UNU BIOLAC

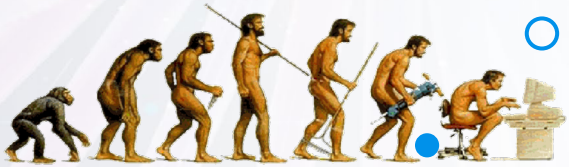
Cursos Internacional .
Secuenciación y Análisis de Datos Genómicos para la Detección Microbiológica de Enfermedades Transmitidas por Alimentos y Aguas

INSTITUTO NACIONAL DE HIGIENE RAFAEL RANGEL

ANÁLISIS DE DATOS PARA SECUENCIACIÓN DE NUEVA GENERACIÓN

Una aproximación a la genómica bacteriana

Rafael Puche Q. rafael.puche@pedeciba.edu.uy @rpucheq
Sección de Bioinformática del Servicio de Secuenciación de ADN
Unidad de Estudios Genéticos y Forenses (UEGF)
Centro de Microbiologia y Biologia Celular
Instituto Venezolano de Investigaciones Científicas (IVIC)



UEGF Unidad de Estudios Genéticos y Forenses

IVIC

Mincyt Ministerio del Poder Popular para Ciencia y Tecnología

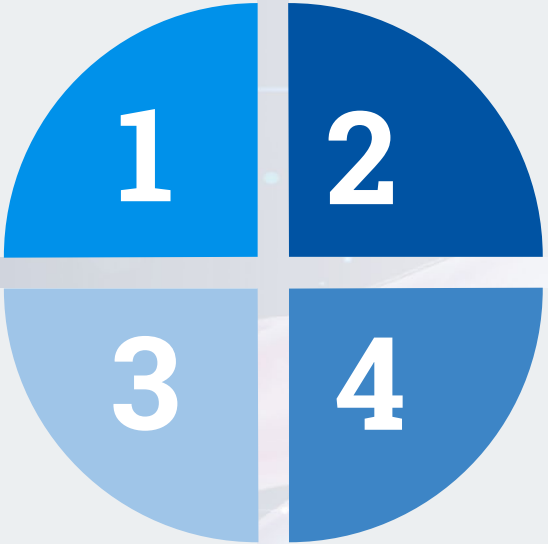
Gente, Ciencia y Tecnología al Servicio de la Salud

- **Biólogo** - LUZ.
- **Master en Bioinformática -** UdelaR. / **PhD** - UCV
- **Profesional Asociado a la Investigación** - IVIC
- **Coordinador de la sección de Bioinformática** - UEGF
- ***Fellow* en Bioinformática:** NGS - UBI, Institut Pasteur de Montevideo
- **Profesor Invitado** - INHRR - Instituto Anatomico, UCV
- ***Fellow* en Bioinformática**: Genómica Bacteriana - CABANA (UK); Universidad de Costa Rica (CIET)
- **Presidente del RSG - Venezuela**, International Society for Computational Biology, Student Council (ISCB-SC)

Sumario

Introducción

Bioinformática - NGS - Genómica



Almacenamiento y Análisis de datos NGS

Workflow - Formatos - Análisis

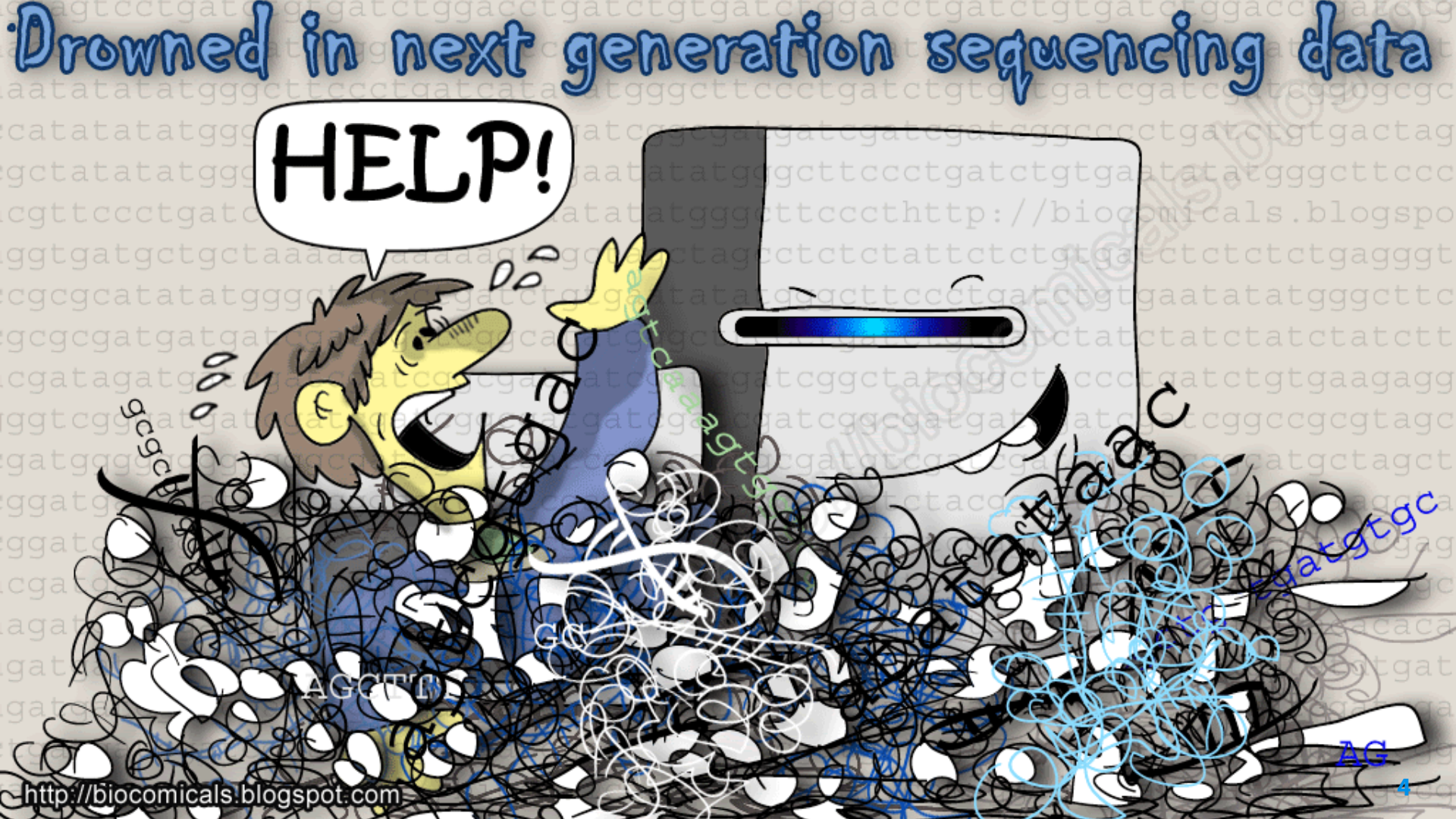
Conceptos-Metodologías - Aplicaciones

Genomica Bacteriana

Aplicaciones - Divulgación- Iniciativas

Contexto Gobal y local de la Bioinformática y Genomica

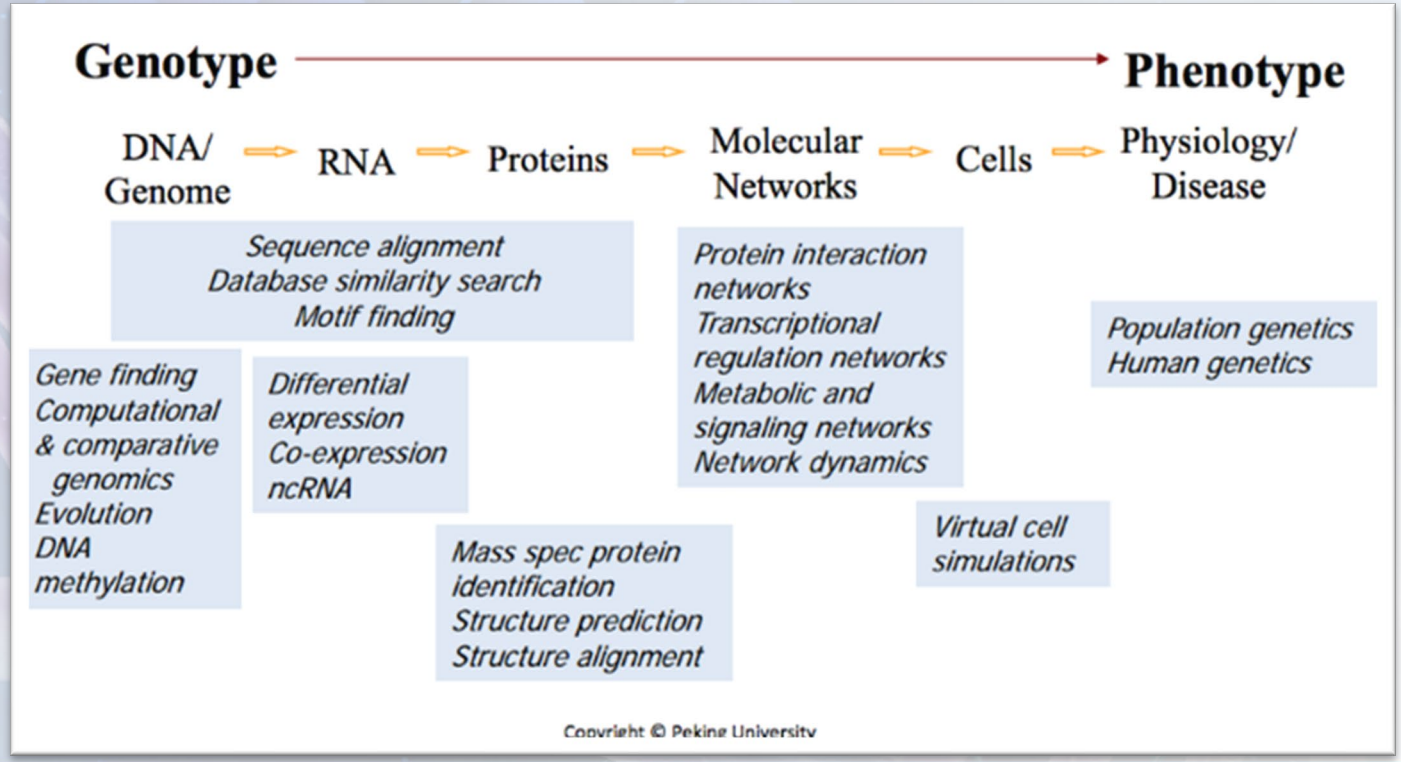
Drowned in next generation sequencing data



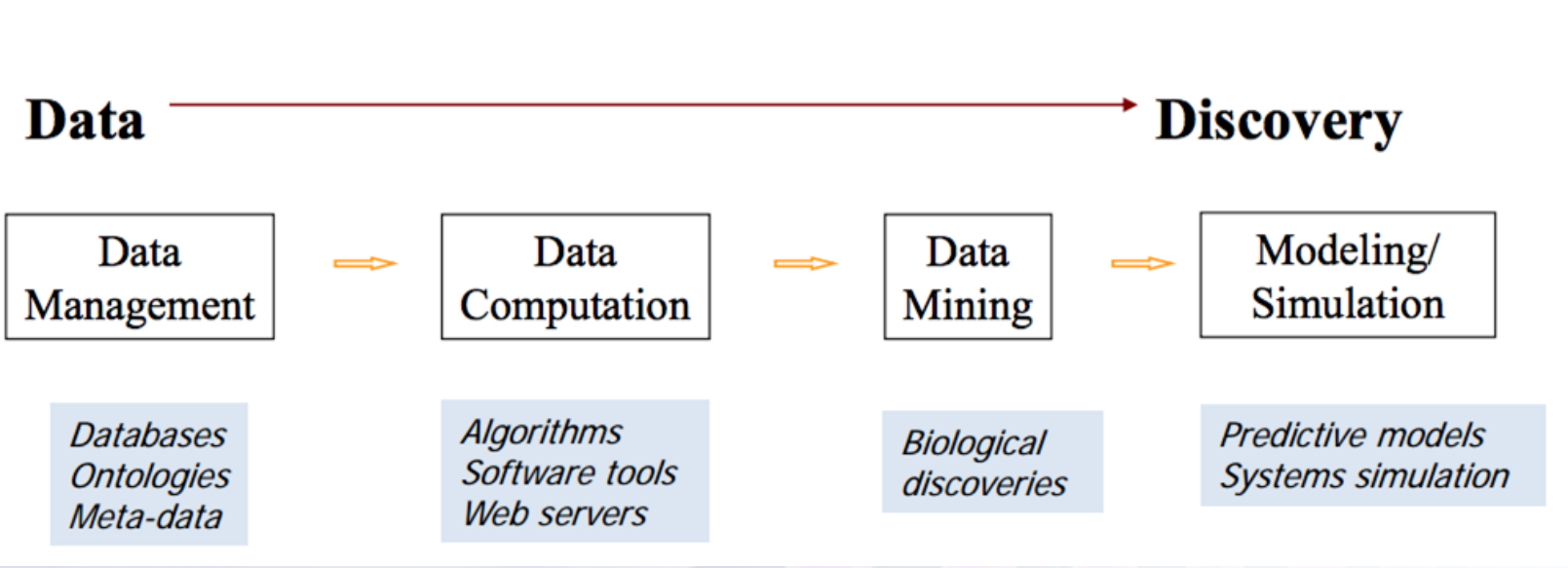
Introduccion

Para entender el contexto del análisis bioinformático

El *prefijo* biológico…



El *sufijo* informático…

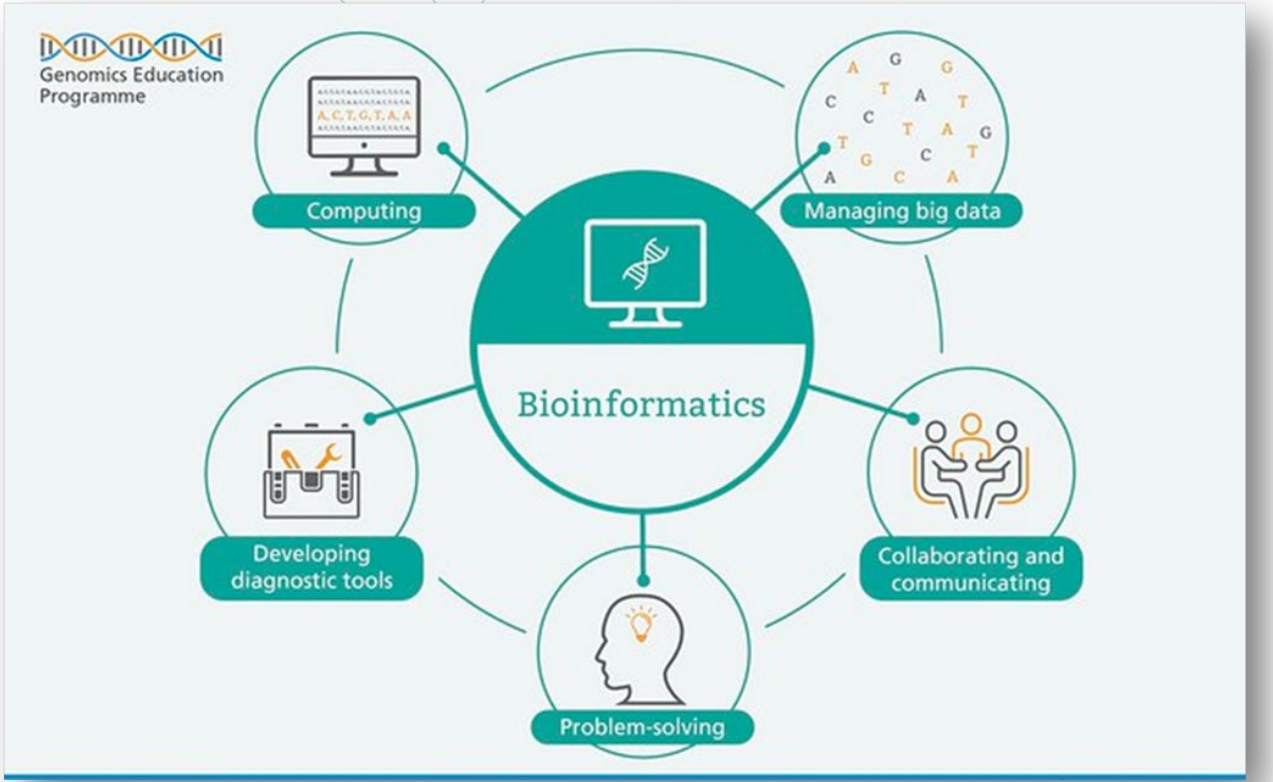


nature

Es un campo interdisciplinario que desarrolla y aplica metodologías computacionales que permiten analizar gran cantidad de datos biológicos, así como secuencias genéticas y proteicas, que permiten hacer nuevas predicciones o descubrir nuevos significados a procesos biológicos.

http://www.nature.com/subjects/computational-biology-and-bioinformatics

NHS England



https://www.genomicseducation.hee.nhs.uk/image-library/

EMBL-EBI

Es la ciencia del almacenamiento, la recuperación y el análisis de grandes cantidades de información biológica.

…

Se trata de un campo interdisciplinario en el que intervienen muchos tipos de especialistas, como biólogos, informáticos y matemáticos.

https://www.ebi.ac.uk/training/online/courses/bioinformatics-terrified/what-bioinformatics/

Finally, a definition for bioinformatics

7TH SEPTEMBER 2015 / BIOMICKWATSON / 5 COMMENTS

Following the trend for ultra-short-but-to-the-point blog posts, I have decided to finally define bioinformatics:

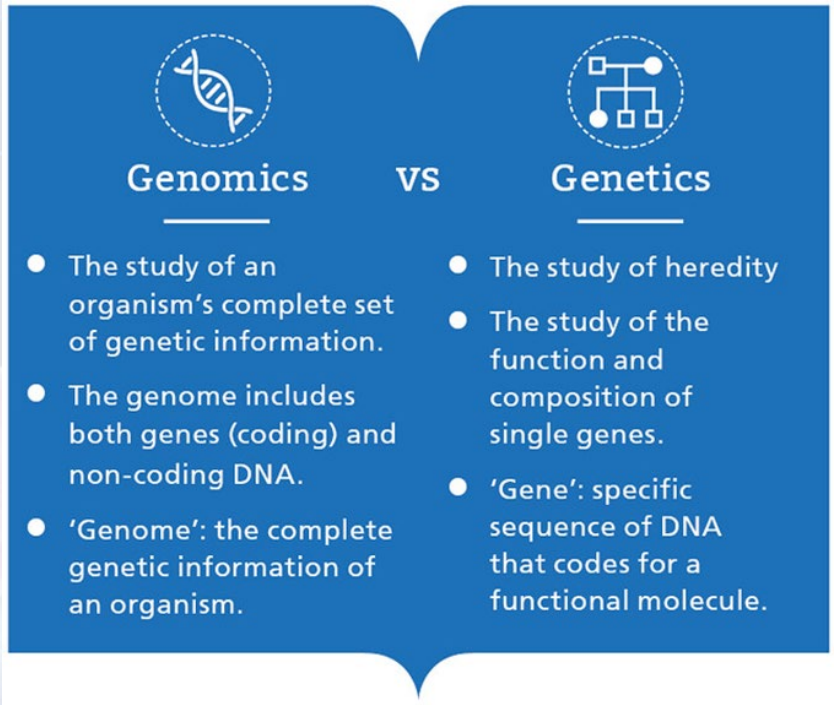
bio-informatics (bi-oh-in-foh-shit-I-don't-understand-how-that-works)

From the word *bio*, meaning "of or related to biology" and *informatics*, meaning "absolutely anything your collaborators or boss don't understand about maths, statistics or computing, including why they can't print and how the internet works"

Personal Chair of
Bioinformatics and
Computational Biology
University of Edinburgh

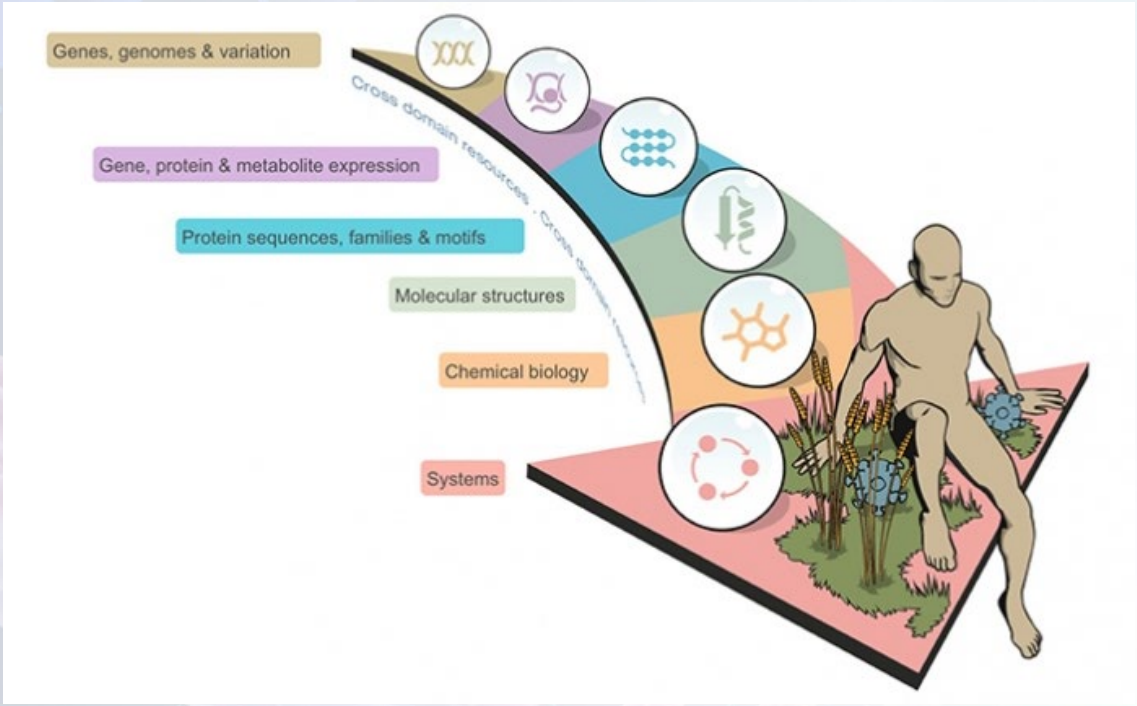
Genomica

- Estudio de un conjunto completo del ADN de un organismo e incorpora elementos de la genética.
- Usa métodos de secuenciación (NGS) y bioinformática para ensamblar y analizar la estructura y función del genoma.
- Se enfoca en la estructura, función, evolución y mapeo de los genomas.
- Es posible que esto lleve a nuevas maneras de diagnosticar, tratar y prevenir enfermedades. (Medicina personalizada)



Diferentes tipos de datos

dentro del ámbito de la bioinformática



¿Cuánto ADN tienen los organismos?

organism	genome size (base pairs)	protein coding genes	number of chromosomes
model organisms			
model bacteria *E. coli*	4.6 Mbp	4,300	1
budding yeast *S. cerevisiae*	12 Mbp	6,600	16
fission yeast *S. pombe*	13 Mbp	4,800	3
amoeba *D. discoideum*	34 Mbp	13,000	6
nematode *C. elegans*	100 Mbp	20,000	12 (2n)
fruit fly *D. melanogaster*	140 Mbp	14,000	8 (2n)
model plant *A. thaliana*	140 Mbp	27,000	10 (2n)
moss *P. patens*	510 Mbp	28,000	27
mouse *M. musculus*	2.8 Gbp	20,000	40 (2n)
human *H. sapiens*	3.2 Gbp	21,000	46 (2n)

¿El tamaño genómico importa?

Genome assembly AmbMex60DD

reference

Download

Submitted sequence	GenBank GCA_002915635.3
Taxon	*Ambystoma mexicanum* (axolotl)
Strain	DD151
WGS project	PGSH02
Submitter	Max Planck Society/University of Kentucky
Date	Apr 1, 2021

View the legacy Assembly page

Assembly statistics

These statistics describe the nuclear genome of the submitted sequence, GCA_002915635.3

Genome size	28.2 Gb
Number of chromosomes	28
Number of scaffolds	27,157
Scaffold N50	1.2 Gb
Scaffold L50	11
Number of contigs	211,437
Contig N50	218 kb
Contig L50	35,415
GC percent	46
Assembly level	Chromosome

Genome assembly GRCh38.p14

reference

Download

Reference sequence	RefSeq GCF_000001405.40
Submitted sequence	GenBank GCA_000001405.29
Taxon	*Homo sapiens* (human)
Submitter	Genome Reference Consortium
Date	Feb 3, 2022

View the legacy Assembly page

Assembly statistics

These statistics describe the nuclear genome of the reference sequence, GCF_000001405.40

Genome size	3.1 Gb
Number of chromosomes	24
Number of scaffolds	470
Scaffold N50	67.8 Mb
Scaffold L50	16
Number of contigs	996
Contig N50	57.9 Mb
Contig L50	18
GC percent	40.5
Assembly level	Chromosome

Leptospira venezuelensis ☆

Leptospira venezuelensis is a species of bacteria in the family *Leptospiraceae*.

Browse taxonomy

Current scientific name	*Leptospira venezuelensis*
Taxonomic rank	species
NCBI Taxonomy ID	1958811

For more details see NCBI Taxonomy

Genome

Browse all 6 genomes

Reference genome ASM215005v1

Venezuelan Scientific Research Institute (2017). Strain: CLM-R50.

RefSeq GCF_002150055.1

Download

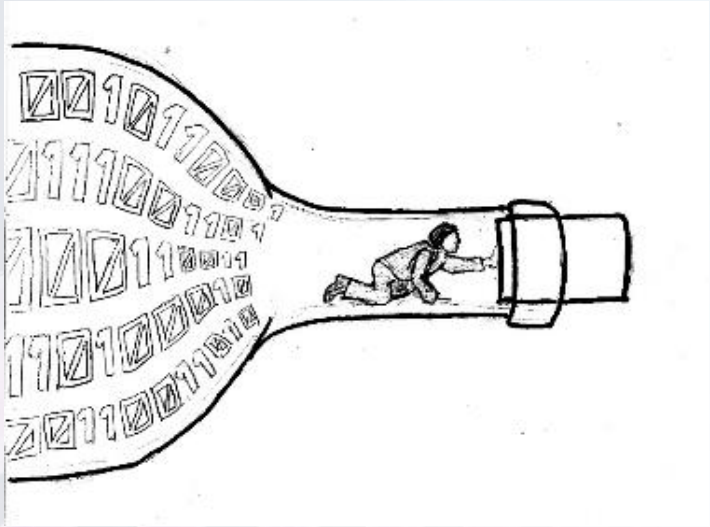
Genome size	4.3 Mb
Contig N50	895.8 kb
Genes	4,018

NCBI Prokaryotic Genome Annotation Pipeline (PGAP)

Secuenciación Masiva

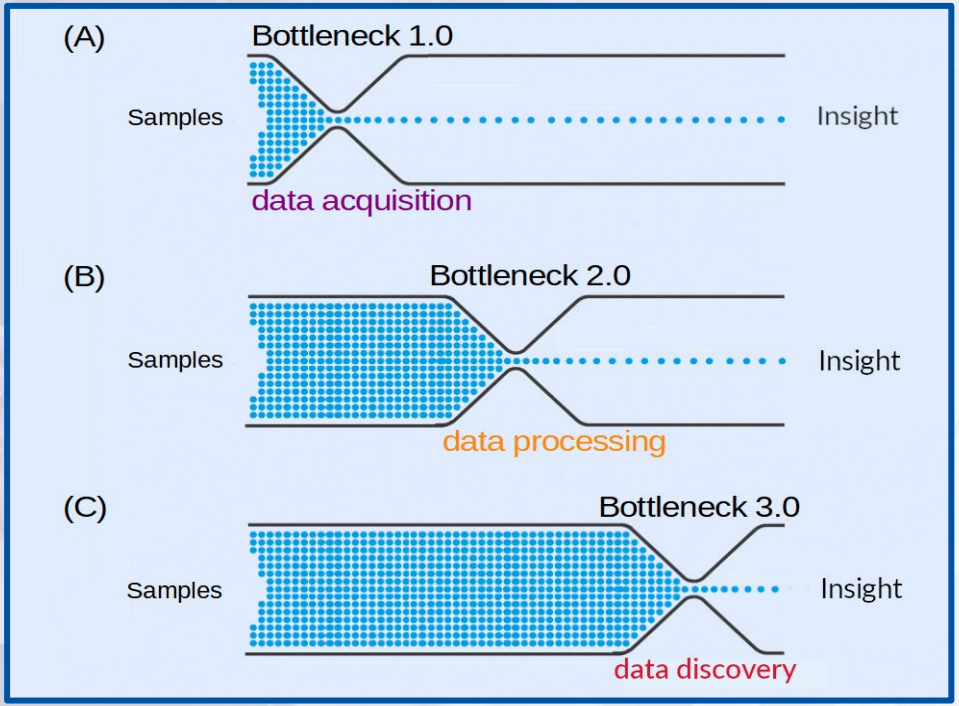
Millones o billones de moléculas de ADN pueden ser secuenciadas en paralelo, mejorando sustancialmente el rendimiento y minimizando el uso de fragmentos clonados, normalmente usados en la secuenciación de genomas en Sanger.

El actual *"cuello de botella"* de los proyectos de secuenciación whole-genome (WGS) y whole-exome (WES), es la forma de estructurar el manejo de la data y los sofisticados análisis computacionales de los datos generados

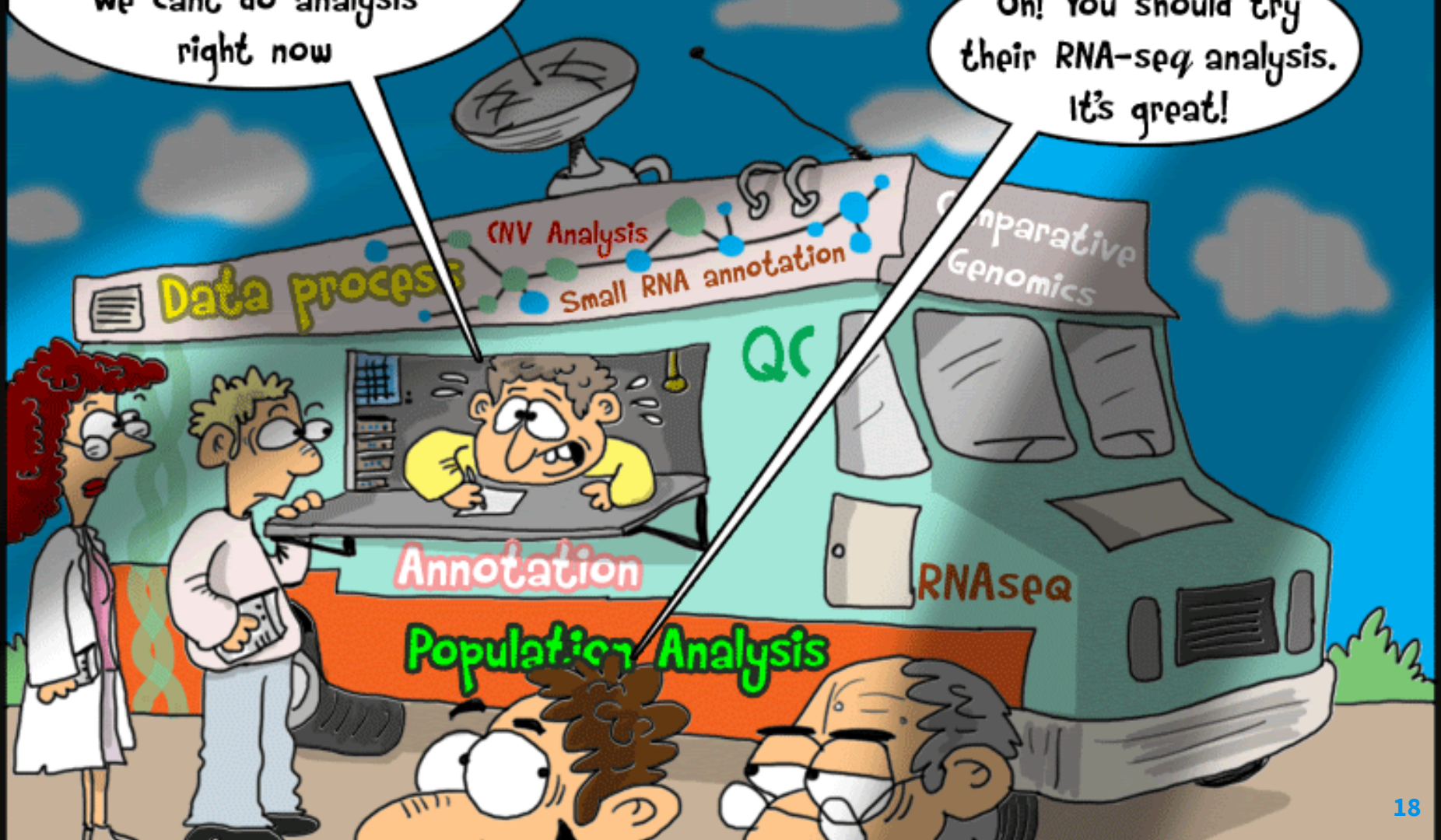


Pabinger et al. *Briefings in Bioinformatics*, 2013

El “cuello de botella”…



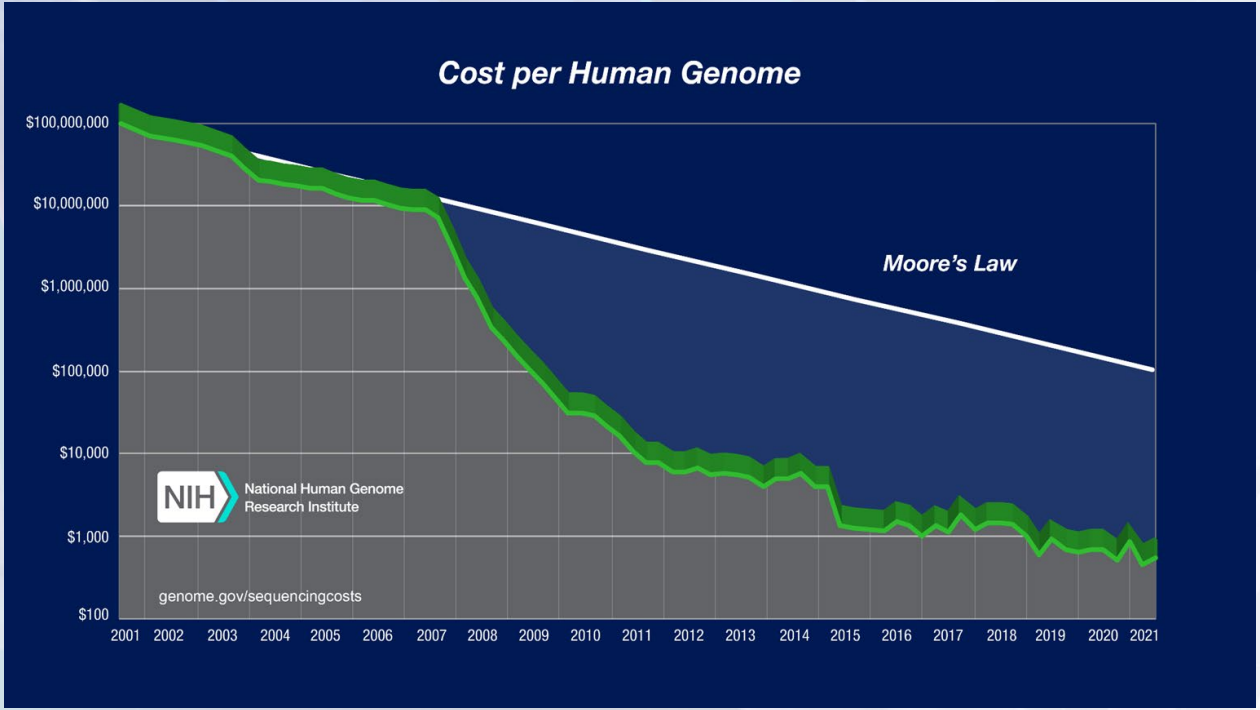
Martinelli, A. The Shifting Bottleneck in Genome Data Research, 2021


we can't do analysis right now
Oh! You should try their RNA-seq analysis. It's great!
Data process
CNV Analysis
Small RNA annotation
Comparative Genomics
QC
Annotation
RNAseq
Population Analysis

Análisis de Datos NGS

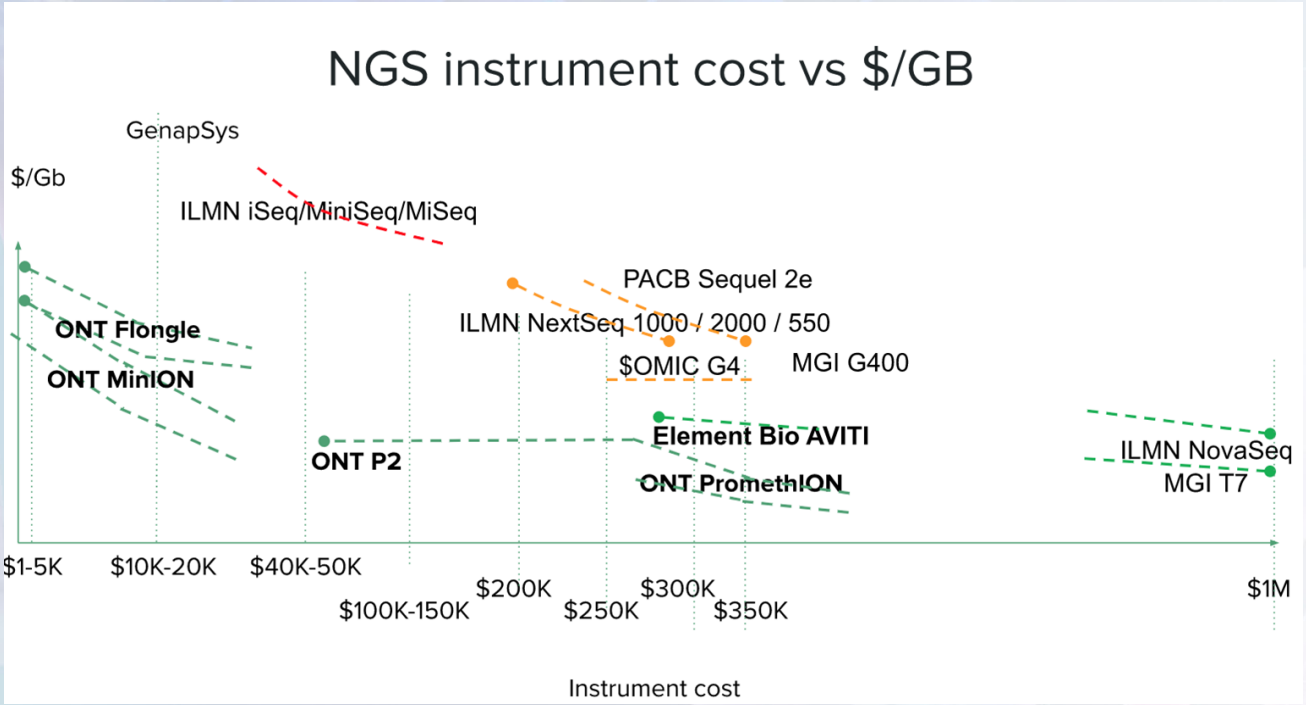
Generación - Formatos - Análisis - Workflow

¿Es la NGS realmente costosa? - Genoma



- **Ley de Moore:** cada dos años se duplica el número de transistores en un microprocesador.
- La consecuencia directa de la ley es que los precios bajan al mismo tiempo que las prestaciones suben

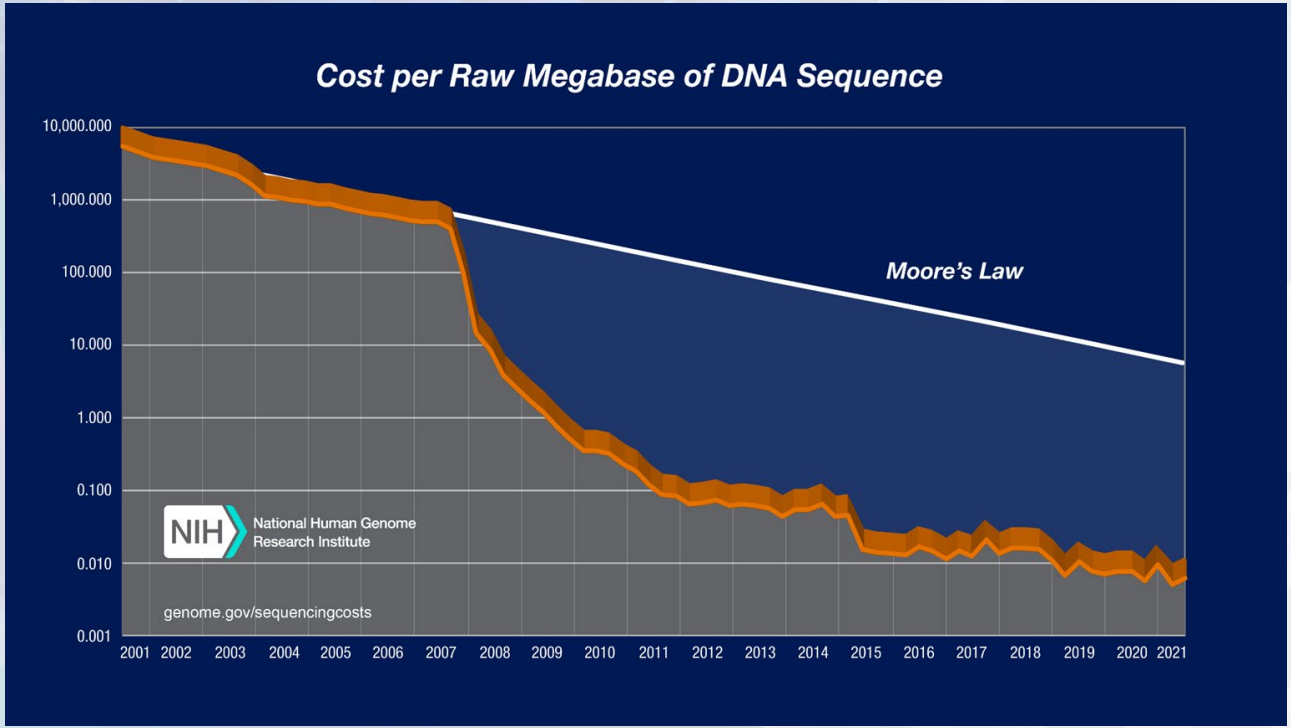
¿Es la NGS realmente costosa? - Secuenciadores



- Precio de los instrumentos *vs* Gb/$ para los principales instrumentos de NGS que existen

Albert Vilella, (2022) https://twitter.com/AlbertVilella/

¿Es la NGS realmente costosa? - Datos



Estructura de Costos:

- Mano de obra, administración, gestión, servicios públicos, reactivos y consumibles
- Instrumentos de secuenciación y otros equipos grandes (amortizados en tres años)
- Actividades informáticas directamente relacionadas con la producción de secuencias (p. ej., sistemas de gestión de información de laboratorio y procesamiento inicial de datos)
- Envío de datos a una base de datos pública

https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

¿Es la NGS realmente costosa? - Datos/Analisis

Rates

Data Analysis (Per Hour)

Custom bioinformatics analysis from experimental design to publication.

UC Campus	$100
Non-UC Academic	$158
Private Enterprise	$195

Biostatistics (Per Hour)

Statistical analysis of bioinformatics data.

UC Campus	$160
Non-UC Academic	$251
Private Enterprise	$310

FLORIDA STATE UNIVERSITY

THE CENTER FOR GENOMICS AND PERSONALIZED MEDICINE

HOME RESEARCH NEWS TRAINING PUBLICATIONS PEOPLE SERVICES & FEES

HOME / SERVICES & FEES

SERVICES & FEES

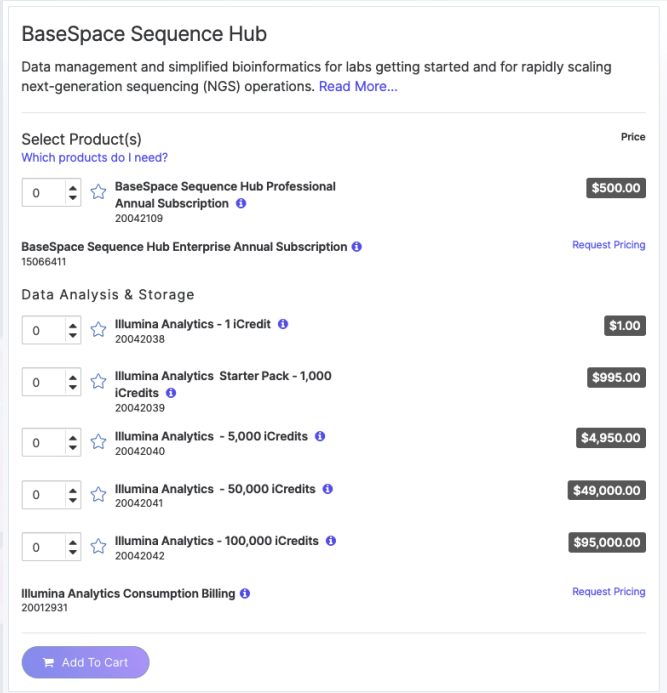
- **The custom services provided by CGPM may include but are not limited to RNA-seq, ChIP-seq, DNA methylation, DNA variation, metagenomics, CRISPR Screening. Since estimating the cost of a bioinformatics project is usually quite difficult, we offer two different ways of working with us. For small, routine projects (e.g. standard RNA-seq analysis), we charge based on a fixed fee schedule which can be found below. For large projects, we charge based on the amount of effort. We will estimate how much effort for each project and notify you when we officially initiate your project.**
- **RNAseq Analysis**
  - Gene counts generation: $200 setup + $40/sample, $25~30/sample if sample size over 20.
  - Statistical analysis: $250 setup + $50/sample, $35~40/sample if sample size over 20.
  - Deliverables:
    - Alignment Files in BAM format
    - Gene / Transcript Expression Level File in MS Excel format
    - BigWig File For count visualization in Integrative Genomics Viewer (IGV)
    - List of Differentially Expressed Gene / Transcript in MS Excel format (includes integration into Geneious software [License held by the Department of Biological Science, FSU] for downstream analysis)
    - List of enrichment GO terms for differentially expressed genes in MS Excel format
    - Volcano plots for pairwise analysis of each sample
    - Barplot for enriched GO terms for differentially expressed genes in PNG/PDF format

FLORIDA STATE UNIVERSITY

https://bioinformatics.ucdavis.edu/services-rates // https://genomics.fsu.edu/services-fees

¿Es la NGS realmente costosa? - Datos/Análisis

COST PER SAMPLE ANALYSIS RATES							
Organization / Assay	Hourly Rate	RNA-seq	Human Exome	Human WGS	Gene Panels	10x Single Cell	ChIP-seq
UNIVERSITY OF NEBRASKA MEDICAL CENTER	$50-$75	$75-$100	$125-$150	$125-$150	$30-$50	$50-$250	
UC DAVIS	$97-$310	$80-$155	$100-$193	$100-$193			
UNIVERSITY OF KANSAS MEDICAL CENTER	$80	$180-$540	$180-$360	$180-$360	$120-$180		$120-$360
ARIZONA STATE UNIVERSITY	$50	$90-$115	$125	$125	$175	$200-$300	$115

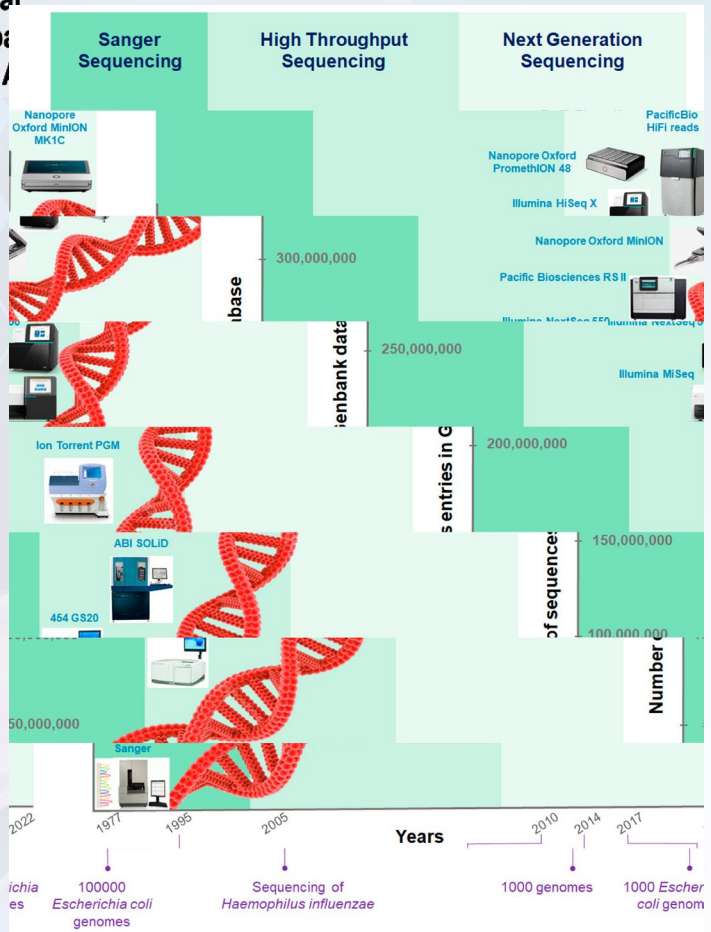


You can make an account on Basespace for free but executing jobs requires a $500 annual subscription fee and purchasing iCredits.

https://medium.com/truwl/what-is-the-cost-of-bioinformatics-a-look-at-bioinformatics-pricing-and-costs

Ecosistema NGS

- **2007** se secuenciaron 1000 genomas bacterianos
- **2014** se secuencio el genoma 1000 de E. coli
- **2017** se alcanzo el genoma 100.000 de *E. coli*
- **2021** mas de 376.000 genomas de *E. coli* secuenciados



Khedher *et al.* (2022), International Journal of Molecular Sciences

Secuenciación genómica

Rápida comparativa de la tecnología de NGS

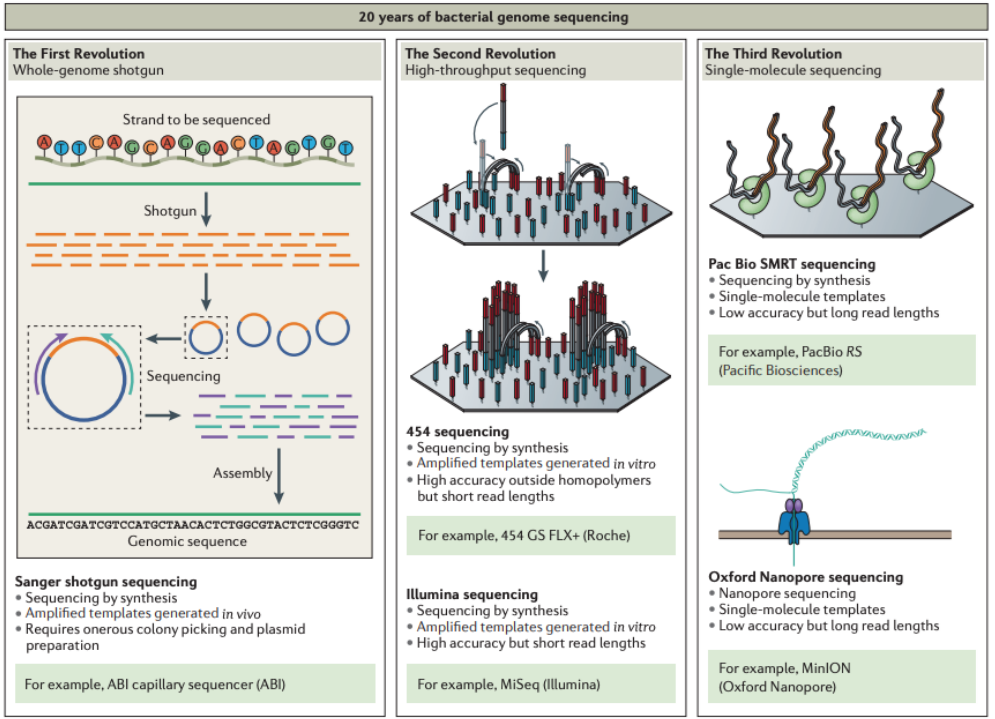


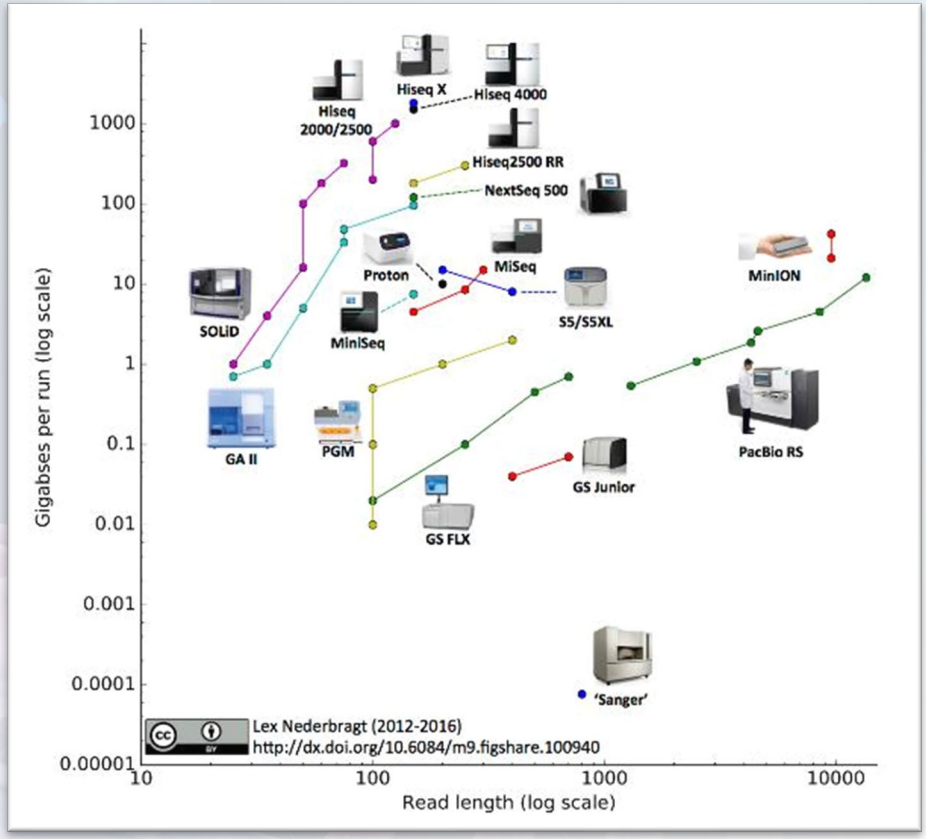
Figure 2 | **Bacterial genomics: the first two decades.** The three revolutions in sequencing technology that have transformed the landscape of bacterial genome sequencing are as follows: whole-genome shotgun sequencing, high-throughput sequencing, and single-molecule long-read sequencing. SMRT, single-molecule real-time.

Loman, N., and Pallen, M. (2015). *Twenty years of bacterial genome sequencing.* Nature

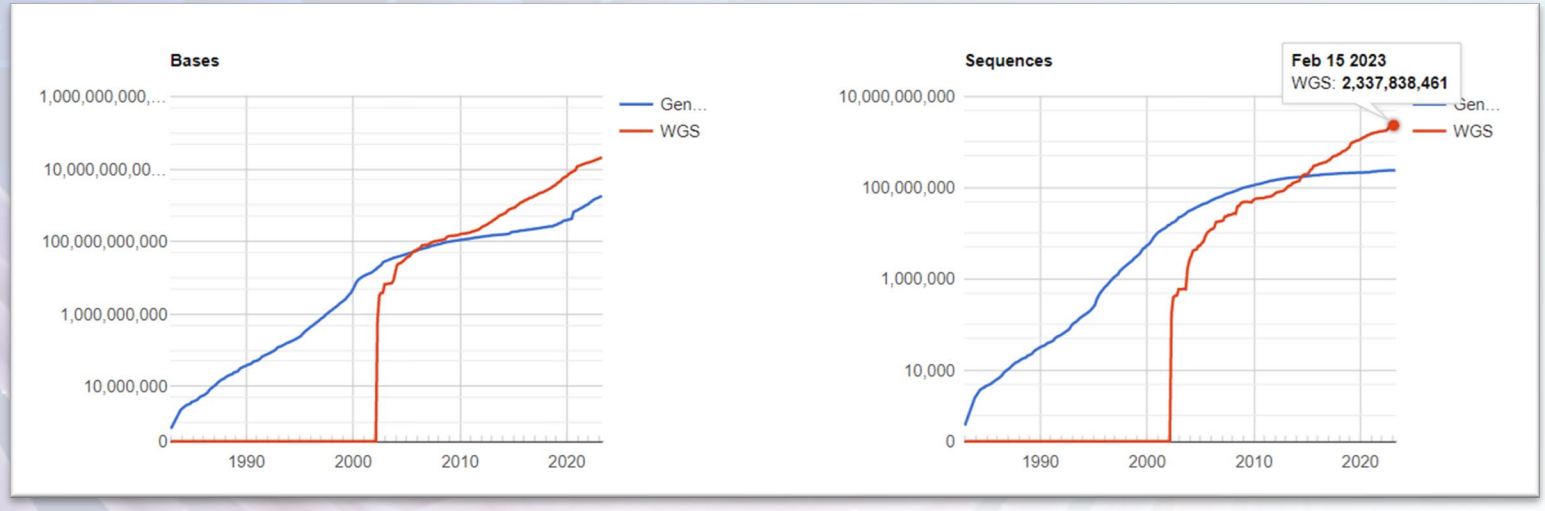
Ecosistema NGS

Rápida comparativa

- **Sanger:** 1 Kb
- **MiSeq:** 10-20 Gb
- **MinION:** >20 Gb*



Estadísticas del Genbank



Comparativa:

- Abril 2017: > 450 millones de secuencias WGS
- Abril 2018: > 600 millones de secuencias WGS
- Febrero 2022: > 1.750 millones de WGS
- Febrero 2023: > 2.300 millones de WGS

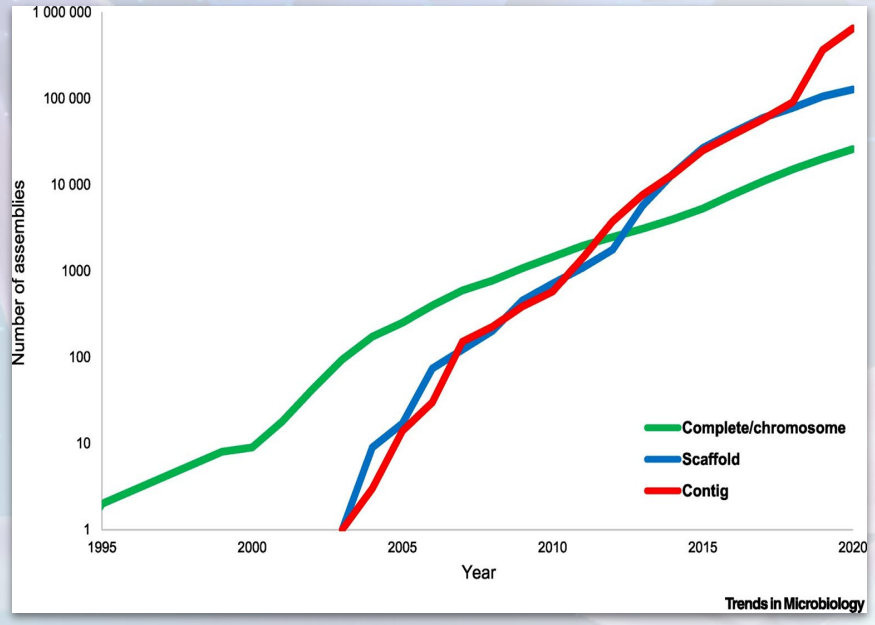
Las secuencias subidas a Genbank se duplican (aprox.) cada 18 meses

- SRA en Junio 2023 alcanzo los 32 Pb

https://www.ncbi.nlm.nih.gov/genbank/statistics/

Estadísticas en Procariotas

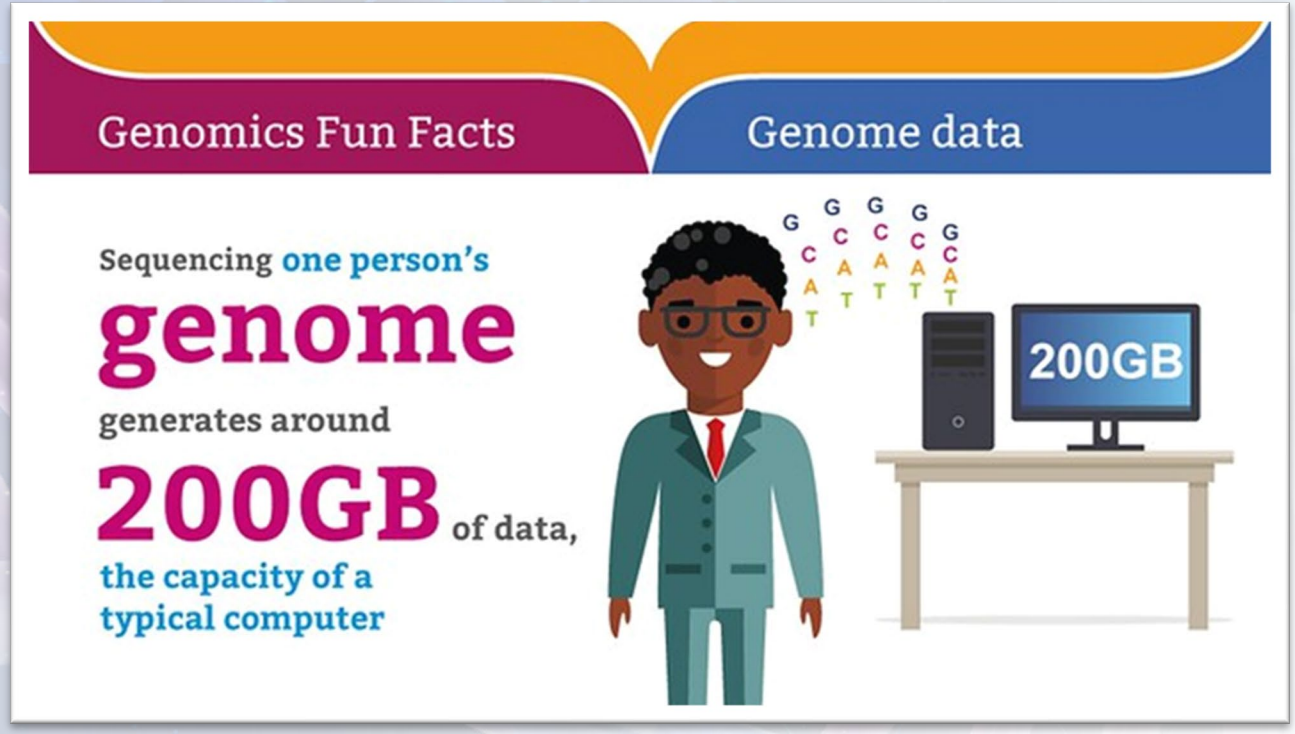
GenBank - diferentes niveles de ensamblaje



- **Genoma completo:** todos los cromosomas están completamente ensamblados con *gaps* que no superan las diez bases.
- **Cromosoma:** todos los cromosomas están completamente ensamblados, pero posiblemente contienen huecos o *scaffolds* no localizados
- **Scaffolds:** los *contigs* de la secuencia están conectados a través de los *gaps*, pero no están organizados en los cromosomas.
- **Contig:** sólo hay contigs no ordenados.

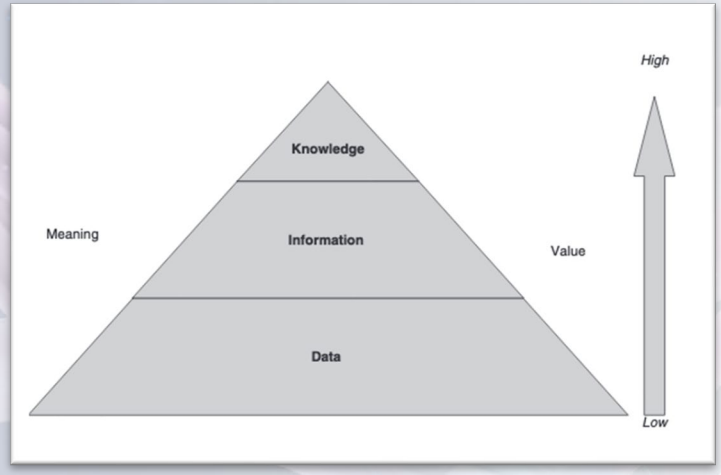
Koonin *et al.* (2021) *Evolution of Microbial Genomics: Conceptual Shifts over a Quarter Century*

Raw Data Genomica



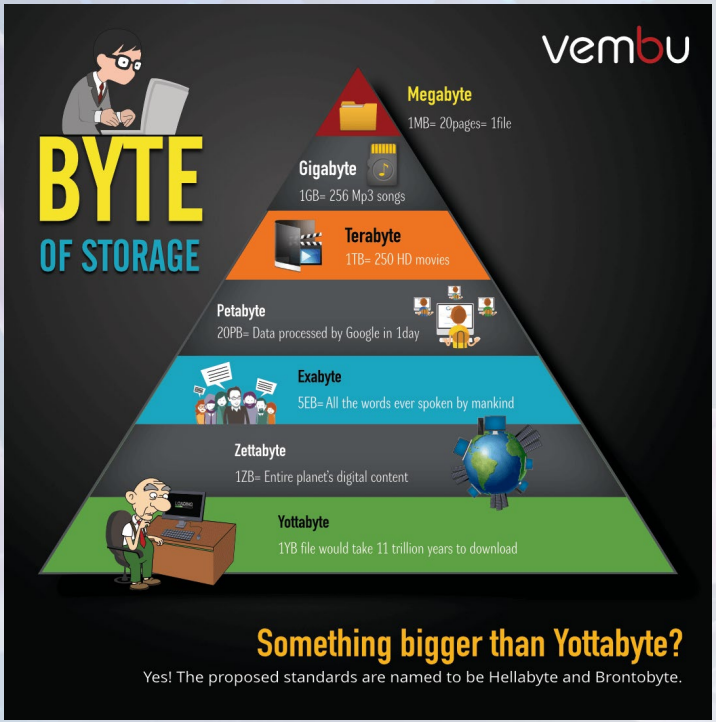
Data, informacion y conocimiento

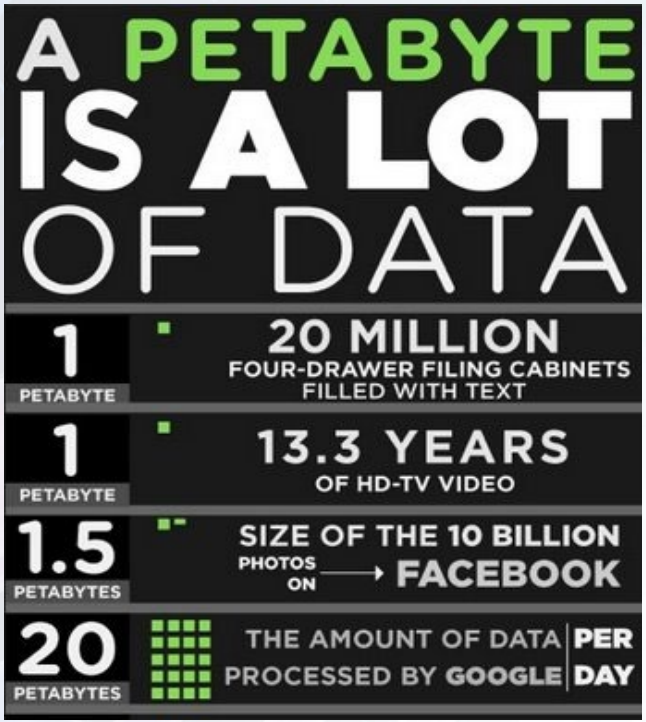
El dato es objetivo y no abstracto , sin embargo la información y el conocimiento son subjetivos y requieren altos grados de abstracción



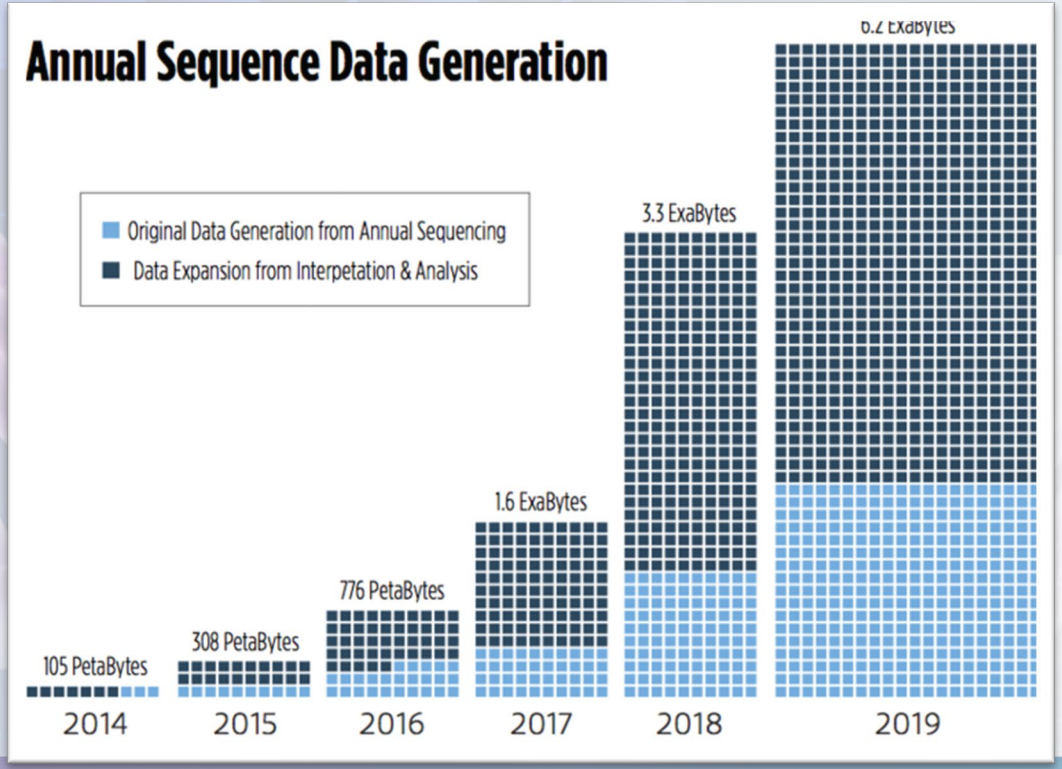
Rowley, Journal of Information Science (2007), || Abascal et al. Bioinformatica con Ñ (2014)

Medidas de almacenamiento de información





Medidas de almacenamiento de información



Tenemos suficiente información?

Algunos ejemplos de cuantos datos se generan

PLOS | BIOLOGY

PERSPECTIVE

Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

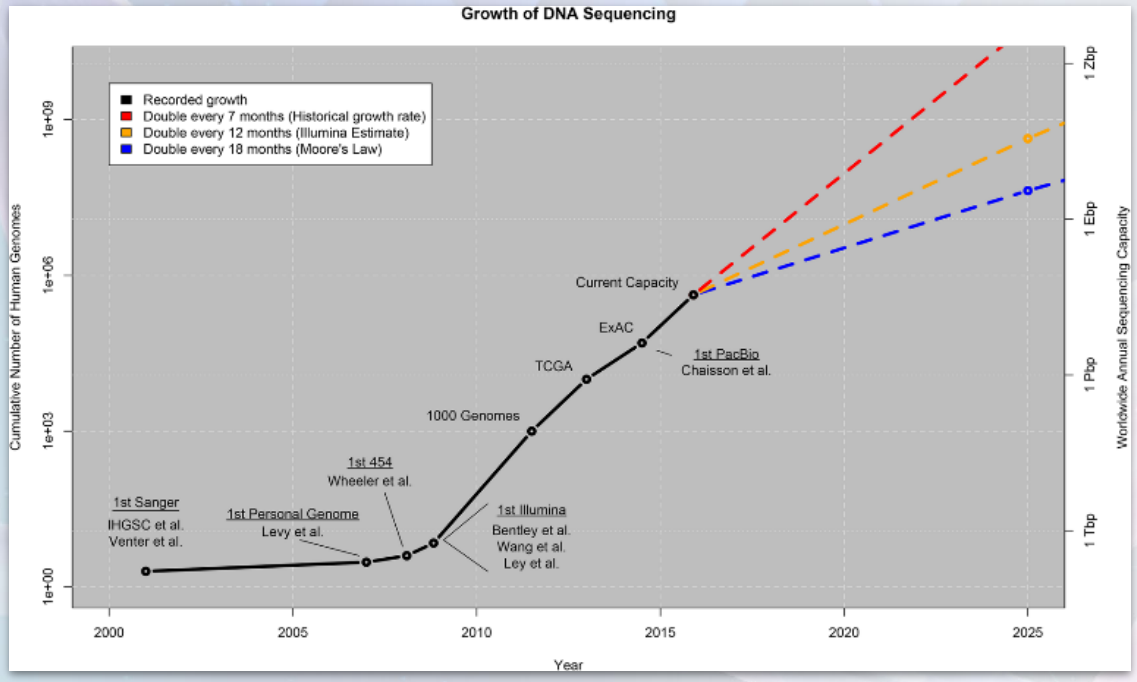
Table 1. Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Tenemos suficiente información?

Algunos ejemplos de cuantos datos se generan



"Genomics clearly poses some of the most severe computational challenges facing us in the next decade.

Genomics is a "four-headed beast"; considering the computational demands across the lifecycle of a dataset—acquisition, storage, distribution, and analysis—genomics is either on par with or the most demanding of the Big Data domains".

Al Petabyte y mas alla...

Article

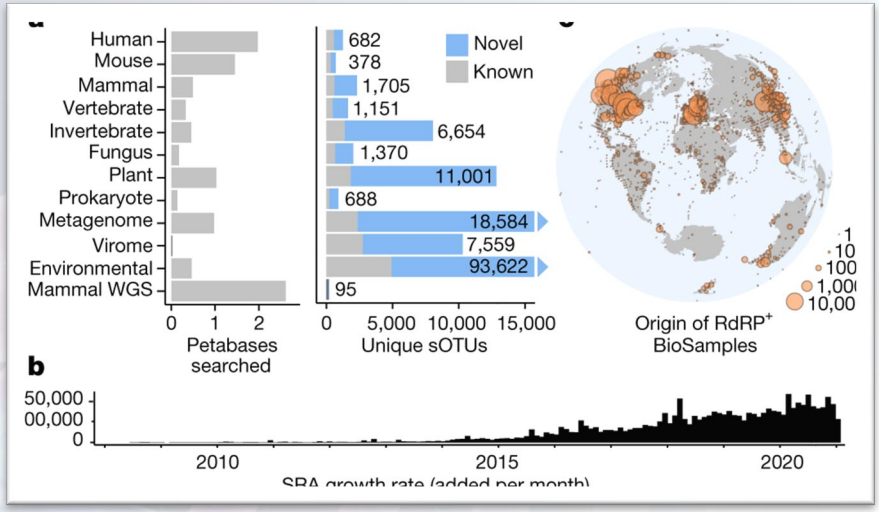
Petabase-scale sequence alignment catalyses viral discovery

https://doi.org/10.1038/s41586-021-04332-2
Received: 10 August 2020
Accepted: 10 December 2021
Published online: 26 January 2022
Check for updates

Robert C. Edgar[1,16], Brie Taylor[2,16], Victor Lin[3,16], Tomer Altman[4,16], Pierre Barbera[5,16], Dmitry Meleshko[6,7,16], Dan Lohr[8,16], Gherman Novakovsky[9,16], Benjamin Buchfink[10,16], Basem Al-Shayeb[11,16], Jillian F. Banfield[12,16], Marcos de la Peña[13,16], Anton Korobeynikov[6,14,16], Rayan Chikhi[15,16] & Artem Babaian[2,16]

Public databases contain a planetary collection of nucleic acid sequences, but their systematic exploration has been inhibited by a lack of efficient methods for searching this corpus, which (at the time of writing) exceeds 20 petabases and is growing exponentially[1]. Here we developed a cloud computing infrastructure, Serratus, to enable ultra-high-throughput sequence alignment at the petabase scale. We searched 5.7 million biologically diverse samples (10.2 petabases) for the hallmark gene RNA-dependent RNA polymerase and identified well over 10^5 novel RNA viruses, thereby expanding the number of known species by roughly an order of magnitude. We characterized novel viruses related to coronaviruses, hepatitis delta virus and huge phages, respectively, and analysed their environmental reservoirs. To catalyse the ongoing revolution of viral discovery, we established a free and comprehensive database of these data and tools. Expanding the known sequence diversity of viruses can reveal the evolutionary origins of emerging pathogens and improve pathogen surveillance for the anticipation and mitigation of future pandemics.

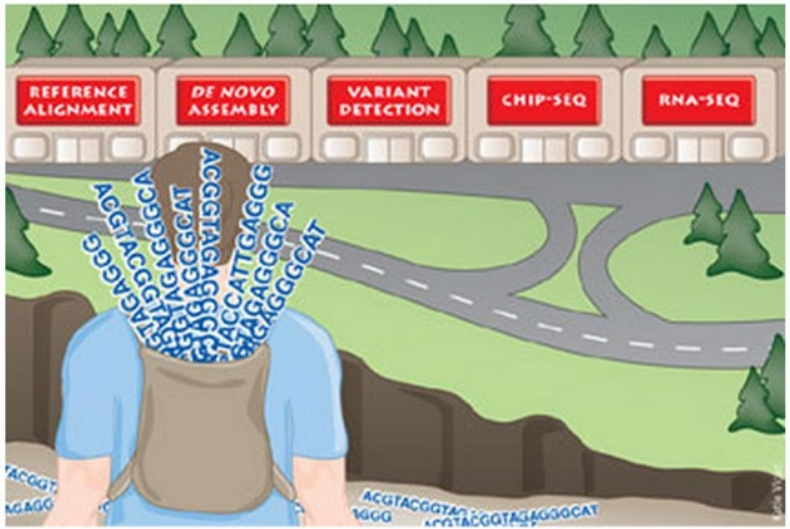
We searched 5.7 million biologically diverse samples (10.2 petabases) for the hallmark gene RNA-dependent RNA polymerase and identified well over 105 novel RNA viruses, thereby expanding the number of known species by roughly an order of magnitude..



https://serratus.io/

Como analizar data generada por NGS?

Podemos abordar estos análisis de la forma clásica?



Katie Vicari

A gap exists between current sequence-generation and data-analysis capabilities.

- *Existe una brecha creciente entre el "output" de la secuenciación masiva y la capacidad de procesar y analizar los datos (McPherson, 2009)*

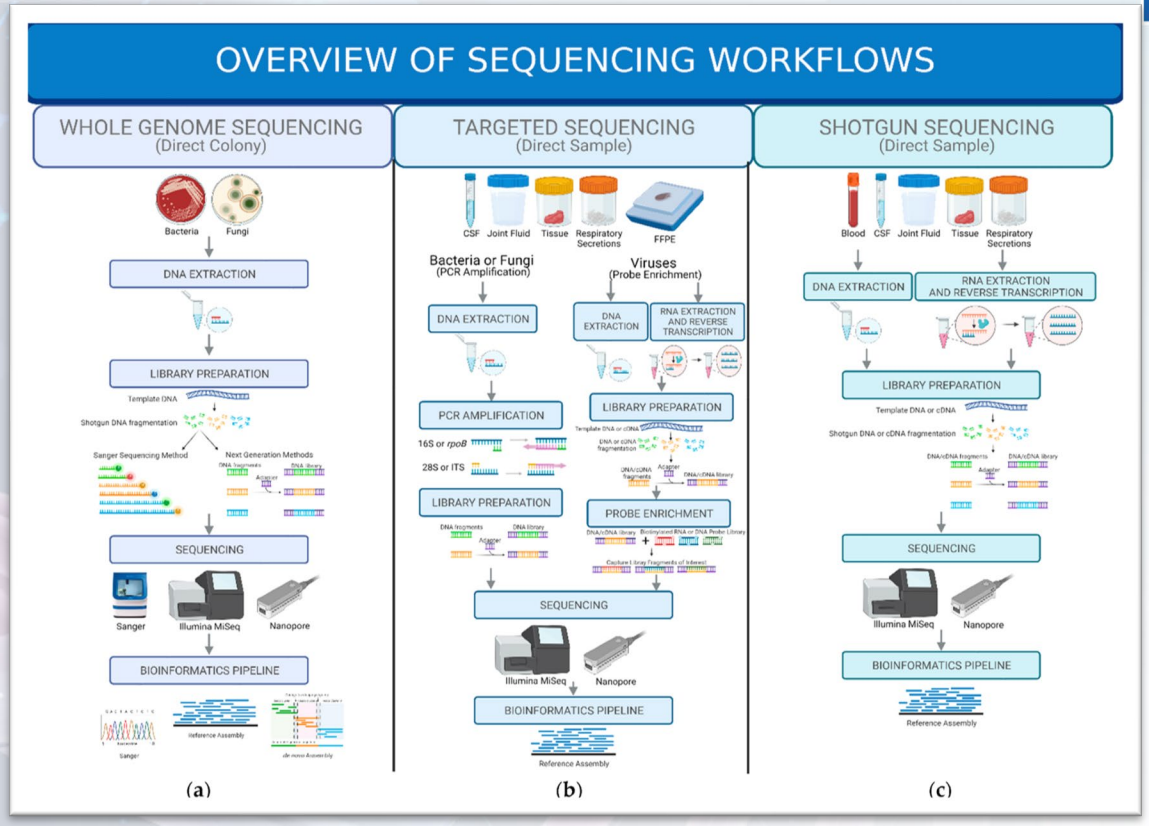
- Desconcierto en el uso de herramientas de *Base Calling,* alineación, ensamblaje y análisis, a menudo incompleta y sin idea de cómo comparar y validar sus resultados.

- US$ 1.000 del genoma vs. US$ 20.000 para el análisis.

McPherson (2009). *Next-generation gap.* Nature Methods

Flujos de trabajo

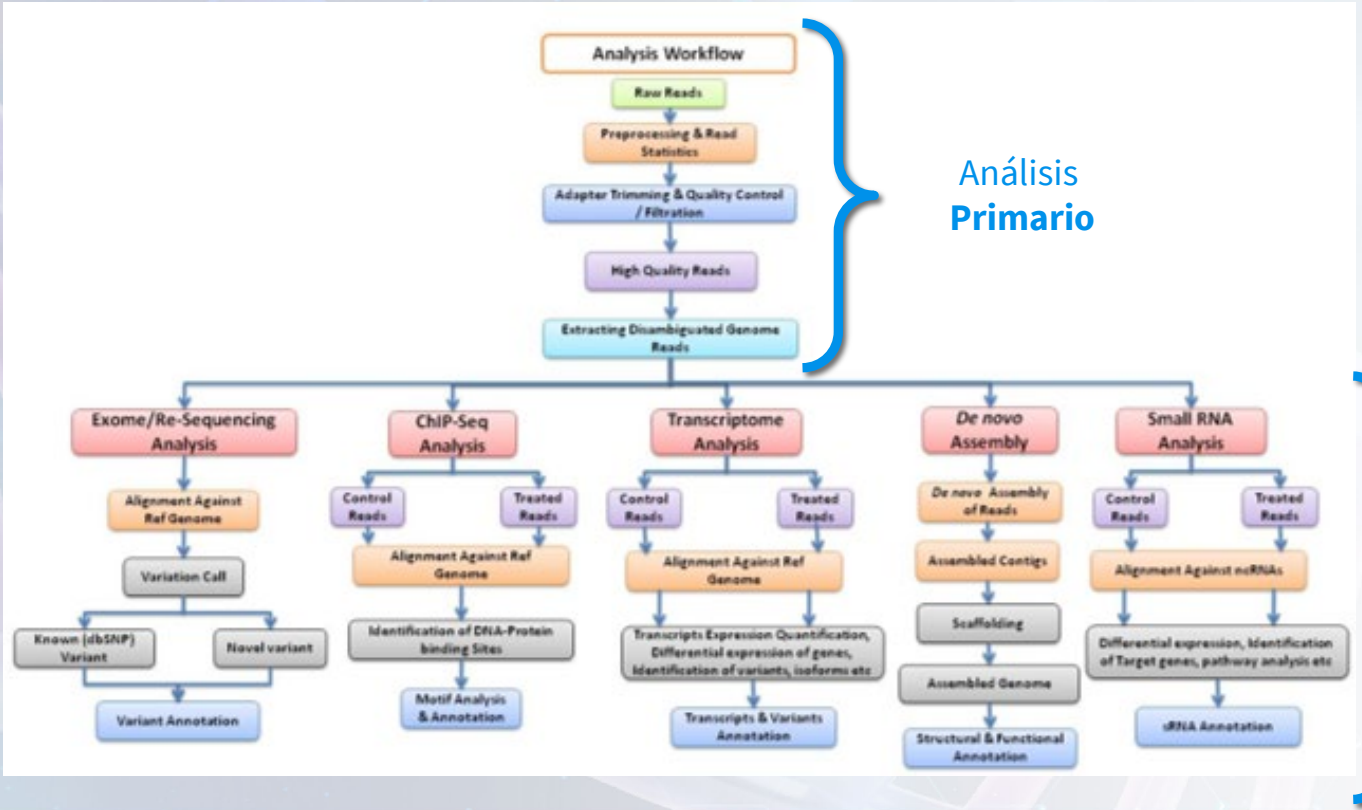
Según nuestro objetivo



Hilt, E.E.; Ferrieri, P. (2022) *NGS and infectious disease.* Gene

Como analizar data generada por NGS?

Podemos abordar estos análisis de la forma clásica?



Análisis
Primario

Análisis
Secundario y Terciario

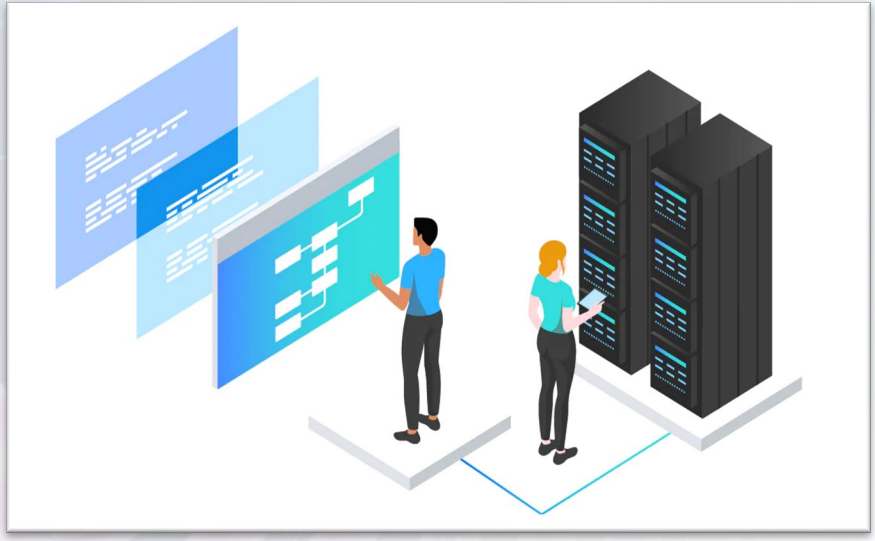
McPherson (2009). *Next-generation gap.* Nature Methods

Formatos

Almacenar secuencias y metadata

¿Por qué diferentes formatos?

- Que se almacena?
- Secuencias de ADN
- Anotaciones de genes
- Información estructural de proteínas
- Datos de alineamiento



Tipo de Información

- Texto Plano
- Formato Binario

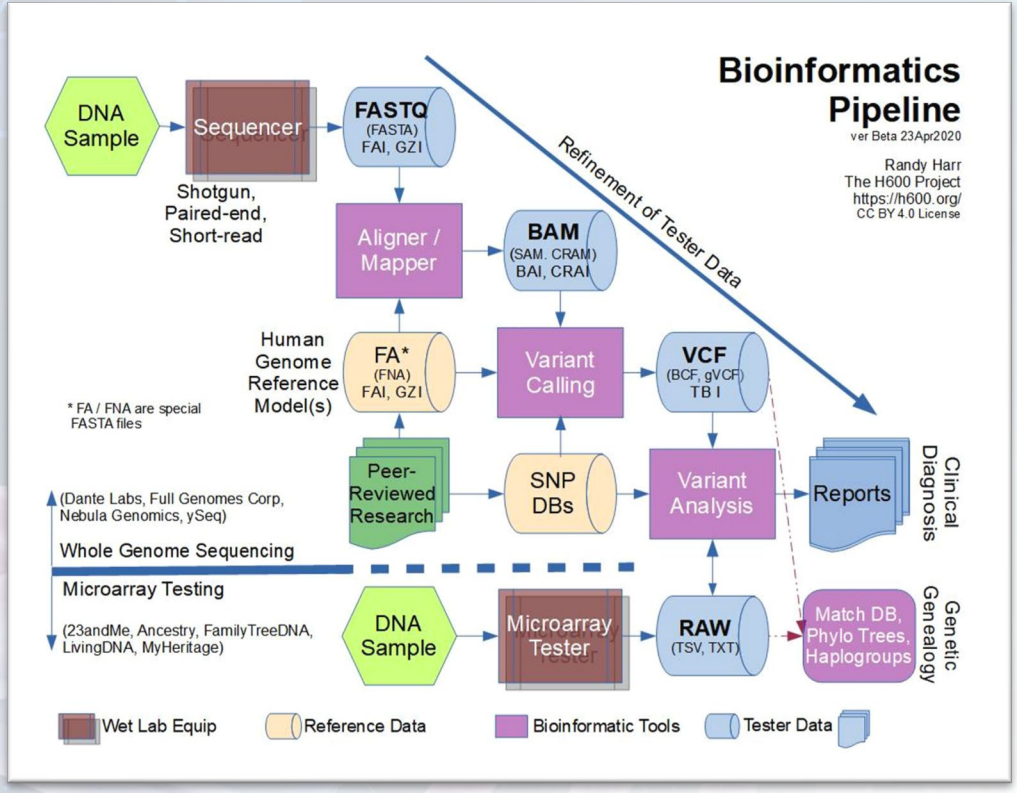
Requerimientos del Software

- Input
- Output

https://www.ensembl.org/info/website/upload/gff.html

Formatos

Almacenar secuencias y metadata



https://h600.org/wiki/Sequencing+File+Formats

Formatos

AB1

AB1: formato binario que almacena los datos de secuenciación Sanger.

- Formato de salida (“output”) del secuenciador
- Contiene información específica de la corrida electroforética
- Propiedad de la compañía Applied Biosystems (Thermo Fisher)
- Extensión del archivo: ***.ab1**

http://www6.appliedbiosystems.com/support/software_community/ABIF_File_Format.pdf

Formatos

FASTA

- No se permiten líneas en blanco dentro del archivo.
- Se recomienda que todas las líneas de texto sean menores de 80 caracteres
- Extensiones del archivo:
  - *.fasta
  - *.seq
  - *.txt
  - *.rafael

FASTA: formato plano que comienza con una sola línea descriptiva, la cual contiene un ">"

```
>C4070_940_003_B11.ab1
CCCCCCTCCCGGGAGAGCCATAGTGGTCTGCGGAACCGGTGAGTACACCGGAATTGCCAGGACGACCGGGTCCTTTCTTG
GATAAACCCTCTCAATGCCTGGAGATTTGGGCGTGCCCCCGCAAGACTGCTAGCCGAGTAGTGTTGGGTCGCGAAAGGCC
TTGTGGTACTGCCTGATAGGGTGCTCGAGAGTGAACCGGGAGGTCTCGTAGACCGTGCACCTCGAGT
```

```
>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

http://www6.appliedbiosystems.com/support/software_community/ABIF_File_Format.pdf

Formatos

BAM

Formato genérico de alineamiento para almacenar reads alineados contra genomas de referencia

- Delimitados por tabulación
- El encabezado comienza por @
- El BAM es el archivo binario del SAM

Each row describes a single alignment of a raw read against the reference genome. Each alignment has 11 mandatory fields, followed by any number of optional fields.

Nucleic Acids Research, 2010, Vol. 38, No. 61767–1771

http://support.illumina.com/help/SequencingAnalysisWorkflow/Content/Vault/Informatics/Sequencing_Analysis/CASAVA/swSEQ_mCA_FASTQFiles.htm

Formatos

FASTQ

Tiene la capacidad de almacenar un puntaje de calidad de forma numérica, asociado a cada nucleótido en una secuencia dada.

Cada archivo FASTQ contiene 4 líneas:

- Identificador de la secuencia
- Secuencia
- Línea del “quality score” (signo +)
- "Quality score” en codigo ASCII

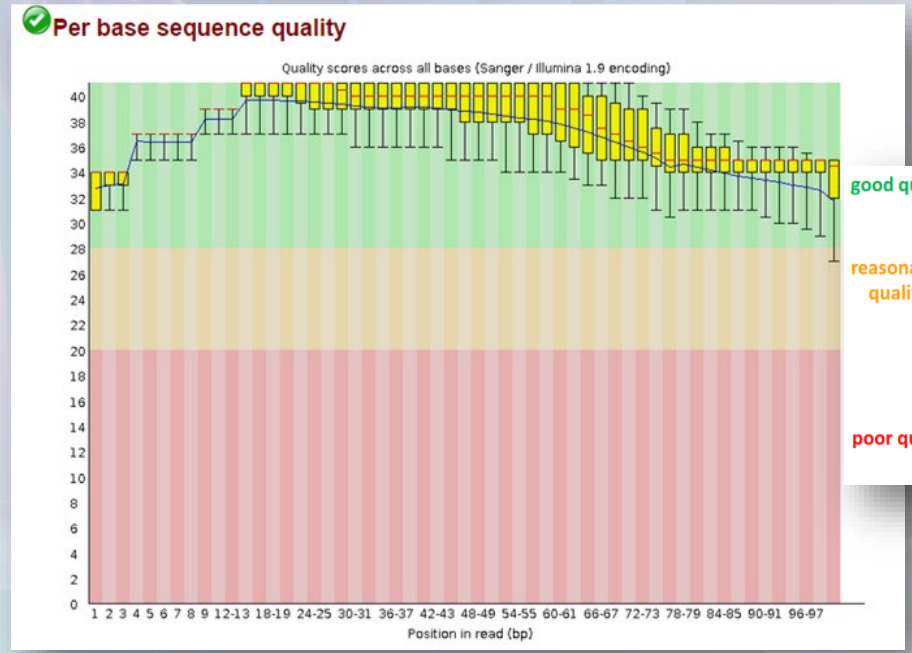
```
@M02189:35:000000000-AD0JN:1:1101:16368:1392 1:N:0:1
ATCCCCTGCTCCTGTGACATCGAATACTTCTTTTGCCACAGTGGGAATATGGTAGGTTTTTGATTCGGAAGAAATATAAAGACTCATTCCTTTTTCCCCACGAGTGATCATCATCGACGGAGAATTTAGATTTTCCGCGATCTTTTTGG
+
AA?3A3FA@FFCFFEGGGGGGGCGHGHHD5GHGHHFFEFGGGHHHEGHHGHH5DGE5FGGGGGHFFGG?FAAGHHFHHHHGG2FG3GGG@DBGHHEFFH2BE>EEHFHGBFGHFEGHCE>EEECHHHFGBGGHHHGDDDDGGEHDG2/<
@M02189:35:000000000-AD0JN:1:1101:14504:1416 1:N:0:1
GATATGTAAAAATGGCCCCCGCCCTGGTTCGGGTGGAGGAGGTGGGCTTGCGGGAGAAGCCTTCTGTCCTATATCATAATATTGCAAAATTCACAACTACAAATTCCTGGAAACTTTGTAGGAGCTCCTACAAGCATTGAACTCACTTC
+
>>A1ADDFFFBFGFGCGGEG?EE0AGEGGDEAEGAEAGC/AEE/EEAB/E>FC?/>/0BF0FGHHHHGGGHFFF22BG2GGHHHDDHG0BBBDFGHEGBGGGHHHEDBD00FD0FGHHHFBDFFGGF@@FCC?FFBGHGGFFGGDGHHH
@M02189:35:000000000-AD0JN:1:1101:13969:1443 1:N:0:1
```

Nucleic Acids Research, 2010, Vol. 38, No. 61767–1771

http://support.illumina.com/help/SequencingAnalysisWorkflow/Content/Vault/Informatics/Sequencing_Analysis/CASAVA/swSEQ_mCA_FASTQFiles.htm

Formatos

Revisando nuestro FASTQ



Basic Statistics

Measure	Value
Filename	SRR611244_1.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1805585
Sequences flagged as poor quality	0
Sequence length	101
%GC	34

- **FastQC:**
  - Revisar FASTQ
  - Evaluar de forma gráfica
  - Primer paso del análisis primario
  - Informacion:
    - Longitud *reads*
    - %GC
    - # *reads*

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Formatos

Formato	Que almacena	Tipo
FASTQ	Formato de almacenamiento de “reads”, incluye *Quality Control*	Binario
FASTA. FNA	Secuencia de nucleotidos, aminoacidos, proteínas	Texto plano
SAM/BAM	Sequence Alignment Map / binario - comprimida	Texto plano (tab) / binario
GFF/GFF3	*General Feature Format* (anotacion)	Texto plano
VCF	*Variant Call Format* (analisis de variantes)	Texto plano (tab)
FASTQ/FAST5	Contiene *raw data* que pueden utilizarse para el *base calling*	HDF5 / Estructurado

https://nanoporetech.com/nanopore-sequencing-data-analysis // https://www.ensembl.org/info/website/upload/gff.html


Just with one button, you do all analysis! Right?
Look! Do you see this red button? I especially made it for you. Everything is just done with this button, so easy!

Flujos de Trabajo en Genómica Bacteriana

Metodologías - Aplicaciones

Flujo de Trabajo

Detras del boton "rojo" existe un flujo de trabajo

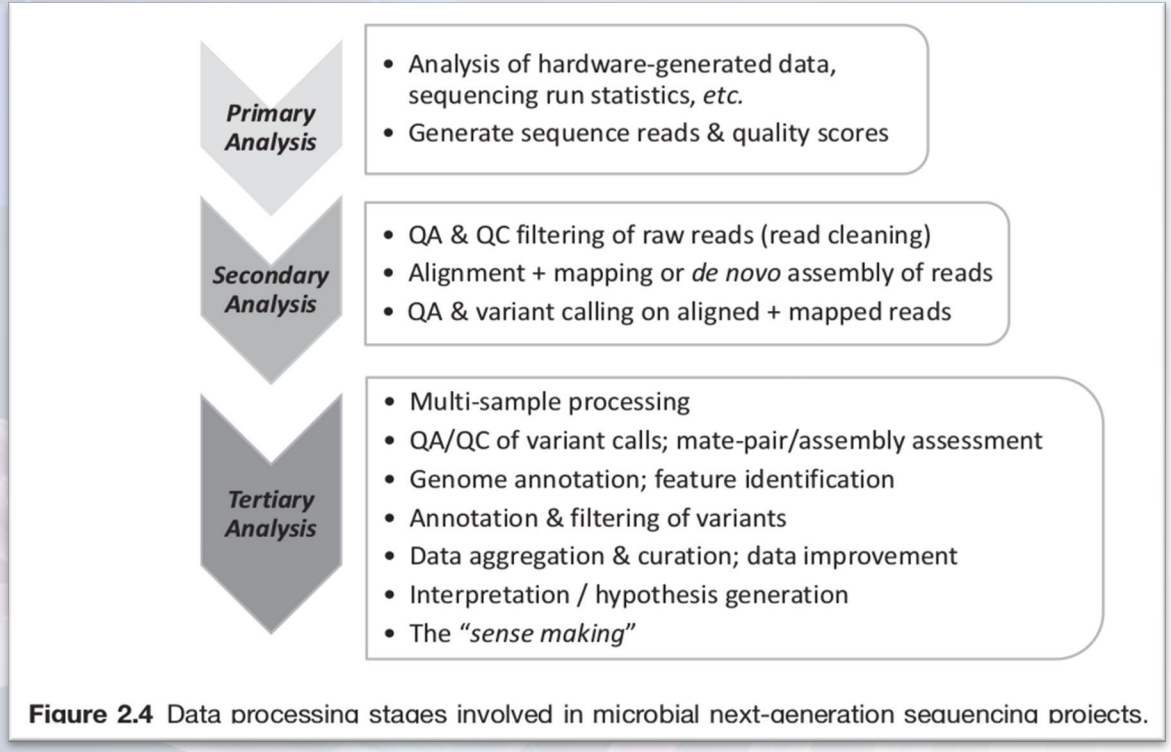
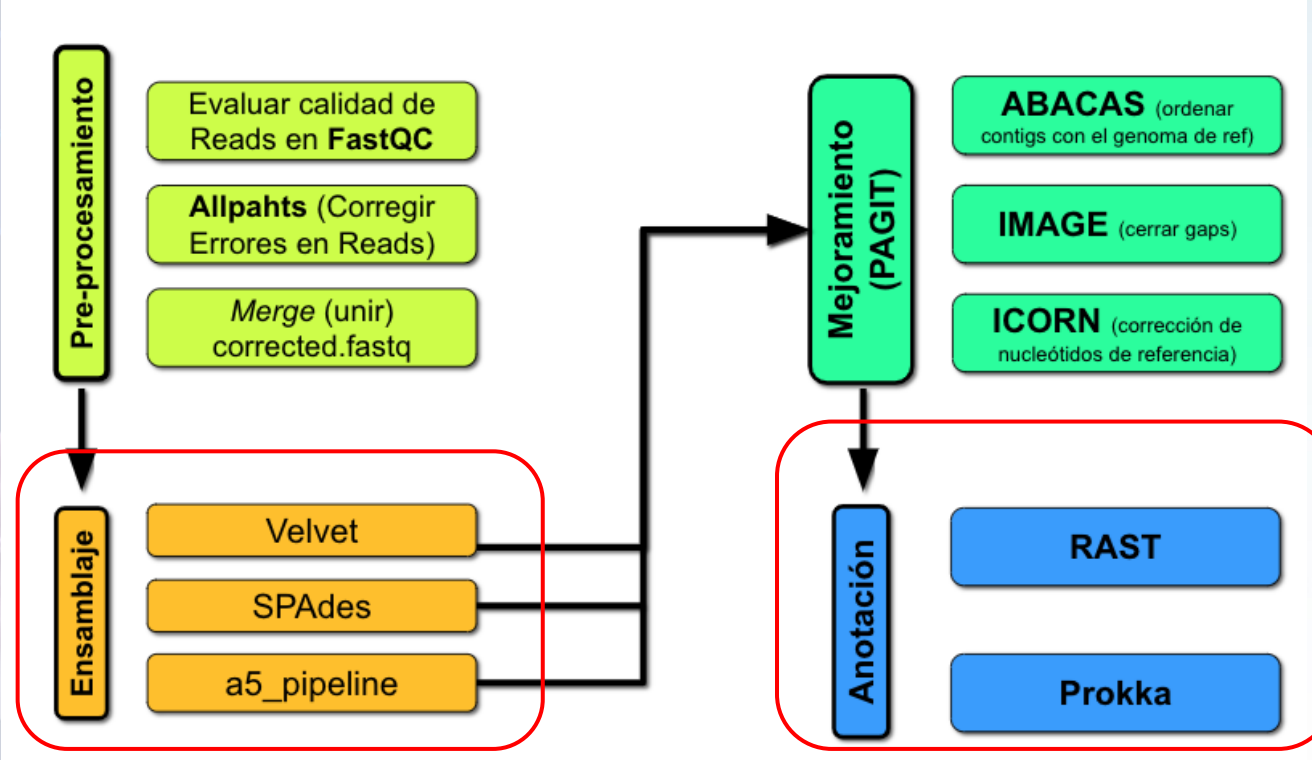


Figure 2.4 Data processing stages involved in microbial next-generation sequencing projects.

Bishop, O. (2016). Bioinformatics and Data Analysis in Microbiology

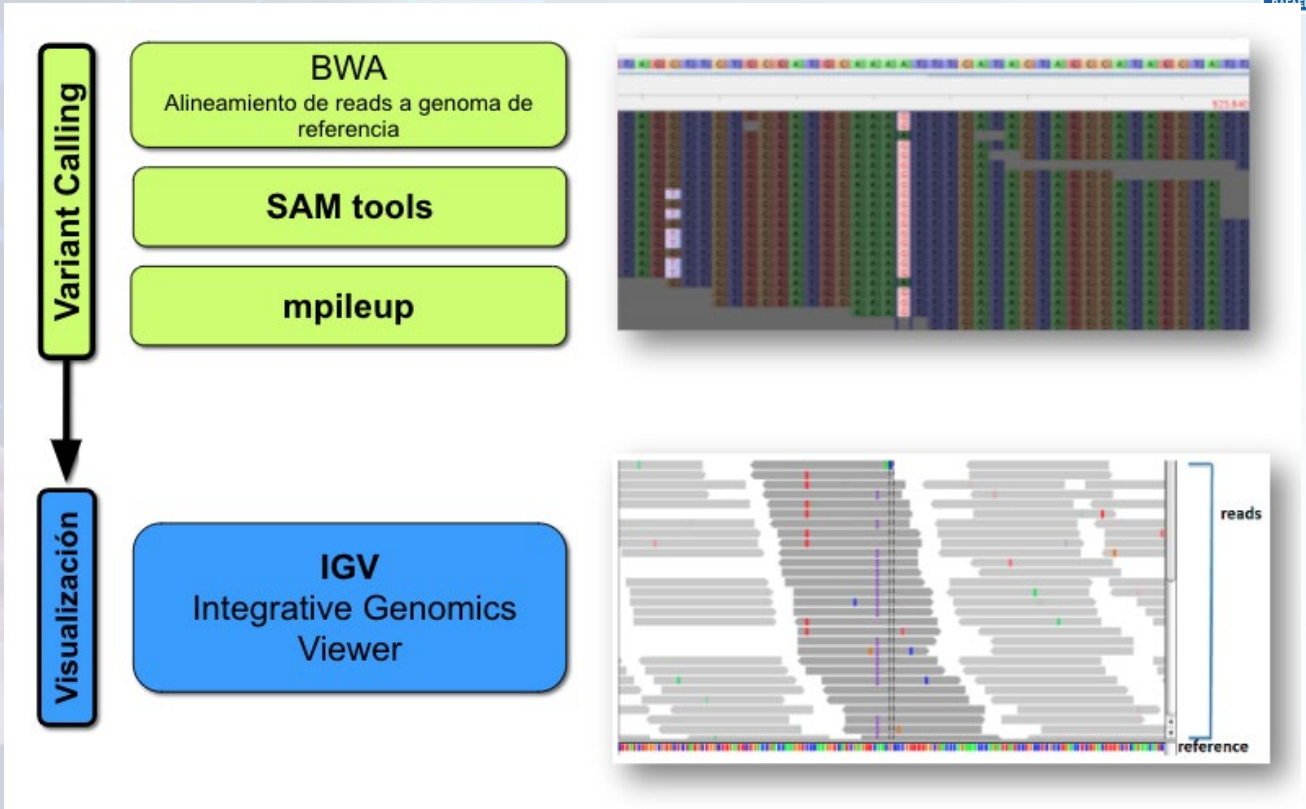
Flujo de Trabajo

Ejemplos del botón "rojo"



Flujo de Trabajo

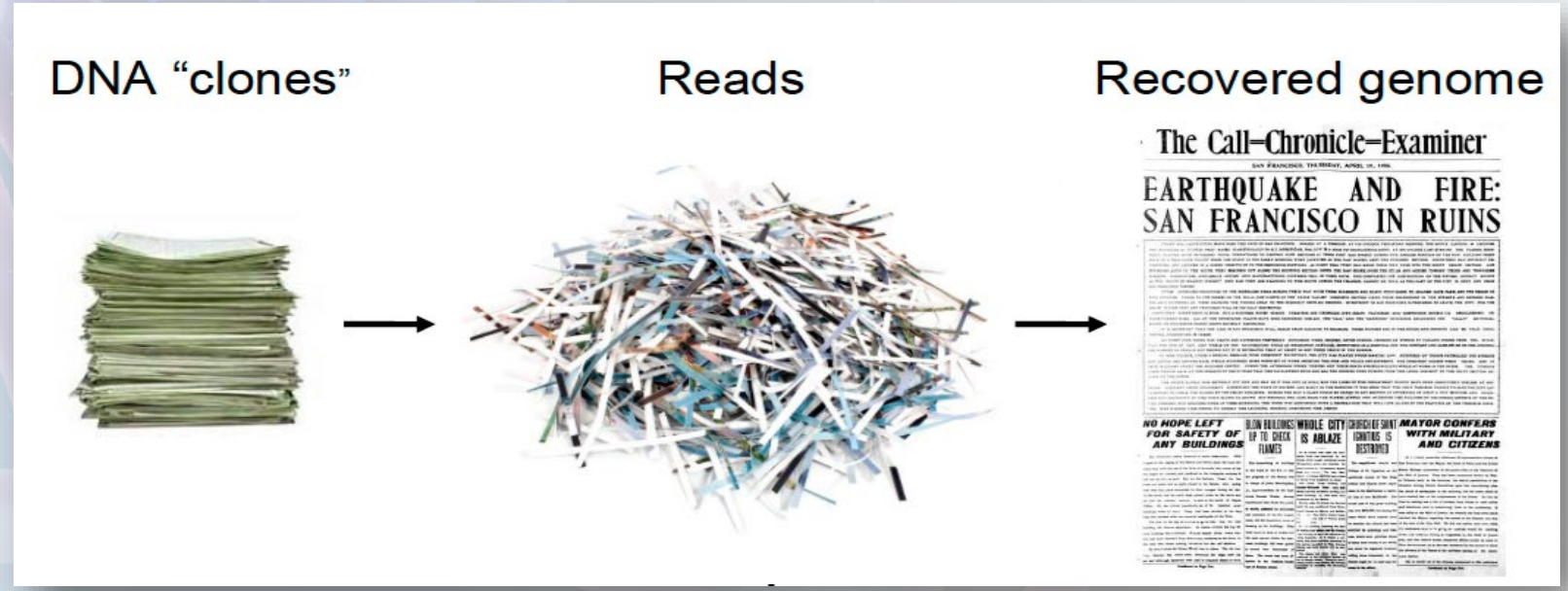
Ejemplos del botón "rojo"



// https://www.ensembl.org/info/website/upload/gff.html

Ensamblaje

Proceso de **reconstrucción una secuencia** de ADN "original" de un organismo a partir de secuencias cortas *(reads)*

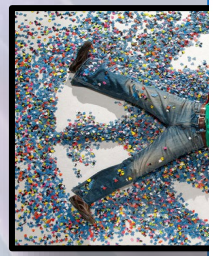


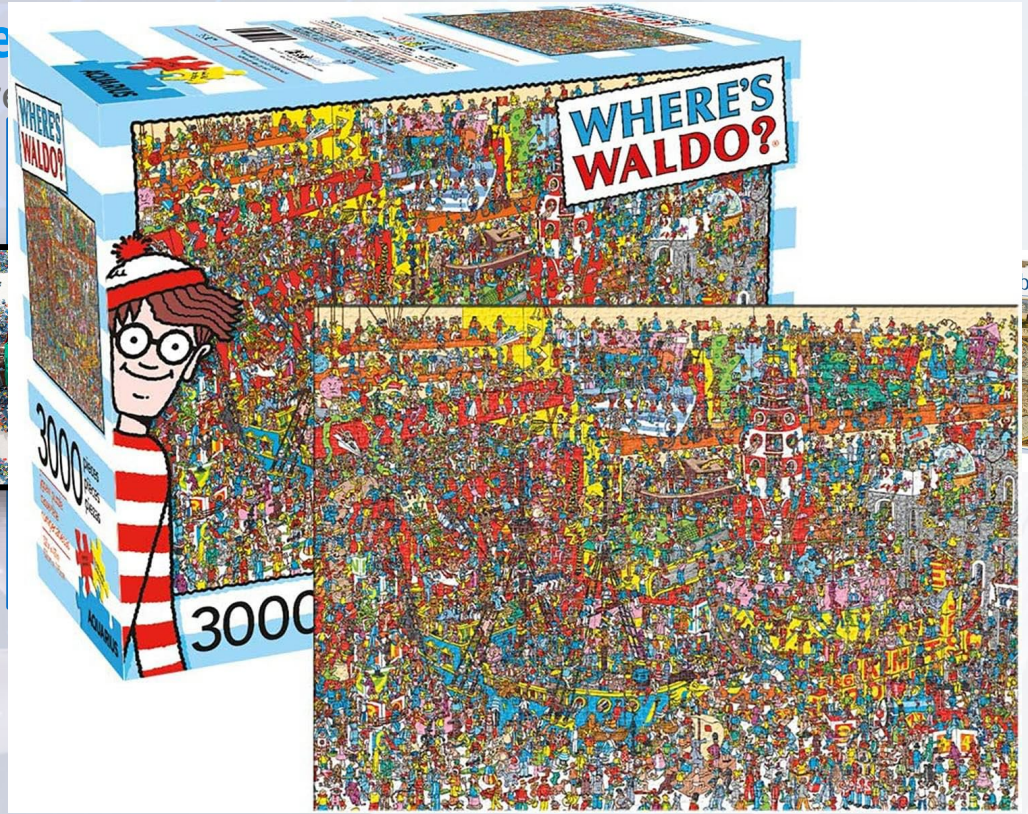
Seemann, (2010). *Genome Assembly Strategies.*

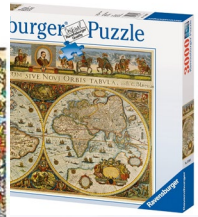
Estrategias de

De novo y con refer

4,5 MB



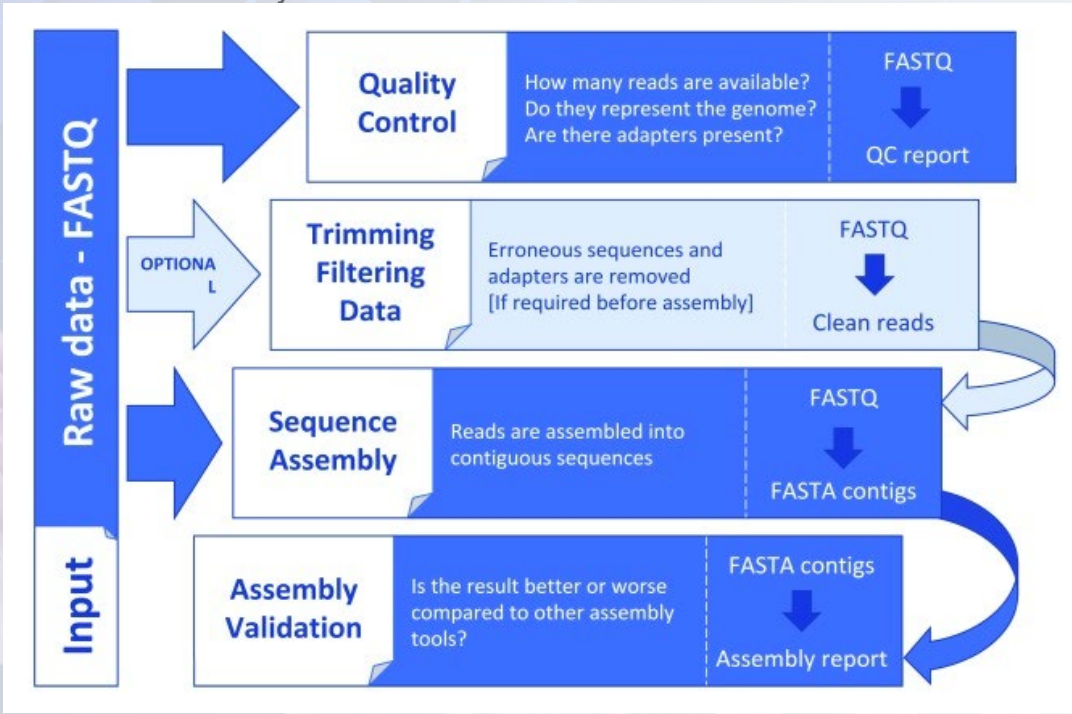




Seemann, (2010). *Genome Assembly Strategies.*

Ensamblaje

Workflow para realizar un "buen" ensamblaje



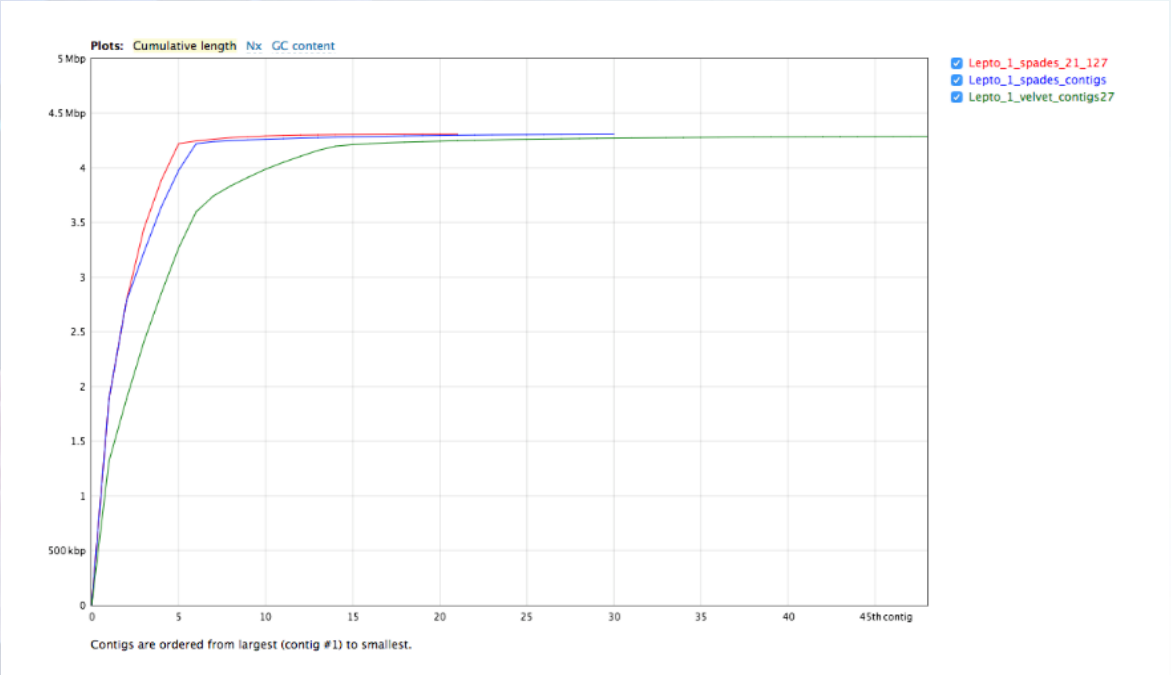
https://f1000research.com/articles/7-148/v1

Comparativa entre Ensambladores

Ensamblaje de Leptospira - Spades *vs* Velvet

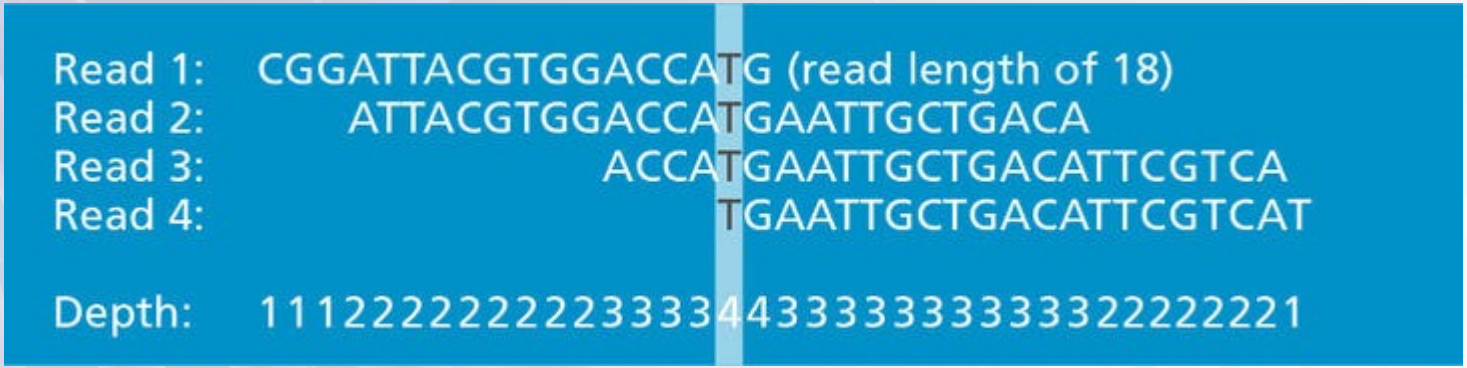
QUAST

Quality Assessment Tool for Genome Assemblies by CAB



Cobertura - *Coverage*

El número de lecturas únicas que contienen un nucleótido determinado en la secuencia reconstruida.

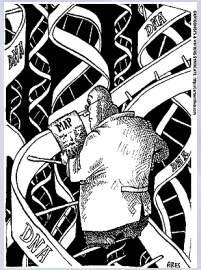


$$\text{Cobertura} = \frac{\#\text{ de reads} * \text{longitud del read}}{\text{tamaño estimado del genoma}}$$

Wee *et al.* (2019). *The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing.* Briefings in Functional Genomics

Anotacion

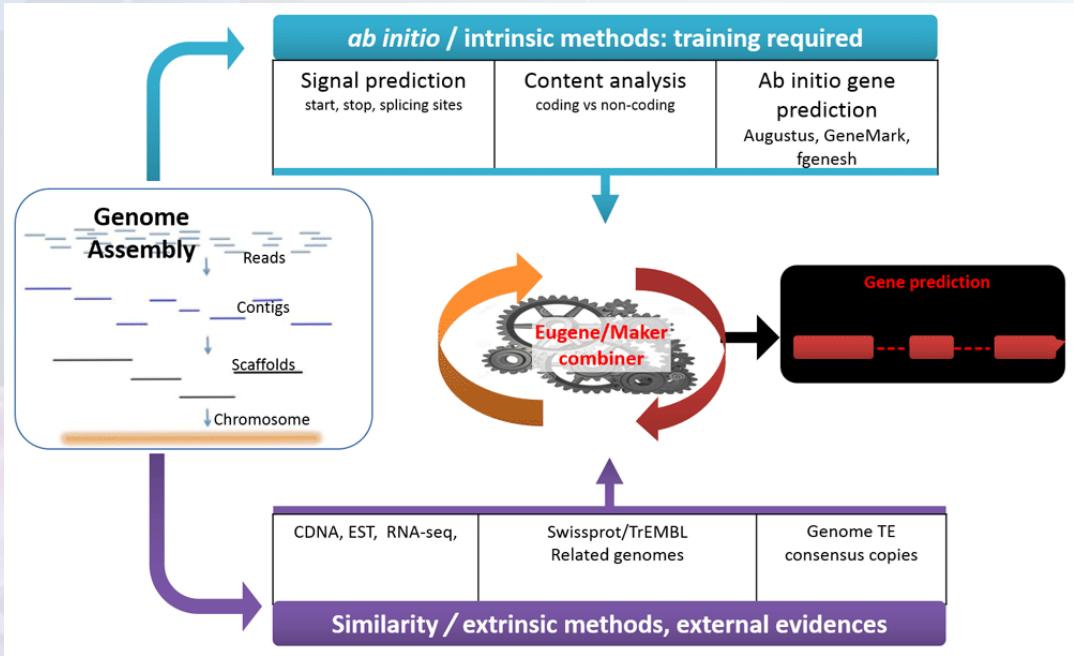
Proceso de identificación de características de interés en un genoma.



Incluye:

- Genes
- RNA no codificantes
- Operones
- Nuevos genes?

La precisión en la anotación genómica es importante y, en algunas veces crítica, en la interpretación biológica *downstream*



Dominguez Del Angel V. *et al. Ten steps to get started in Genome Assembly and Annotation.* F1000

Flujo de trabajo "perfecto"?

Estrategia de Ensamblaje

PLOS COMPUTATIONAL BIOLOGY

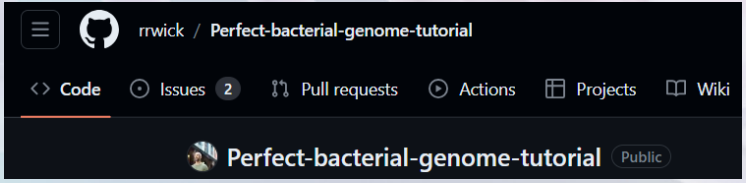
EDUCATION

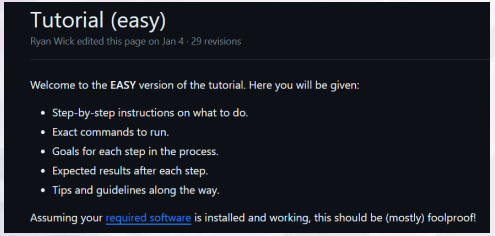
Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing

Ryan R. Wick[1]*, **Louise M. Judd**[2], **Kathryn E. Holt**[1,3]

1 Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Australia, **2** Department of Microbiology and Immunology, University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia, **3** Department of Infection Biology, London School of Hygiene & Tropical Medicine, London, United Kingdom

Published: March 2, 2023





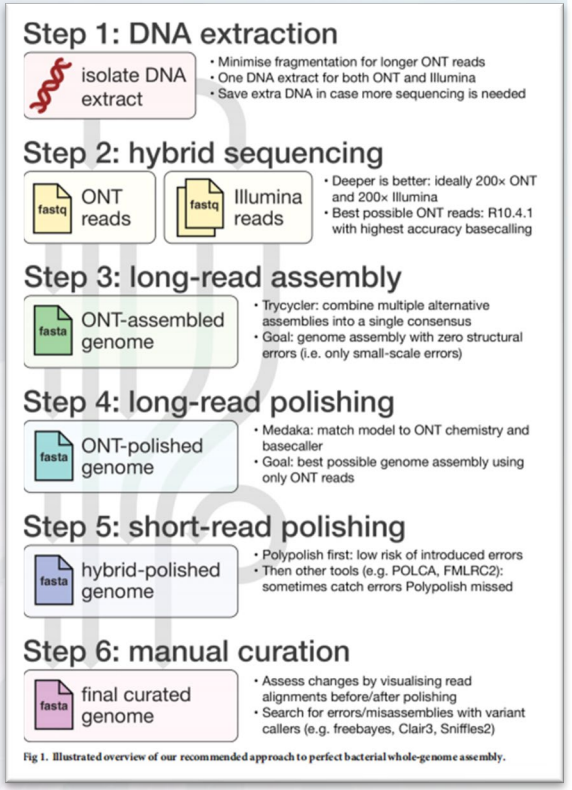


Fig 1. Illustrated overview of our recommended approach to perfect bacterial whole-genome assembly.

https://doi.org/10.1371/journal.pcbi.1010905

Flujo de trabajo "perfecto"?

Estrategia de Ensamblaje

nature communications

Article

https://doi.org/10.1038/s41467-022-35713-4

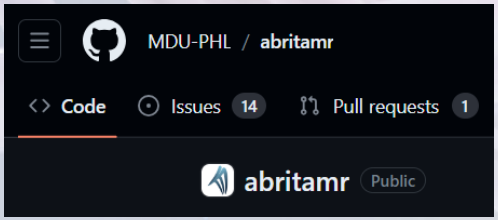
An ISO-certified genomics workflow for identification and surveillance of antimicrobial resistance

Received: 1 June 2022

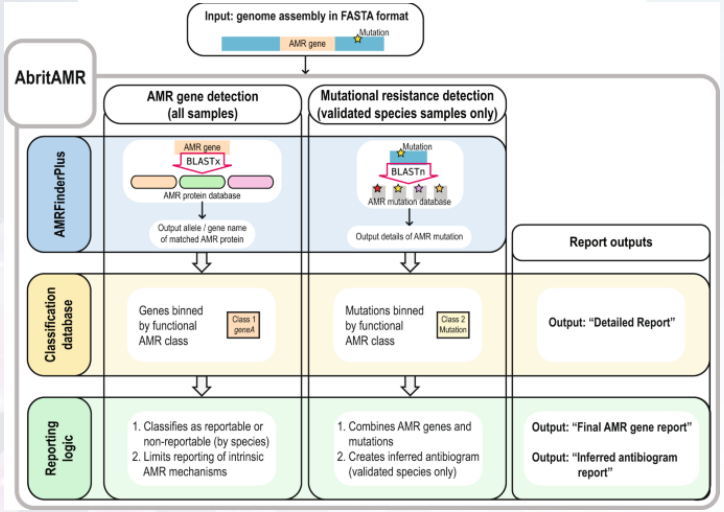
Accepted: 21 December 2022

Published online: 04 January 2023

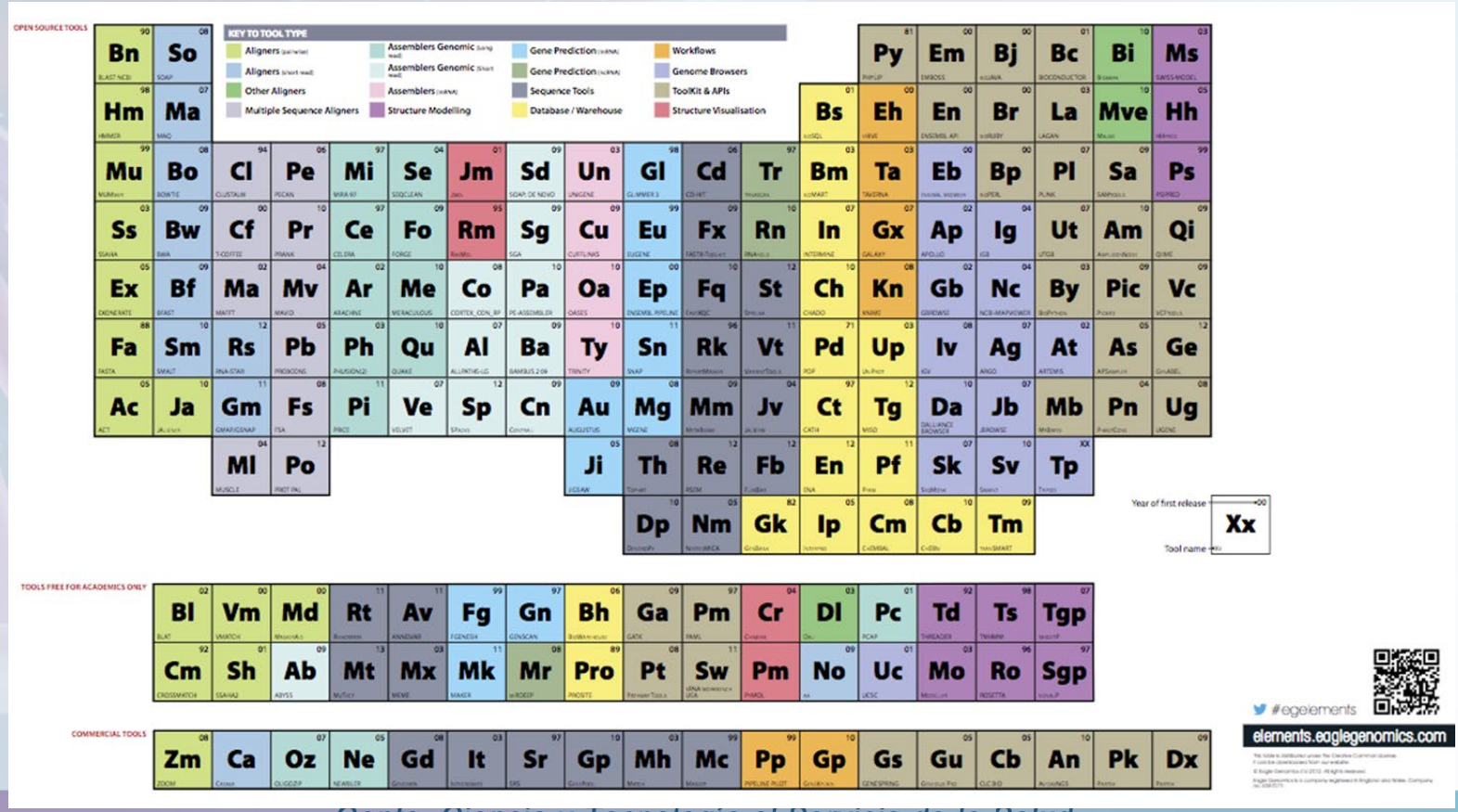
Norelle L. Sherry[1,2,3], Kristy A. Horan[1], Susan A. Ballard[1], Anders Gonçalves da Silva[1], Claire L. Gorrie[3], Mark B. Schultz[1], Kerrie Stevens[1], Mary Valcanis[1], Michelle L. Sait[1], Timothy P. Stinear[3], Benjamin P. Howden[1,2,3,4] & Torsten Seemann[1,3,4]



abriTAMR



https://doi.org/10.1038/s41467-022-35713-4 / https://github.com/MDU-PHL/abritAMR


OPEN SOURCE TOOLS
KEY TO TOOL TYPE
Aligners
Aligners
Other Aligners
Multiple Sequence Aligners
Assemblers Genomic
Assemblers Genomic
Assemblers
Structure Modelling
Gene Prediction
Gene Prediction
Sequence Tools
Database / Warehouse
Workflows
Genome Browsers
ToolKit & APIs
Structure Visualisation
Year of first release
Xx
Tool name
TOOLS FREE FOR ACADEMICS ONLY
COMMERCIAL TOOLS
#egeiements
elements.eaglegenomics.com

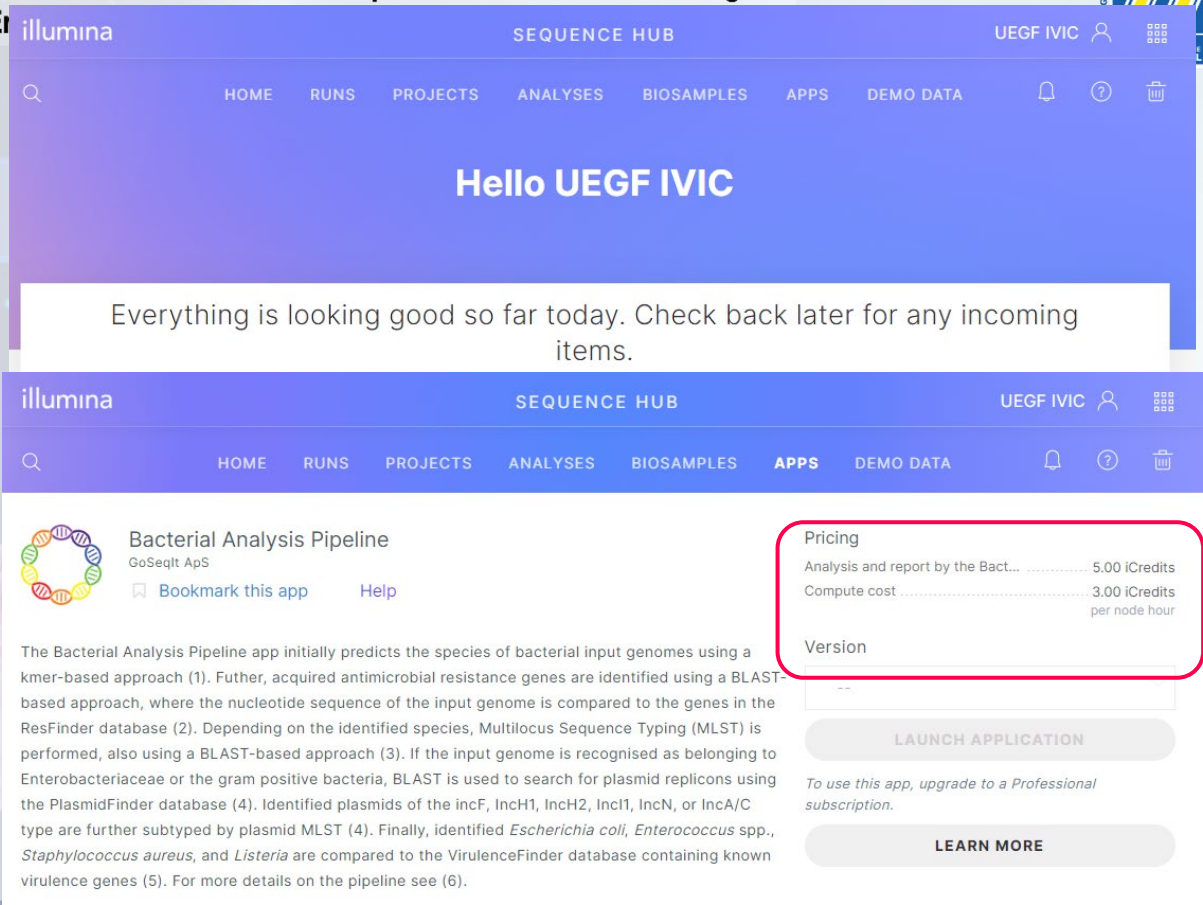
Software multipropósito con interfaz gráfica

Como realizar los diferentes tipos de análisis

Software	Tipo	Aplicacion	Sistema Operativo	Link
BaseSpace	Cloud Computing	NGS	Web	https://basespace.illumina.com
UGENE	Análisis Primario y Secundario	Sanger / “Toolkit”(+NGS)	Linux, Windows y Mac OSX	http://ugene.net
MEGA 11	Análisis secundario	Sanger / Detección mutaciones / Filogenia	Linux, Windows y Mac OSX	https://www.megasoftware.net/
Galaxy	Cloud Computing	NGS	Web	https://usegalaxy.org/

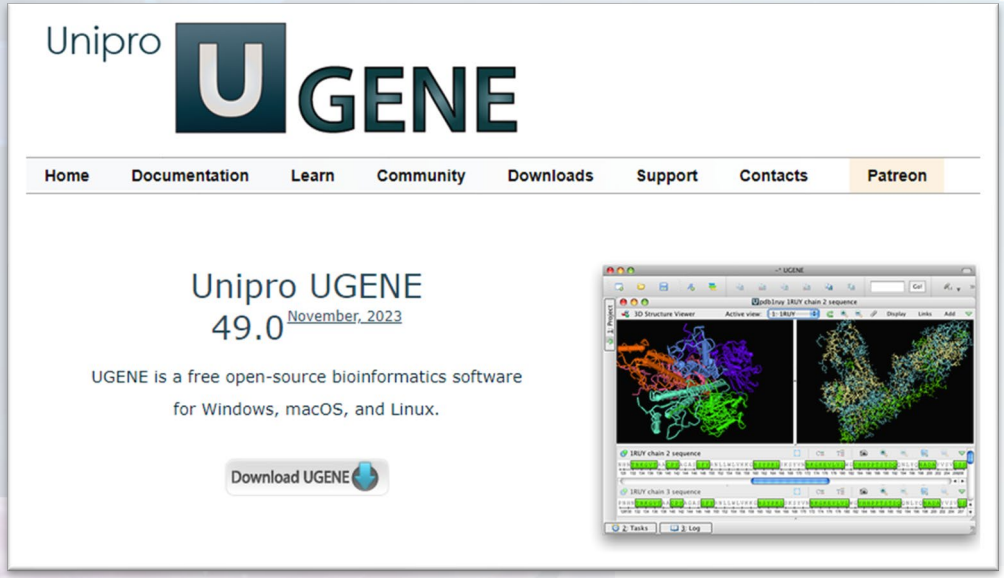
BaseSpace - illumina

- Plataforma informática que integra secuenciador y workflows
- Datos de corridas, almacenamiento y análisis de datos FASTQ
- Compartir data con colegas
- Análisis en forma de “apps” (aplicaciones)
- Limitaciones: capacidad de análisis en cuenta gratis, seguridad?, control del análisis



UGENE - Unipro

- Software Libre Multiplataforma
- Revisión, anotación y edición de ADN y proteínas
- Alineamientos Múltiples de secuencias
- Diseño de primers
- Metrica básica
- Manejo de datos de NGS (illumina)



Cite Us

Okonechnikov K, Golosova O, Fursov M, the UGENE team. **Unipro UGENE: a unified bioinformatics toolkit** . *Bioinformatics* 2012 28: 1166-1167. doi:10.1093/bioinformatics/bts091

Golosova O, Henderson R, Vaskin Y, Gabrielian A, Grekhov G, Nagarajan V, Oler AJ, Quiñones M, Hurt D, Fursov M, Huyen Y. **Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses**. *PeerJ 2014* 2:e644. doi:10.7717/peerj.644

Rose R, Golosova O, Sukhomlinov D, Tiunov A, Prosperi M. **Flexible design of multiple metagenomics classification pipelines with UGENE.** *Bioinformatics*, bty901, 2018/10/25. doi:10.1093/bioinformatics/bty901

Galaxy

- *Web based*
- Accesibilidad
- Reproducibilidad
- Colaborativo
- Comunidad de entrenamiento - GTN
- Manejo de datos de NGS

Galaxy PROJECT

Es una plataforma científica de *workflow* para análisis de datos que tiene como objetivo hacer que la biología computacional sea más accesible para los científicos sin experiencia en programación de computadoras

The Galaxy Community, (2022), The Galaxy platform for accessible, reproducible and collaborative biomedical analyses. NAR / https://usegalaxy.org/

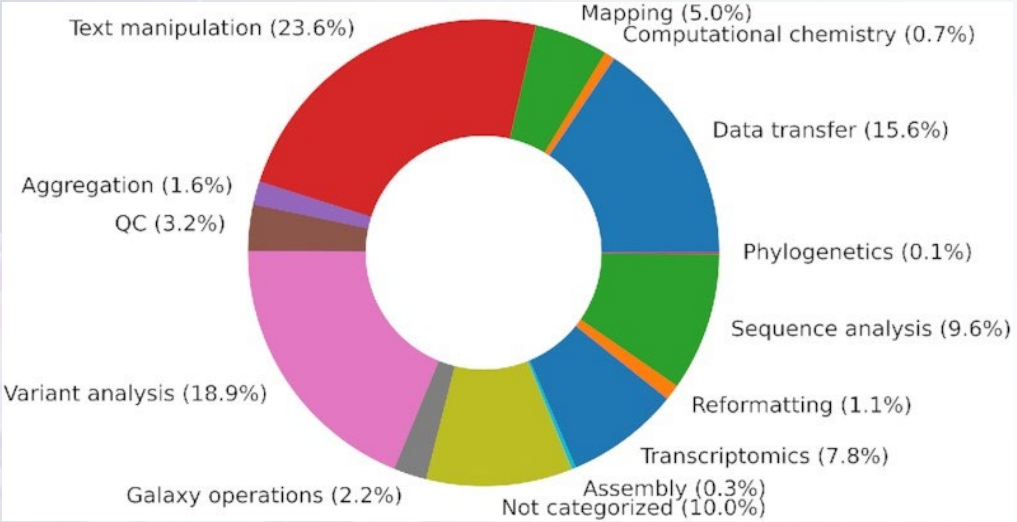
Nucleic Acids Research, 2022 1
https://doi.org/10.1093/nar/gkac247

The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update

The Galaxy Community*

Received February 03, 2022; Revised March 17, 2022; Editorial Decision March 28, 2022; Accepted March 30, 2022

Categorización del tipo de herramientas ejecutadas por los usuarios en los tres servidores de usegalaxy más populares.
(actualización 21 de Abril 2022)



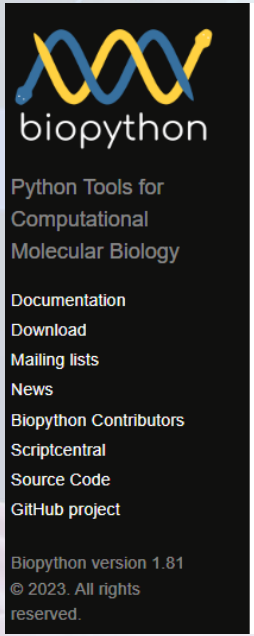
The Galaxy Community, (2022), The Galaxy platform for accessible, reproducible and collaborative biomedical analyses. NAR / https://usegalaxy.org/

Biopython - v 1.79

Línea de comandos

Conjunto de **aplicaciones y programas escritos en Python** para análisis de datos biológicos, programados por una comunidad internacional.

- Manipulación de secuencias de ADN, AA, Proteinas
- Cálculo de estructuras proteicas,
- Genética de poblaciones
- Filogenia e inteligencia artificial



Edit this page on GitHub

Biopython

See also our News feed and Twitter.

Introduction

Biopython is a set of freely available tools for biological computation written in Python by an international team of developers.

It is a distributed collaborative effort to develop Python libraries and applications which address the needs of current and future work in bioinformatics. The source code is made available under the Biopython License, which is extremely liberal and compatible with almost every license in the world.

We are a member project of the Open Bioinformatics Foundation (OBF), who take care of our domain name and hosting for our mailing list etc. The OBF used to host our development repository, issue tracker and website but these are now on GitHub.

Examining your GFF file

Since GFF is a very general format, it is extremely useful to start by getting a sense of the type of data in the file and how it is structured. `GFFExaminer` provides an interface to examine and query the file. To examine relationships between features, examine a dictionary mapping parent to child features:

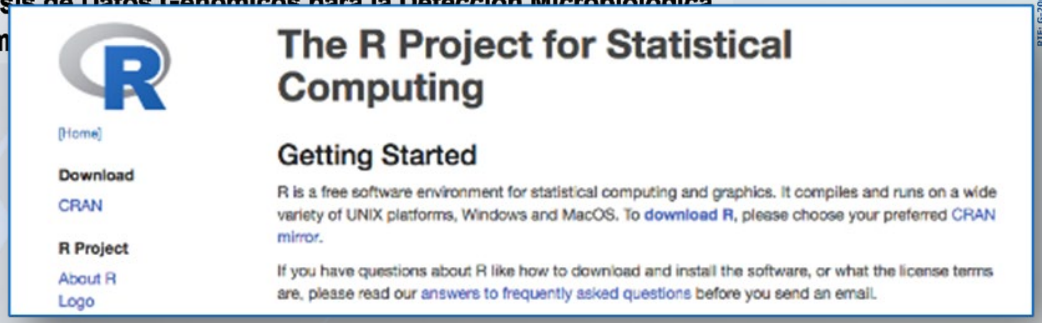
```
import pprint
from BCBio.GFF import GFFExaminer

in_file = "your_file.gff"
examiner = GFFExaminer()
in_handle = open(in_file)
pprint.pprint(examiner.parent_child_map(in_handle))
in_handle.close()
```

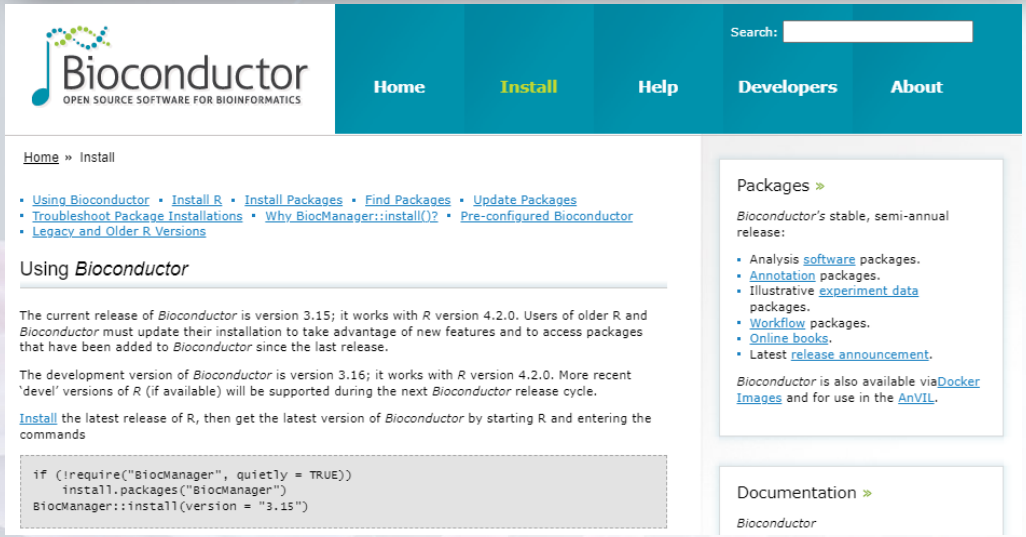
R

Línea de comandos

Ambiente (lenguaje) para computación estadística y gráficos

- Manipulación de datos,
- Cálculo estadísticos y matemáticos
- Visualización gráfica, (base, ggplot, ggtree)
- **Bioconductor** software libre de código abierto que facilita el análisis riguroso y reproducible de los datos de los ensayos biológicos actuales y emergentes.

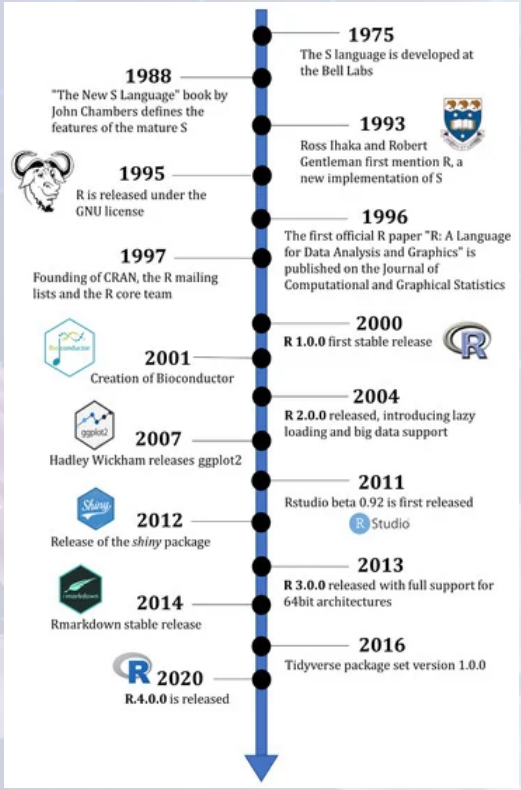




https://www.r-project.org/ // https://www.bioconductor.org/

R

Línea de comandos



Using *Bioconductor*

The current release of *Bioconductor* is version 3.15; it works with *R* version 4.2.0. Users of older R and *Bioconductor* must update their installation to take advantage of new features and to access packages that have been added to *Bioconductor* since the last release.

The development version of *Bioconductor* is version 3.16; it works with *R* version 4.2.0. More recent 'devel' versions of *R* (if available) will be supported during the next *Bioconductor* release cycle.

Install the latest release of R, then get the latest version of *Bioconductor* by starting R and entering the commands

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install(version = "3.15")
```

Courses & Conferences

Bioconductor provides training in computational and statistical methods for the analysis of genomic data. You are welcome to use material from previous courses. However, you may not include these in separately published works (articles, books, websites). When using all or parts of the Bioconductor course materials (slides, vignettes, scripts) please cite the authors and refer your audience to the Bioconductor website.

Upcoming events are advertised 6 to 8 weeks in advance.

Show 25 entries Search:

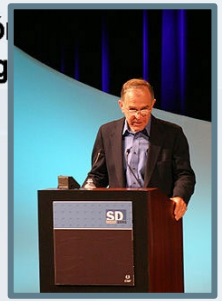
Keyword	Title	Course	Materials	Date	Bioc/R Version
Talk	Visualization and analysis of highly multiplexed imaging data, Nils Eling	SpatialSeminar	slides, video	2022-03-28	3.15/4.2
Talk	Analysing Spatially Resolved Transcriptomics Data with Bioconductor, Helena Crowell	SpatialSeminar	slides, video	2022-01-31	3.15/4.2
Talk	Spatial Transcriptomics Technologies and Analysis Tools, Dario Righelli	SpatialSeminar	slides, video	2022-01-31	3.15/4.2
Talk	The R-universe Build Infrastructure, Jeroen Ooms	BiocDevelForum	slides, video	2021-11-18	3.15/4.2
Talk	Translating R package documentation, Martin Morgan	BiocDevelForum	slides, video	2021-06-17	3.14/4.1
Talk	HCA / Bioconductor Seed Network Symposium, Various	HCA	video	2021-06-15	3.13/4.1
Workshop	Week 7: Participant stories, Various	AnVILpopup	material, video	2021-06-14	3.13/4.1
Workshop	Week 6: Reproducible research with AnVILPublish, Martin Morgan	AnVILpopup	material, video	2021-06-07	3.13/4.1
Talk	Discussion of changes in R-4.1, Martin Morgan, Mike Smith	BiocDevelForum	video	2021-05-27	3.13/4.1
Workshop	Week 5: Using AnVIL for teaching R / Bioconductor, Levi Waldron	AnVILpopup	material, video	2021-05-24	3.13/4.1

https://www.tidyverse.org/ // https://www.r-project.org/ // https://www.bioconductor.org/

Terminal - bash - shell

La caja negra

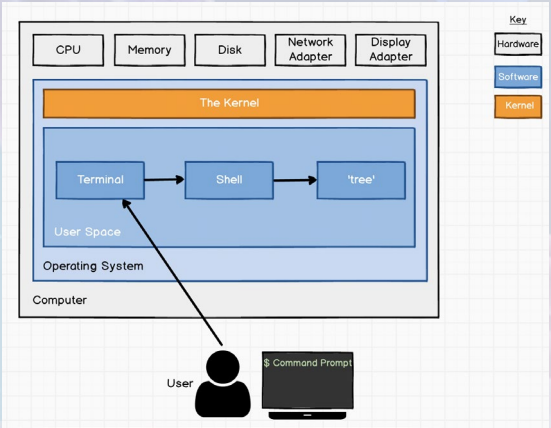
Popular interfaz de usuario de línea de comandos



Stephen Bourne
(1977)

Brian Fox
(1989)

Bourne-again shell




```
rpuche@Lenovo-G40:~$ ssh kabre.cenat.ac.cr
rpuche@kabre.cenat.ac.cr's password:
Last login: Sun Jun  5 08:21:25 2022 from meta2.cnca
================================================
Welcome to Kabré Supercomputer!
Costa Rica National High Technology Center (CeNAT)
Contact and inquiries: cluster@cenat.ac.cr
================================================

(base) [rpuche@login-3 ~]$ cd rpuche/
(base) [rpuche@login-3 rpuche]$ ls
Assemblies                                ST_11_Raw_data_knight_list       repair.RP_Oct_21.slurm
Bakta_Nov_21.sh                           Shovill_Thai_Nov_21.sh           slurm-51012.out
ChangeNameContigs_unicycler_Batchs.sh     Unicycler.RP_Dic_21.slurm        slurm-51015.out
ChangeNamesAssembly_unicycler_Batch_6.sh  Unicycler.RP_Ene_22.slurm        slurm-51177.out
ChangeNamesUnicycler.sh                   Unicycler.RP_Nov_21_Mar.slurm    slurm-51179.out
ChangeNamesdedupe.sh                      Unicycler.RP_Oct_21.slurm        slurm-51180.out
```

https://es.wikipedia.org/wiki/Bash

Terminal - bash - shell

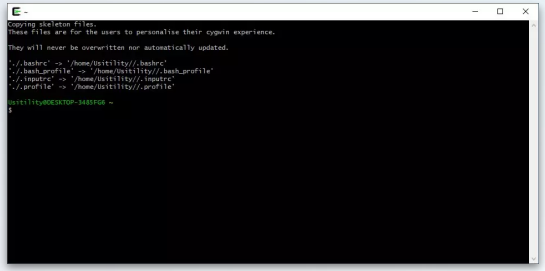
La caja negra

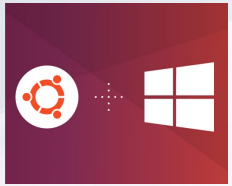
Popular interfaz de usuario de línea de comandos

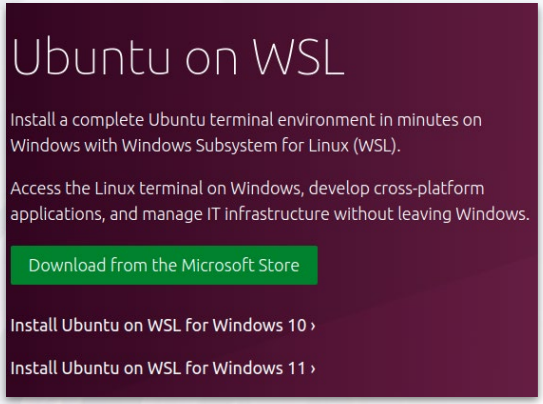
- Preparar y conservar datos.
- Ordenar y filtrar datos.
- Limpiar y actualizar datos.
- Gran cantidad de secuencias

```
- print the current directory
- list the contents
- make a new directory
- copy and rename files
- examine the content of (text) files
- Use tab completion
- Use meta characters
```

Cygwin







https://hackernoon.com/how-to-install-bash-on-windows-10-lqb73yj3 // https://www.codecademy.com/resources/blog/what-is-bash-used-for/

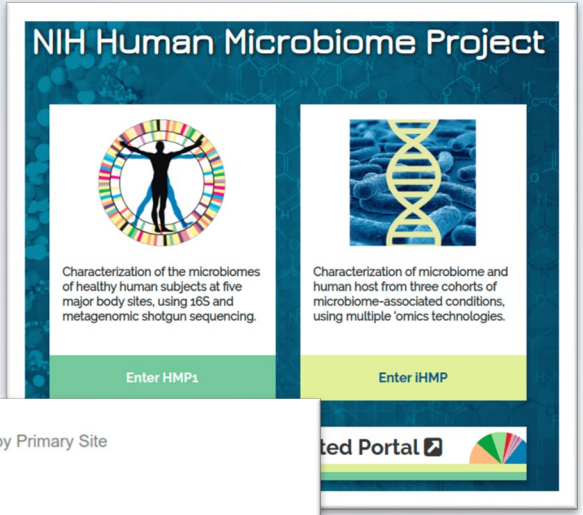
Contexto global-local de la Bioinformática

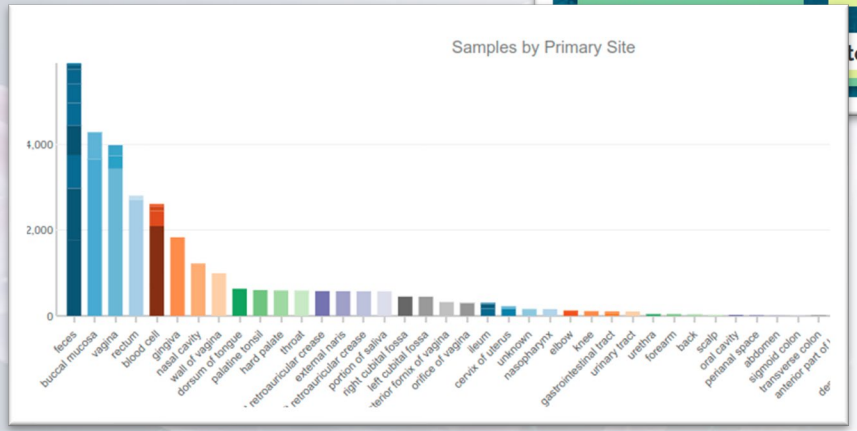
Aplicaciones - Divulgación - Iniciativas

Proyectos actuales en Genomica

NIH Human Microbiome Project

- El microbioma humano es el conjunto de todos los microorganismos que viven asociados al cuerpo humano.
- Cuerpo humano: fosas nasales, cavidad oral, piel, tracto gastrointestinal y tracto urogenital.
- Desarrollo de un conjunto de referencia de 3.000 genomas de aislados microbianos



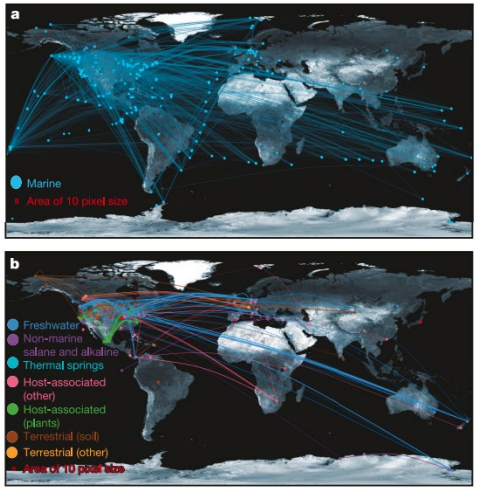


Lloyd-Price, J. et al. (2017). *Strains, functions and dynamics in the expanded Human Microbiome Project.* Nature // https://hmpdacc.org/

Proyectos actuales en Genomica

Uncovering earth's virome

- *5 Tb of metagenomic sequence data*
- *3,042 geographically diverse samples to assess the global distribution*
- *We discovered over 125,000 partial DNA viral genomes*
- Publicado en 2016



ARTICLE

Uncovering Earth's virome

David Paez-Espino[1], Emiley A. Eloe-Fadrosh[1], Georgios A. Pavlopoulos[1], Alex D. Thomas[1], Marcel Hu
Natalia Mikhailova[1], Edward Rubin[1,2,3], Natalia N. Ivanova[1] & Nikos C. Kyrpides[1]

Viruses are the most abundant biological entities on Earth, but challenges in detecting, isolating, and classifying unknown viruses have prevented exhaustive surveys of the global virome. Here we analysed over 5 Tb of metagenomic sequence data from 3,042 geographically diverse samples to assess the global distribution, phylogenetic diversity, and host specificity of viruses. We discovered over 125,000 partial DNA viral genomes, including the largest phage yet identified, and increased the number of known viral genes by 16-fold. Half of the predicted partial viral genomes were clustered into genetically distinct groups, most of which included genes unrelated to those in known viruses. Using CRISPR spacers and transfer RNA matches to link viral groups to microbial host(s), we doubled the number of microbial phyla known to be infected by viruses, and identified viruses that can infect organisms from different phyla. Analysis of viral distribution across diverse ecosystems revealed strong habitat-type specificity for the vast majority of viruses, but also identified some cosmopolitan groups. Our results highlight an extensive global viral diversity and provide detailed insight into viral habitat distribution and host–virus interactions.

Paez-Espino, D., Eloe-Fadrosh, E., Pavlopoulos, G. et al. *Uncovering Earth's virome*. Nature (2016)

Proyectos actuales en Genomica

Genome data to malaria control

- 1895 muestras de P. vivax
- Colectadas en 88 sitios alrededor del mundo
- 20.000 muestras de los plasmodium más letales
- Abril 2022

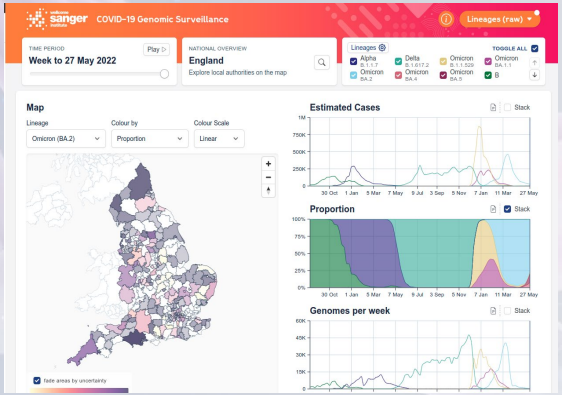
Genomic surveillance of malaria is vital in the fight to control and eliminate it. As is the case for COVID-19, it can be used to identify new variants, track the spread of strains of the pathogen in space and time, and pick up emerging resistance to drugs, diagnostic tests and vaccines

https://www.malariagen.net/ // https://sangerinstitute.blog/2022/04/25/from-genome-data-to-malaria-control/

Proyectos actuales en Genomica

SARS-Cov-2 genomic data for surveillance

- Sanger Institute ha secuenciado 2,5 millones de genomas desde 2020
- Representa un 20% del total global



- 30 Petabytes de datos genómicos ~ 60 millones de laptops

https://covid19.sanger.ac.uk/lineages/raw // https://sangerinstitute.blog/2020/10/22/sequencing-covid-19-at-the-sanger-institute/

Proyectos actuales en Genómica

Latinbiota - Estudio de Microbioma en Latinoamérica

El consorcio internacional Latinbiota tiene como objetivo comprender la composición y variabilidad del microbioma humano en los países de América Latina.

Inicio en 2019.
La incorporación de Venezuela esta en fase de planificacion.

Nuestro enfoque se basa en herramientas metagenómicas y bioinformáticas de última generación para descubrir y caracterizar los microbios en condiciones de salud y enfermedad.

¿Quiénes somos?
EQUIPO

Dr. Gregorio Iraola
Uruguay
Lider del consorcio

Institut Pasteur de Montevideo

wellcome sanger institute

EMBL-EBI

https://www.latinbiota.net/ // https://www.bbc.com/mundo/noticias-59524325

Y en Venezuela?

Estamos haciendo bioinformática orientada al análisis, procesamiento y almacenamiento de datos generados por NGS

Unidad de Estudios Genéticos y Forenses (UEGF)

Secuenciación de genoma completo con MiSeq

Febrero-Marzo 2016

- Capacitación de Personal por illumina
- Primera corrida de NGS en Venezuela
- Se secuenciaron genomas bacterianos, virales y de protozooarios
- Se generó más 15 Gb de datos

Abril 2016

- Ensamblaje y anotación del primer genoma bacteriano en el país.

Run 160302 M04136 0002 000000000-AM5LY: Run & Lane Metrics

	CYCLES	YIELD	PROJECTED YIELD
Read 1	301	8.29 Gbp	8.29 Gbp
Read 2 (I)	8	193.54 Mbp	193.54 Mbp
Read 3 (I)	8	193.54 Mbp	193.54 Mbp
Read 4	301	8.29 Gbp	8.29 Gbp
Non-Index Reads Total	602	16.59 Gbp	16.59 Gbp
Totals	618	16.98 Gbp	16.98 Gbp

Hacia dónde vamos...

Algunos objetivos!

Creación y consolidación de un grupo interdisciplinario de trabajo e investigación en Bioinformática que apoye a estudiantes e investigadores con el procesamiento y análisis de datos generados por NGS.

En la UEGF

Campos de accion

- Formación y Divulgación
- Infraestructura Computacional
- Producción Científica
- Asesorías y Proyectos Colaborativos Nacionales e Internacionales

En la UEGF

Recursos humanos e Infraestructura computacional

Hardware

Dell Precision Tower

- 24 Núcleos
- 64 Gb de memoria RAM
- 6 TB de Disco Duro

Software

GNU/Linux

Ensambladores

- Velvet
- SPAdes
- Unicycler

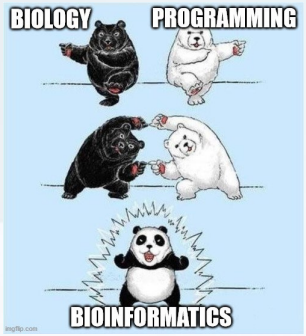
Anotación

- RAST
- Bakta
- Prokka

Genomica Comparativa

- fastANI
- Panaroo

Humanware



- Bioinformáticos
- Biólogos
- Informáticos
- Matemáticos
- Estadisticos
- Científicos de Datos
- Fisicos

UNU BIOLAC

Cursos Internacional .
Secuenciación y Análisis de Datos Genómicos para la Detección Microbiológica de Enfermedades Transmitidas por Alimentos y Aguas

INSTITUTO NACIONAL DE HIGIENE RAFAEL RANGEL

Formacion y Divulgacion

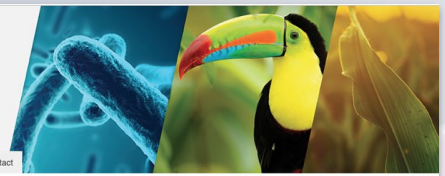
Actividades organizadas e impulsadas desde la Unidad y RSG

Ciclo de conferencias en BioInformática y Biología Computacional 2016

III Ciclo de Conferencias en BIOINFORMÁTICA 2019

IV Ciclo de Conferencias en BIOINFORMÁTICA 2020 Y BIOLOGÍA COMPUTACIONAL

CABANA
Capacity building for bioinformatics in Latin America



Home | About | Workshops | Research Secondments | Train the Trainer | eLearning Resources | News | Contact

III Cycle of Conferences in Bioinformatics and Computational Biology 2019

Dates: 06/21/2019 to 10/25/2019
Genetic and Forensic Studies Unit (UEGF, in Spanish) organises the third edition of Conferences on Bioinformatics and Computational Biology, with the aim of create a friendly environment for discussion, dissemination and formation in different topics oriented to the research and analysis of massive data generated from Next Generation Sequencing (NGS). This new edition, will make emphasis in showing the diverse applications of computational biology in Genomics, Biodiversity, Public Health, Agronomy, Data Science and Machine Learning. Including an array of specialist in bioinformatics and computational biology.
Confirmed speakers:
June 21, 2019, Marco Cristancho, Universidad de los Andes, Colombia, 'Genomics and Bioinformatics in Latin America' 10:00 (Venezuelan Time, GMT -04:00)
July 19, 2019, Cesar Rodriguez, Universidad de Costa Rica / CIET, 'Bacterial Genomics and antibiotic resistance' 12:00 Venezuelan Time, GMT -04:00)
Other speakers, to be announced.

Location: Online
URL: https://sites.google.com/view/conferenciasbioinformatica2019
Application procedure: Registration to attend the presentations via Youtube Live is free. https://sites.google.com/view/conferenciasbioinformatica2019/registro

UNESCO
Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

Comisión Nacional de Cooperación con la UNESCO
República Bolivariana de Venezuela

COMISIÓN NACIONAL VENEZOLANA DE COOPERACIÓN CON LA UNESCO
CNCU

UNU BIOLAC

Cursos Internacional .
Secuenciación y Análisis de Datos Genómicos para la Detección Microbiológica de Enfermedades Transmitidas por Alimentos y Aguas

INSTITUTO NACIONAL DE HIGIENE RAFAEL RANGEL

Formacion y Divulgacion - 2023

Actividades organizadas e impulsadas desde la UEGF y RSG-Venezuela

UNU BIOLAC

Cursos Internacional .
Secuenciación y Análisis de Datos Genómicos para la Detección Microbiológica de Enfermedades Transmitidas por Alimentos y Aguas

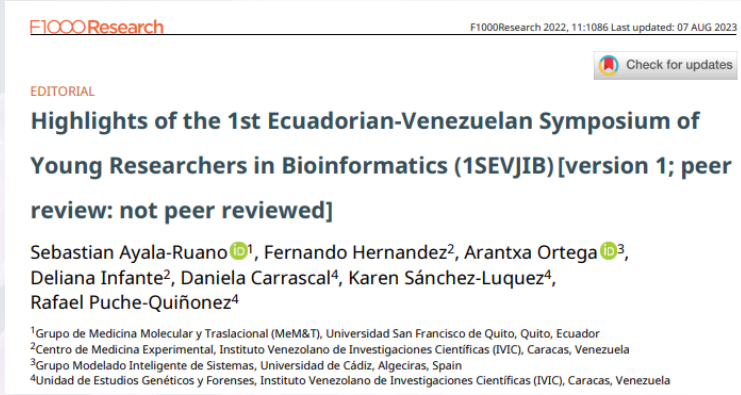
Eventos Cientificos

Actividades organizadas e impulsadas desde el RSG

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

ISCB REGIONAL Student GROUP Venezuela

El Grupo Regional de Estudiantes (RSG) de Bioinformática es una organización sin fines de lucro que tiene como objetivo consolidar y formalizar los trabajos e investigaciones en bioinformática y biología computacional que realizan los estudiantes de pregrado/posgrado que pertenecen a diferentes instituciones del país.



F1000Research F1000Research 2022, 11:1086 Last updated: 07 AUG 2023

Check for updates

EDITORIAL

Highlights of the 1st Ecuadorian-Venezuelan Symposium of Young Researchers in Bioinformatics (1SEVJIB) [version 1; peer review: not peer reviewed]

Sebastian Ayala-Ruano[1], Fernando Hernandez[2], Arantxa Ortega[3], Deliana Infante[2], Daniela Carrascal[4], Karen Sánchez-Luquez[4], Rafael Puche-Quiñonez[4]

[1]Grupo de Medicina Molecular y Traslacional (MeM&T), Universidad San Francisco de Quito, Quito, Ecuador
[2]Centro de Medicina Experimental, Instituto Venezolano de Investigaciones Científicas (IVIC), Caracas, Venezuela
[3]Grupo Modelado Inteligente de Sistemas, Universidad de Cádiz, Algeciras, Spain
[4]Unidad de Estudios Genéticos y Forenses, Instituto Venezolano de Investigaciones Científicas (IVIC), Caracas, Venezuela

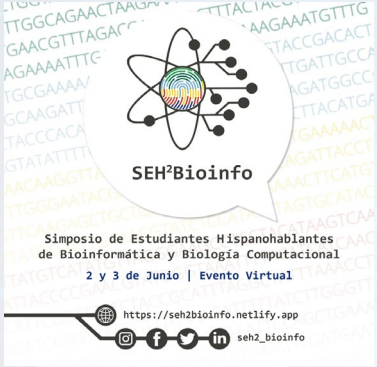
UNU BIOLAC

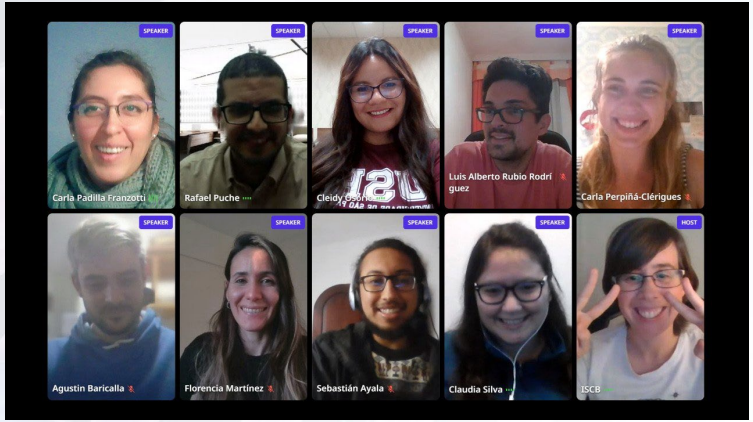
Cursos Internacional .
Secuenciación y Análisis de Datos Genómicos para la Detección Microbiológica de Enfermedades Transmitidas por Alimentos y Aguas

INSTITUTO NACIONAL DE HIGIENE RAFAEL RANGEL

Eventos Cientificos

Actividades organizadas e impulsadas desde el RSG





Gente, Ciencia y Tecnología al Servicio de la Salud

Producción científica

Participacion Internacional

FOI

ENA
European Nucleotide Archive

DTU Food
National Food Institute

1 INTRODUCTION AND OBJECTIVES

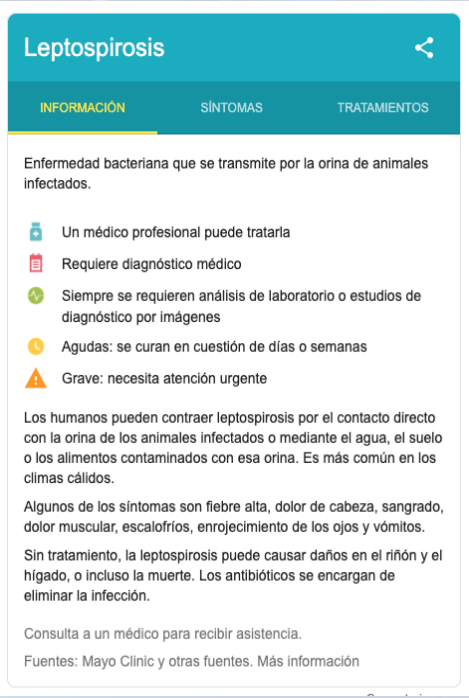
This is a joint proficiency test for the UNSGM, Compare and GMI Networks, and the component 1 (PT1) is a simulation exercise aiming to strengthen the ability to detect a biological threat based on genomic analysis. A biological agent is a bacterium, virus, protozoan, parasite, or fungus that can be used purposefully as a weapon in bioterrorism or biological warfare. In addition to these living and/or replicating pathogens, biological toxins are also included among the bio-agents.

Organismos:

- National Food Institute, Technical University of Denmark (DTU Food)
- United Nations Office of Disarmament Affairs (UNODA)
- Swedish Defense Research Agency (FOI)
- European Nucleotide Archive (ENA).

Producción científica

Estudiando la leptospirosis



Expertos del Instituto Venezolano de Investigaciones Científicas (IVIC) y del Institut Pasteur de Montevideo y París, identificaron una nueva especie, detectada en un paciente humano, una rata y una vaca, todos residentes en zonas cercanas en Venezuela.

Según el trabajo publicado en el International **Journal of Systematic and Evolutionary Microbiology**, la variedad de huéspedes de la bacteria —bautizada Leptospira venezuelensis— evidencia que el escenario es favorable para su diseminación.

UNU BIOLAC

Cursos Internacional .
Secuenciación y Análisis de Datos Genómicos para la Detección Microbiológica de Enfermedades Transmitidas por Alimentos y Aguas

INSTITUTO NACIONAL DE HIGIENE RAFAEL RANGEL

RIF: G-20000101-1

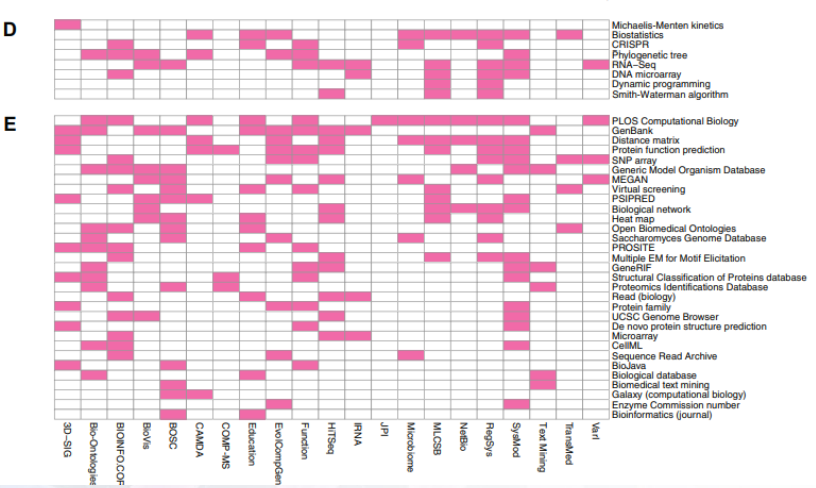
Producción científica

Bioinformatics, 38, 2022, i19–i27
https://doi.org/10.1093/bioinformatics/btac236
ISCB/ISMB 2022

OXFORD

Characterizing domain-specific open educational resources by linking ISCB Communities of Special Interest to Wikipedia

Alastair M. Kilpatrick[1,*,†], Farzana Rahman[2,3,*,†], Audra Anjum[4], Sayane Shome[5], K. M. Salim Andalib[6], Shrabonti Banik[7], Sanjana F. Chowdhury[8], Peter Coombe[9], Yesid Cuesta Astroz[10], J. Maxwell Douglas[11], Pradeep Eranti[12], Aleyna D. Kiran[13], Sachendra Kumar[14], Hyeri Lim[15], Valentina Lorenzi[16,17], Tiago Lubiana[18,19], Sakib Mahmud[20], Rafael Puche[21], Agnieszka Rybarczyk[22], Syed Muktadir Al Sium[8], David Twesigomwe[23,24], Tomasz Zok[22], Christine A. Orengo[25], Iddo Friedberg[26,27], Janet F. Kelso[28] and Lonnie Welch[29]

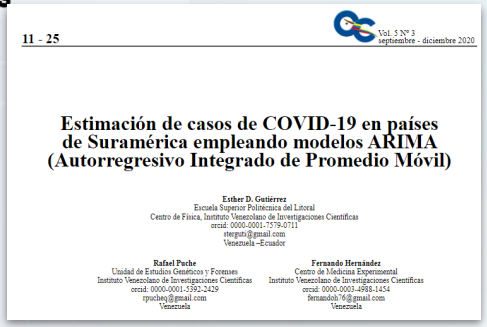


Producción científica

ARIMA - Covid - Ciencias de Datos - Interdisciplinario

ARIMA:

- ARIMA es un modelo estadístico utilizado para analizar series de tiempo, y ayuda a predecir valores futuros, como la temperatura o el precio de una acción.
- Objetivo: estimar nuevos contagios usando datos públicos disponibles para Venezuela y la región suramericana, realizando la predicción a 30 días del número de casos de Covid-19 en países suramericanos



11 - 25

Vol. 5 Nº 3
septiembre - diciembre 2020

Estimación de casos de COVID-19 en países de Suramérica empleando modelos ARIMA (Autorregresivo Integrado de Promedio Móvil)

Esther D. Gutiérrez
Escuela Superior Politécnica del Litoral
Centro de Física, Instituto Venezolano de Investigaciones Científicas
orcid: 0000-0001-7579-0711
stegut@gmail.com
Venezuela –Ecuador

Rafael Puche
Unidad de Estudios Genéticos y Forenses
Instituto Venezolano de Investigaciones Científicas
orcid: 0000-0001-5392-2429
rpuche@gmail.com
Venezuela

Fernando Hernández
Centro de Medicina Experimental
Instituto Venezolano de Investigaciones Científicas
orcid: 0000-0003-4988-1454
fernandoh76@gmail.com
Venezuela

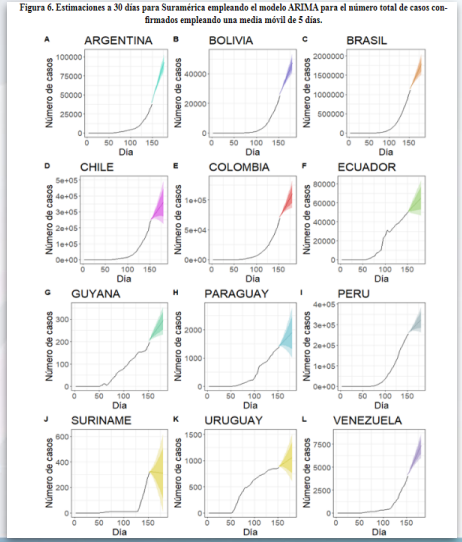


Figura 6. Estimaciones a 30 días para Suramérica empleando el modelo ARIMA para el número total de casos confirmados empleando una media móvil de 5 días.

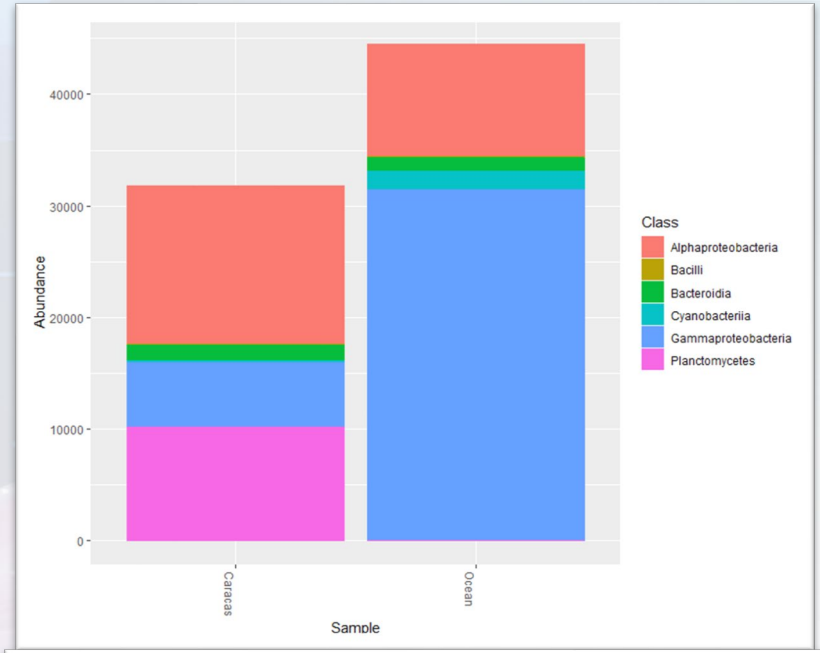
Gutiérrez, E. D., Puche, R., & Hernández, F. (2020). *Observador Del Conocimiento*, *5*(3), 11–25.

Proyectos Colaborativos

Metagenomica

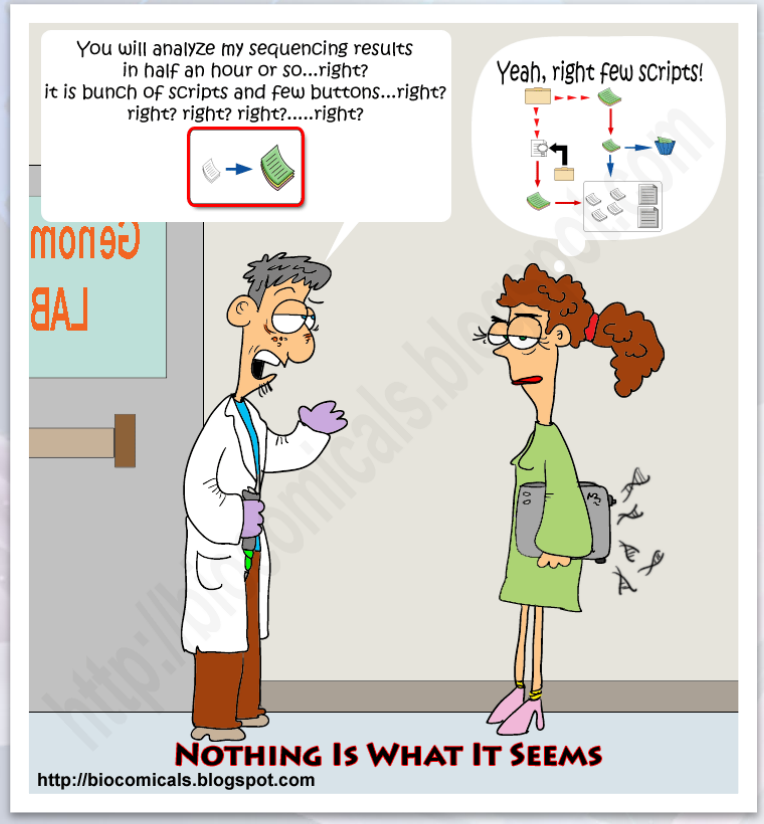
- Caracterización de comunidades bacterianas presentes en ambientes acuáticos de la región Capital sujetos a impacto antrópico.

- Dra. Christine Cagnon (Universidad de Pau, Francia)
- Dra. Milagro Fernández (IVIC)
- Dra. Paula Suarez (USB)




```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 2532 taxa and 31 samples ]
sample_data() Sample Data:       [ 31 samples by 3 sample variables ]
tax_table()   Taxonomy Table:    [ 2532 taxa by 7 taxonomic ranks ]
```

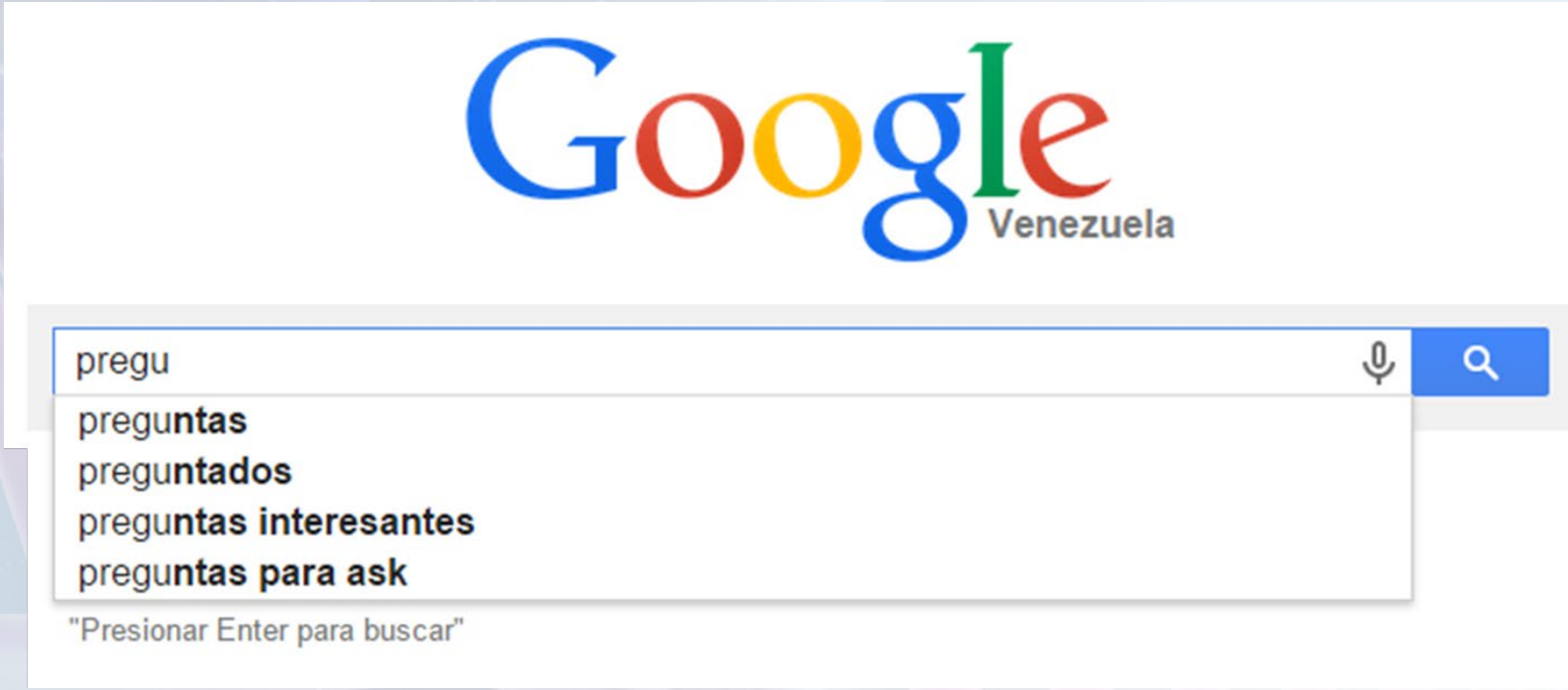
No es tan simple como se aparenta...



...pero no imposible

Muchas preguntas por responder...

Muchas preguntas por hacer...


Google
Venezuela
pregu
preguntas
preguntados
preguntas interesantes
preguntas para ask
"Presionar Enter para buscar"

Gracias!

¿Mas preguntas?

0212-5041622

rafael.puche@pedeciba.edu.uy

Twitter: @rpucheq

Telegram: t.me/BioinformaticaVZLA