

# **Intro to bioinformatics, foundations and applications**

**BioData 2025**



Delivering healthier lives through  
innovation in gut health,  
microbiology and food

[Read more about the Quadram Institute](#)

Science ◀ Health ◀ Food ◀ Innovation

# From pipettes to metagenomes

**Bachelor and Msc in microbiology**

**2014-17: PhD in molecular biology**

I2BC, CEA Saclay



Thesis: Redox control of protein secretion in the yeast *S. cereviseae*

# From pipettes to metagenomes

**Bachelor and Msc in microbiology**

**2014-17: PhD in molecular biology**

I2BC, CEA Saclay



Thesis: Redox control of protein secretion in the yeast *S. cereviseae*



**2016-17: Msc in Computer science**

→ Transition from a purely biological background to bioinformatics



# From pipettes to metagenomes



**2017-20: Postdoctoral fellow in computational biology for metagenomics**  
Hurwitz lab, University of Arizona, USA

**2019-20: Data Science Fellow**  
Data science Institute, University of Arizona, USA

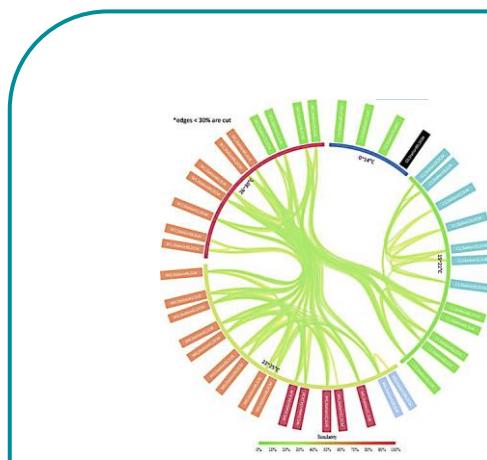


**iMicrobe**

**Cyberinfrastructures**  
to help the open  
sharing and analysis  
of microbiome data

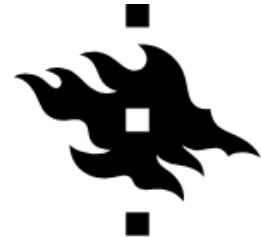


Inclusive **communities** that  
facilitate collaboration  
between domain and data  
scientists



**Content-based analytics** for  
comparative metagenomics  
and  
sequence classification

# From pipettes to metagenomes



HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI

# 2021-24: Postdoctoral fellow in computational biology for metagenomics Salonen lab



# Impact of the **early-life gut microbiome** **and virome** on infant health and wellbeing



## Data curation and stewardship for large-scale metagenomic projects

# From pipettes to metagenomes

## MOLECULAR BIOLOGY

**Sample size:**

10s to 100s of samples

**Daily work:**

- Molecular biology
- Culturing & Microbiology
- Lab organization

**Timeline:**

Weeks to months

## BIOINFORMATICS

**Sample size:**

1,000s of samples

**Daily work:**

- Writing code & development
- Statistical analysis
- Data stewardship

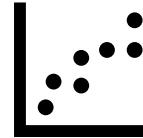
**Timeline:**

Hours to weeks

# Overview workshop



What is Bioinformatics?



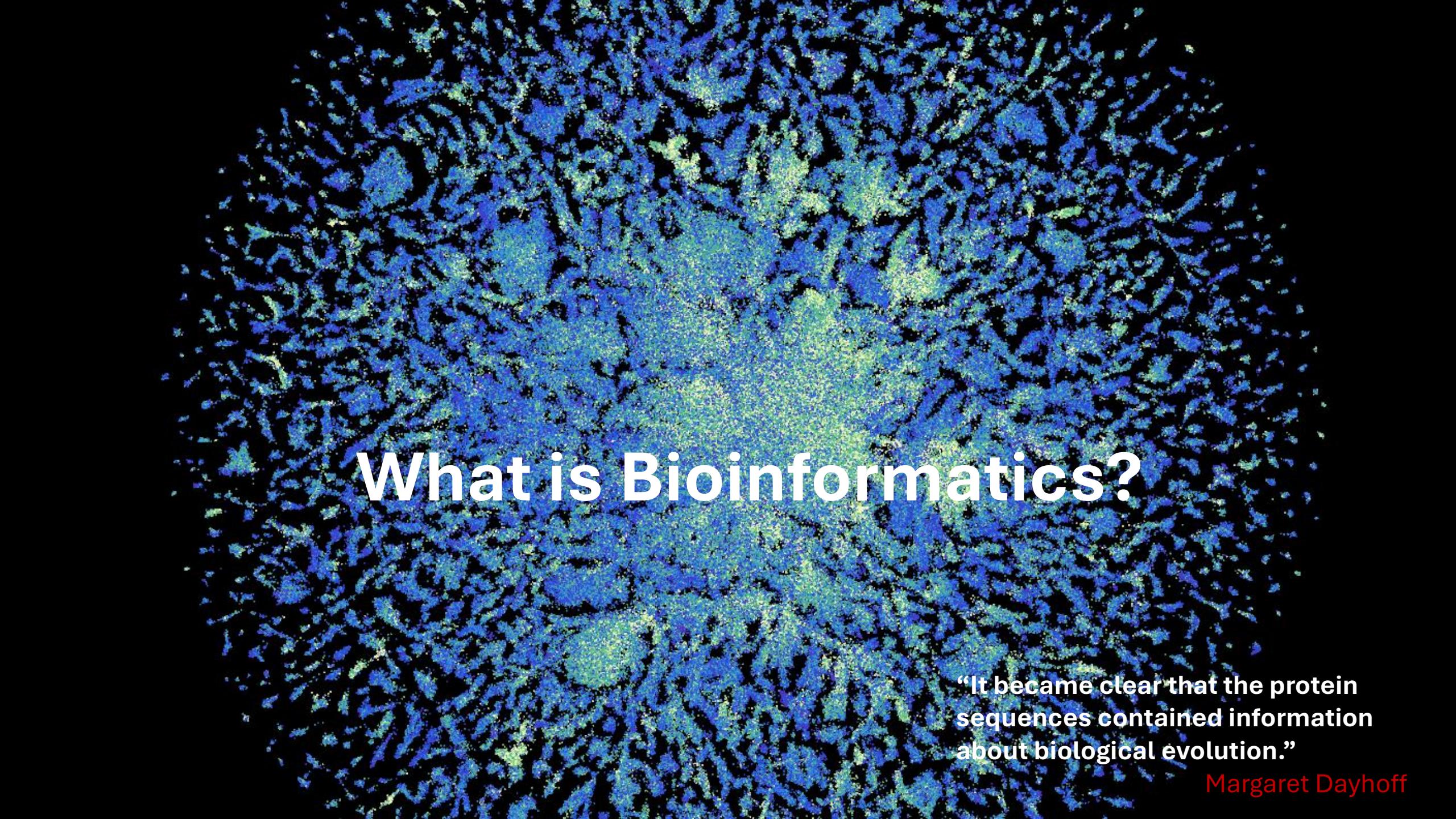
What is the difference between bioinformatics and biostatistics?



What is the history of bioinformatics and why NGS are central to bioinformatics?



What are the role and career paths for bioinformaticians?



# What is Bioinformatics?

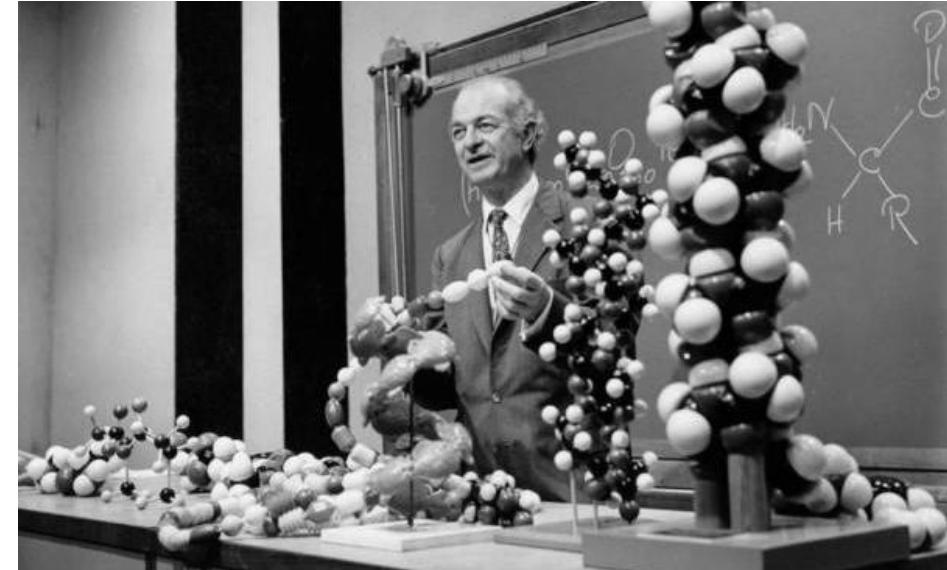
**“It became clear that the protein sequences contained information about biological evolution.”**

Margaret Dayhoff

# The birth of Bioinformatics

## 1960s: The Problem

- Protein sequences being discovered (insulin, hemoglobin...)
- Each lab keeps own records
- No way to compare sequences
- Evolutionary relationships

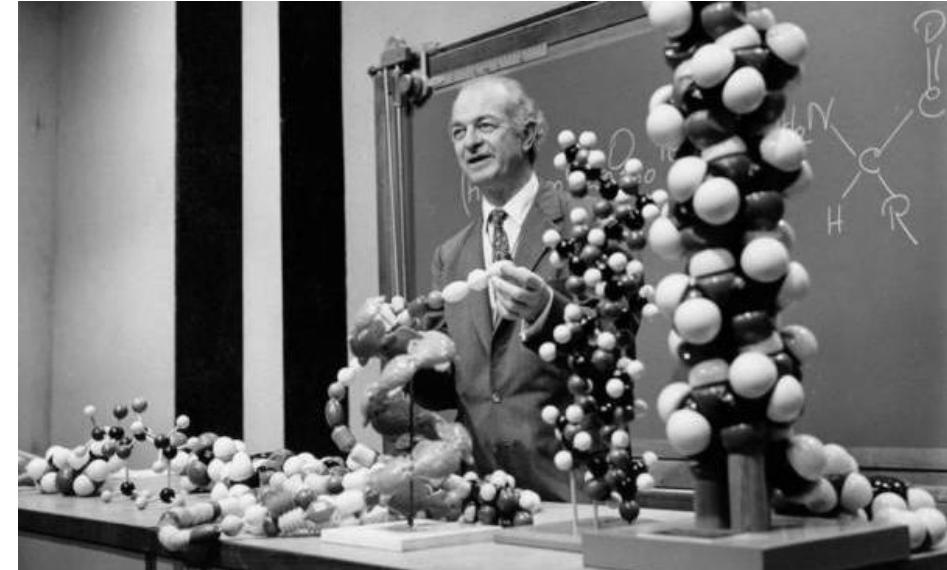


<https://scarc.library.oregonstate.edu/>

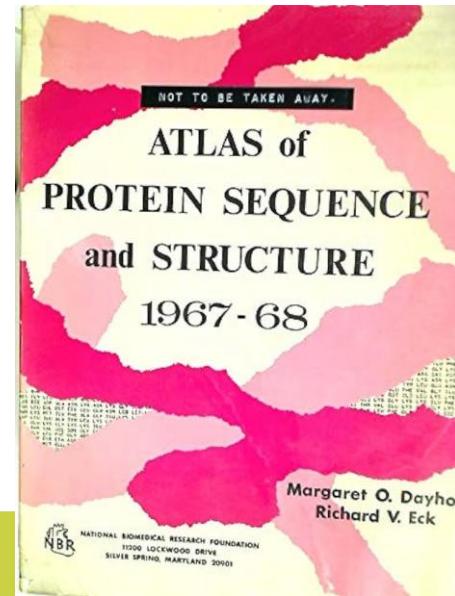
# The birth of Bioinformatics

## 1960s: The Problem

- Protein sequences being discovered (insulin, hemoglobin...)
- Each lab keeps own records
- No way to compare sequences
- Evolutionary relationships



<https://scarc.library.oregonstate.edu/>



<https://en.wikipedia.org/>

Margaret Dayhoff et al

**Enter Margaret Dayhoff :**  
**ATLAS OF PROTEIN SEQUENCE AND STRUCTURE**

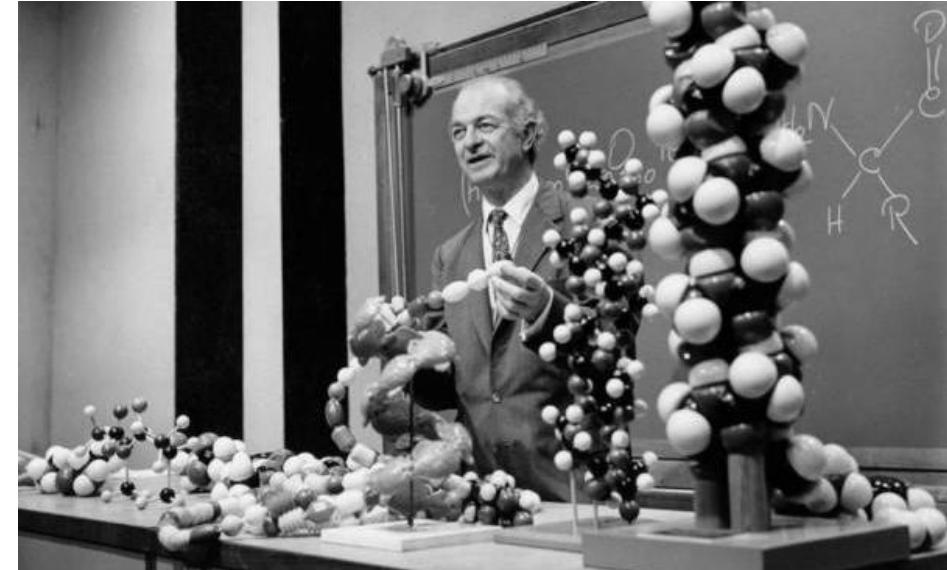
**What made it revolutionary:**

- ✓ First centralized sequence database
- ✓ Sequences in standardized format (one-letter aa codes)
- ✓ Updated annually with community contributions

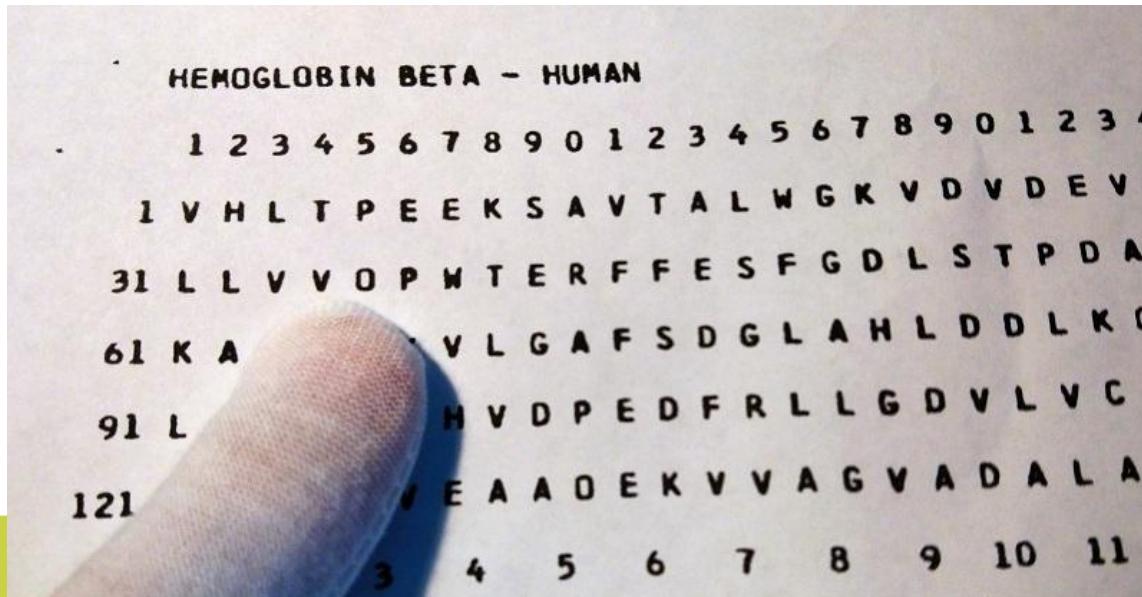
# The birth of Bioinformatics

## 1960s: The Problem

- Protein sequences being discovered (insulin, hemoglobin...)
- Each lab keeps own records
- No way to compare sequences
- Evolutionary relationships



<https://scarc.library.oregonstate.edu/>



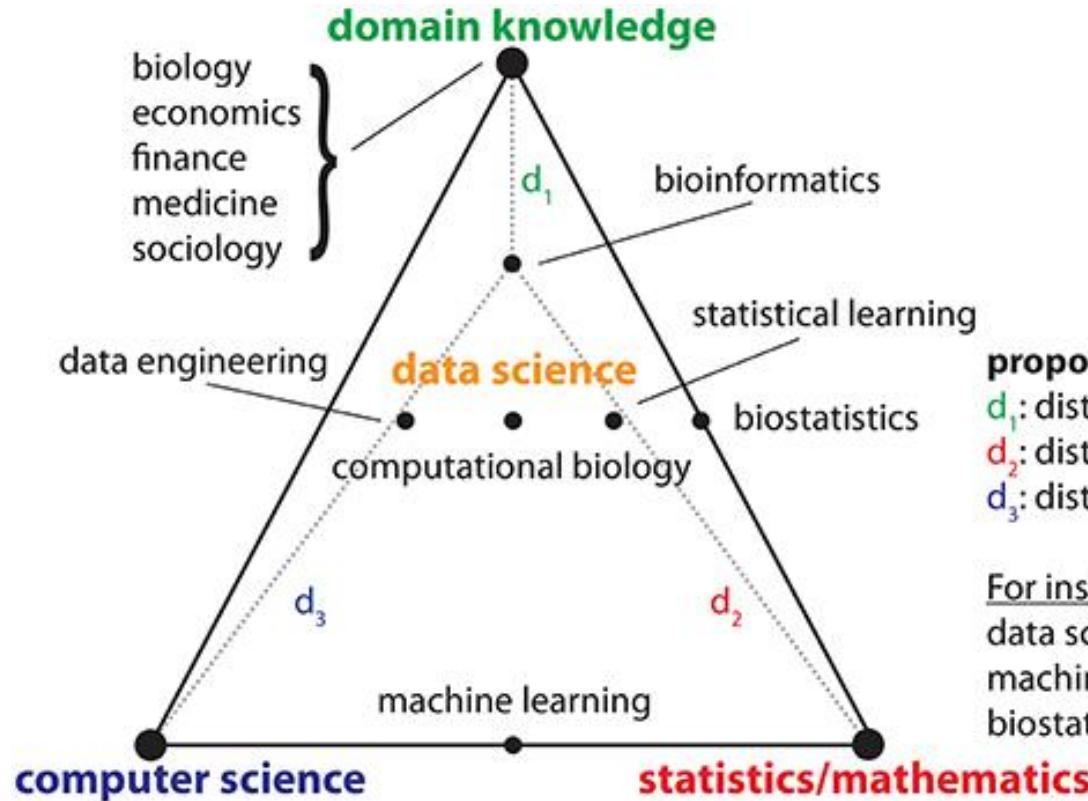
Margaret Dayhoff et al

**Enter Margaret Dayhoff :**  
ATLAS OF PROTEIN SEQUENCE AND STRUCTURE

**What made it revolutionary:**

- ✓ First centralized sequence database
- ✓ Sequences in standardized format (one-letter aa codes)
- ✓ Updated annually with community contributions

# What Makes Bioinformatics Different from biostatistics?



**proportion vector:**  $(d_1, d_2, d_3)$

$d_1$ : distance to 'domain knowledge'

$d_2$ : distance to 'statistics/mathematics'

$d_3$ : distance to 'computer science'

For instance:

data science:  $(1/3, 1/3, 1/3)$

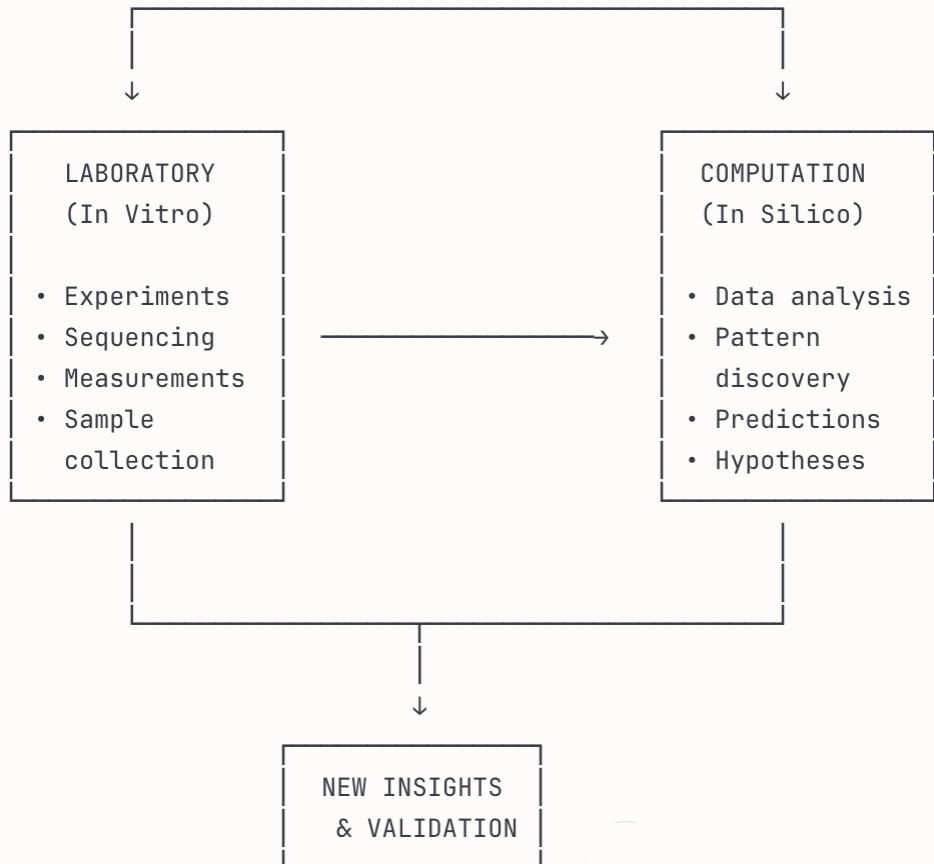
machine learning:  $(0, 1/2, 1/2)$

biostatistics:  $(1/2, 1/2, 0)$

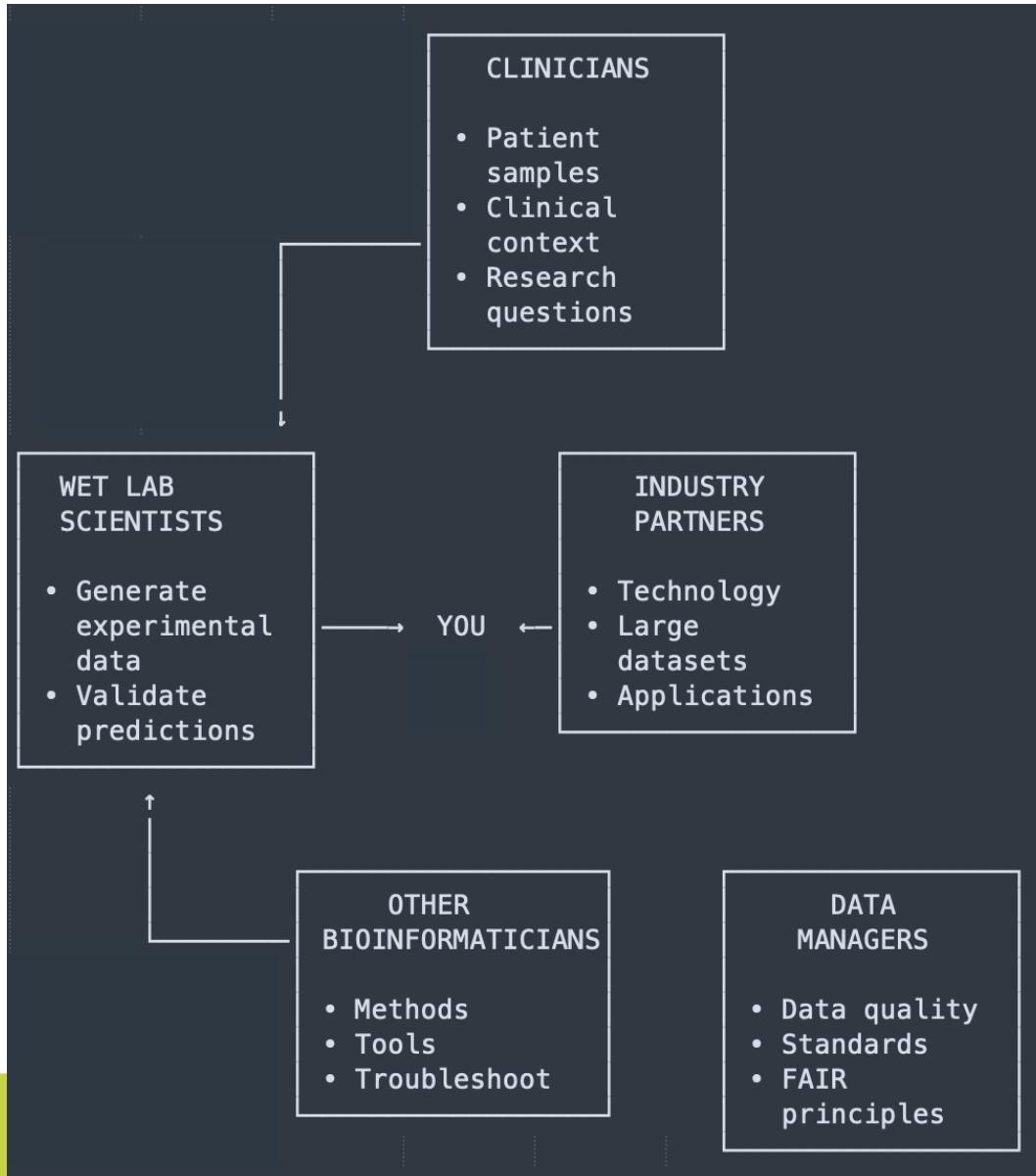
# How do Bioinformatics fit into research?

NATURAL WORLD  
(In Vivo)

- Clinical questions
- Ecological observations
- Real-world phenomena



# Bioinformatics is a team sport



## My first questions with any dataset:

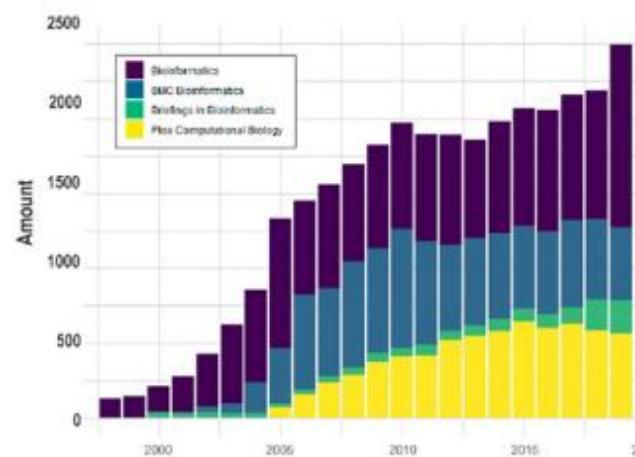
- What question are we trying to answer?
- How was the data generated?
- What potential biases should I take into account in the analysis...

This means having a clear communication with multiple partners & stakeholders

# Bioinformatics is a team sport

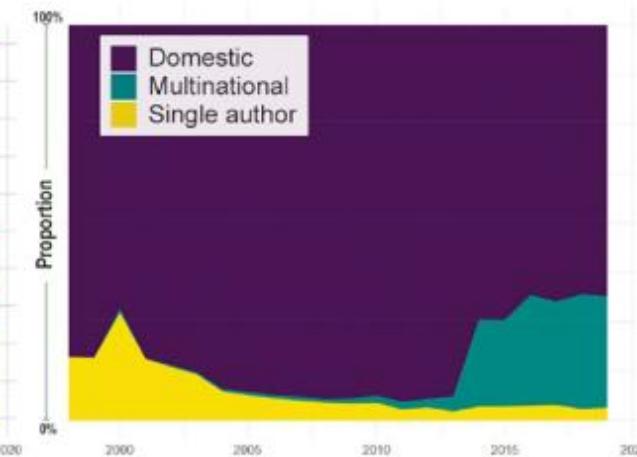
A

Publications by journal



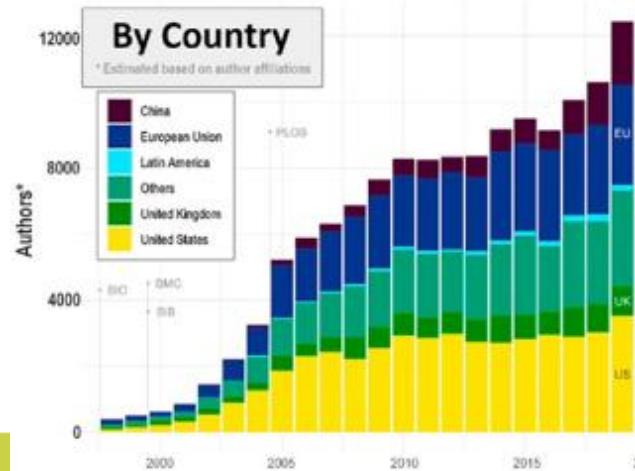
B

International Collaborations



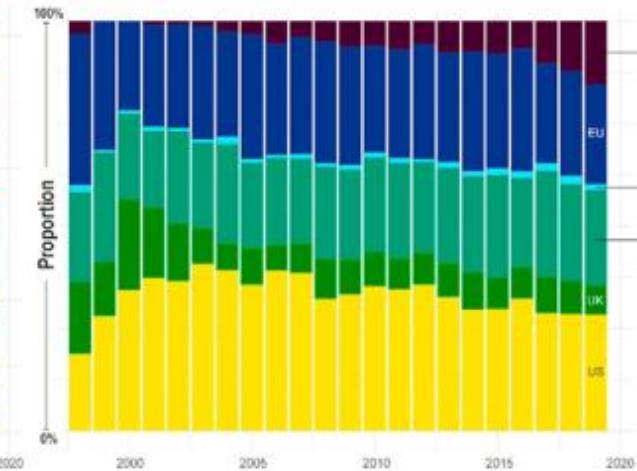
C

Authors by nationality



D

Authors by nationality (%)

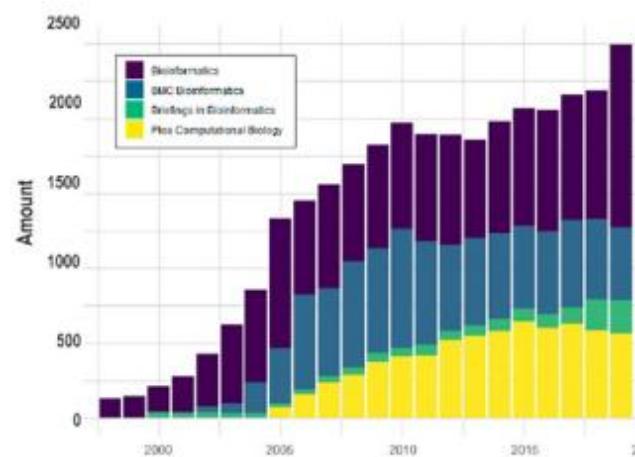


A Brief History of Bioinformatics Told by Data Visualization

# Bioinformatics is a team sport

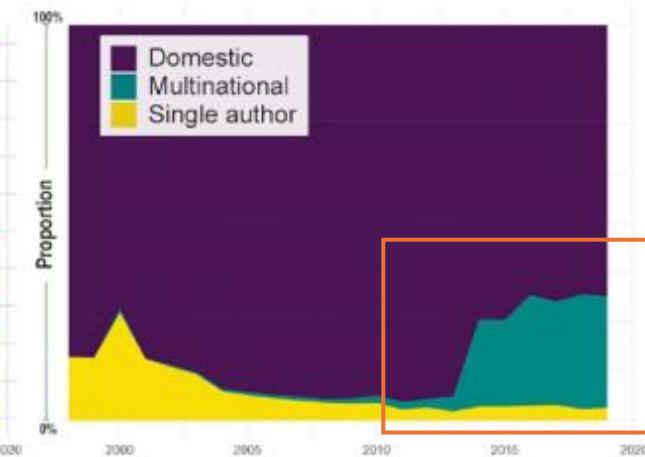
A

Publications by journal



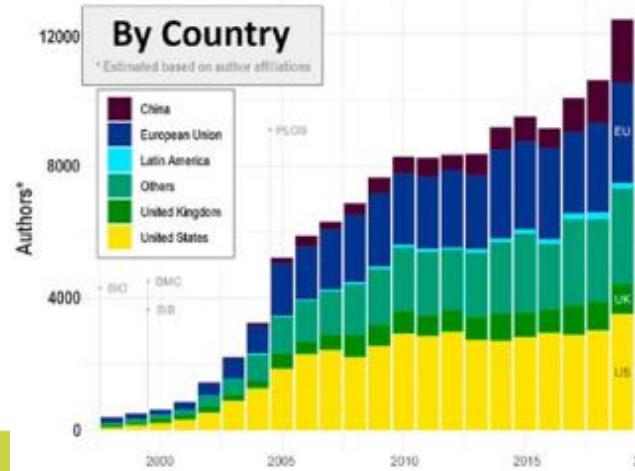
B

International Collaborations



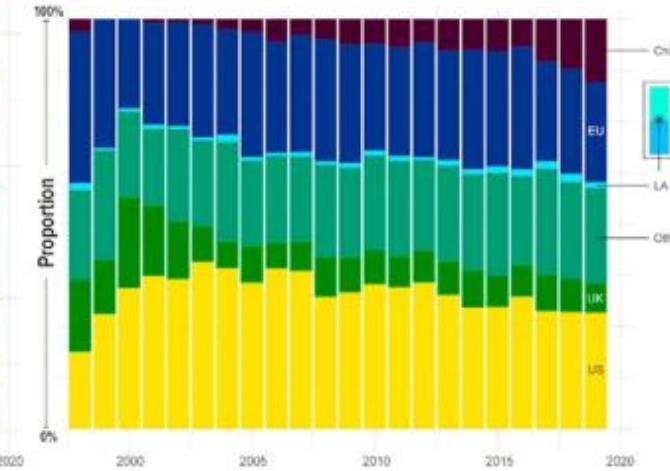
C

Authors by nationality



D

Authors by nationality (%)

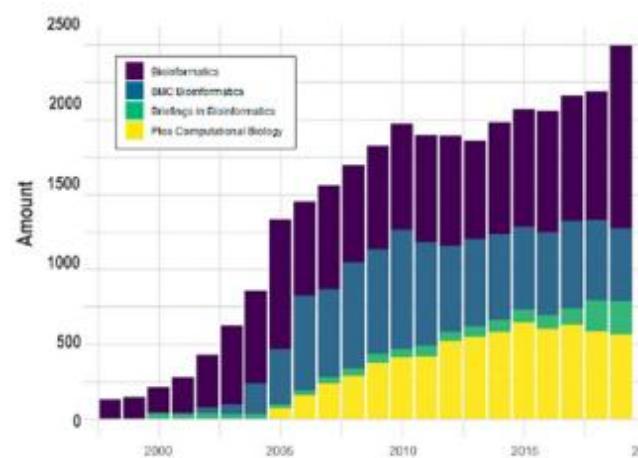


Bioinformatic projects more and more rely on international large-scale collaborations

# Bioinformatics is a team sport

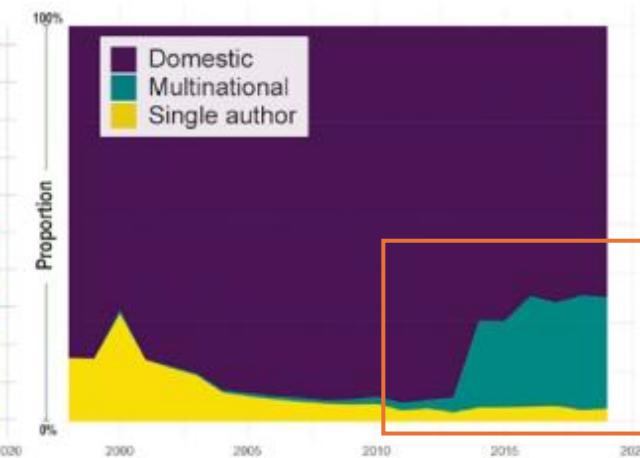
A

Publications by journal



B

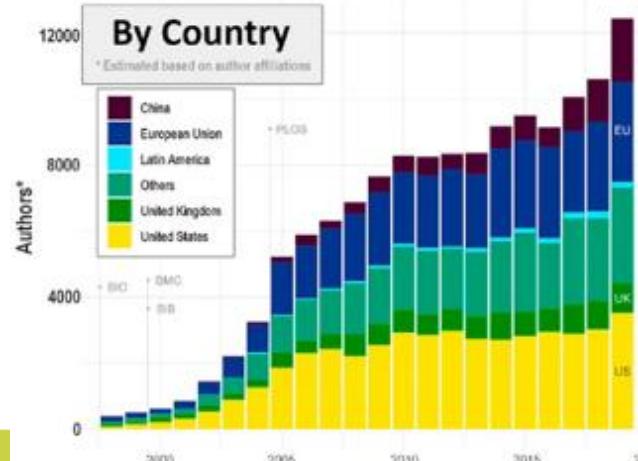
International Collaborations



Bioinformatic projects more and more rely on international large-scale collaborations

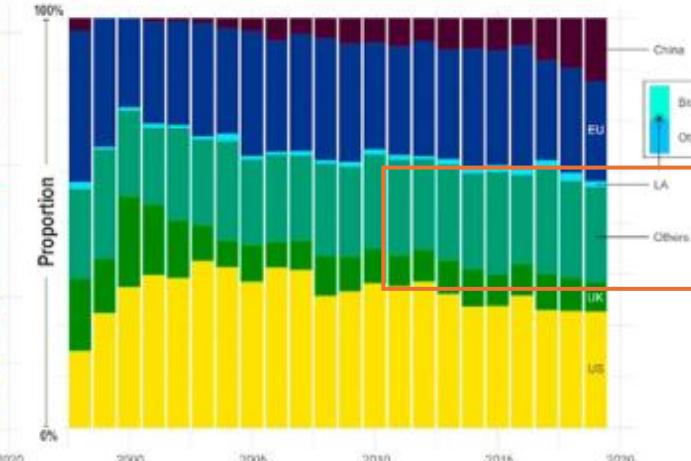
C

Authors by nationality



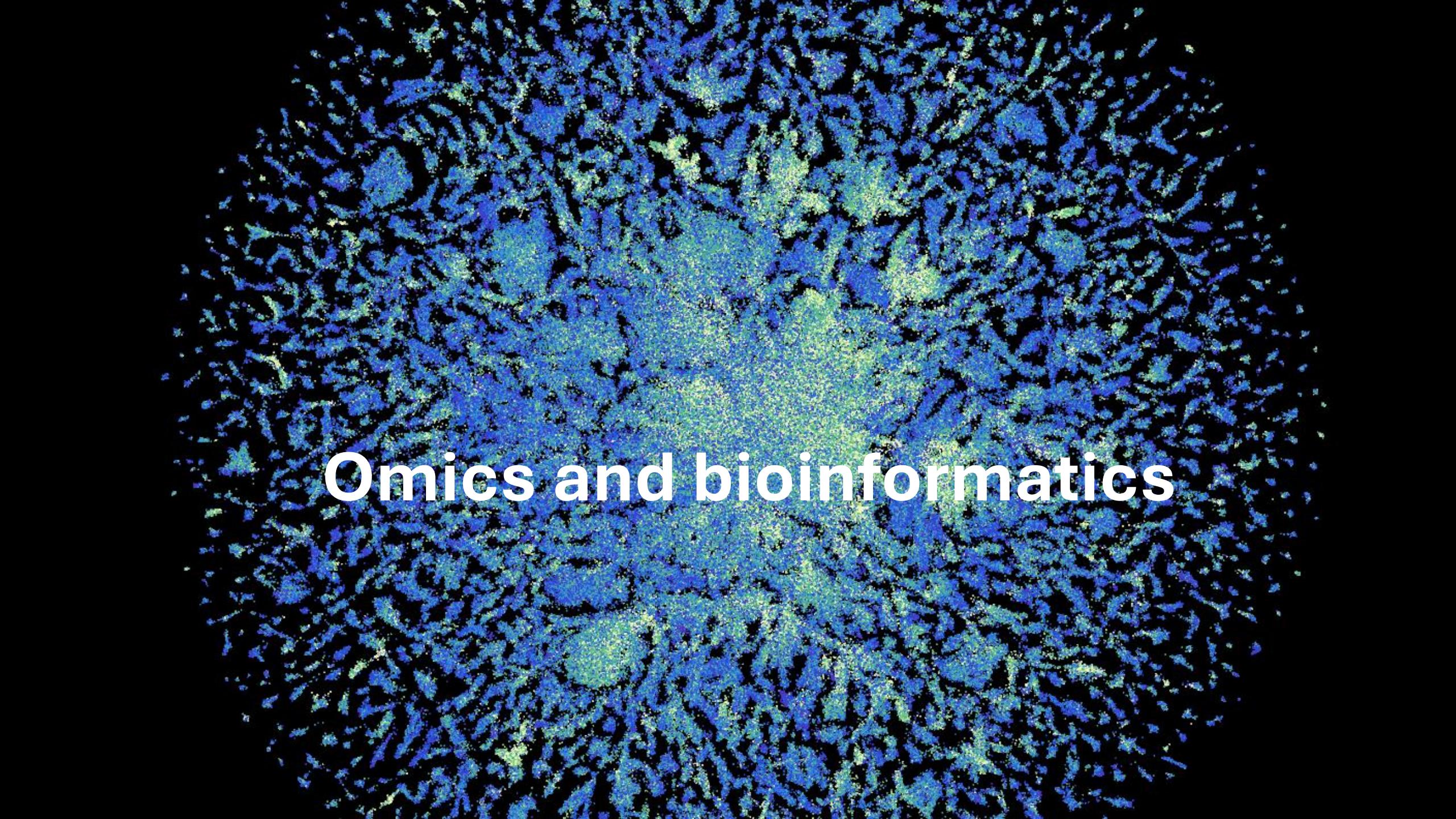
D

Authors by nationality (%)



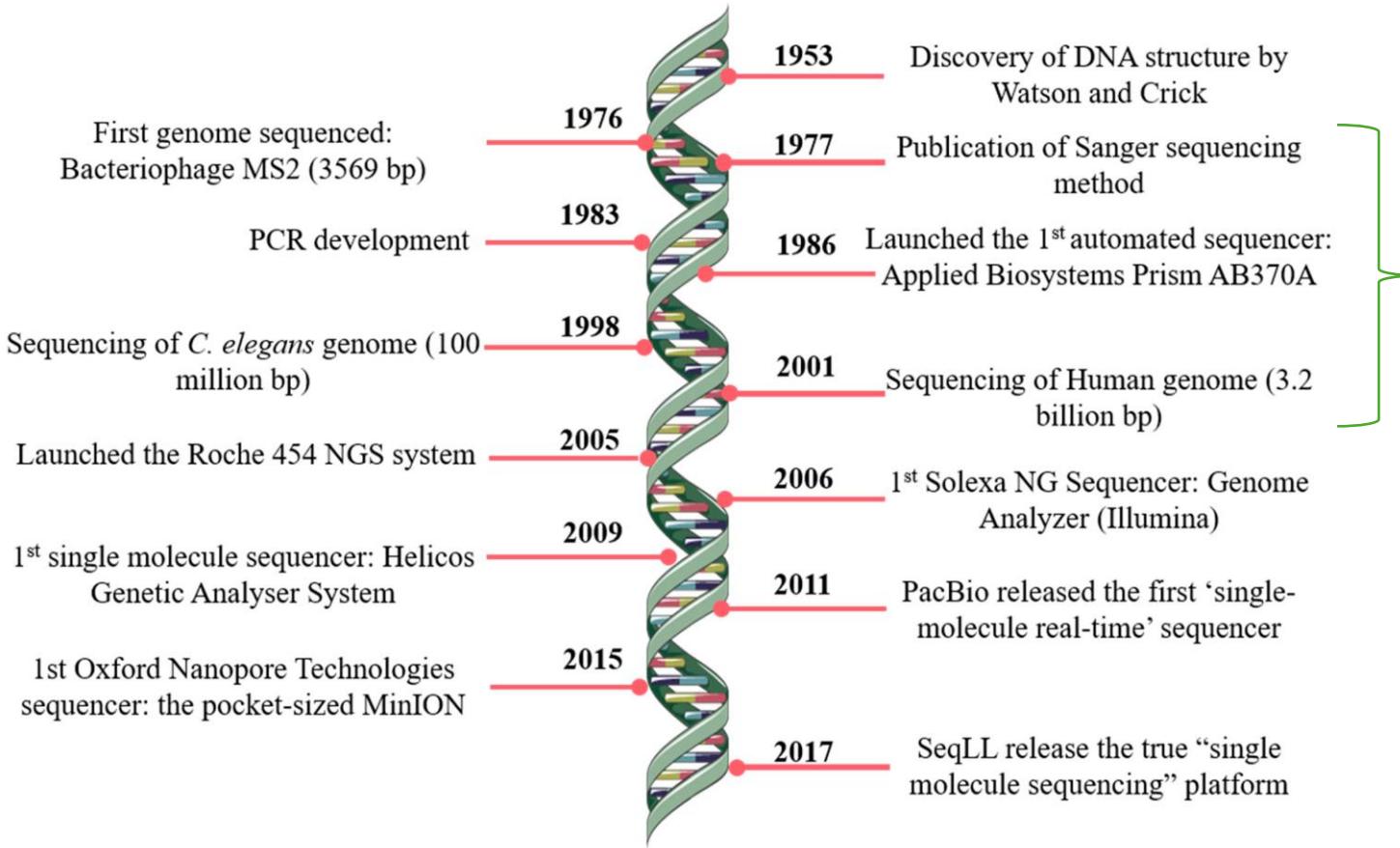
Yet, bioinformatics publications are still dominated by researches from US, UK, EU and China...

A Brief History of Bioinformatics Told by Data Visualization



**Omics and bioinformatics**

# The central role of NGS in bioinformatics



## BEFORE NGS: SANGER SEQUENCING (1977-2005)

**COST:** \$100 million per genome  
(Human Genome Project completed in 2003)

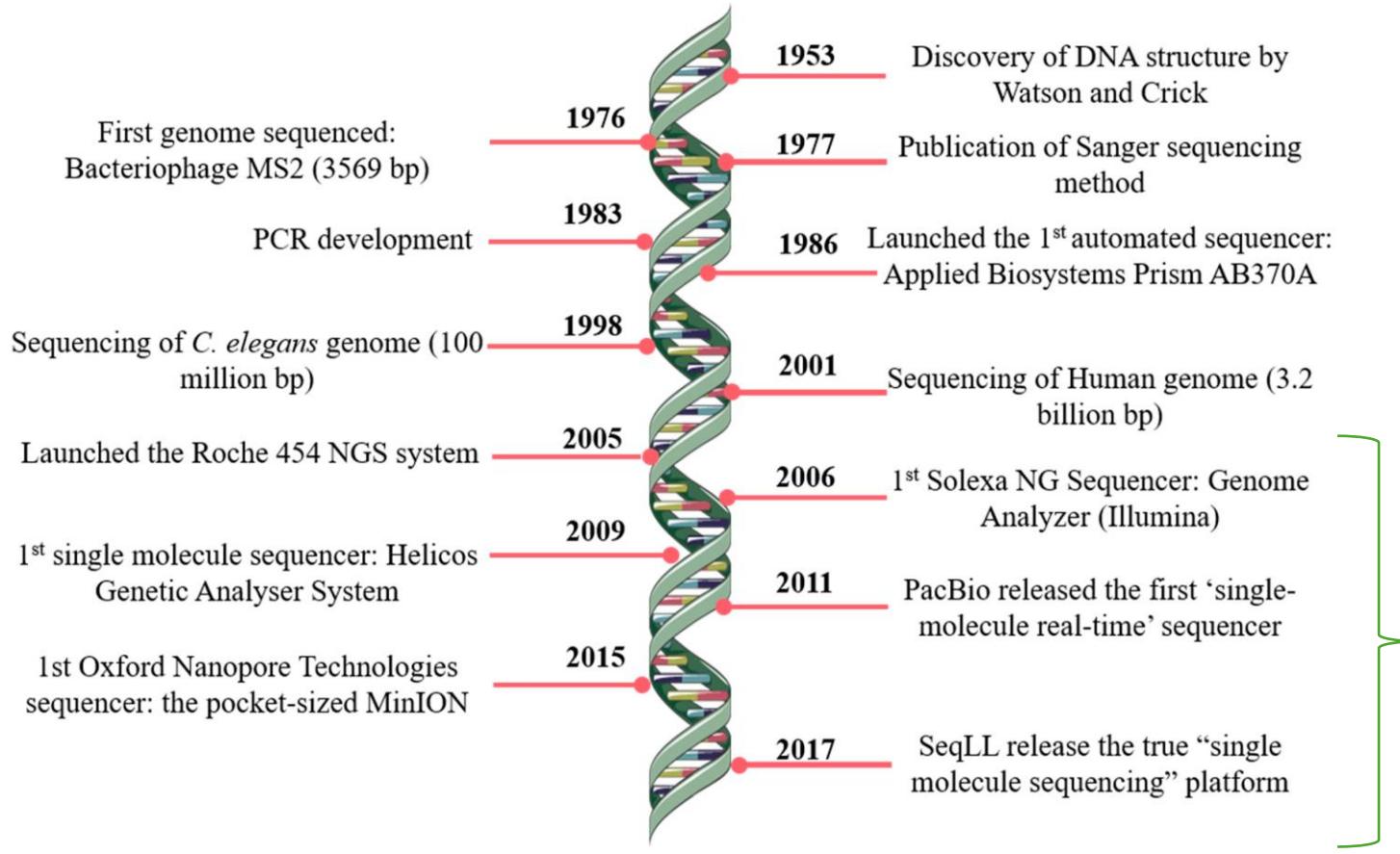
**TIME:** 13 years for the first human genome

**DATA SCALE:** ~1 Megabase per week

### WHAT WAS POSSIBLE:

- Study one gene at a time
- Sequence only cultured organisms
- Process individual samples
- Single data types

# The central role of NGS in bioinformatics



## NGS: NEXT-GENERATION SEQUENCING (2005+)

COST: \$1,000 per genome today (100,000× cheaper!)

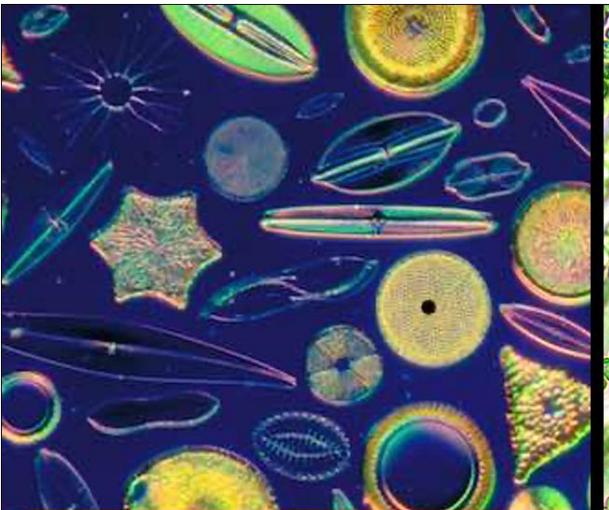
TIME: 1-2 days for a human genome (2,500× faster!)

DATA SCALE: ~1 Terabase per week (1,000,000× more data!)

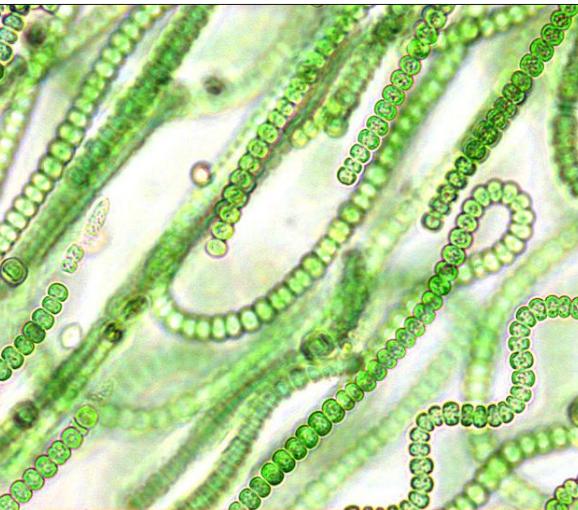
### WHAT BECAME POSSIBLE:

- Study ALL genes simultaneously
- Sequence entire microbial communities
- Process thousands of samples in parallel
- Integrate multiple data types (multi-omics)

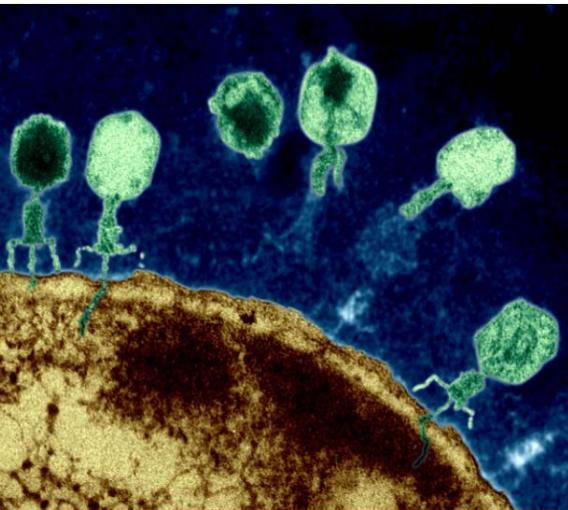
# What is a microbiome?



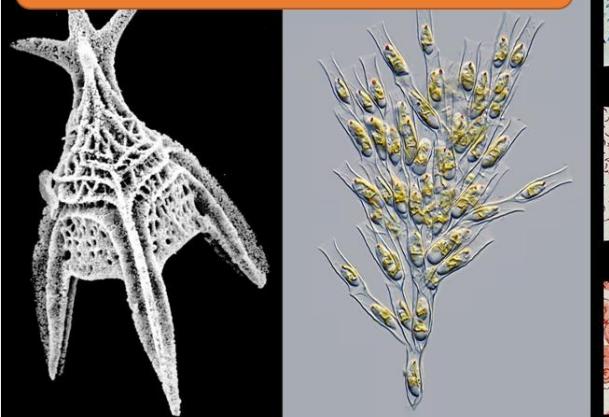
Micro-eukaryotes



Bacteria and Archaea



Viruses



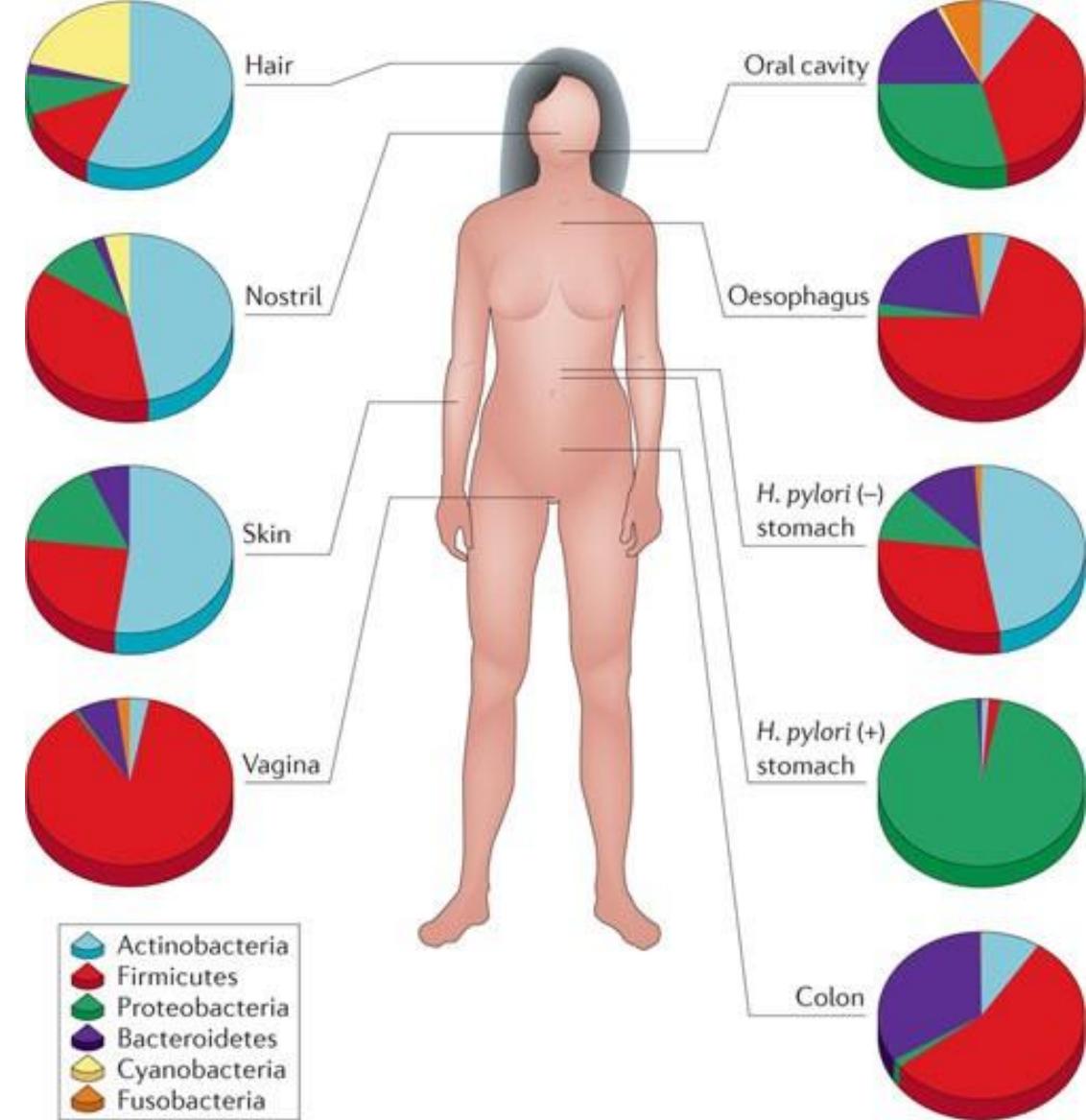
A **microbiome** is a complex mixture of microorganisms that reside in a specific environmental niche.

OHMI: <https://doi.org/10.1186/s13326-019-0217-1>

# What is the human microbiome ?

The human microbiome is the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space.

HMP: <https://doi.org/10.1038/nature06244>



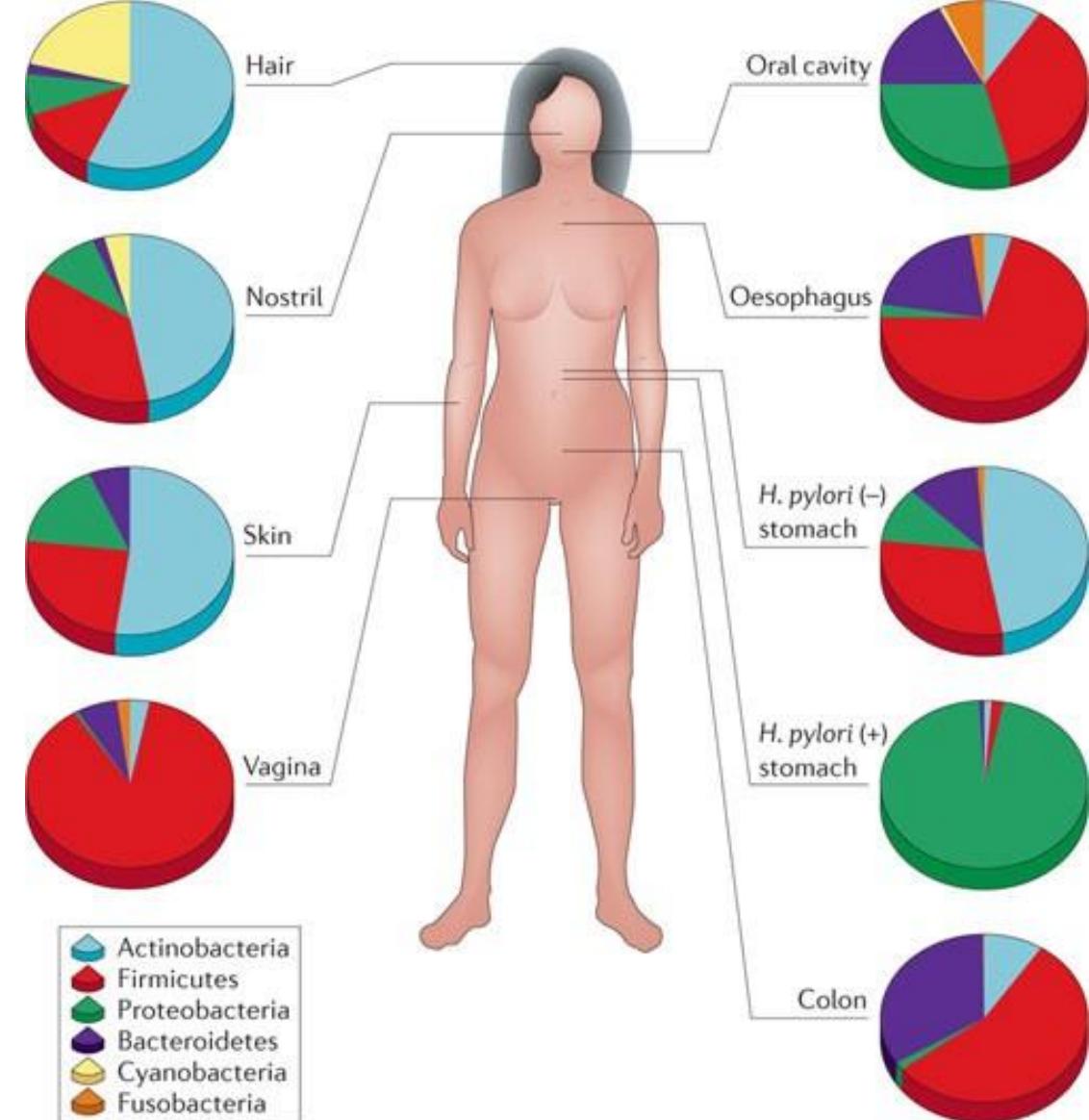
Nature Reviews | Genetics

# What is the human microbiome ?

## The human body contains

- at least 1000 different species of prokaryotes
- carries 150 times more microbial genes than human genes

Microbiota composition differ according to different locations, age, sex, genetic background, and diet of the host

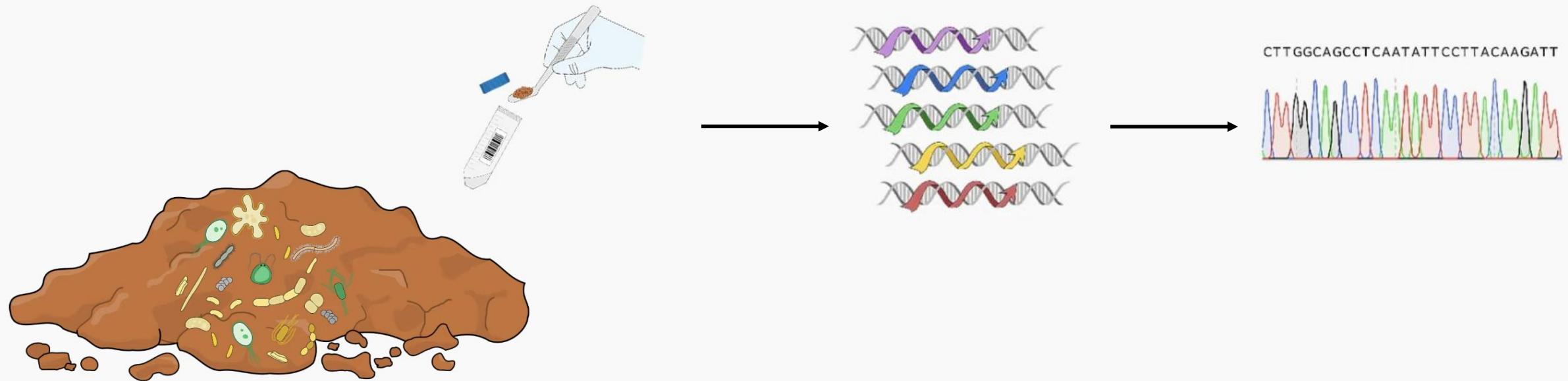


The human microbiome is an essential organ whose function are central to human health

Microbial metagenomics allows to explore the complex interaction between microbiota and host health

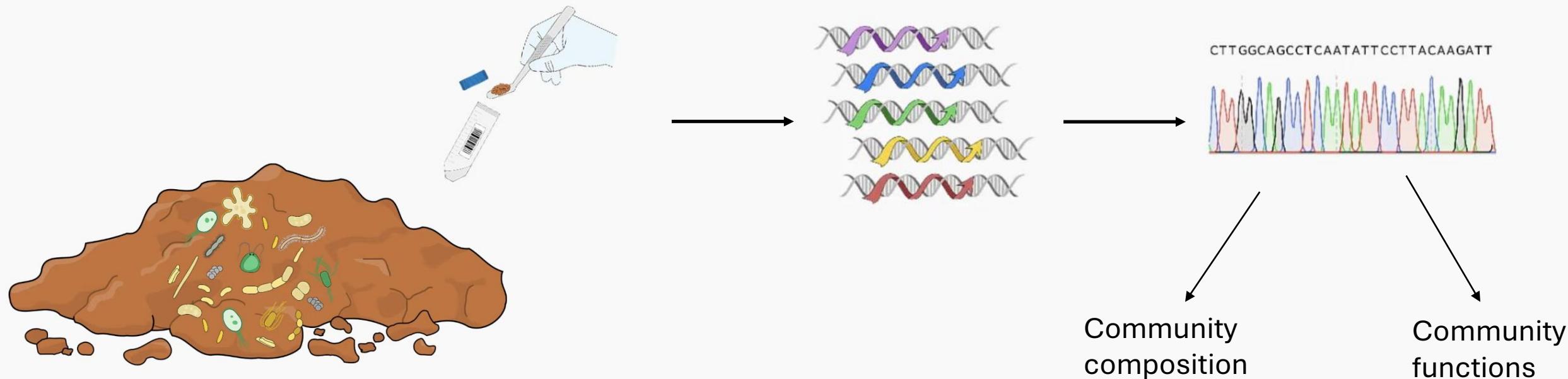


# Metagenomics approaches



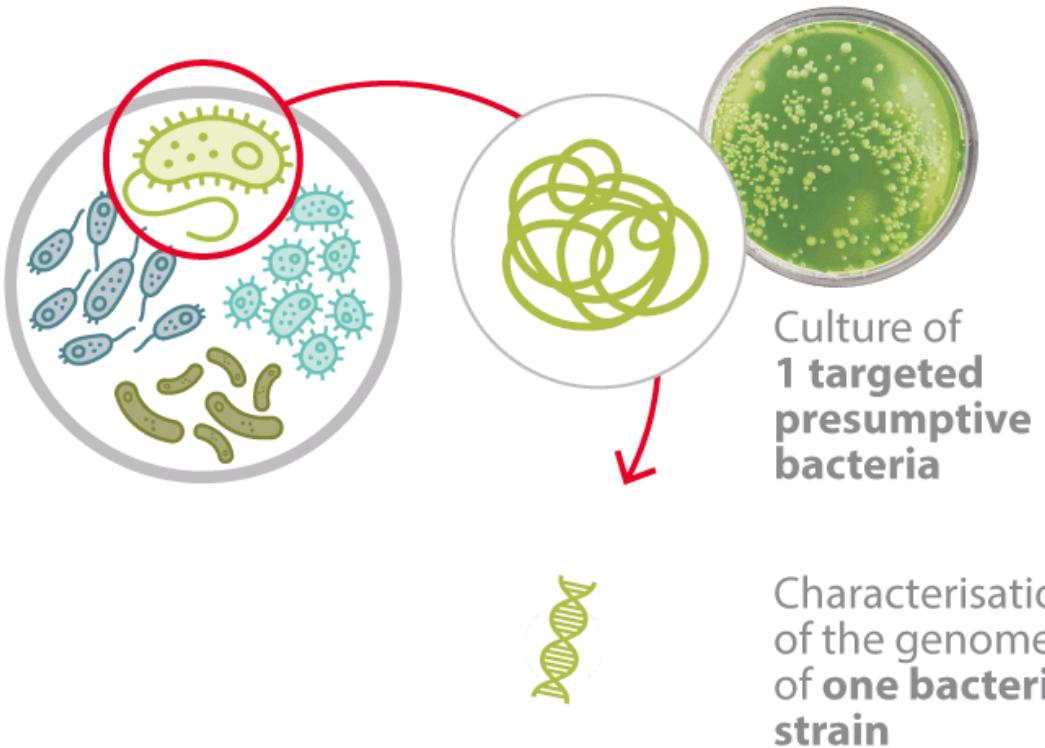
Metagenomics is the genomic analysis of microbial communities by extracting and sequencing their DNA. Metagenomics allows to study complex communities of microorganisms directly in their natural environment

# Metagenomics approaches

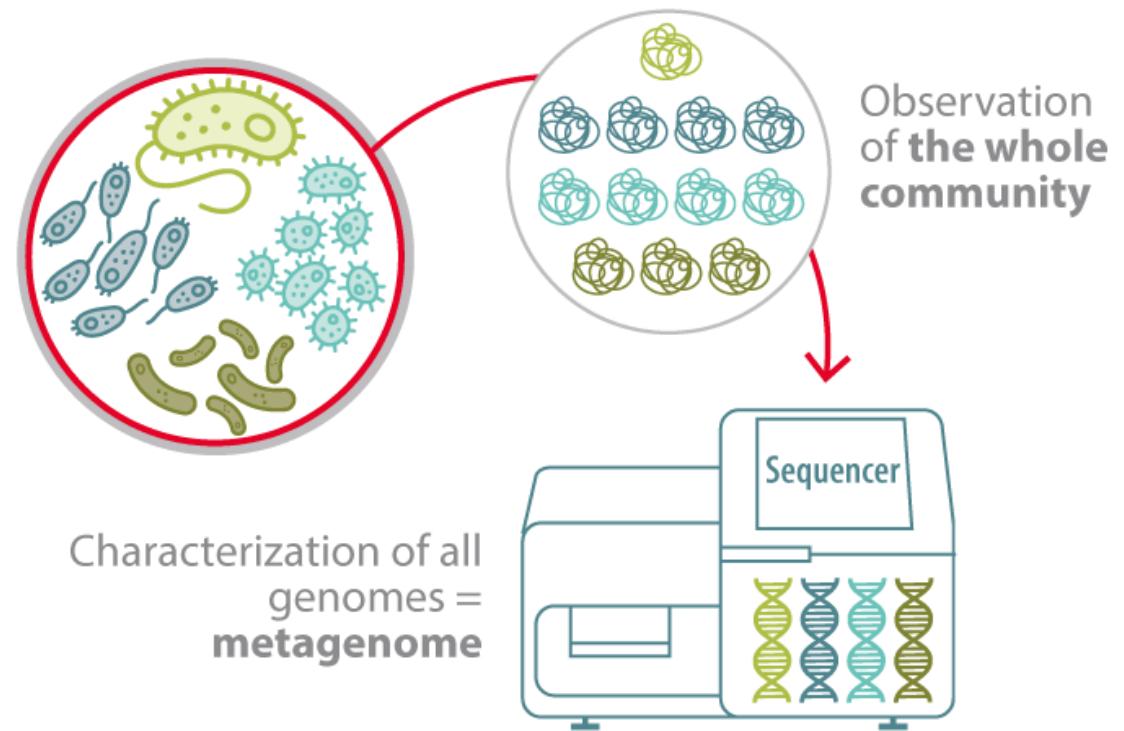


Metagenomics is the genomic analysis of microbial communities by extracting and sequencing their DNA. Metagenomics allows to study complex communities of microorganisms directly in their natural environment

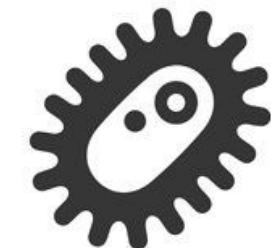
## CULTURE



## METAGENOMICS



# Why do we care about the human microbiome?



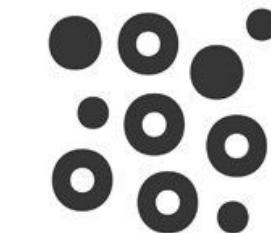
**Detection of an infection** leading to disease state  
**Assessment of treatment** method efficiency

e.g. Diabetic foot ulcer



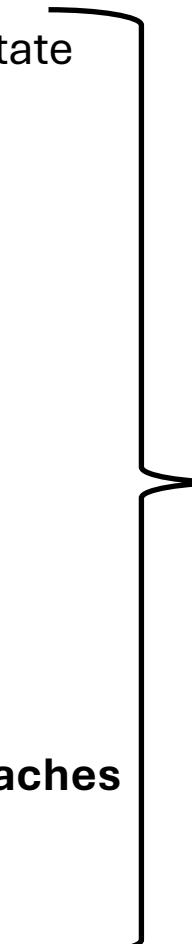
**Novel biomarker of diseases**

e.g. Colo-Rectal Cancer



**Central role in personalized medicine approaches**

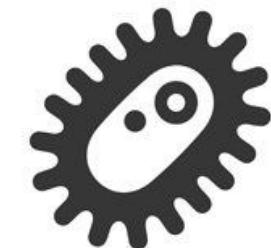
e.g. Response to treatments



## **Who is there?**

- What is the microbial species composition?
- How dynamic through time?
- How different the composition is between healthy and non-healthy states?

# Why do we care about the human microbiome?



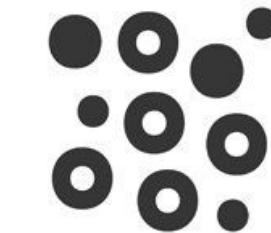
**Detection of an infection** leading to disease state  
**Assessment of treatment** method efficiency

e.g. Diabetic foot ulcer



**Novel biomarker of diseases**

e.g. Colo-Rectal Cancer



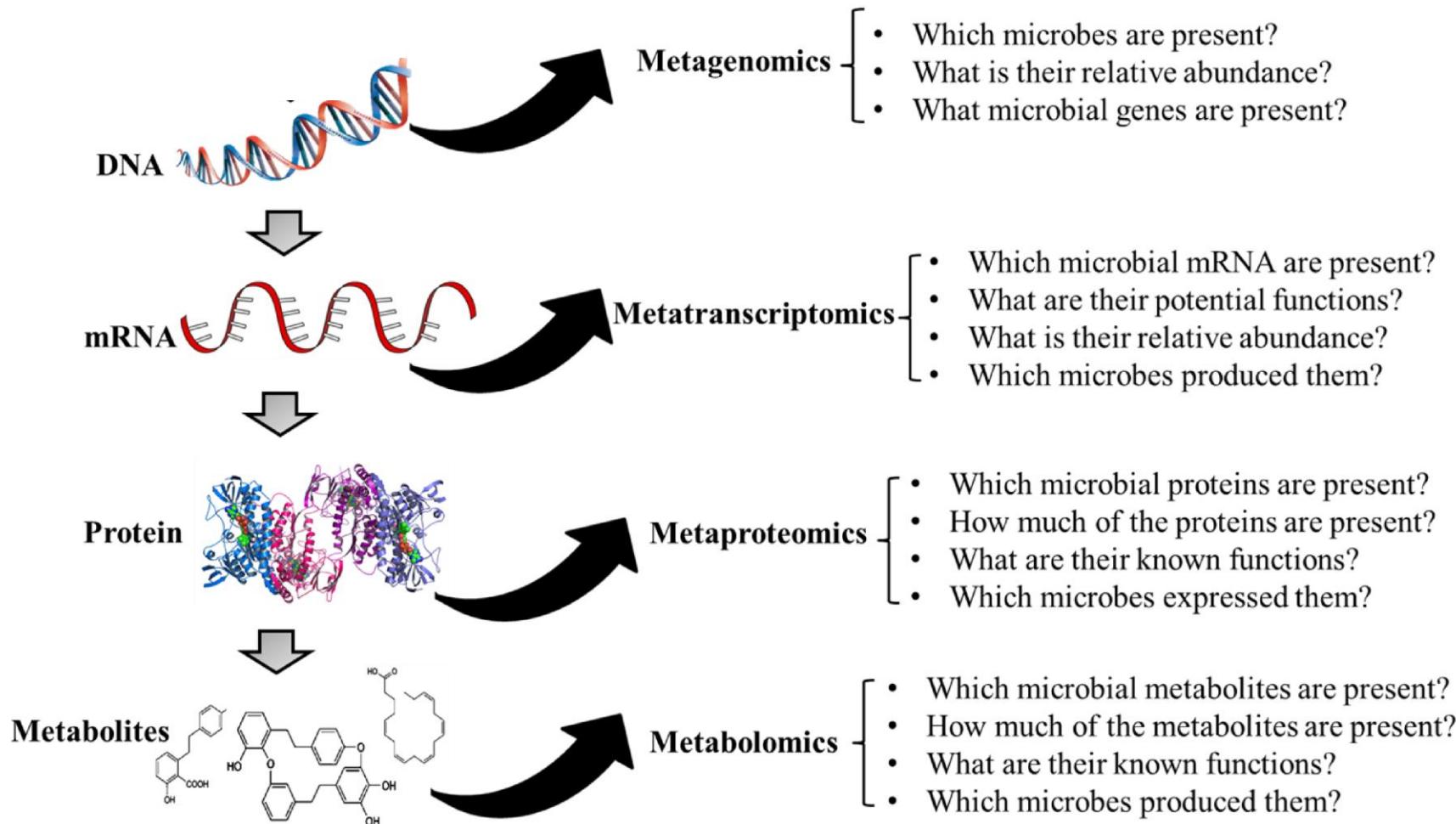
**Central role in personalized medicine approaches**

e.g. Response to treatments

## What are they doing?

- What functions these species carries?
- What virulence factors can be identified?
- What metabolites are produced by the community?

# The Many Flavors of Omics



**Cost increases** dramatically with each layer

**Complexity** of analysis grows exponentially

# Bioinformatics beyond sequence data

Imaging analysis

## MEDICAL IMAGING & DIAGNOSIS

- AI-powered X-ray and MRI interpretation
- Automated tumor detection and classification

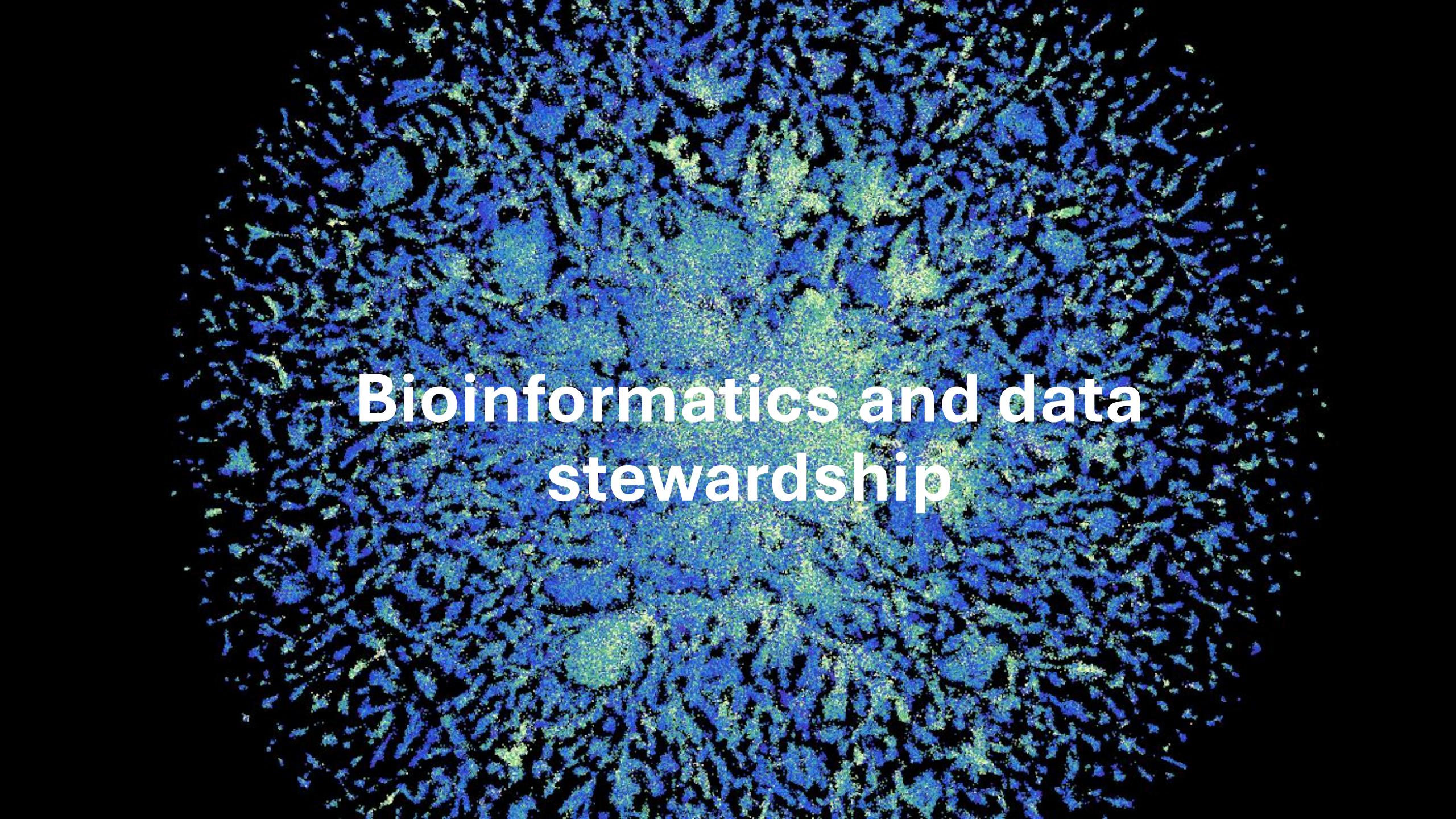
## ECOLOGY & CONSERVATION

- Camera trap data + species identification
- Biodiversity monitoring at scale

Sensor data

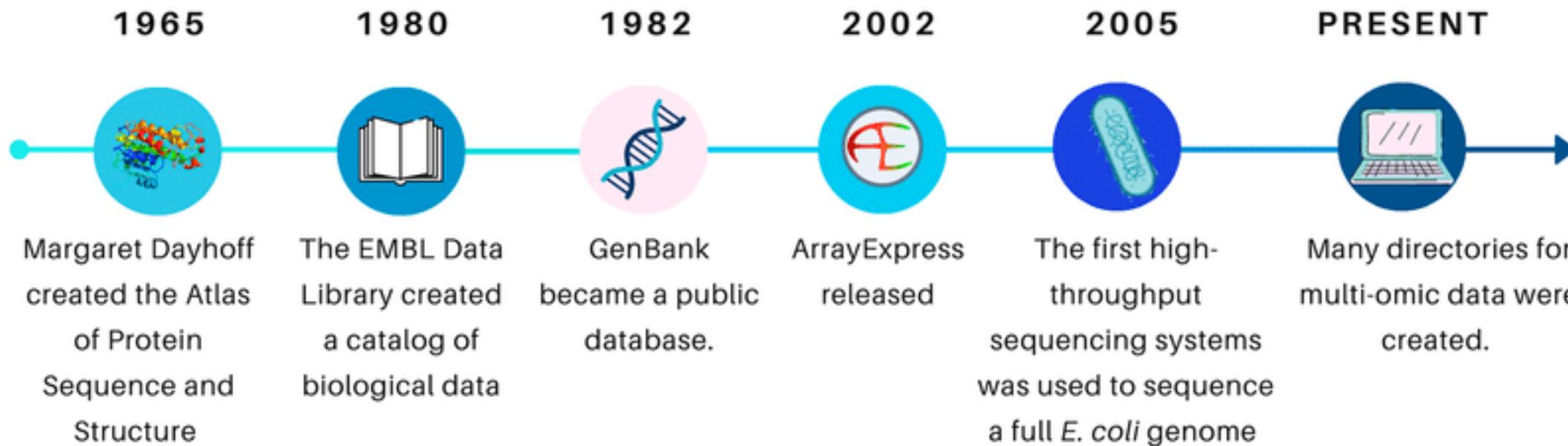
## CLINICAL APPLICATIONS

- Wearable sensor data analysis
- Integrated patient decision support systems



# Bioinformatics and data stewardship

## A brief history of biological databases



# The research data loss problem

- Data stored on temporary devices
- Poor organization and file naming
- Missing metadata and context
- Obsolete file formats

→ Inability to verify analyses and claims

Editorial | [Open Access](#) | Published: 21 February 2020

## No raw data, no science: another possible source of the reproducibility crisis

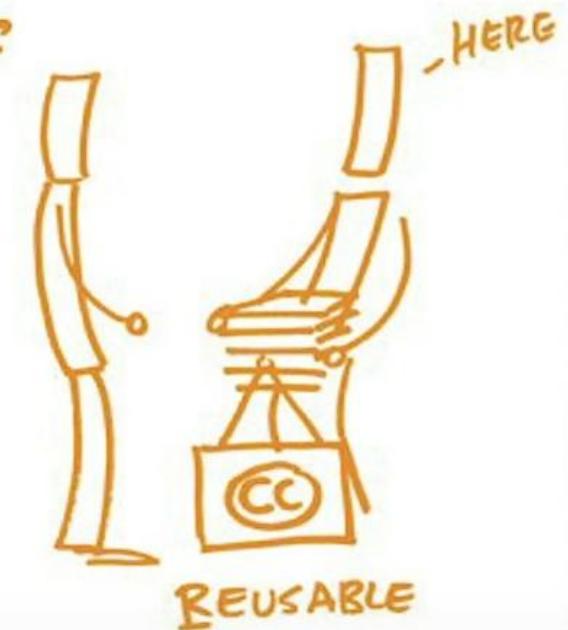
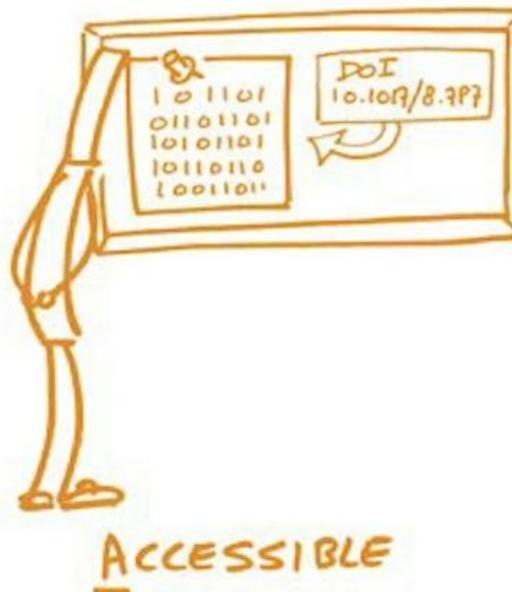
[Tsuyoshi Miyakawa](#) 

[Molecular Brain](#) **13**, Article number: 24 (2020) | [Cite this article](#)

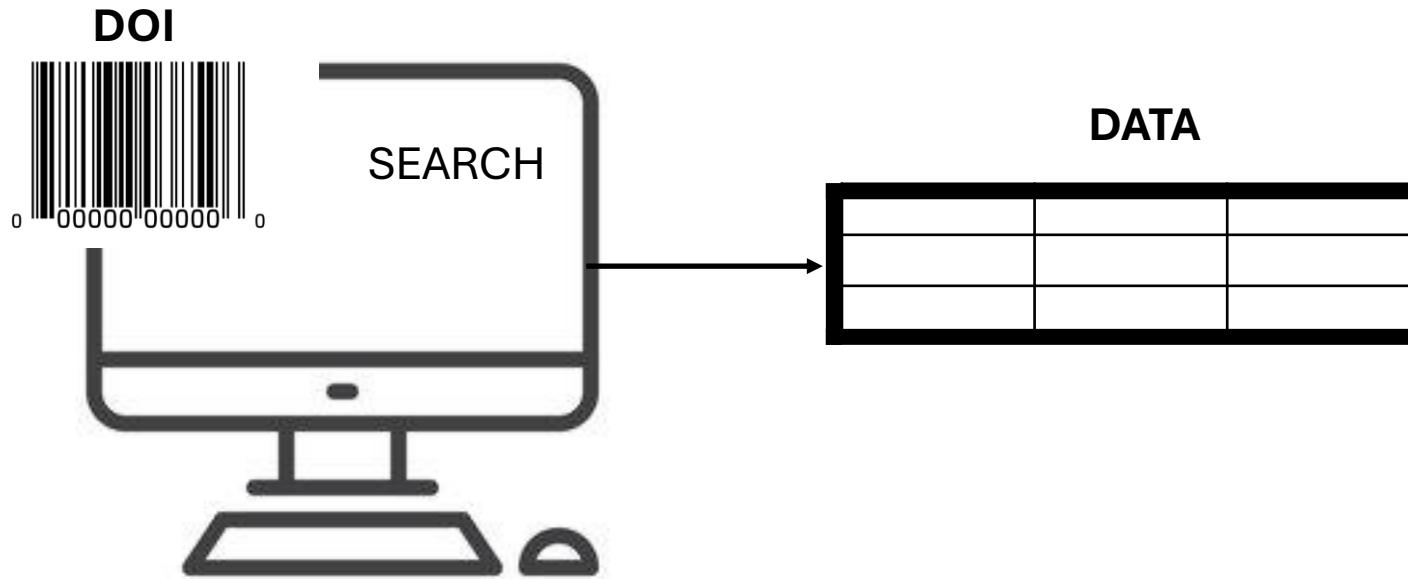
**36k** Accesses | **2** Citations | **2180** Altmetric | [Metrics](#)

# The FAIR data principles

## FAIR DATA PRINCIPLES



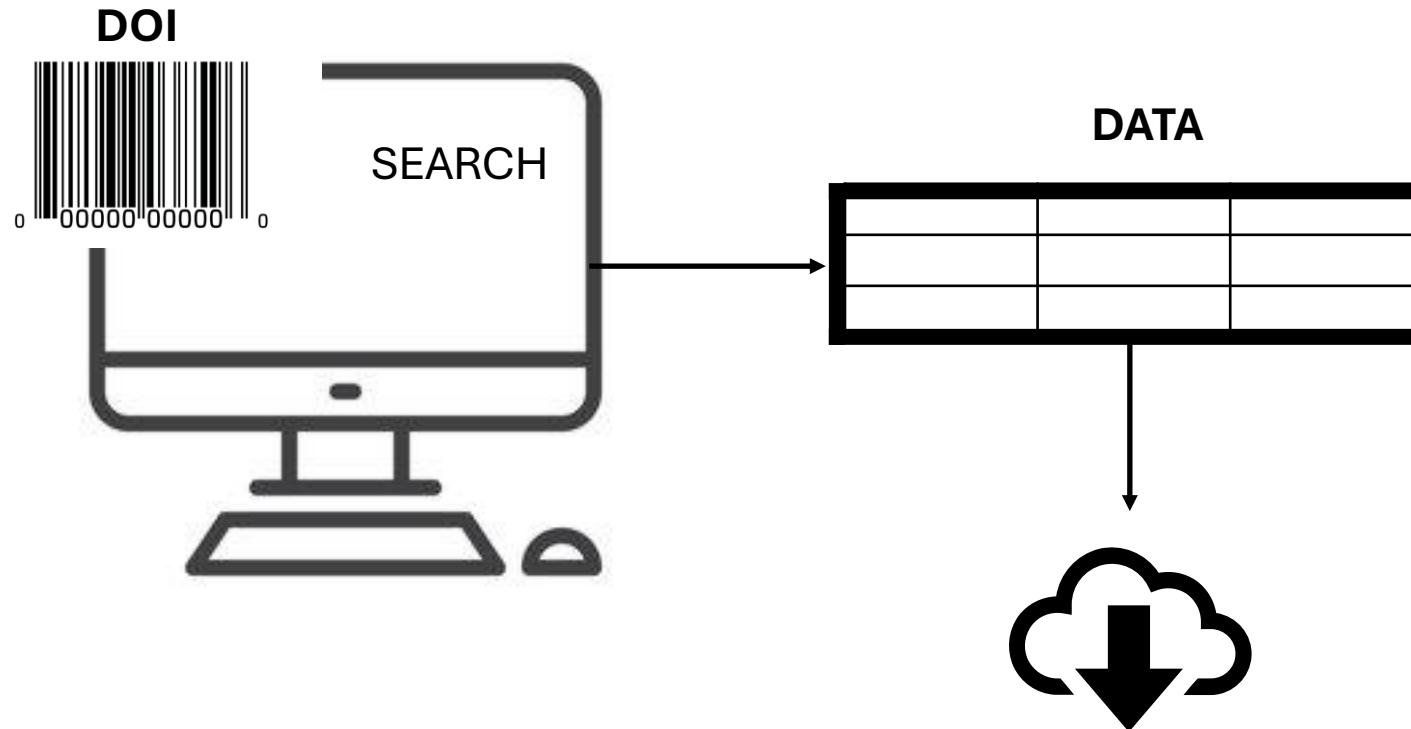
# The FAIR data principles : FINDABLE



Make all relevant raw data findable for future readers:

- Supplementary material
- Dedicated databases
- Data archives (Zenodo, FigShare)
- Data can be associated to a DOI

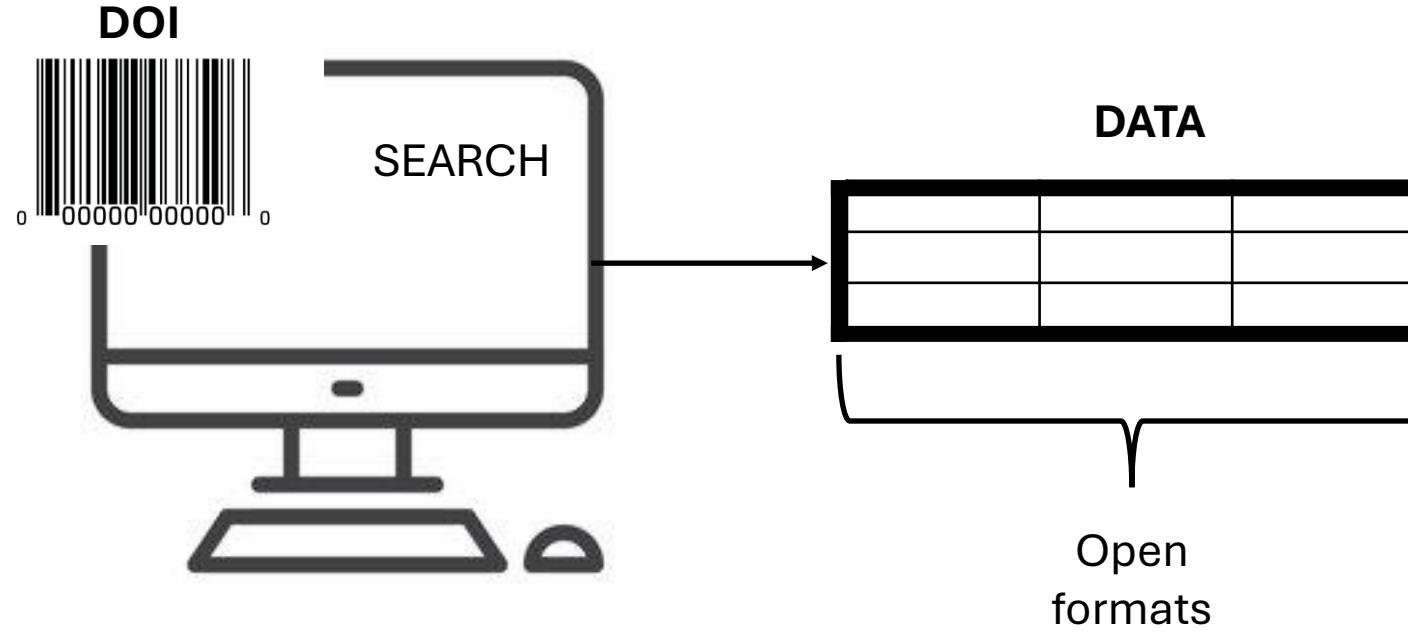
# The FAIR data principles : ACCESSIBLE



Make all relevant raw data accessible for future readers:

- Data not behind a paywall
- Clear data access conditions  
(avoid data accessible “upon request”)

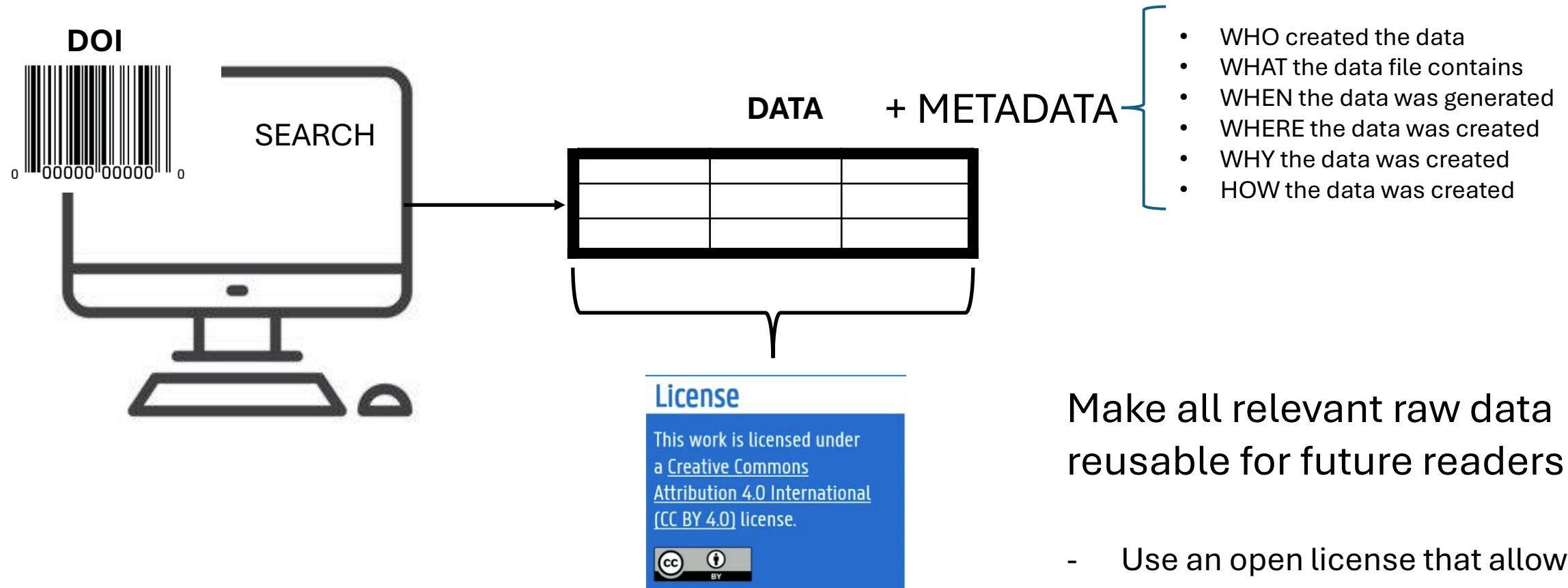
# The FAIR data principles : INTEROPERABLE



Make all relevant raw data interoperable for future readers:

- Use an open file format
- Using ontologies and controlled vocabulary

# The FAIR data principles : REUSABLE



# The bioinformatician as data steward



## WHAT PROPER DATA MANAGEMENT ENABLES

### 1. REUSE FOR NEW QUESTIONS

Data outlives the original study

Example: Training ML models on public data

### 3. REPRODUCIBILITY

Verify and build on published work

Example: Re-run analyses with updated tools

### 4. COLLABORATION

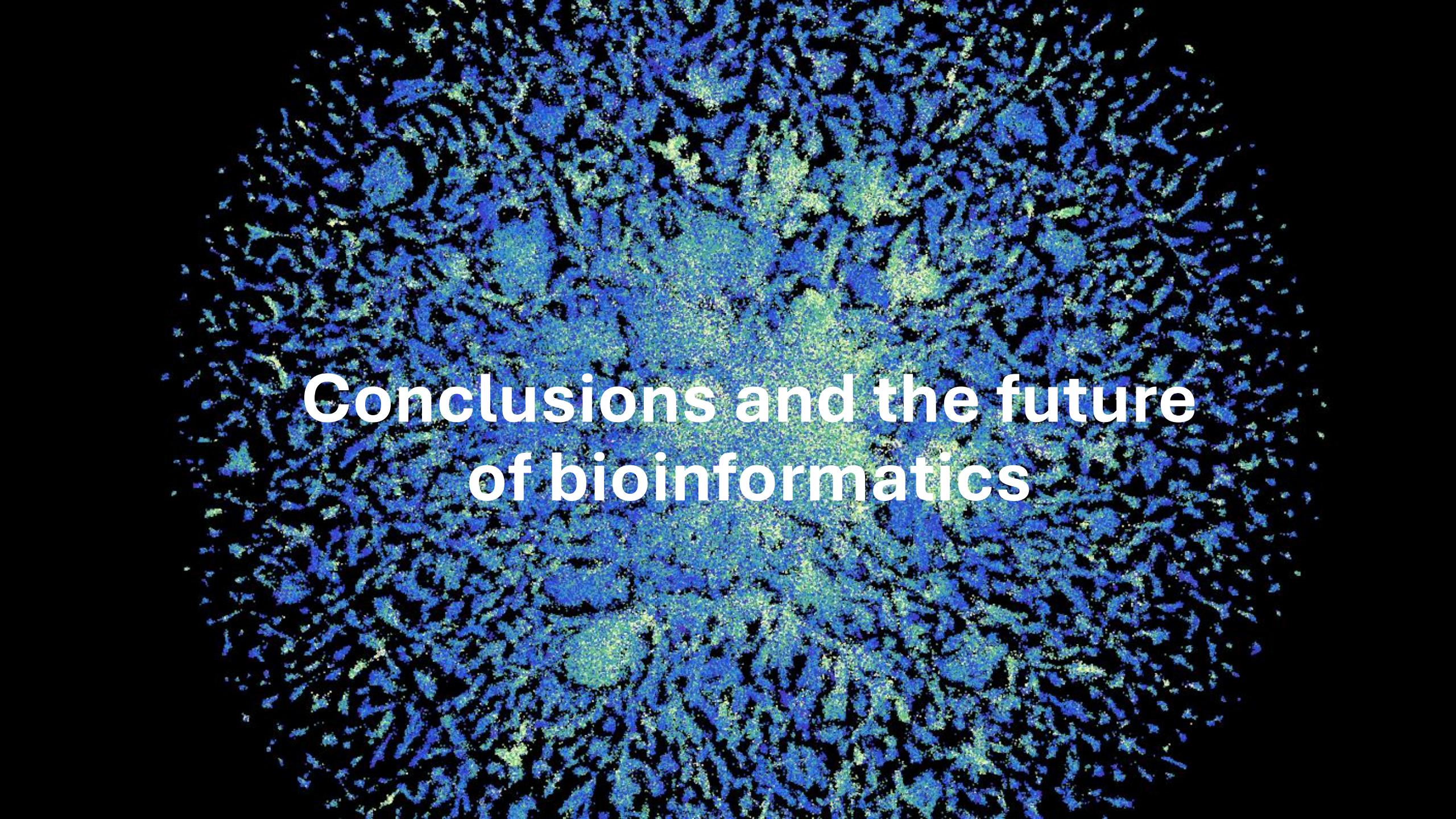
Share data across labs and countries

Example: International consortium

### 5. TRAINING & EDUCATION

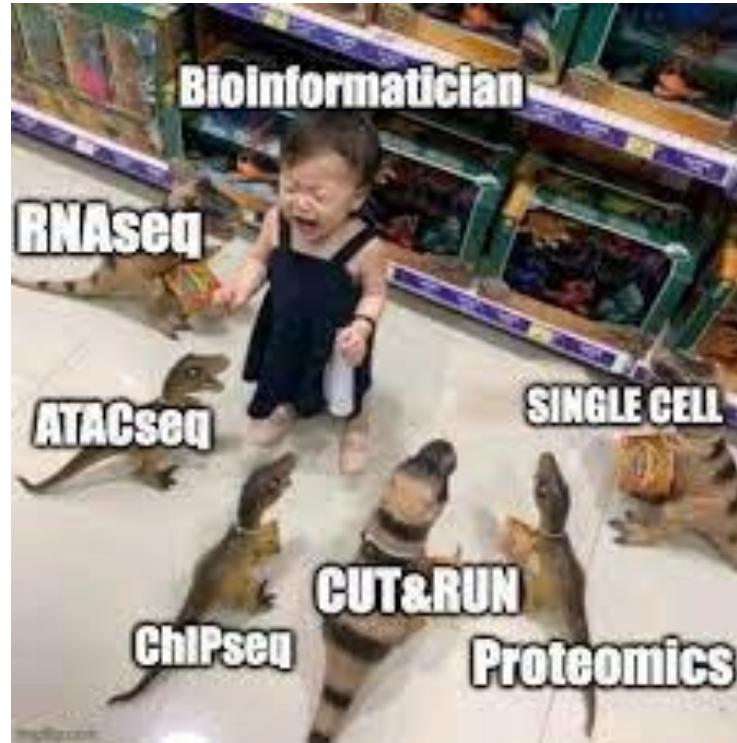
Students learn on real datasets

Example: Using public data in workshops



# **Conclusions and the future of bioinformatics**

# What future for bioinformatics?



## HOW FAST IS BIOINFORMATICS MOVING?

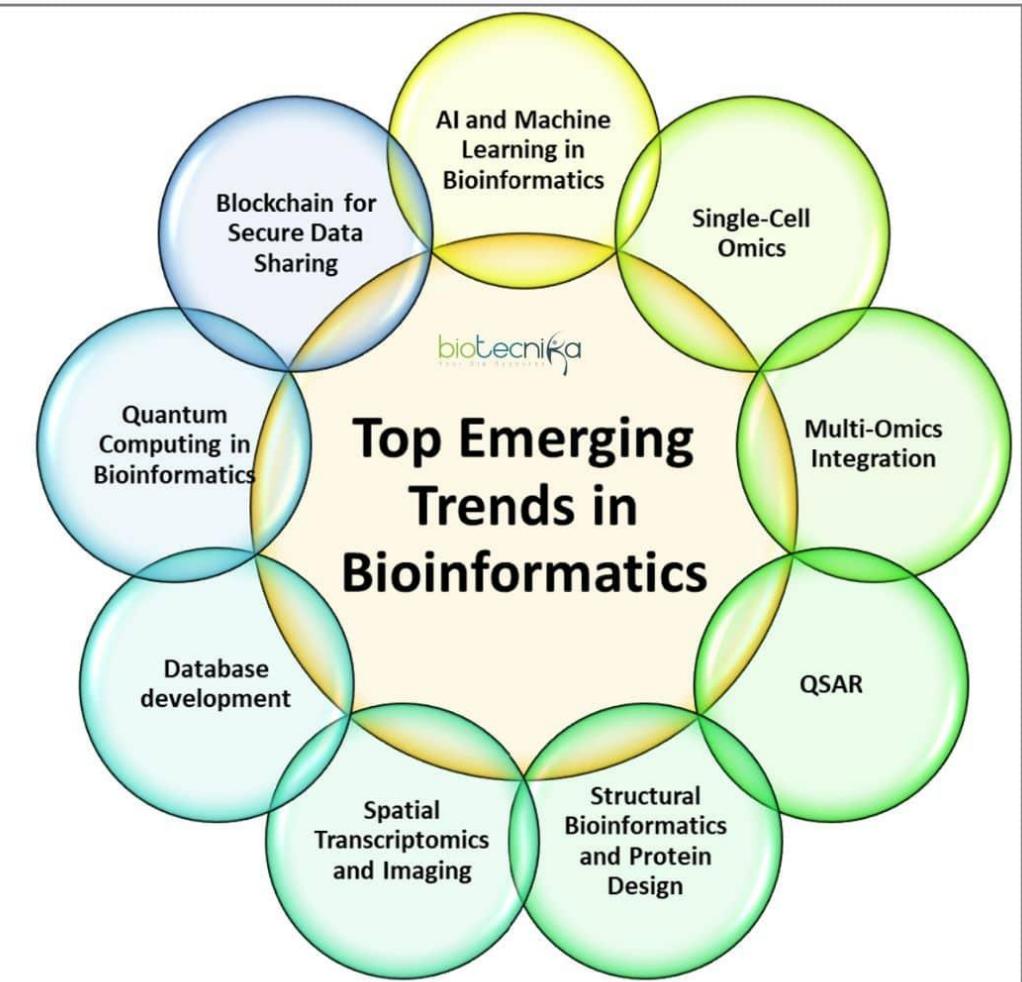
- 2014: Metagenomics emerges as a field
- 2018: Single-cell sequencing becomes routine
- 2020: Spatial transcriptomics wins Method of the Year
- 2022: AlphaFold solves protein structure prediction
- 2024: Multi-omics integration is standard
- 2025: ???

The field reinvents itself every 3-5 years

### What this means for YOU:

- The tools you learn today will be outdated soon
- Adaptability > mastering one tool
- Stay curious, keep learning

# What future for bioinformatics?



## EMERGING FRONTIERS

### Beyond Omics:

- Wearable sensors generating continuous health data
- Electronic health records + genomics integration
- Real-time environmental monitoring

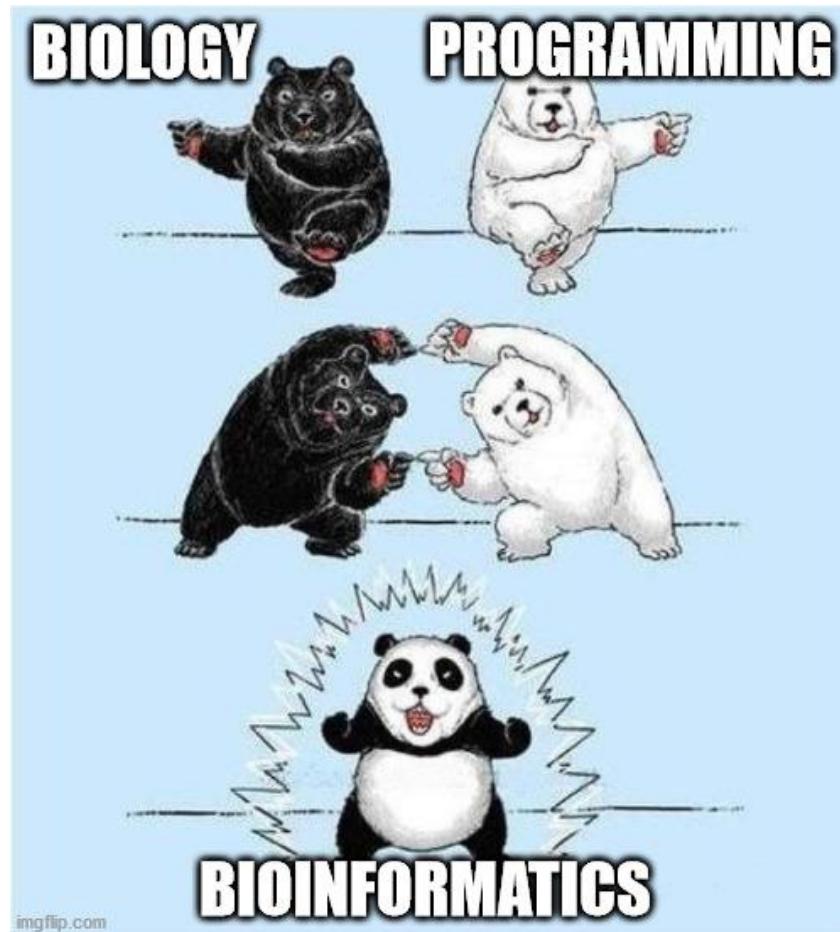
### New Technologies:

- Long-read sequencing
- Spatial omics (where things are in tissues matters)
- Single-cell resolution (every cell is different)

### New Applications:

- Precision medicine
- Microbiome therapeutics
- Climate & conservation
- Pandemic preparedness

# What future for bioinformatics?



## EMERGING FRONTIERS

### The Challenges Ahead:

- ⚠ Data privacy and ethics in an AI world
- ⚠ Computational infrastructure  
(where do we store petabytes?)
- ⚠ Making data and computational resources accessible globally (not just rich countries)
- ⚠ Training the next generation (that's YOU!)

# What future for bioinformatics?



## ARTIFICIAL INTELLIGENCE IN BIOINFORMATICS

What AI is already doing:

- ✓ Protein structure prediction (AlphaFold)
- ✓ Variant interpretation
- ✓ Image analysis (pathology, microscopy)
- ✓ Generating analysis code

Bioinformatics is already impacted by AI:

- ⚠ Some routine tasks will be automated
- ⚠ Entry-level positions may change

# What future for bioinformatics?



## ARTIFICIAL INTELLIGENCE IN BIOINFORMATICS

What AI is already doing:

- ✓ Protein structure prediction (AlphaFold)
- ✓ Variant interpretation
- ✓ Image analysis (pathology, microscopy)
- ✓ Generating analysis code

Bioinformatics is already impacted by AI:

- ⚠ Some routine tasks will be automated
- ⚠ Entry-level positions may change

The future bioinformatician:

- Works WITH AI tools
- Focuses on interpretation, not just analysis
- Brings biological insight
- Asks better questions than AI can generate

# What future for bioinformatics?



## The Opportunity:

- Biology is generating more data than ever
- Clinical applications are exploding
- Global health needs bioinformatics (pandemic prep, AMR)
- Conservation needs bioinformatics (climate change)
- The questions are getting MORE interesting, not less

## MY ADVICE:

- 
- Don't try to master everything → Focus on fundamentals
  - Don't fear AI → You have to learn to work with it
  - Don't work in isolation → Collaboration is key
  - Don't ever stop learning → This field rewards curiosity



**My contact:**  
[alise.ponsero@quadram.ac.uk](mailto:alise.ponsero@quadram.ac.uk)

**LinkedIn:**  
alise-ponsero-843b953b

**WE MUST HAVE**



# The Many Flavors of Omics

