Session 3.1 – Análisis de datos genómicos en bacteriología, opciones y metodologías

Isabel Cuesta

BU-ISCIII

Unidades Centrales Científico Técnicas – SGSAFI-ISCIII

3 al 10 noviembre 2022
CURSO FORMACIÓN AESAN-CNA

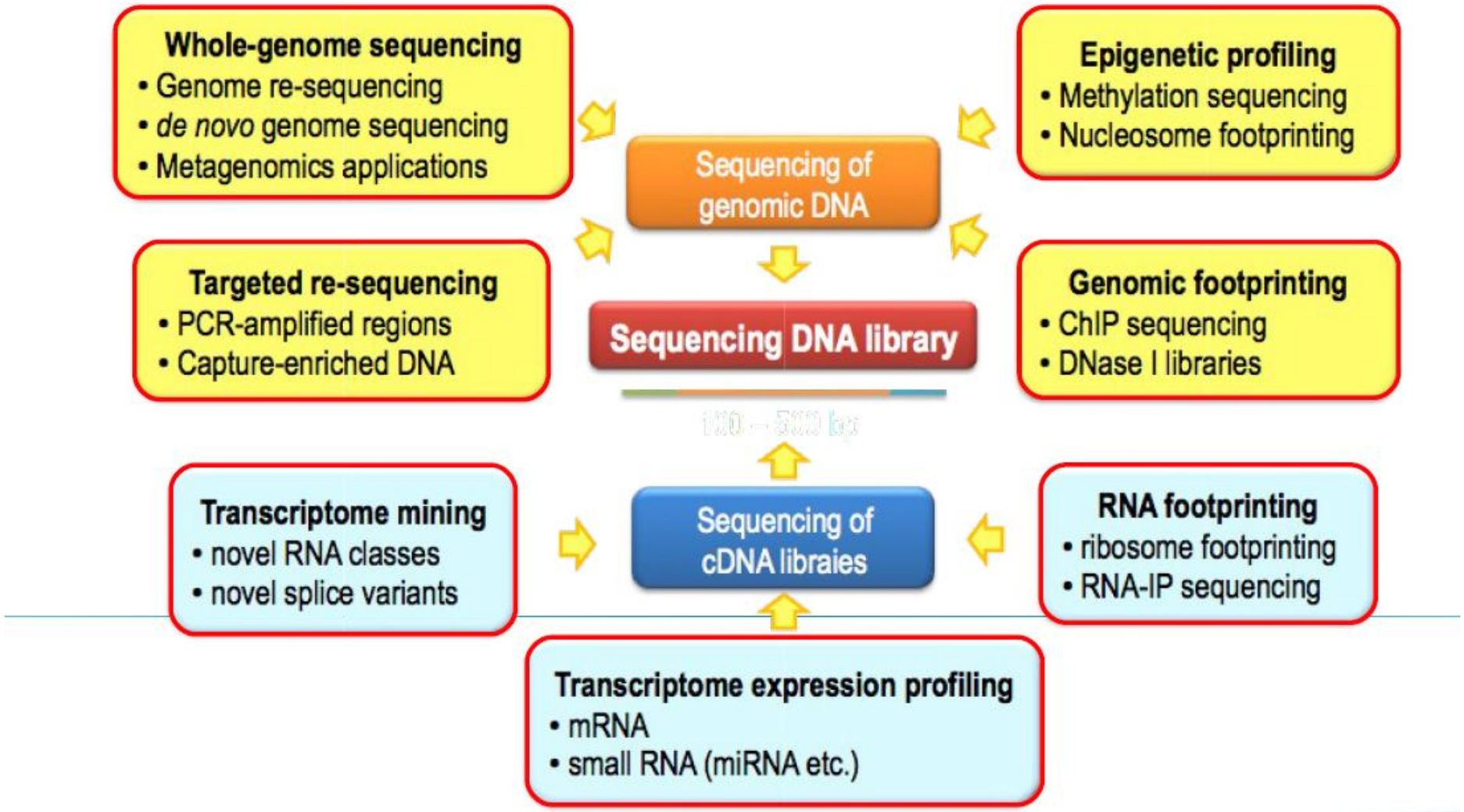
Index

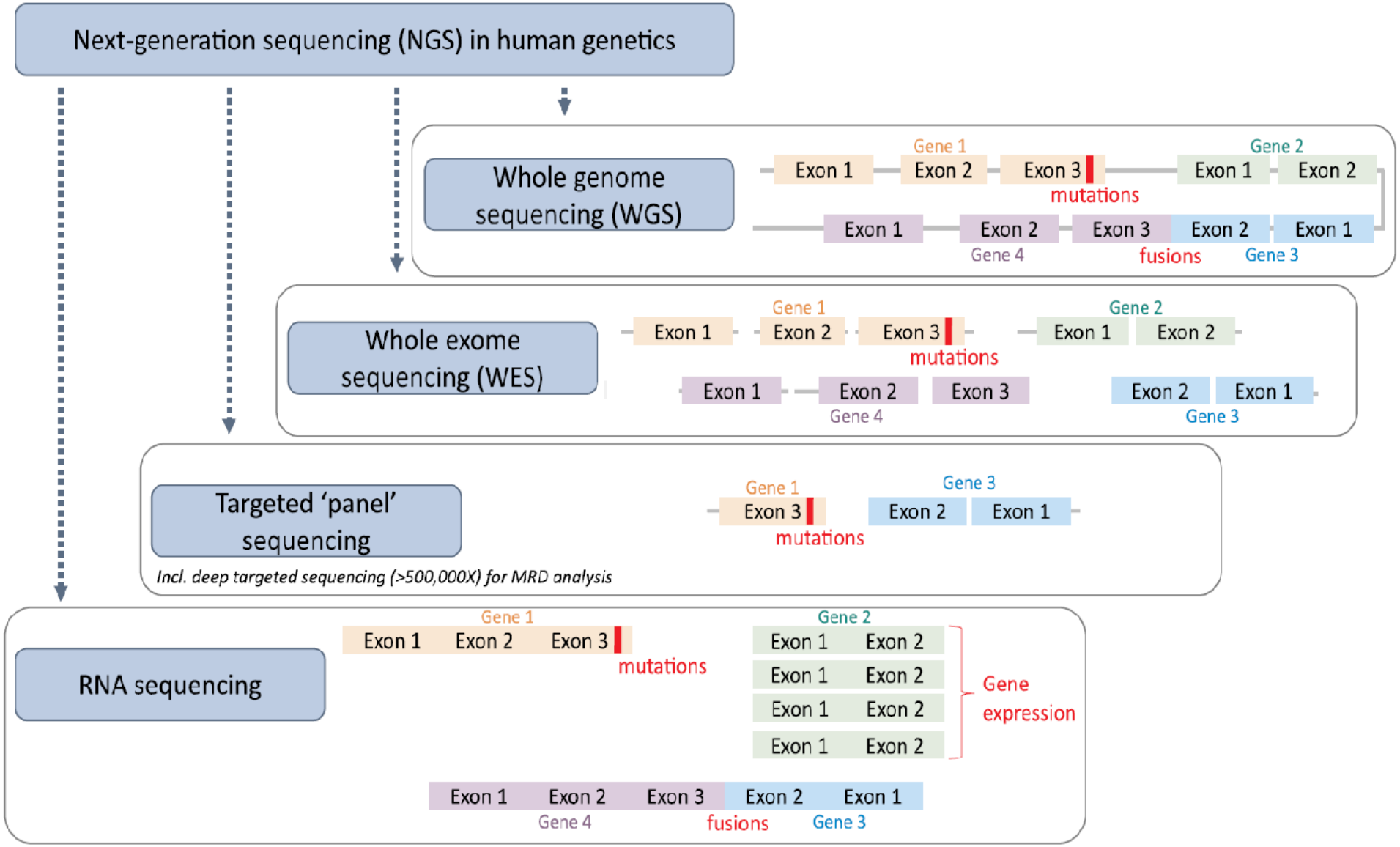
- NGS applications
- Microbial genomics
- Library strategies
- Bioinformatics analysis
- Challenges in Bioinformatics

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

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

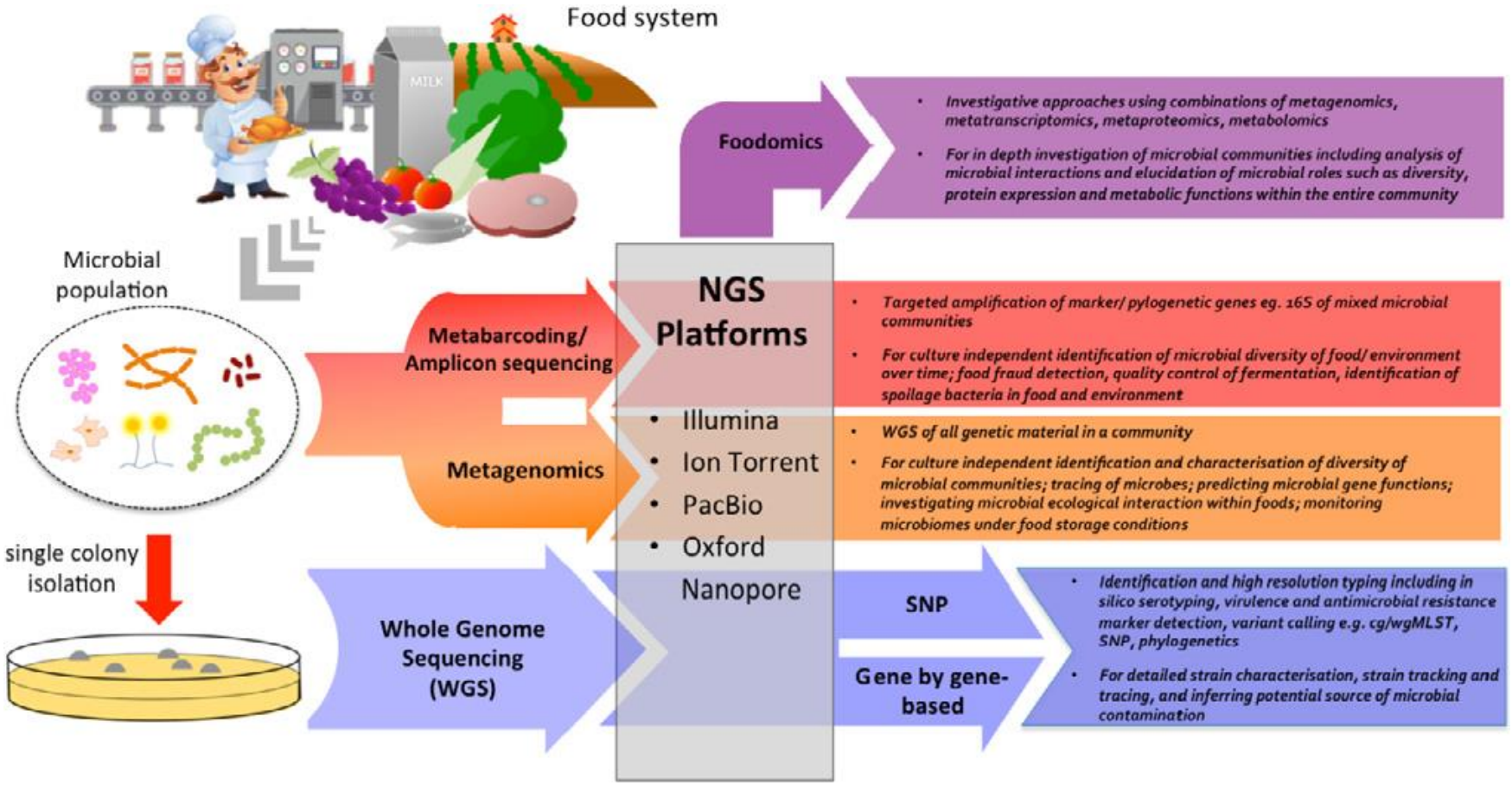
- Maxima complejidad en secuenciación y coste
- Información de regiones no codificantes
- Análisis de variaciones estructurales
- Elevado volumen de datos

Index

- NGS applications
- Microbial genomics
- Library strategies
- Bioinformatics analysis
- Challenges in Bioinformatics

Summary of potential NGS use by the food industry

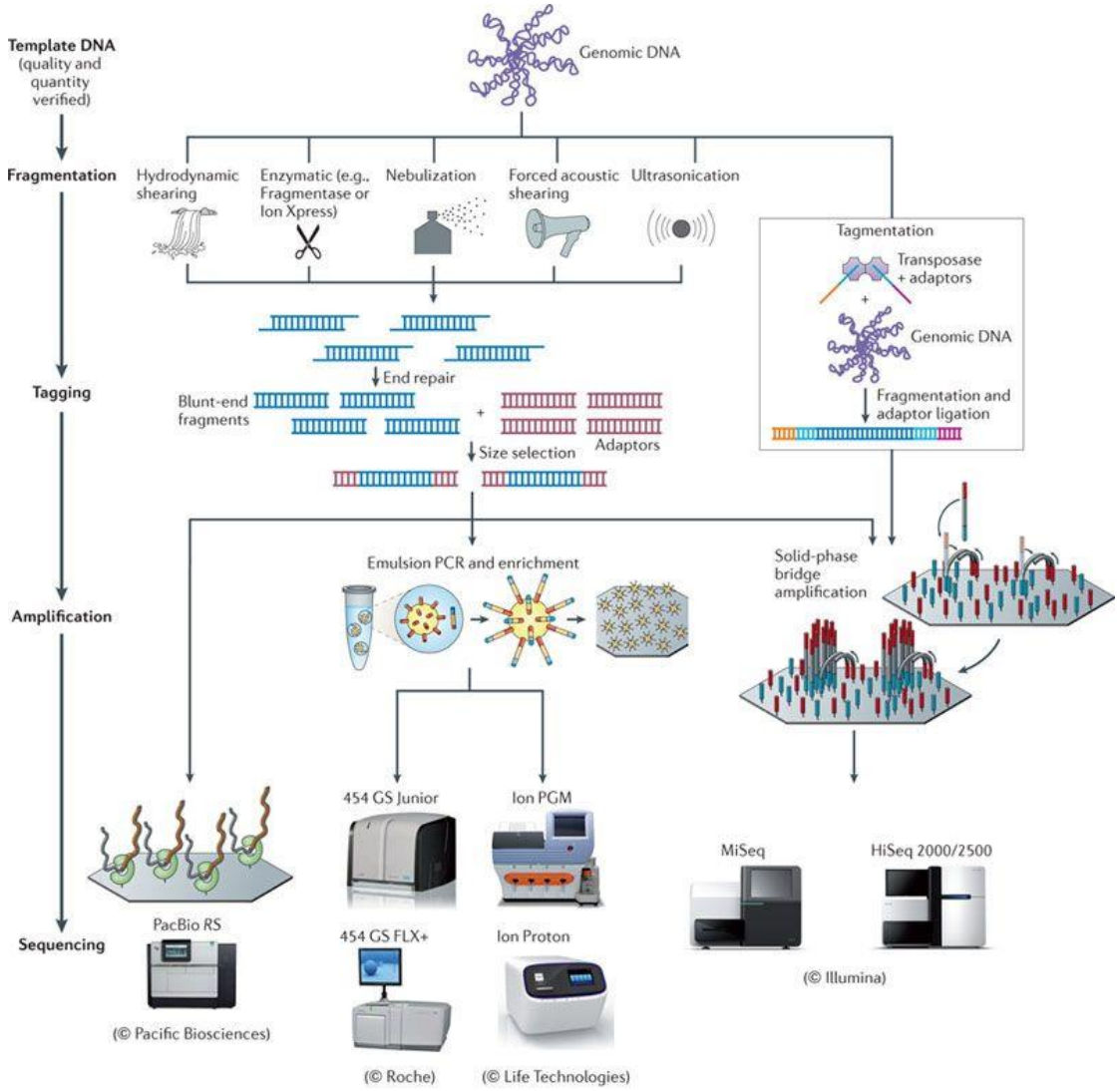
Jagadeesan et al., Food Microbiology 79 (2019) 96-115



Index

- NGS applications
- Microbial genomics
- Library strategies
- Bioinformatics analysis
- Challenges in Bioinformatics

High-throughput sequencing platforms



Nature Reviews | Microbiology Loman et al, 2012

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

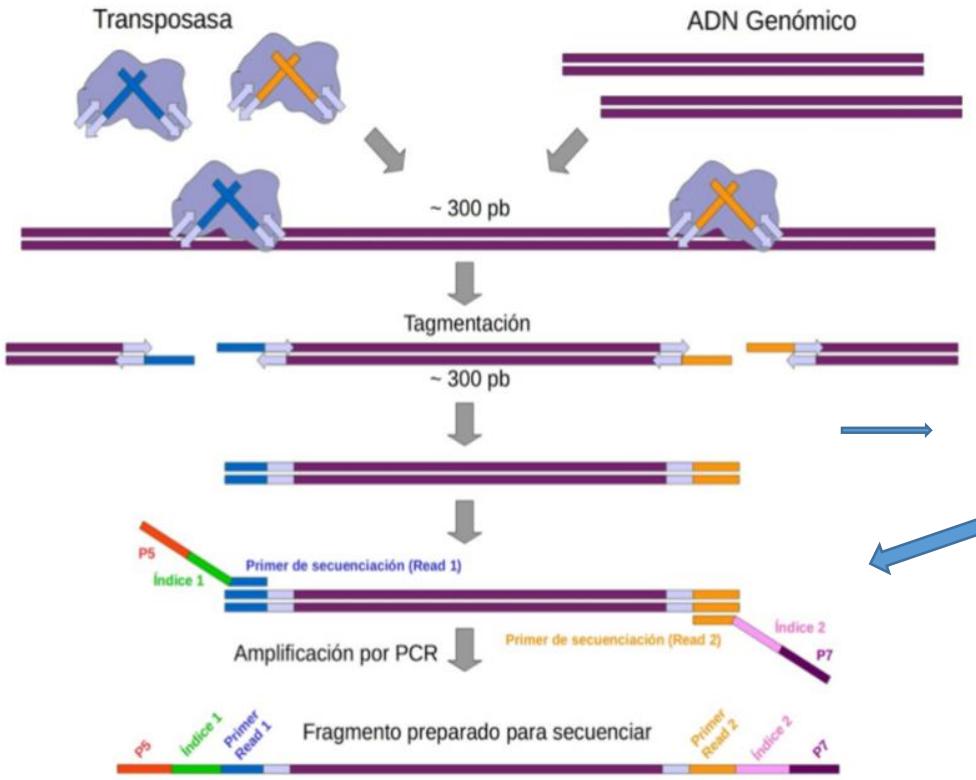
1. Metagenómica

IDENTIFICACIÓN MICROORGANISMOS

1. Metataxonomía

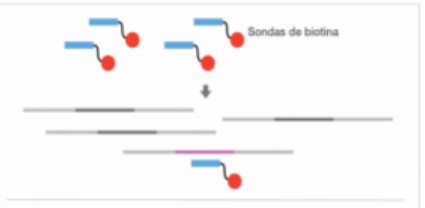
PREPARACIÓN LIBRERÍA

ENZIMÁTICA
FÍSICA



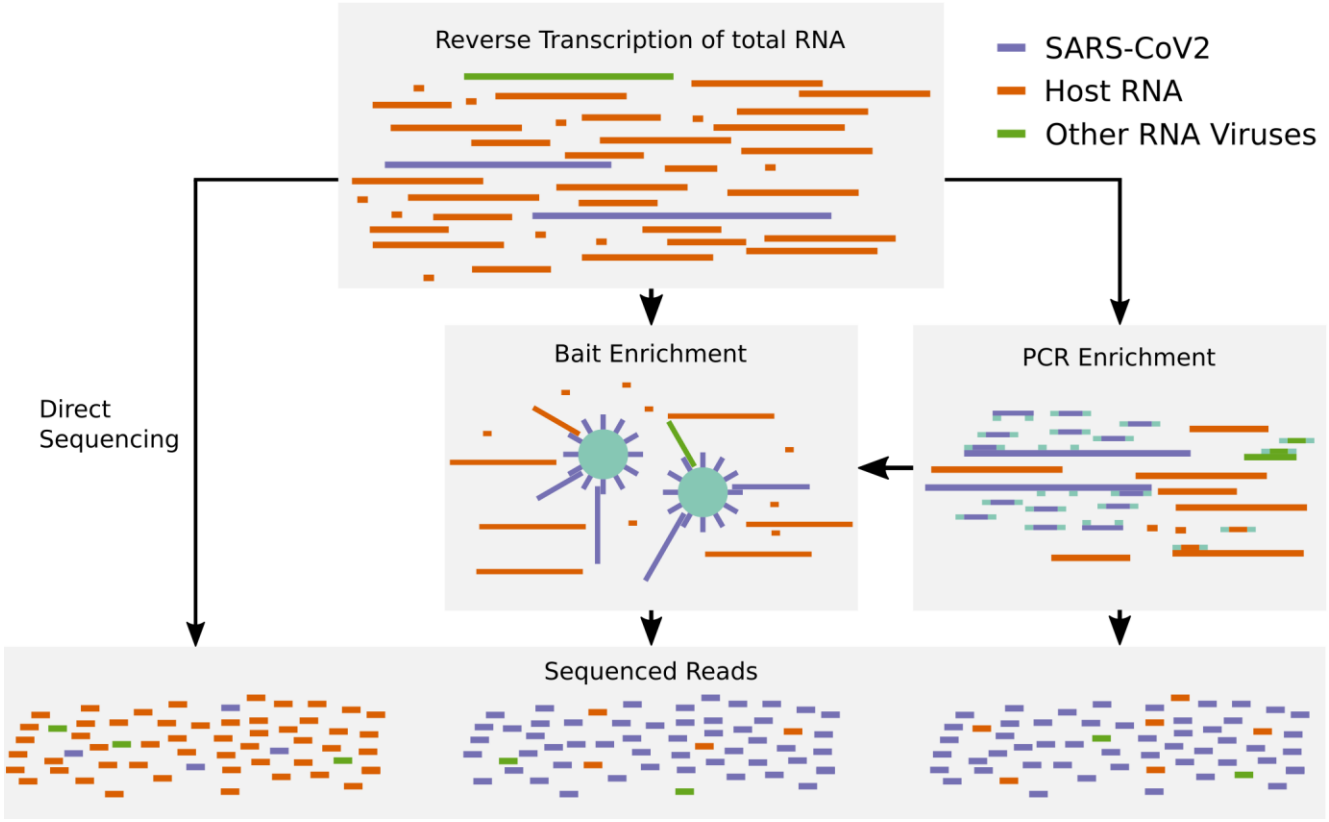
RNA -> cDNA

ENRIQUECIMIENTO:
PCR
CAPTURA SONDAS

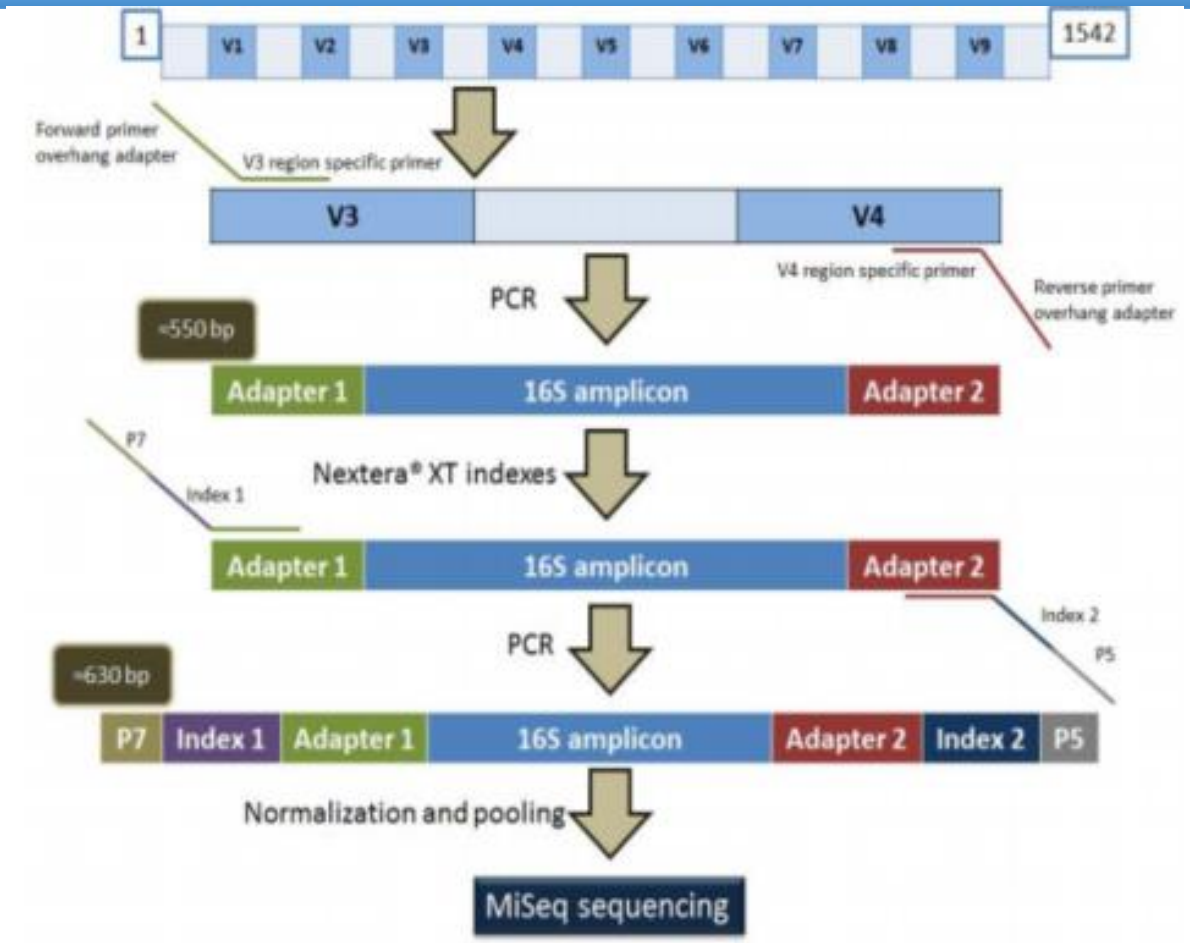


Guia Práctica Genómica https://www.uv.es/varnau/GM_Cap%C3%ADtulo_2.pdf

PREPARACIÓN LIBRERÍA



PREPARACIÓN LIBRERÍA, rRNA 16S, caracterización microbiota



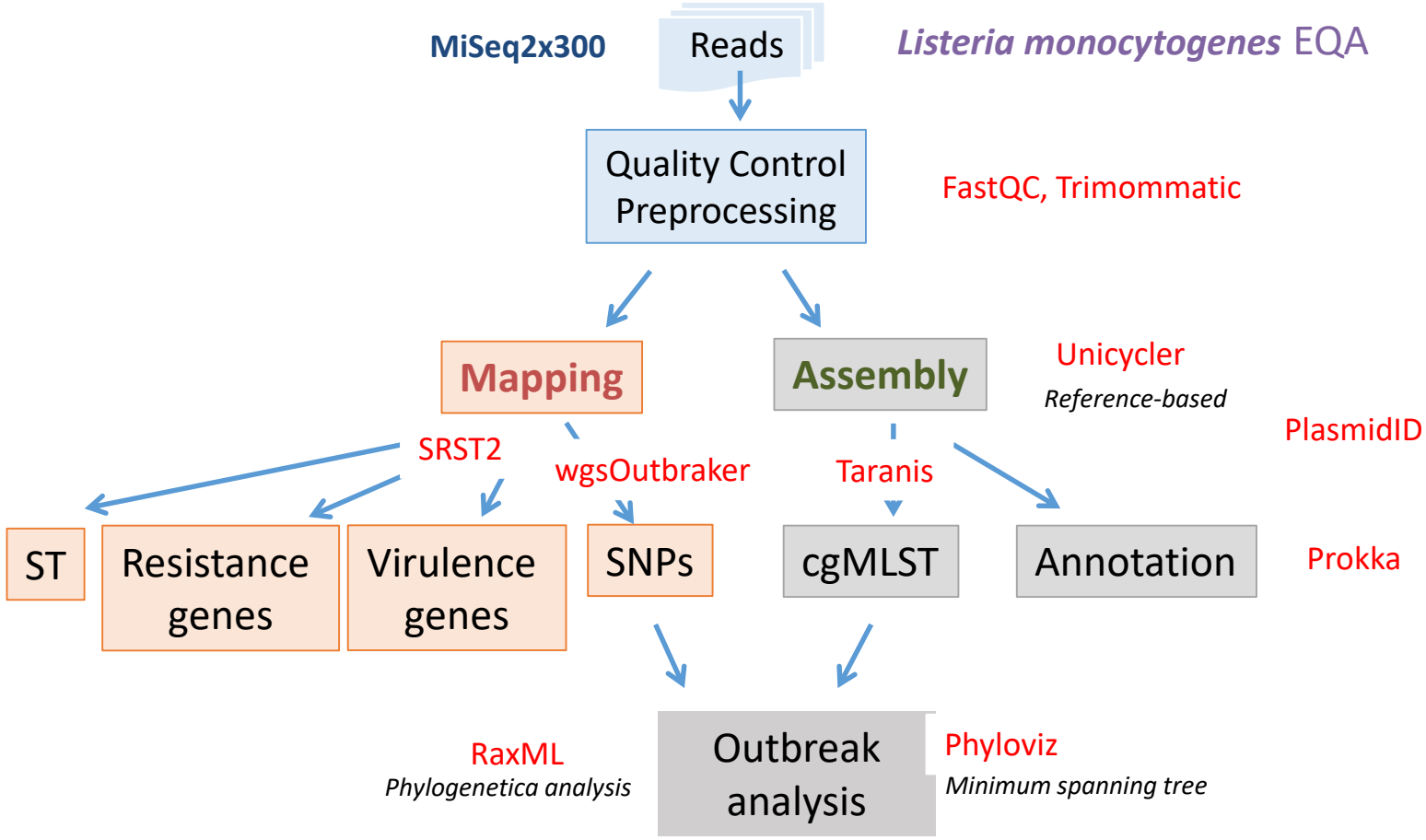
Index

- NGS applications
- Microbial genomics
- Library strategies
- **Bioinformatics analysis**
- Challenges in Bioinformatics

Bioinformatics analysis in microbial genomics

- SPECIE IDENTIFICATION
  - WGS - Kmers analysis
  - TARGET 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

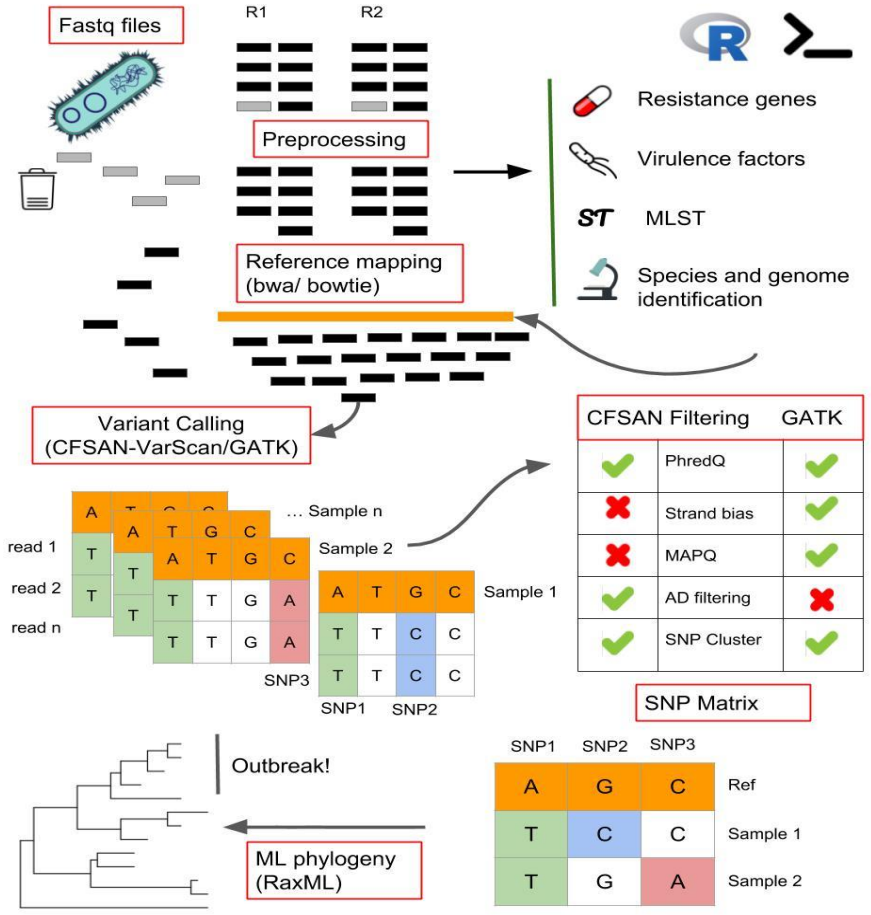
- Live-SET

High-quality SNPs para crear filogenia para investigación de brotes

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

- WGS-Outbraker

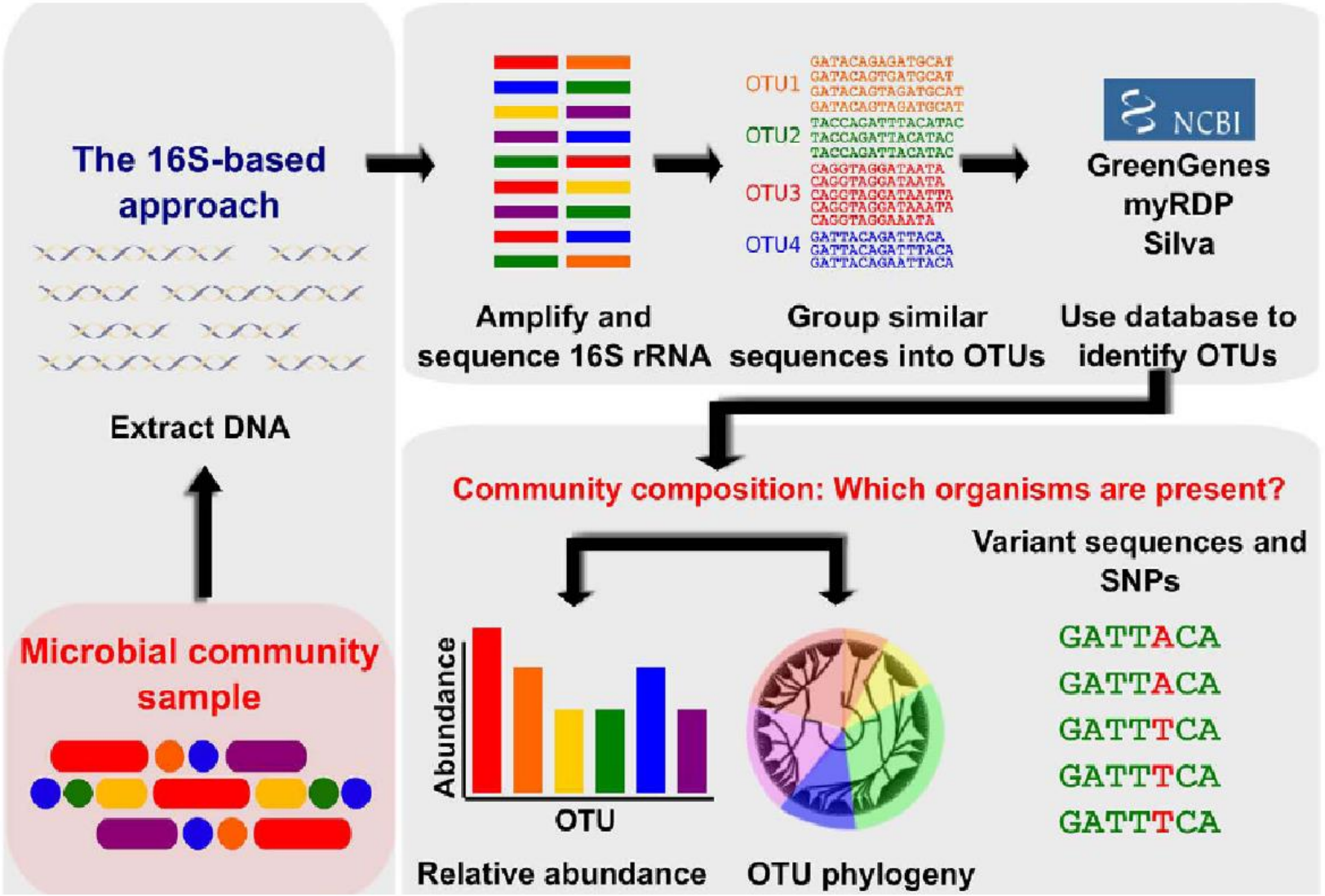
WGS-Outbreaker https://github.com/BU-ISCIII/WGS-Outbreaker



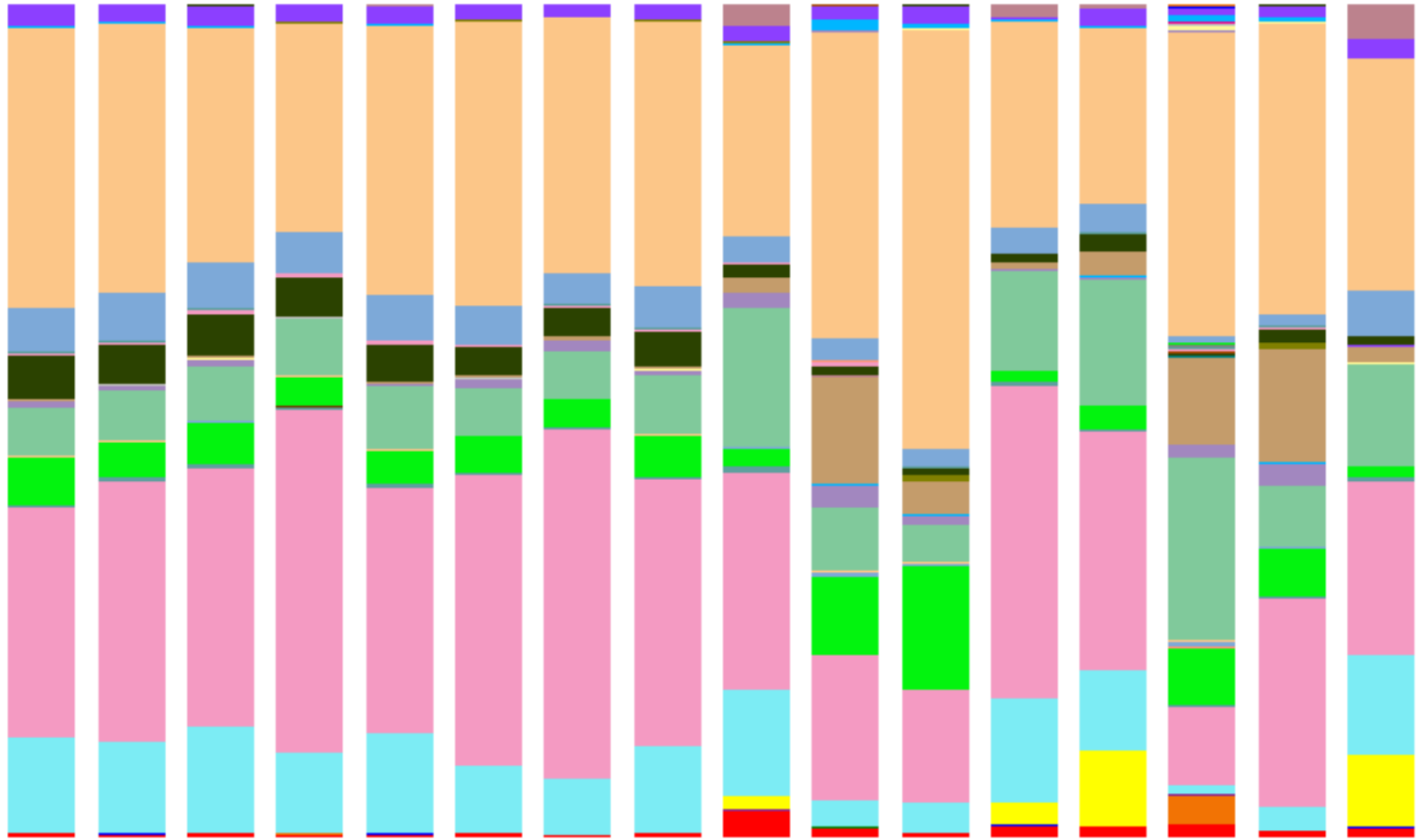
Metataxonomics vs Metagenomics (16S vs Shotgun)

	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



Taxonomy summary (i.e. phylum level)

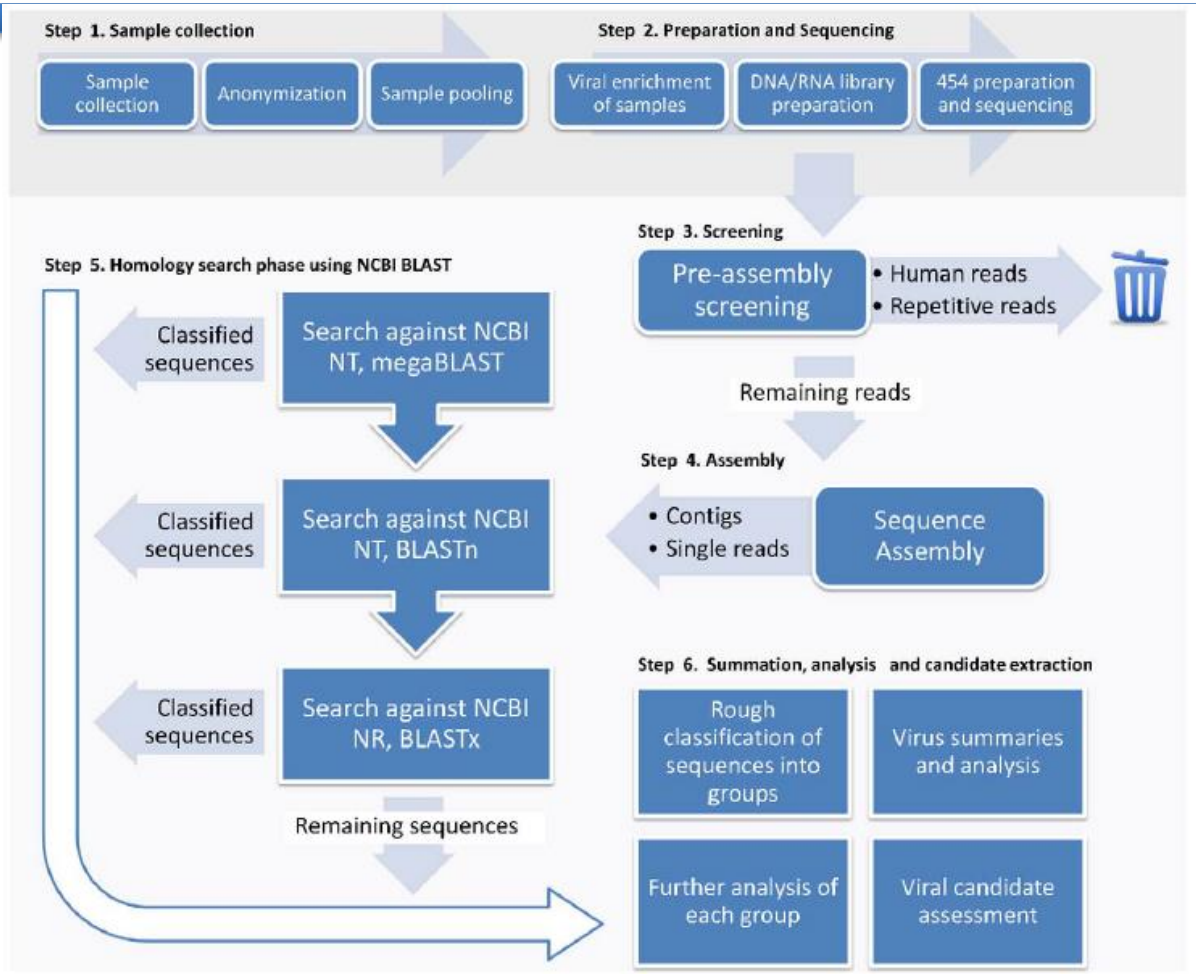


Metataxonomics

Problemas:

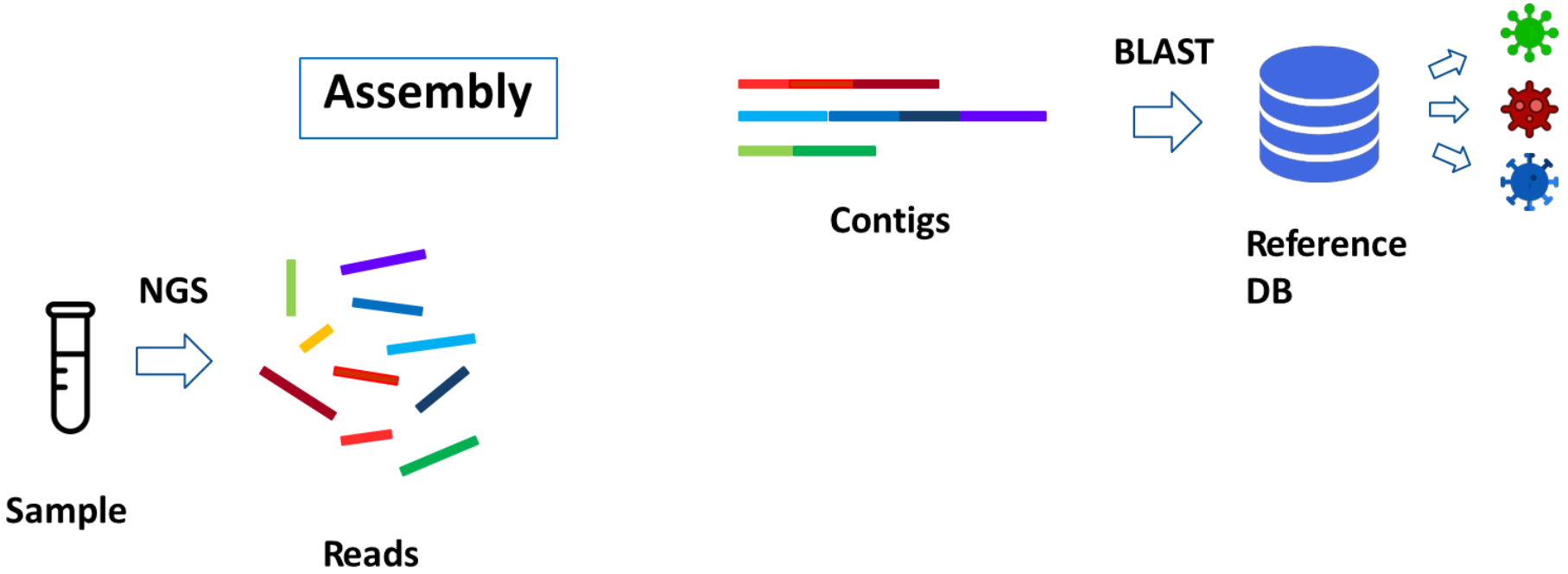
- Raros en el genoma ($< 0.1\%$)
- Los trozos similares dificultan el ensamblado correcto de lecturas pequeñas
- No todos los rRNA se amplifican en la misma medida con los *primers* universales
- Especies con diversas copias de sus genes rRNA
- No se conoce un umbral fijo de similitud que separe especies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

Metagenómica, pipeline de análisis

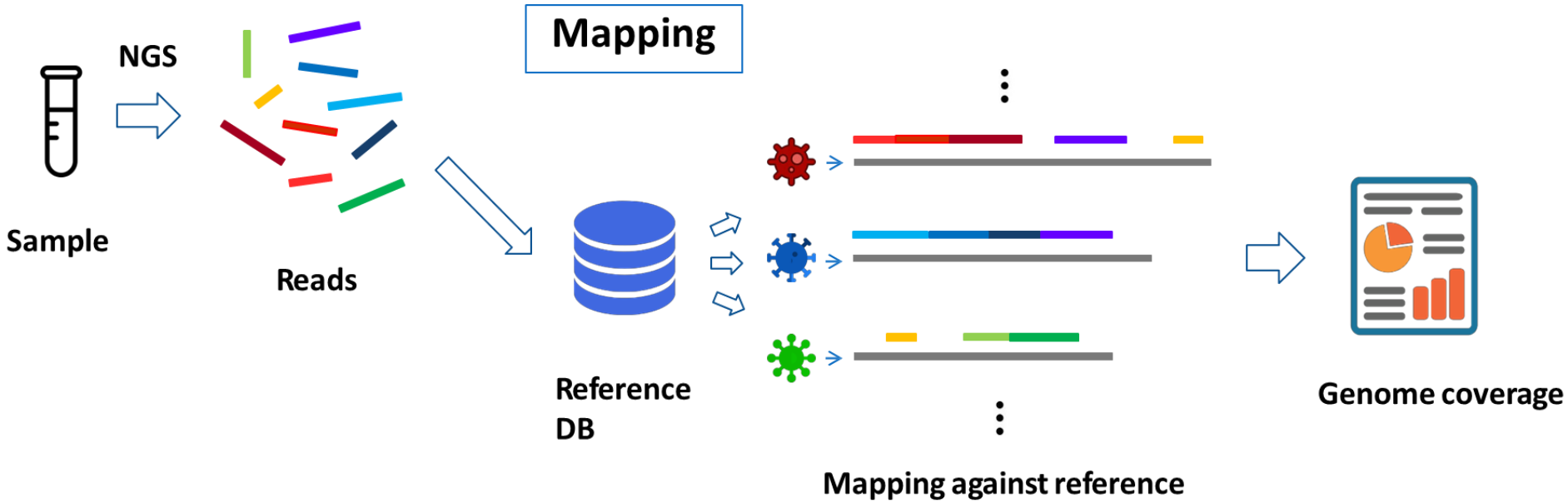


Lysholm et al., Plos One 2012:7,2, e30875

Metagenomic analysis approaches



Metagenomic analysis approaches

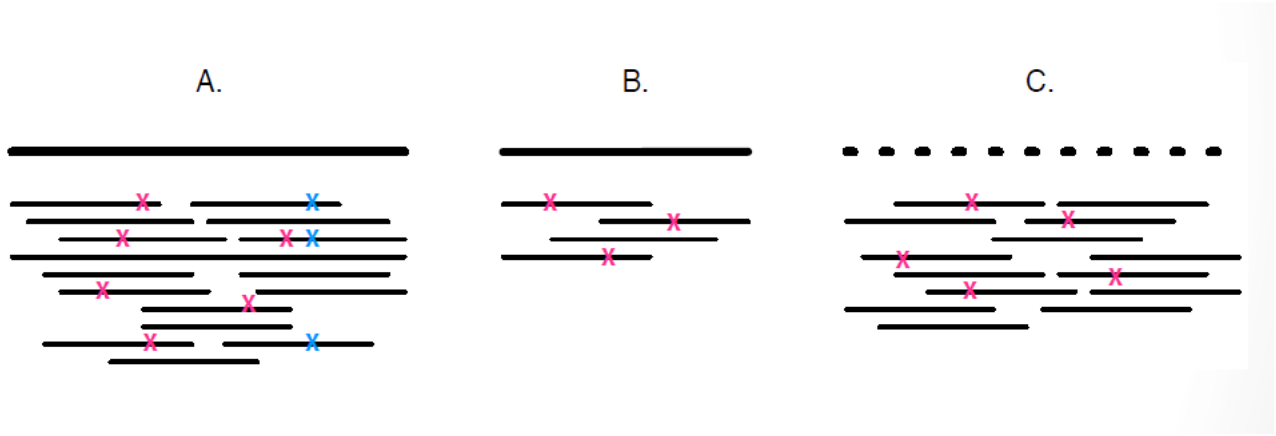


Metataxonomics vs Metagenomics (16S vs Shotgun)

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

Básicamente tres problemas

Resecuenciación, Conteo y ensamblado

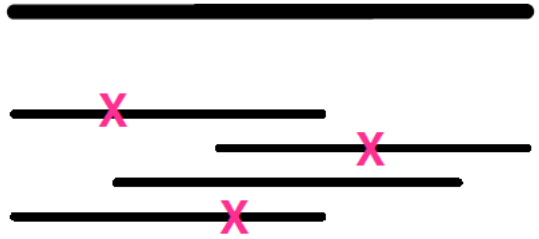


Resecuenciación

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

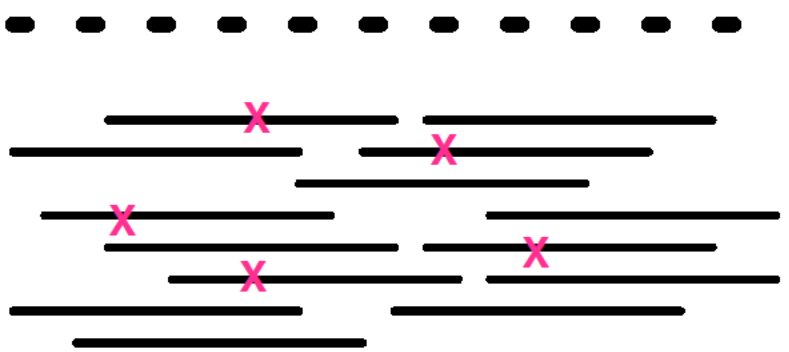
Conteo

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

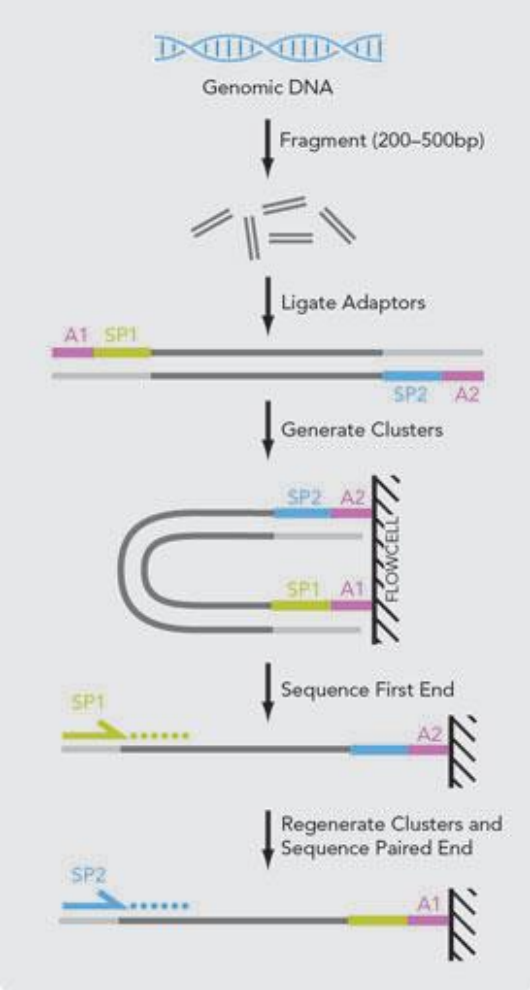


Ensamblado

No hay genoma de referencia y lo construimos de novo



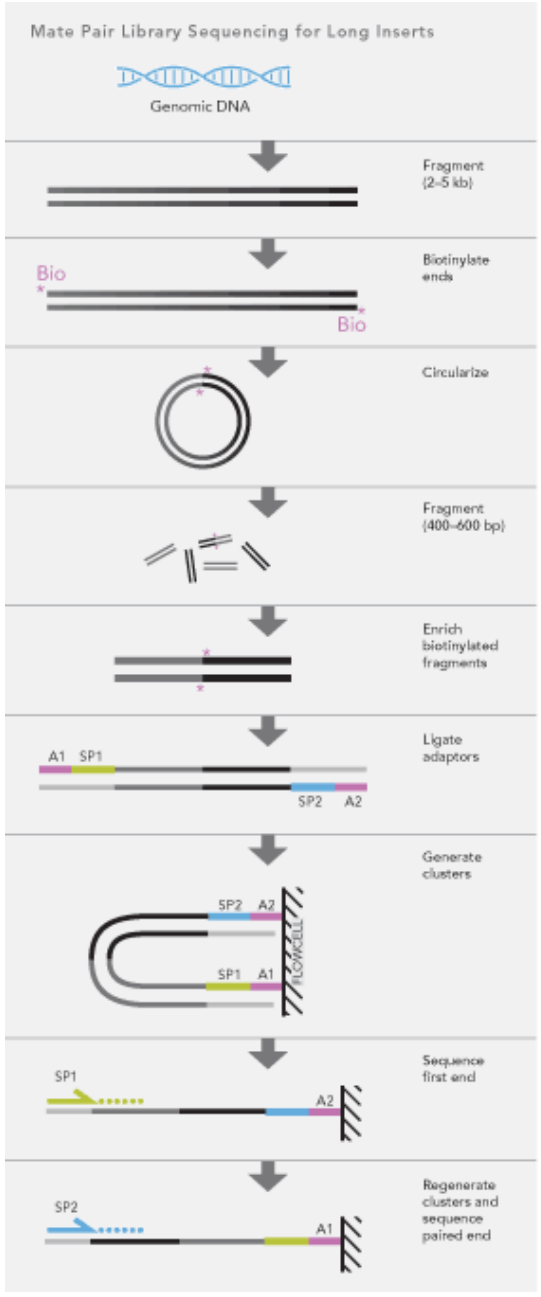
Que es Pair-end?



Secuenciación de un fragmento (bp)

Modificación de single-read DNA,
Leyendo por ambos extremos, forward y reverse

Que es Mate-pair?



Mate Pair library preparation is designed to generate short fragments that consist of two segments that originally had a separation of several kilobases in the genome. Fragments of sample genomic DNA are end-biotinylated to tag the eventual mate pair segments.
Self-circularization and refragmentation of these large fragments generates a population of small fragments, some of which contain both mate pair segments with no intervening sequence. These Mate Pair fragments are enriched using their biotin tag. Mate Pairs are sequenced using a similar two-adapter strategy as described for paired-end sequencing.

Secuenciación de dos fragmentos separados kb.

Util:
Secuenciación de un Genoma de novo
Finalizar un genoma
Detección de variantes estructurales

Sequencing terms

Depth of coverage

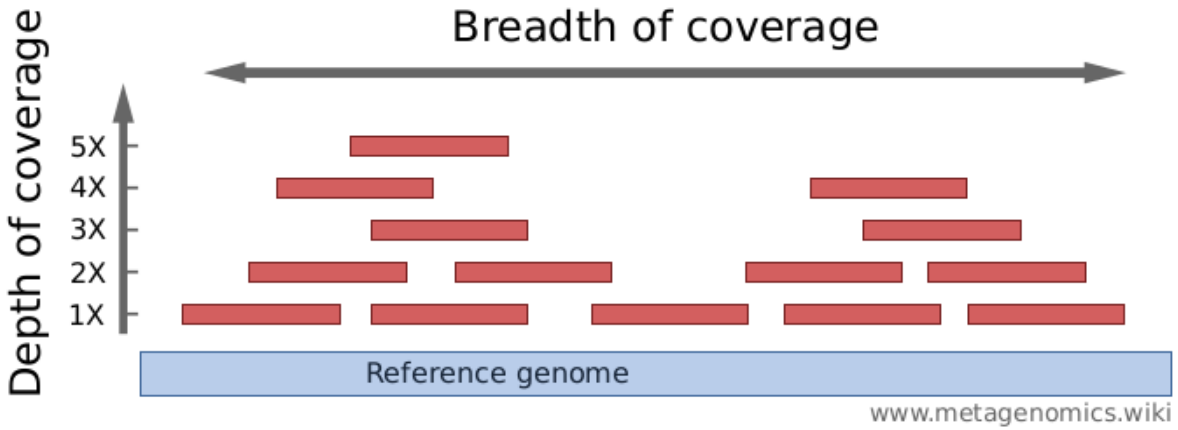
How strong is a genome "covered" by sequenced fragments (short reads)?

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).

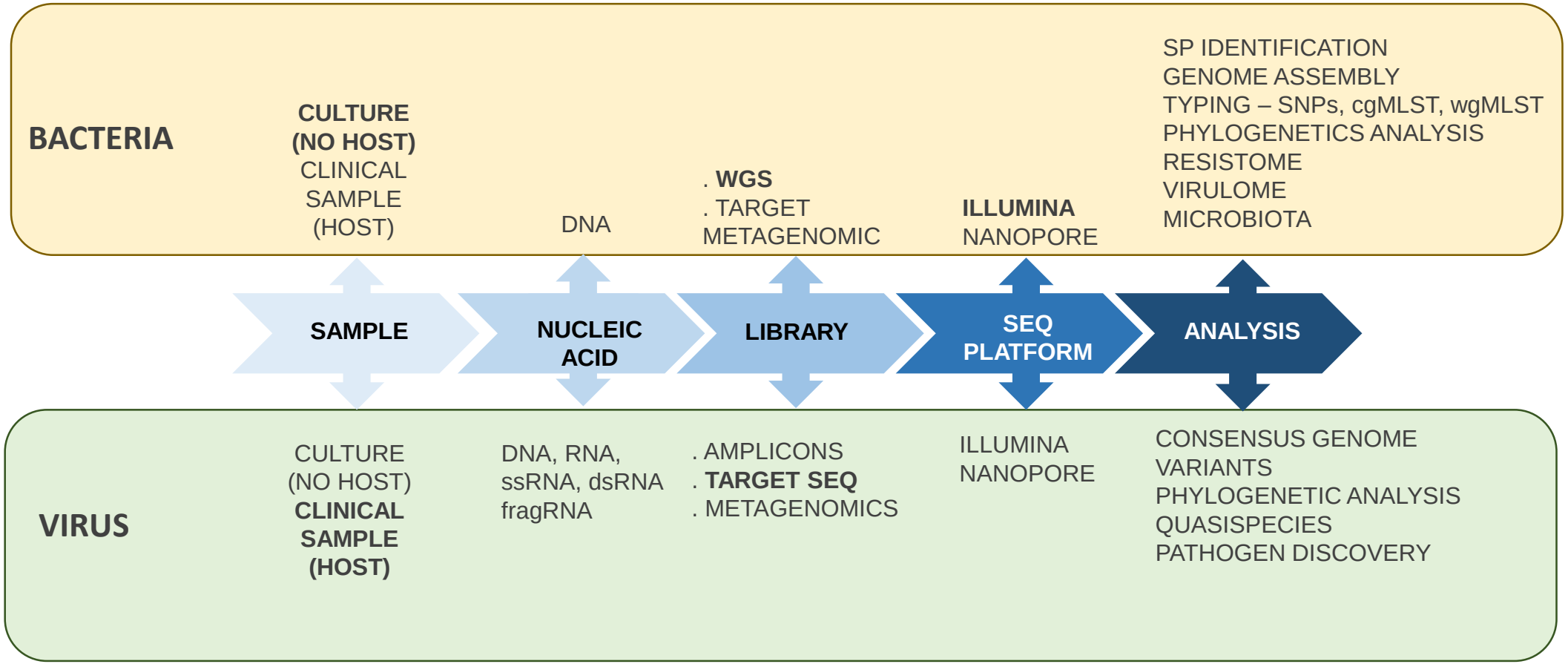
Breadth of coverage

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.



Bacterial and Viral Genome Sequencing



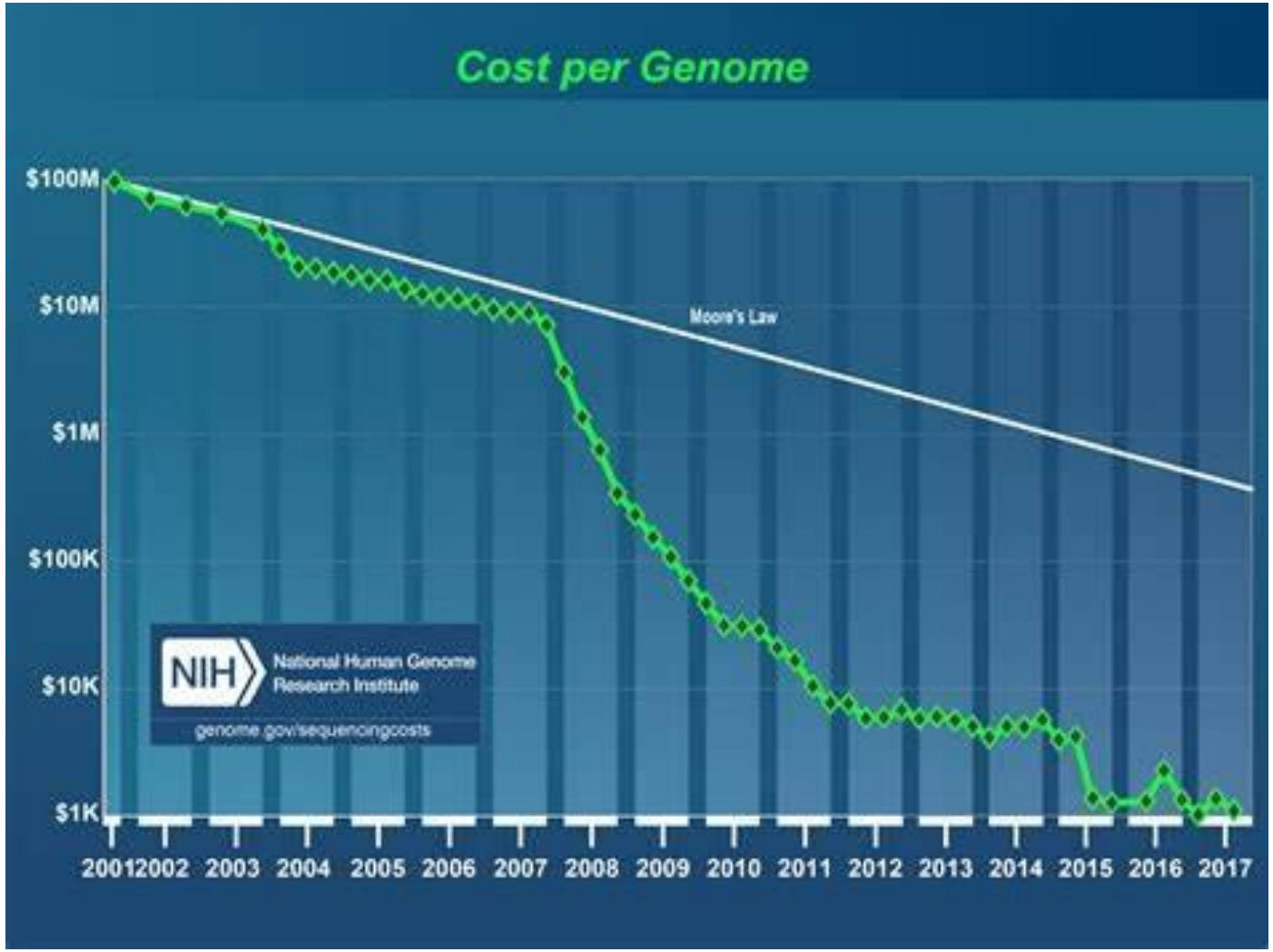
Index

- NGS applications
- Microbial genomics
- Library strategies
- Bioinformatics analysis
- Challenges in Bioinformatics

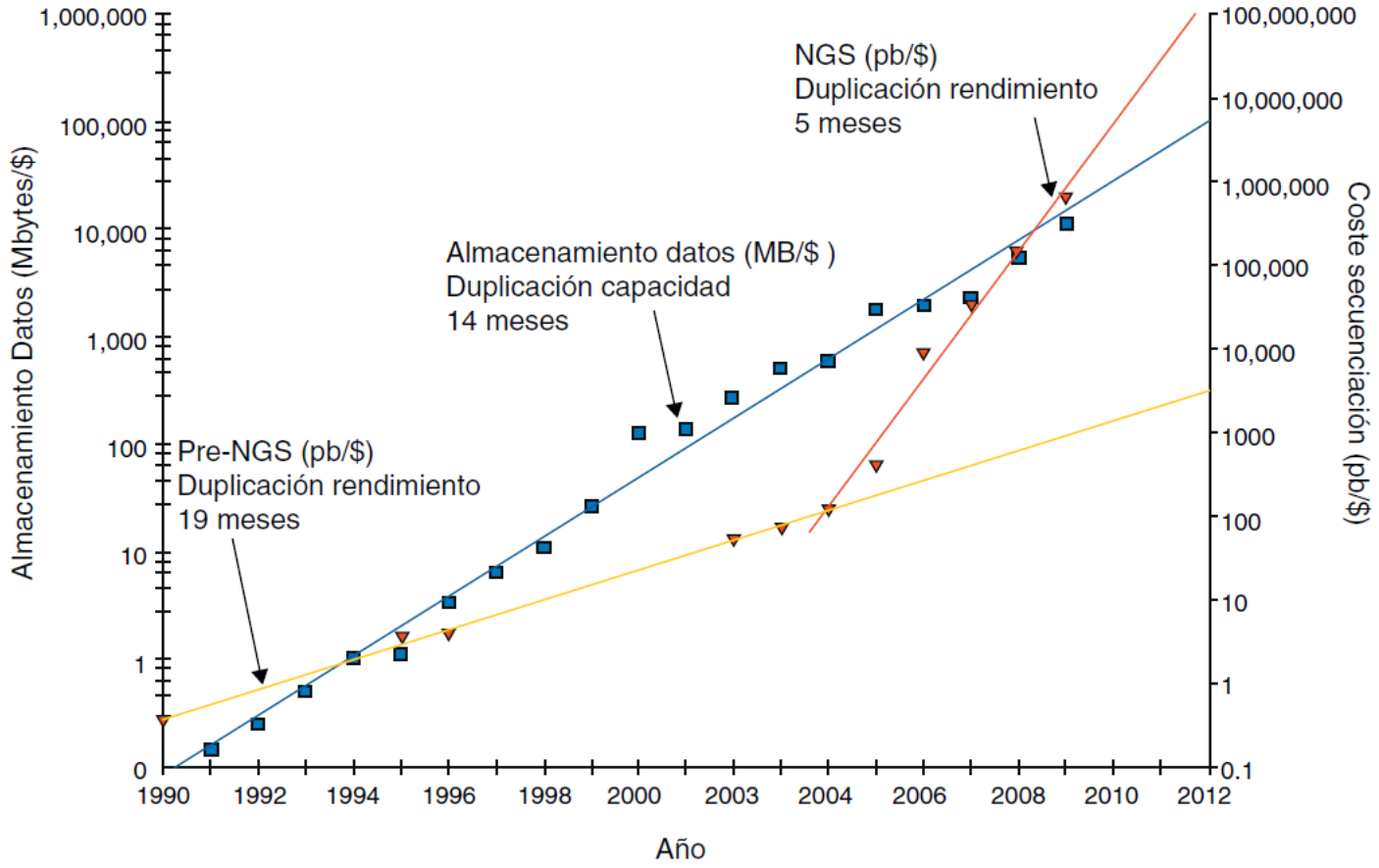
Retos de la Bioinformática en NGS

- 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 1: 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)

Thanks for your attention!

Questions???