GreenBIM : Environmental impact of Bioinformatics

Asmae Bachr[1], Alizée Bardon[1], Sara Bencheikh[1], Fiona Bottin[1], Justine Flipo[1], Coline Gardou[1], Sidonie Halluin[1], Bryce Leterrier[1], Meije Mathé[1], Louis Ollivier[1], Solène Pety[1], Marie Gspann[2] and Hélène Dauchel[2]@ | [1]Master's Degree of Bioinformatics, Modeling and Statistics - alphabetical order | 2020-2023 course | University of Rouen Normandy, [2]Supervisors

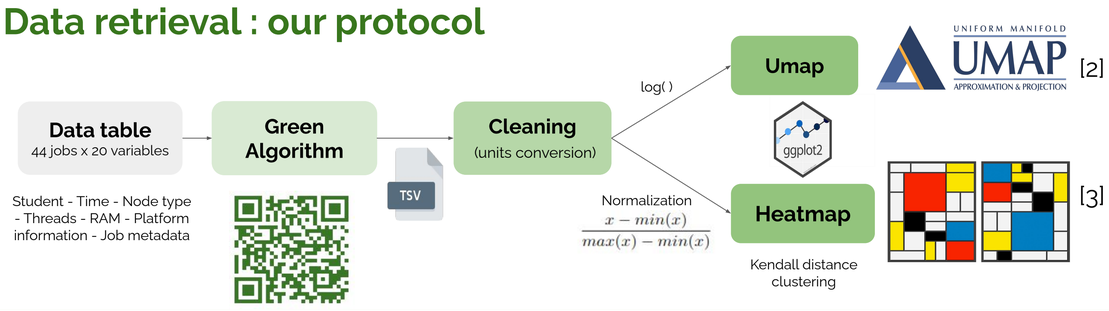
Background

Climate change involves world wide challenges for nearly all aspects of life, including health, economy and science. The University of Rouen Normandy has been awarded the **"Sustainable Development and Social Responsibility" label** for its commitment to the ecological transition. In this context, the students of the Master's Degree in Bioinformatics have questioned the environmental impact of their field of study. Actually, the impact of bioinformatic computations on global warming has generally been underappreciated despite the current climate crisis [1].

Here we present the **Bioinformatics Master students' carbon consumption** during their apprenticeship. Consumption data encompass the activities led in their bioinformatics domains (e.g. "omics", structural biology) and their range of analysis (*e.g.* genome assembly, quality control).

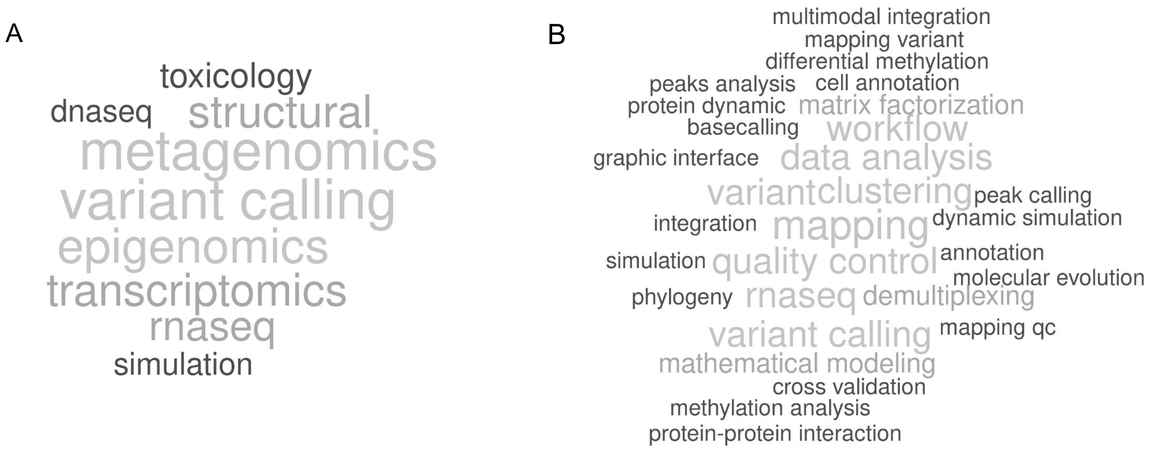
To address our question, we gathered all jobs carried out by the students for a month and determined their environmental impact using **the Green Algorithms [1]**, as described in the top right poster frame. This is a reliable proxy, as it takes into account the main computational parameters - RAM, nodes, type of computing machine - and provides the estimated carbon footprint (CF) for each computational task. We studied how these results break down through the bioinformatics domains under scrutiny.

1. Data retrieval : our protocol



2. Description of our dataset

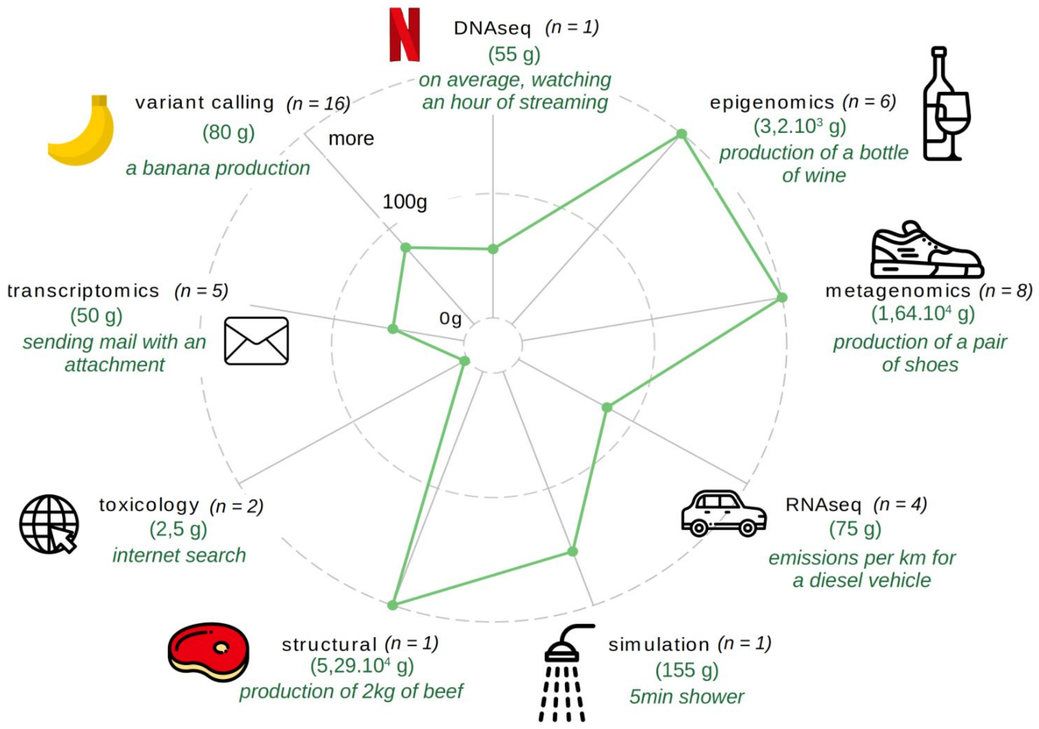
In our dataset (n=44 jobs), the main areas of bioinformatics are depicted (A). The associated analyses are diverse and represent the ones that require most resources (B).



Word clouds [4] of the (A) bioinformatics domains and (B) analyses present in the dataset. *Carbon footprint and energy consumption were measured for all the words displayed. Word size is proportional to the number of jobs for each domain or analysis.*

4. Daily life comparisons

The jobs carried out by all the students required disparate quantities of resources and spanned a wide range of environmental impacts: from 2.5g to 52.9kg of CO_2 equivalent.



Radar chart [5] of cumulative carbon footprint for each bioinformatics domain. *Equivalents in daily life activities are described across all jobs (n=44).*

Conclusion

These results should be placed in a broader context of energy consumption. Indeed, an individual's emissions are **not limited to their professional consumption**. The latter represents only a small part of the emissions: commuting, food consumption or housing generally exceed workplace consumption.

However, being mindful of the environmental impact of our activity is crucial in order to aim for a sustainable future. Indeed, the field of **big data** is expanding with the implied **consequences** - *e.g.* energy consumption and data storage. Now comes the time to **set up guidelines for a better future**. The collective "**labos 1point5**" (*https://labos1point5.org/*) is pioneering the approach by conducting a national study on the environmental impact of research activities. The group investigations will help finding levers for climate action and ecological transition.

Considering how **urgent the situation is**, making the research stakeholders aware of these issues is key to achieve the necessary sustainable goals for humanity to survive.

Carbon footprint computing [1]

Carbon footprint C (gCO2e) is computed by:

$$C = (n_c \times P_c \times u_c + n_m \times P_m) \times t \times PUE \times CI \times 0.001$$

Core energy (W)
n_c : number of cores
P_c : power used by "one" core (Watt)
u_c : core usage factor

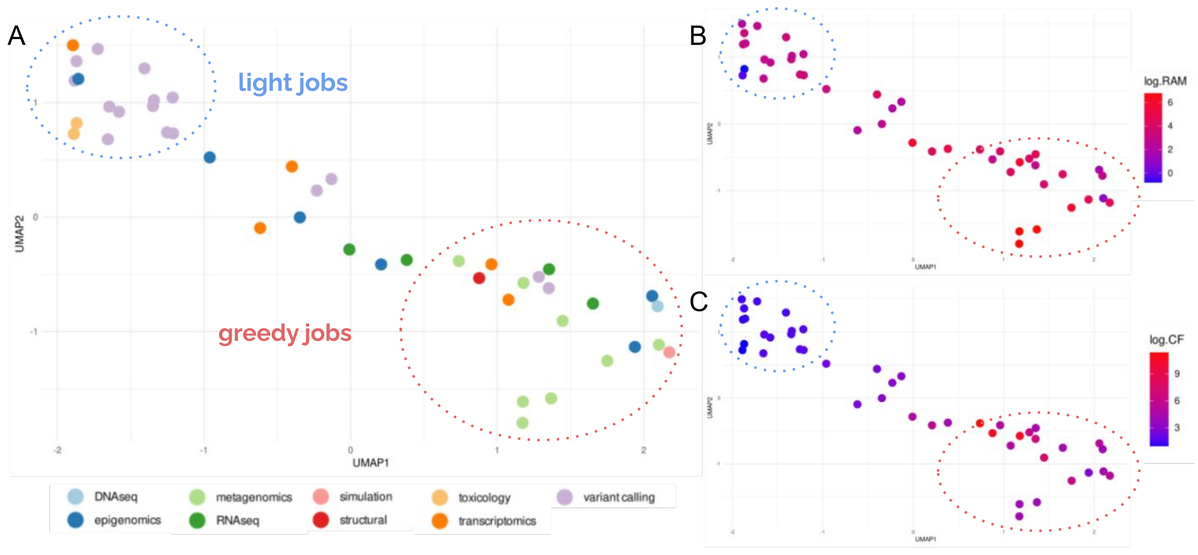
Memory energy (W)
n_m : memory available for computing (GB)
P_m : energy used by memory (Watt)

Time (h)

Power Usage Efficiency
Total power drawn by the facility on the power used by computing equipment

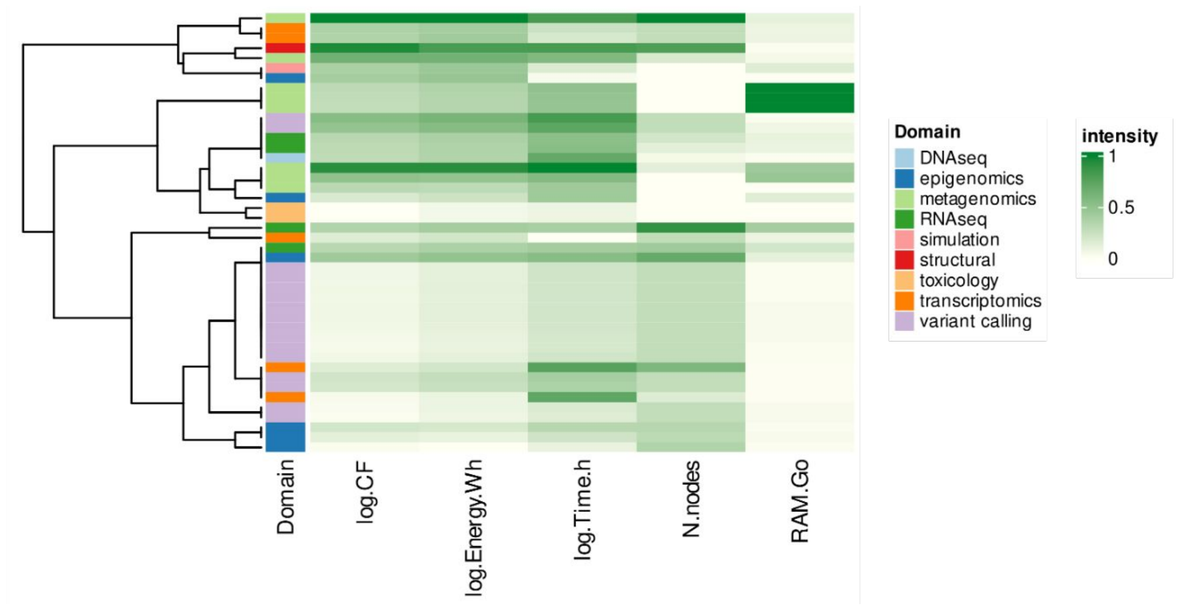
Carbon Intensity (gCO2e kWh^{-1})
estimated CO2 equivalent to produce 1 kWh

3. Results : technical resources and carbon footprint



UMAP of the 44 jobs. *Each job is colored according to (A) its attached bioinformatics domain, (B) the RAM (GB) requested and (C) its carbon footprint (CF, gCO2e). RAM and carbon footprint values were log-scaled to ease wide-ranged data visualization.*

Jobs within a given domain have varied technical requirements and carbon footprints. The environmental impact of a job depends on the type of analysis performed and on the tools used, resource allocation can therefore range very differently for each bioinformatic task.



Heatmap view of the classification for the 44 jobs according to their environmental impact and technical properties. *The data was normalized (value between 0 and 1) and the clustering was performed using Kendall distance (cf. protocol).*

The jobs having the highest carbon footprint are not necessarily the ones demanding the most resources. The balance between the amount of RAM requested and the number of nodes allocated is the main factor in the environmental impact of the jobs (*e.g.* metagenomics jobs).

5. Towards a more sustainable environment [6]

Limiting the device production footprint: keeping, maintaining and using the same device as long as possible as well as investing in more energy efficient hardware can drastically reduce our environmental impact.

Increasing code efficiency: checkpointing the code and minimizing the number of analyses. Using optimal settings and recent optimised libraries. Testing on a smaller dataset. When releasing software, making sure to advise for the best hardware choice.

Location of data centers: it is the actionable factor with the greatest impact. Choosing facilities willing to share their energy efficiency and where electricity is produced. Offsetting carbon footprint could also be a possibility.

References

[1] Lannelongue, L., Grealey, J., Inouye, M., *Green Algorithms: Quantifying the Carbon Footprint of Computation.* Adv. Sci. 2021, 2100707. https://doi.org/10.1002/advs.202100707

[2] umap: McInnes, Leland, and John Healy. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* arXiv:1802.03426.

[3] ComplexHeatmap: Gu, Z. (2016) *Complex heatmaps reveal patterns and correlations in multidimensional genomic data.* Bioinformatics. DOI: 10.1093/bioinformatics/btw313.

[4] wordcloud: Ian Fellows (2018). *wordcloud: Word Clouds.*R package version 2.6. https://CRAN.R-project.org/package=wordcloud

[5] Ricardo Bion (2022). *ggradar: Create radar charts using ggplot2. R package version 0.2.*

[6] Lannelongue L, Grealey J, Bateman A, Inouye M (2021) *Ten simple rules to make your computing more environmentally sustainable.* PLoS Comput Biol 17(9): e1009324. https://doi.org/10.1371/journal.pcbi.1009324

Acknowledgements : The students were supported by the CEA/CNRGH, the CHU Caen, the IFREMER, the INRAE, the INERIS, the Pasteur Institute of Paris, the University Paris Saclay and the companies hosting them for their apprenticeship during the Master Degree in bioinformatics at URN.