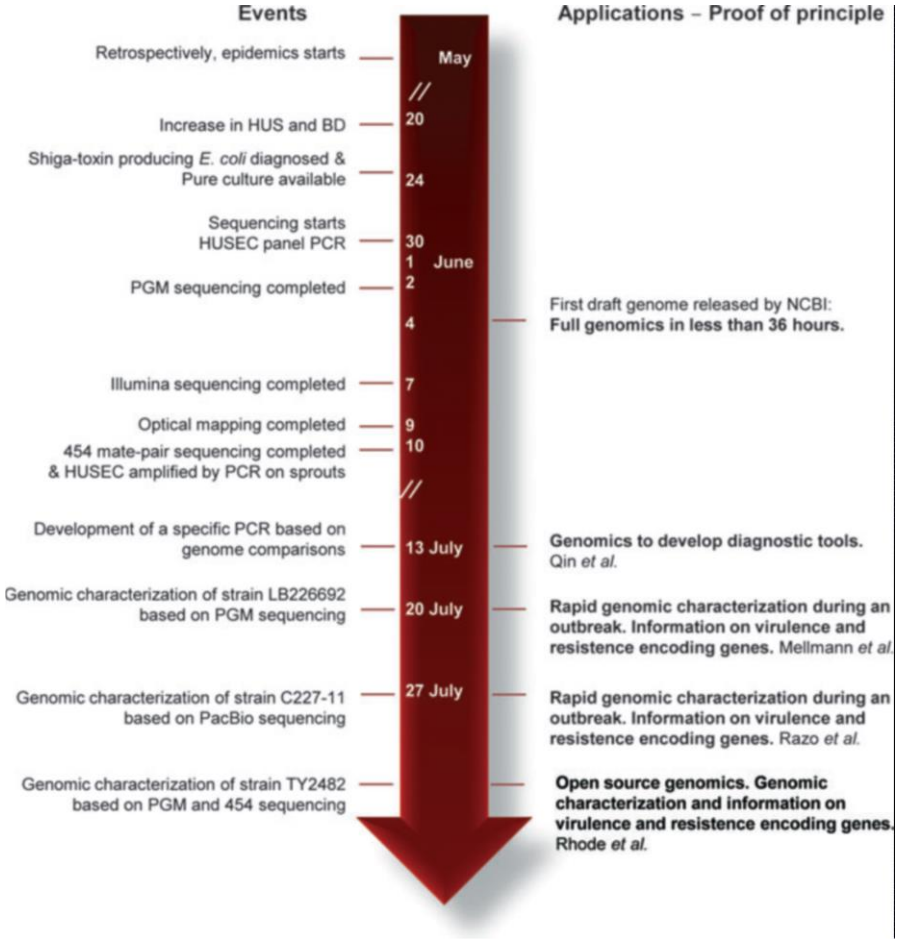
Crisis Política y Económica Europa: Alto impacto en la Economía Europea, mayor afectación en la Española

Secuenciación Genoma

华大基因 BGI

Universitätsklinikum Hamburg-Eppendorf

The Escherichia coli O104:H4 epidemics: event timeline and major outputs



Bertelli and Greub, Clin Microb and Infect, 2013

Foodborne outbreak identification “Crisis del pepino”

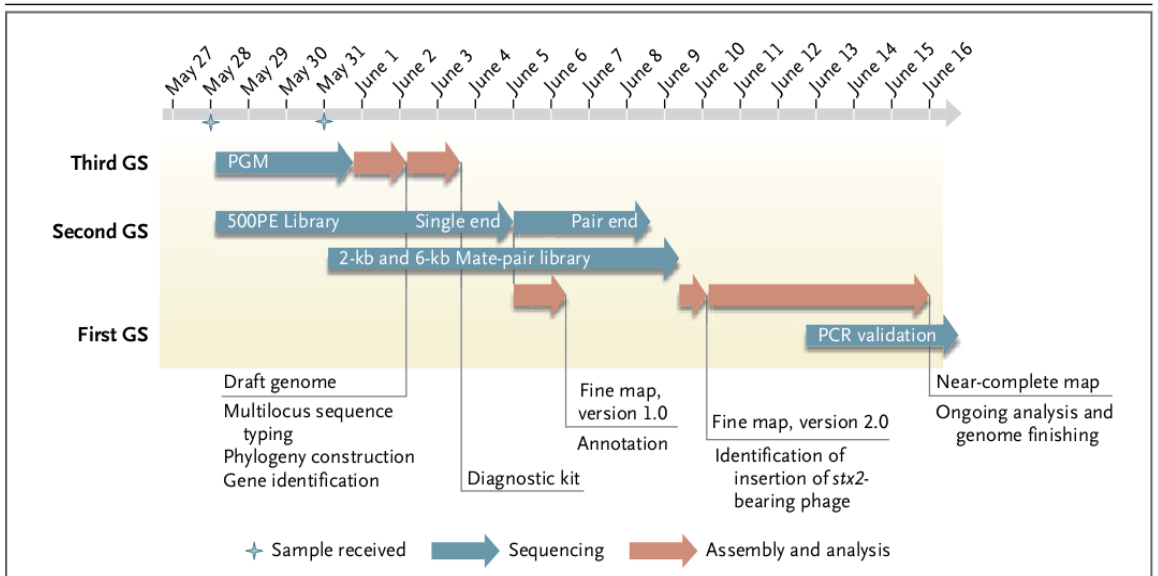


Figure 1. Timeline of the Open-Source Genomics Program.

After receiving the first batch of DNA samples on May 28, 2011, sequencing runs with the use of the Ion Torrent Personal Genome Machine (PGM) and Illumina (small-insert library) were initiated simultaneously. On May 31, the second batch of DNA was received and used for Illumina large-insert sequencing. An assembly of the Ion Torrent reads was released on June 2, which enabled subsequent analyses (multilocus sequence typing, phylogenetic analysis, and genome comparisons). Errors in the Ion Torrent data were corrected with the use of later Illumina data, and a high-quality draft genome sequence was created. GS denotes generation of sequencing technology. The symbols at May 28 and May 31 in the timeline indicate the arrival of DNA samples.

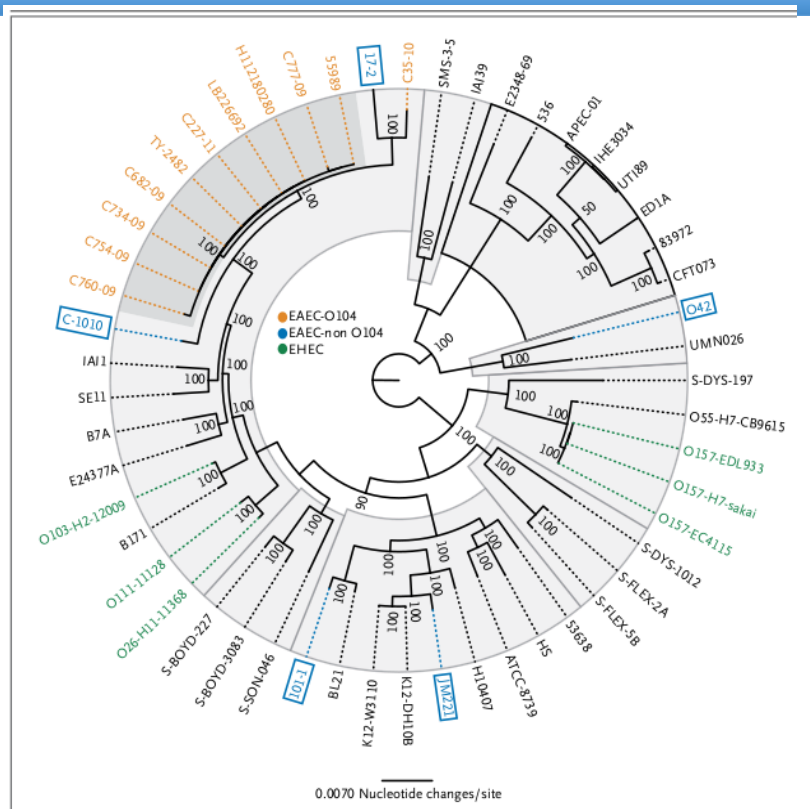


Figure 2. Phylogenetic Comparisons of 53 *Escherichia coli* and Shigella Isolates.

Genomic sequences were compared with the use of 100 bootstrap calculations, as described by Sahl et al.[35] The species-based phylogeny was inferred with the use of 2.56 Mbp of the conserved core genome. The O104:H4 isolates are shown in orange, the reference enteroaggregative *E. coli* (EAEC) isolates in blue, and the enterohemorrhagic *E. coli* isolates in green. (The classification of the other strains is shown in Fig. 4 and Table 4 in the Supplementary Appendix.) The O104:H4 isolates cluster into a single clade (dark gray); in contrast, the reference EAEC isolates are extremely divergent and are represented throughout the phylogeny.

Rohde et al NEJM 2011, 365:718-24

Andalusian Listeria Outbreak

Actualización de información sobre el brote de intoxicación alimentaria causado por Listeria monocytogenes.

Publica: Agencia Española Seguridad alimentaria y Nutrición
Fecha: 29 agosto 2019
Sección: Seguridad Alimentaria

Jueves 29 de agosto de 2019, 12.00 horas

ACTUALIZACIÓN EN RELACIÓN CON LA DISTRIBUCIÓN DE PRODUCTOS RELACIONADOS CON LA ALERTA.

La Agencia Española de Seguridad Alimentaria y Nutrición (AESAN) recomienda a las personas que tengan en su domicilio algún producto de la marca "La Mechá" se abstengan de consumirlo. Si se dispone del producto se debe devolver al punto de compra y, de no ser posible, desecharlo.

Brote de listeriosis: sube el número d afectados y se apunta a la falta de higiene en la carne como causa

EFE 25.08.2019

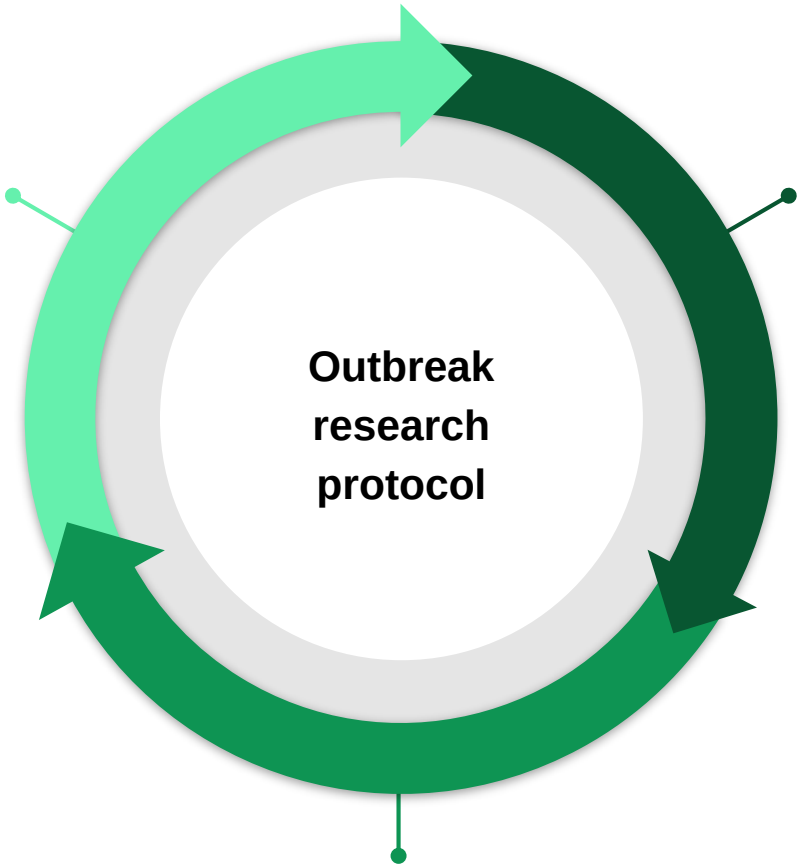
- Tres nuevos casos, en Sevilla y Cádiz, dejan el número de personas afectadas en Andalucía en 192.
- La carne con listeria de la marca blanca se vendió en los municipios de Sevilla.
- La empresa que vendió la marca blanca de Magrudis dice que cumple los protocolos.

- Meat “La Mechá”. Margulis S.L.
- 250 cases related.
- Meat “"La Montanera del Sur".INCARYBE S.L”, suspicion. (Cádiz)
- Meat “Sabores de Paterna” (Málaga)

Andalusian Listeria Outbreak

Transversal:

- Data management
- Sample tracking
- Infrastructure

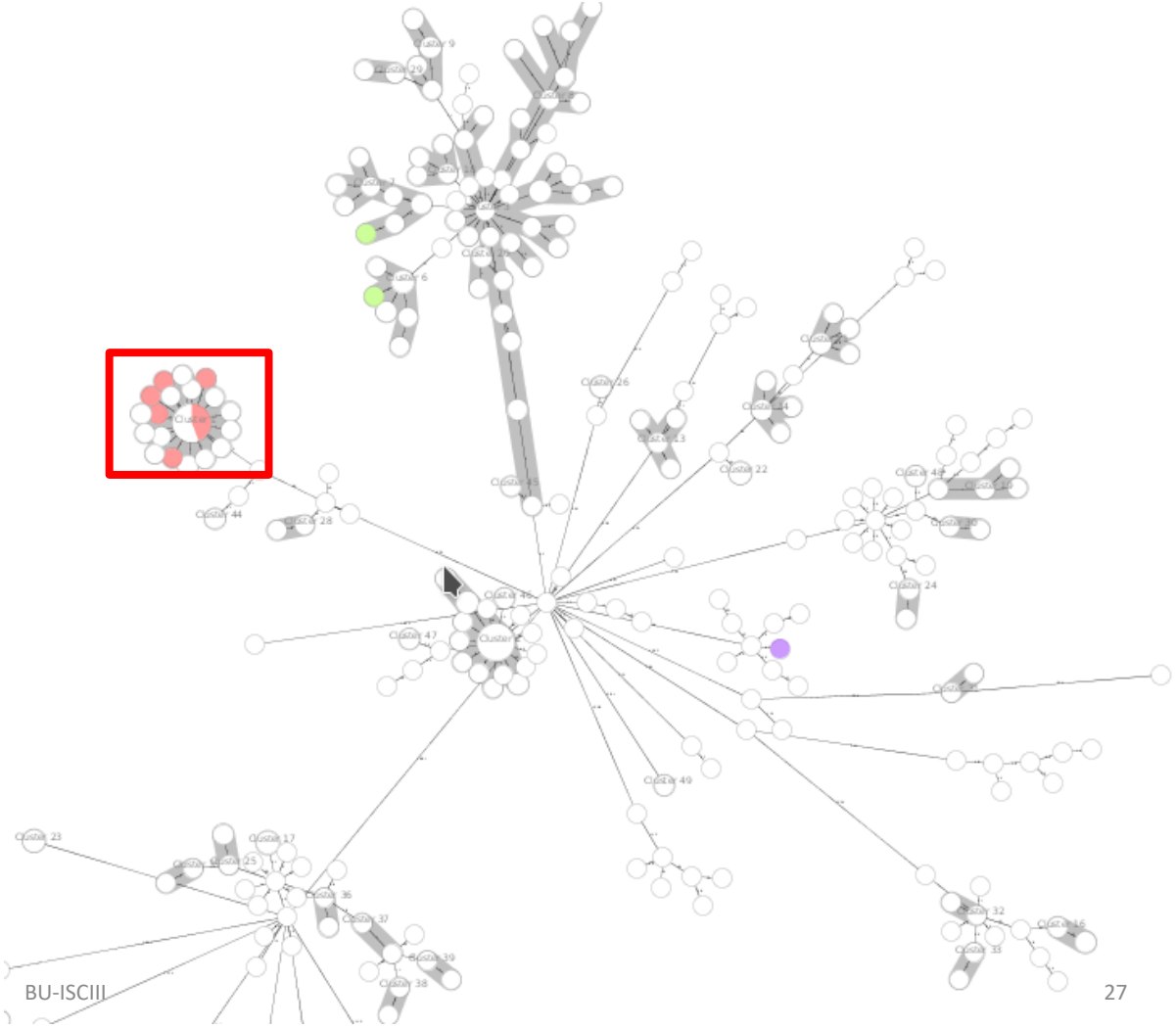


Andalusian Listeria Outbreak

- 625 listeria samples already sequenced
- 258 suspected to be related to the outbreak (mid august to mid september)

Results:

- 233 related to the outbreak, confirmed to be caused by the meat “La Mechá”
- 25 sporadic cases not related to the outbreak.



Pathogen discovery: new virus – SARS-CoV-2

Deep Meta-Transcriptomic Sequencing

bronchoalveolar lavage fluid (BALF)

Meta-transcriptomic library

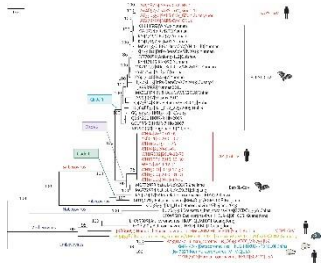
2x150 MiniSeq 56,565,928 sequences reads

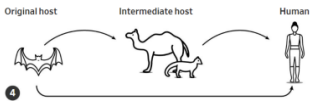
De novo-assembled - Megahit

384,096 Contigs

Screened for potential aetiological agents

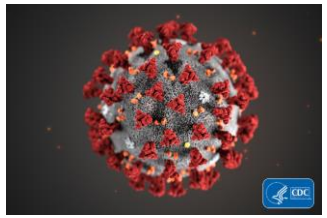
The longest 30,474 nt





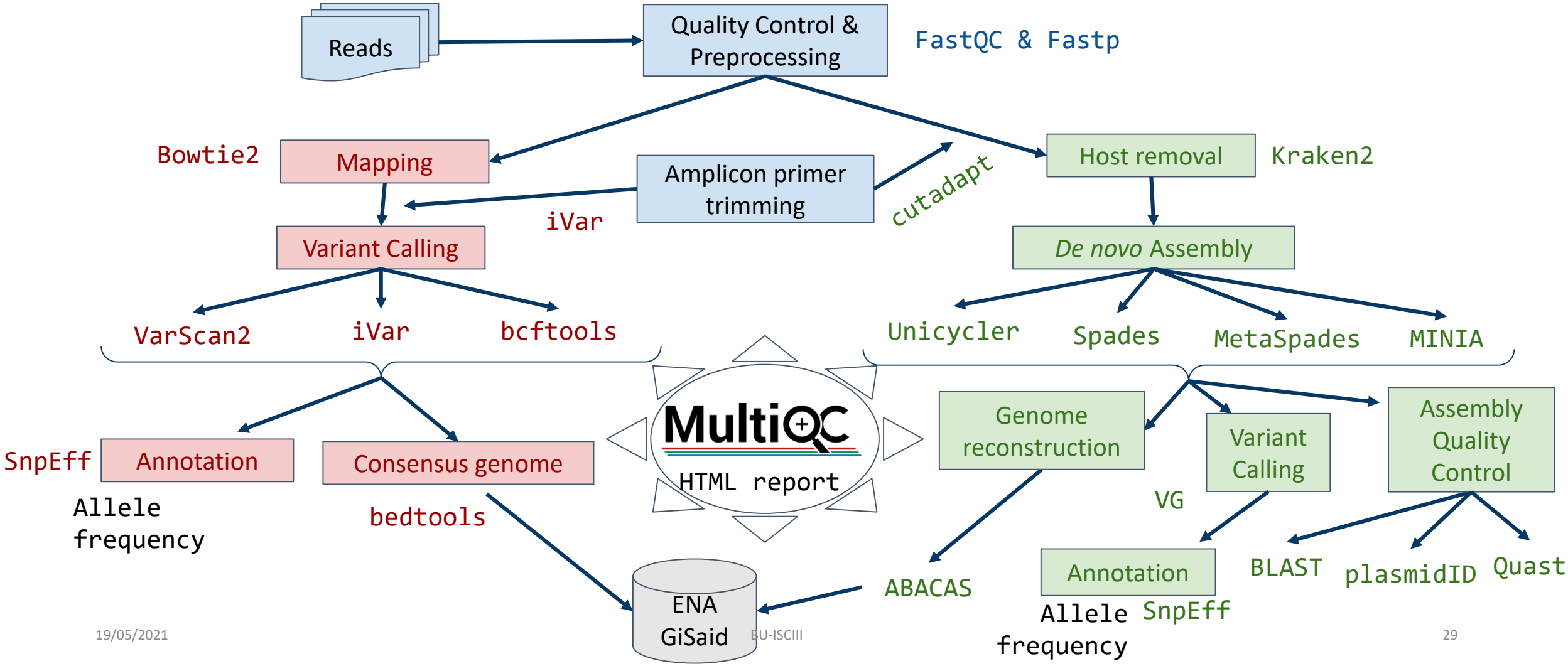
89.1% identity

Closely related to a bat SARS-like coronavirus



Wu et al., Nature 2020

Viralrecon


Reads
Quality Control & Preprocessing
FastQC & Fastp
Bowtie2
Mapping
Amplicon primer trimming
cutadapt
Host removal
Kraken2
iVar
Variant Calling
De novo Assembly
VarScan2
iVar
bcftools
Unicycler
Spades
MetaSpades
MINIA
MultiQC
HTML report
SnpEff
Annotation
Consensus genome
Genome reconstruction
Variant Calling
Assembly Quality Control
Allele frequency
bedtools
VG
ENA
GiSaid
ABACAS
Annotation
BLAST
plasmidID
Quast
Allele frequency
SnpEff

Surveillance-based reporting

Virussequenc	sample	totalreac	readshostR	readsho	%readsho	readsviru	%readsviru	unmappedreac	%unmapedreac	meanDPcoverageviru	Coverage>10x(%	Variantsinconsensusx1	MissenseVarian	%Ns1C	Lineages
NC_045512.2	211524	2040680	24760	49520	2.43	1980480	97.05	10680	0.52	4607.76	99.79	41	28	0,17	B.1.1.7
NC_045512.2	210710-2	2166652	12091	24182	1.12	2140652	98.80	1818	0.08	5331.12	99.77	40	25	0,18	B.1.1.7
NC_045512.2	211496	2092444	2299	4598	0.22	2081982	99.50	5864	0.28	4684.85	99.55	35	23	0,22	B.1.1.7
NC_045512.2	211429	1739402	1643	3286	0.19	1730009	99.46	6107	0.35	3891.88	98.84	46	27	0,24	B.1.1.7
NC_045512.2	211507	1394744	1500	3000	0.22	1386515	99.41	5229	0.37	3438.01	99.05	39	26	0,25	B.1.1.7
NC_045512.2	211517	2284388	6870	13740	0.60	2263829	99.10	6819	0.30	4882.96	99.66	40	26	0,26	B.1.1.7
NC_045512.2	211405	1946156	19779	39558	2.03	1900227	97.64	6371	0.33	4174.87	99.57	37	25	0,32	B.1.1.7
NC_045512.2	211497	1823806	104	208	0.01	1821435	99.87	2163	0.12	4889.73	99.65	39	25	0,33	B.1.1.7
NC_045512.2	211430	1859150	378	756	0.04	1853758	99.71	4636	0.25	4475.72	99.66	20	7	0,34	B.1.177
NC_045512.2	211498	1974394	713	1426	0.07	1963930	99.47	9038	0.46	4331.72	99.63	38	25	0,34	B.1.1.7
NC_045512.2	211506	2054730	28336	56672	2.76	1987951	96.75	10107	0.49	3822.90	99.15	41	27	0,49	B.1.1.7
NC_045512.2	211516	2240664	446	892	0.04	2232598	99.64	7174	0.32	4427.42	99.16	19	8	0,61	B.1.177
NC_045512.2	211495	2245802	4263	8526	0.38	2218628	98.79	18648	0.83	3620.33	98.60	35	23	0,62	B.1.1.7
NC_045512.2	211513	2191920	1071	2142	0.10	2183372	99.61	6406	0.29	4394.22	99.17	18	8	0,66	B.1.177
NC_045512.2	211489	1995358	619	1238	0.06	1990170	99.74	3950	0.20	4938.41	99.14	39	27	0,7	B.1.1.7
NC_045512.2	211512	2032422	11430	22860	1.12	1985067	97.67	24495	1.21	3070.94	98.21	39	26	0,7	B.1.1.7
NC_045512.2	211419	2143186	2827	5654	0.26	2125398	99.17	12134	0.57	4100.12	99.08	41	24	0,77	B.1.1.7
NC_045512.2	211509	1905366	1412	2824	0.15	1895839	99.50	6703	0.35	4146.72	98.79	40	26	0,78	B.1.1.7
NC_045512.2	211411	2192376	1420	2840	0.13	2180537	99.46	8999	0.41	3849.04	98.07	29	18	0,96	B.1.351
NC_045512.2	211519	2406852	382	764	0.03	2390245	99.31	15843	0.66	3833.59	98.42	27	12	0,99	B.1.177
NC_045512.2	210790-2	2042452	32448	64896	3.18	1928075	94.40	49481	2.42	3288.34	98.38	32	18	1,08	B.1.160
NC_045512.2	211428	2055490	33405	66810	3.25	1975531	96.11	13149	0.64	3604.64	97.95	37	23	1,22	B.1.1.7
NC_045512.2	211511	1989742	549	1098	0.06	1980788	99.55	7856	0.39	4077.32	98.31	38	26	1,35	B.1.1.7
NC_045512.2	211425	1761330	11924	23848	1.35	1714479	97.34	23003	1.31	2725.52	97.38	39	26	1,55	B.1.1.7
NC_045512.2	211508	1711536	547	1094	0.06	1701951	99.44	8491	0.50	3551.33	98.30	20	13	1,57	C.4
NC_045512.2	211518	2215498	22494	44988	2.03	2158117	97.41	12393	0.56	3630.94	98.03	39	26	1,58	B.1.1.7

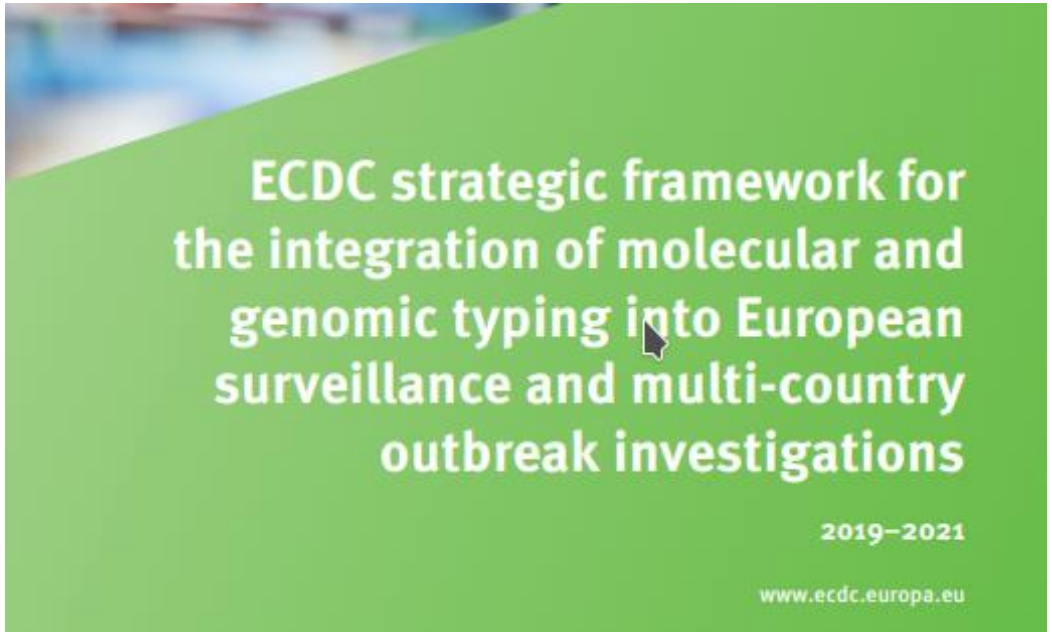
One Health approach, *infectious diseases could be better controlled and prevented*

Spanish National Microbiology Center (CNM)

Mission: Provide support to the National Health System and the different Spanish Regions in the diagnosis and control of infectious diseases. In order to fulfill this mission it acts as Reference center offering a series of scientific activities:

- Diagnosis
- **Surveillance** → Outbreak research: Molecular source detection
- Infectious diseases research
- Training

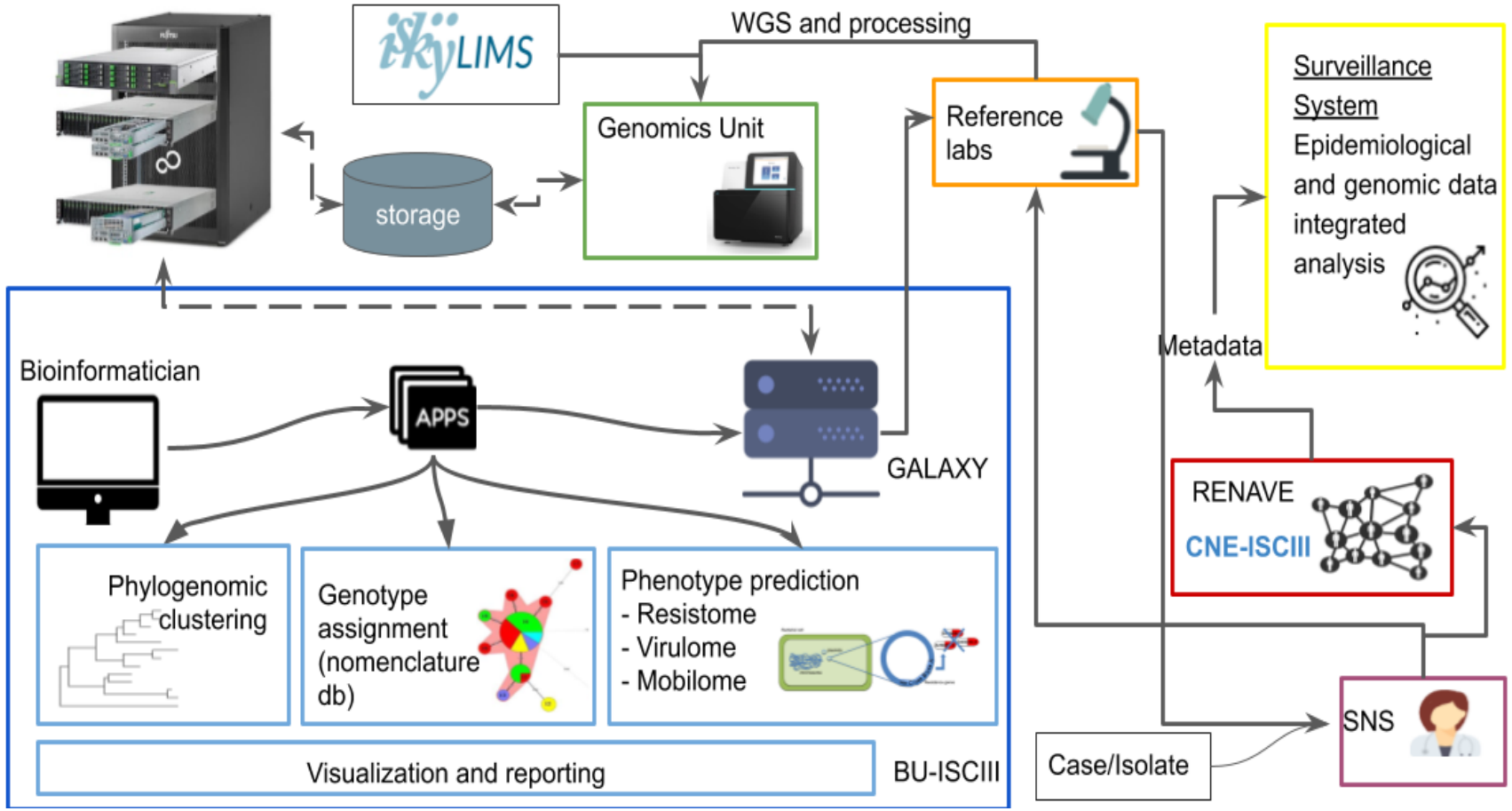
ECDC roadmap and international commitment



- **Operationalisation of EU-wide WGS-based surveillance systems in the near term:** start implementation of WGS-based surveillance for *Listeria monocytogenes*, *Neisseria meningitidis*, Carbapenemase-producing *Enterobacteriaceae* and antibiotic-resistant *Neisseria gonorrhoeae*;

2018

Bioinformatics Unit preparedness



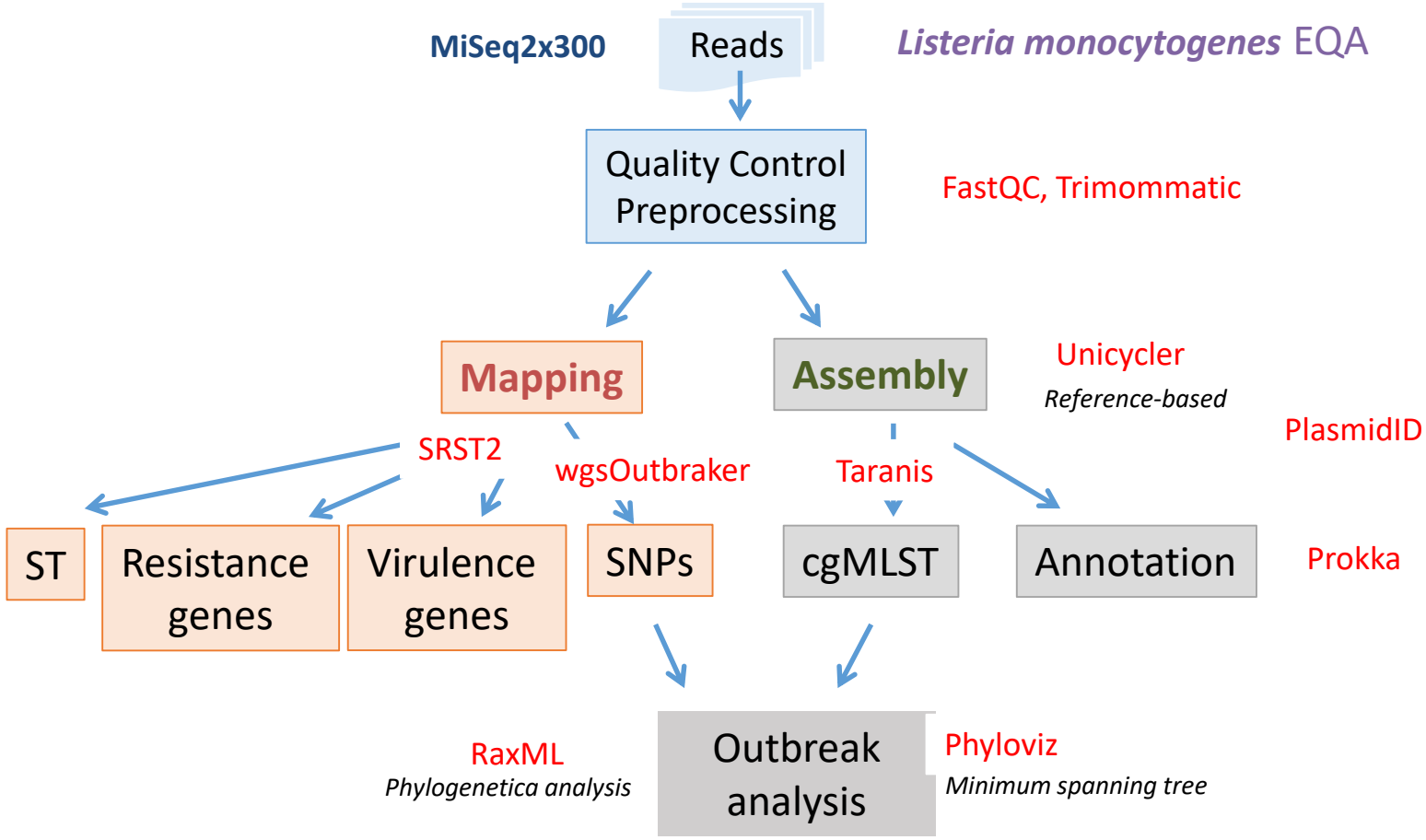
- Infrastructure
- Data
- Sample tracking
- Sharing and intercommunication
- Bioinformatics analysis

Adapted and extended from Expert opinion on whole genome sequencing for public health surveillance. ECDC

Bioinformatics analysis in microbial genomics

- SPECIE IDENTIFICATION
  - WGS - Kmers analysis
  - TARGETED METAGENOMIC, rRNA - MICROBIOTA
- ASSEMBLY GENOME
  - de NOVO or REFERENCE –BASED
  - cgMLST, wgMLST – MINIMUM SPANING TREE
  - METAGENOMIC – HOMOLOGY -BASED
- VARIANT CALLING
  - REFERENCE GENOME SELECTION
  - HAPLOYD GENOME
  - LOW FREQUENCY VARIANT – QUASISPECIES
  - SNPs MATRIX – PHYLOGENETIC ANALYSIS
- STRUCTURAL AND FUNCTIONAL ANNOTATION
  - RESISTOME, VIRULOME, SEQUENCE-TYPE

Workflow example



Software disponible – VARIANT CALLING

- CFSAN SNP Pipeline

Extracción de SNPs de alta calidad de aislados relacionados

http://snppipeline.readthedocs.io/en/latest/

- GATK, modo haploide
- Samtools
- Varscan
- Snippy

Identificación de variantes haploides y construcción de filogenia usando core genome SNPs

http://github.com/tseemann/snippy

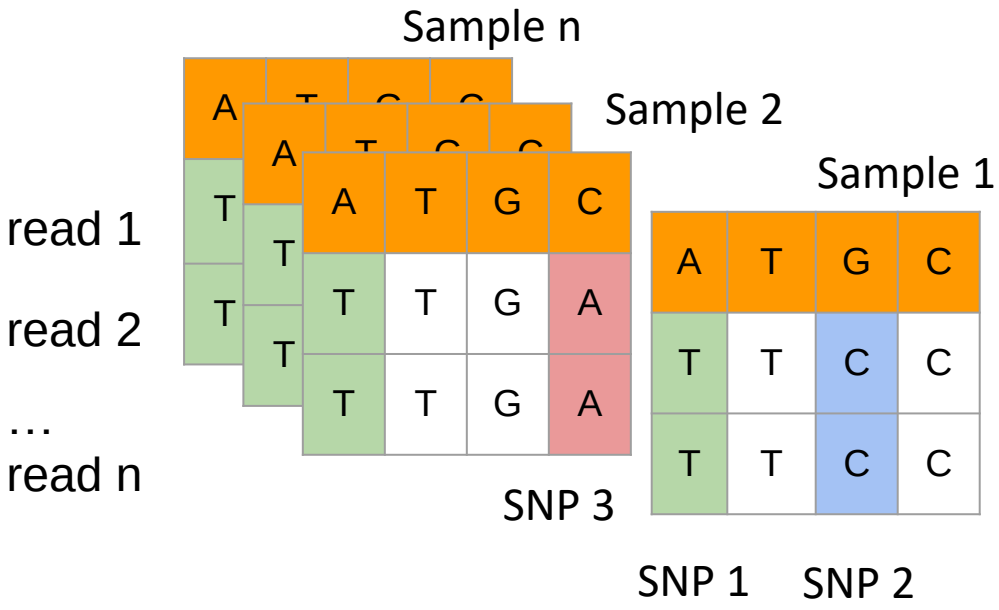
- Live-SET

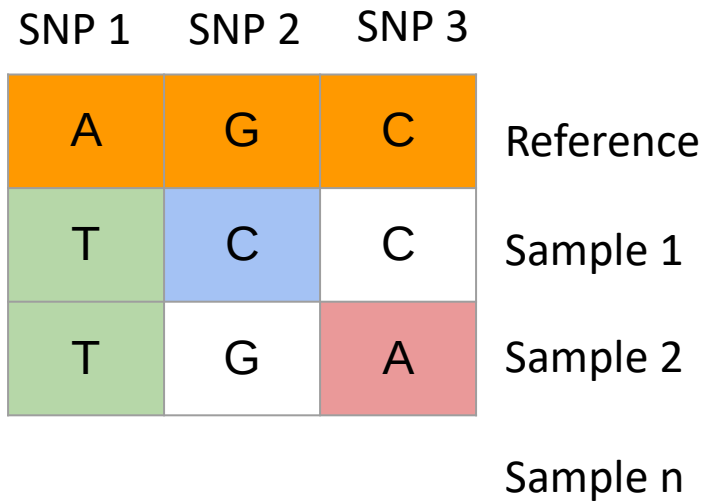
High-quality SNPs para crear filogenia para investigación de brotes

https://github.com/lskatz/lyve-SET

- WGS-Outbraker

Generación de matriz de SNPs – BACTERIA –OUTBREAK ANALYSIS



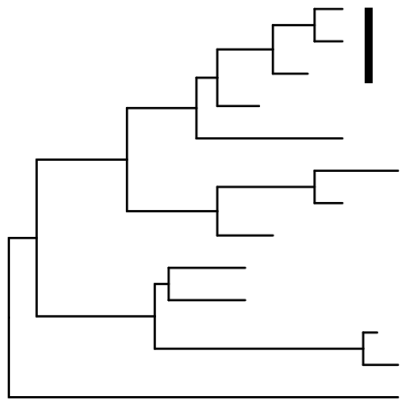


Generación de matriz de SNPs – BACTERIA –OUTBREAK ANALYSIS

SNP matrix

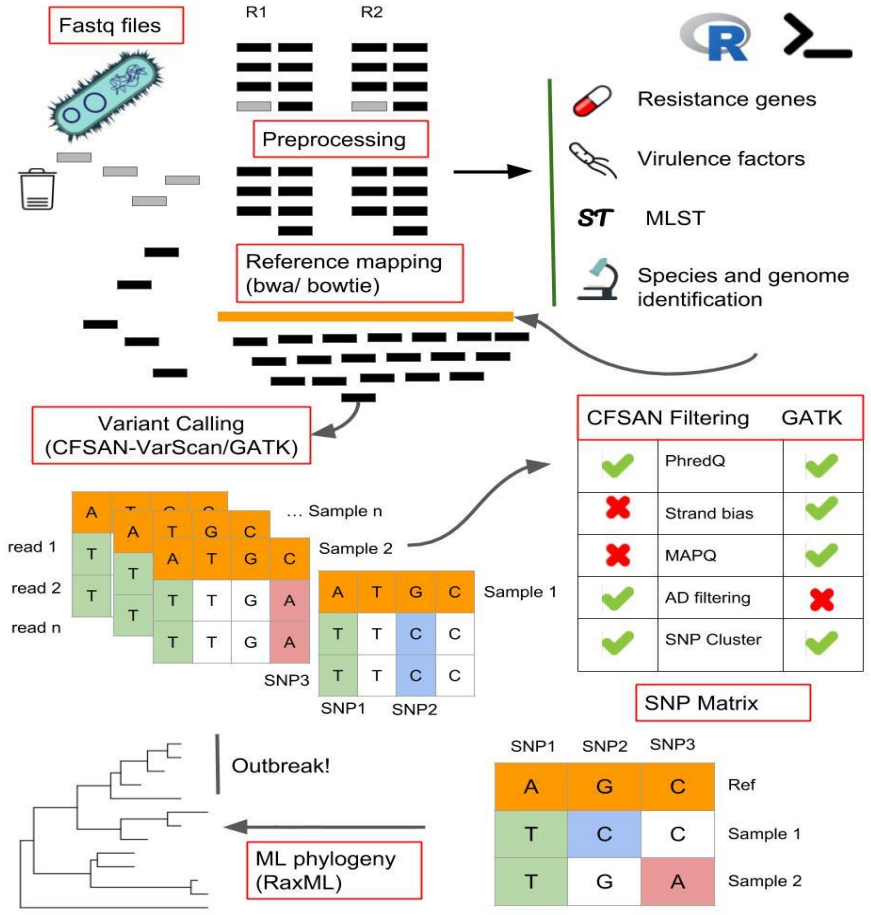
SNP 1	SNP 2	SNP 3	
A	G	C	Reference
T	C	C	Sample 1
T	G	A	Sample 2
			Sample n

Phylogeny



Outbreak!

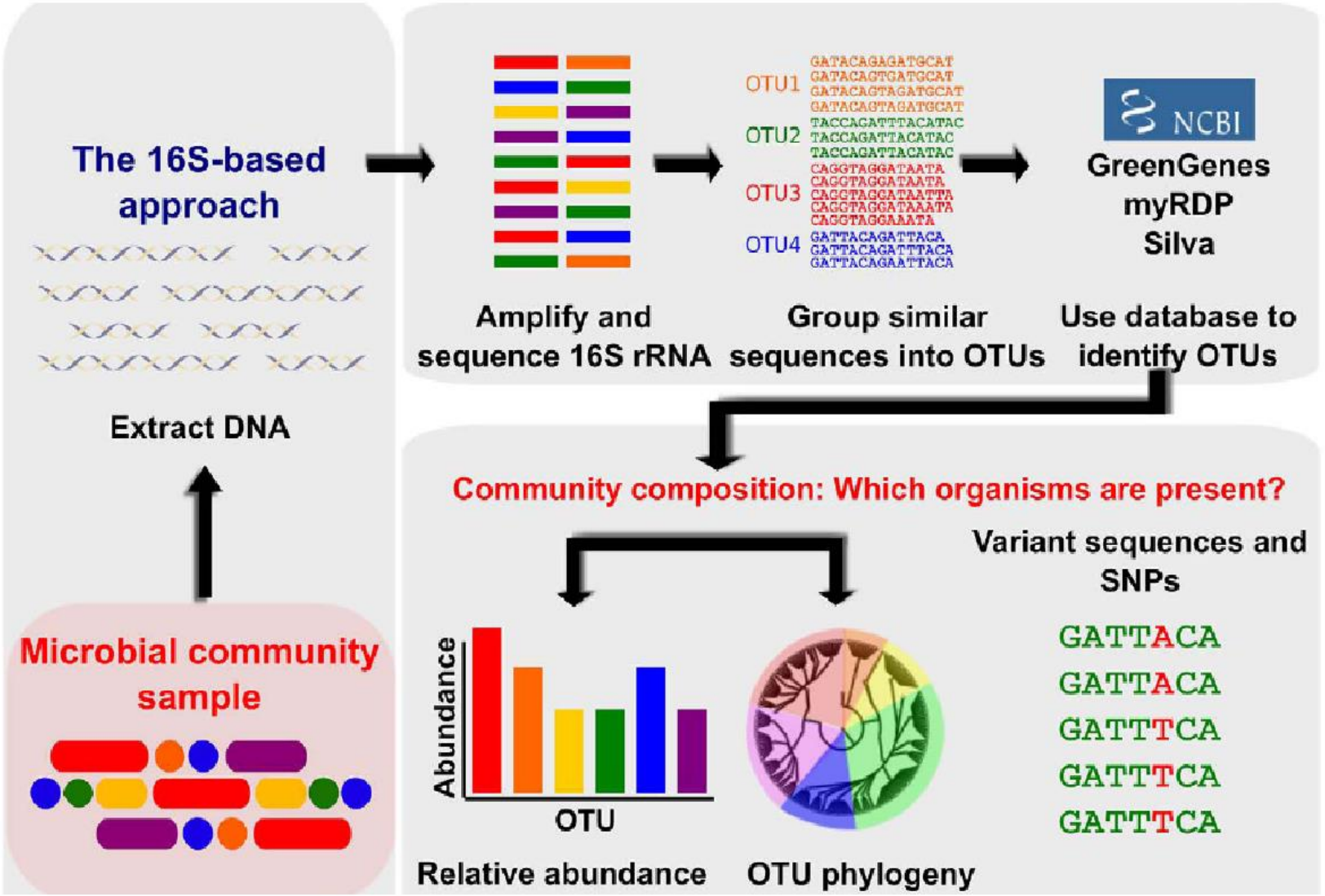
WGS-Outbreaker https://github.com/BU-ISCIII/WGS-Outbreaker



Metataxonomics or targeted metagenomics vs Metagenomics (16S - targeted 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



Metataxonomics

Etapas:

1. Filtrado
2. Eliminación de quimeras y otras anomalías
3. Formación de OTU
4. Identificación de los OTU con organismos en bases de datos

Algunos paquetes permiten llevar a cabo todo el proceso:

- mothur: `http://www.mothur.org`
- QIIME: `http://qiime.org`

Metataxonomics

Problemas:

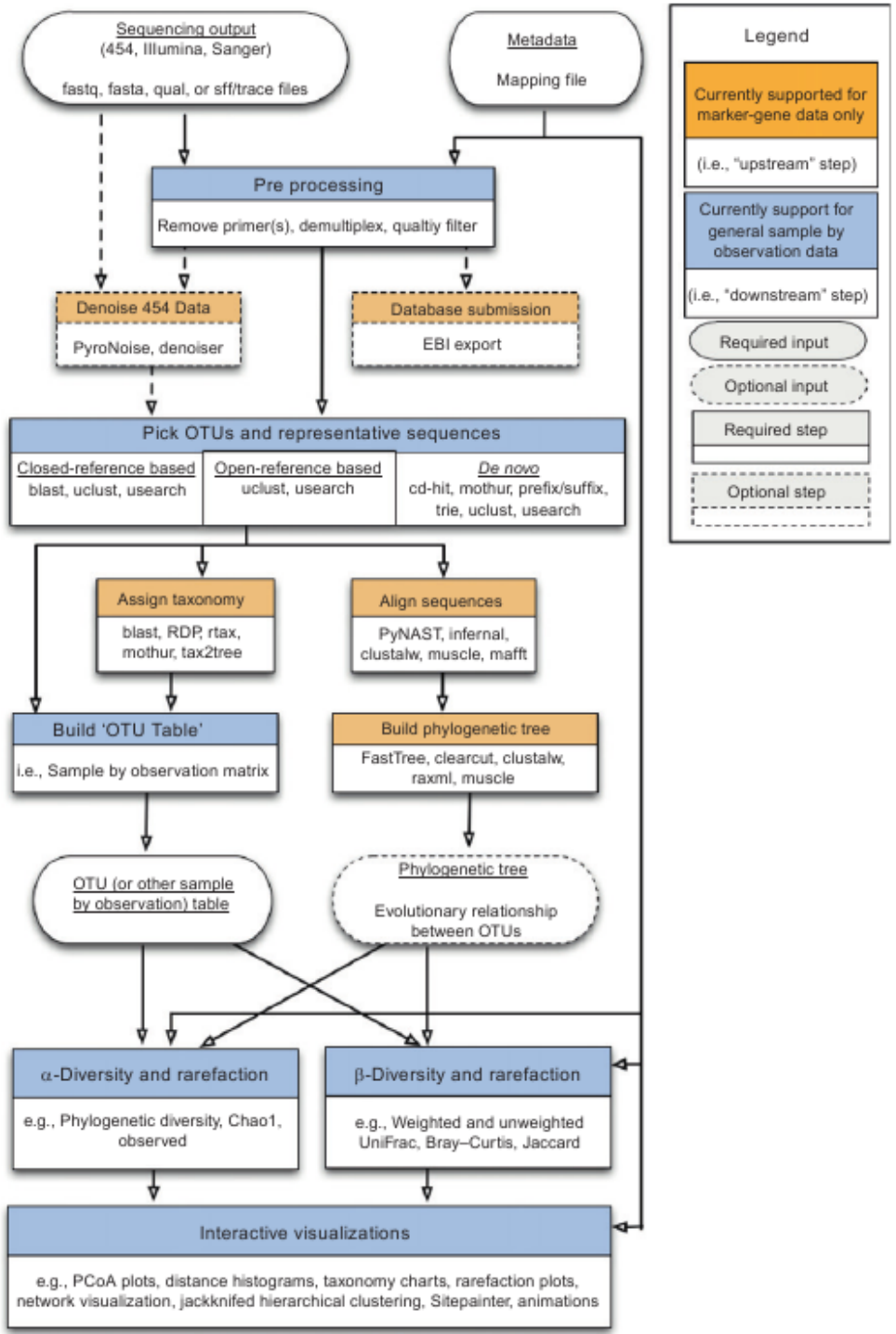
- Raros en el genoma ($< 0.1\%$)
- Los trozos similares dificultan el ensamblado correcto de lecturas pequeñas
- No todos los rRNA se amplifican en la misma medida con los *primers* universales
- Especies con diversas copias de sus genes rRNA
- No se conoce un umbral fijo de similitud que separe especies
- Tendencia a producirse quimeras en la PCR

Targeted metagenomics

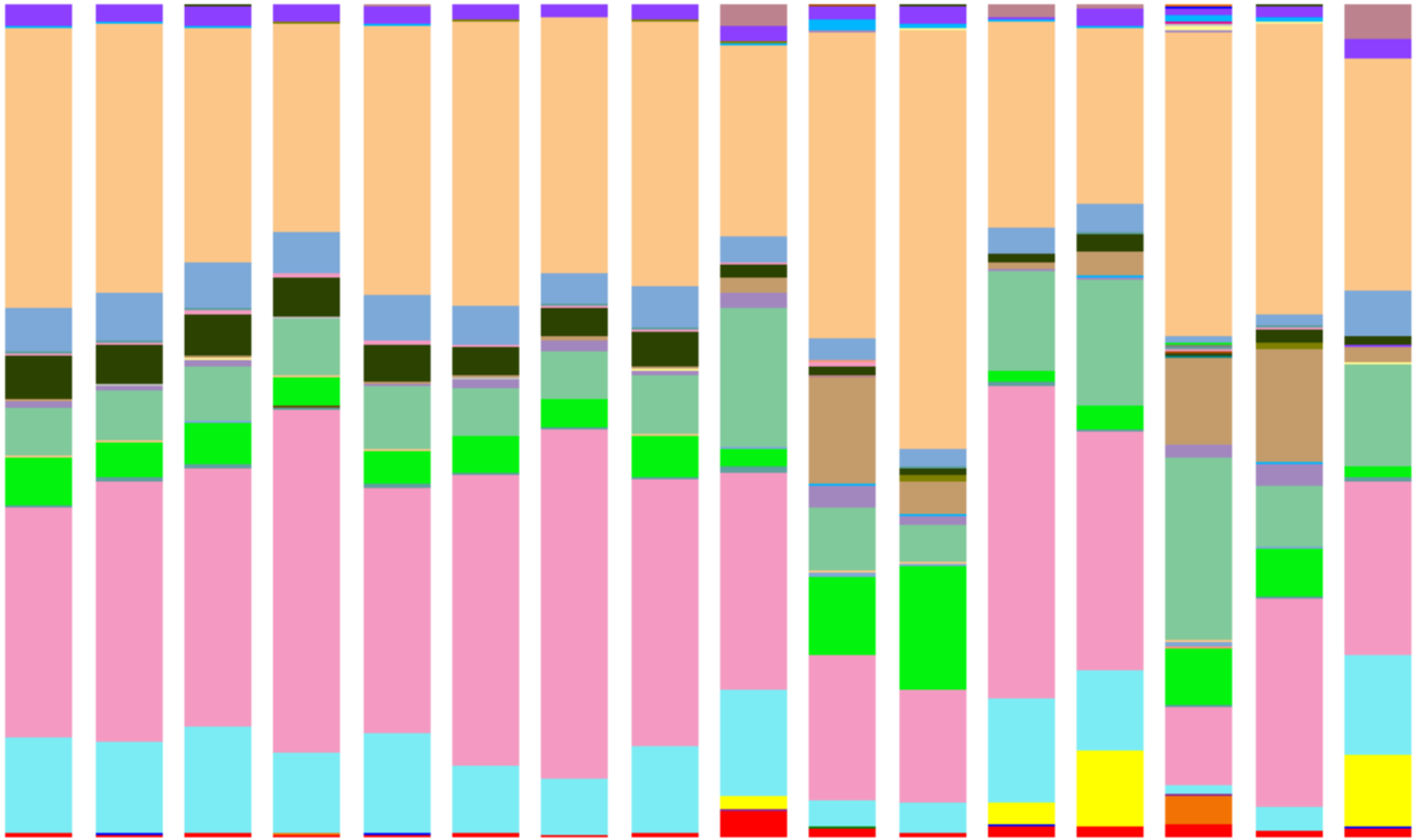
QIIME
WORKFLOW

Integrated pipeline of third-party tools

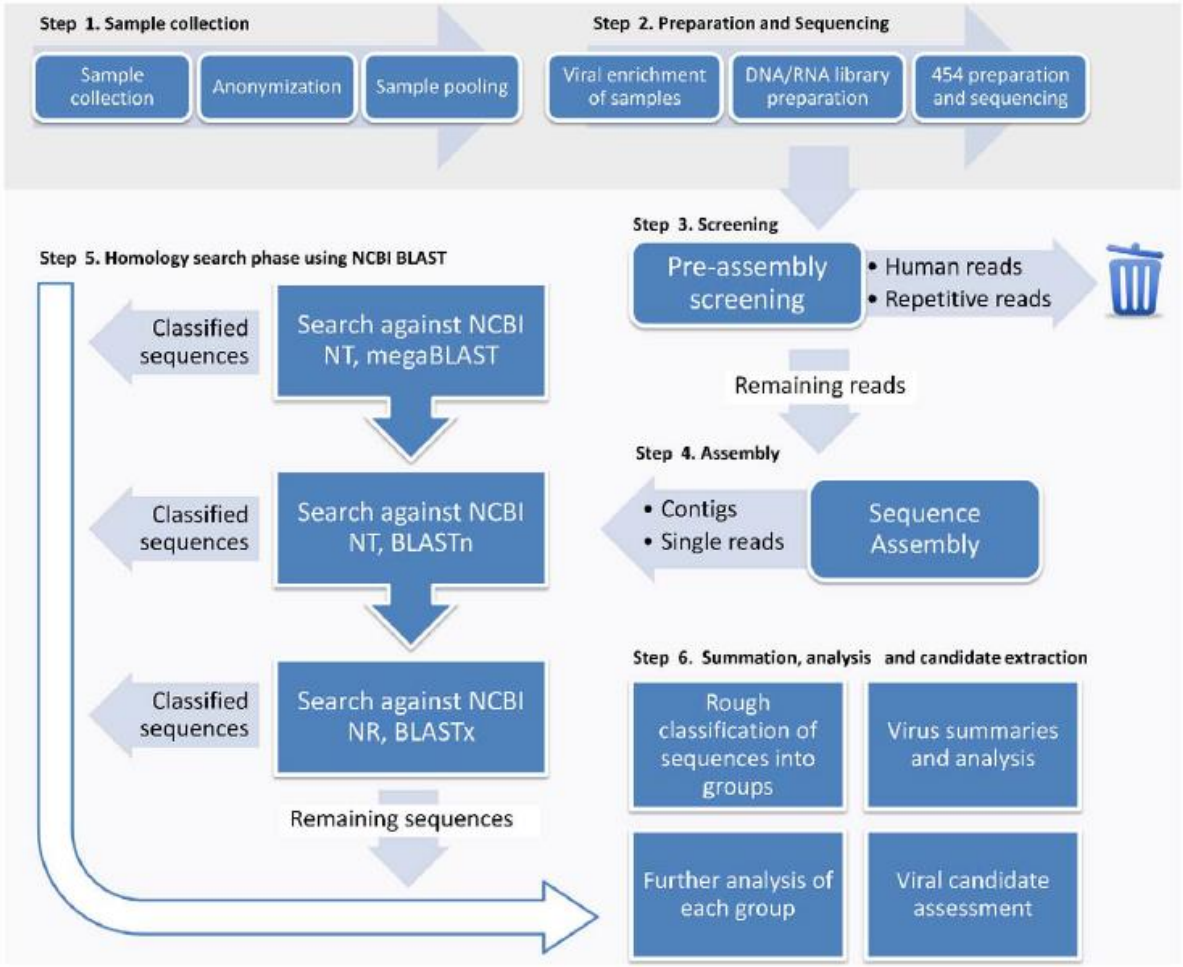
Navas-Molina et al., Methods in enzymology 2013, 531: 371-439



Taxonomy summary (i.e. phylum level)

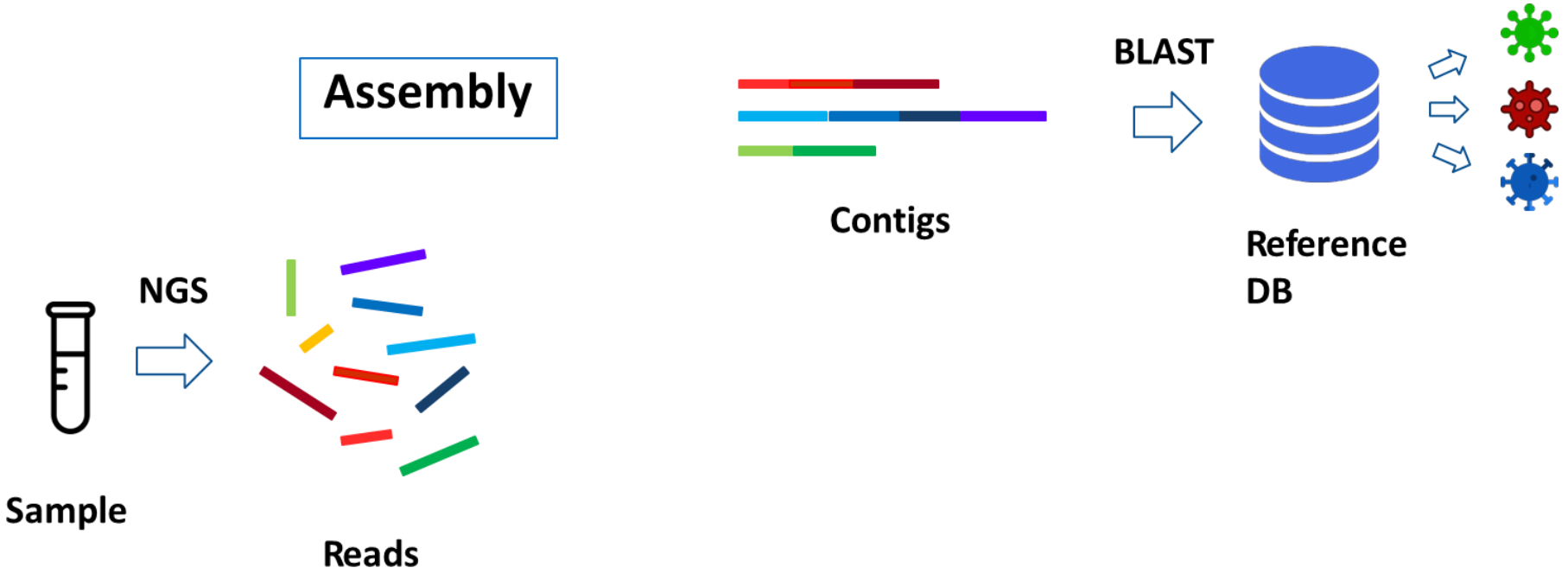


Metagenómica, pipeline de análisis

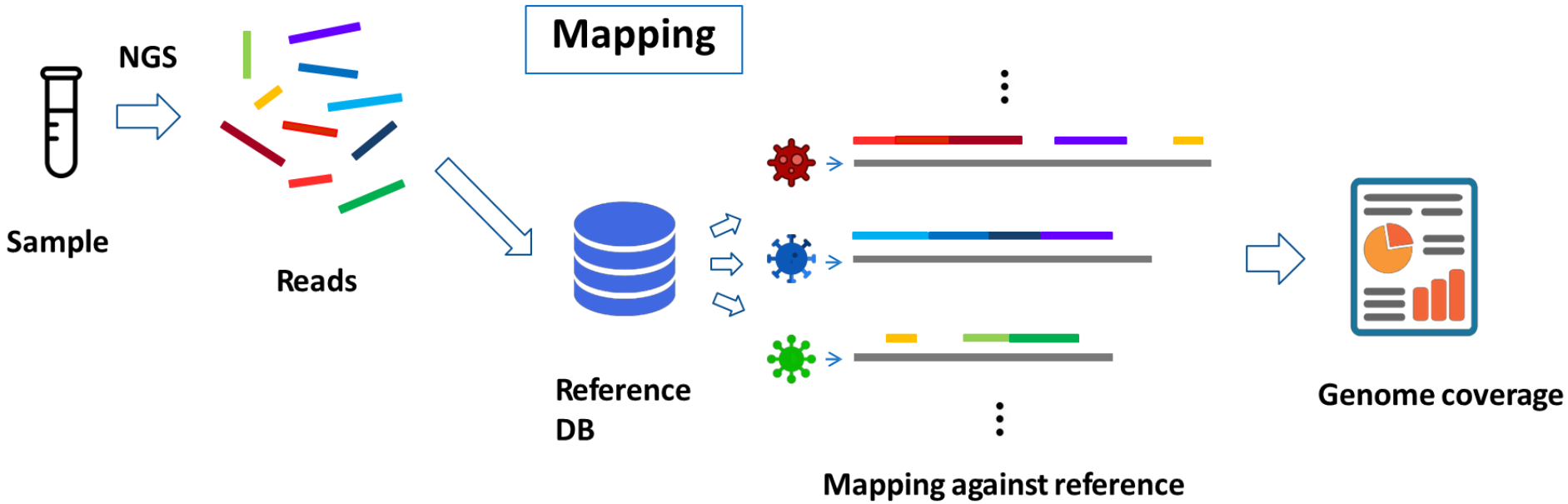


Lysholm et al., Plos One 2012:7,2, e30875

Metagenomic analysis approaches



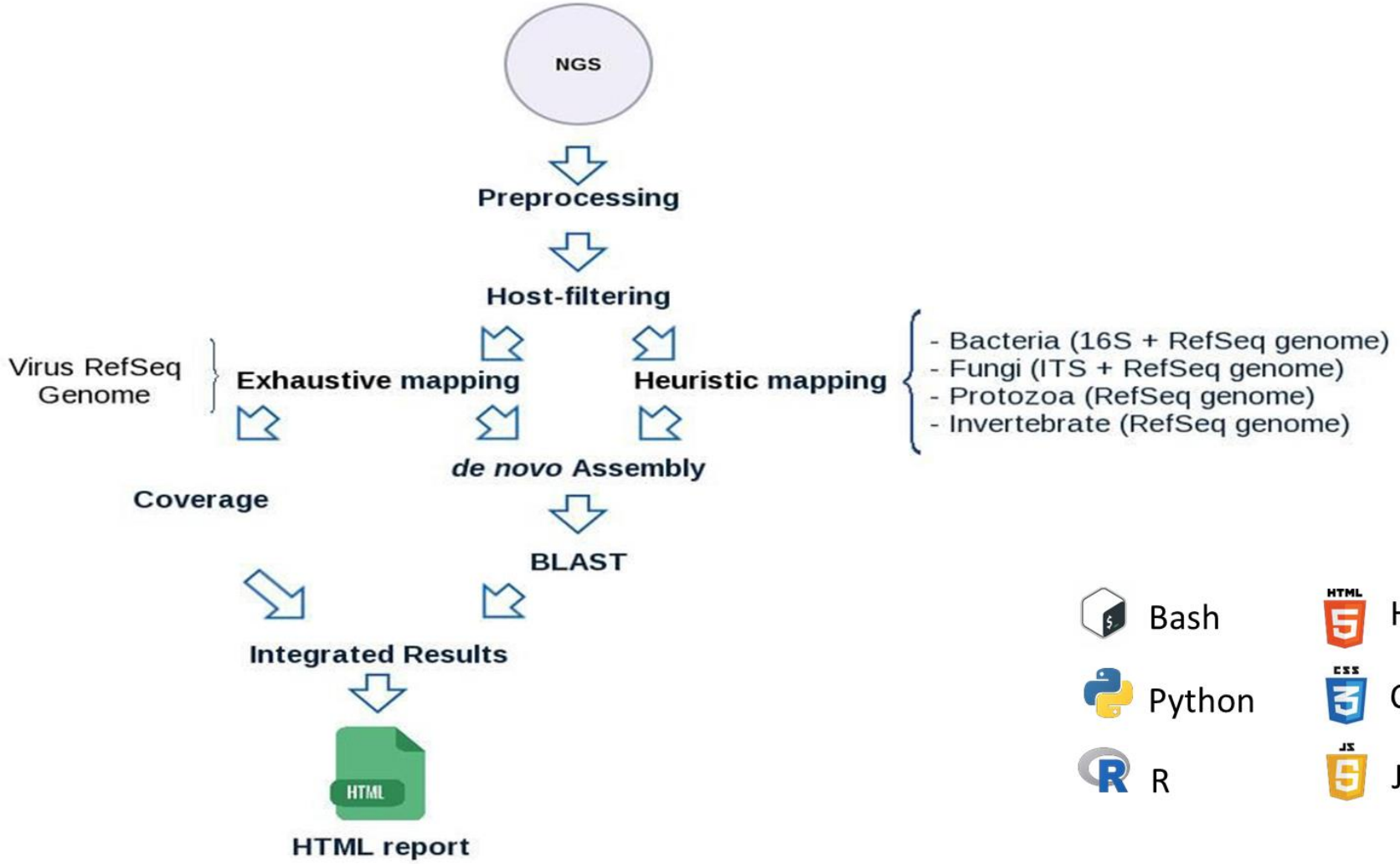
Metagenomic analysis approaches



Metataxonomics vs Metagenomics (16S vs Shotgun)

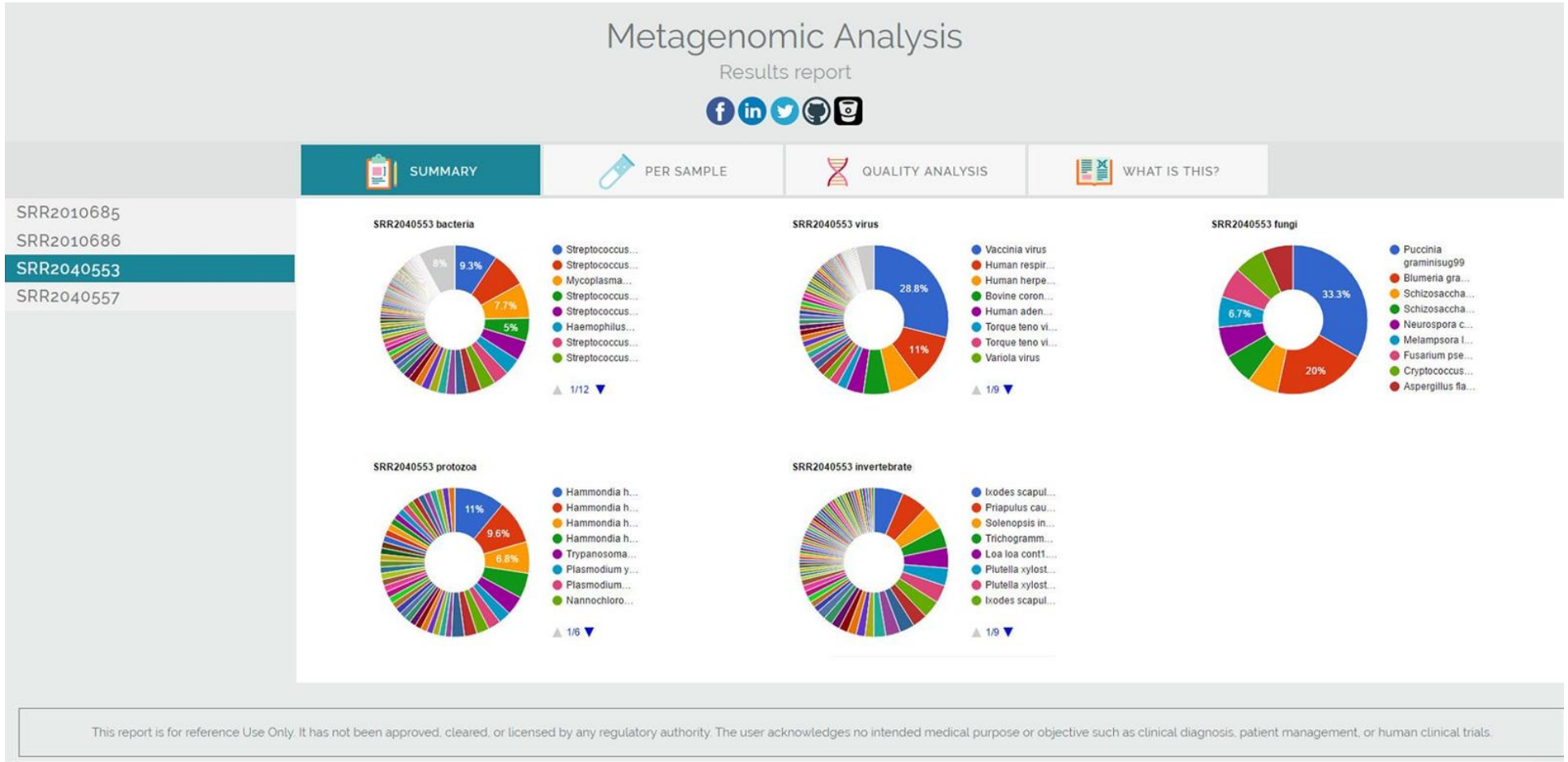
Software	Organism	Genetic portion used		Binning algorithm used			Genome coverage	Novel pathogen discovery
		Genetic markers	Whole Genome	Clustering	Mapping	Assembly		
Mothur	Bacteria	X		X			No	No
QIIME	Bacteria	X		X		X	No	No
MEGAN	Bacteria		X			X	No	No
Platypus	Bacteria		X		X		No	No
SURPI	Virus		X			X	No	Yes
Virus-TAP	Virus		X			X	No	Yes
VIP	Virus		X		X		No	Yes
Pathosphere	Virus, Bacteria, Eukarya		X			X	No	Yes

Metagenomic analysis: PikaVirus

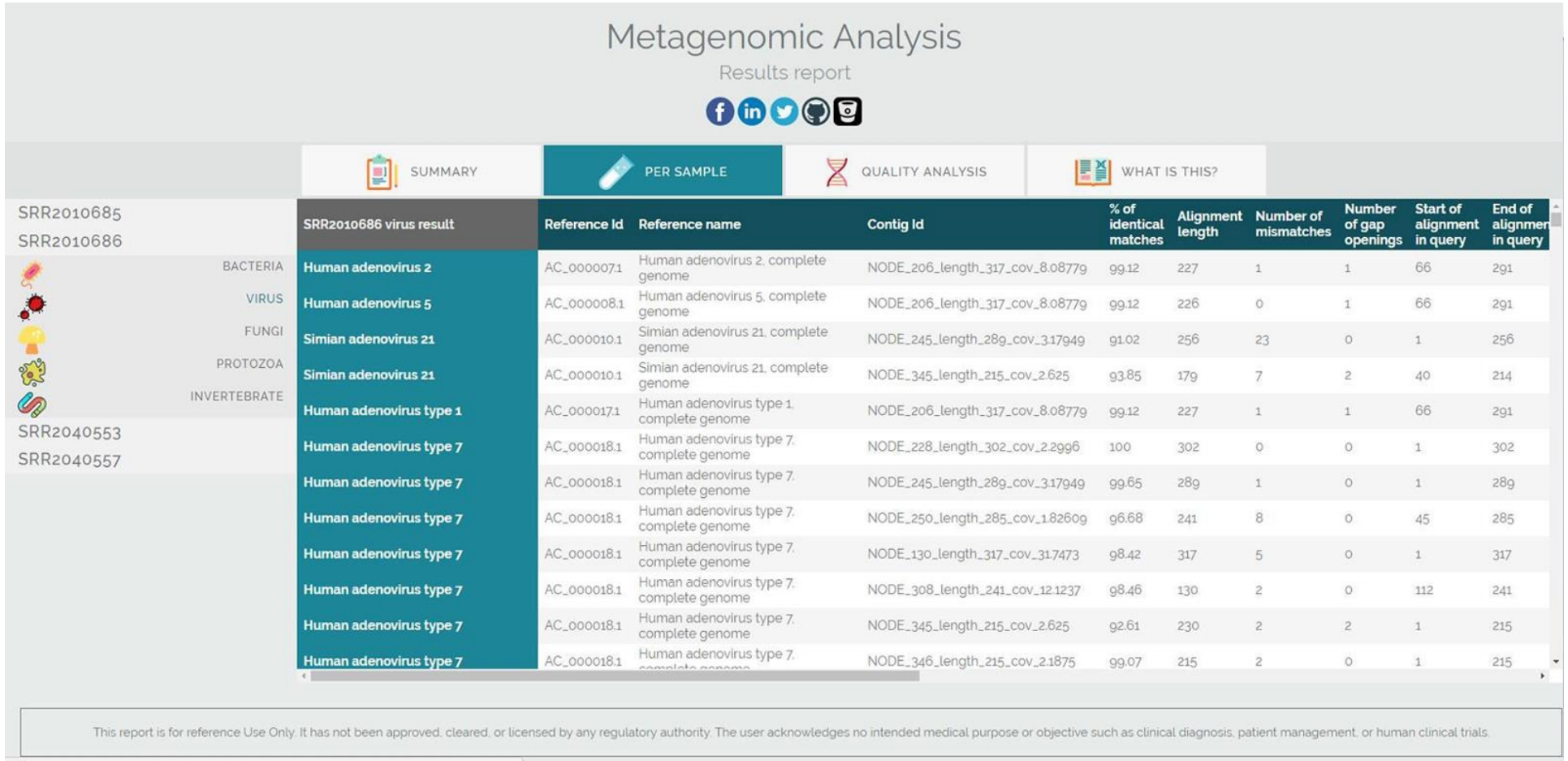


Bash
Python
R
HTML5
CSS3
JavaScript

Metagenomic analysis: PikaVirus



Metagenomic analysis: PikaVirus



Metagenomic Analysis

Results report

SUMMARY | PER SAMPLE | QUALITY ANALYSIS | WHAT IS THIS?

SRR2010685
SRR2010686

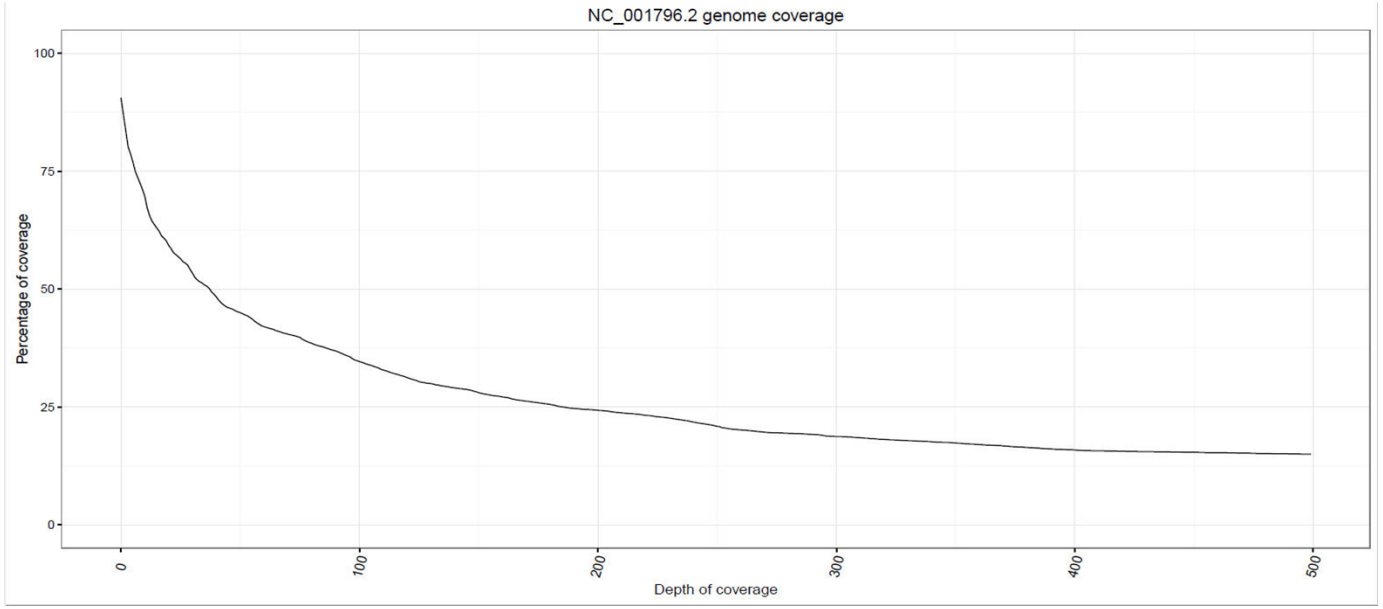
BACTERIA
VIRUS
FUNGI
PROTOZOA
INVERTEBRATE

SRR2040553
SRR2040557

SRR2010686 virus result	Reference Id	Reference name	Contig Id	% of identical matches	Alignment length	Number of mismatches	Number of gap openings	Start of alignment in query	End of alignment in query
Human adenovirus 2	AC_000007.1	Human adenovirus 2, complete genome	NODE_206_length_317_cov_8.08779	99.12	227	1	1	66	291
Human adenovirus 5	AC_000008.1	Human adenovirus 5, complete genome	NODE_206_length_317_cov_8.08779	99.12	226	0	1	66	291
Simian adenovirus 21	AC_000010.1	Simian adenovirus 21, complete genome	NODE_245_length_289_cov_3.17949	91.02	256	23	0	1	256
Simian adenovirus 21	AC_000010.1	Simian adenovirus 21, complete genome	NODE_345_length_215_cov_2.625	93.85	179	7	2	40	214
Human adenovirus type 1	AC_000017.1	Human adenovirus type 1, complete genome	NODE_206_length_317_cov_8.08779	99.12	227	1	1	66	291
Human adenovirus type 7	AC_000018.1	Human adenovirus type 7, complete genome	NODE_228_length_302_cov_2.2996	100	302	0	0	1	302
Human adenovirus type 7	AC_000018.1	Human adenovirus type 7, complete genome	NODE_245_length_289_cov_3.17949	99.65	289	1	0	1	289
Human adenovirus type 7	AC_000018.1	Human adenovirus type 7, complete genome	NODE_250_length_285_cov_1.82609	96.68	241	8	0	45	285
Human adenovirus type 7	AC_000018.1	Human adenovirus type 7, complete genome	NODE_130_length_317_cov_31.7473	98.42	317	5	0	1	317
Human adenovirus type 7	AC_000018.1	Human adenovirus type 7, complete genome	NODE_308_length_241_cov_12.1237	98.46	130	2	0	112	241
Human adenovirus type 7	AC_000018.1	Human adenovirus type 7, complete genome	NODE_345_length_215_cov_2.625	92.61	230	2	2	1	215
Human adenovirus type 7	AC_000018.1	Human adenovirus type 7, complete genome	NODE_346_length_215_cov_2.1875	99.07	215	2	0	1	215

This report is for reference Use Only. It has not been approved, cleared, or licensed by any regulatory authority. The user acknowledges no intended medical purpose or objective such as clinical diagnosis, patient management, or human clinical trials.

Metagenomic analysis: PikaVirus

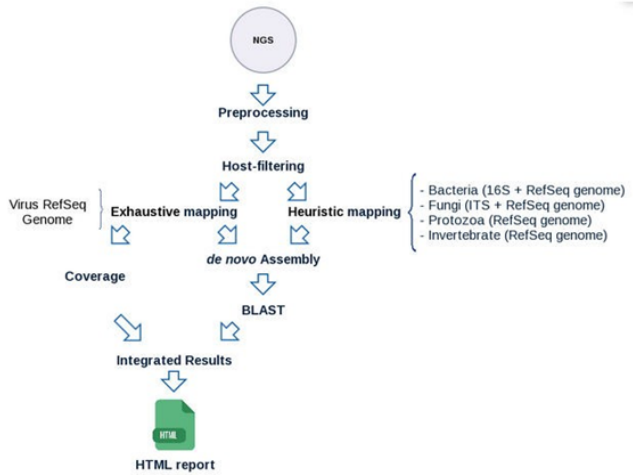


Wiki page and Project code

Welcome to pikaVIRUS

This project includes scripts to run metagenomic analysis on a single or several samples.

Workflow



Important!

First things first, for this to work there are a few dependencies you need to have installed. Also, it is necessary to have a refseq DB for every organism group you want to search for in your samples. You can find the the dependency list in Dependencies and the procedure to generate the DB files in References.

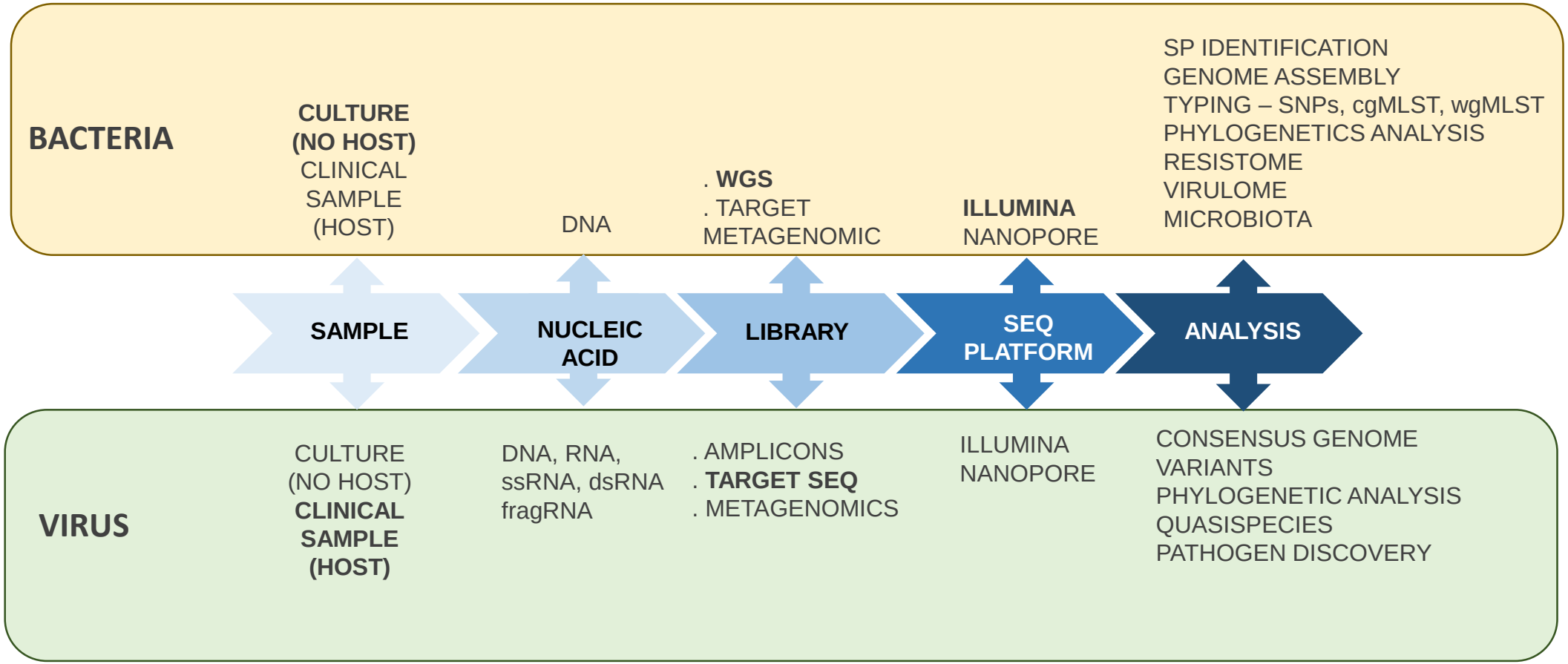
Clone this wiki locally

https://github.com/AndreaRP/

Clone in Desktop

https://github.com/BU-ISCIII

Bacterial and Viral Genome Sequencing



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 Paired-end
- Número de bases por muestra
- Profundidad de cobertura x

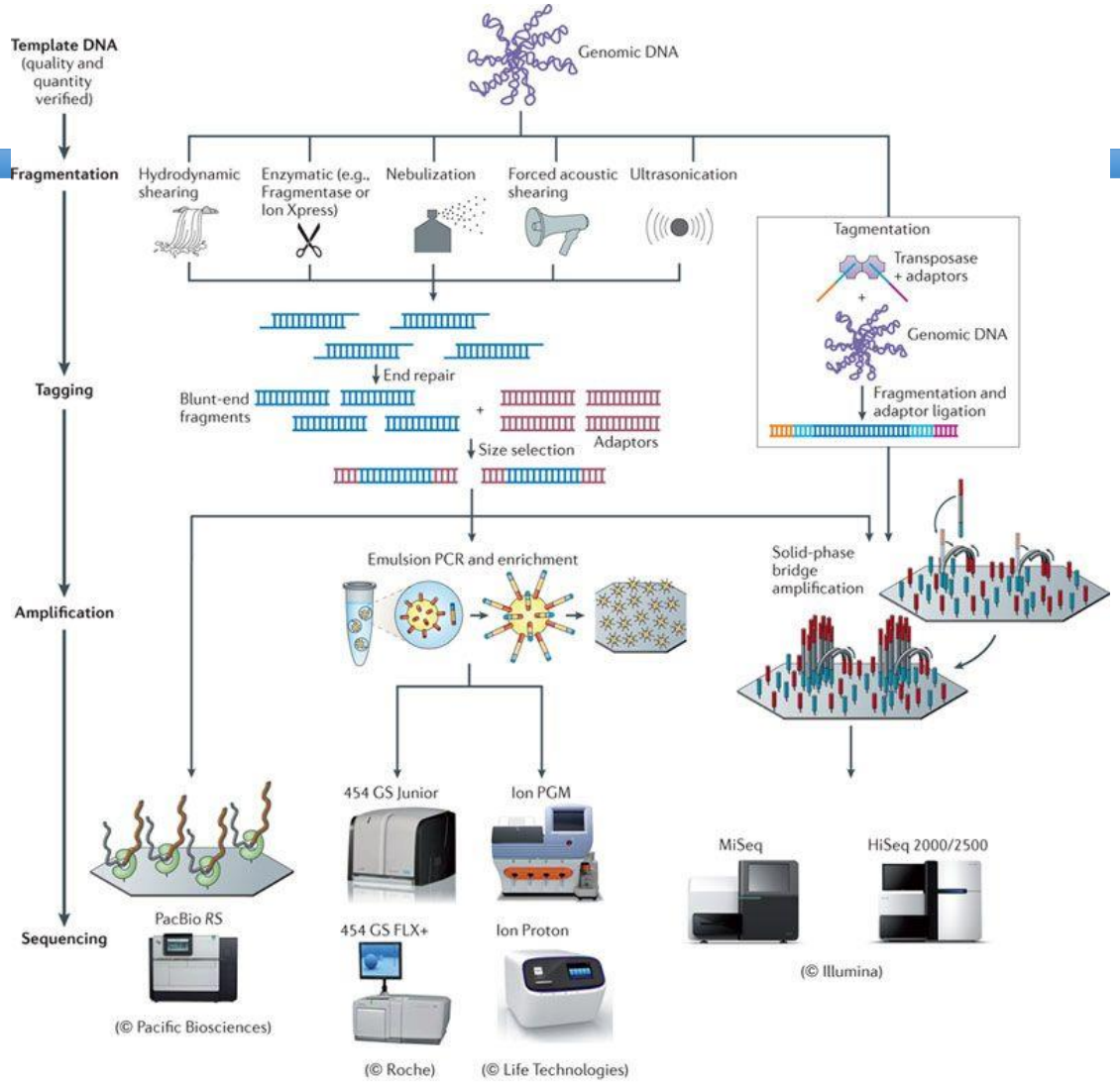
SECUENCIACIÓN GENOMAS

1. Metagenómica

IDENTIFICACIÓN MICROORGANISMOS

1. Metataxonomía

High-throughput sequencing platforms

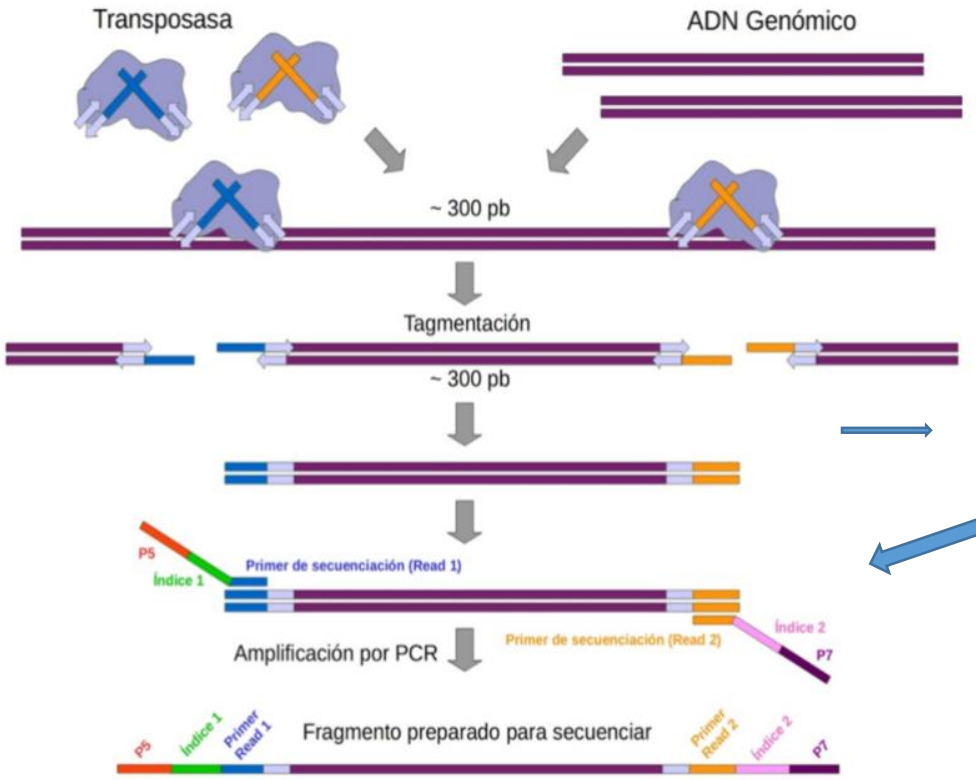


Nature Reviews | Microbiology

Loman et al, 2012

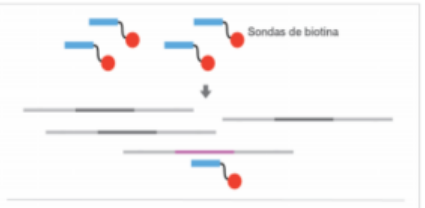
PREPARACIÓN LIBRERÍA

ENZIMÁTICA FÍSICA



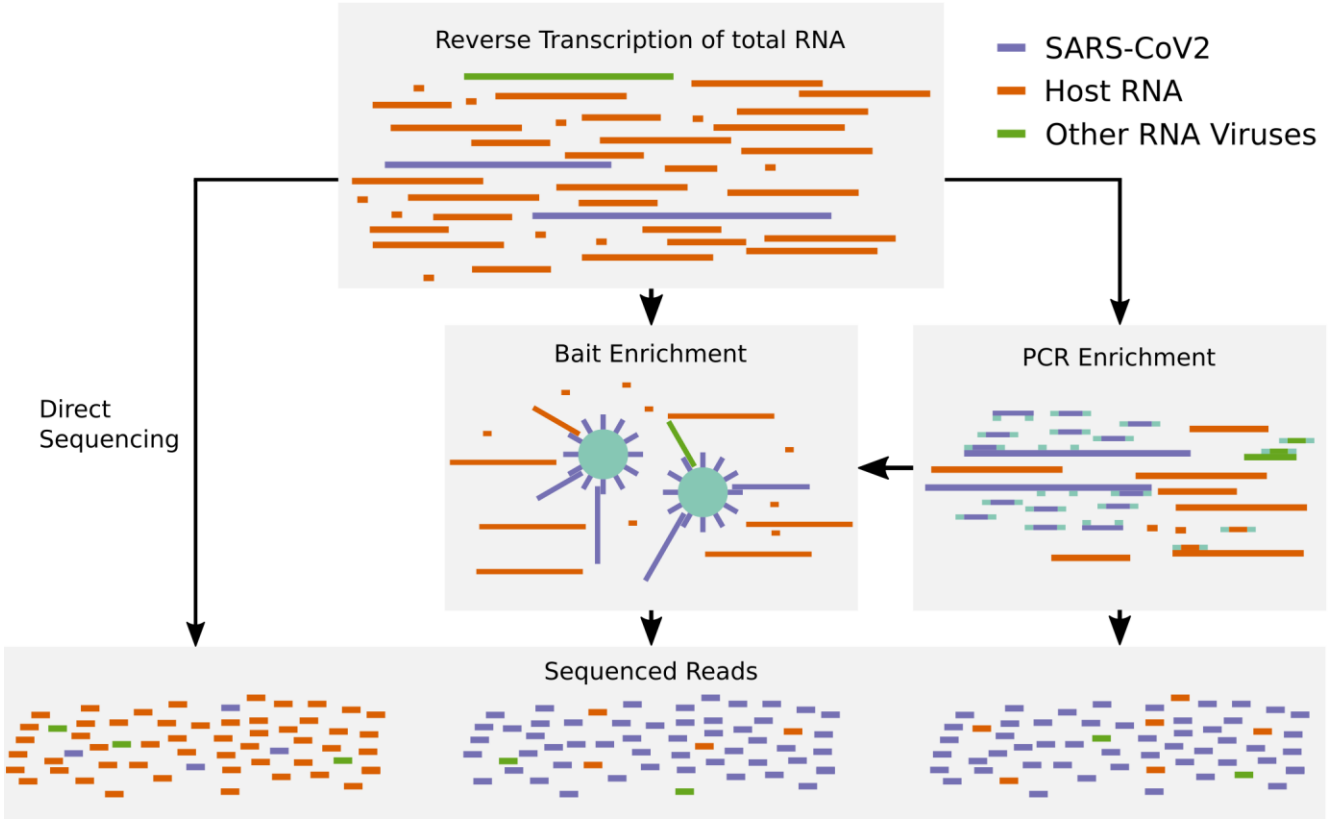
RNA -> cDNA

ENRIQUECIMIENTO:
PCR
CAPTURA SONDAS

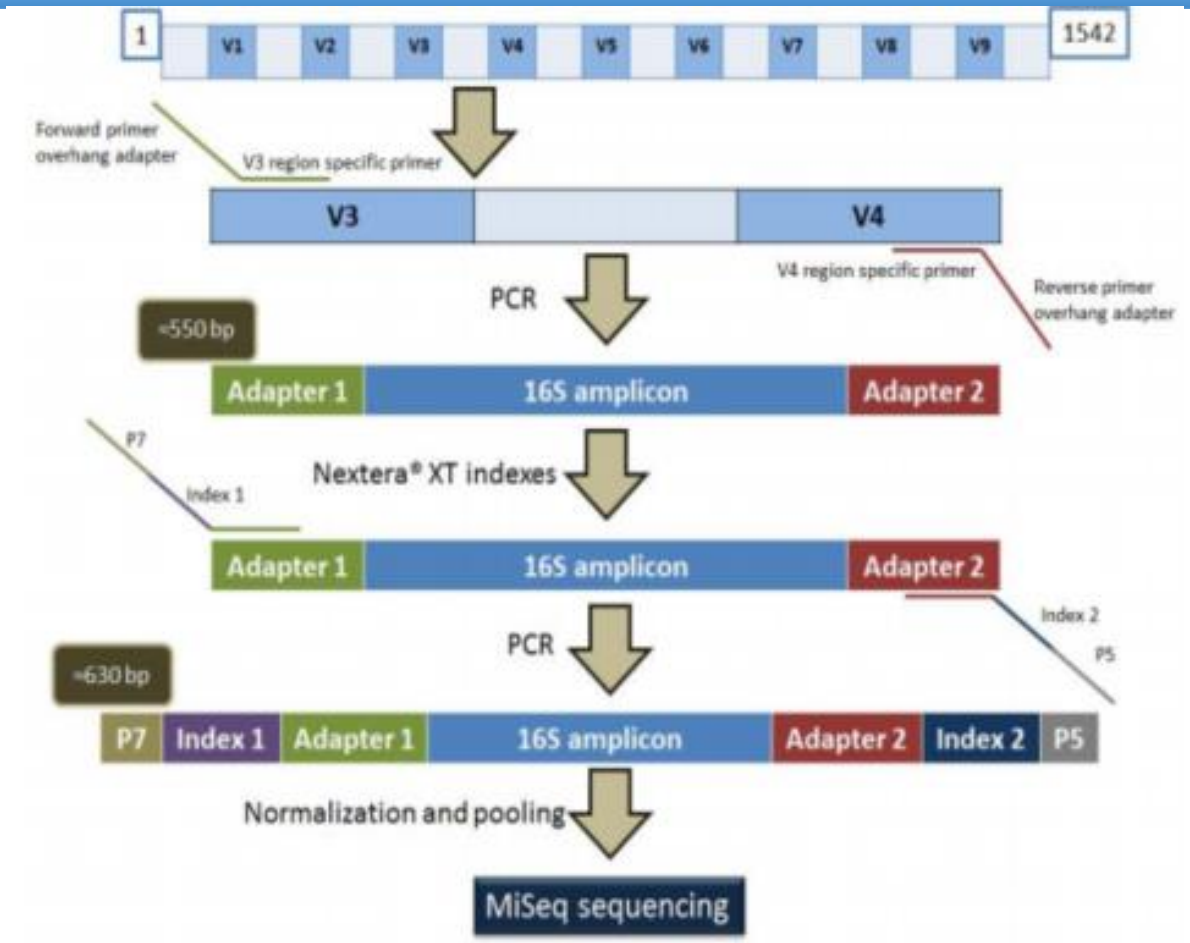


Adaptado de Guia Práctica Genómica https://www.uv.es/varnau/GM_Cap%C3%ADtulo_2.pdf

PREPARACIÓN LIBRERÍA



PREPARACIÓN LIBRERÍA, rRNA 16S, caracterización microbiota



Aplicaciones basadas en la preparación de librería

Sequencing Methods Review

A review of publications featuring Illumina® Technology

Aplicaciones basadas en la preparación de librería

Sequencing Methods Review

A review of publications featuring Illumina® Technology

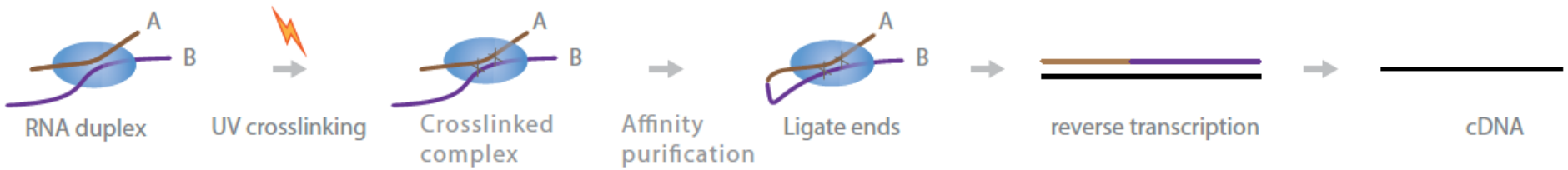
Aplicaciones basadas en la preparación de librería

Sequencing Methods Review

A review of publications featuring Illumina® Technology

CROSSLINKING, LIGATION, AND SEQUENCING OF HYBRIDS (CLASH-SEQ)

Crosslinking, ligation, and sequencing of hybrids (CLASH-Seq) maps RNA-RNA interactions[18]. In this method RNA-protein complexes are UV crosslinked and affinity-purified. RNA-RNA hybrids are then ligated, isolated, and reverse-transcribed to cDNA. Deep sequencing of the cDNA provides high-resolution chimeric reads of RNA-RNA interactions.



Pros	Cons
• *Maps RNA-RNA interactions* • *Performed in vivo*	• *Hybrid ligation may be difficult between short RNA fragments*

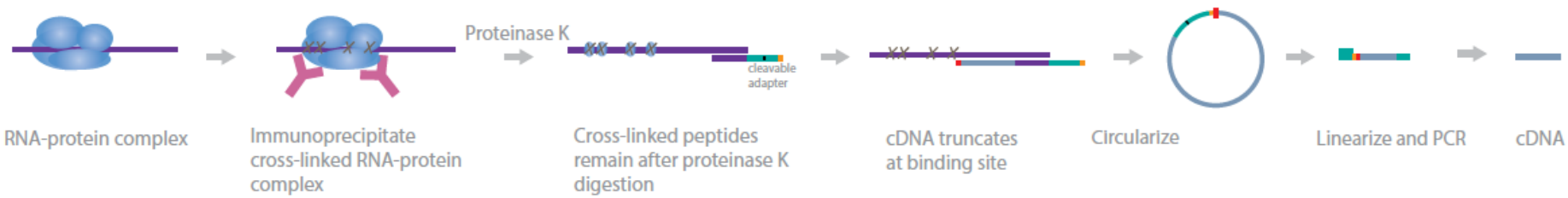
CHROMATIN ISOLATION BY RNA PURIFICATION (CHIRP-SEQ)

Chromatin isolation by RNA purification (ChIRP-Seq) is a protocol to detect the locations on the genome where non-coding RNAs (ncRNAs), such as long non-coding RNAs (lncRNAs), and their proteins are bound[7]. In this method, samples are first crosslinked and sonicated. Biotinylated tiling oligos are hybridized to the RNAs of interest, and the complexes are captured with streptavidin magnetic beads. After treatment with RNase H the DNA is extracted and sequenced. With deep sequencing the lncRNA/protein interaction site can be determined at single-base resolution.

Pros	Cons
• *Binding sites can be found anywhere on the genome* • *New binding sites can be discovered* • *Specific RNAs of interest can be selected*	• *Nonspecific oligo interactions can lead to misinterpretation of binding sites* • *Chromatin can be disrupted during the preparation stage* • *The sequence of the RNA of interest must be known*

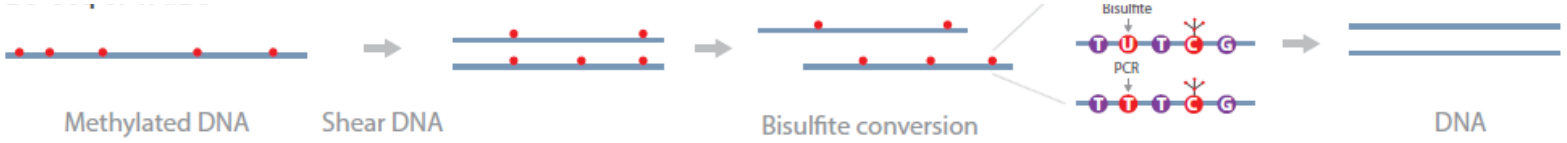
INDIVIDUAL NUCLEOTIDE RESOLUTION CLIP (ICLIP)

Individual nucleotide resolution CLIP (iCLIP) maps protein-RNA interactions similar to HITS-CLIP and PAR-CLIP[15]. This approach includes additional steps to digest the proteins after crosslinking and to map the crosslink sites with reverse transcriptase. In this method specific crosslinked RNA-protein complexes are immunoprecipitated. The complexes are then treated with proteinase K, as the protein crosslinked at the binding site remains undigested. Upon reverse transcription, cDNA truncates at the binding site and is circularized. These circularized fragments are then linearized and PCR-amplified. Deep sequencing of these amplified fragments provides nucleotide resolution of protein-binding site.

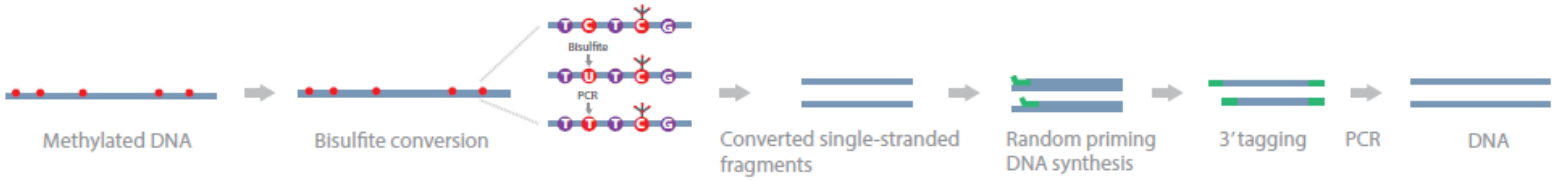


Pros	Cons
• *Nucleotide resolution of protein-binding site* • *Avoids the use of nucleases* • *Amplification allows the detection of rare events*	• *Antibodies not specific to target will precipitate nonspecific complexes* • *Non-linear PCR amplification can lead to biases affecting reproducibility* • *Artifacts may be introduced in the circularization step*

BISULFITE SEQUENCING (BS-SEQ)



EpiGnome Methyl-Seq



Pros	Cons
BS-Seq or WGBS	
• *CpG and non-CpG methylation throughout the genome is covered at single-base resolution* • *5mC in dense, less dense, and repeat regions are covered*	• *Bisulfite converts unmethylated cytosines to thymidines, reducing sequence complexity, which can make it difficult to create alignments* • *NPs where a cytosine is converted to thymidine will be missed upon bisulfite conversion* • *Bisulfite conversion does not distinguish between 5mC and 5hmC*
EpiGnome	
• *Pre-library bisulfite conversion* • *Low input gDNA (50 ng)* • *Uniform CpG, CHG, and CHH coverage* • *No fragmentation and no methylated adapters* • *Retention of sample diversity*	• *Bisulfite converts unmethylated cytosines to thymidines,reducing sequence complexity, which can make it difficult to create alignments* • *SNPs where a cytosine is converted to thymidine will be missed upon bisulfite conversion* • *Bisulfite conversion does not distinguish between 5mC and 5hmC* • *Higher duplicate percentage*

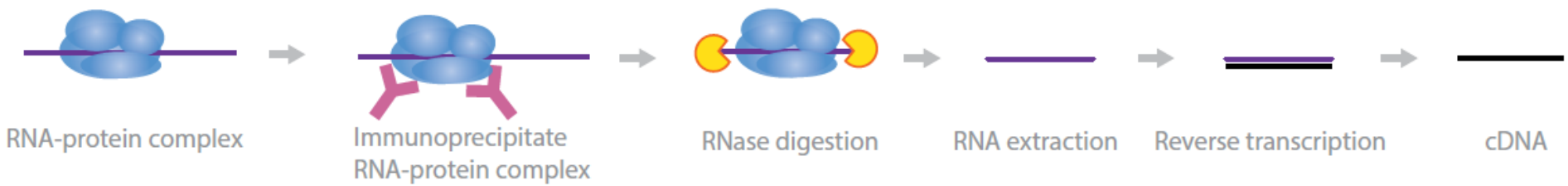
RIBOSOME PROFILING SEQUENCING (RIBO-SEQ)/ARTSEQ™

Active mRNA Translation Sequencing (ARTseq), also called ribosome profiling (Ribo-Seq), isolates RNA that is being processed by the ribosome in order to monitor the translation process[10]. In this method ribosome-bound RNA first undergoes digestion. The RNA is then extracted and the rRNA is depleted. Extracted RNA is reverse-transcribed to cDNA. Deep sequencing of the cDNA provides the sequences of RNAs bound by ribosomes during translation. This method has been refined to improve the quality and quantitative nature of the results. Careful attention should be paid to: (1) generation of cell extracts in which ribosomes have been faithfully halted along the mRNA they are translating in vivo; (2) nuclease digestion of RNAs that are not protected by the ribosome followed by recovery of the ribosome-protected mRNA fragments; (3) quantitative conversion of the protected RNA fragments into a DNA library that can be analyzed by deep sequencing[11]. The addition of harringtonine (an alkaloid that inhibits protein biosynthesis) causes ribosomes to accumulate precisely at initiation codons and assists in their detection.

Pros	Cons
• *Reveals a snapshot with the precise location of ribosomes on the RNA* • *Ribosome profiling more closely reflects the rate of protein synthesis than mRNA levels* • *No prior knowledge of the RNA or ORFs is required* • *The whole genome is surveyed* • *Can be used to identify protein-coding regions*	• *Initiation from multiple sites within a single transcript makes it challenging to define all ORFs* • *Does not provide the kinetics of translational elongation*

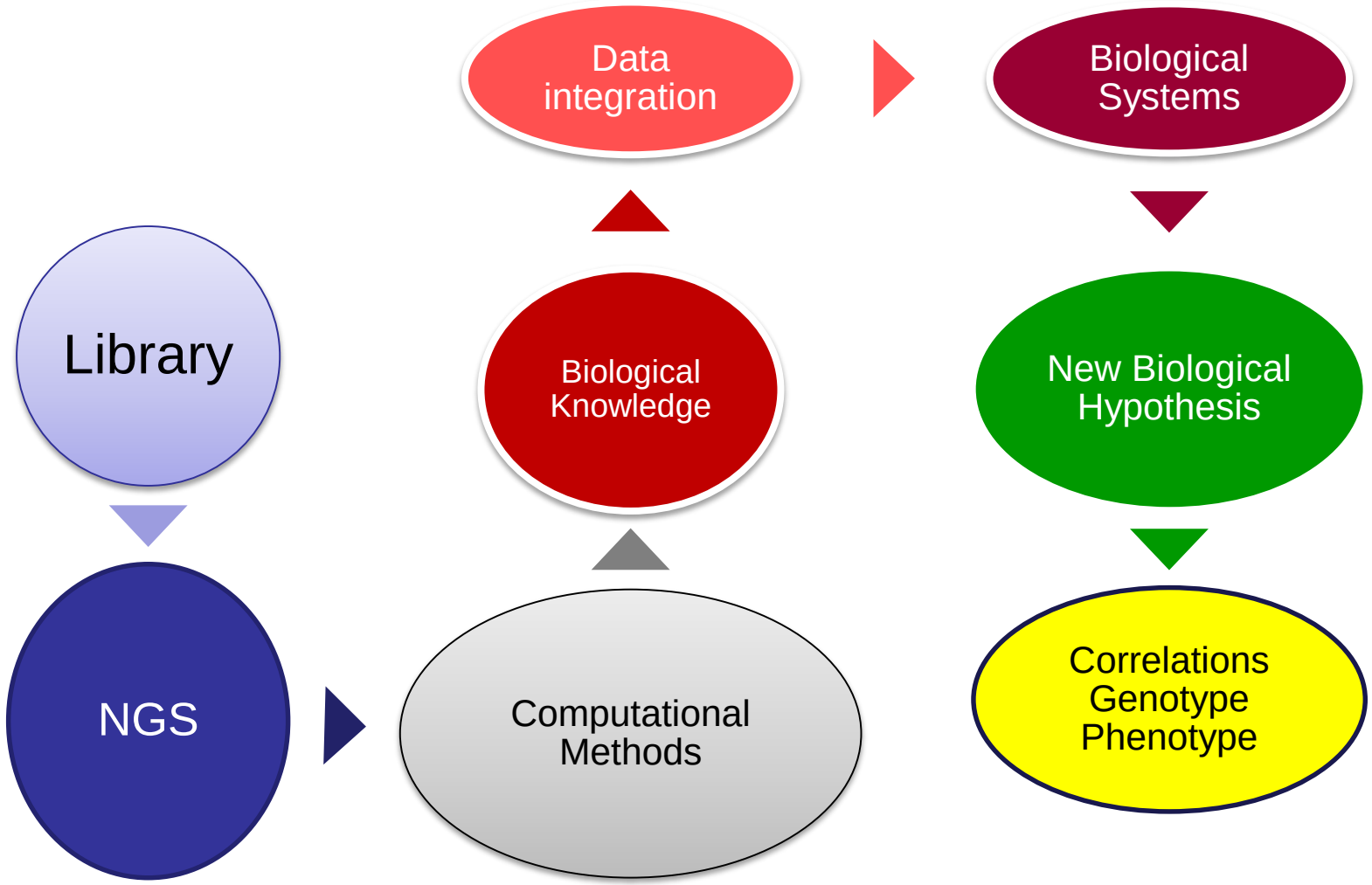
RNA IMMUNOPRECIPITATION SEQUENCING (RIP-SEQ)

RNA immunoprecipitation sequencing (RIP-Seq) maps the sites where proteins are bound to the RNA within RNA-protein complexes[12]. In this method, RNA-protein complexes are immunoprecipitated with antibodies targeted to the protein of interest. After RNase digestion, RNA covered by protein is extracted and reverse-transcribed to cDNA. The locations can then be mapped back to the genome. Deep sequencing of cDNA provides single-base resolution of bound RNA.



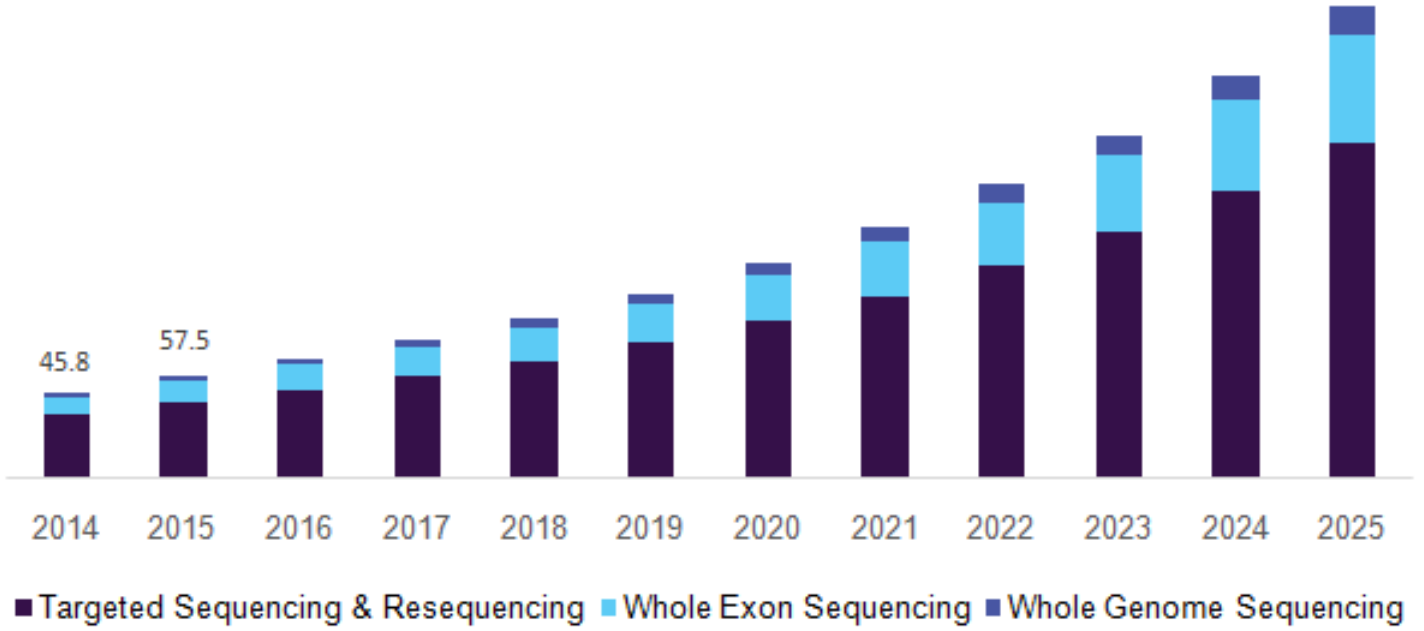
Pros	Cons
• *Maps specific protein-RNA complexes, such as polycomb-associated RNAs* • *Low background and higher resolution of binding site due to RNase digestion* • *No prior knowledge of the RNA is required* • *Genome-wide RNA screen*	• *Requires antibodies to the targeted proteins* • *Nonspecific antibodies will precipitate nonspecific complexes* • *Lack of crosslinking or stabilization of the complexes may lead to false negatives* • *RNase digestion must be carefully controlled*

IMPACT OF MASSIVE SEQUENCING

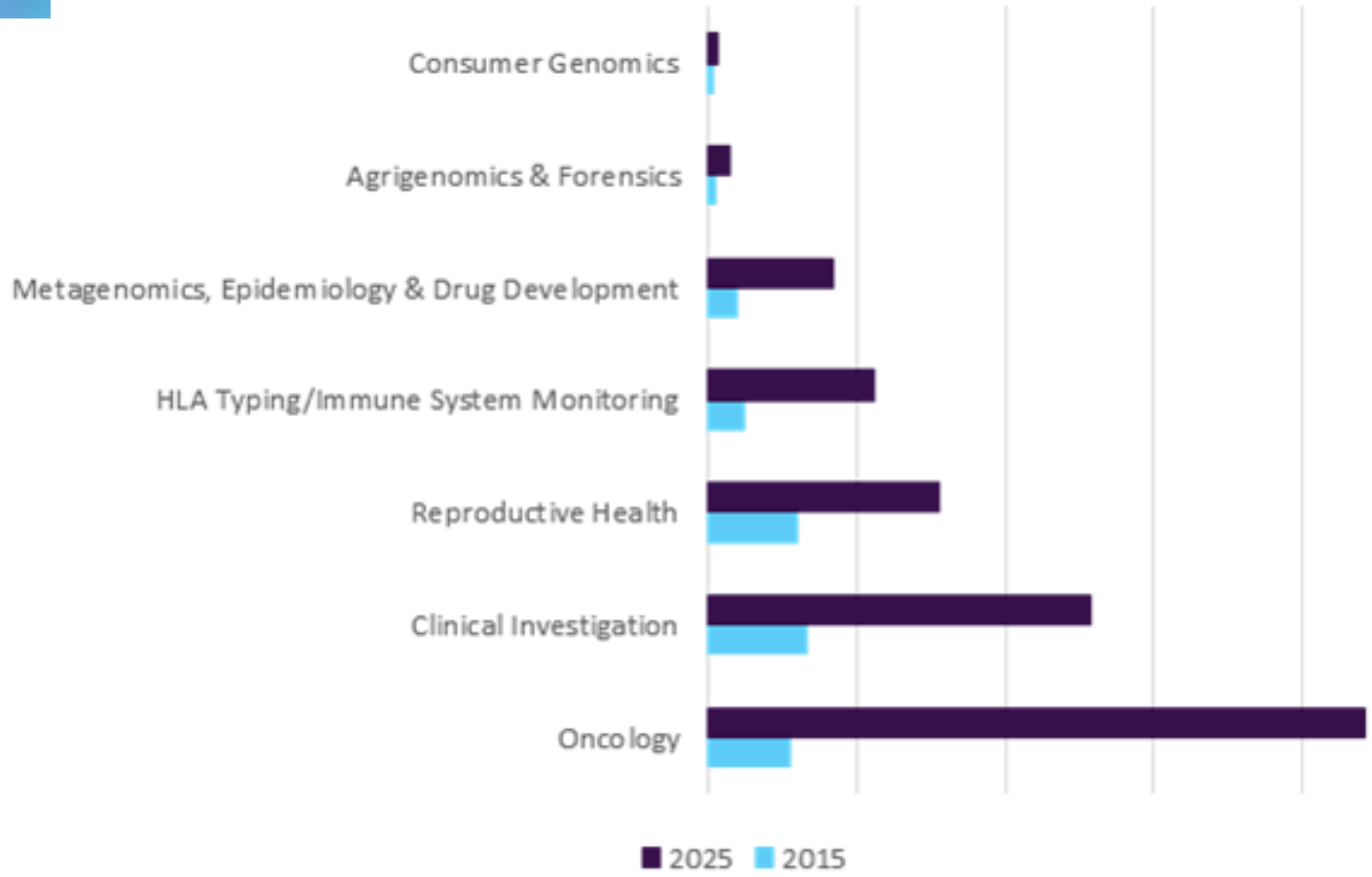


Next Generation Sequencing (NGS) Market Size & Forecast By Application (Oncology, Reproductive Health), By Technology (Targeted, WGS, WES), By Workflow (Data Analysis), By end-use (Academic & Clinical Research), And Trend Analysis, 2014 - 2025

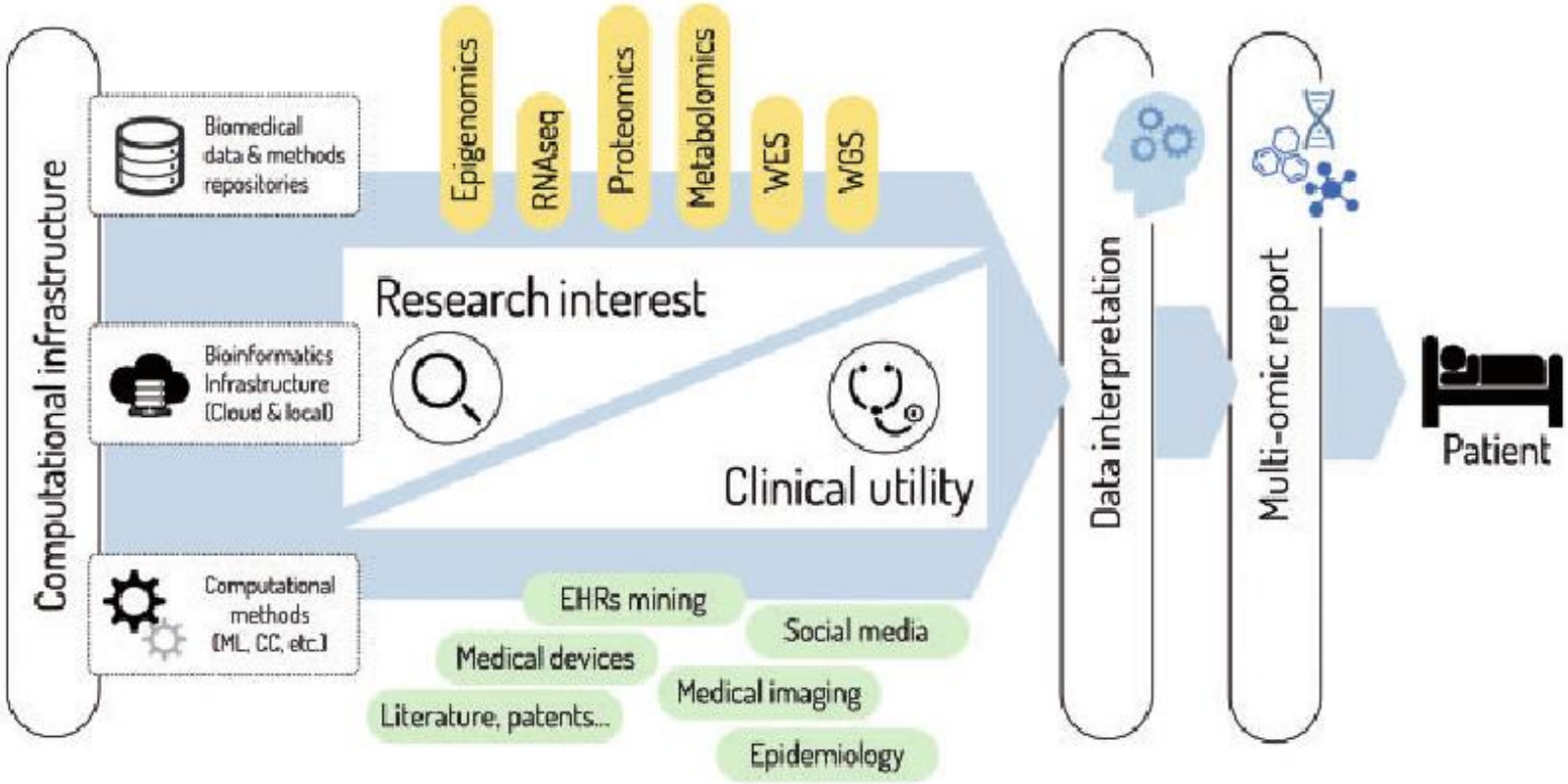
UK NGS market revenue, by technology, 2013 - 2024 (USD Million)



Global next generation sequencing market by applications, 2015 & 2025 (USD Million)

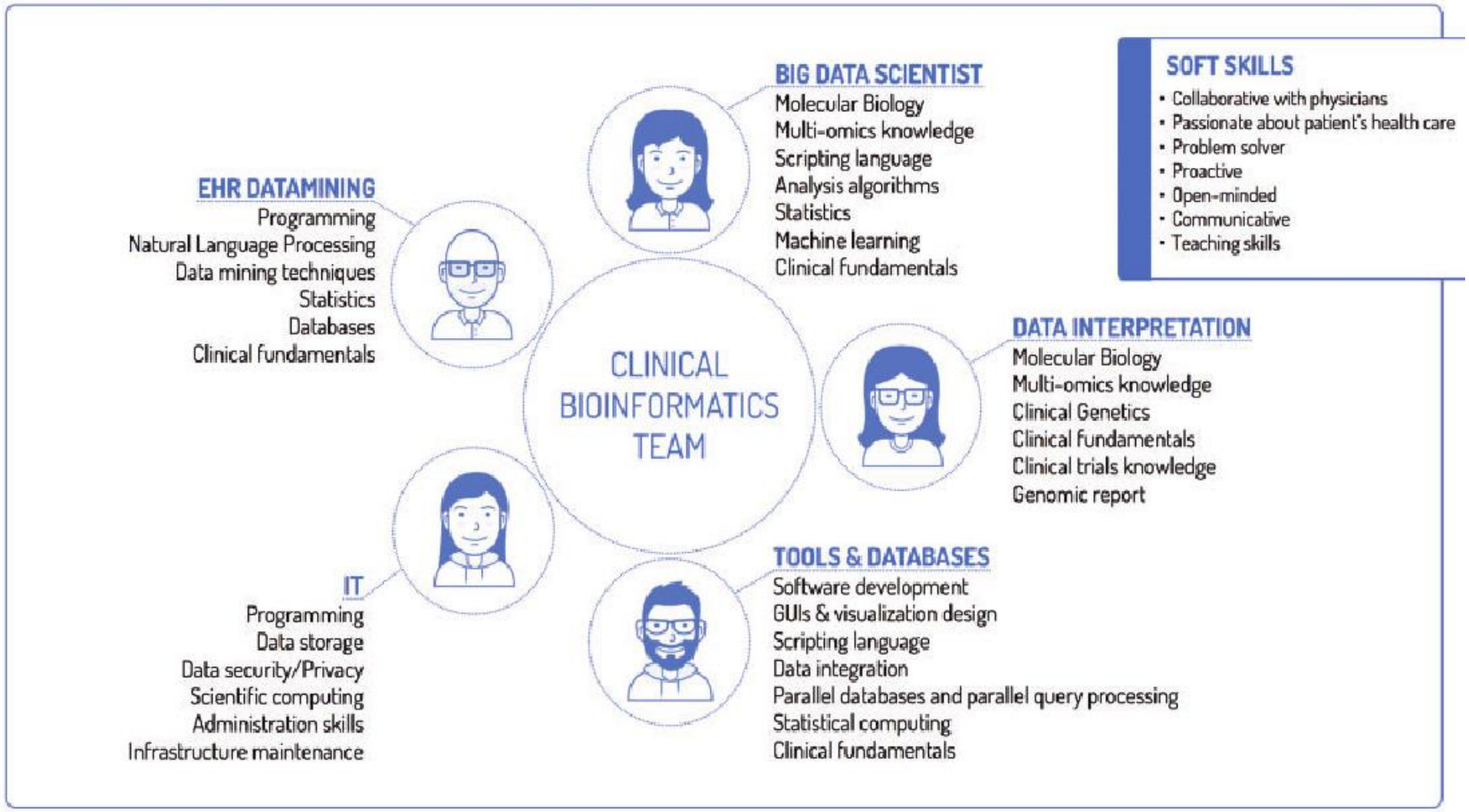


Precision Medicine workflow: from data to patient care



Gómez-López et al., Brief Bioinf 2017, 1-15

Clinical bioinformatics laboratory profile



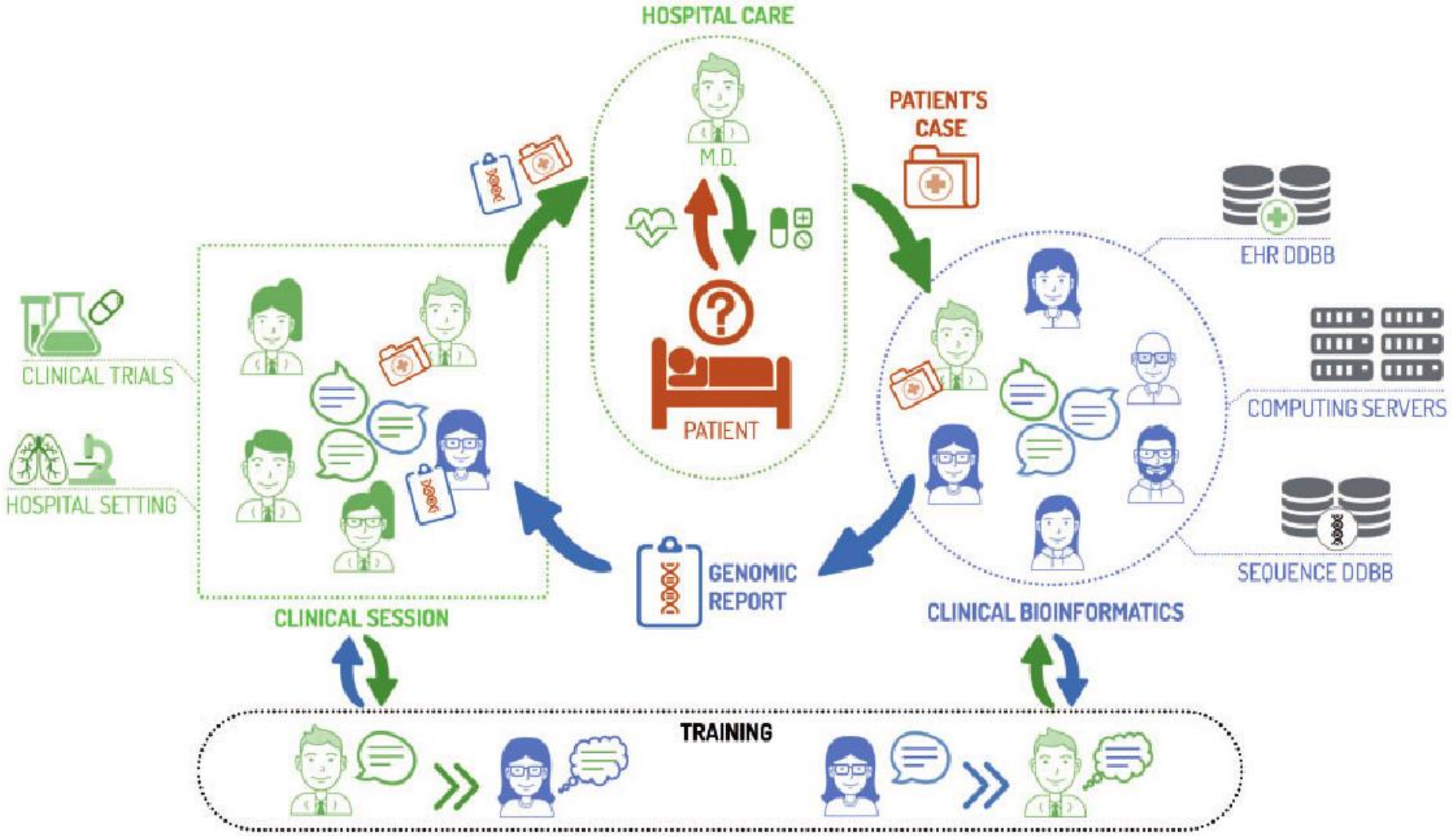
Skills for Clinical Bioinformaticians

Box 1. Fundamental technical skills for clinical bioinformaticians

1. Informatics
○ Experience in UNIX command line.
○ A basic programming language (i.e. Python). R as a useful language for handling statistics.
○ Knowledge of big data environments.
2. Life sciences
○ Understand the different types of biological data and databases.
○ Comprehend HTP data analysis methods.
○ Multi-omics data integration and interpretation.
3. Clinical scenario
○ Be familiar with EHRs, clinical terminology and medical procedures and protocols.
○ Get to know medical genomics: diagnosis, predictive and prognosis biomarkers.
○ Understand clinical trial design and monitoring.

Gómez-López et al., Brief Bioinf 2017, 1-15

Precision Medicine Workflow in Hospitals



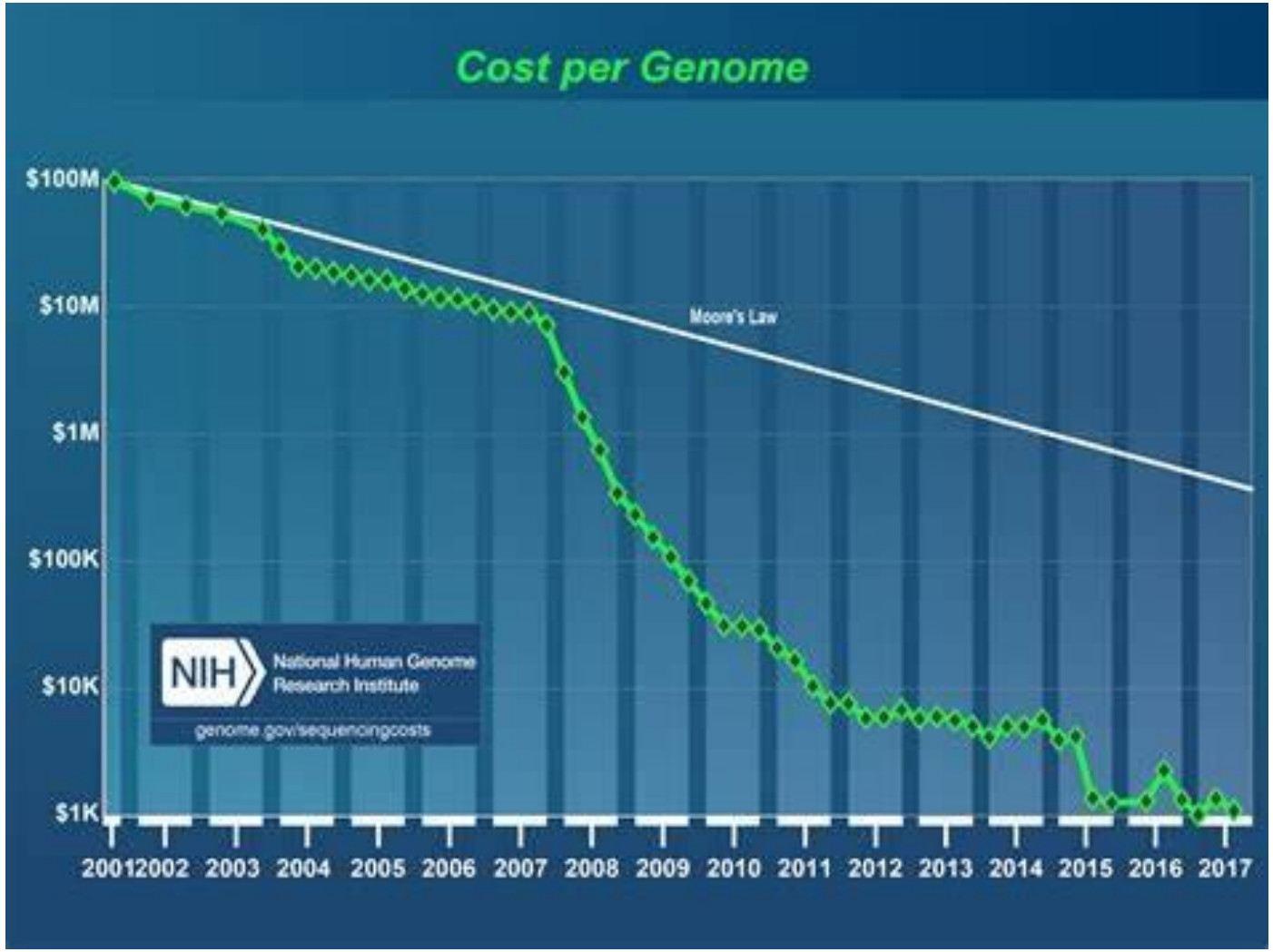
Retos de la Bioinformática en NGS

- Tecnología que evoluciona muy rápido
  - nuevos formatos de ficheros
  - nuevas aplicaciones
  - nuevos análisis

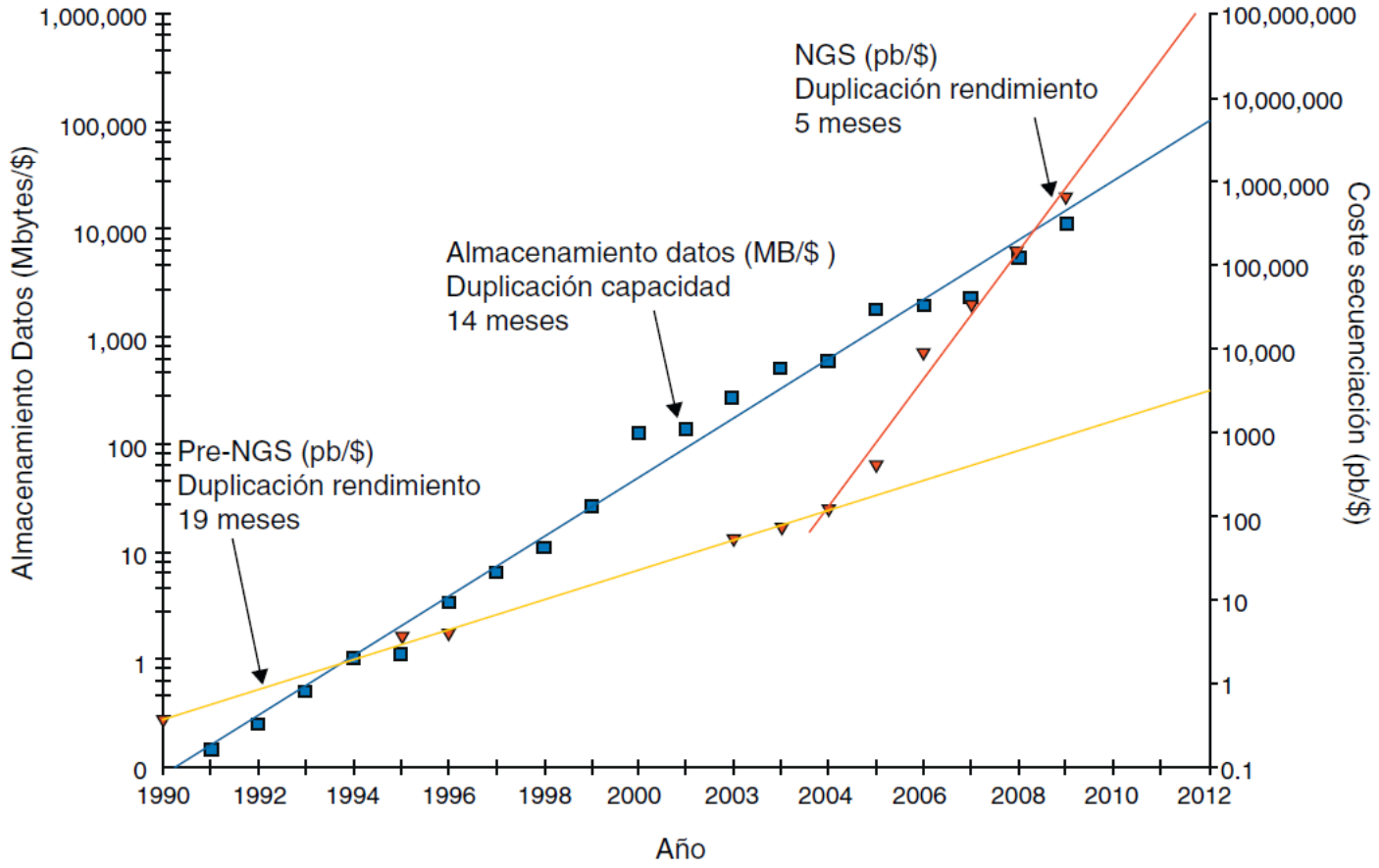
- Coste de la secuenciación disminuye
  - el embudo es el análisis de datos

- Adquisición de secuenciador debe ir ligado
  - a la compra de computo y contratación de bioinformático

Coste actual de la secuenciación



Costes del almacenamiento vs secuenciación



Adaptada de Stein, Genome Biology 2010, 11:207

Retos de la Bioinformática en NGS

- Necesidades de computo
  - ficheros de gran volumen (10Gb)
  - elevado uso de CPU y/o memoria
  - software no comercial en SO Unix

- Necesidades son dependientes de proyecto
  - No es lo mismo secuenciar un genoma 500Gb
  - que 50 genomas 25Tb

- Si el proyecto es la aplicación en clínica
  - Las necesidades de almacenamiento aumentan
  - por número de pacientes y por tiempo

Retos de la Bioinformática en NGS

- Desarrollo de BD curadas (confianza = reference)
- Algoritmos que resuelvan el problema biológico planteado.
- Necesidades de Bioinformáticos
  Análisis de los datos

Softwares comerciales en Bioinformática y NGS

Table I: Examples, features and comparisons of some commonly used commercial bioinformatics software suites

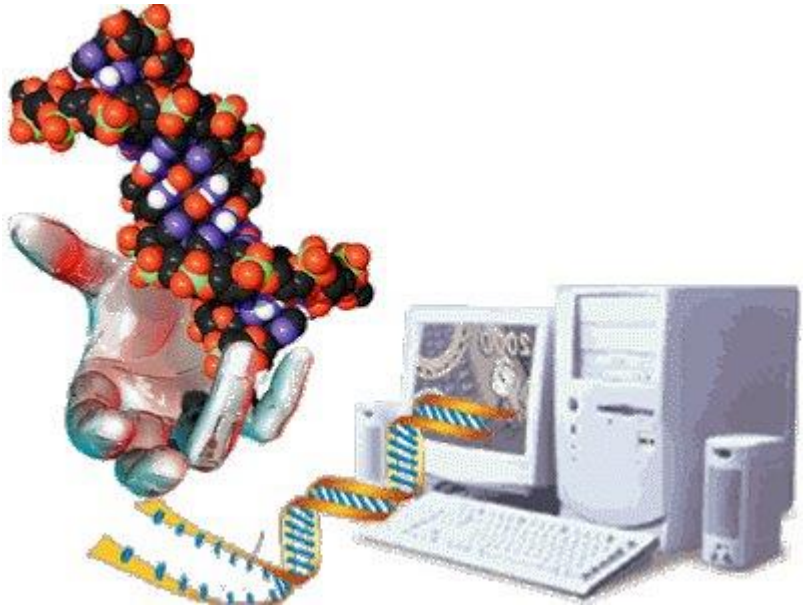
Software	Company	Cost (USD)[a]	Free trial (days)	Platform[b]	NGS analyses[c]	Evolutionary analyses[d]	Database searching[e]	Plug-ins	Workflows	Teaching suitability
Avadis NGS	Strand Scientific Intelligence	$4500	20	M, W, L	✓	×	×	×	✓	×
CLC Genomics Workbench	CIC bio, Qiagen	$5500	30	M, W, L	✓	✓	✓	✓	✓	✓
CodonCode Aligner	CodonCode	$720	30	M, W	✓	✓	×	×	×	✓
Genamics Expression	Genamics	$295	30	W	×	✓	✓	✓	×	×
Geneious	Biomatters	$795	14	M, W, L	✓	✓	✓	✓	✓	✓
Full Lasergene Suite	DNASTAR	$5950	30	M, W	✓	✓	✓	✓	✓	✓
MacVector & Assembler	MacVector	$300	21	M	✓	✓	✓	×	×	✓
NextGENe	Softgenetics	$4049	35	W	✓	×	×	×	×	×
Sequencher	Gene Codes	$2500	30	M, W	✓	✓	✓	✓	×	✓
VectorNTI Advance	Life Technologies	$600	30	W	×	✓	✓	×	✓	✓

Softwares en Bioinformática y NGS

- Tecnología que evoluciona muy rápido
  - nuevos formatos de ficheros
  - nuevas aplicaciones
  - nuevos análisis
  - nuevos algoritmos

- Software en continuo desarrollo (Unix)

Gracias por la atención
Preguntas ???



Isabel Cuesta

Unidad de Bioinformática – Unidades Científico Técnicas - ISCIII

isabel.cuesta@isciii.es