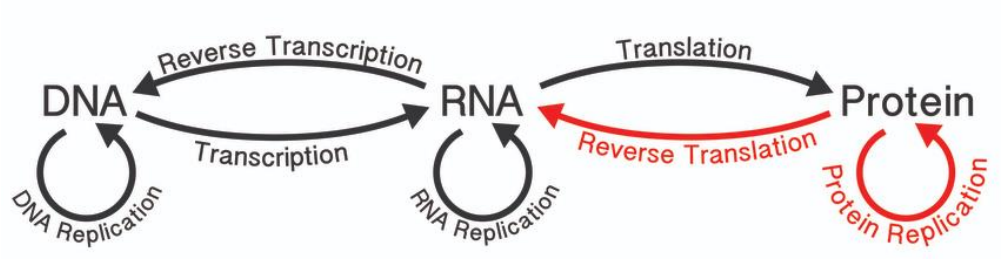
Introducción
Análisis de Secuencias

Guillermo Barturen
gbarturen@ugr.es

Orígenes de la bioinformática

- La bioinformática que hoy conocemos ha surgido en torno a la genómica, jugando un papel clave en su desarrollo
- Pero sus **fundamentos conceptuales** vienen de mucho más atrás: Biología teórica, Biología computacional, Biocomputación…
- La **información** es un concepto fundamental en Biología
- Primero se descubrió su **transmisión**: leyes de Mendel
- El contenido en **información es lo que más distingue a un ser vivo** de algo que no lo es (*Müller, 1916*)
- What is Life? (*Schrödinger, 1943*): **Cristal aperiódico, portador de información**



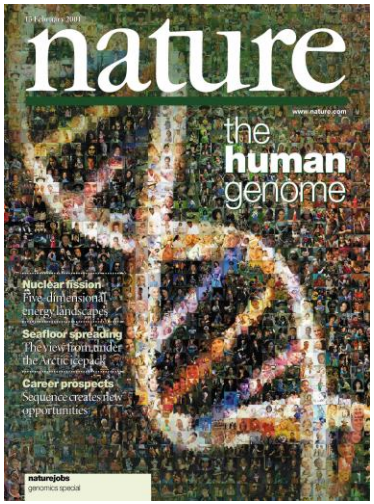
Orígenes de la bioinformática

- Biología Molecular: **moléculas informacionales** (ADN y proteínas): la información biológica es **digital**, no analógica

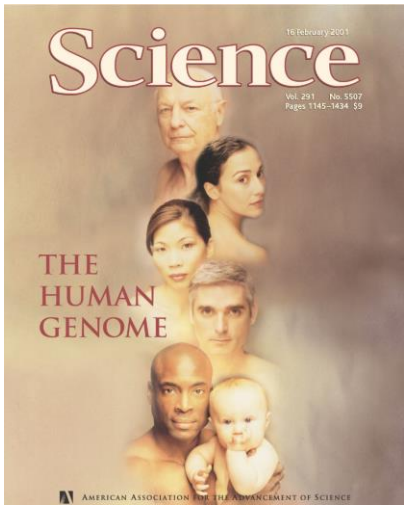
- Un ser vivo no es más que la **realización de un programa genético** (Waddington, Longuet-Higgins, Jacob, Luria…)

- **Secuenciación de ADN y proteínas**: el estudio del contenido en información de los seres vivos se convirtió en un tópico del mayor interés

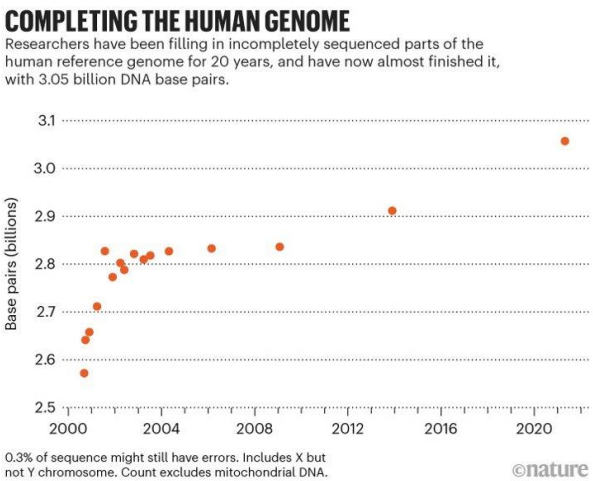
- El **Proyecto Genoma** pronto produjo una acumulación explosiva de secuencias que reclamó la atención prioritaria de los bioinformáticos…



Primer draft del genoma humano 2003



El genoma humano en la actualidad



- Último ensamblado GRCh38, 2013
- Genoma completo T2T-CHM13, 2022
- Cromosoma Y, 2023

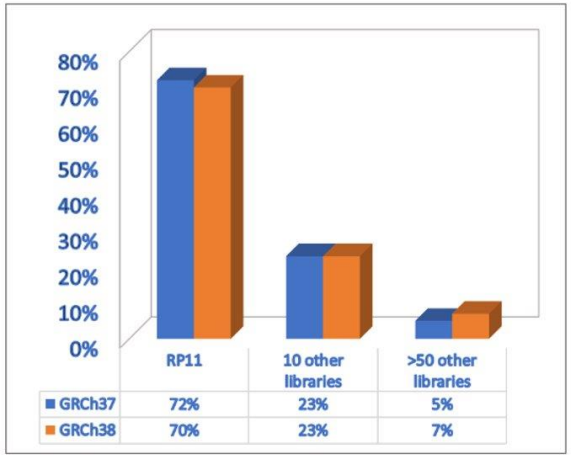


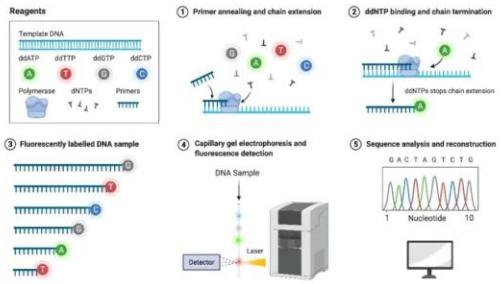
Figure 1. Contribution of genomic libraries to GRCh37 and GRCh38.

Referencia actual: 20 individuos y 70% de la secuencia de uno solo

Objetivo: ~350 individuos (700 haplotipos de altísima calidad) que refleje la diversidad de las 26 poblaciones incluidas en el proyecto 1000 genomas

Desarrollo explosivo de la secuenciación masiva

Secuenciación Sanger



Illumina / SOLiD

PacBio / Nanopore

	First generation	Second generation	Third generation
Fundamental technology	Size-separation of specifically end-labeled DNA fragments	Wash-and-scan SBS	Single molecule real time sequencing
Resolution	Averaged across many copies of the DNA molecule	Averaged across many copies of the DNA molecule	Single DNA molecule
Current raw read accuracy	High	High	Lower
Current read length	Moderate (800-1000 bp)	Short (generally much shorter than Sanger sequencing)	> 1000 bp
Current throughput	Low	High	High
Current cost	High cost per base, Low cost per run	Low cost per base, High cost per run	Low cost per base, High cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing
Time to result	Hours	Days	< 1 day
Sample preparation	Moderately complex, PCR amplification is not required	Complex, PCR amplification is required	Various
Data analysis	Routine	Complex (due to large data volumes & short reads)	Complex
Primary results	Base calls with quality values	Base calls with quality values	Base calls with quality values

Adapted from Schadt, et al. Hum Mol Genet 2010[13]

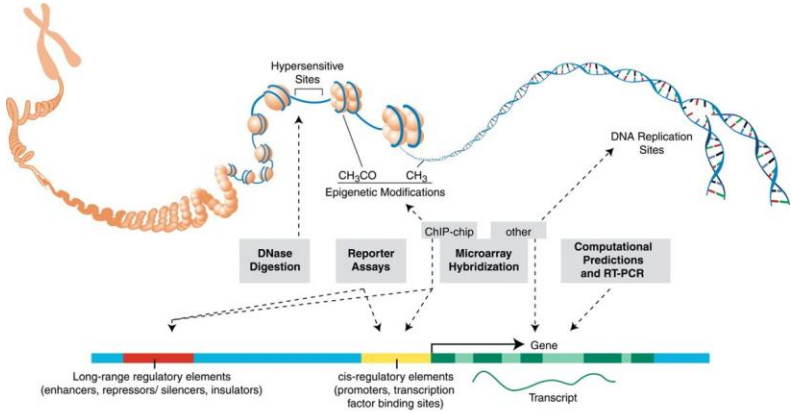
Proyecto 1000 genomas

- 2506 individuos secuenciados de 26 poblaciones en 2015
- 88 millones de variantes: 84.7 millones de SNPs, 2.6 millones de insercions/deleciones cortas y 60,000 variantes estructurales
- 4-5 millones de variantes de media entre un genoma individual y la referencia

- La mayor parte de las variantes genéticas se localizan en las regiones no-codificadoras del genoma (elementos reguladores)
- Cada ser humano lleva por término medio 200-300 variantes de pérdida de función (LOF)
- Una variante cada 20 bp!

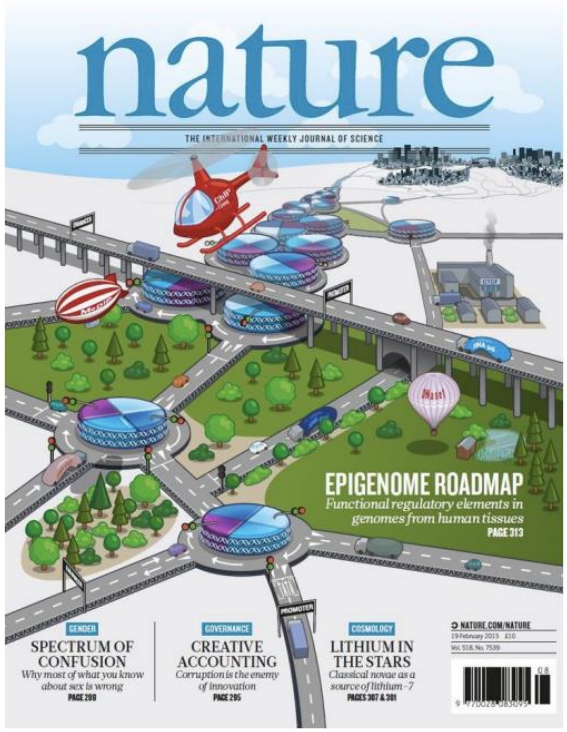
Proyecto ENCODE

- El objetivo de ENCODE (2003) es elaborar un catálogo exhaustivo de todos los elementos funcionales en el genoma humano
- La suma de los exones de los aprox. 21.000 genes humanos no llegan al 2% del genoma ¿para que sirve el 98% restante? ¿es ADN basura?

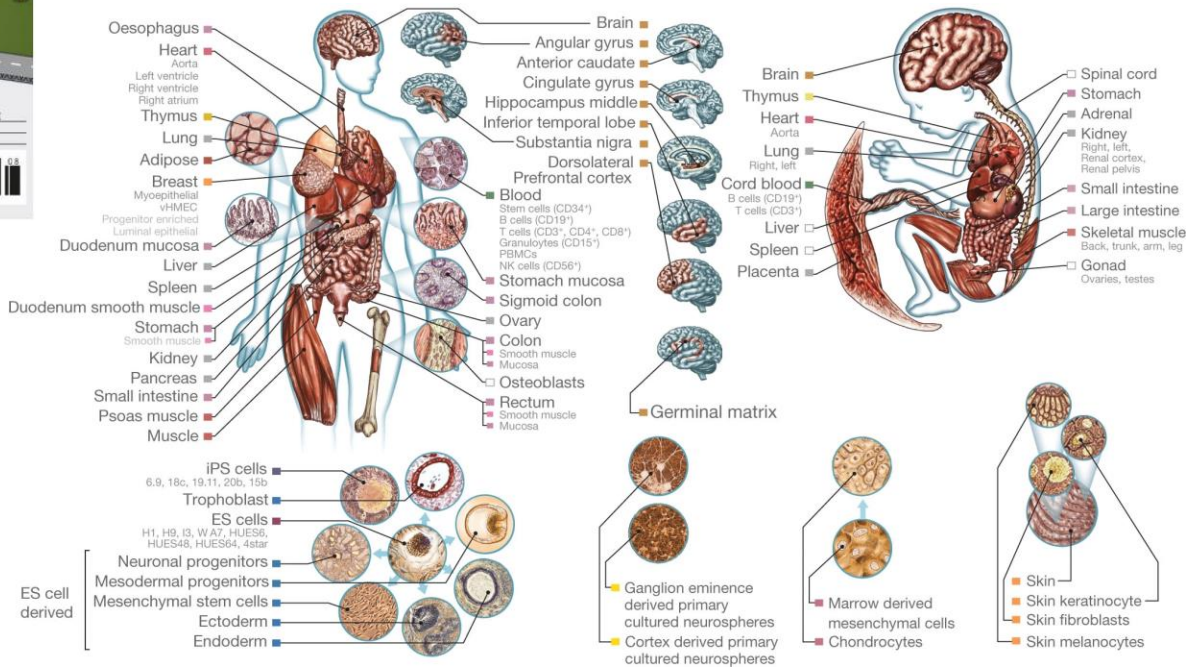


- La mayor parte del genoma (80.4%) se puede asociar con al menos una función en alguno de los 147 tipos celulares analizados. Puesto que puede haber hasta 2.000 tipos celulares, este porcentaje podría llegar a ser mucho más alto!
- Muchas de los elementos funcionales encontrados se localizan en las regiones no-codificadoras (fuera de los genes).
- Muchas enfermedades se asocian con un determinado factor de transcripción que varía entre tipos celulares.

ROADMAP EPIGENOMICS

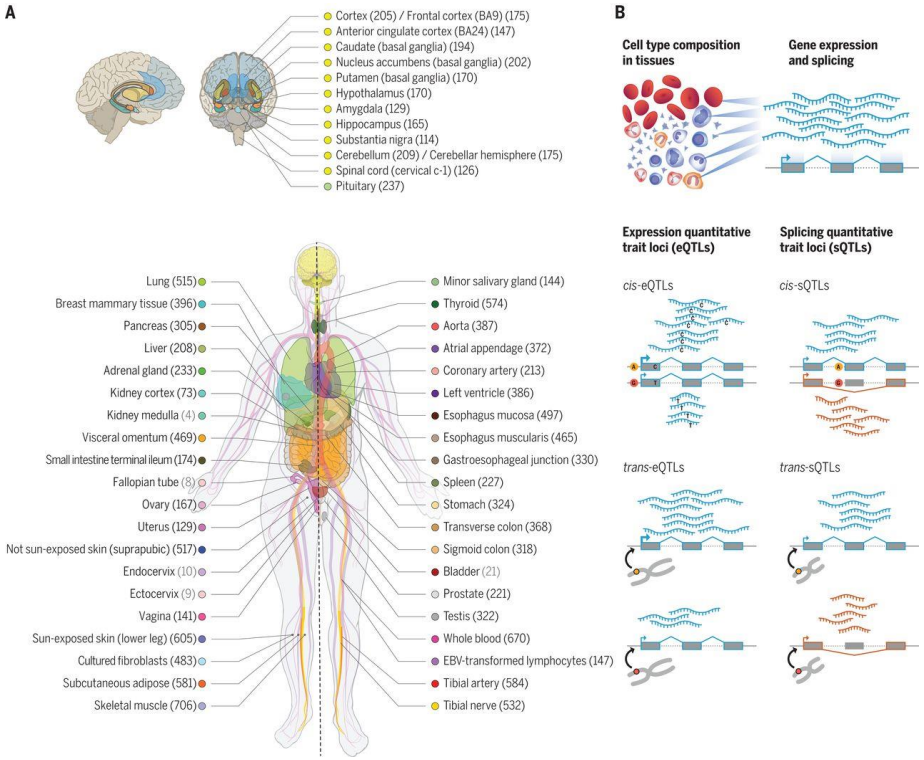


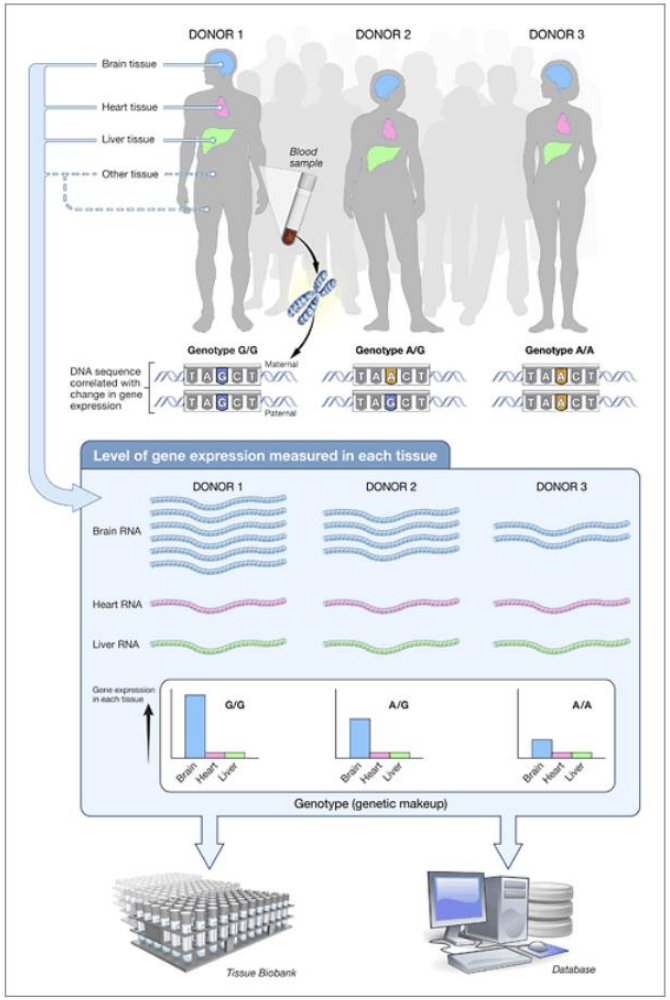
Análisis integrado de 111 epigenomas humanos



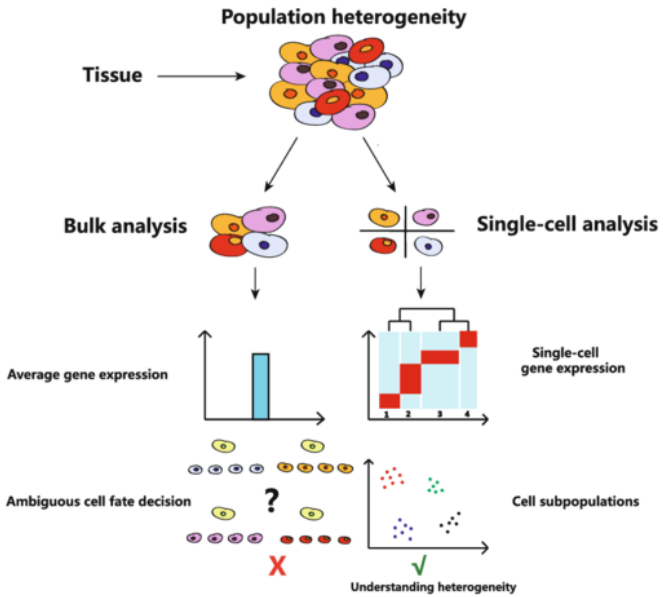
Proyecto GTEx

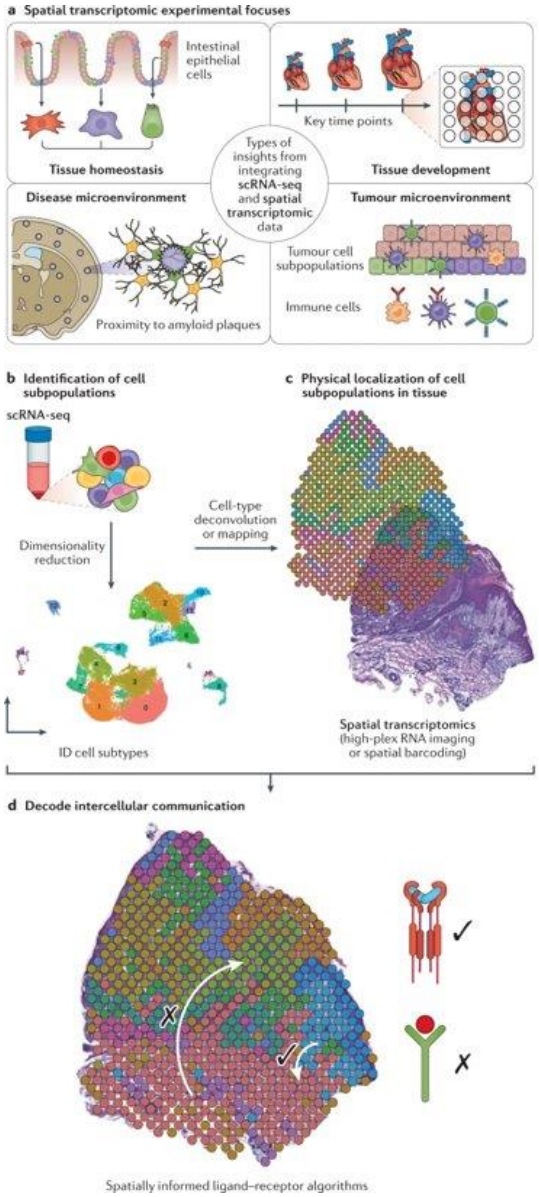
Estudio de la regulación génica en la expresión de los genes





Análisis de célula única y espacial

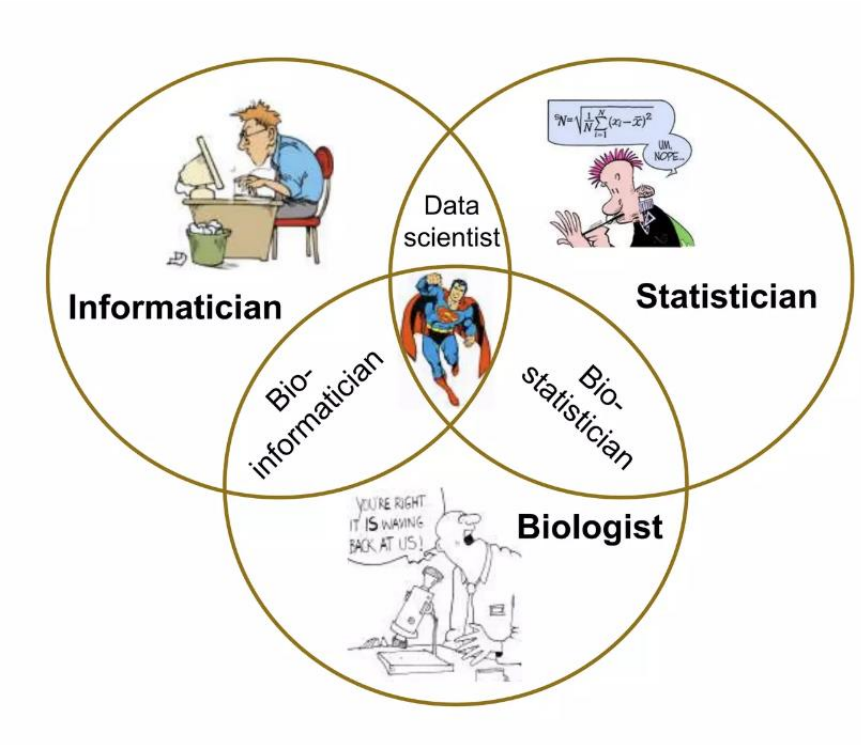




- Las últimas tecnologías nos permiten obtener información individual de cada célula incluyendo transcriptoma, proteómica y accesibilidad.
- Además, podemos incluso recuperar la información espacial de las células en el tejido de origen.

¿Qué es un bioinformático?

La bioinformática actualmente es un campo muy amplio que engloba múltiples perfiles como análisis de datos ómicos, bioinformática estructural, desarrollador de algoritmos...



Pero todos ellos necesitan una serie de aptitudes y conocimientos muy diversos: programación, estadística, conocimiento de bases de datos y algoritmos y por supuesto un buen conocimiento de los procesos biológicos.

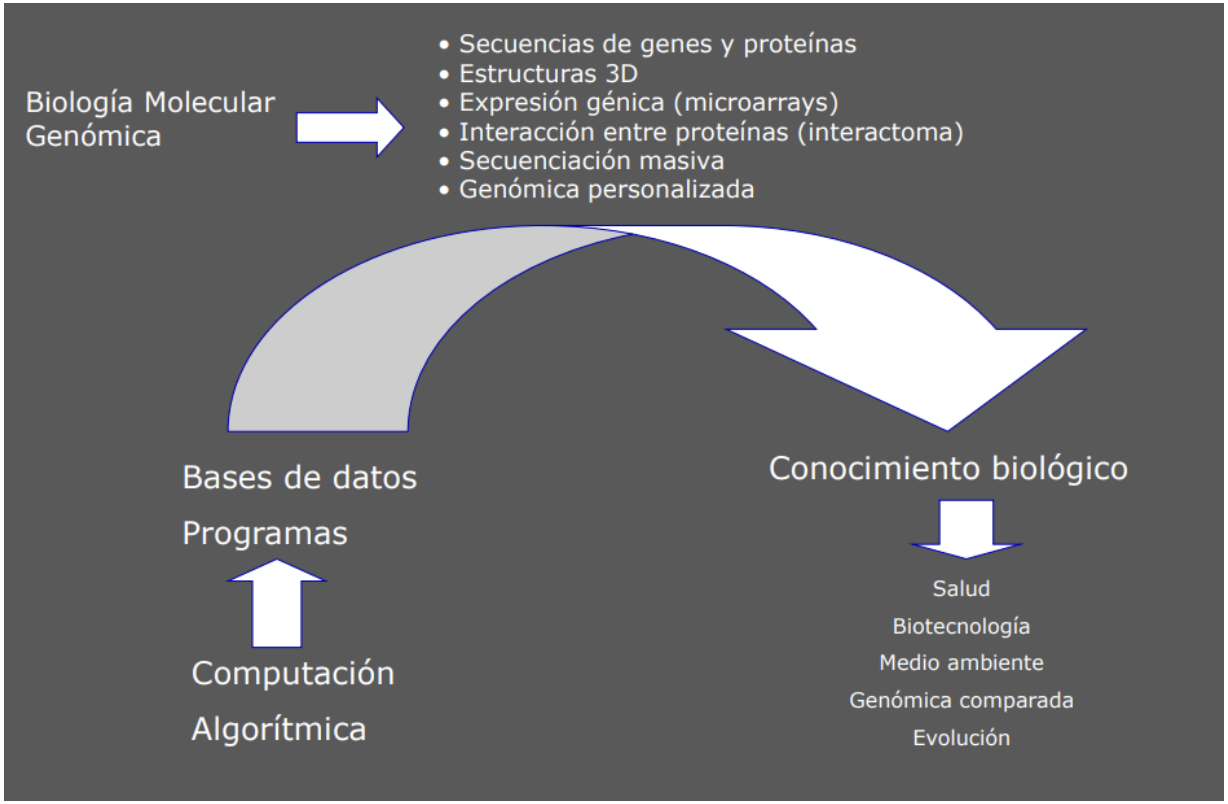
¿Qué es la vida?

- La característica fundamental de la vida es el **flujo de información a lo largo del tiempo**: transcripción, traducción, código genético, sensores…

- Pero ¿qué es la información?
- ¿Puede tener efectos causales?
- ¿Se conserva, como la materia y la energía?
- ¿Aumenta o disminuye a lo largo del proceso evolutivo?

- El origen de la vida o el de la consciencia, no podrán resolverse sin responder esas preguntas. Y en ello la **bioinformática está llamada a jugar un papel clave.**

La información (la vida) está en las secuencias



¿Dónde se almacenan las secuencias?

EMBL-EBI

National Center for Biotechnology Information
NCBI

UniProt