FAIR_bioinfo : a software perspective of reproducibility of bioinformatics analyses

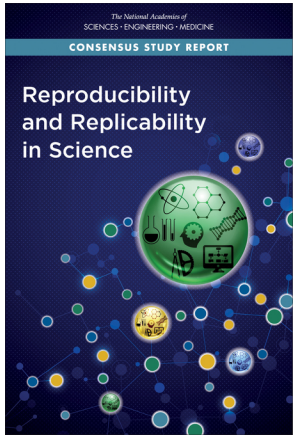
Thomas Denecker, Céline Hernandez, Claire Toffano-Nioche

ANR-11-INBS-0013

FAIR_bioinfo: a software perspective of reproducibility of bioinformatics analyses

Thomas DENECKER, Céline HERNANDEZ, Claire TOFFANO-NIOCHE

Thomas DENECKER



Reproducibility context

[1] Baker M. *Nature*, 533(7604):452-4, 2016
[2] National Academies of Sciences, Engineering, and Medicine. 2019. Washington, DC: *The National Academies Press*. https://doi.org/10.17226/25303.
[3] Wilkinson MD, *et al*. *Sci Data*, 3:160018, 2016
[4] Grüning B, *et al*. *Cell Syst*. 6(6):631-63, 2018

There is currently a reproducibility crisis[1] that challenges the robustness of scientific results[2] ➡ FAIR data principles[3] used to ensure data integrity

Apply similar guideline to code development of bioinformatics processes through the use of 3rd party software tools[4]

Findable

3rd party tools = references
Analysis pipelines = easy to find (GitHub pages)

Accessible

Available codes (Github, dockerhub)
3rd party tools = *open source* (conda)

Interoperable

Cooperation of tools (snakemake, docker) both locally and on servers (cloud or cluster)

Reusable

Protocol simply replayable (snakemake) identically (R package Shiny) in a virtual environment (docker)

FAIR bioinfo

Introduction training to a range of software to makes a complete bioinformatics analysis reproducible from the same data set over time

https://github.com/thomasdenecker/FAIR_Bioinfo

A multi-faceted training

Adapted to various audiences
● Biologists (8 x 1,5 hours/month)
● Bioinformatician (2 days)

Bioinformatics know-hows:
Data processing and automation (+env.) ★
Data analysis, reporting & sharing ★

Sessions	Learning objectives
Reproducibility introduction ●●	
It's not magic ●	★★ Command line usage
The code memory ●●	★★ Code versioning
Install & play with analysis tools ●●	★★ Environments
A trip to the sea ●●	★ Containerization
I've got the power! ●●	★ Parallelization, computing cluster
LoveR ●●	★ Dynamic visualization
Notebooks ●●	★★ Literate programming
Completing a reproducible project ●●	★ DOI, license

A step-by-step RNA-seq example used throughout the training

Raw data download

⇩

RNA-seq processing pipeline

⇩

Table of counts by genes

⇩

Differential analysis
&
"user" results exploration

⇩

Analysis report

Increase reproducibility with a 7-step solution

FAIR bioinfo

Reproducibility

1 - Version the code and share it

2 - Create a virtual environment

3 – Install the tools by a manager

4 - Create an analysis script

5 - Transfert the analysis to a server

6 – Facilitate the portability of results exploration

7 - Edit an analysis report

Alernative tools:

A virtuous cycle

FAIR bioinfo

thomasdenecker

FAIR raw data

+

FAIR_bioinfo scripts/protocols

=

FAIR processed data

Some limits:

FAIR_bioinfo training:

- ✗ use of an already instantiated VM
- ✓ create your own VM image

Reproducibility to the exact bit?

- ✗ container uses some resources of the support machine
- ✓ version control of the env. (Nix, Guix)

Parallelization:

- ✗ loss of computational order, multi-threading, same hardware?
- ✓ …?

Poster 213

FAIR_bioinfo : a software perspective of reproducibility of bioinformatics analyses

Thomas Denecker[1], Céline Hernandez[2] and Claire Toffano-Nioche[3]

[1] Fungal Epigenomics and Development
[2] Next Generation Sequencing core facility
[3] RNA Sequence,S tructure, and Function
Institute for Integrative Biology of the Cell (I2BC)
Université Paris-Saclay, CEA, CNRS, 91198, Gif-sur-Yvette, France

Corresponding Author: claire.toffano-nioche@u-psud.fr

1 Reproducibility context

Recent studies have established that a reproducibility crisis challenges robustness of scientific results [1] including computational biology. In this context FAIR data principles are increasingly being used to ensure data integrity [2]. To complement this principles, we introduce FAIR_bioinfo, to apply similar guideline to code development and ensure reproducibility of results obtained from the same data set over time.

Computer tools do exist that can be applied in bioinformatics [3]. Convinced of their usefulness we propose an initiation to a range of software to make a complete bioinformatics analysis reproducible.

2 The FAIR_bioinfo Training

This training is based on a concrete example of classical data analysis, a differential gene expression between two RNAseq conditions. It deals with two of the know-how of the bioinformatician's job: the automated processing of raw data (through virtualization and pipeline development) and the analysis of processed data (with environment management and "notebooks" writing).

Improvement in reproducibility can be achieved in several steps, where each step brings an additional degree of reproducibility through a specific family of software tools.

We offer training for people from different backgrounds: on the one hand, over a long period of time, for learners who do not necessarily know programming (8 months, at a rate of 1h30 per month) so that they can replay the presentations at their own pace [4], and on the other hand, for bioinformaticians, in a format condensed into 2 days. In both solutions, the training is focused on general concepts with practical illustrations.

As the importance of reproducibility is no longer to be proven, the main interests of this training is to provide practical guidelines for its daily implementation with the long-term objective for everyone to gradually adopt good practices to overcome the challenge of reproducibility in science.

Acknowledgements

This work was supported by the French Infrastructure Institut Français de Bioinformatique (IFB) ANR-11-INBS-0013

References

[1] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452-4, 2016
[2] Mark D Wilkinson, Michel Dumontier, Ijsbrand Joan Aalbersberg, Gabirelle Appleton, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3:160018, 2016
[3] Björn Grüning, John Chilton, Johannes Köster, Ryan Dale, *et al*. Practical Computational Reproducibility in the Life Sciences. *Cell Syst.* 6(6):631-63, 2018
[4] Thomas Denecker, Claire Toffano-Nioche, https://github.com/thomasdenecker/FAIR_Bioinfo, 2019