The Genomics' BIG DATA Problem

UNIVERSITY OF CAMBRIDGE

Rosana Cortés Garrido
Bachelor's Degree in Genetics

UAB Universitat Autònoma de Barcelona

INTRODUCTION

BIG DATA is a phenomenon described by the three "Vs": **"Volume"** (massive in amount), **"Variety"** (heterogeneous and complex), and **"Velocity"** (high speed of generation) (1).

Next-generation sequencing, imaging systems and mass spectrometry-based flow cytometry are generating data at **super-exponential rate**, outpacing Moore's Law by a factor of 4 (2). One personal genome represents ~100 GB of data, and sequencing costs have dropped from $100,000 in 2001 to $1,000 in 2015 (Fig. 1) .

OBJECTIVE: to provide a synthesis of the main computing, statistical and ethical issues raised by the Big Data challenge in the field of genomics, as well as of the main approaches adopted to deal with them.

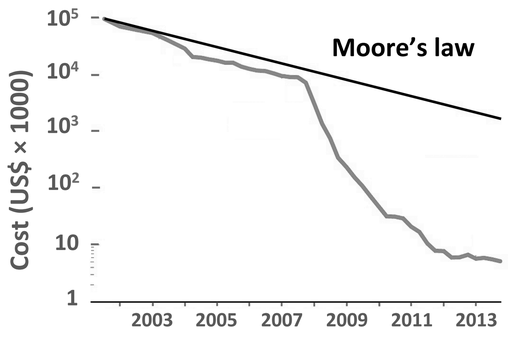

Figure 1. Evolution of the "Cost per Genome" (3).

METHODS

- Literature search on Google Scholar, Institute for Scientific Information (ISI) Web of Knowledge, NCBI PubMed and the Cambridge University Library databases, using the key words *big data*, *genetics*, *genomics*, *challenges*, *mining*, *computing*, *visualization*, *correlation*, *causation*, *statistical power*, *medicine* and *ethics*, along with keyword variations using the asterisk (*), and combinations thereof using the boolean operators AND, OR and NOT, and parentheses.
- Reading, self-reflection, information extraction and integration, and writing of a critical review.

RESULTS I. Current Trends in Data Handling

1. Data Storage: File Systems

Distributed → Single server

Cluster → Several servers + Software

Parallel → Several servers + Software + Redundancy

2. Data Processing

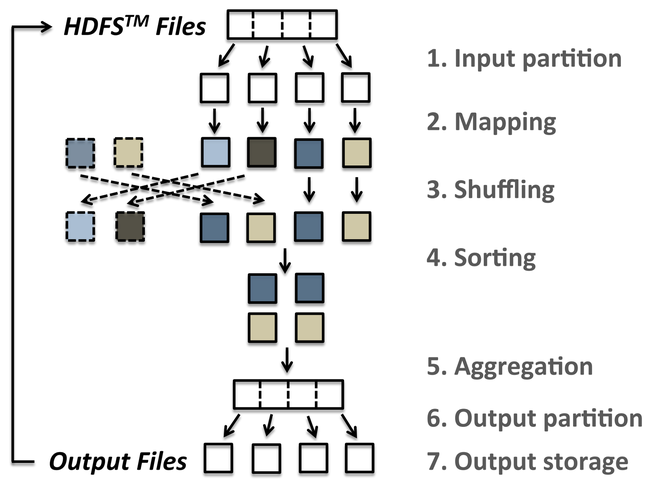

Figure 2. *Apache™ Hadoop®* strategy (4).

3. Computing approaches

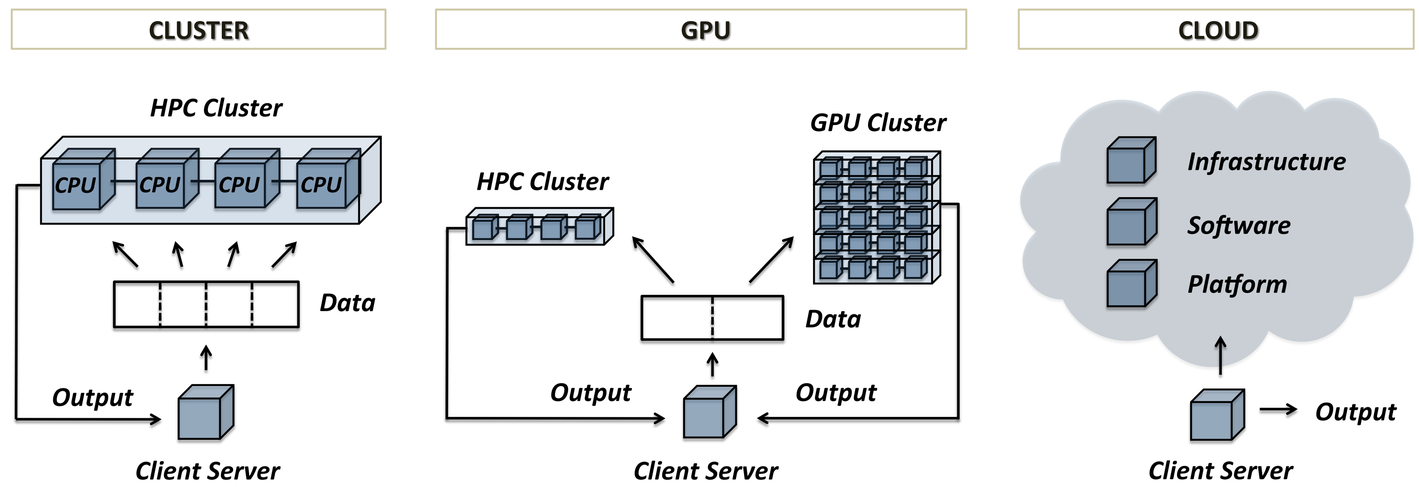

Figure 3. Computing approaches

RESULTS II. The Challenges

Computing Issues

Topic	Description	Solutions
Expertise	Lack of expertise (4).	Training.
Data transfer	Too slow to transfer terabytes (2).	Moving the computing to the data (2).
Data Analysis	Too complex.	*i. Structured Query Language* (4). *ii.Python* (4).
Applications	Lack of analytics/visualization technologies (4).	*Graphical User Interfaces (GUIs).*
Security	Loss and corruption of data due to the high distribution.	i. Digital signatures (1). ii. Username and password (1). iii.Identity and Profile as a Service.
Re-identification	Using Y-chromosome *short tandem repeats* and public information.	De-identification algorithms.
Data standardization	i. Several sources, levels and formats (2). ii. Inconsistencies between records (5).	Standard.

Statistical Issues

Topic	Description	Solutions
Correlation *vs* Causation	Correlations do not allow making predictions.	Theories.
Sampling Bias	Non-random samples.	*i. "N = All".* ii. Understanding what we ask.
Multiple comparisons	Signal:noise tends to zero (6).	Transparency.
Theory formation	Correlation-causation iterative model (Fig. 4).	Tools for theory developing (6).
Visualization Bias	False patterns appear.	Modelling (6).

Ethical Issues

Topic	Description	Solutions
Data Withholding	Secrecy (Fig. 5) (9).	i. Publication embargoes. ii. Cooperative.
Security concerns	Cloud Computing perceived as insecure.	Change of practice.
Data tenancy	i. Data trading: yes or no? ii. Limited interoperability between clouds (4).	Debate.
Health risk	Identification of mutations in public data (5).	Appropriate consent models.
Communication	Incidental finding.	Debate.
Consent models	Participant's opinions are irrelevant once the data is public.	i. Appropriate language. ii. "Citizen Science" models.

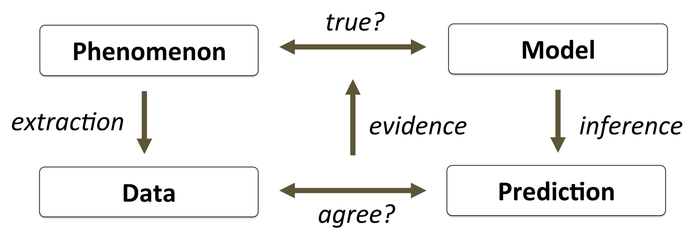

Figure 4. Model of scientific reasoning (7).

Data Withholding
Male gender
Formal instruction
Discouragement during training

Verbal
Commercial activities
Negative experiences

Publishing
Competitiveness
Industry support

Figure 5. Positively correlated factors with Data Withholding (8).

CONCLUSIONS

- **Fourth Paradigm** emerges in science: capture, curation and analysis.
- **Cloud Computing** and **parallelization strategies** are needed to move computation to the data.
- Computing Issues demand efficient use of limited money, power, space and people.
- Statistical Issues call for correlation-causation **iterative models**, and new statistical methods.
- Ethical Issues call for computational solutions, bioethical debate and **cooperation.**
- The Big Data phenomenon extends to other science fields, medicine, economy and society.

REFERENCES

1. Merelli, I., Pérez-Sánchez, H., Gesing, S. & D'Agostino, D. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. *Biomed. Res. Int.* **2014**, 134023 (2014).
2. Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L. & Nolan, G. P. Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.* **11**, 647-57 (2010).
3. Technology: The $1,000 genome; http://www.nature.com/news/technology-the-1-000-genome-1.14901 (2014).
4. O'Driscoll, A., Daugelaite, J. & Sleator, R. D. 'Big data', Hadoop and cloud computing in genomics. *J. Biomed. Inform.* **46**, 774-81 (2013).
5. Simpson, C. L. *et al.* Practical barriers and ethical challenges in genetic data sharing. *Int. J. Environ. Res. Public. Health*. **11**, 8383-8398 (2014).
6. Bollier, D. *The Promise and Peril of Big Data* (The Aspen Institute, Washington, 2010).
7. Callebaut, W. Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Stud. Hist. Philos. Biol. Biomed. Sci.* **43**, 69-80 (2012).
8. Blumenthal, D. *et al.* Data withholding in genetics and the other life sciences: prevalences and predictors. *Acad. Med.* **81**, 137-45 (2006).