

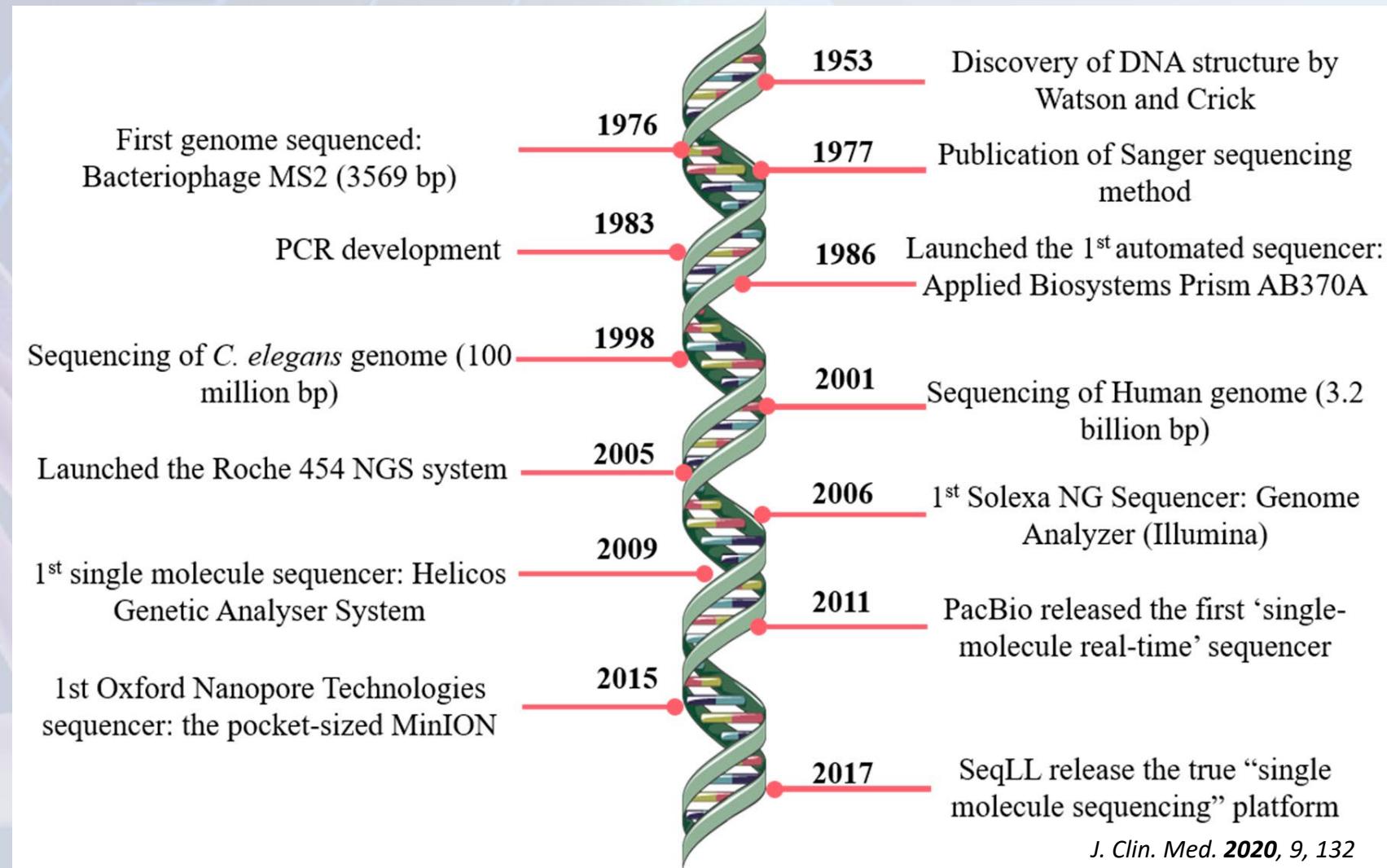
Herramientas computacionales para el análisis de datos genómicos

Fernando Hernandez, IVIC



Los Avances!!!

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



- 1976:** Virus de ARN -- Fago MS2 (3 kbp)
- 1977:** Virus de ADN -- Fago Φ -X174 (6 kbp)
- 1995:** Bacteria -- *Haemophilus influenzae* (1.8 Mbp)
- 1995:** Eukarya -- *Saccharomyces cerevisiae* (12 Mbp)
- 1996:** Arquea -- *Methanococcus jannaschii* (1.6 Mbp)
- 2000:** Primer borrador del genoma humano -- J. Craig Venter (3 Gbp)

La ola de secuenciación del ADN

- 1953: Estructura del ADN
- 1972: ADN recombinante
- 1977: Secuenciación de Sanger
- 1985: PCR
- 1988: NCBI
- 1990: BLAST



THE JOURNAL OF BIOLOGICAL CHEMISTRY
 Vol. 248, No. 11, Issue of June 10, pp. 3860-3875, 1973
Printed in U.S.A.

The Nucleotide Sequence of *Saccharomyces cerevisiae* 5.8 S Ribosomal Ribonucleic Acid

(Received for publication, November 20, 1972)

GERALD M. RUBIN*

From the Medical Research Council Laboratory of Molecular Biology, Cambridge, CB2 2QH, England

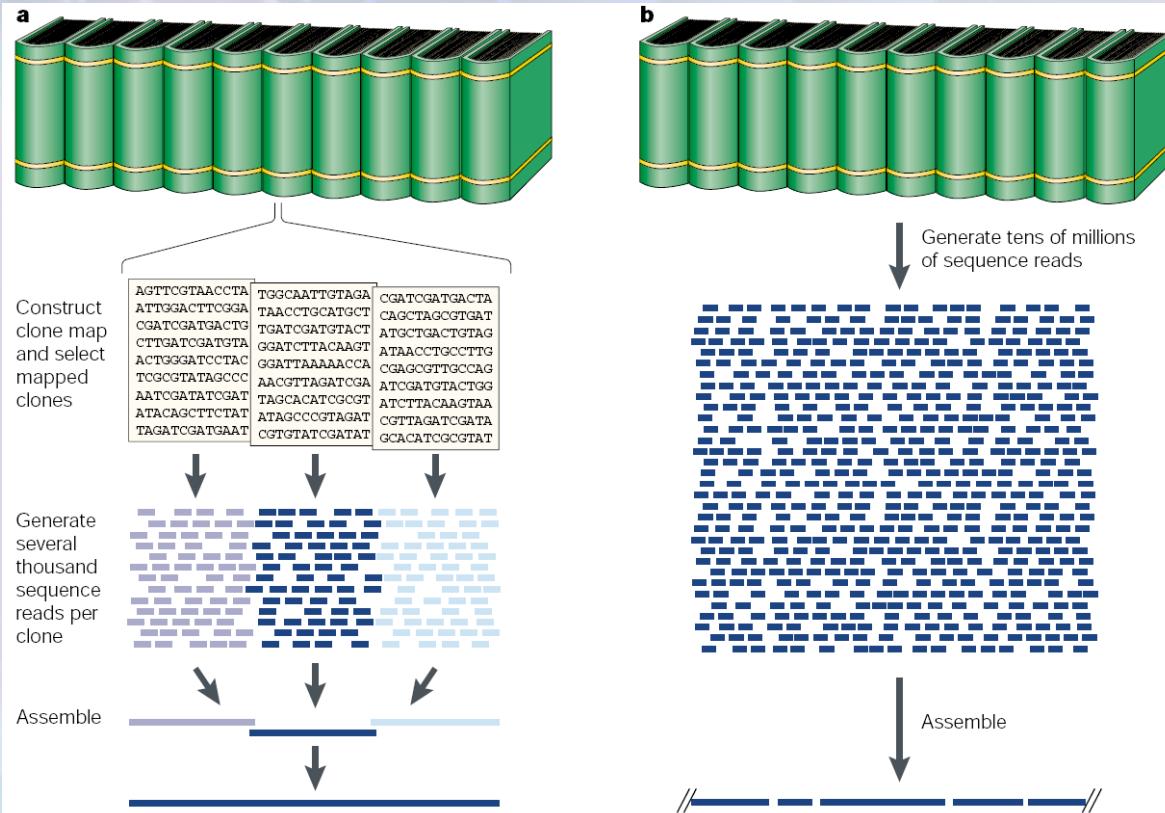
SUMMARY

The nucleotide sequence of *Saccharomyces cerevisiae* 5.8 S ribosomal RNA (also known as the 7 S or 1RNA species) has been determined to be pApApApCpUpUpUpCpApApCpA pApCpGpGpApUpCpUpCpUpUpGpGpUpUpCpUpCpGpC pApUpCpGpApUpGpApApGpApApCpGpCpApGpCpGpApA pApUpGpCpGpApUpApCpGpUpApApUpGpUpGpApAΨpUpG pCpApGpApApUpUpCpCpGpUpGpApApUpCpApUpCpGpA pApUpCpUpUpUpGpApApCpCpApGpGpGpGpCpA pCpCpCpUpUpGpGpUpApUpCpCpApGpGpGpGpCpA pUpGpCpCpUpGpUpUpGpApGpCpGpUpCpApUpUpU.

Low Phosphate Medium—Inorganic phosphate was precipitated (as $MgNH_4PO_4$) from 10% Bacto-yeast extract and 20% Bacto-peptone by the addition of 10 ml of 1 M $MgSO_4$ and 10 ml of concentrated aqueous ammonia per liter. The phosphates were allowed to precipitate at room temperature for 30 min, and the precipitate was removed by filtration through Whatman No. 1 filter paper. The filtrate was adjusted to pH 5.8 with HCl and autoclaved. Sterile glucose was added to a final concentration of 2%.

Secuenciación en 1970

La carrera del genoma humano



- **Clone-by-clone and**
- **Whole-genome shotgun**

La carrera del genoma humano

- Proyecto **Genoma Humano**: 1990-2003

Originalmente 1990-2005

Impulsado por la mejora tecnológica y la automatización.

Competencia de Celera

- La **informática** es esencial para los esfuerzos de secuenciación públicos y privados.

Ensamblaje de secuencias y predicción de genes.

Borrador de trabajo terminado simultáneamente en la primavera de 2000.

Genoma humano completo 2003



PRODUCTION

Rooms of equipment
Sample preparations
35 people
3-4 weeks



SEQUENCING

74x Capillary Sequencers
10 people
15-40 runs per day
1-2Mb per instrument per day
120Mb total capacity per day

Secuenciación en el 2001



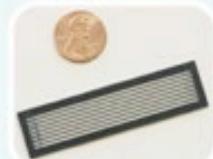
PRODUCTION

1x Cluster Station
1 person
1 day



SEQUENCING

1x Genome Analyzer
Same person as above
1 run per 3-5 days
0.5Gb per day per instrument



Secuenciación en el 2007

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

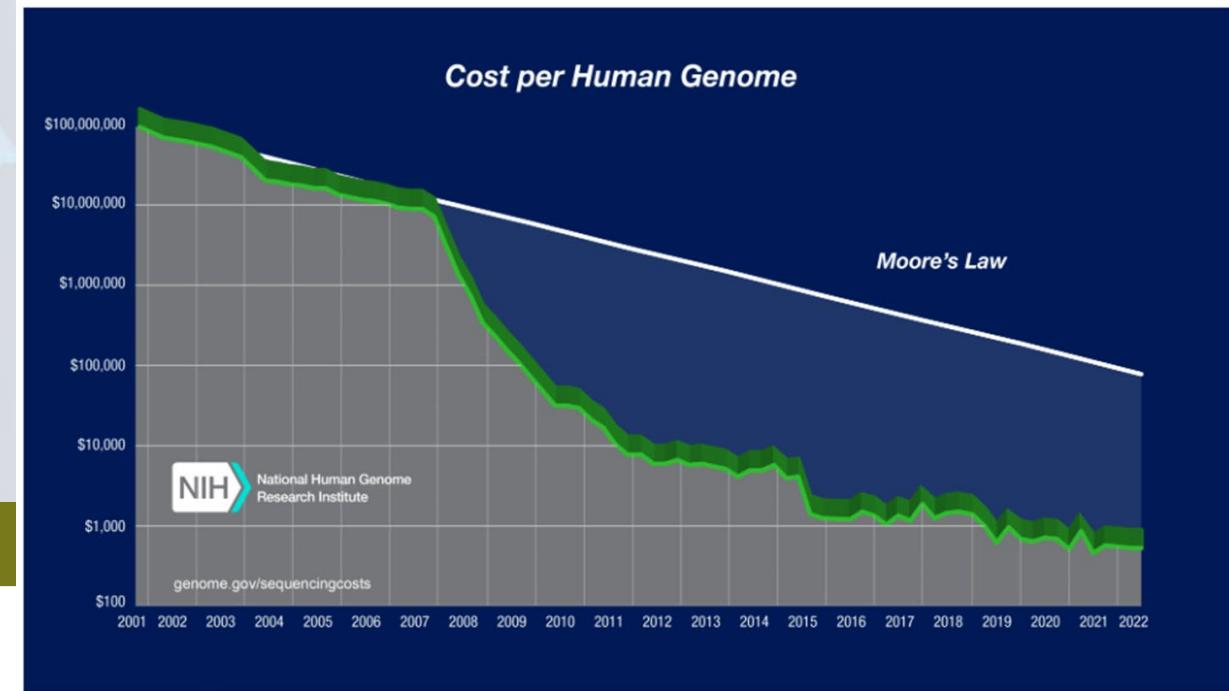
| | | iSeq 100 System | MiniSeq System | MiSeq Series + | NextSeq Series + |
|-----------------------|--|---|---|---|--|
| Run Time | | 9–17.5 hours | 4–24 hours | 4–55 hours | 12–30 hours |
| Maximum Output | | 1.2 Gb | 7.5 Gb | 15 Gb | 120 Gb |
| Maximum Reads Per Run | | 4 million | 25 million | 25 million † | 400 million |
| Maximum Read Length | | 2 × 150 bp | 2 × 150 bp | 2 × 300 bp | 2 × 150 bp |
| | |  |  |  |  |
| | | NextSeq Series + | HiSeq 4000 System | HiSeq X Series‡ | NovaSeq 6000 System |
| Run Time | | 12–30 hours | < 1–3.5 days | < 3 days | ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells) |
| Maximum Output | | 120 Gb | 1500 Gb | 1800 Gb | 6000 Gb |
| Maximum Reads Per Run | | 400 million | 5 billion | 6 billion | 20 billion |
| Maximum Read Length | | 2 × 150 bp | 2 × 150 bp | 2 × 150 bp | 2 × 150 bp |

Secuenciación hoy!



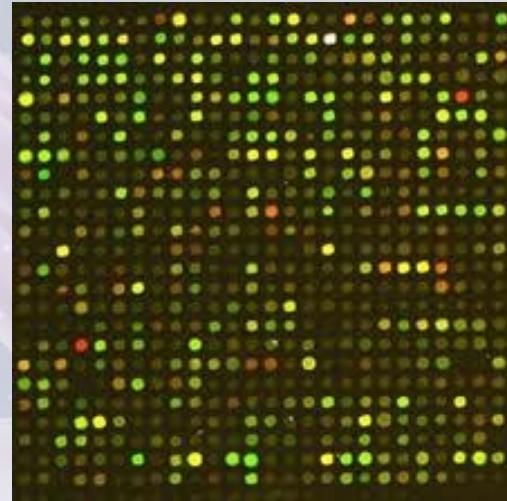
Letter | Published: 24 September 2018

A universal SNP and small-indel variant caller using deep neural networks

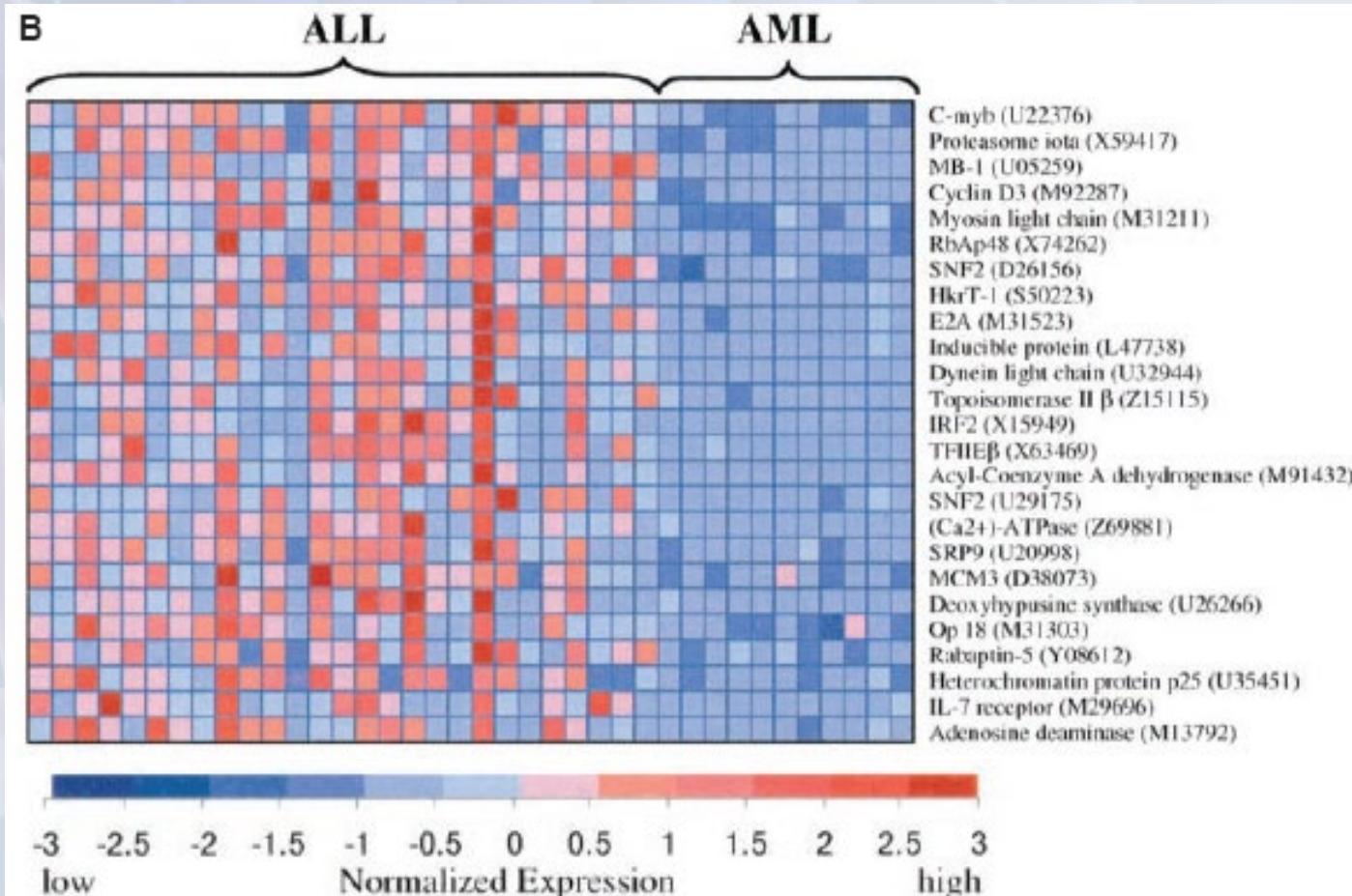


La ola de los **microarreglos!!**

- Los microarreglos contienen de cientos a millones de sondas
- Detectar simultáneamente que tanto cada gen se está expresando

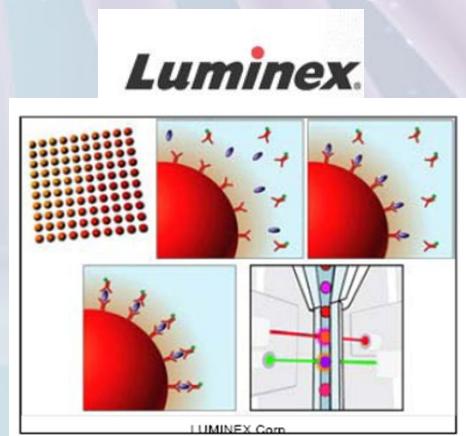


- Golub et al, Science 1999.



LLA vs LMA

- Inferir niveles de expresión de un gran conjunto de genes
- **Mapa de conectividad:** 1.5M perfiles de expresión para mas de 3 mil perturbaciones en genes y 5 mil drogas



“Microarreglos” hoy



Retos!!!!

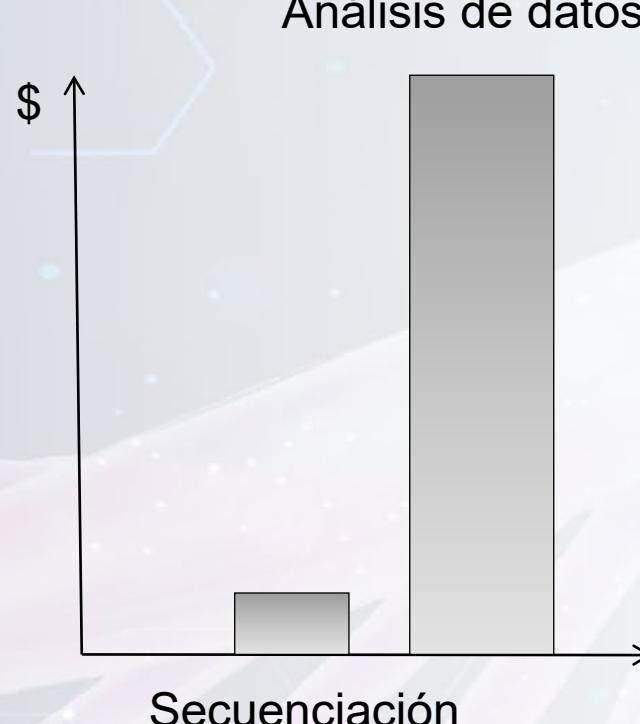
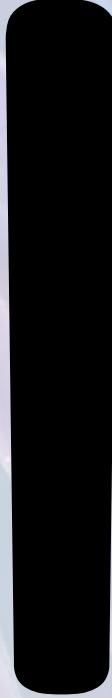
Cambio de **paradigma**

- De **genes** individuales a **genomas** completos
- De **transcritos** individuales a **transcriptomas** completos
- De **organismos** individuales a **grupos metagenómicos** complejos
- De **organismos** modelo hasta las **especies** particulares.

Presenta importantes **retos** en análisis de **grandes datos** (Big data)



Análisis de datos!



Aun se requieren + bioinformáticos!



CABANA
Capacity building for bioinformatics in Latin America

Home About Workshops Research secondments Train the trainer eLearning Webinars News Contact us

Fortalecer capacidades en bioinformática para Latinoamérica



Alto poder computacional
Una infraestructura IT dedicada

Tecnologías de NGS

| Compañía | Plataforma | Amplificación | Método de secuenciación |
|---------------------------|------------------------|-----------------|-------------------------|
| Roche | 454 | emPCR | Pirosecuenciación |
| Illumina | HiSeq MiSeq | Bridge PCR | Síntesis |
| LifeTech | SOLID | emPCR/ Wildfire | Ligación |
| Thermo | Ion Torrent/Ion Proton | emPCR | Síntesis (pH) |
| Pacific Bioscience | RSII | Ninguna | Síntesis |
| Complete genomics | Nanoballs | Ninguna | Ligación |
| Oxford Nanopore | minION | Ninguna | Flujo |

Tecnologías fallidas: Helicos, Polonator, etc.



En desarrollo

Conferences > 2012 12th IEEE International ... 

Tunnel-current based single-molecule identification of DNA/RNA oligomer by using nano-MCBJ

Publisher: IEEE

[Cite This](#)

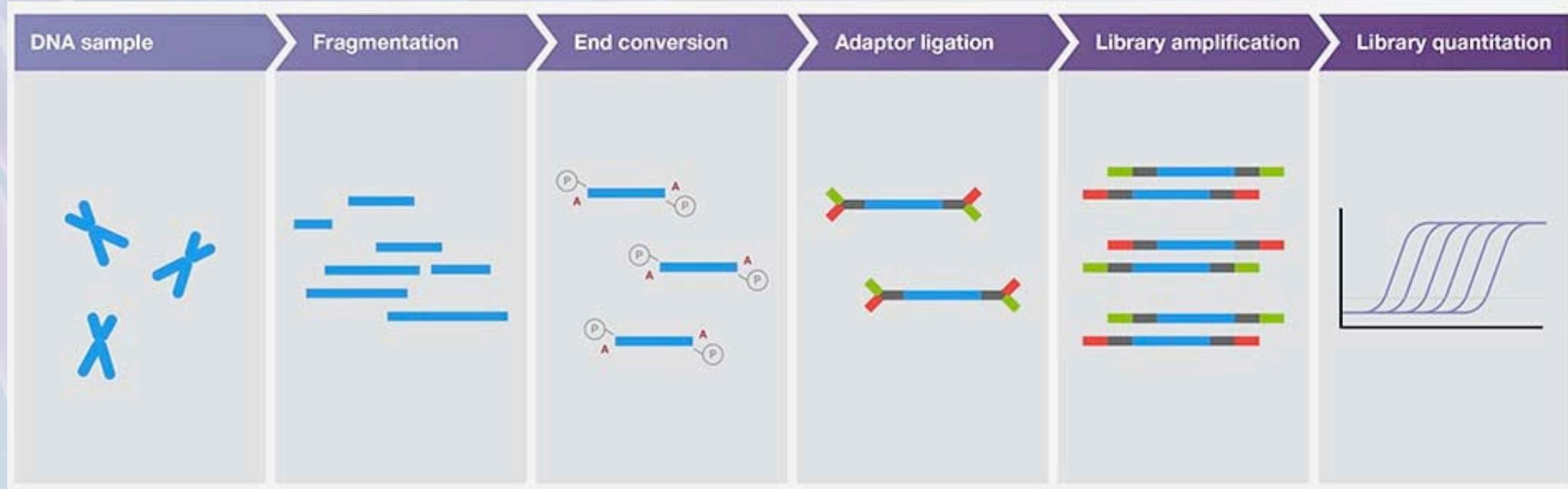
 [PDF](#)



En **NGS**, una librería se define como una colección de fragmentos de **ADN/ARN** que representan la totalidad del **genoma/transcriptoma** o una **región en específico**

Una buena librería debe tener una alta sensibilidad y especificidad

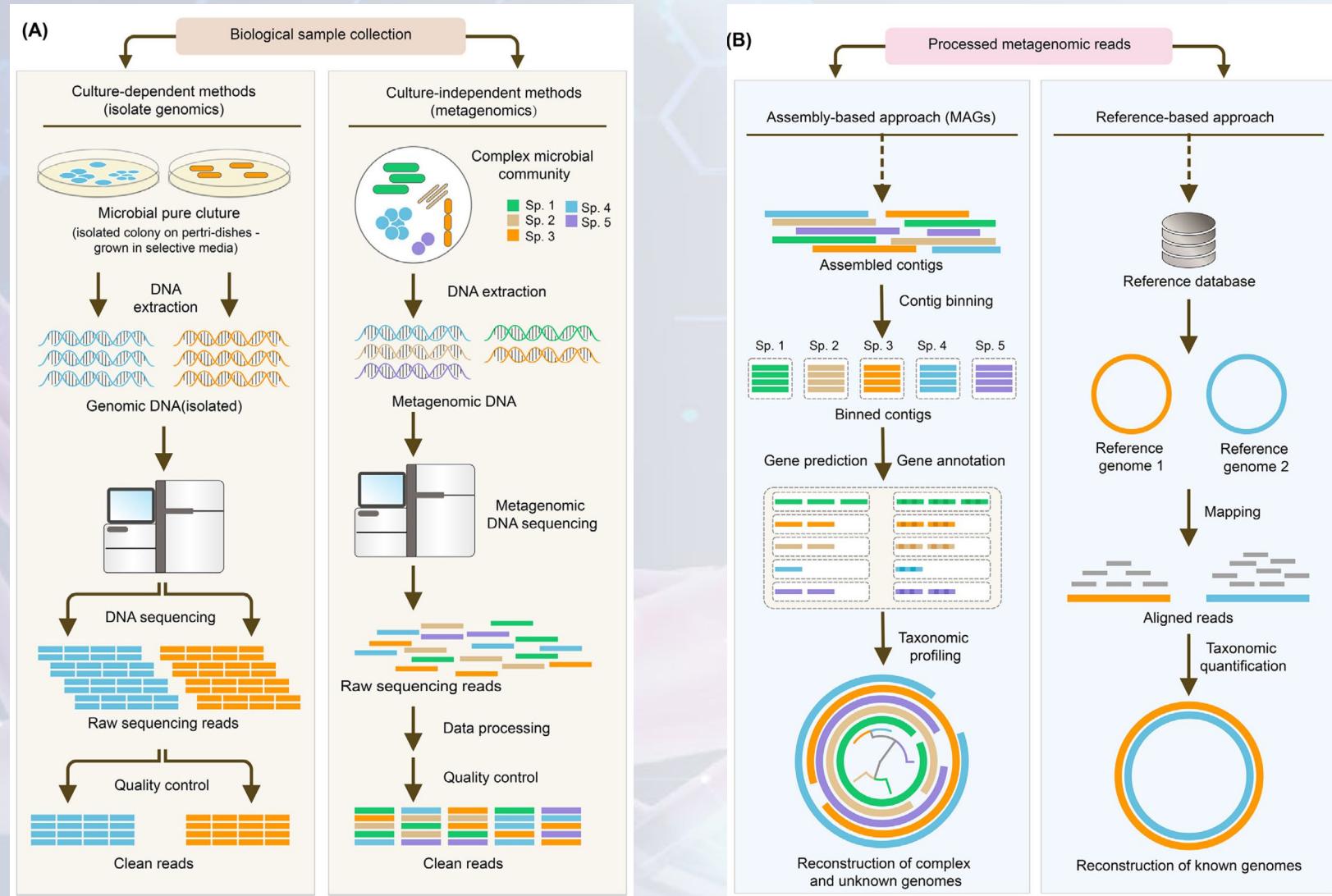
Cursos Internacional .
Secuenciación y Análisis de Datos Genómicos para la Detección Microbiológica
de Enfermedades Transmitidas por Alimentos y Aguas



<https://www.thermofisher.com/>

Cursos Internacionales .

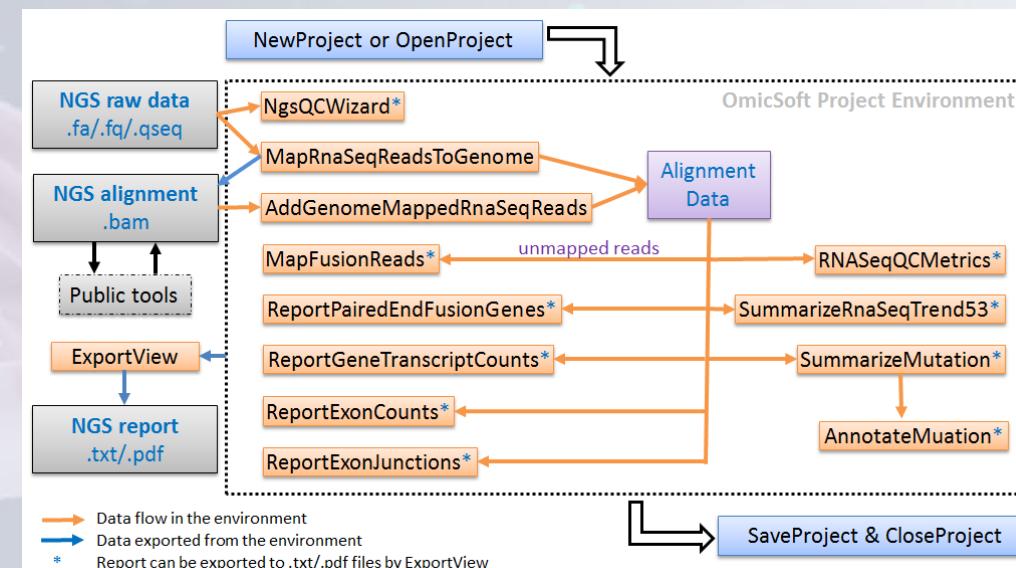
Secuenciación y Análisis de Datos Genómicos para la Detección Microbiológica de Enfermedades Transmitidas por Alimentos y Aguas



- Básicamente un conjunto de pasos para analizar datos



- Los flujos de trabajo pueden ser muy **sencillos** o muy **complejos!!!...**





BIOINFORMÁTICA
322 millones de resultados

HERRAMIENTAS BIOINFORMÁTICAS
175 millones de resultados

Noviembre 2023

Diferencia entre bioinformática y biología computacional

Biología computacional = el estudio de la **biología** utilizando **técnicas computacionales**. El objetivo es aprender nueva biología, conocimientos sobre los sistemas vivos. Es acerca de la **ciencia**.

Bioinformática = la creación de herramientas (**algoritmos, bases de datos**) que resuelven problemas. El objetivo es construir herramientas útiles que funcionen con datos biológicos. Es acerca de la **ingeniería**

Bioinformática y biología computacional

- **Interdisciplinaria**

Estadística, Biología, Informática
Cambio de énfasis con el tiempo

- **Aplicada**

Desde estudiantes de pregrado hasta posdoctorantes
Formación útil para muchos
Cuanto mas practiques, mejor lo harás

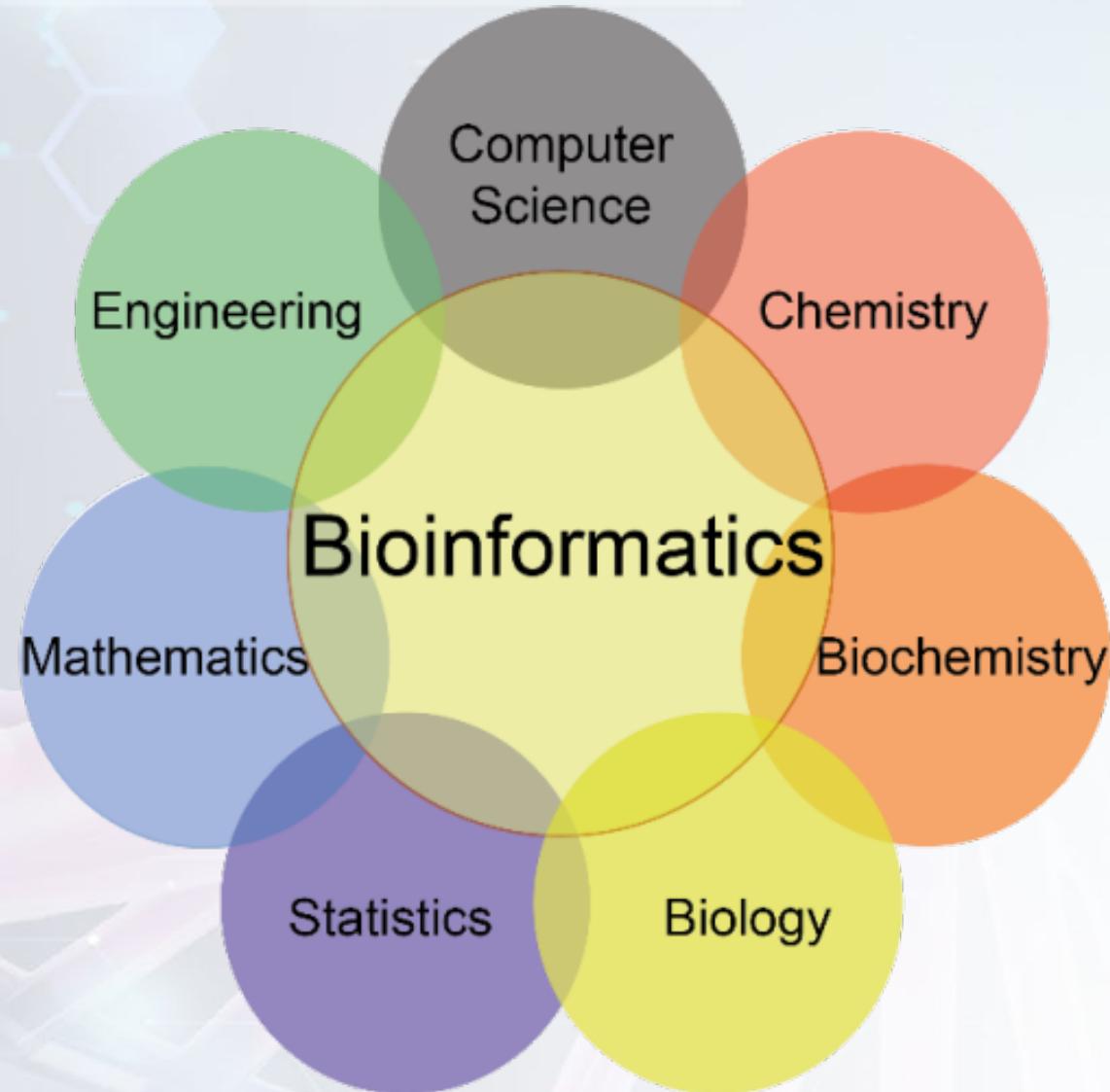
- **Avanza con los desarrollos tecnológicos**



Programas/ Repositorios

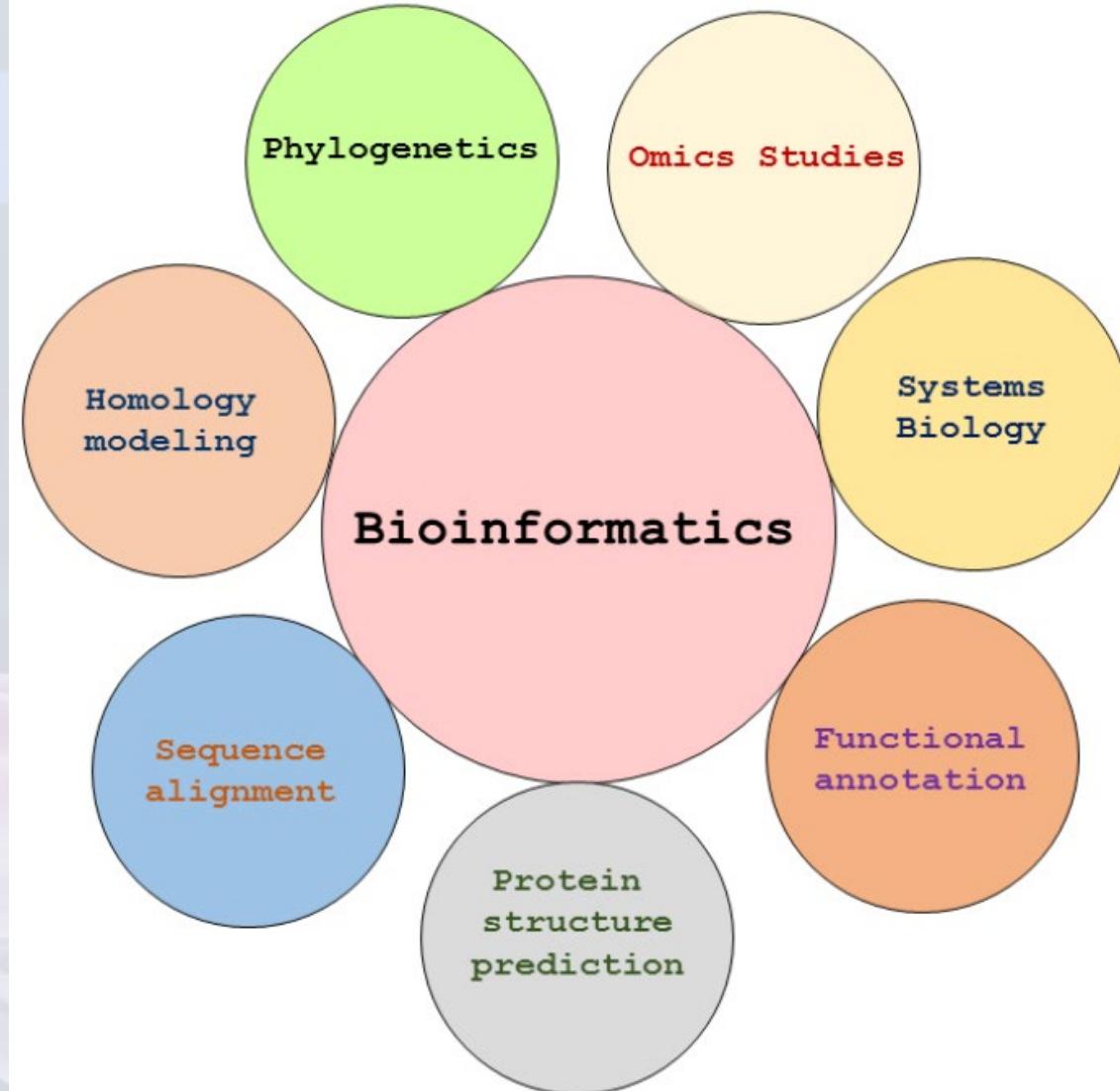


Ciencia multidisciplinaria



Múltiples enfoques

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



No tener miedo de realizar análisis de datos y algo de programación



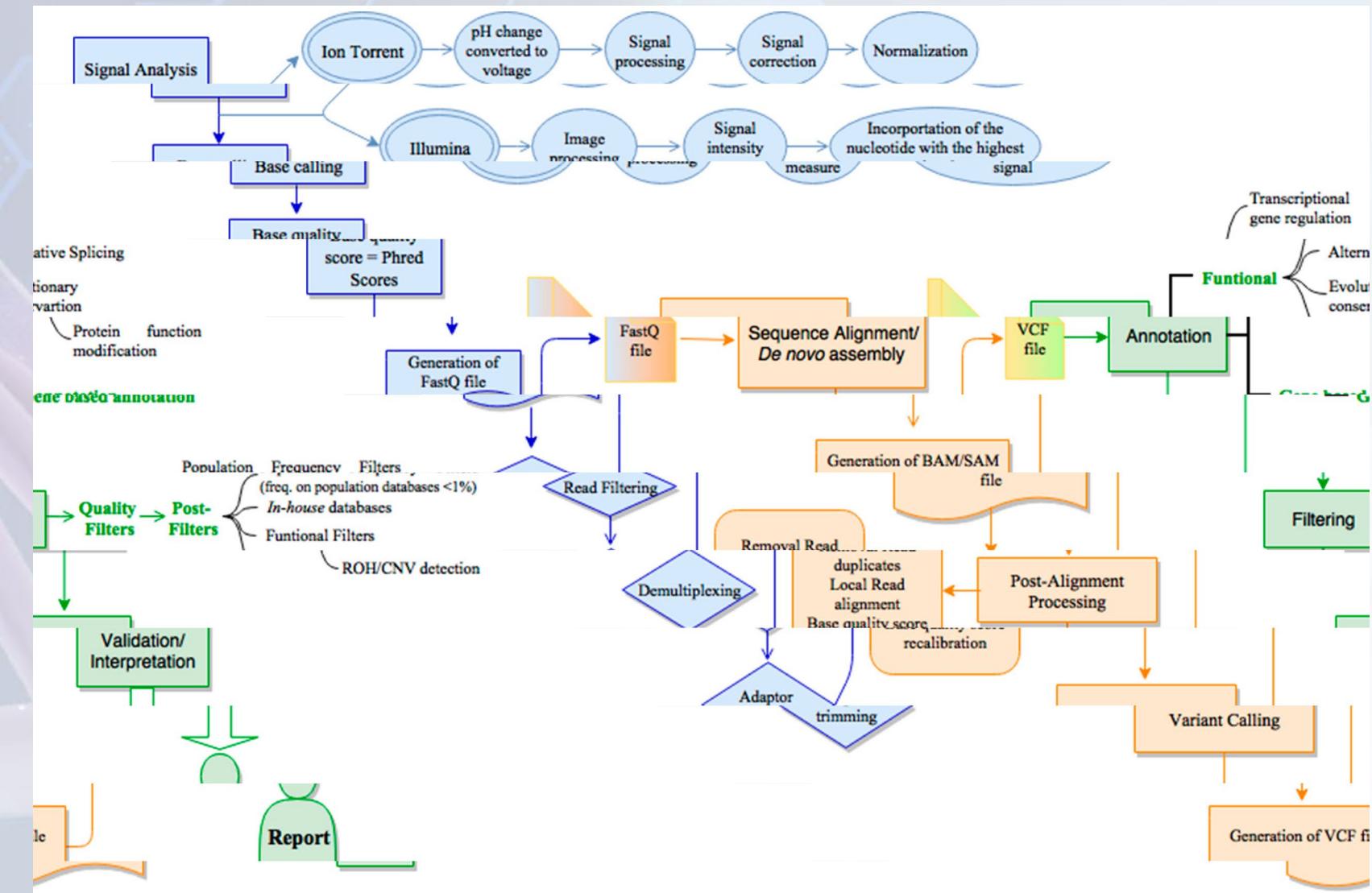
Herramientas



- Control de calidad
- Análisis primarios
- Análisis secundarios
- Análisis terciarios

Cursos Internacionales .

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



Análisis primarios

Trimmomatic

FastQC

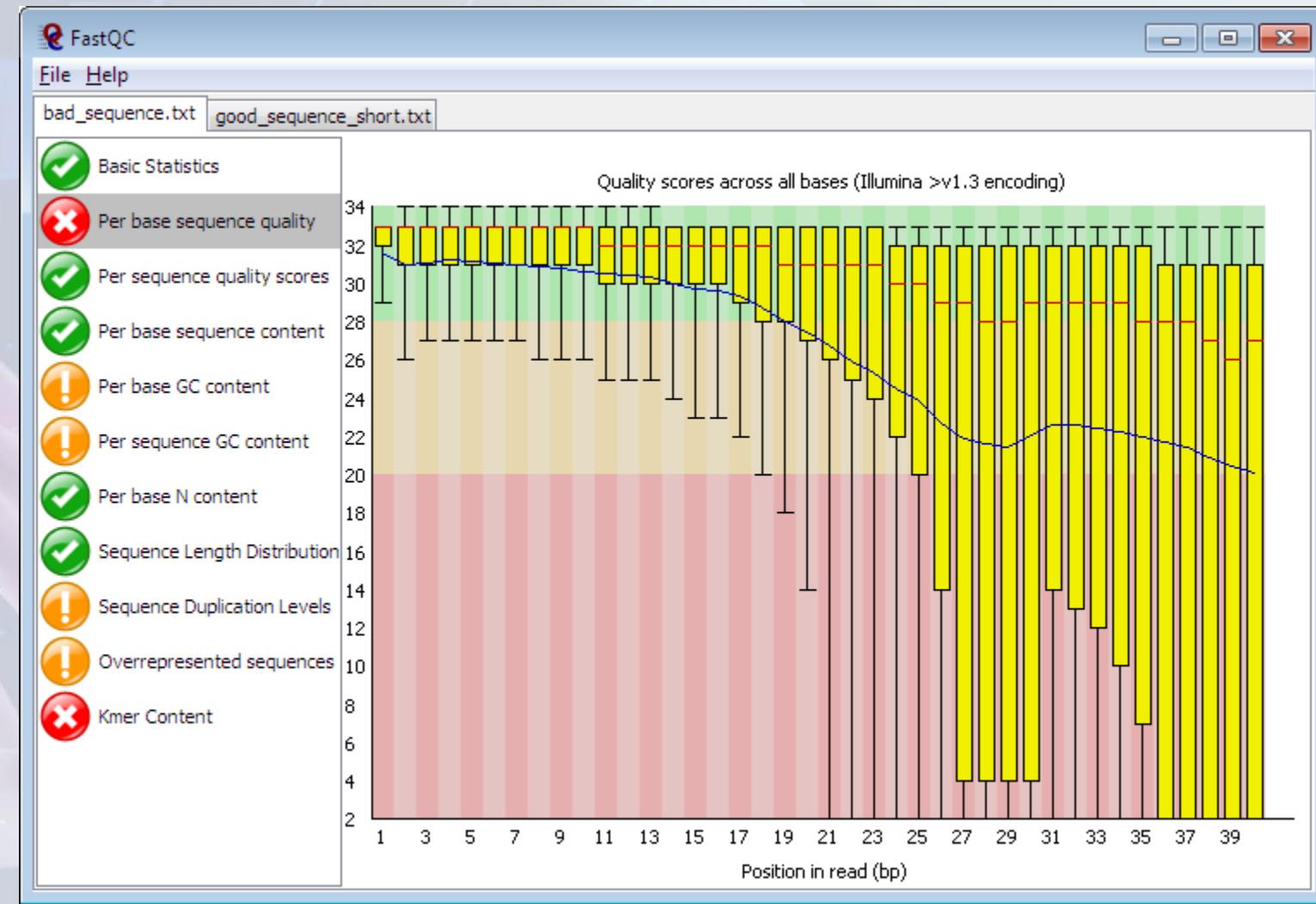
QC/Trimming

SequelTools (Long reads)

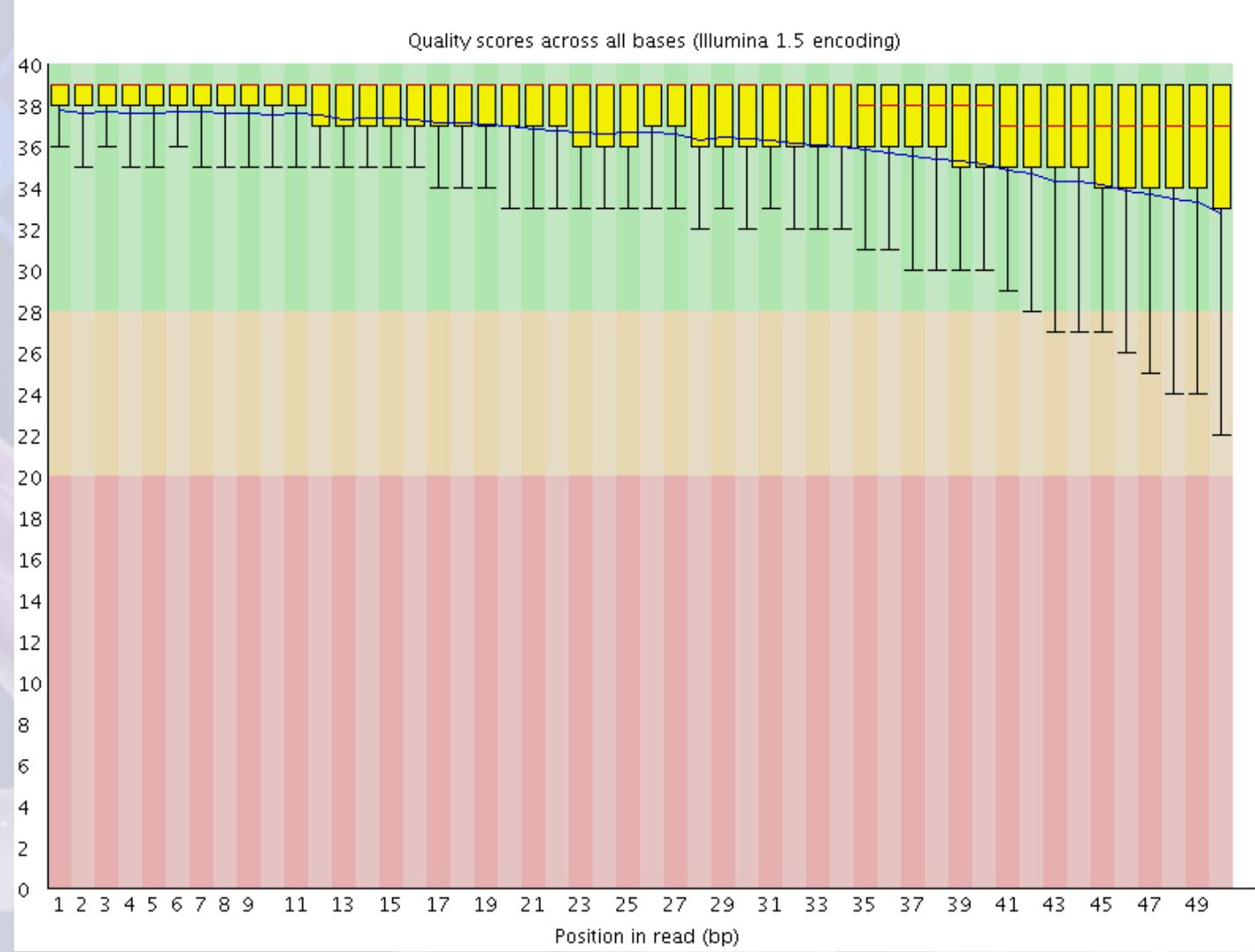
fastp

Cursos Internacionales .

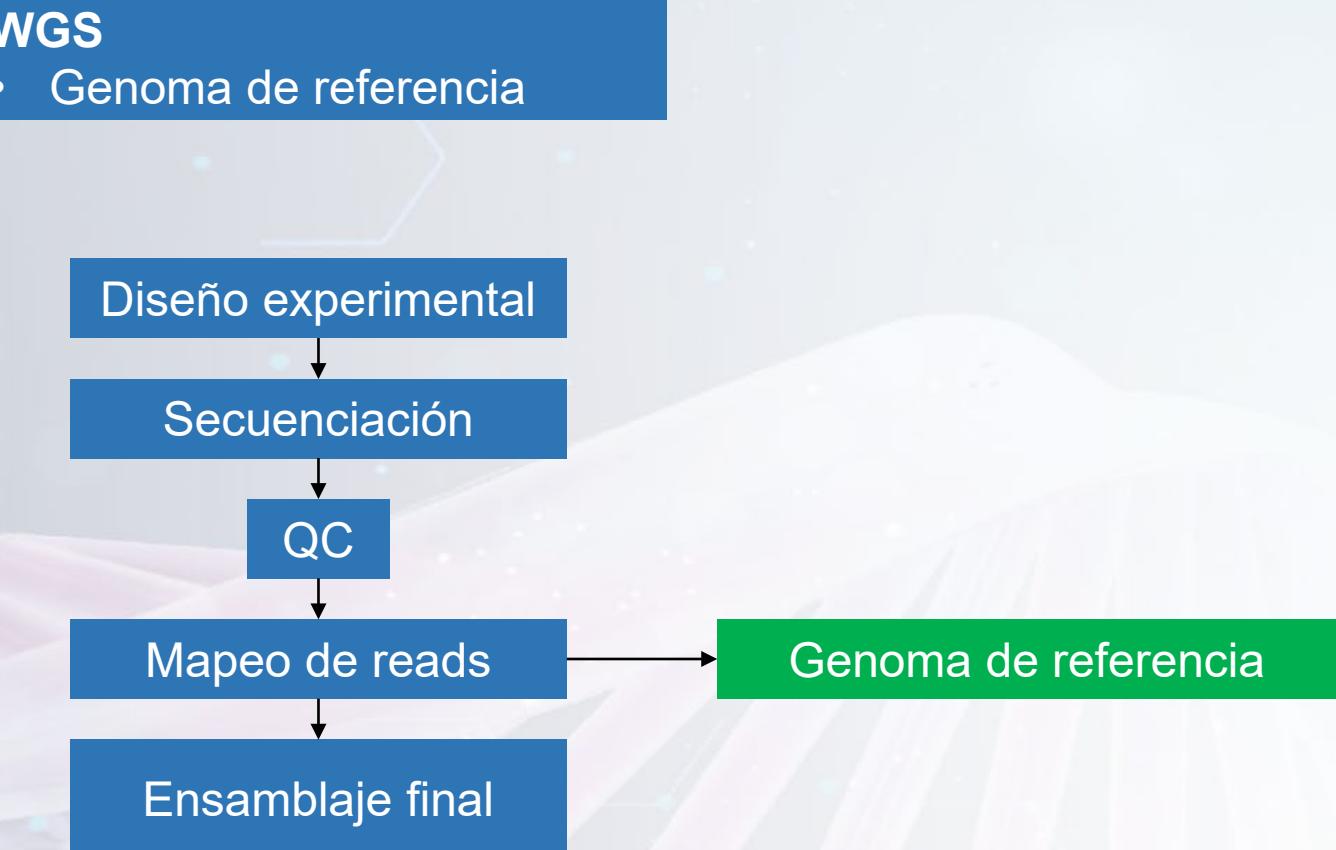
Secuenciación y Análisis de Datos Genómicos para la Detección Microbiológica de Enfermedades Transmitidas por Alimentos y Aguas



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

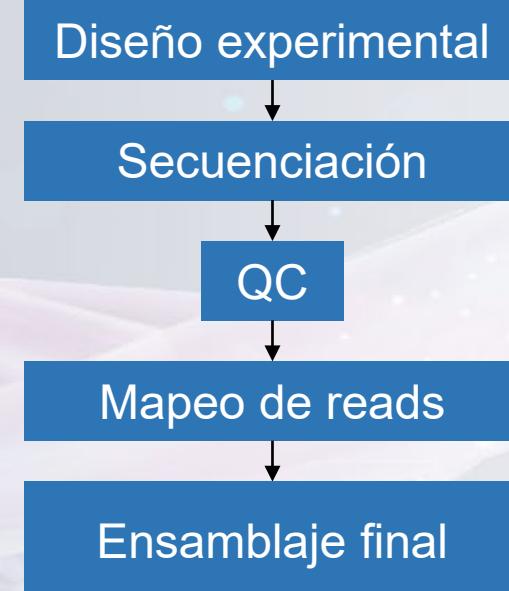


Análisis secundarios



Análisis secundarios

WGS
• Sin genoma de referencia



BWA/Bowtie2

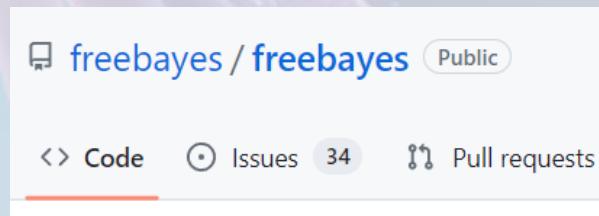
QUAST
Quality Assessment Tool for Genome Assemblies



SAMtools

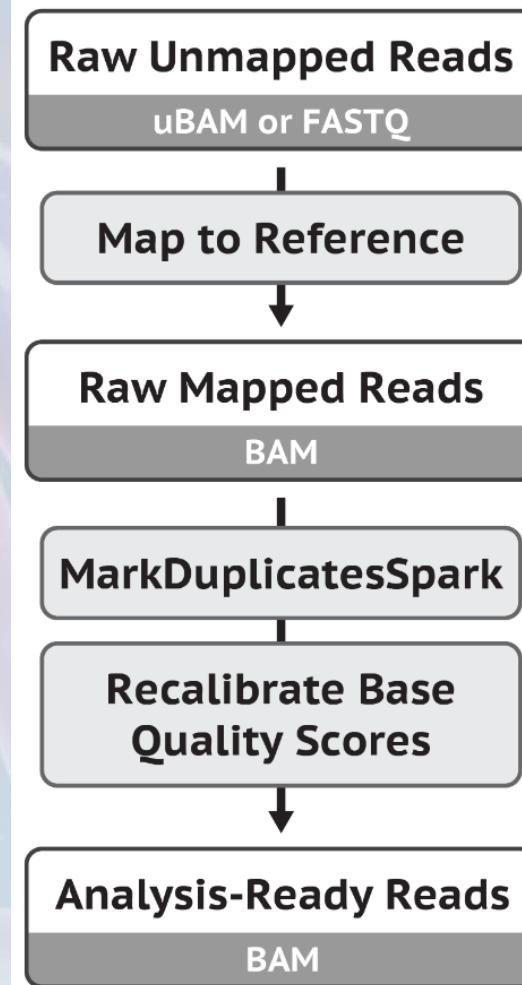
Escrito en **C**

Formato SAM/BAM

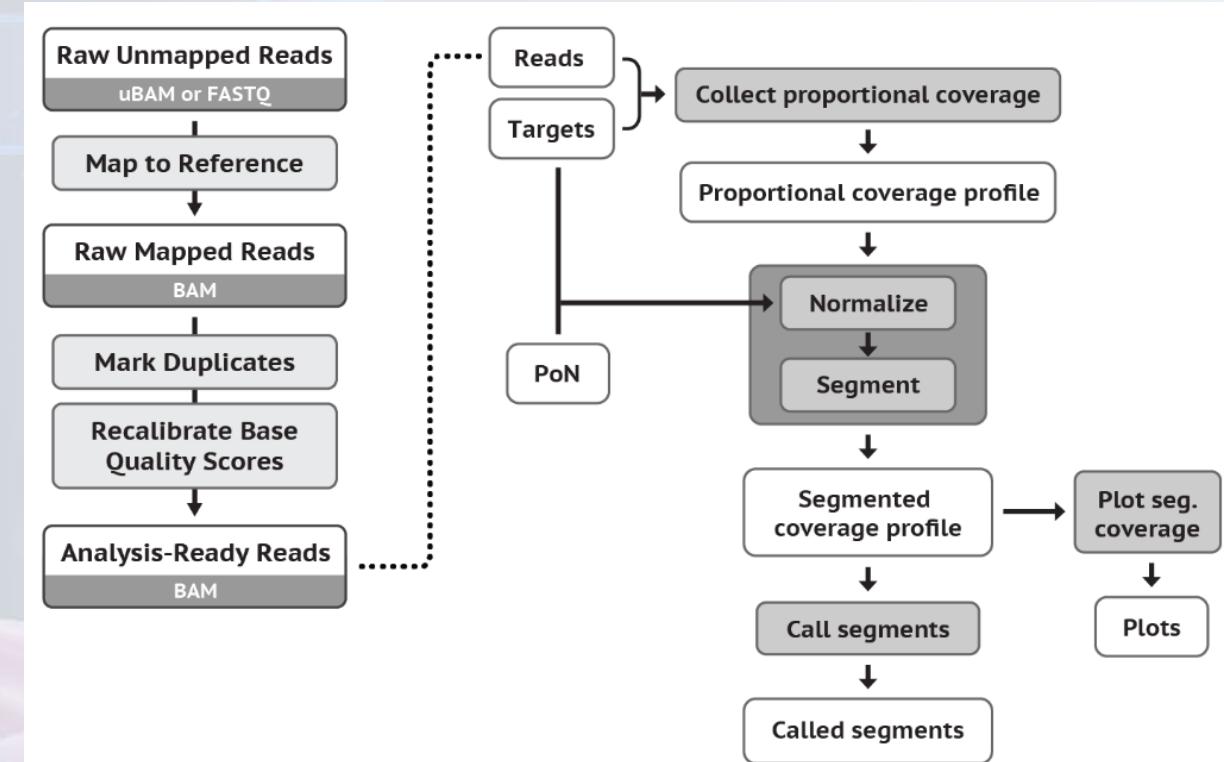


SAMtools

Archivo .VCF



Data pre-processing for variant discovery



Somatic copy number variant discovery (CNVs)

ANNOVAR

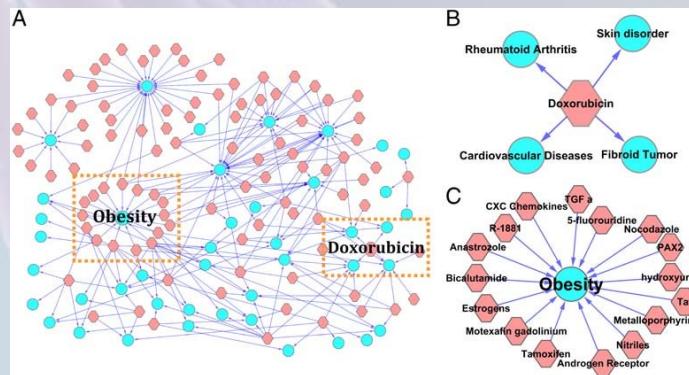
SIFT
Polyphen-2
CADD
Condel

SnpEff & SnpSift

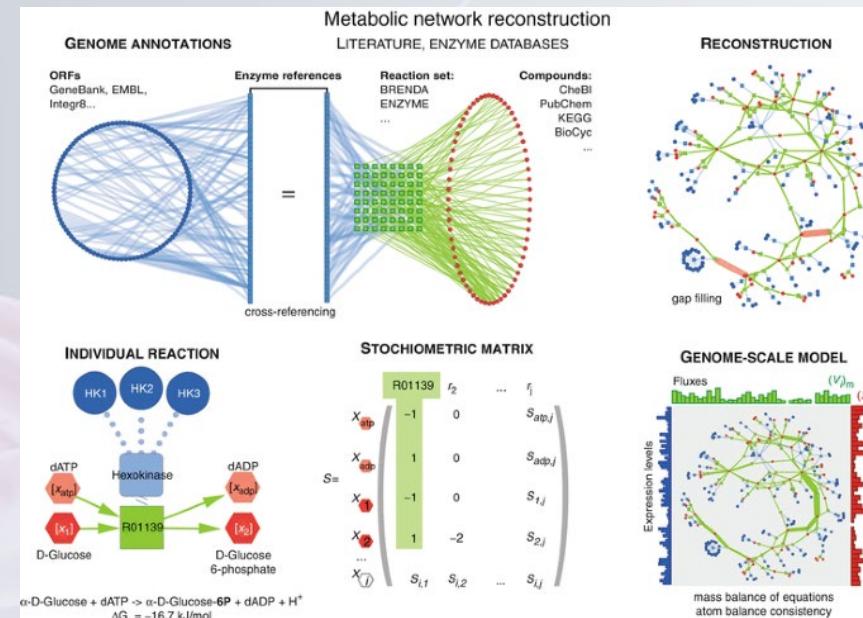
Ensembl Variant Effect Predictor (VEP)

Análisis terciarios

Mapa de conectividad



Reconstrucción metabólica



Ontologías génicas



Una pregunta de interés



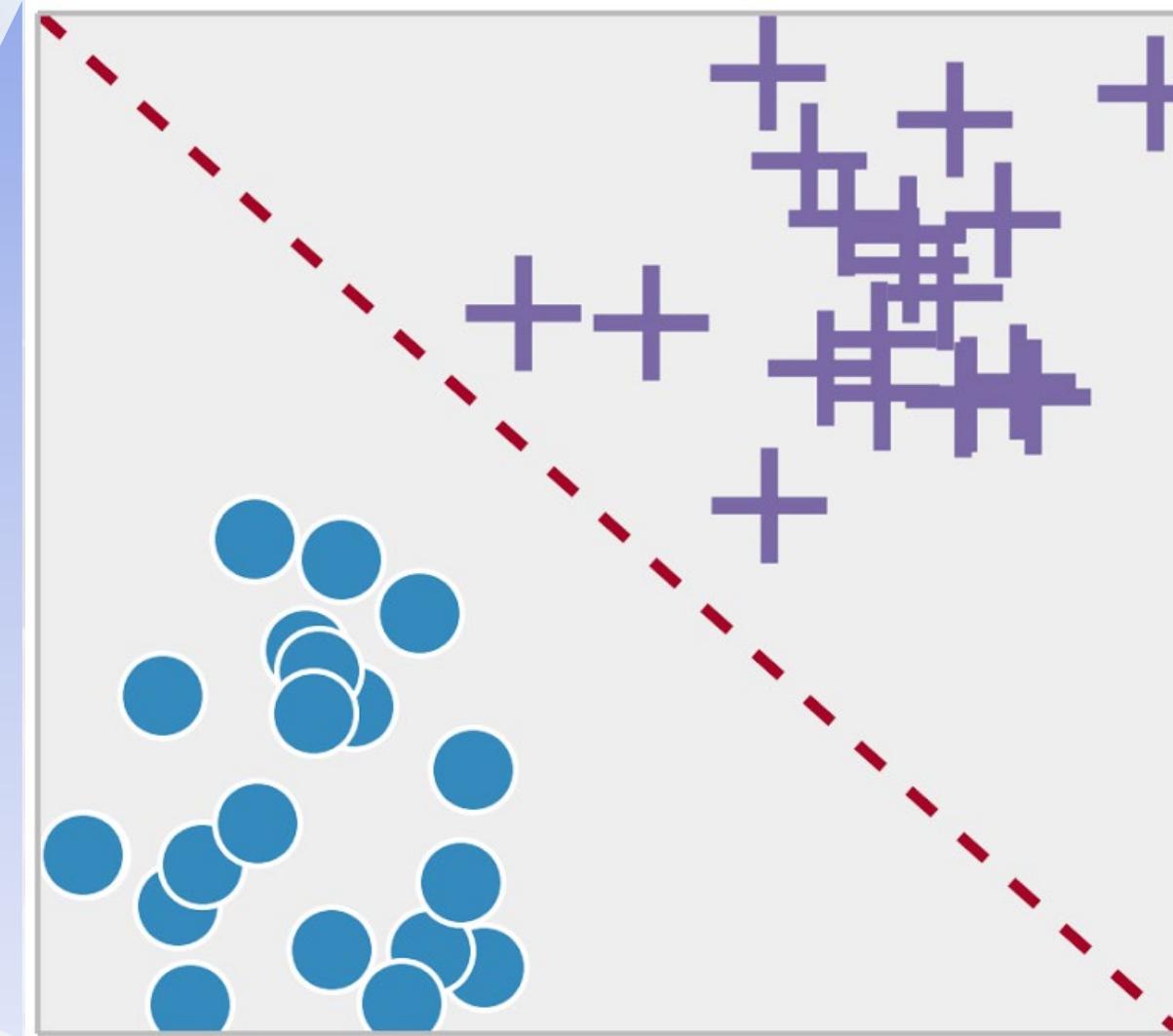
¿Sería posible establecer una correlación entre la respuesta transcripcional de los Macrófagos de individuos que sufrieron Chagas de transmisión oral y la severidad de la enfermedad?

Colaboración: Dr. Juan Luis Cabrera/Esther Gutiérrez

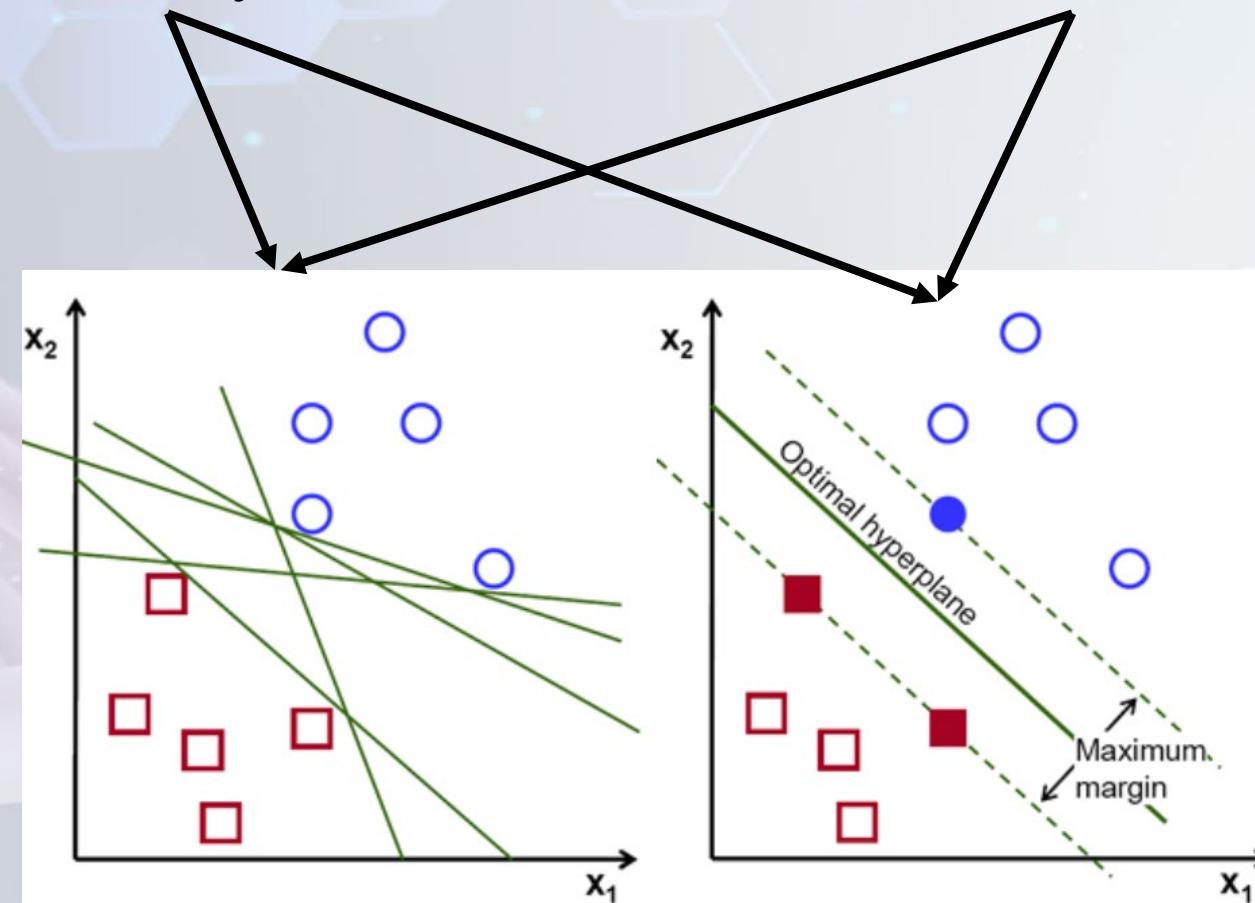
Herramientas de aprendizaje automático

k-NN, SVM, redes neurales,
reglas de clasificación,
regresión lineal, etc

Datos masivos



GDE Pacientes y Controles Datos clínicos de severidad



| Grado | N° |
|---|-----------|
| Asintomáticos (Grupo 0) | 5 |
| Fiebre, dolor de cabeza, taquicardia, decaimiento (Grupo 1). | 14 |
| Hospitalización, miocarditis, derrame pericárdico (Grupo 2) | 11 |
| Controles (Base) | 8 |



Entrada a la MSV



**UNU
BIOLAC**

Grado 0

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

Grado 1

Grado 2

Control

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32

Aleatorización

60/40

Entrenamiento

1 2 4 11 24
22 6 3 23 30
29 18 27 32 25
26
14 7 5

Prueba

16 15 8
19 21 31 12
17 20
10 13 9 28

| | | Referencia | | |
|----------------|----------|------------|----------|----------|
| Predictión | Grado 0 | Grado 1 | Grado 2 | Base |
| Grado 0 | 1 | 0 | 0 | 0 |
| Grado 1 | 1 | 3 | 2 | 0 |
| Grado 2 | 0 | 0 | 3 | 0 |
| Base | 0 | 0 | 0 | 3 |

76% de precisión

Clasificación necesita más información

**1**

Datos de Microarreglos/NGS

2

Datos clínicos

3

Predicción(Predictor)

Sintomatología

Controles médicos

Exámenes de laboratorio

El laboratorio de la Dra. Petsalaki tiene amplia experiencia en la utilización de algoritmos de inferencia en redes biológicas

Evangelia Petsalaki



Group Leader - Petsalaki research group

Evangelia Petsalaki's research group studies human cell signalling in healthy and disease conditions. The group uses interdisciplinary approaches, including data-driven network inference, modelling of cell processes and data integration, to understand how different environmental or genetic conditions affect cell signalling responses leading to diverse cell phenotypes. Evangelia has a PhD in structural bioinformatics from EMBL and the University of Heidelberg (2009) and did her post doctoral work at the Lunenfeld-Tanenbaum Research Institute in Toronto, Canada (2010-2016).

Identification of phenotype-specific networks from paired gene expression-cell shape imaging data

Barker CG , Petsalaki E, Giudice G , Ekpenyong EN, Bakal C, Petsalaki E

Method



molecular
systems
biology

CEN-tools: an integrative platform to identify the contexts of essential genes

Sumana Sharma^{1,2,*†‡} , Cansu Dincer^{1,†} , Paula Weidemüller¹ , Gavin J Wright² & Evangelia Petsalaki^{1,**} 

Cell Systems

Volume 10, Issue 5, 20 May 2020, Pages 384-396.e9



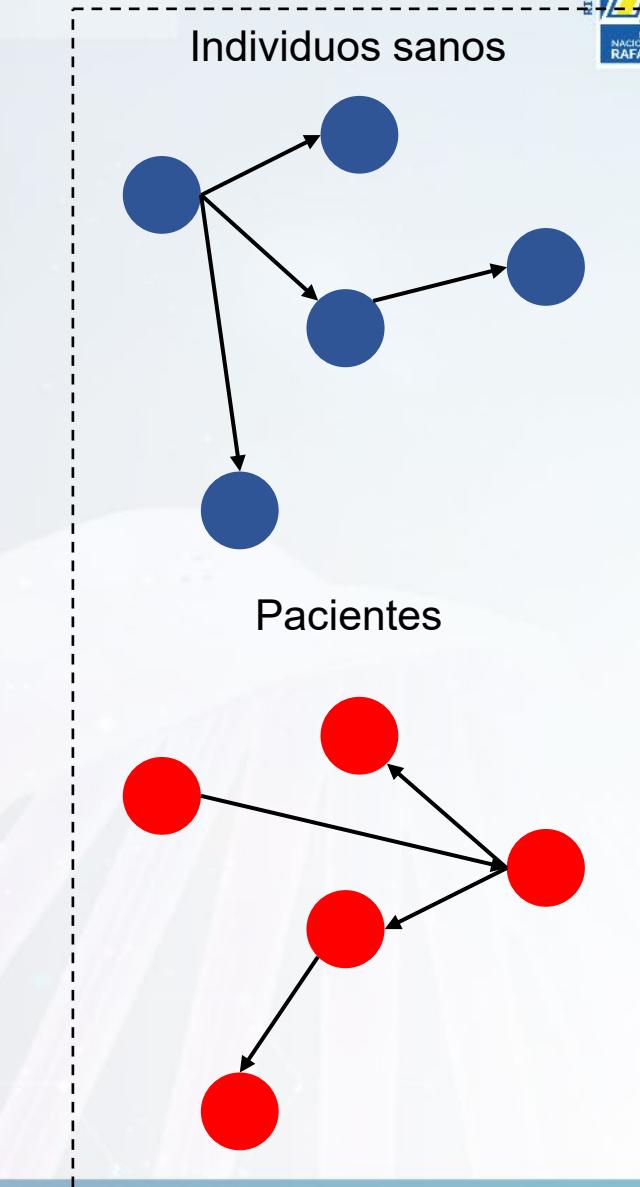
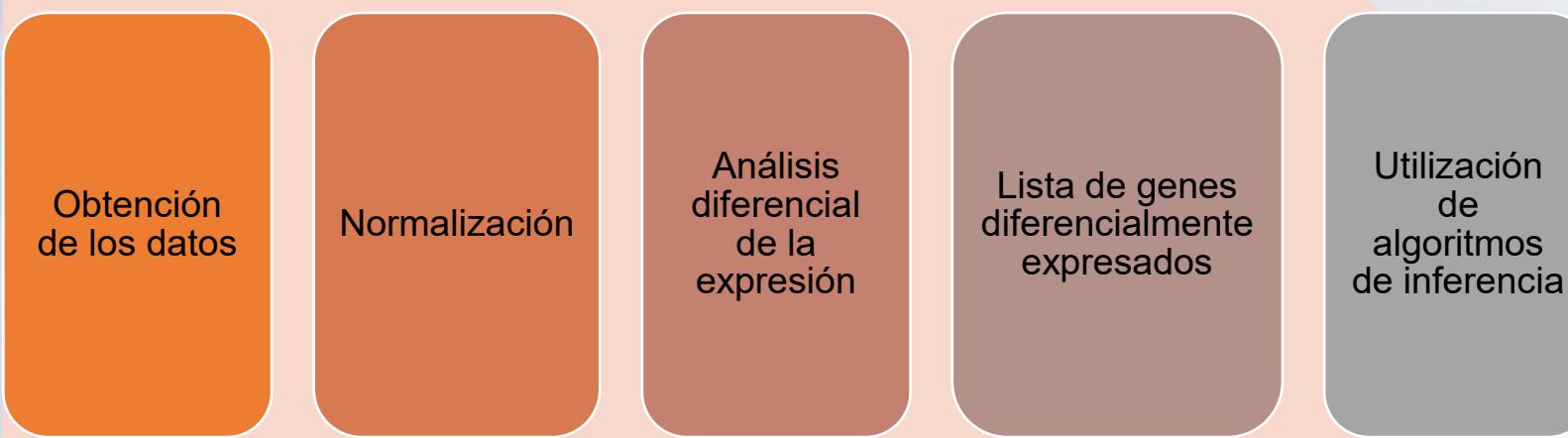
Article

Prediction of Signed Protein Kinase Regulatory Circuits

Brandon M. Invergo^{1, 3, 4}, Borgthor Petursson^{1, 3}, Nosheen Akhtar², David Bradley¹, Girolamo Giudice¹, Maruan Hijazi², Pedro Cutillas²  , Evangelia Petsalaki¹  , Pedro Beltrao^{1, 5}  

Cursos, Seminarios y Talleres en Biología al Servicio de la Salud

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



Trabajo de grado de maestría

Deliana Infante (Bióloga USB)

Mercedes Fernández, Rafael Puche, Carlos Ramírez

Identificación de genes centrales (nodales) y vías de señalización involucrados en la respuesta inmunitaria en la infección de macrófagos con *Trypanosoma cruzi* derivados de pacientes que sufrieron enfermedad de Chagas por transmisión oral.



IMPLEMENTACIÓN DE UNA PLATAFORMA BIOINFORMÁTICA PARA ANÁLISIS DE DATOS GENÓMICOS DE BACTERIAS PATÓGENAS DE IMPORTANCIA EN SALUD PÚBLICA

Puche, Rafael; Hernandez, Fernando; Fernandez-Mestre, Mercedes; Ramírez, Carlos; Esther Gutiérrez; Patricio Yankilevich, Campos, Josefina; Iraola, Gregorio

