Towards the integration of energy footprint issues in genomics

Christophe Battail, Ph.D., Pharm.

Genetics and Chemogenomics team
Biosciences and Bioengineering for Health lab (UA 13 INSERM-CEA-UGA)
IRIG – CEA Grenoble

Aristote seminar - Ecole Polytechnique - 13/04/2023

Human genomic medicine

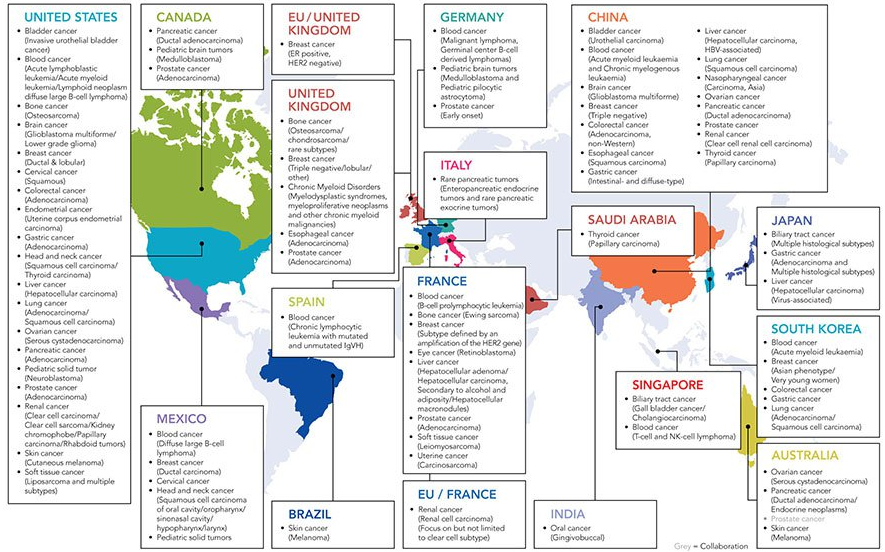
- Generation of **large-scale molecular and cellular profiling data** (genome and other omics) from large cohorts of patients by international research programs.

- Leading to significant progress in understanding the **molecular diversity of diseases** and in defining **subtypes**.

- Ongoing translation of these scientific advances into **precision medicine** procedures:
  - use of specific molecular signatures to refine the **diagnosis**;
  - consideration of genomic alterations to anticipate **patient prognosis**;
  - better predict the **therapeutic response** of patients to specific drugs.

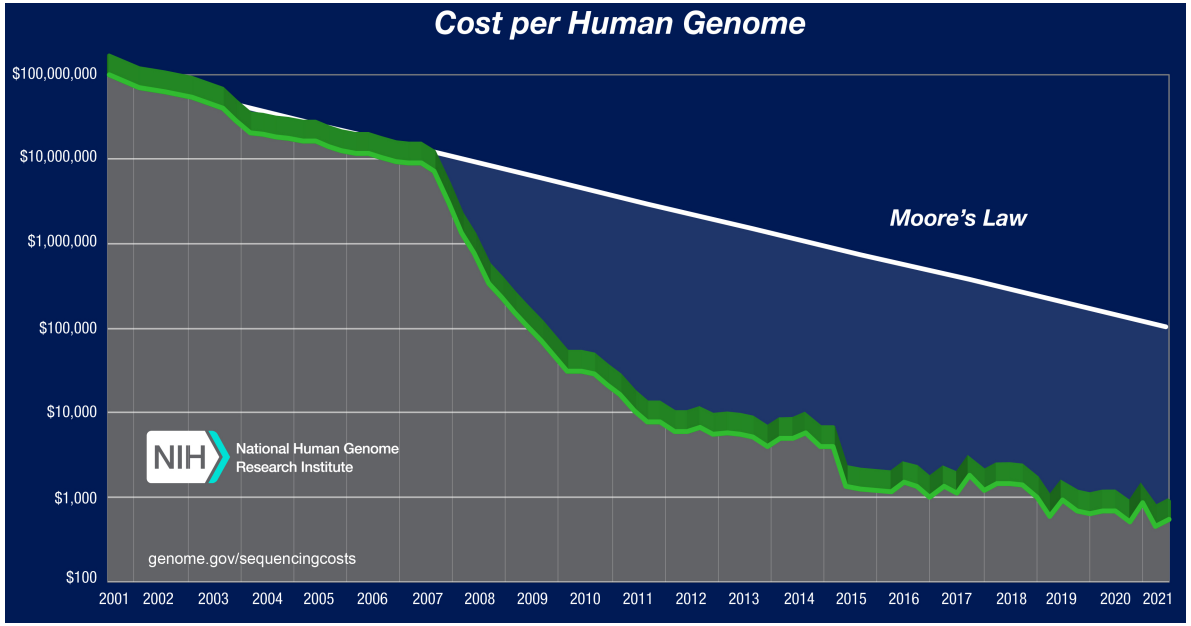
ARGO Accelerating Research in Genomic Oncology

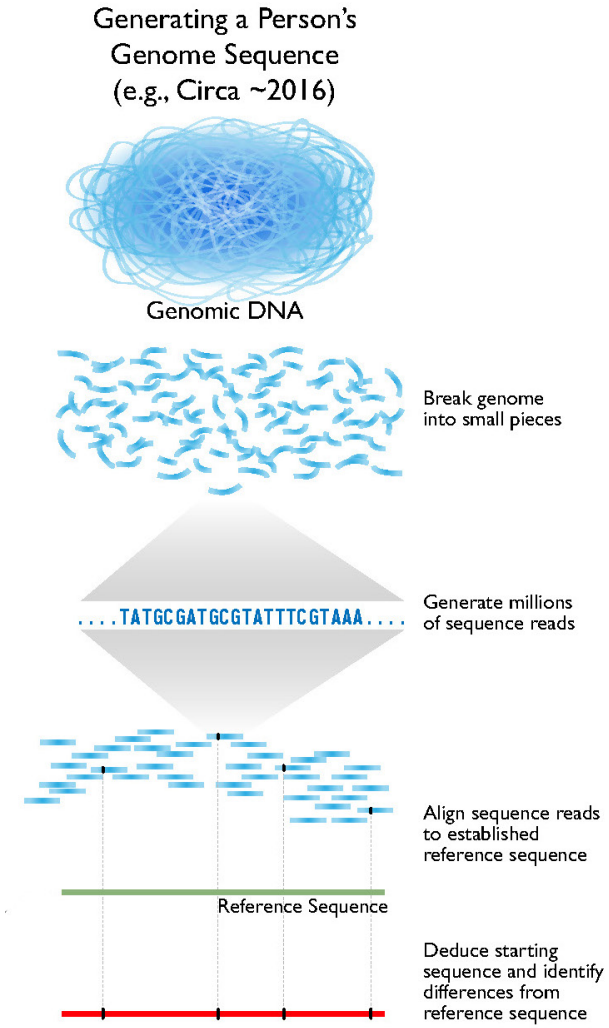
International Cancer Genome Consortium



Human genomic technology

- **High-throughput DNA sequencing**, the major experimental technology allowing the rapid and cost-effective generation of data for precision medicine.



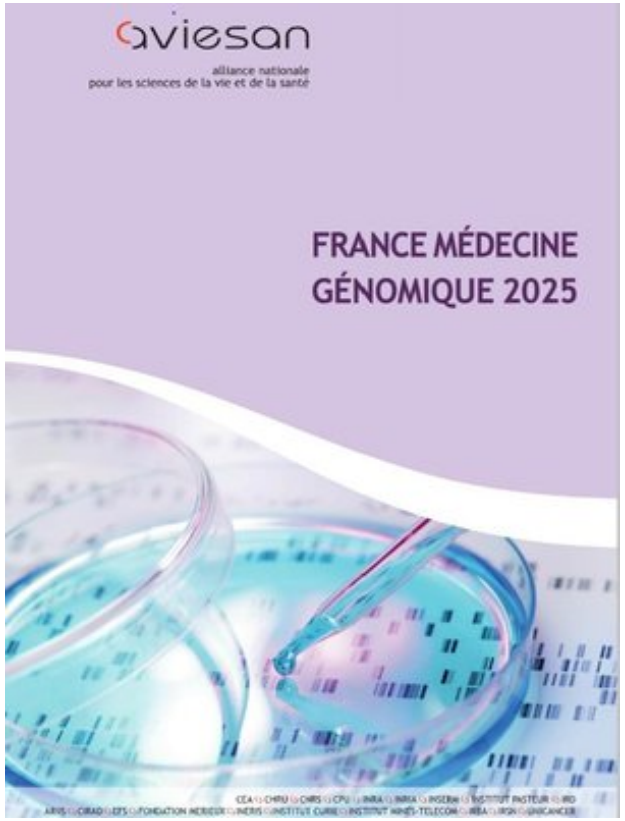


Illumina NovaSeq
(20 billion sequences / 13-38 hours)

Adapted from NHGRI resources

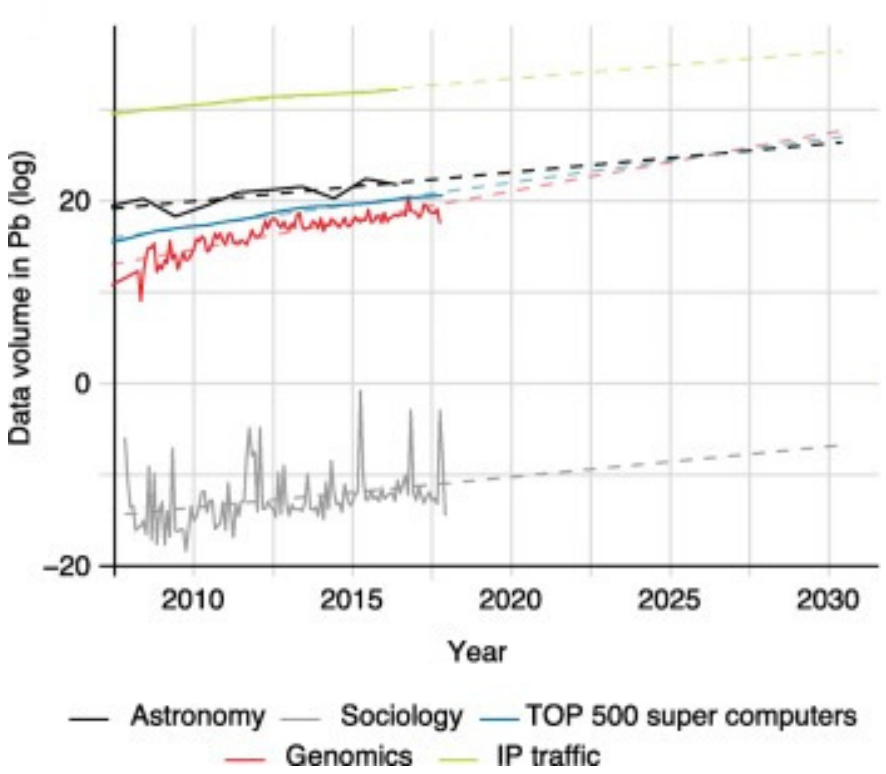
French genomic medicine initiative

- The **2025 French genomic medicine plan**, started in 2016:
  - **integrate genomic technologies** into the patient care system;
  - enable **equitable access** to DNA sequencing in clinical practice for rare diseases and cancers;
  - develop an **industrial sector** in genomic medicine;
  - **protect patient health data** and provide citizen with information on precision medicine.

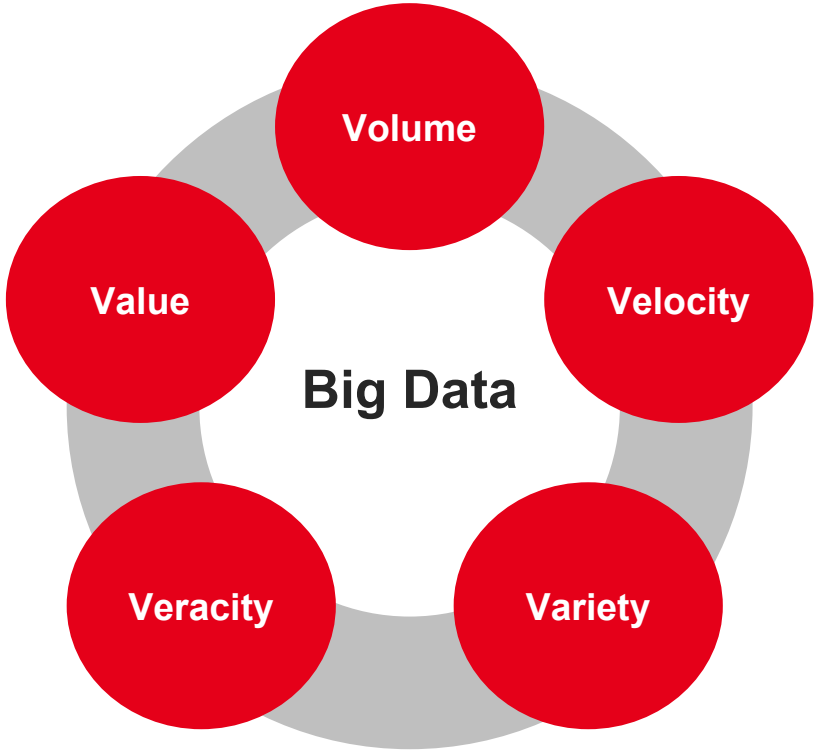


Data topology in genomic science

- **Volume/Velocity**: total data volume smaller than in earth science, but the growth trend of genomics may lead to overtake it.

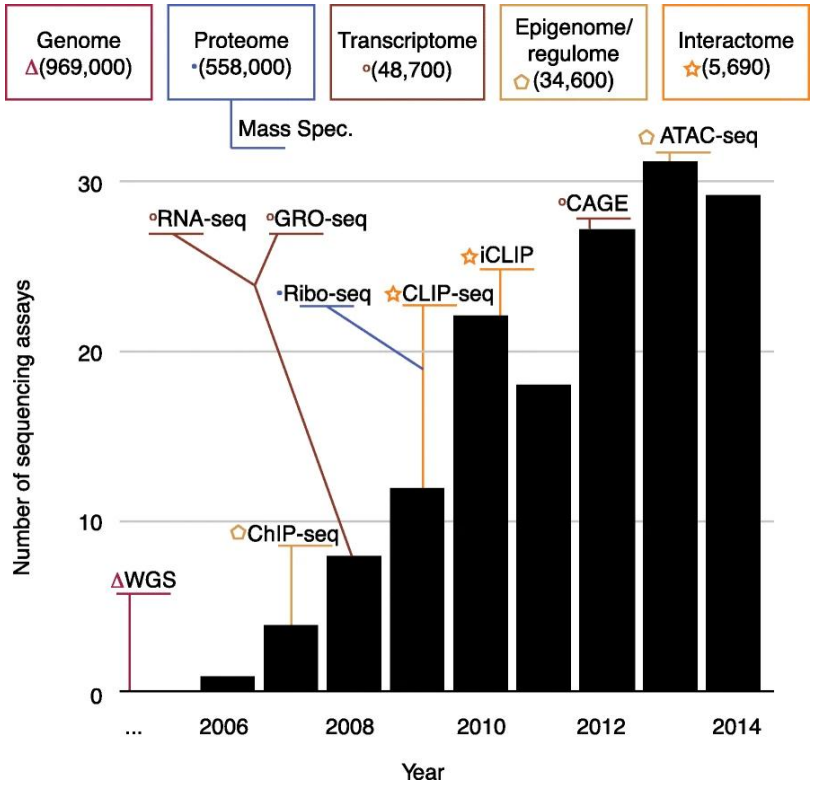


Source from Navarro et al., Genome Biology 2019.

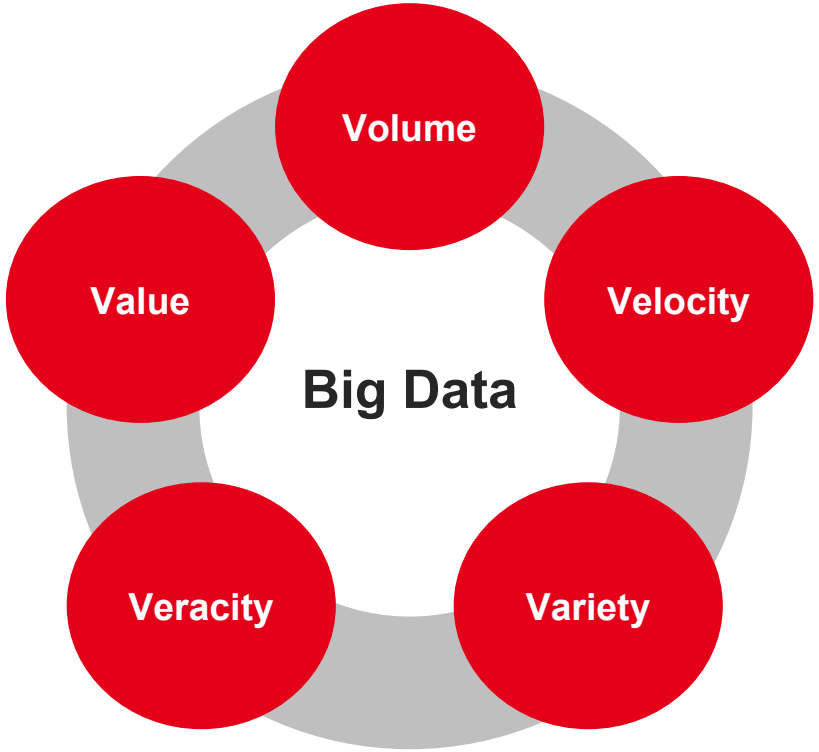


Data topology in genomic science

- **Variety**: the apparent monolithic nature of sequencing output hides diverse set of assays used to measure many aspects of genomes.



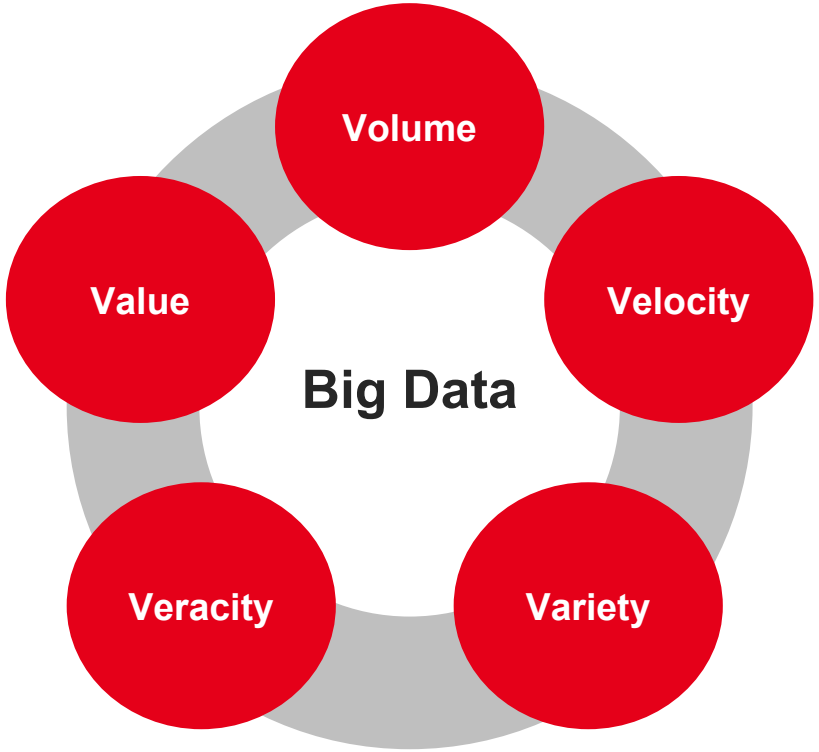
Source from Navarro et al., Genome Biology 2019.



Data topology in genomic science

- **Veracity**: the assurance of **quality and integrity of (confidential & sensitive) genomic data** is crucial given their use in the context of patient care.
  - **technical bias** between technologies and experimental protocols;
  - **biological and ethnic biases** of patient cohorts on which clinical decision rules are based;
  - **data poisoning attacks** aiming to bias patient care (health risk, trust in precision medicine).

- **Value**: far from being based solely on their generation, the value of genomic data depends on the **meaningful and innovative analyzes** carried out.

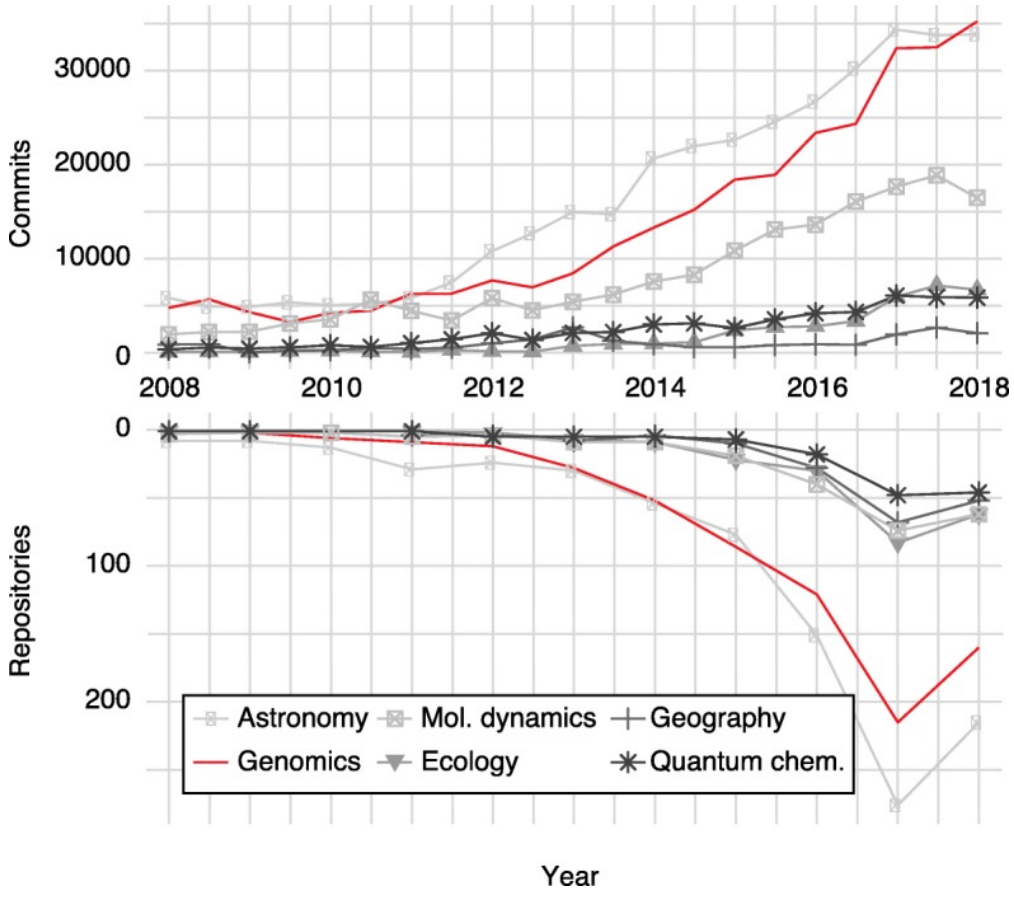


Computational tools in genomic science

- **Open source tools**:
  - developed by **academic laboratories**, currently considered to be those capable of better managing genomic data due to their precision and their innovative property;
  - commonly used in the clinics, despite the problem of **software certification**. All results must go through interpretation and validation by a clinician.

- **Abundant software ecosystem**:
  - the **relative youth** of the genomic field;
  - **variety and complexity** of the aspects of the genomes investigated.

- **One-to-many relationship**: a sample of raw data will be **transformed into several processed data** to address all the molecular questions raised in clinic.



Source from Navarro et al., Genome Biology 2019.

Implementation of a research project based on genomic data

01

IDENTIFY KNOWLEDGE AND DATA

A **multi-disciplinary consortium** of biomedical experts, in biology and clinics, and data experts, in bioinformatics and AI, defines the knowledge and data foundations necessary to implement a project.

02

COLLECT DATA

Data experts, assisted by **digital rights lawyers**, carry out the procedures for accessing the data, their conditions of use to meet the objectives of the project, and assess the dimensionality of the storage IT infrastructure.

03

HARMONIZE DATA

Data experts reprocess large-scale genomic data using workflows of bioinformatics methods and high-performance computing infrastructures.

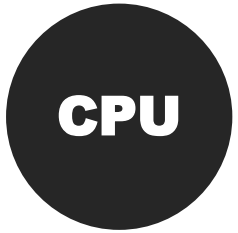
Genomic data harmonization issues

- **Sources of technical variability between data sets:**
  - preparation of the biological sample
    - sample preservation (fresh frozen / fixed with formaldehyde)
    - DNA / RNA extraction protocol
    - DNA / RNA quality and quantity
    - sequencing library preparation protocol
    - high-throughput sequencer used to generate the data
  - heterogeneous processing (different tools and parameters) of sequencing data

- **Bioinformatics reprocessing of raw sequencing data** from data collections gathered within a research project remains the best way to eliminate technical biases at source.

- **Challenges** related to the choice of **IT infrastructure** and **data flow strategy**:
  - data volume;
  - compatibility between bioinformatics methods and hardware;
  - energy footprint.

Hardware infrastructures in computational genomics



CENTRAL PROCESSING UNITS

- **Gold reference implementation** of algorithms in computational genomics.
- Fit the technical skills of young researchers who supply the majority of algorithms used in research.
- Versatile deployment of calculations locally and on high-performance infrastructures.

GRAPHICAL PROCESSING UNITS

- First ports of methods not adopted by the research community because of lack of reproducibility with results obtained by CPU implementation.
- **Nvidia Parabricks** suite offers reimplementation of popular bioinformatics algorithms initially developed for CPUs.
- Popularity of this hardware due to its ease of use, wide applicability and cost effectiveness.

FIELD PROGRAMMABLE GATE ARRAYS

- **Illumina Dragen**, an edge computing platform, coupled with the DNA sequencer, offers accurate and fast bioinformatics algorithms.
- Large raw data files being processed on the fly, not need to be transferred to another IT infrastructure.
- The weakness comes from the software licensing costs which prevent its adoption by academic researchers and companies with limited resources.

Energy footprint in computational genomics

- **Computational genomics** requires the **analysis of large and complex datasets**, requiring increasingly large computational resources.

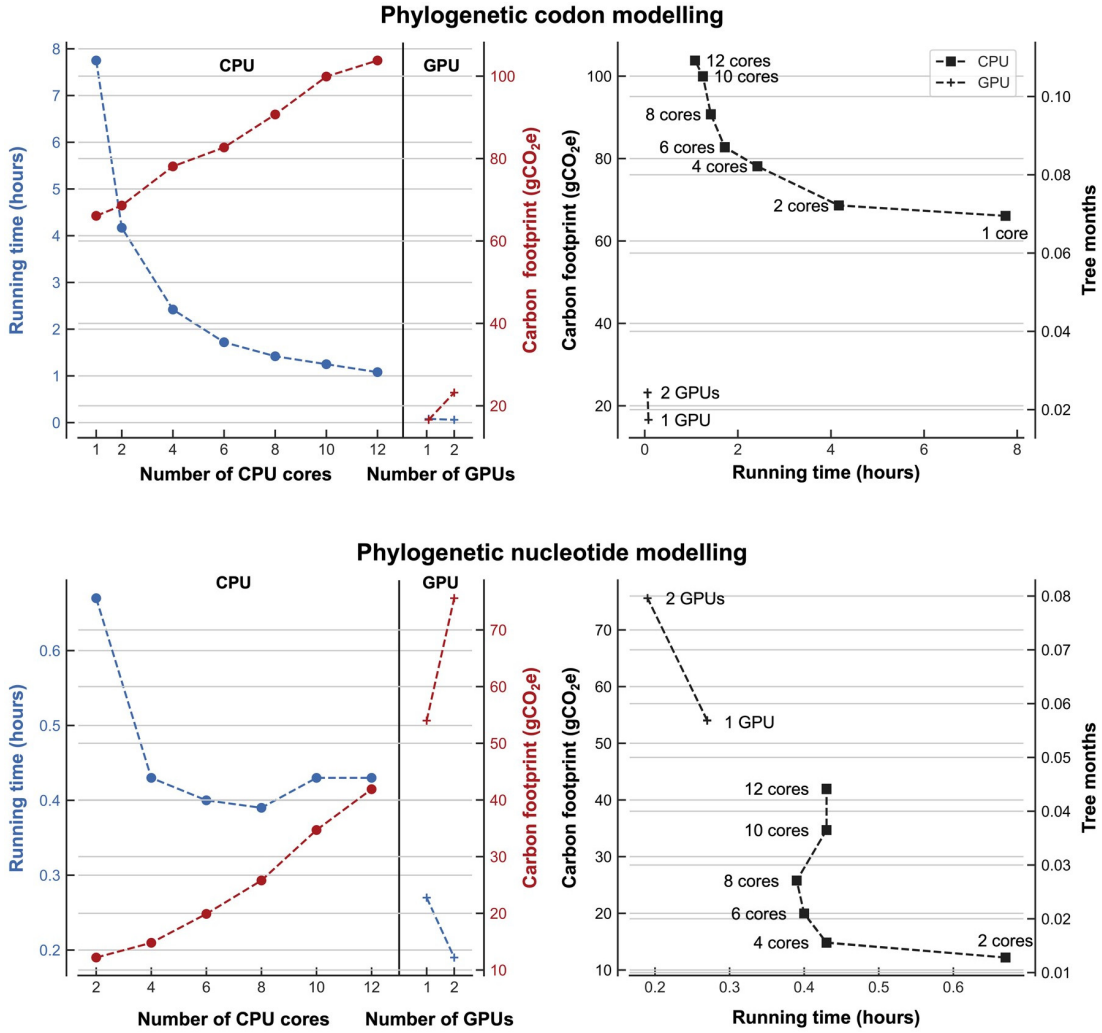
- Although genomics research enables better understanding and treatment of diseases, IT infrastructure energy consumption causes **CO_2 emissions** that have a detrimental impact on **global warming and human health**.

- Research studies have started to investigate the **environmental impact of calculations in several scientific fields**: since 2019 for machine learning, since 2020 for astrophysics and **in 2022 for computational genomics** (Grealy et al., Mol. Biol. Evol 2022).

- To **facilitate results interpretation**, carbon footprints were often compared to:
  - carbon emission produced by an average European car (0.175 kgCO_2e/km)
  - amounts of carbon sequestered by a tree (0.917 kgCO_2e per month)

Energy footprint in computational genomics

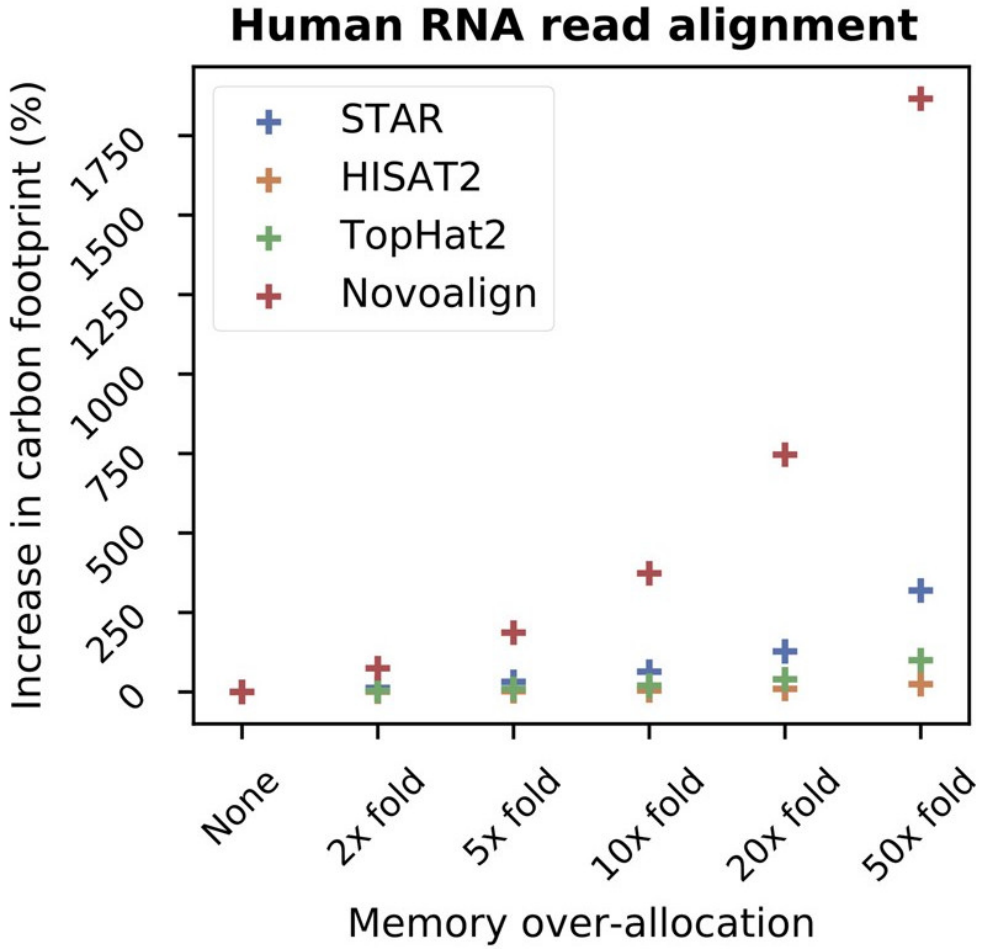
- **The impact of processors and parallelization:**
  - often results in **tradeoffs** between runtime and carbon footprint;
  - **identify on a case-by-case basis** the best association between the carbon footprint and the number of cores used (non-linear relationship).
  - the impact of the **GPU** computing on the carbon footprint is highly variable depending on the algorithm and its optimization.



Source from Grealey et al.,Mol. Biol. Evol. 2022.

Energy footprint in computational genomics

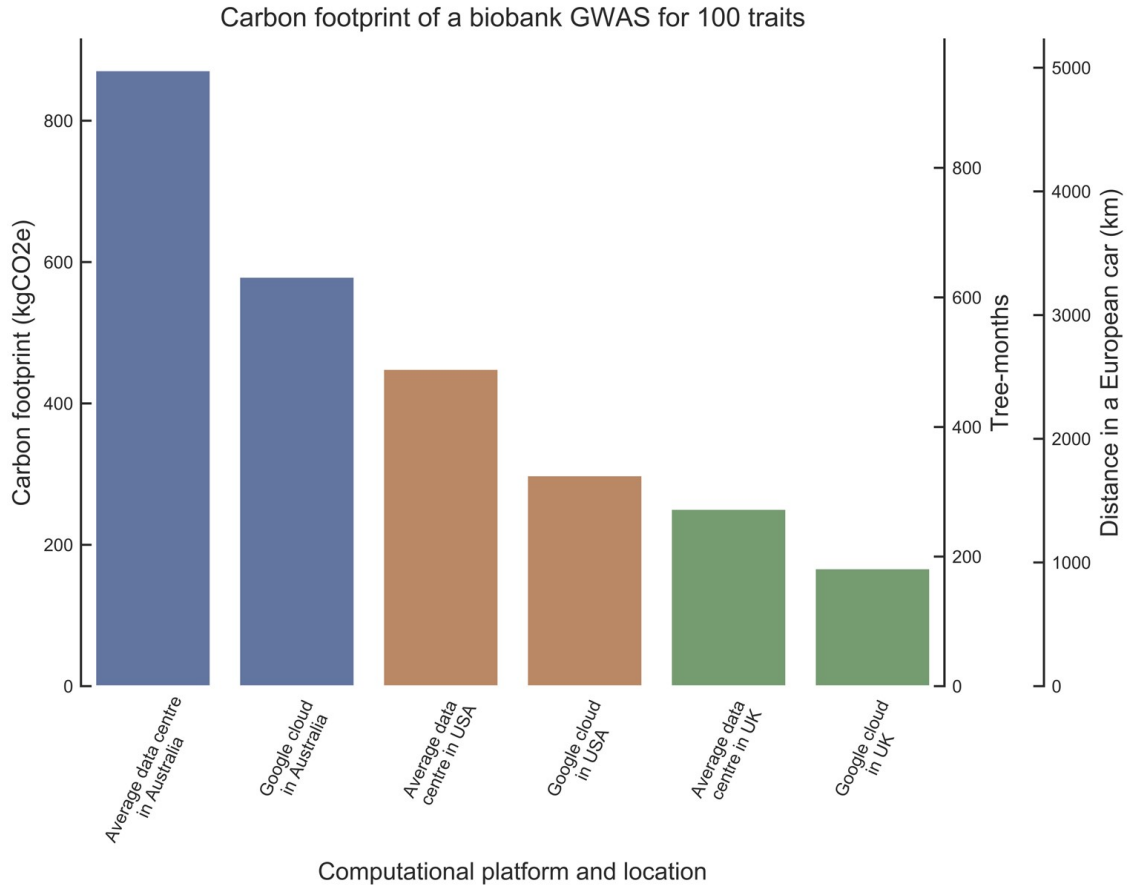
- **The impact of memory:**
  - energy consumption depends mainly on **available memory, not on used memory**. Thus, having too much memory available for a task leads to an unnecessary increase in the carbon footprint.
  - **identify on a case-by-case basis** the minimum memory allocation needed to perform a given calculation.



Source from Grealey et al.,Mol. Biol. Evol. 2022.

Energy footprint in computational genomics

- **The impact of geography:**
  - Since **large data centers** are optimized to reduce overall power consumption, such as cooling, they are therefore often more energy efficient than smaller installations.
  - The **energy mix** of the country producing the electricity supplying the computing center on which the calculations will take place will influence the carbon footprint (www.green-algorithms.org).



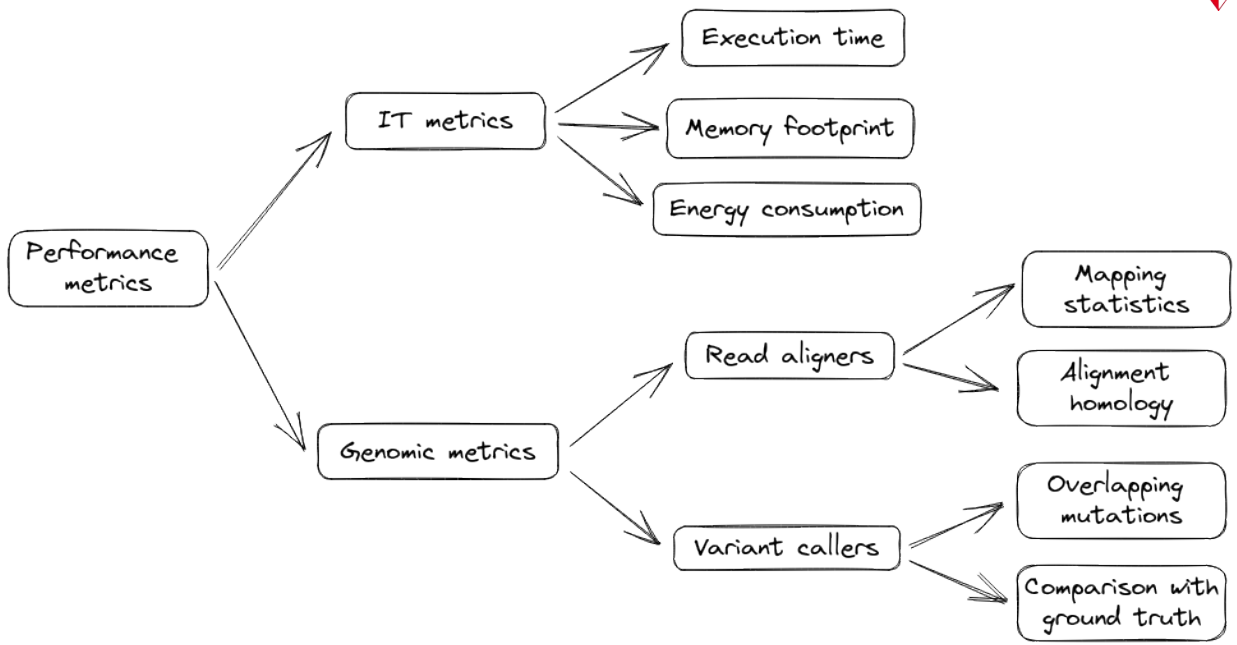
Source from Grealey et al.,Mol. Biol. Evol. 2022.

Energy footprint in computational genomics

- **Comprehensive evaluation** of bioinformatics methods for processing raw transcriptome sequencing data:
  - execution time
  - memory footprint
  - energy consumption

- **Comparison of original academic CPU versions to NVIDIA's GPU implementations** available in their Clara Parabricks suite.

- Work group:
  - **CEA**: P. Bazelle (Ing. bioinfo.), E. Bardet (M2 bioinfo.)
  - **NVIDIA**: R. Abdelkhalek (Ing. HPC), DGX A100 server
  - **DENERGIUM**: H. Mathieu



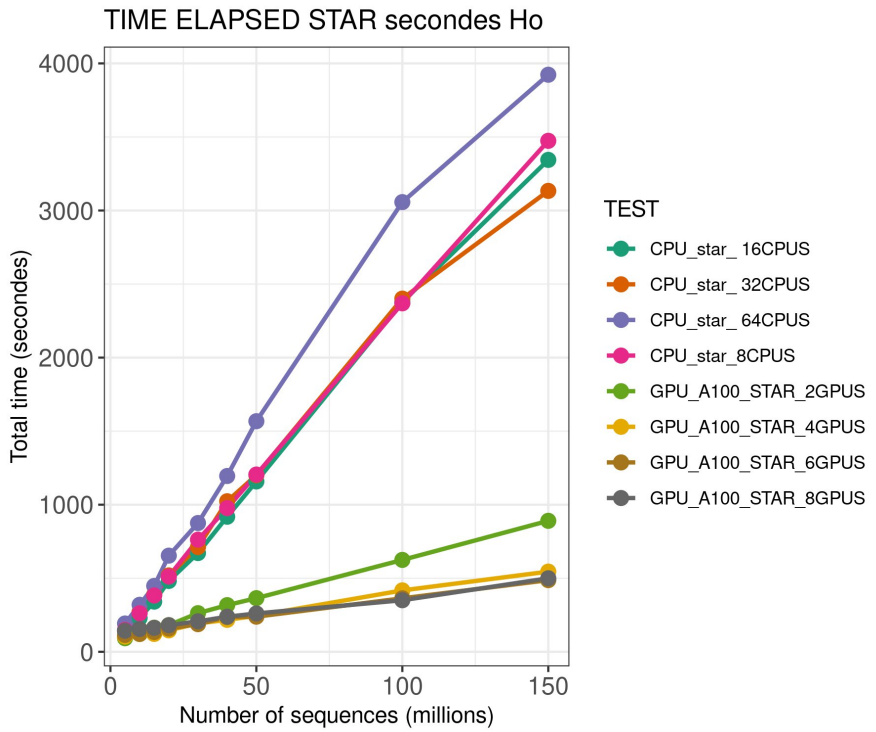
Pauline Bazelle

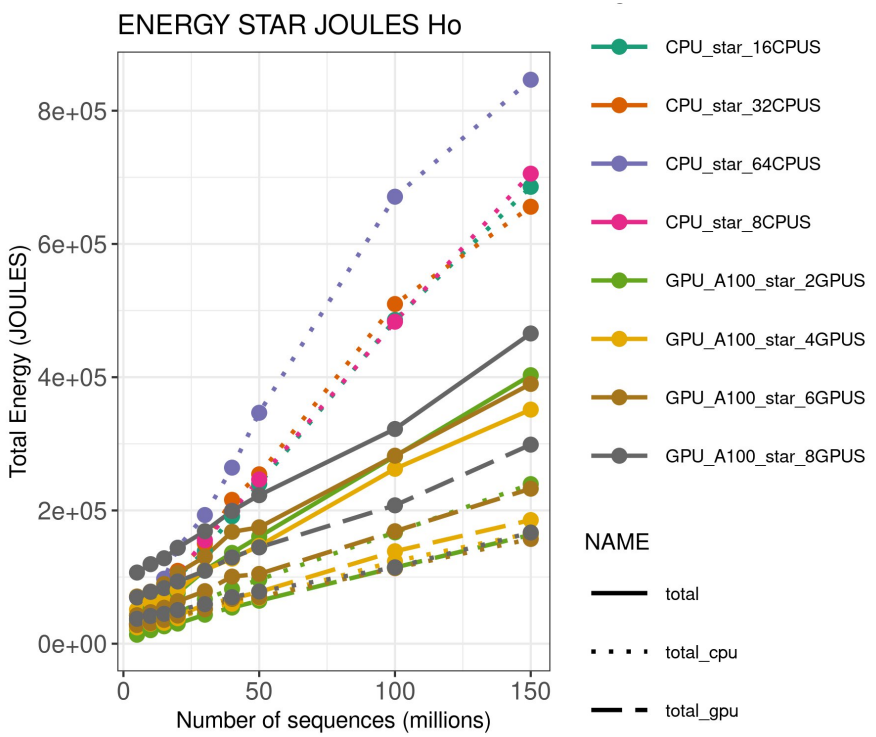
Etienne Bardet

NVIDIA

DENERGIUM

Energy footprint in computational genomics





GPU versions of **bioinformatics alignment and variant caller methods for transcriptome** sequencing data allow:

- considerably reduced calculation times;
- more frugal energy consumption;
- reproducibility with CPU methods reference outputs.

Conclusions

- **Genomic sciences** are now in the **Big Data** field with the prospect of becoming in 10 years one of the most data-producing scientific fields.

- The strong activity of the computational genomics research community results in an **avalanche of published methods** exclusively evaluated, for the moment, on their **innovative character** and **not on their energy footprint**.

- Recent interest, over the past two years, of energy issues in computational genomics. Work must be done to **identify and implement good practices** on data and calculation distribution strategies (centralized / distributed, choice of hardware).

- An evolution of the **Data Management Plan**, a document now systematically requested by funding agencies at the start of a research project, **integrating energy footprint aspects** could **raise research stakeholders' awareness** of this issue and encourage them to take them into account in their projects.

Merci

Christophe Battail

- christophe.battail@cea.fr
- https://www.bge-lab.fr/genchem

CEA GRENOBLE

17 avenue des Martyrs

38000 Grenoble, France