Sesión 2 - Secuenciación Masiva Aplicaciones

Isabel Cuesta

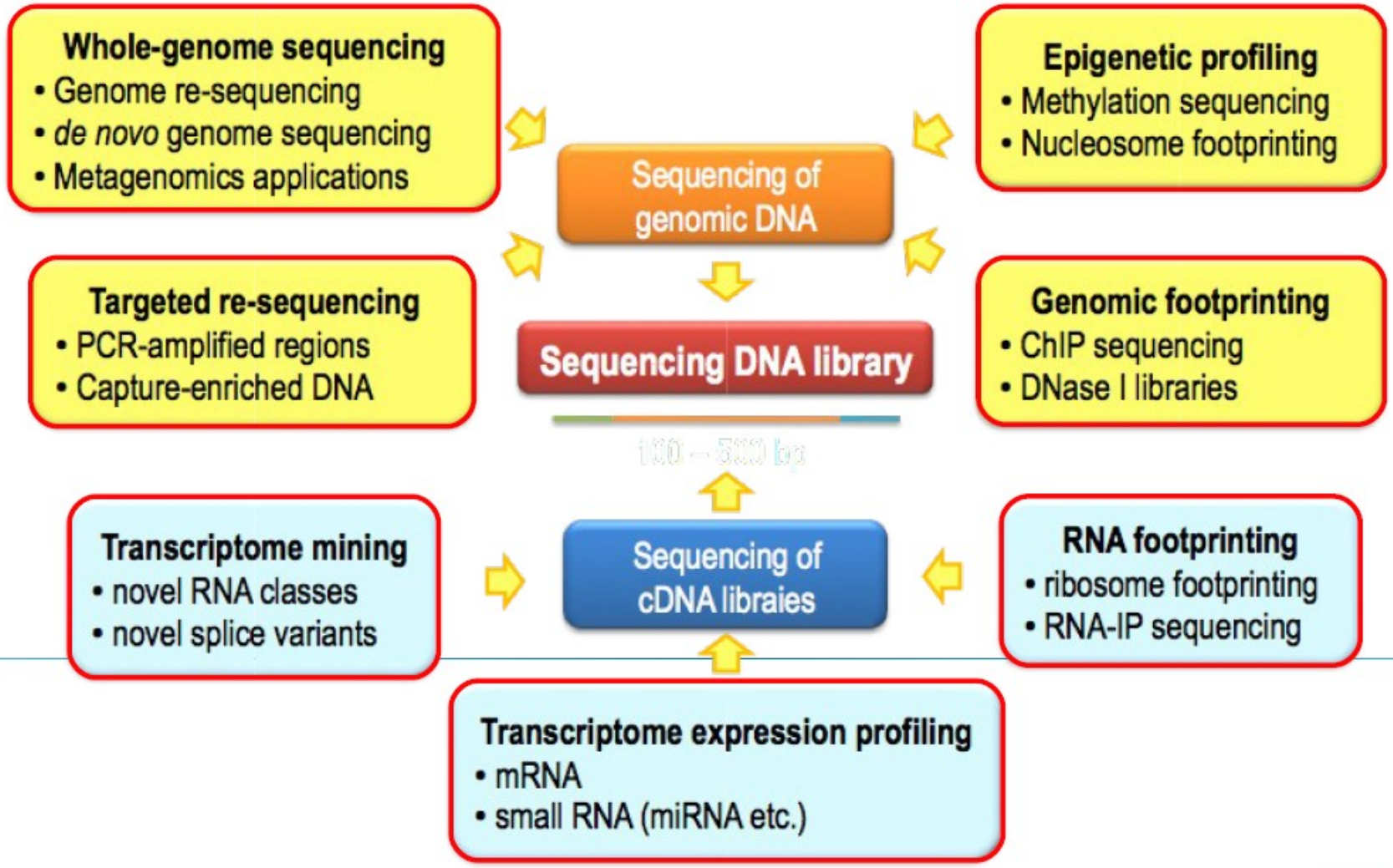
Unidad de Bioinformática (BU-ISCIII)
Unidades Comunes Científico Técnicas – SGAFI-ISCIII

17-21 Junio 2019, 7ª 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 secuenciació



Aplicaciones basadas en la prepa

Sequencing Methods Review

A review of publications featuring Illumina® Technology

Aplicaciones basadas en la prepa

Sequencing Methods Review

A review of publications featuring Illumina® Technology

Aplicaciones basadas en la prepa

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.

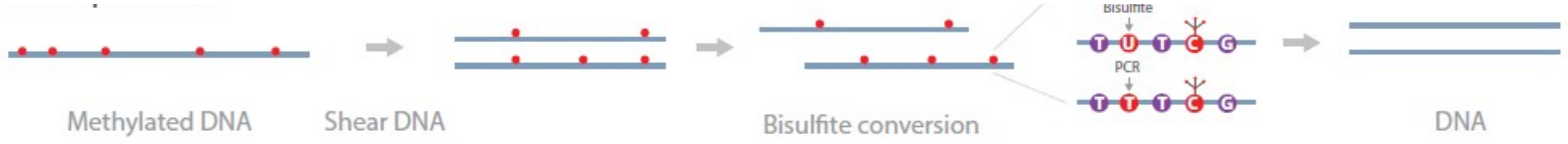
Pros	Cons
• *Binding sites can be found anywhere on the genome* • *New binding sites can be discovered* • *Specific RNAs of interest can be selected*	• *Nonspecific oligo interactions can lead to misinterpretation of binding sites* • *Chromatin can be disrupted during the preparation stage* • *The sequence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

INDIVIDUAL NUCLEOTIDE RESOLUTION CLIP (ICLIP)

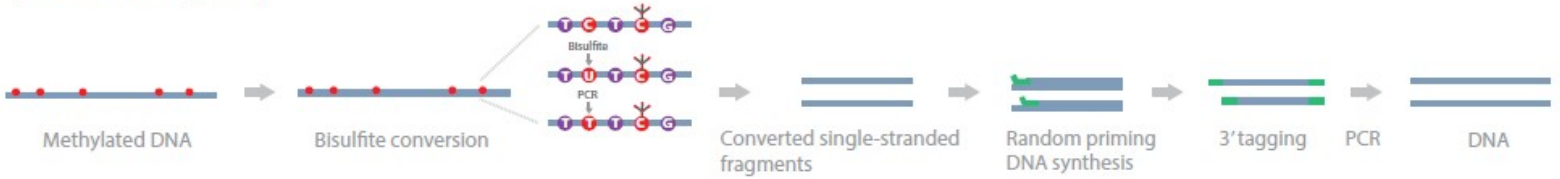
Individual nucleotide resolution CLIP (iCLIP) maps protein-RNA interactions similar to HITS-CLIP and PAR-CLIP[15]. This approach includes additional steps to digest the proteins after crosslinking and to map the crosslink sites with reverse transcriptase. In this method specific crosslinked RNA-protein 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*	• *Bisulfite converts unmethylated cytosines to thymidines, reducing sequence complexity, which can make it difficult to create alignments*
• *5mC in dense, less dense, and repeat regions are covered*	• *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*	• *Bisulfite converts unmethylated cytosines to thymidines,reducing sequence complexity, which can make it difficult to create alignments*
• *Low input gDNA (50 ng)*	• *SNPs where a cytosine is converted to thymidine will be missed upon bisulfite conversion*
• *Uniform CpG, CHG, and CHH coverage*	• *Bisulfite conversion does not distinguish between 5mC and 5hmC*
• *No fragmentation and no methylated adapters*	• *Higher duplicate percentage*
• *Retention of sample diversity*	

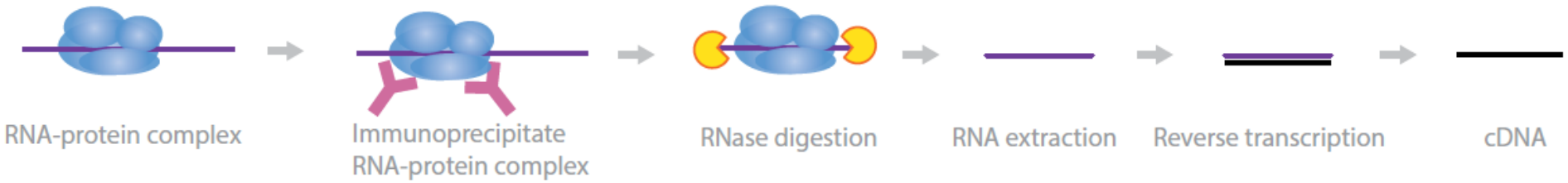
RIBOSOME PROFILING SEQUENCING (RIBO-SEQ)/ARTSEQ™

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

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.



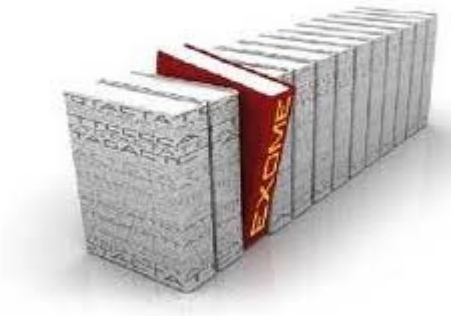
Pros	Cons
• *Maps specific protein-RNA complexes, such as polycomb-associated RNAs* • *Low background and higher resolution of binding site due to RNase digestion* • *No prior knowledge of the RNA is required* • *Genome-wide RNA screen*	• *Requires antibodies to the targeted proteins* • *Nonspecific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

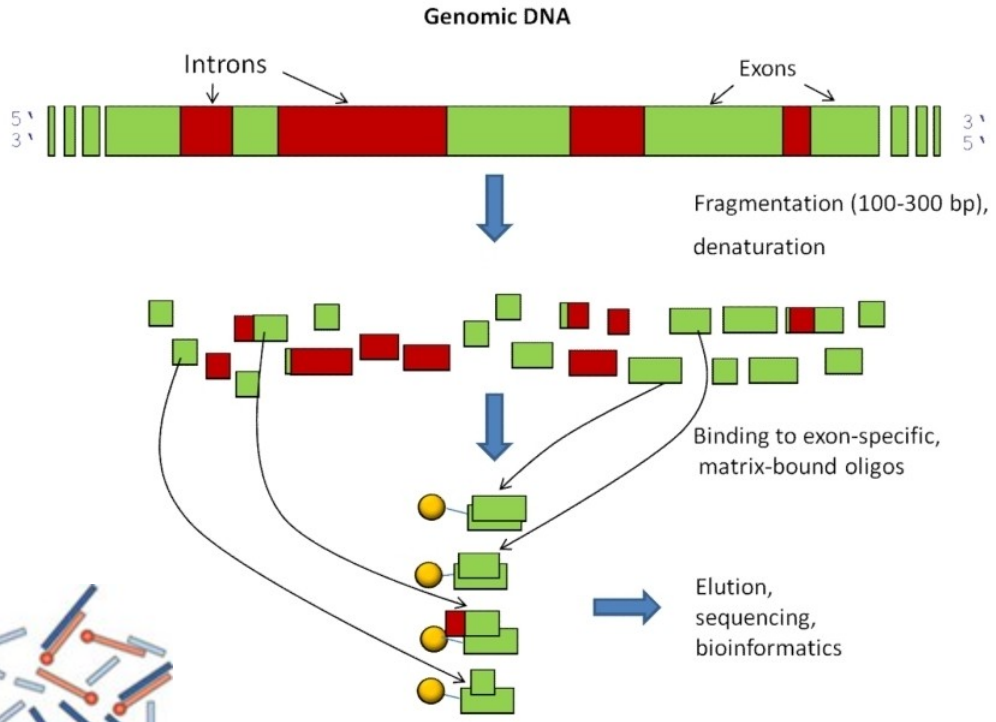
WHOLE-TRANSCRIPT AMPLIFICATION FOR SINGLE CELLS (QUARTZ-SEQ)

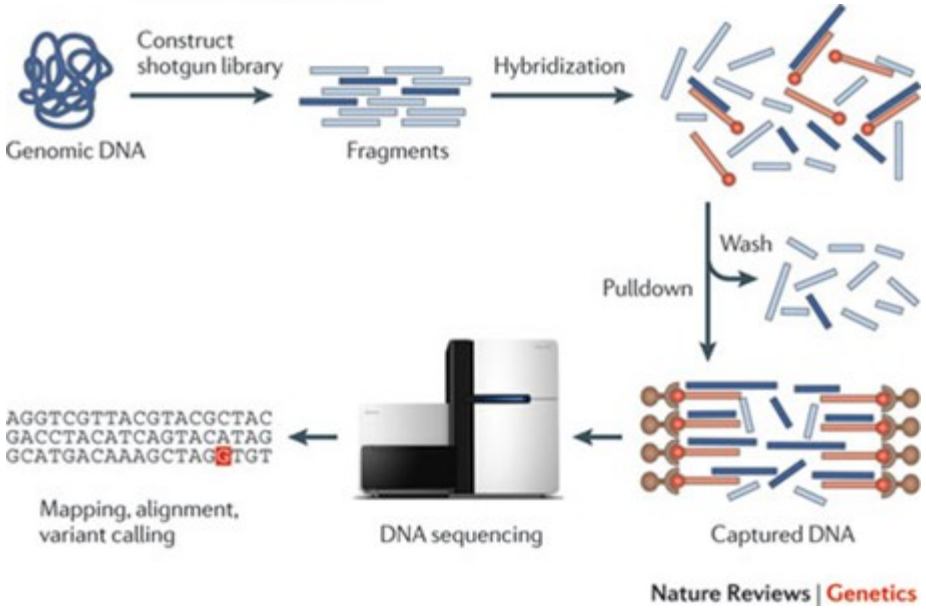
The Quartz-Seq method optimizes whole-transcript amplification (WTA) of single cells[35]. In this method, a reverse-transcription (RT) primer with a T7 promoter and PCR target is first added to extracted mRNA. Reverse transcription synthesizes first-strand cDNA, after which the RT primer is digested by exonuclease I. A poly(A) tail is then added to the 3' ends of first-strand cDNA, along with a dT primer containing a PCR target. After second-strand generation, a blocking primer is added to ensure PCR enrichment in sufficient quantity for sequencing. Deep sequencing allows for accurate, high-resolution representation of the whole transcriptome of a single cell.

Pros	Cons
• *Single-tube reaction suitable for automation*	• *PCR biases can underrepresent GC-rich templates*
• *Digestion of RT primers by exonuclease I eliminates amplification of byproducts*	• *Amplification errors caused by polymerases will be represented and sequenced incorrectly*
• *Short fragments and byproducts are suppressed during enrichment*	• *Targets smaller than 500 bp are preferentially amplified by polymerases during PCR*

EXOME







30-50 Mb
Human Exome
$1000

Genoma, Exoma, Panel? desde

PANEL

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

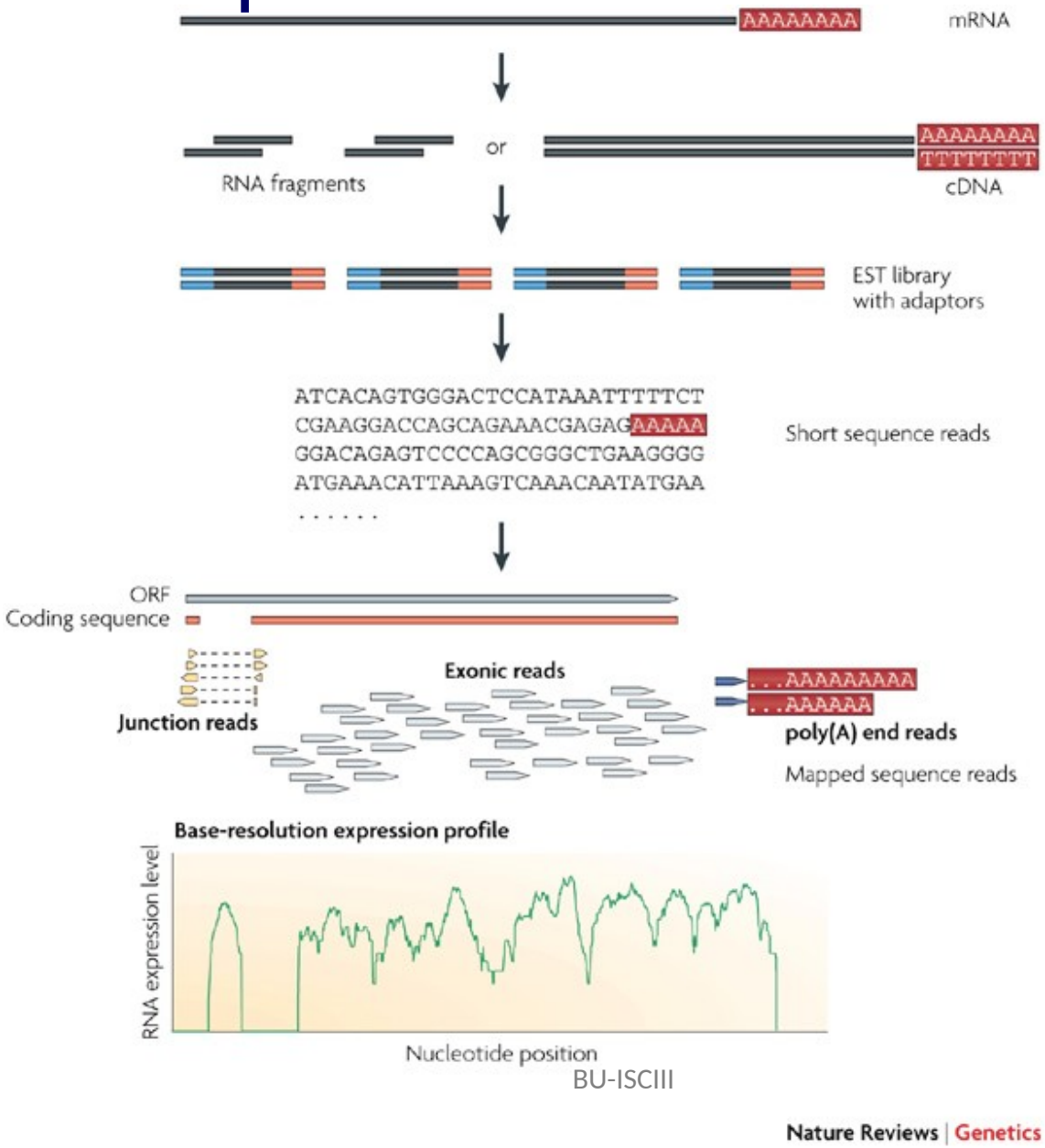
EXOMA

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

GENOMA

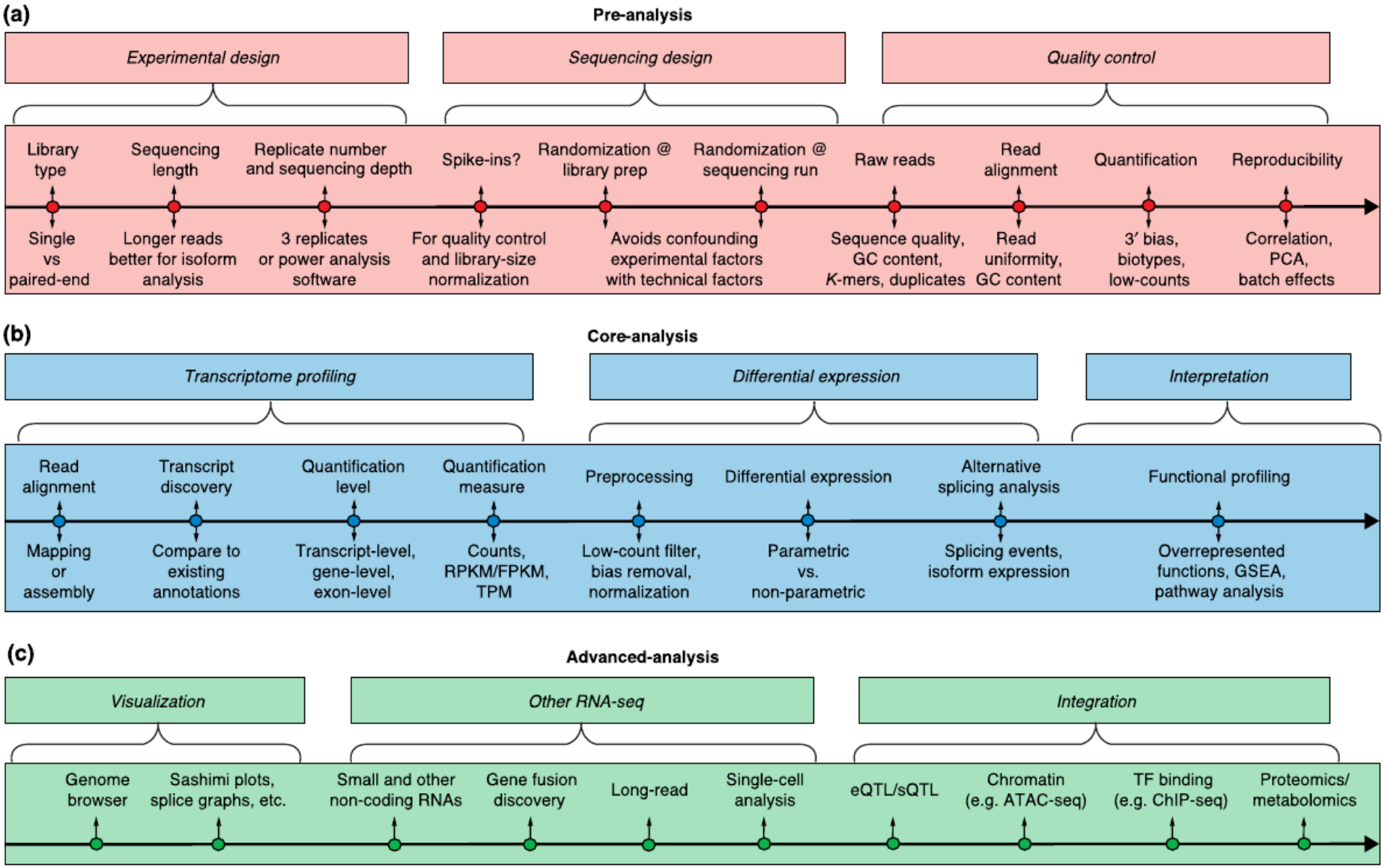
- Maxima complejidad en secuenciación y coste
- Información de regiones no codificantes
- Análisis de variaciones estructurales
- Elevado volumen de datos

RNA seq



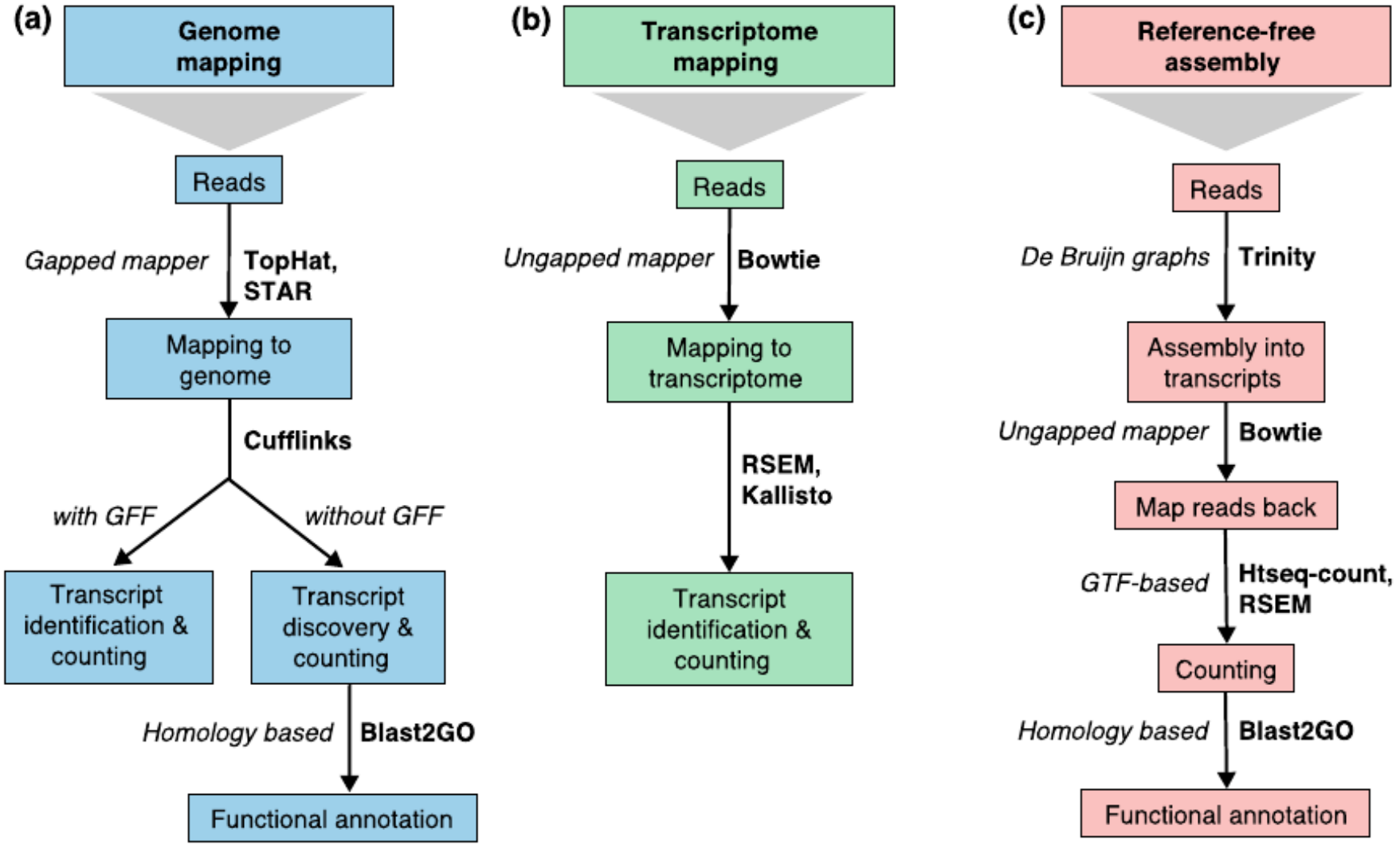
Nature Reviews | Genetics

RNA seq



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

RNA seq: transcript identification strategies

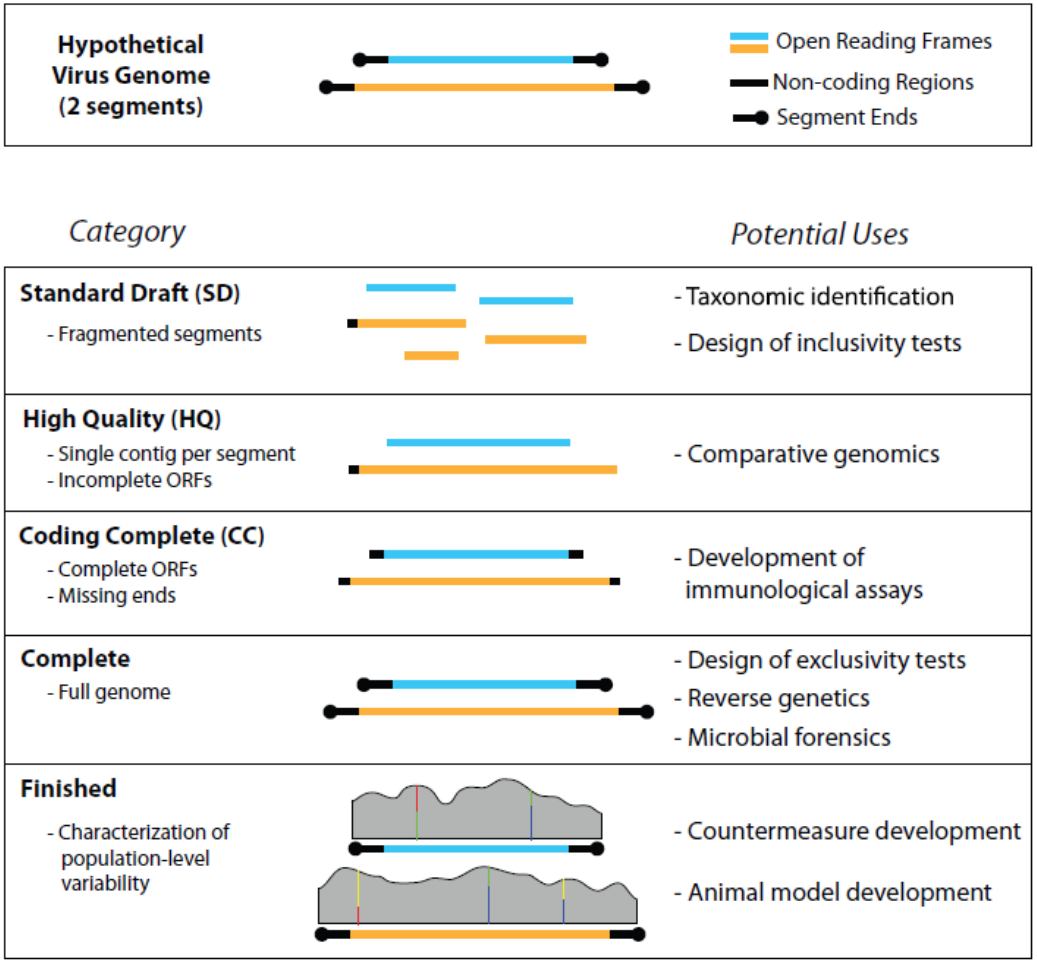


Conesa et al., Genome Biology (2016) 17:13

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

Secuenciación genomas, estar

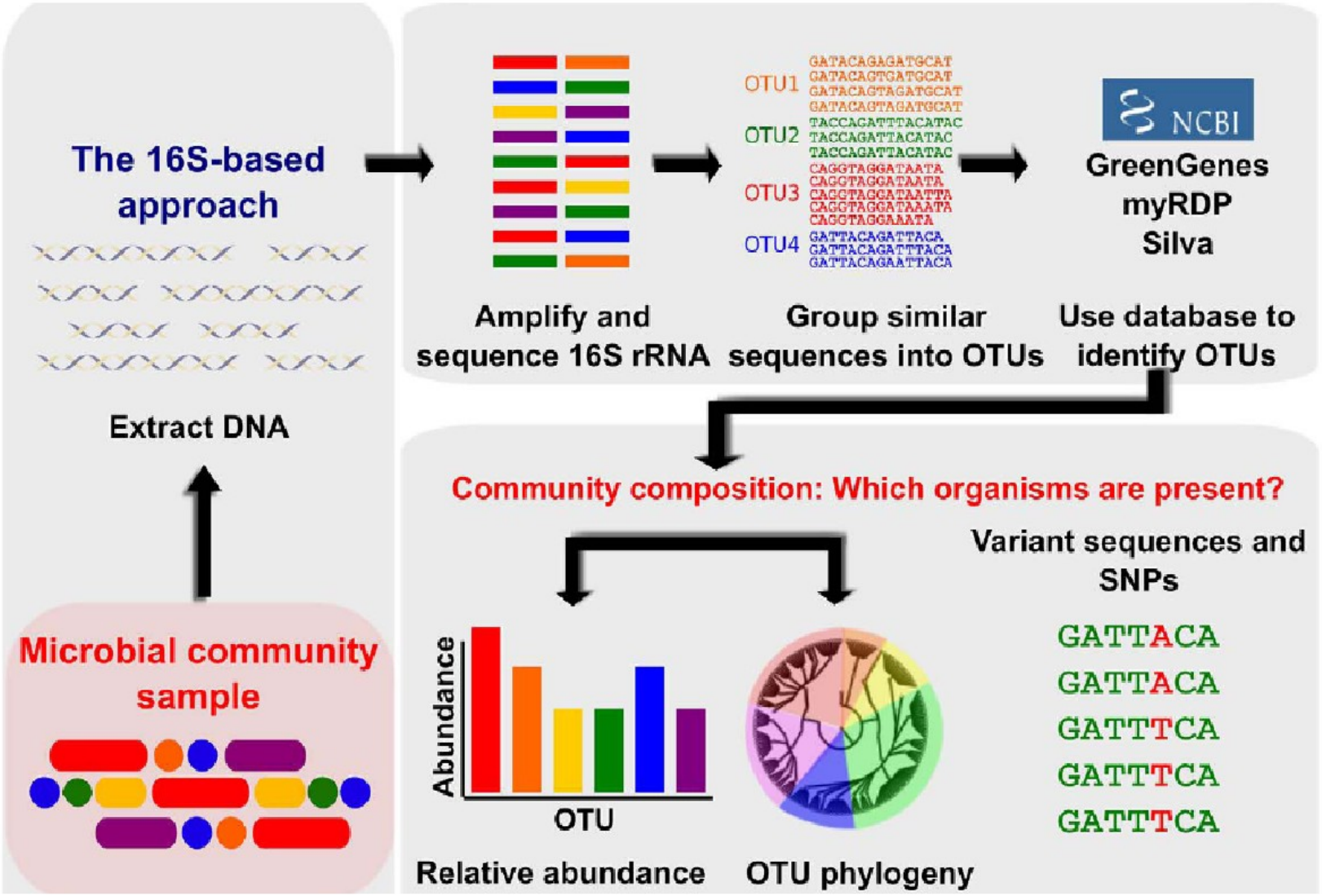


Ladner et al., mbio May / June 2014, 5:3, e01360-14

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



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

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:

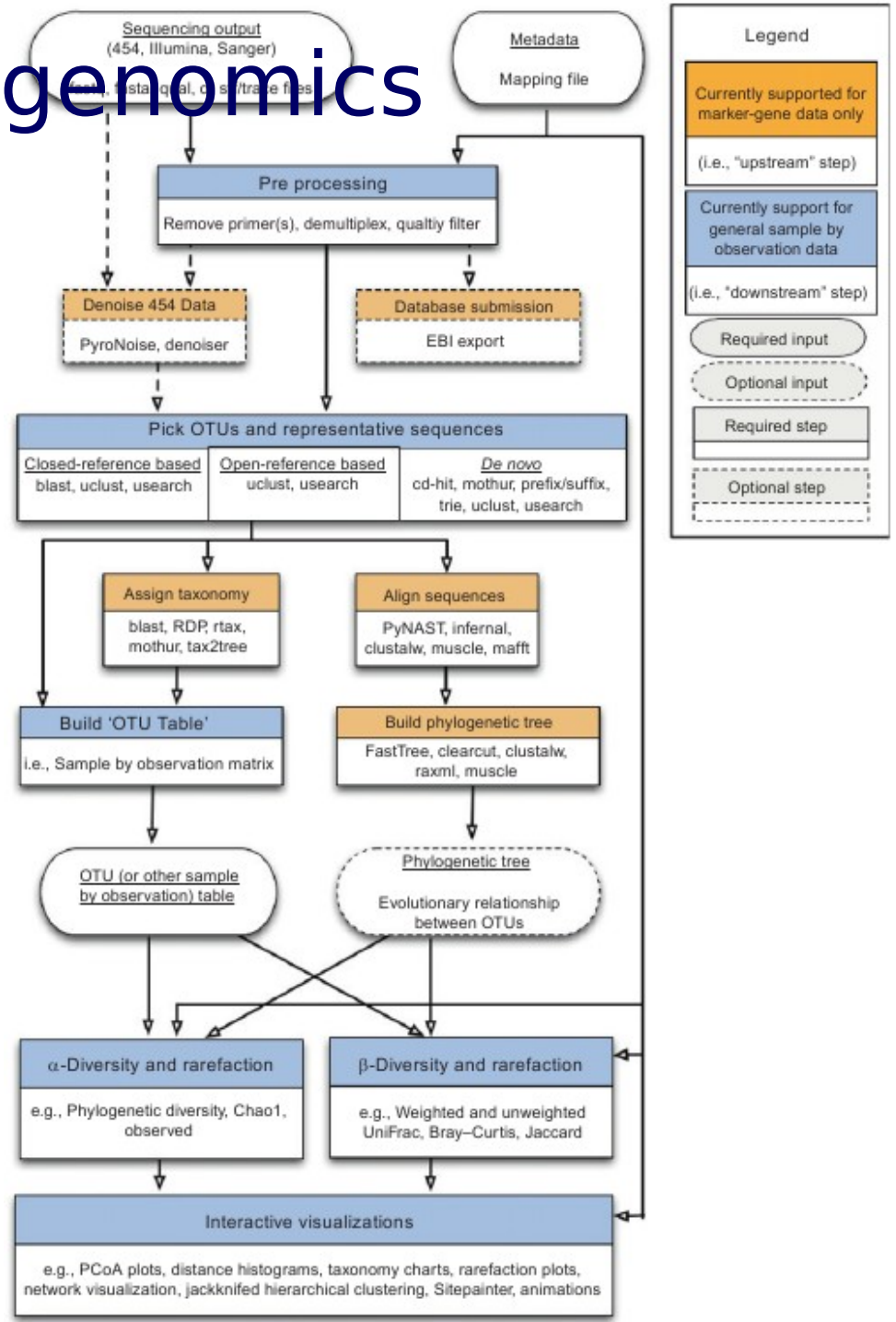
- mothur: `http://www.mothur.org`
- QIIME: `http://qiime.org`

Targeted metagenomics

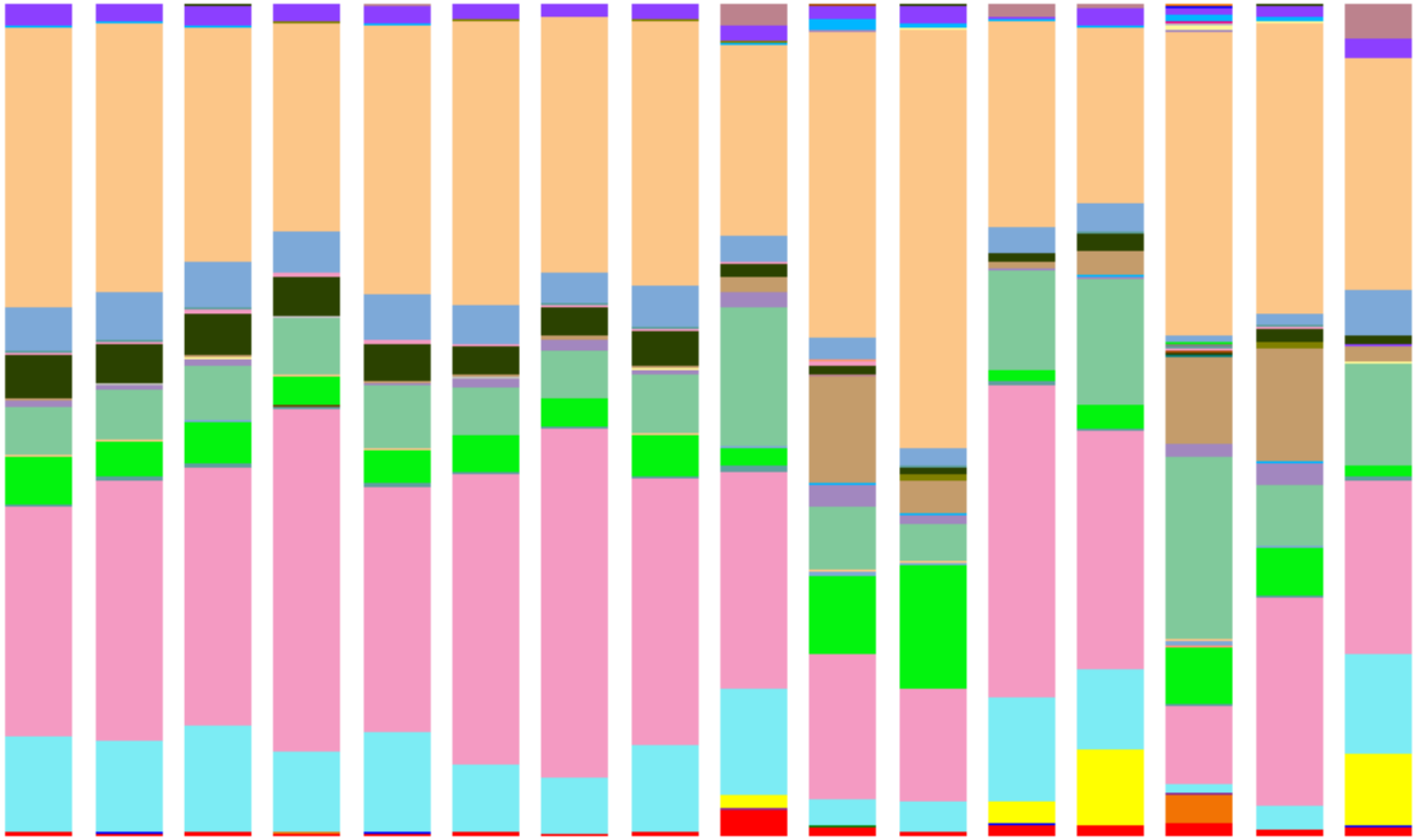
QIIME WORKFLOW

Integrated pipeline of third-party tools

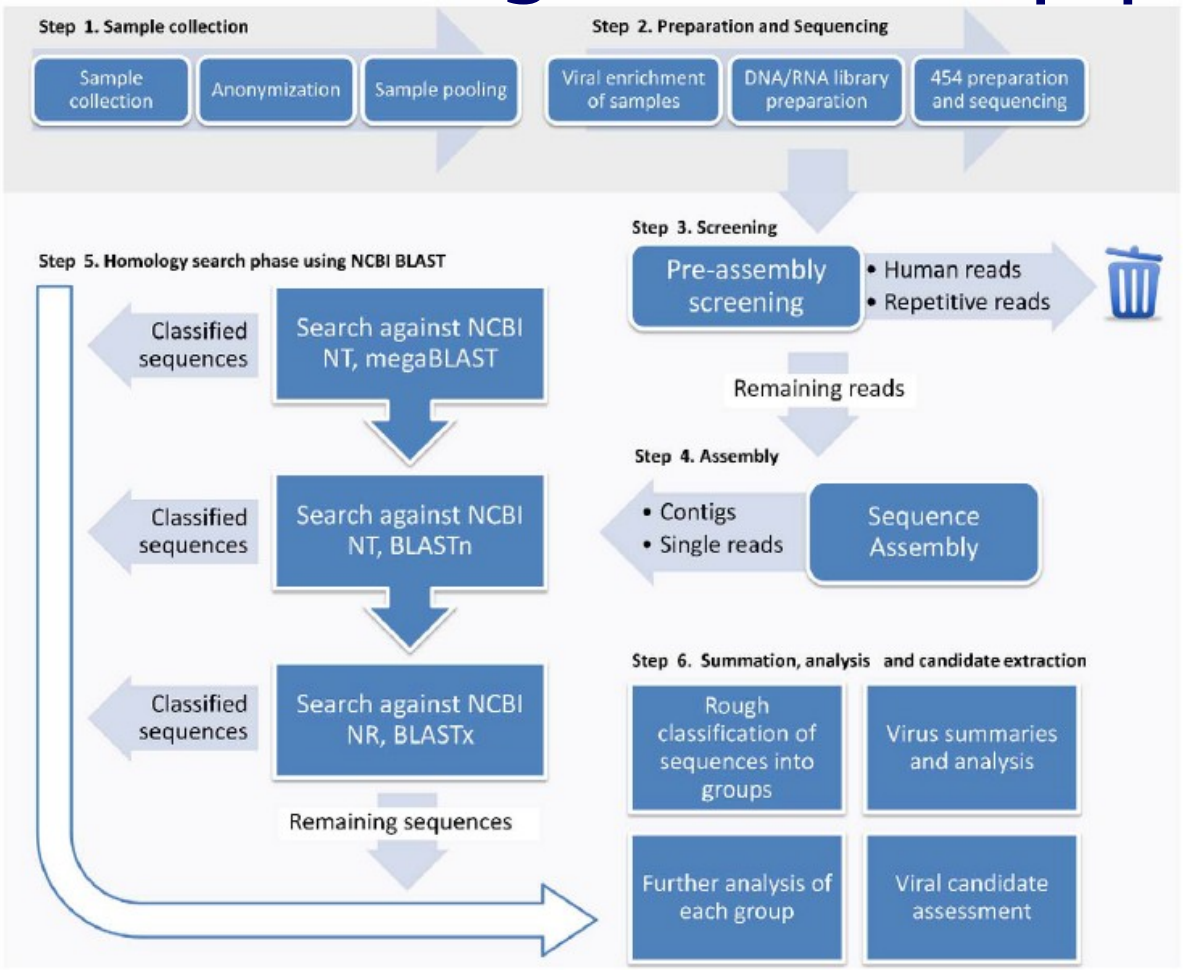
Navas-Molina et al., Methods in enzymology 2013, 531: 371-439



Taxonomy summary (i.e. phylum level)

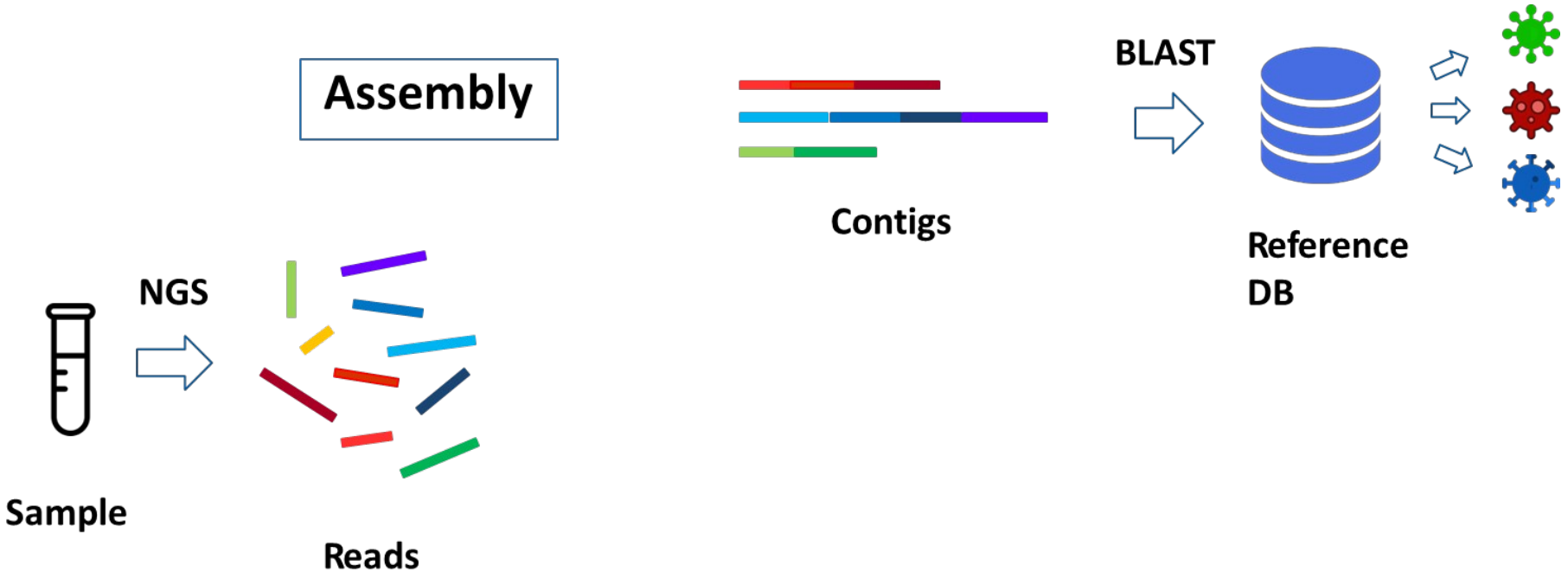


Metagenómica, pipeline de

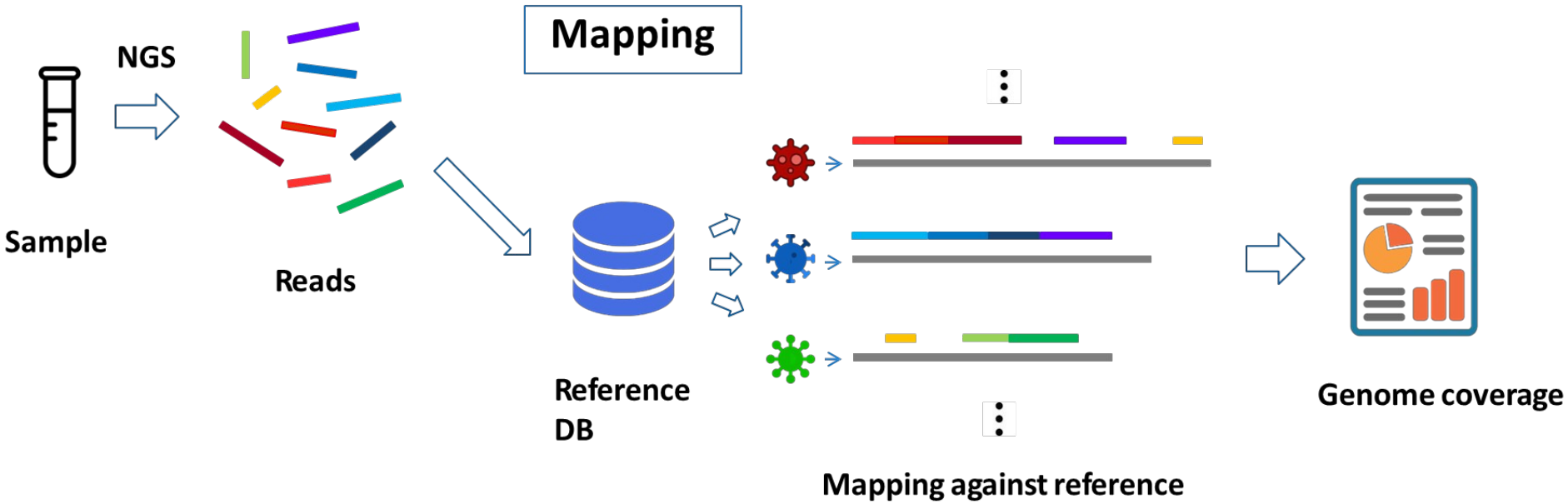


Lysholm et al., Plos One 2012:7,2, e30875

Metagenomic analysis approaches



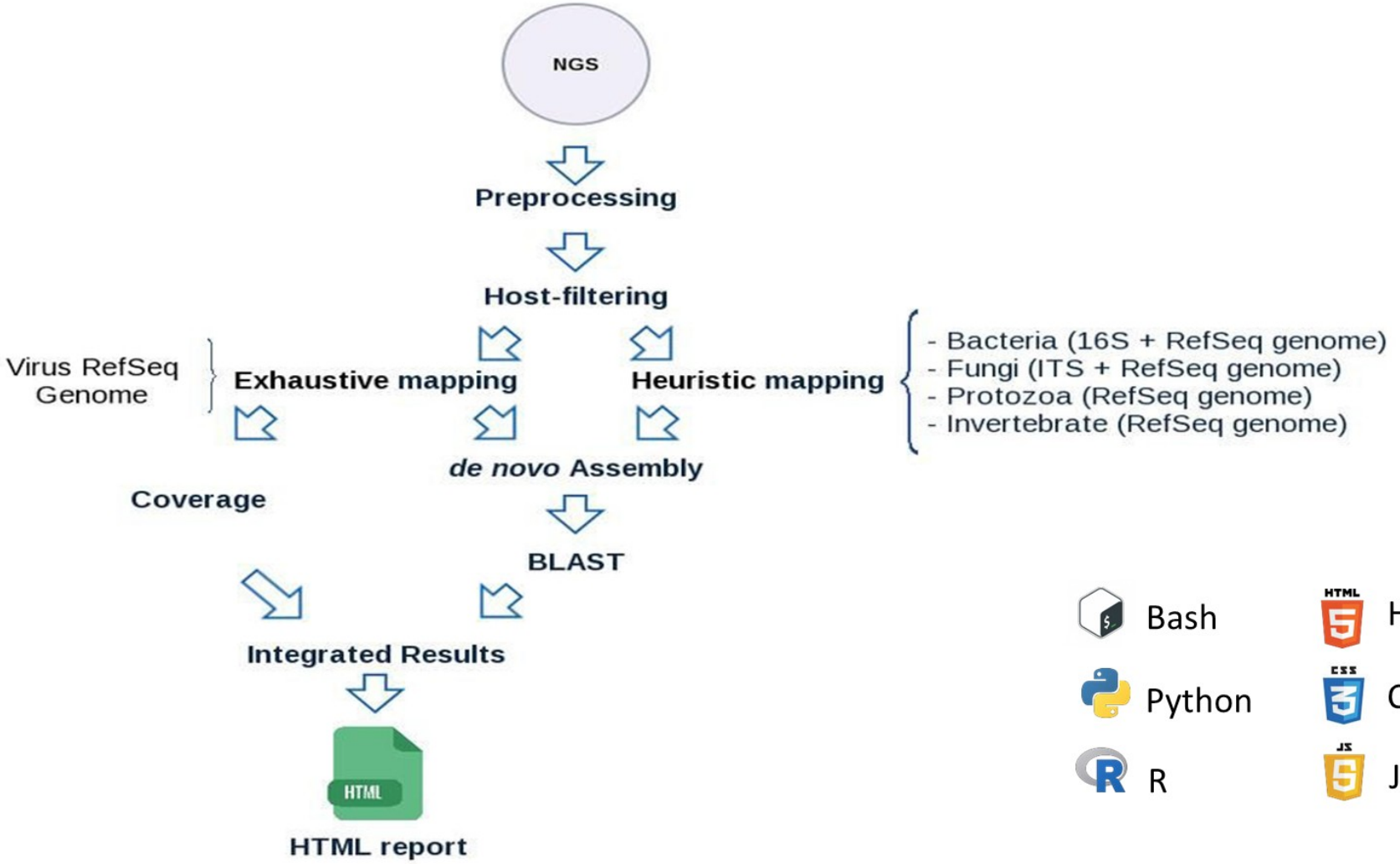
Metagenomic analysis approaches



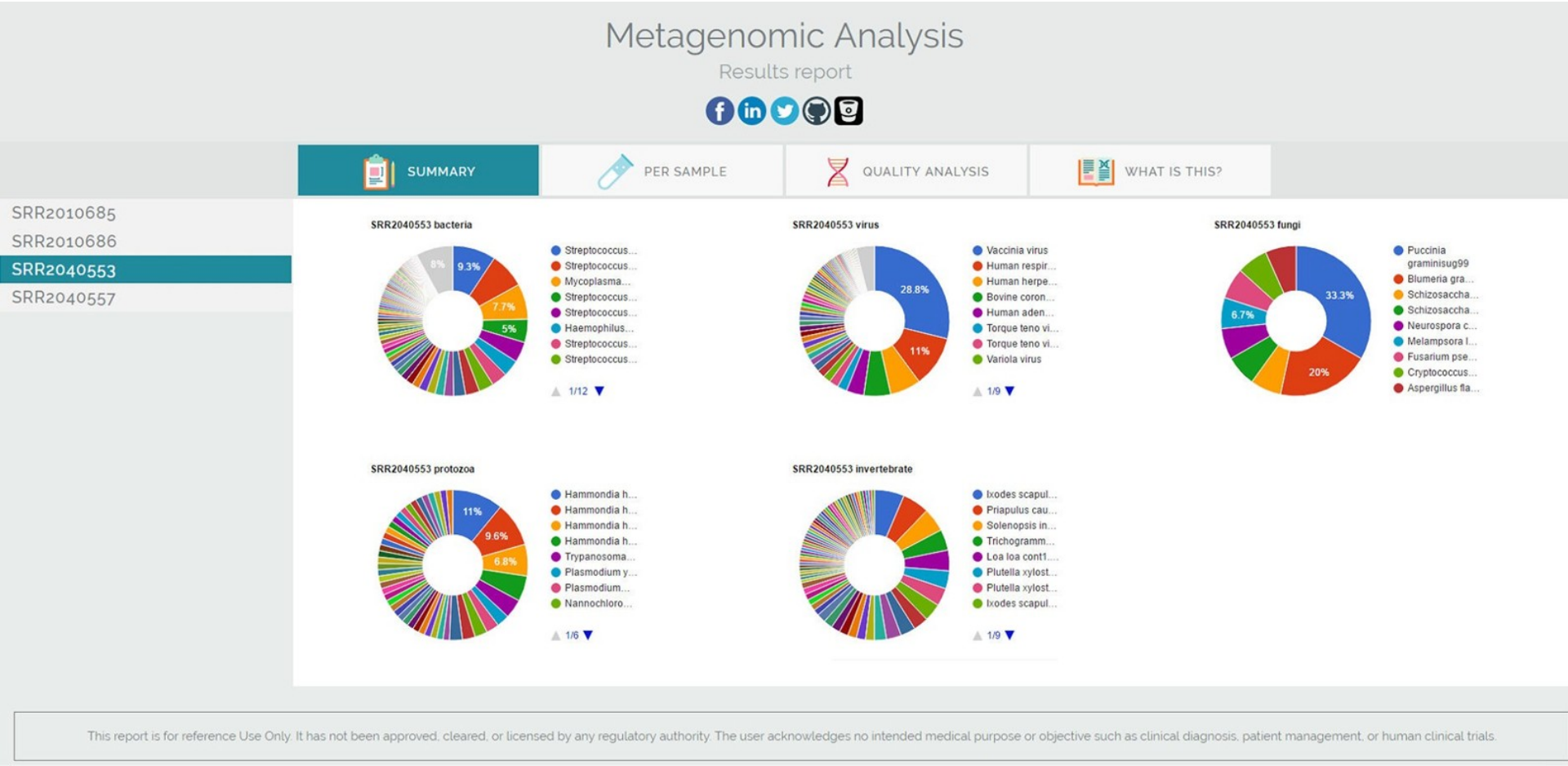
Metataxonomics vs Metagenomics (16S vs Shotgun)

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

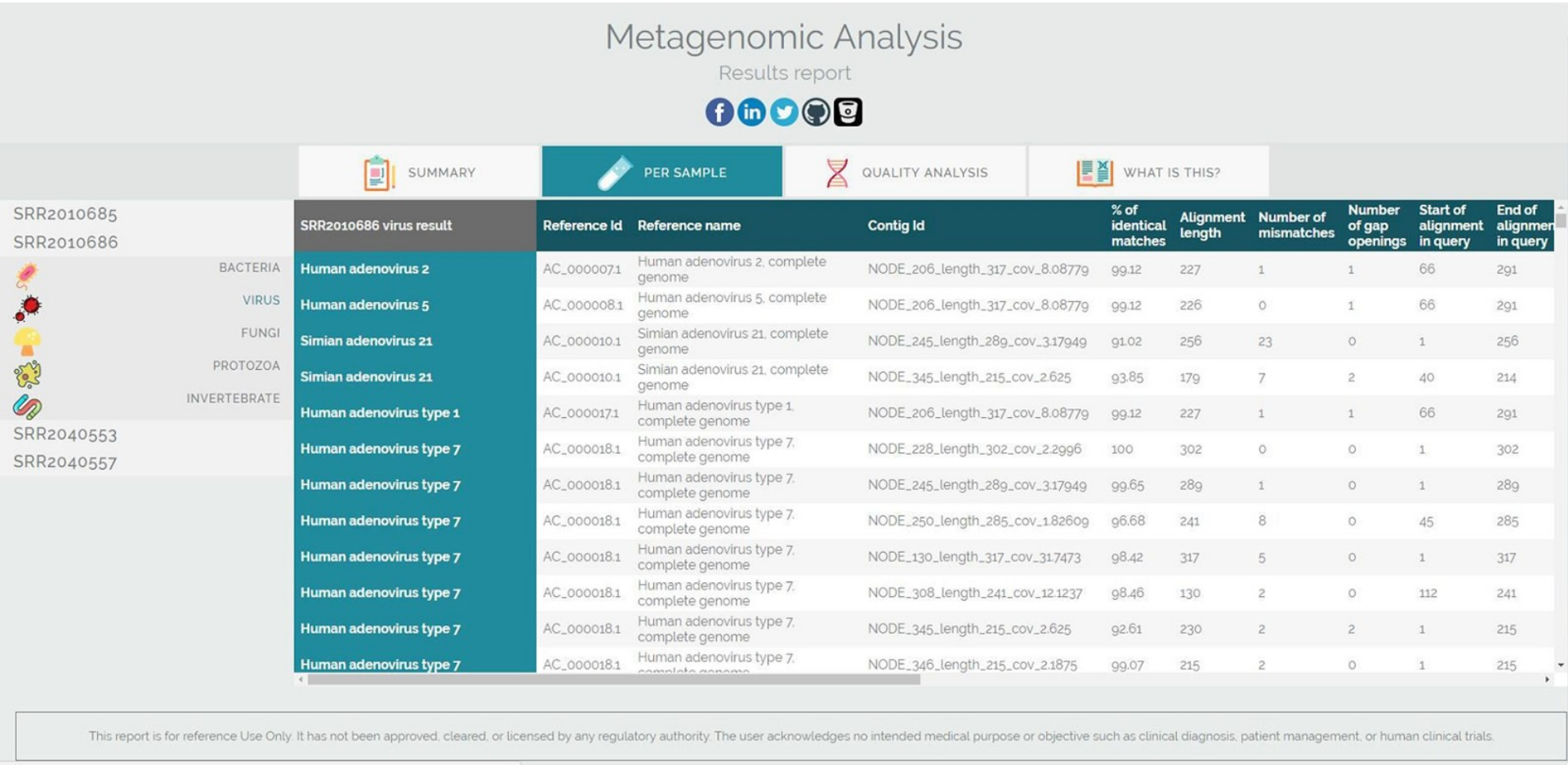
Metagenomic analysis: PikaVirus



Metagenomic analysis: PikaVirus

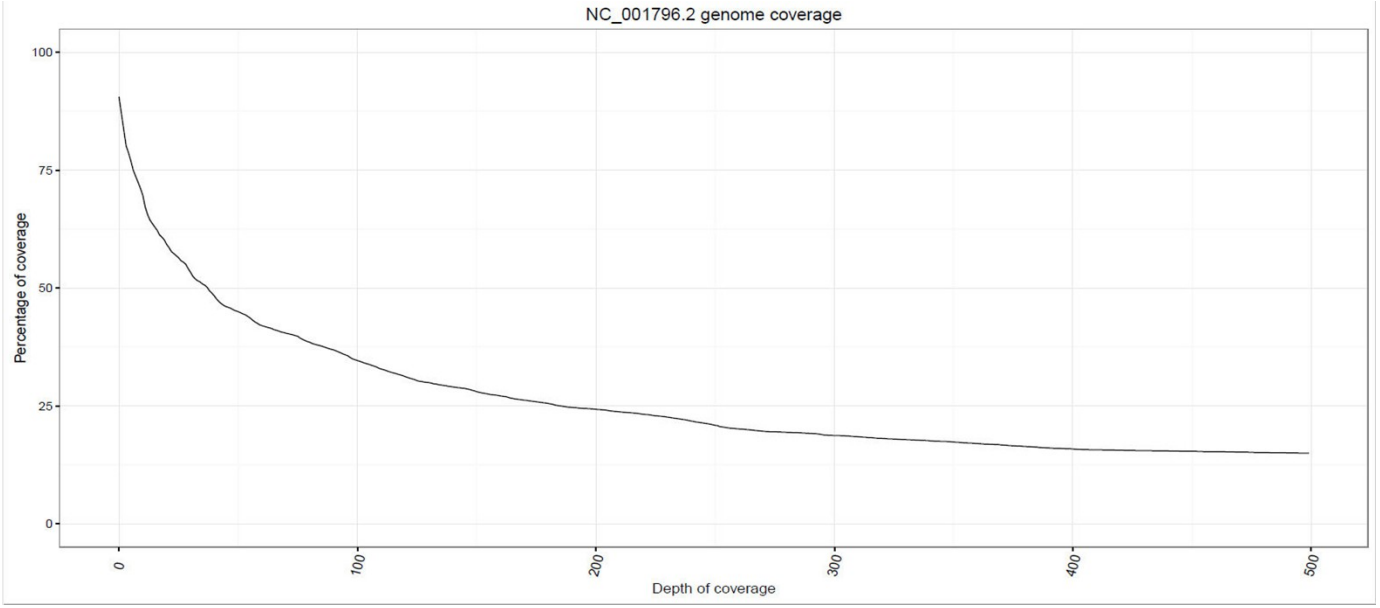


Metagenomic analysis: PikaVirus



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 alignmen 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,	NODE_346_length_215_cov_2.1875	99.07	215	2	0	1	215

Metagenomic analysis: PikaVirus

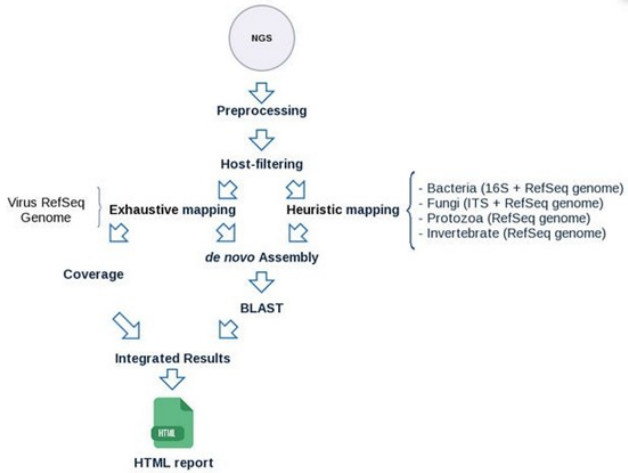


Wiki page and Project code

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.

Pages 31

- Home
- Dependencies
- References
  - Host
  - Bacteria
  - Virus
  - Fungi
  - Invertebrate
  - Protozoa
- Usage
  - i. Configuration file
  - ii. Quality control
  - iii. Mapping
    - Host filtering
    - Bacteria
    - Virus
    - Fungi
    - Parasite
  - iv. Assembly
  - v. BLAST
  - vi. Coverage
  - vii. Results
    - Summary
    - By Sample
    - Quality
    - Info
- Directory Structure

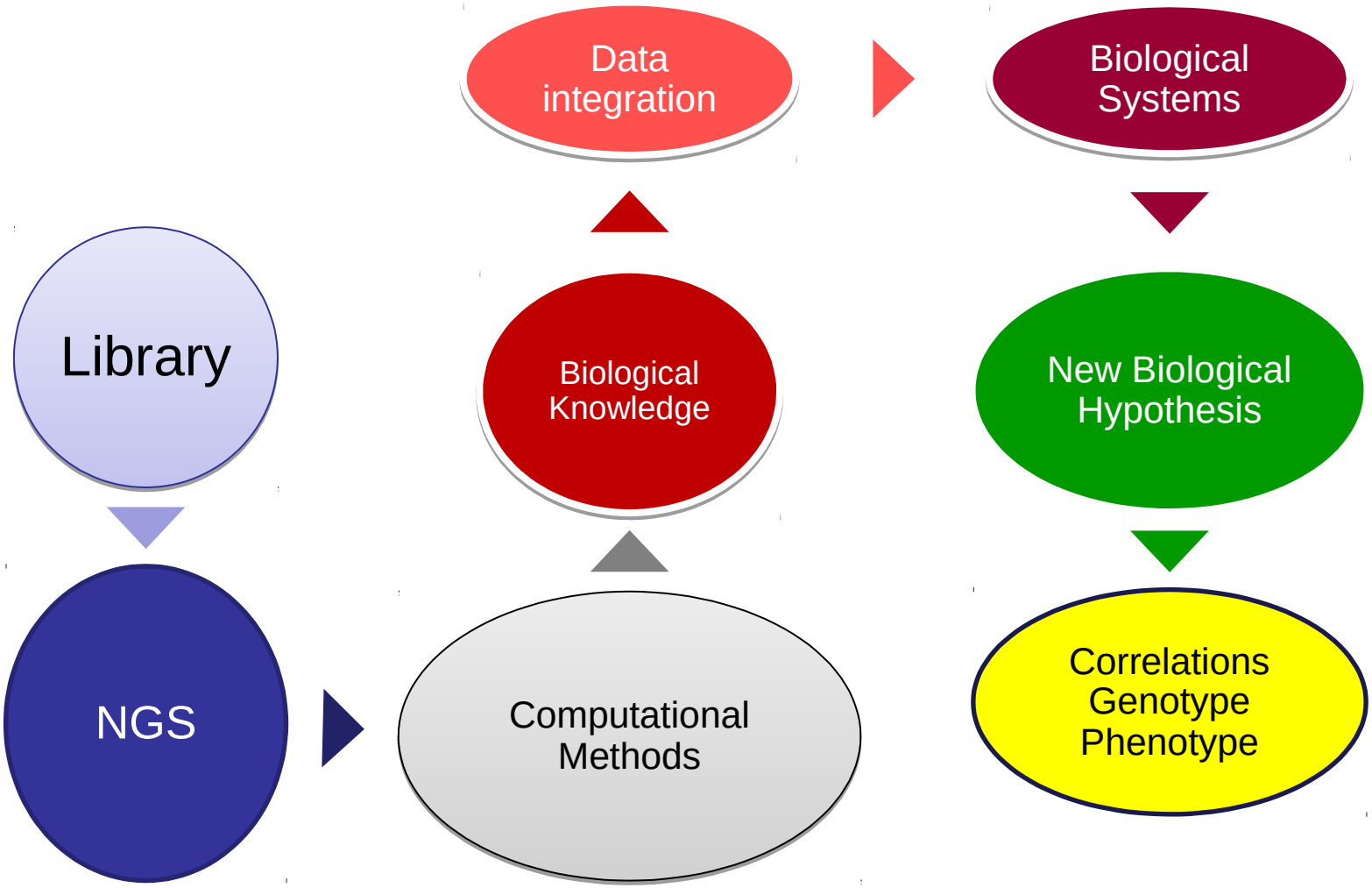
Clone this wiki locally

https://github.com/AndreaRP/

Clone in Desktop

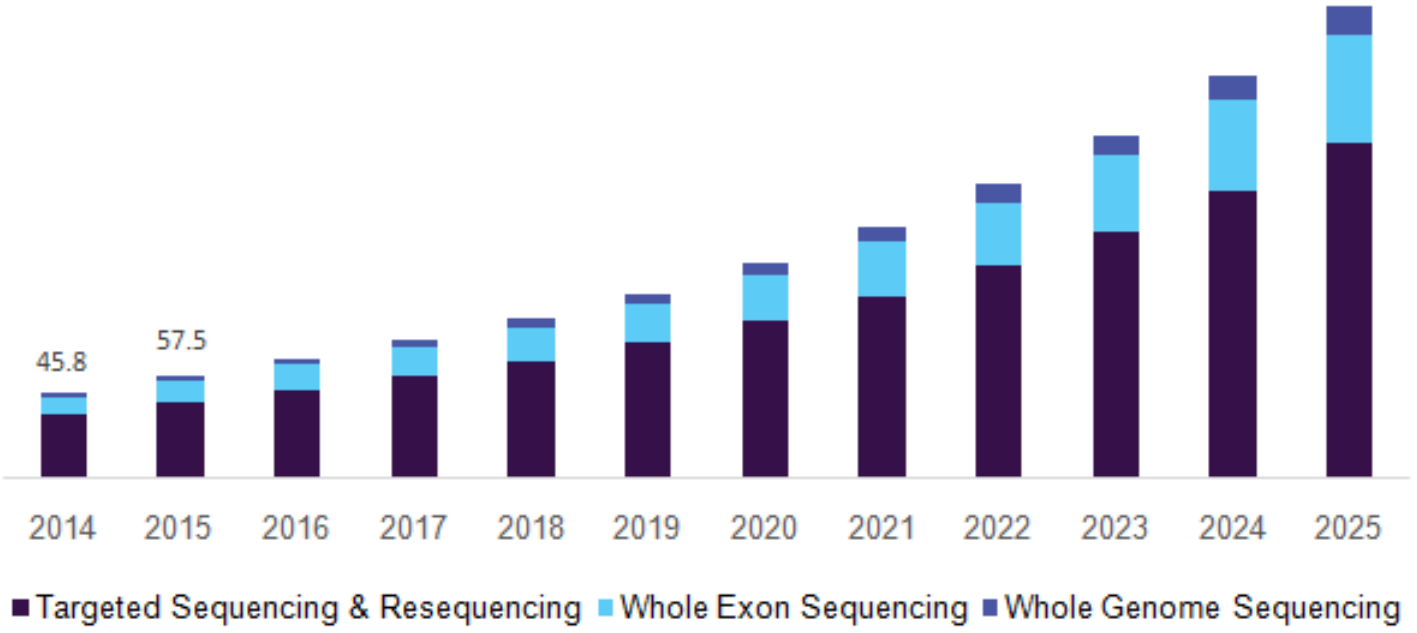
https://github.com/BU-ISCIII

IMPACT OF MASSIVE SEQUENCING

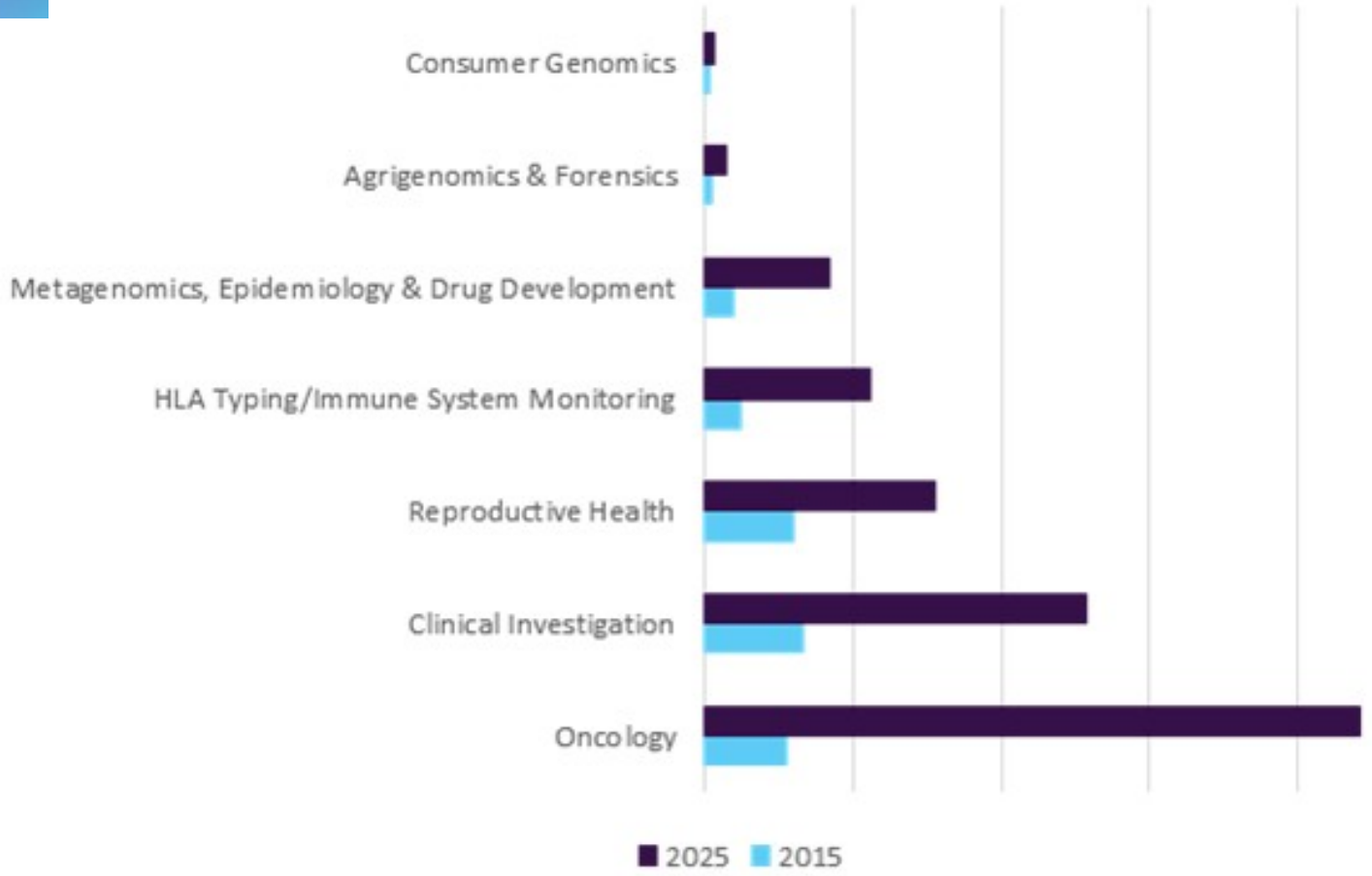


Next Generation Sequencing (NGS) Market Size & Forecast By Application (Oncology, Reproductive Health), By Technology (Targeted, WGS, WES), By Workflow (Data Analysis), By end-use (Academic & Clinical Research), And Trend Analysis, 2014 - 2025

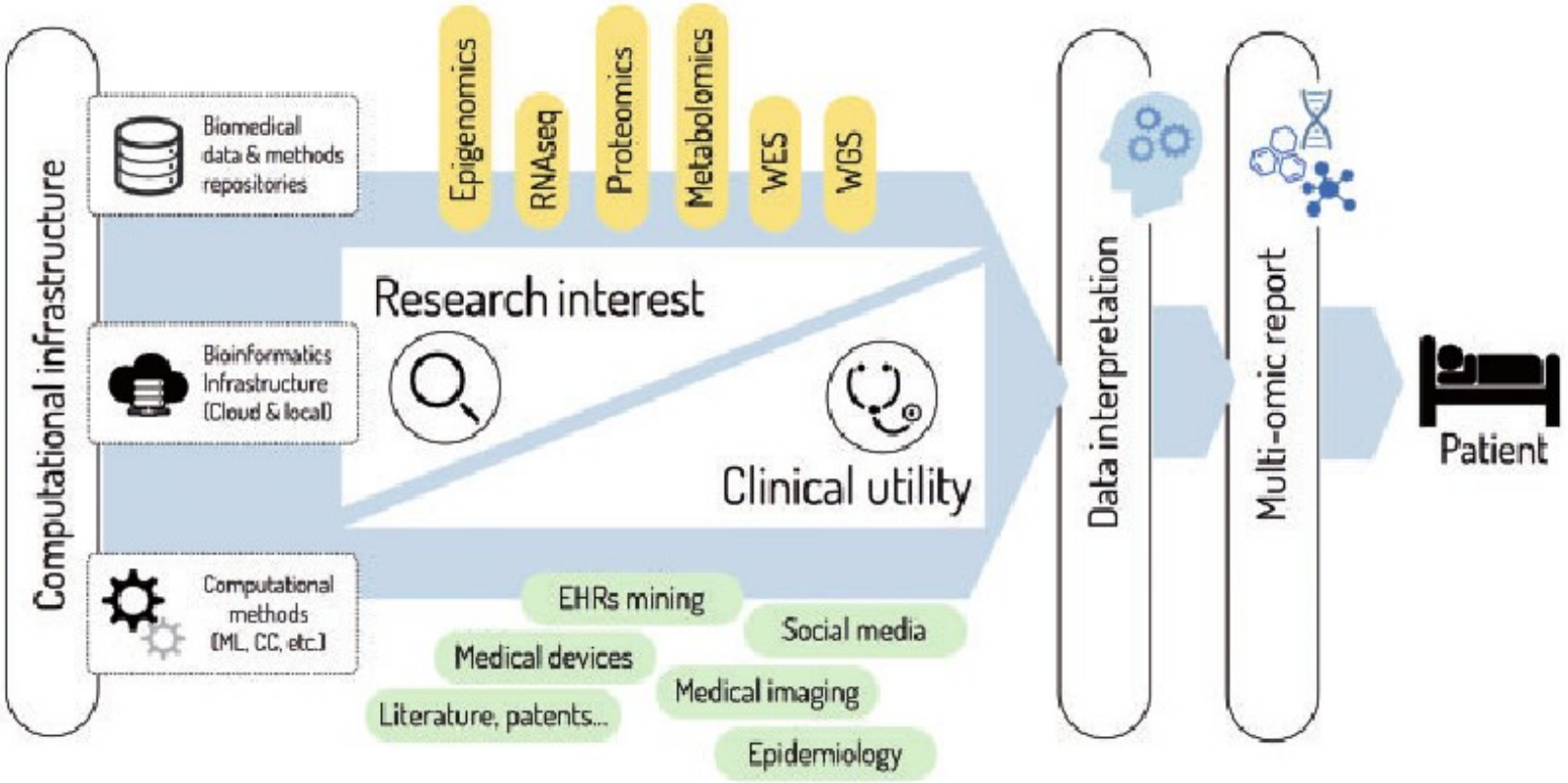
UK NGS market revenue, by technology, 2013 - 2024 (USD Million)



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

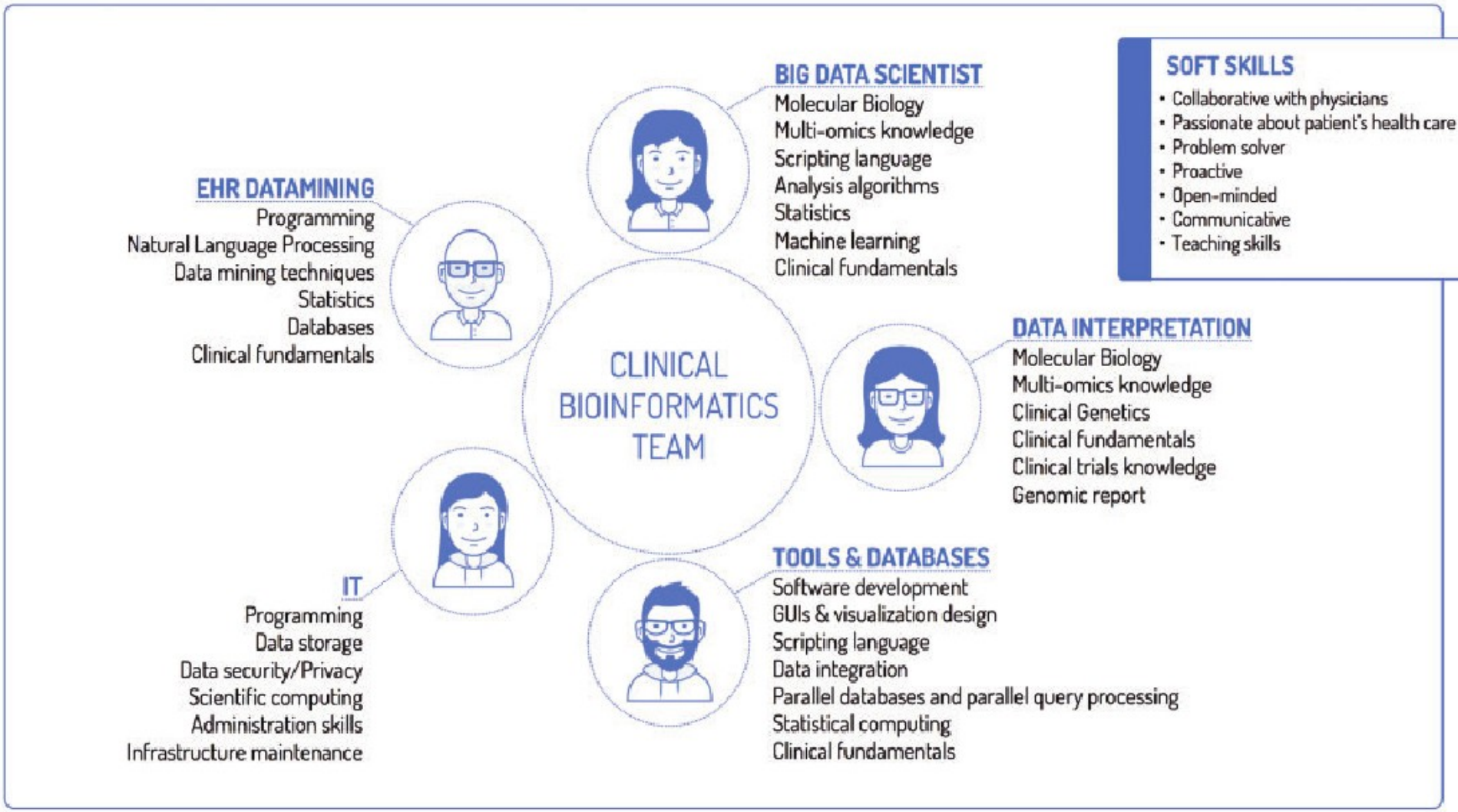


Precision Medicine workflow: from data to patient care



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

Clinical bioinformatics laboratory profile



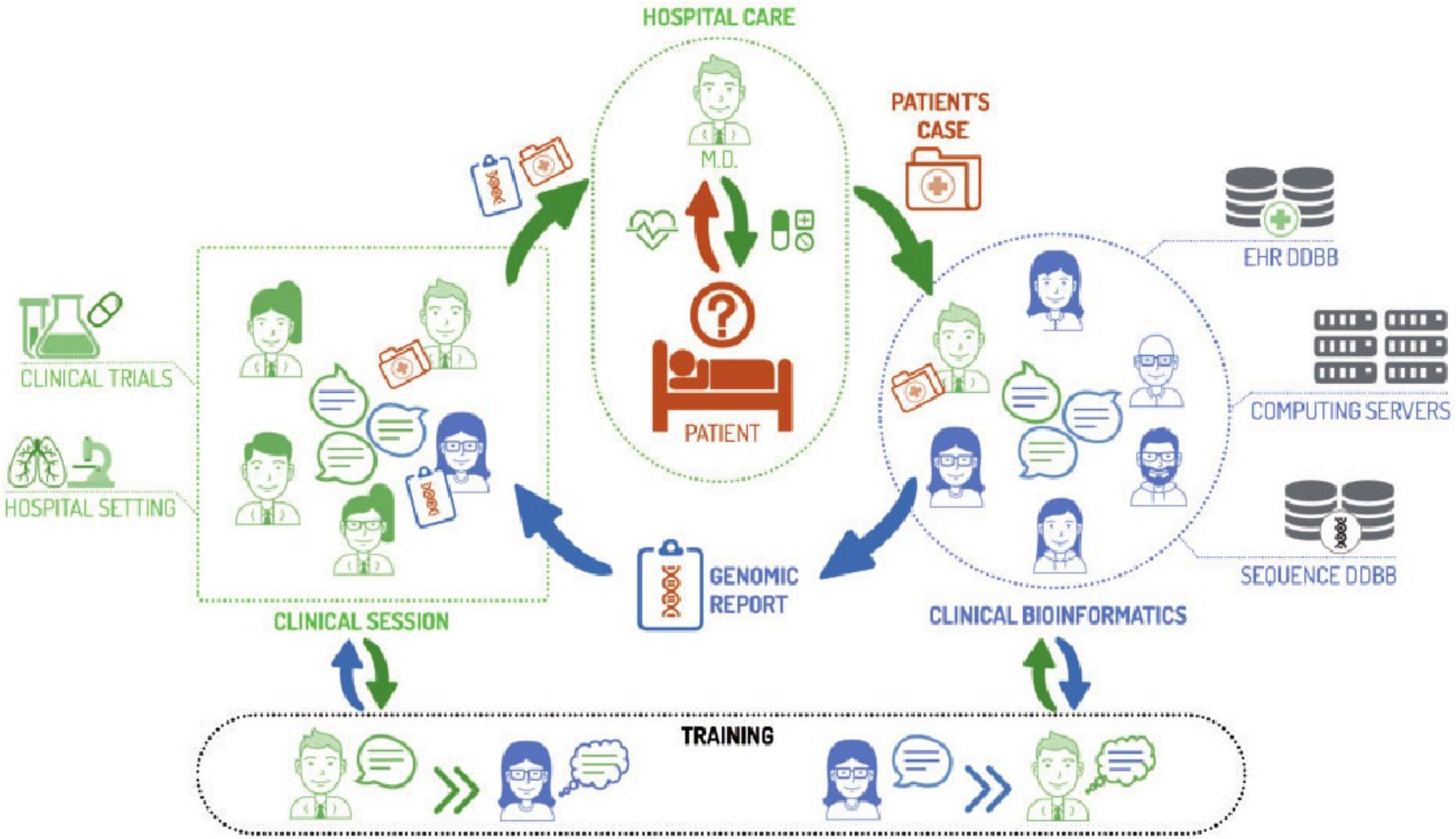
Skills for Clinical Bioinformaticians

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



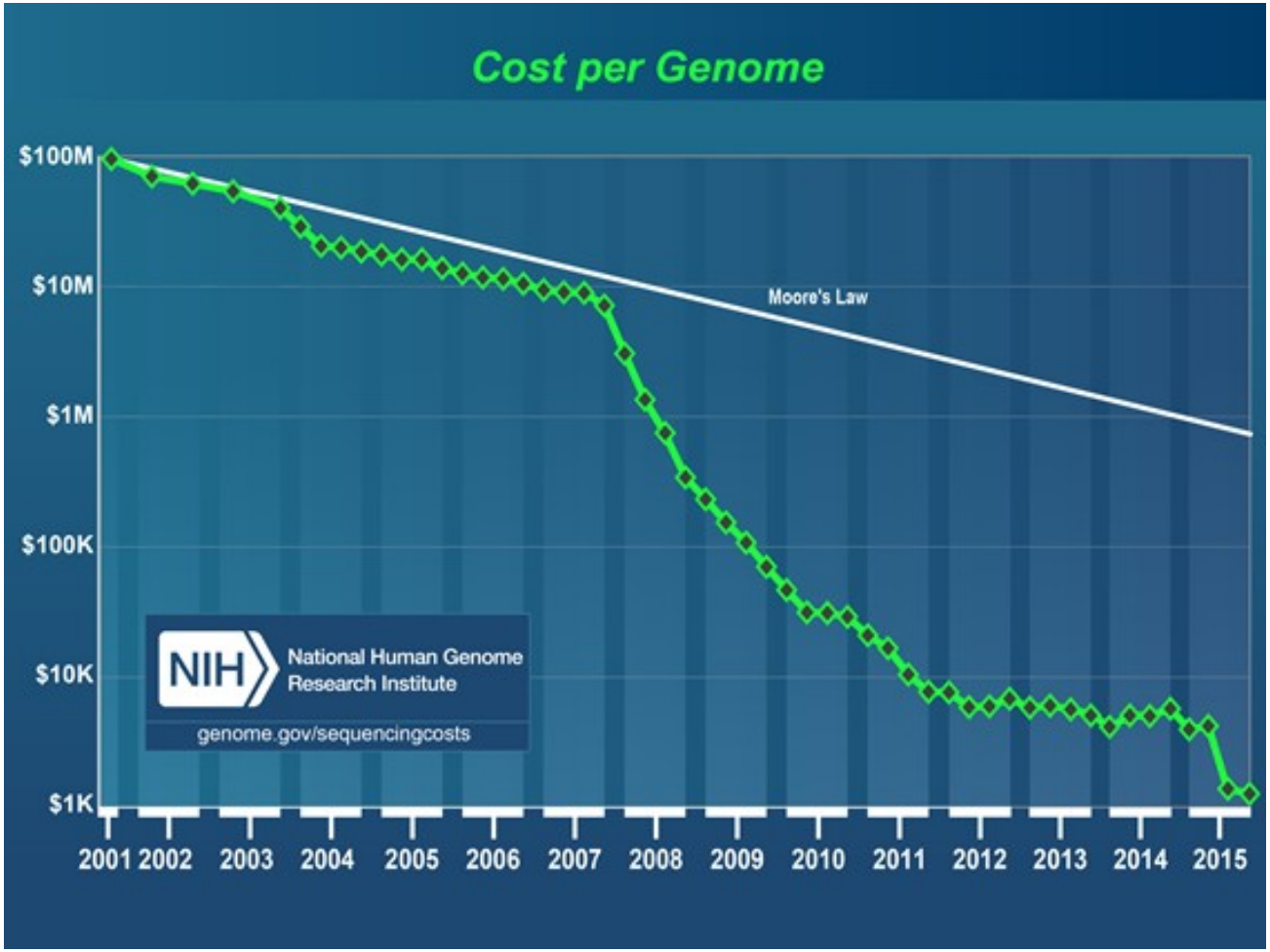
Retos de la Bioinformática en NG

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

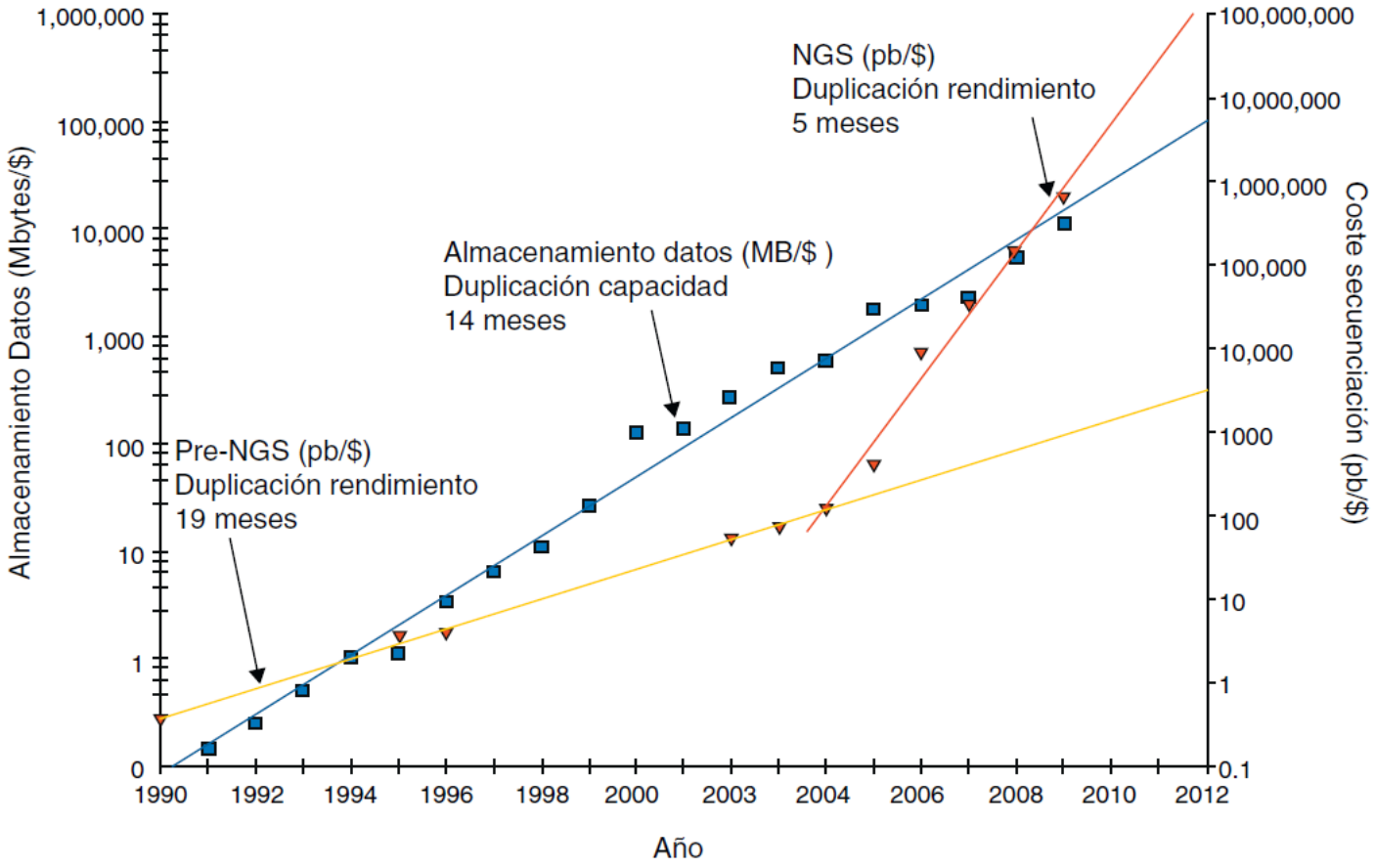
- Coste de la secuenciación disminuye
  el embudo es el análisis de datos

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

Coste actual de la secuenciación



Costes del almacenamiento vs se



Adaptada de Stein, Genome Biology 2010, 11:207

Retos de la Bioinformática en NG

- 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 NG

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

Softwares comerciales en Bioinfo

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 M

- 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