FAIR_bioinfo : Open Science and FAIR principles in a bioinformatics project

How to make a bioinformatics project more reproducible

C. Hernandez[1] T. Denecker[2] J. Sellier[2] G. Le Corguillé[2]
C. Toffano-Nioche[1]

[1]Institute for Integrative Biology of the Cell (I2BC)
UMR 9198, Université Paris-Sud, CNRS, CEA
91190 - Gif-sur-Yvette, France

[2]IFB Core Cluster taskforce

June 2021

General information

Practical information:

- Dates: June 28th - 30th
- Location: Institut des Systèmes Complexes, 113 rue Nationale, 75013-Paris
- Courses: 9:00 to 17:30
- Meal: 12:30-14:00
- Pauses: 10:30-11:00 + 15:30-16:00
- 2 days of courses + 1 day of course building

Round table:

- Teachers
- Learners

Ressources:

- cc BY NC SA
- GitLab
- LaTeX

Training schedule

Day 1:

- Introduction to reproducibility
- History management (3 Practical Sessions, git, GitHub)
- Control your development environment (1 PS, CONDA)
- Encapsulation (2 PS, docker)

Day 2:

- Workflow (2 PS, SNAKEMAKE)
- Traceability with notebooks (2 PS, jupyter, markdown)
- IFB resources (2 PS, slurm, S)
- Sharing and disseminating (GitHub, zenodo)
- Conclusion

Day 3:

- Empowerment and improvement of resources

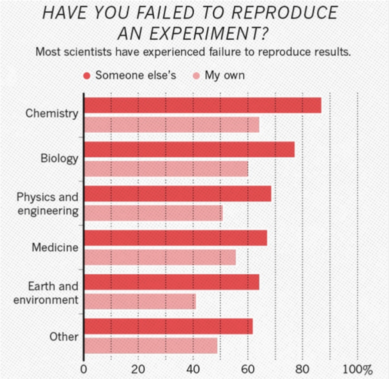
Table of contents

1. Introduction to reproducibility
   - Reproducibility
   - A solution
2. History management
3. Control your development environment
4. Workflow
5. Tracability with Notebook
6. IFB resources
7. Sharing and dissemination
8. Conclusion
9. 3rd Day

Reproducibility

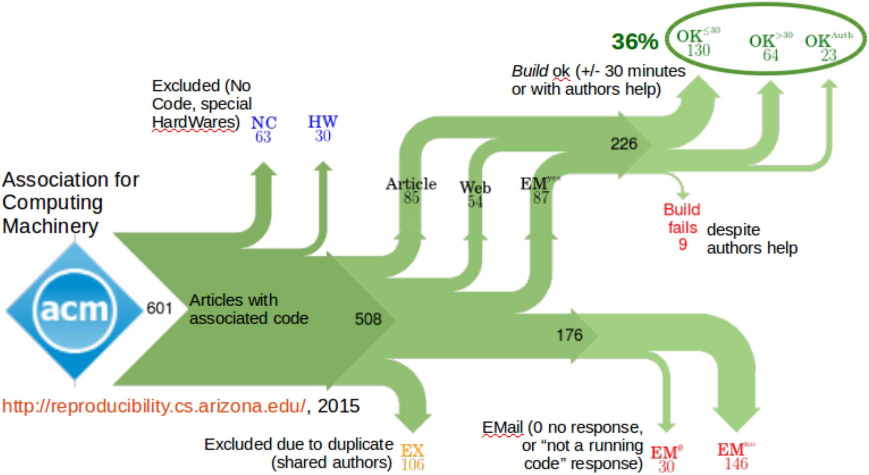
A reproducibility problem, Biology

70% of the analyses in Experimental Biology are not reproducible



Monya Baker, 1,500 scientists lift the lid on reproducibility, *Nature*, 2016

A reproducibility problem, Computer Sciences



Association for Computing Machinery

acm

601

Articles with associated code

508

Excluded (No Code, special HardWares) NC 63 HW 30

Article 85

Web 54

EM^{yes} 87

226

Build ok (+/- 30 minutes or with authors help)

36% $OK^{\leq 30}$ 130 $OK^{>30}$ 64 OK^{Auth} 23

Build fails despite authors help 9

176

EMail (0 no response, or "not a running code" response) EM^{0} 30 EM^{no} 146

Excluded due to duplicate (shared authors) EX 106

http://reproducibility.cs.arizona.edu/, 2015

A reproducibility problem, Bioinformatics

TEN YEARS REPRODUCIBILITY CHALLENGE
RESCIENCE SPECIAL ISSUE
FREE TO READ - FREE TO PUBLISH
Workshop
June 22, 2020
BORDEAUX
Would you dare to run the
code from your past self ?
(the one that does not answer mail)
SUBMISSION DEADLINE 01/04/2020
http://rescience.github.io/ten-years
In association with Inria, CNRS, Software Heritage, ReScience, Comité pour la Science Ouverte,
URFIST Bordeaux & Mission de la pédagogie et du numérique pour l'enseignement supérieur.
Contact: nicolas.rougier@inria.fr

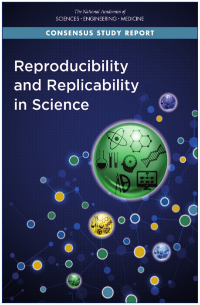
Ten-Year Reproducibility Challenge, Konrad Hinsen
Can your 2009 code still run?
special issue of ReScience and result comments in *Nature*

Who's never wanted to take over a protocol, a pipeline, or a tool without running into it?

- unable to install tools: not compatible OS, not availability of dependencies
- tool update ⇒ codes unusable: python 2 vs. 3, change of function arguments (R)
- inability to reproduce the results of computational analysis: package versions, IDE: stable version of the language different according to the OS (Rstudio)

Reproducibility in science

Reproducible research, Repeatability, Replicability, Reproducibility, Replication: overlapping semantics ⇒ a plethora of definitions![a]



The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

Reproducibility and Replicability in Science

National Academies of Sciences, Engineering, and Medicine (2019).[b]

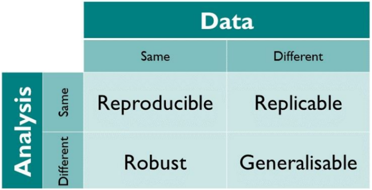
ACM definition (2016):

Repeatability Same team, same exp. setup

Replicability Different team, same exp. setup

Reproducibility Different team, different exp. setup

Whitaker's matrix of reproducibility (2017):[c]



		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

[a] https://www.researchgate.net/publication/323118701_Terminologies_for_Reproducible_Research
[b] National Academies of Sciences, Engineering, and Medicine. 2019. Washington DC. The National Academies Press, https://www.nap.edu/read/25303/chapter/1
[c] https://doi.org/10.6084/m9.figshare.5443201.v1, Slide number 7

FAIR_bionfo's finding

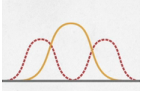
Depends on the object of study x

what needs to be "memorized" to replay the experience:

Raw Data

FAIR data principles & Data Management Plans

→



Statistical or bioinformatic analysis

Codes - parameters - workflows

→

Validation

Publication: thesis, article, report, etc

How to gain in reproductibility?

Focus on codes, parameters, and workflows used throughout the analysis process

Monya Baker, 1,500 scientists lift the lid on reproducibility, *Nature*, 2016

A solution

Divert FAIR data principles towards processes

Findable

Third party tools used = ref. in their field

.

Easy to find analysis protocol (Github pages)

Accessible

Available codes (Github, dockerhub)

.

Third party open source tools

Interoperable

Cooperation of tools (snakemake, docker) as well as locally than on servers (cloud or cluster)

Reusable

Protocol replayable (snakemake) identically (Rshiny) in a virtual environment (docker)

Promote learning

FAIR bioinfo

Our objective

FAIR raw data

$+$

FAIR scripts

$=$

FAIR processed data

Course

Take your first steps with several companion tools to gain in reproducibility

Example based

Just the beginning of an NGS analysis
A full analysis is given as bonus (The NGS analysis is simply used as an example and not explained)

Ressources

- awesome a curated list of reproducible research case studies, projects, tutorials, and media
- The Role of Metadata in Reproducible Computational Research
- Towards reproducible computational biology
- A very similar sweden courses with git, conda, snakemake, jupyter, r-markdown, docker, singularity