Management of large genomics data with free software

Michele Finelli
m@biodec.com
BioDec

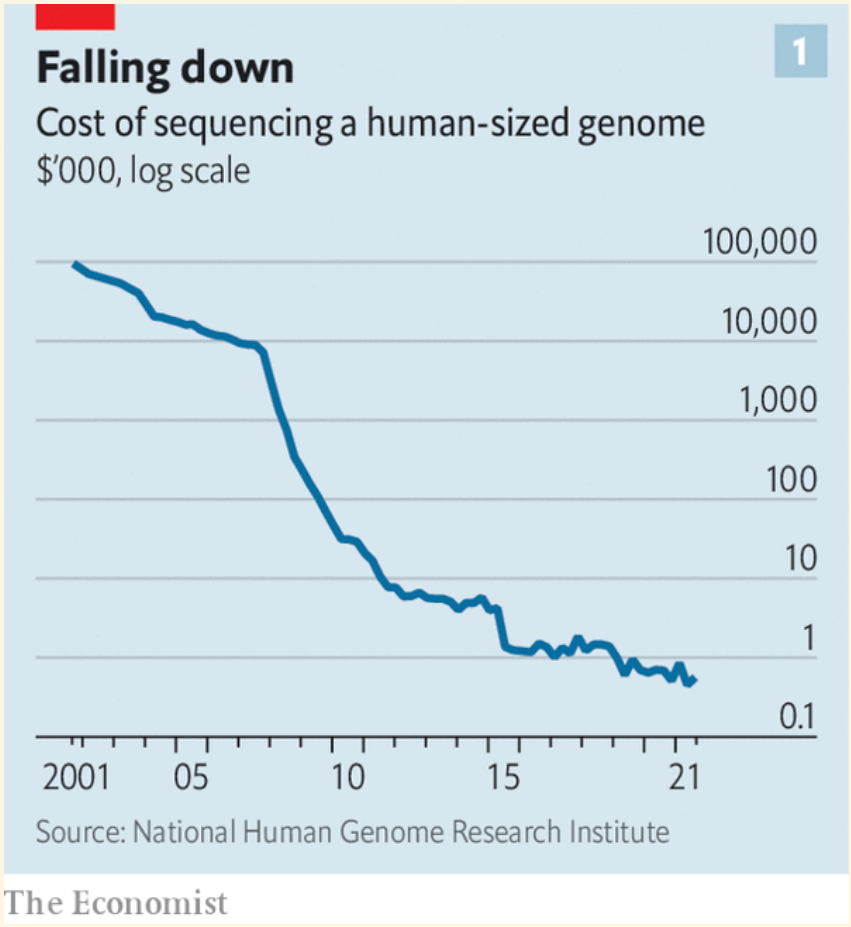
BIODEC
Evolving ICT Infrastructures

Index

The NGS revolution

- The last twenty years have seen an amazing improvement in the capability to perform human (and not-human) genome sequencing.
- These so called *Next Generation Sequencing* (NGS) technologies have revolutionized the world of biology and medicine.
- The Apr 13th 2023 Economist article "Epic ambition" summarizes the current status and shows how the cost of the complete sequence of a human genome went from over the 50.000.000$ of the Human Genome Project (HGP) to few hundreds of dollars.

Cost of sequencing a human-sized genome



The sequencing process

- Using physical properties, the NGS device converts the information of a biological sample (blood or tissue, for example) into text sequences (*reads*) representing elements (*bases*) of the genome or of other genetic material.
- Different sequencing techniques create reads of different lengths: some technologies are more suited to shorter reads, others to longer ones.
- Given the error rate of sequencing technologies — which is a *physical process* — the same part of the genome is read multiple times, leading to many reads for the same region (that is the *coverage* of the genome).

Size of sequences

Application	Estimated output
Human whole-genome sequencing (at 40x coverage)	~120 Gbases
Human exome sequencing (at 100x coverage)	~8 Gbases
Microbial whole-genome sequencing	~300 Mbases
16S rRNA sequencing	~60 Mbases

Table: Data output for common NGS applications. 1 megabase (Mbases) = 1.000.000 bases; 1 gigabase (Gbases) = 1.000.000.000 bases

Index

Storage and computation

- Genomes must be **stored** in secure and **meaningful** way.
- The value of data is in the **analysis** that can be done on them.
- There are computational processes — called **pipelines** — that produce the **annotations** that people will use to interpret the genomics data.
- Annotations (*i.e.* the genome metadata) must be accessed, often with **graphical interfaces**.

Meaningful storage

- Genomics data is *large* (*i.e.* many TBs) and in some cases even *huge* (many PBs).
- Structuring a large amount of data is a **complex task**.
- Also, you have the usual issues of large data management: backup, disaster recovery, etcetera.

Pipelines execution

- Tools that analyze the data *take a long time*: it is not unusual that they take days.
- Access to computational data must be as fast as possible: but fast storage is small, *i.e.* it cannot accommodate large data.
- Balancing this requirement is a **complex task**.

Access to annotations

- Annotations created by the computational process are intrinsically complicated to visualize and explore, because the genomics strands are *very long sequences*, with a rich and heterogeneous ecosystem of different metadata, depending on the biological or the medical questions that are asked.
- The metadata of one person's genome are *personal data* so they are subject to privacy regulations.
- Balancing UX with privacy guarantees is another **complex task**.

Index

Two suggestions

FREE SOFTWARE AT THE RESCUE !

1. The OpenCB project.
2. The Galaxy project.

The OpenCB project

- OpenCB (Open Computational Biology - https://github.com/opencb) is a project that aims to create tools to efficiently manage large genomics databases.
- These tools are not widely known or used:
  1. the software stack is complex (and it is getting even more complex with more recent releases);
  2. steep learning curve.

The OpenCB tools

OpenCGA	The storage engine: emphasis on privacy, performances and scalability (the older version uses MongoDB, the newest is Hadoop/HBase based).
IVA	The user interface (*Interactive Variant Analyser*) to OpenCGA data.
Cellbase	An integration database to manage annotations from several genomics and biological databases.

The Galaxy project

Galaxy (https://galaxyproject.org/) is an open-source platform for FAIR[1] data analysis.

- Use tools from various domains (that can be plugged into workflows) through its graphical web interface.
- Run code in interactive environments (RStudio, Jupyter, . . .) along with other tools or workflows.
- Manage data by sharing and publishing results, workflows, and visualizations.
- Ensure reproducibility by capturing the necessary information to repeat and understand data analyses.

[1] FAIR: Findability, Accessibility, Interoperability and Reusability.

The Galaxy tool

- Galaxy is a Python/Django-based web application that uses a PostgreSQL database.
- The main use case of Galaxy is as a web frontend to manage, schedule and execute computational pipelines.
- Galaxy has a large number of integrations, ranging from queuing systems (SLURM) to containers (Docker), identity management systems (LDAP, AD) etcetera.
- Galaxy has a pretty solid user base and an active community.

Summing up

Storage OpenCGA and Cellbase *as an improvement to . . .* **files** and **custom integration** (even Excel files[2]).

Pipelines Galaxy *as an improvement to . . .* **shell** and the **CLI**.

Visualization IVA and Galaxy *as an improvement to . . .* **any random tool**.

Remember: the use cases of the Galaxy and the IVA web interface are **different** so you will probably need **both**.

[2] I wish I were joking.

BioDec projects

OpenCB BioDec is currently deploying a large installation for the Genetic Unit of a main Italian hospital: data in the order of the hundreds of terabytes will be managed and analyzed by bioinformaticians.

Galaxy BioDec did many installations in the past ten years, ranging from research institutions to companies involved in drug-design and gene therapy, hospitals etcetera.

Index

Get in touch

Who BioDec S.r.l. — a bioinformatics company.

Where Physical and digital places.

Address Via Calzavecchio 20/2, Casalecchio di Reno (BO).

Web https://www.biodec.com/

Email info@biodec.com

Phone +39 051 0548263, ask about me (Michele Finelli).

THANKS !