Promoting reproducibility in bioinformatics with Reprohackathons

Master AMI2B - Paris Saclay University

Frédéric Lemoine, Thomas Cokelaer, Sarah Cohen-Boulakia
GEVA / HUB Bioinformatics and Biostatistics, Institut Pasteur

2023/03/09

INSTITUT PASTEUR

Experimental variability

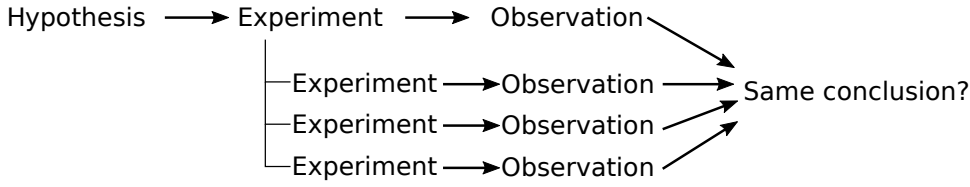
In experimental sciences, variablility of the results is mainly due to:

- Biological variations
  - Random nature of measured phenomena;
  - Different subjects, organisms, samples.
- Technical variations
  - Small changes in experimental conditions;
  - Noise of measurement tool;
  - Sample preparation.

Even with all things equal otherwise

Experimental variability: Conséquences?

- The same experiment gives different results
- The same experiment leads to the same scientific interpretation (hopefuly)



Computational variability

In data analysis: computers and programs are supposed to be exact!
⇒ perfect reproducibility? (Hint: No)

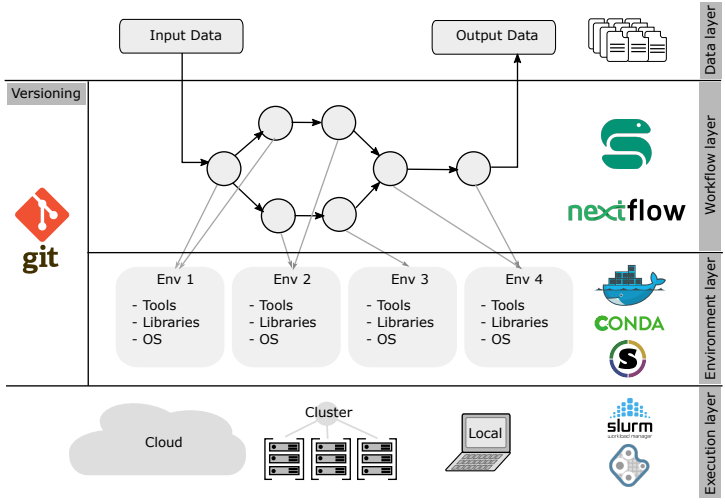
- Different versions of operating system;
- Different versions of tools used;
- Different hardware;
- Random nature of some algorithms (simulations, etc.);
- Numerical instability;
- Parallel algorithms;
- Poor method description;
- etc.

Computational reproducibility

We define several *levels* of reproducibility (Cohen-Boulakia et al., FGCS, 2017):

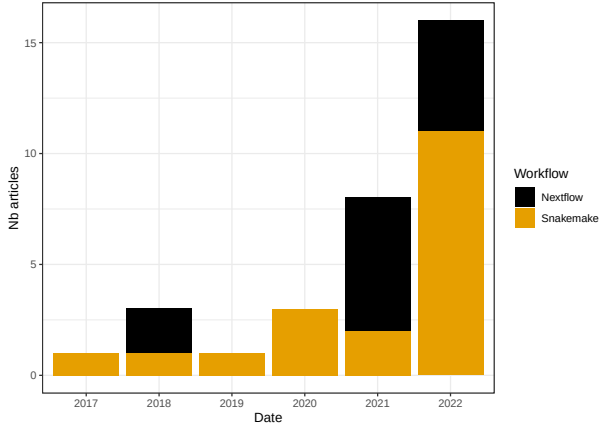
- Repeat: The data analysis experiment is performed in the exact same computational setting as the original experiment. In that case, results should be exactly the same without any variation. This necessitate to gather as many information as possible about the initial experiment, i.e. all tools versions, all operating system library versions, the state of the random number generator, etc.;
- Replicate: The data analysis experiment can be performed in a slightly different environment (different tool versions, different library versions, different random seeds, etc.), but the general protocol remains the same. In that case, results are not exactly the same, but scientific interpretation should be identical;
- Reproduce: The data analysis experiment aims at validating the scientific hypothesis, and can be performed in a different environment and with a different protocol (different tools, different workflow, etc.). This level of reproducibility gives us the best level of confidence about the quality of the results.

Available tools



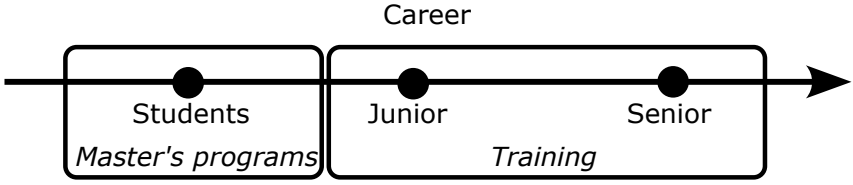
Actual usage?

Number of mentions of "Snakemake" or "Nextflow" in Nature or Science:



More and more data intensive biology + more and more adoption of workflow systems.
But : Not there yet!!!

How to promote better practices for data analysis?



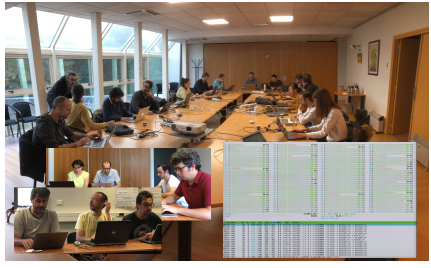
- Trainings for biologists & bioinformaticians practitioners (CNRS courses, Pasteur course, etc.)
- Workshops, Hackathons
- Dedicated Master's programs!

First reprohackathons (GDR MADICS)

https://ifb-elixirfr.github.io/ReproHackathon/
Goal: To test the capacity of current workflows to reproduce a published scientific experiment, in **2 days**.

- 2017 - Gif sur Yvette: RNA-Seq data analysis workflow;
- 2018 - Lyon : Comparison of phylogenetic tree inference programs;
- 2019 - Montpellier: High-throughput plant phenotyping image analysis.

Workflows: Snakemake, Nextflow, CWL, Galaxy, etc.



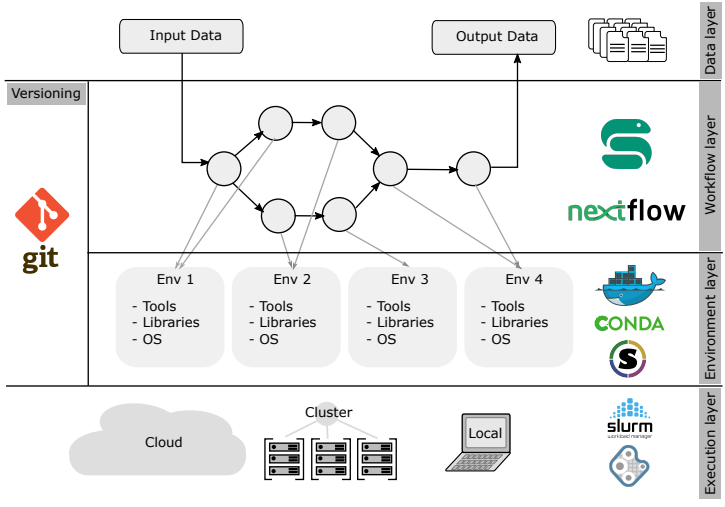
Reprohackathons: program for Master's students

Master Program since 2020

Reprohackathon

- 3-4 months program (Sept-Dec)
- 1st part: Theoretical courses + Practicals (Containers, Workflows, versioning, etc.)
- 2nd part: Intense project (Reprohackathon)

Theory + practicals



Project

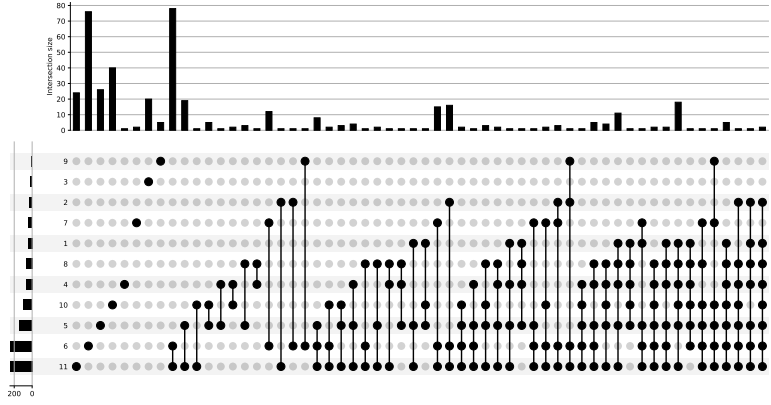
- We give an article analyzing large sequencing dataset (RNA-Seq)
- Students need to read and understand the article and the data analysis
- Then in autonomy and in small groups, they have to :
  - Re-implement it as a reproducible workflow using the ecosystem (Containers, HPC, Nextflow/Snakemake, git, etc.)
  - Run it on a cloud infrastructure (IFB Cloud)
  - Interpret and compare their results

Evaluation is based on:

- Code readability
- Our ability to run the data analysis without issue
- Final defense

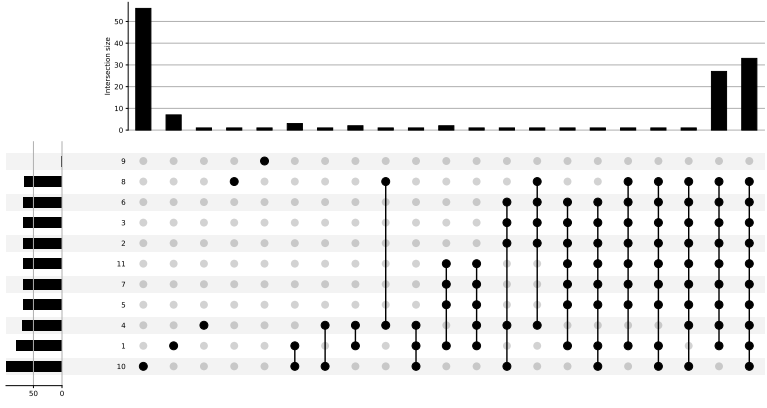
Interesting findings

At a first look, students results were poorly reproducible:



Interesting findings

But if we take the final statistical analysis apart (a lot of decisions to take here):



Thank you for your attention