

FAIR_bioinfo for bioinformaticians

Introduction to the tools of reproducibility in bioinformatics

C. Hernandez¹ T. Denecker¹ J. Sellier² G. Le Corguillé²
C. Toffano-Nioche¹

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

²IFB Core Cluster taskforce





Sept. 2020










Conclusion

Training schedule

Day 1:

- Introduction to FAIR_bioinfo
- Encapsulation ( docker)
- Workflow ()
- IFB resources (, )

Day 2:

- History management (,  GitHub)
- Software environment management ()
- Traceability with notebooks (, )
- Sharing and disseminating (, )

Let's take a step back.

Findable



Easy to find
protocols



with DOI ()

Accessible



Open source




CONDA, ...)

Interoperable



Think "workflow"

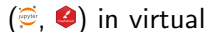


locally or on
servers ()

Reusable



Replayable
protocols



in virtual
environments
( / )

A virtuous cycle



FAIR raw data

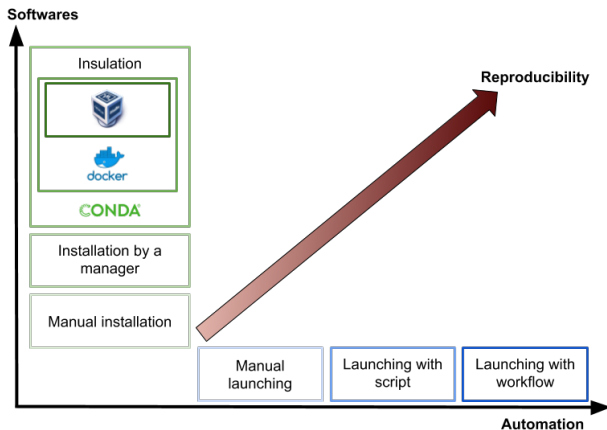
+

FAIR_bioinfo scripts/protocols

=

FAIR processed data

Reproducibility is a multi-dimensional process

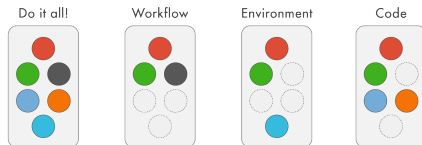




Swedish similar tutorial



[nbis-reproducible-research.readthedocs.io:](https://nbis-reproducible-research.readthedocs.io/)
Tutorials, Introduction
to the tutorial



So... What now?

What now?

Automation

- Manual
- Write a script
- Use a workflow manager

Software

- Local installation
- Package manager
- Conda environment
- Image / container
- Virtual machine

Adding Tests

Unit test: test a part of the code

```
1 ## module 1
2 sum <- function(x, y){
3     return (x+y)
4 }
5
6 # Unit test
7 sum(2,2) == 4
```

```
1 ## module 2
2 power <- function(x, y){
3     return (x**y)
4 }
5
6 # Unit test
7 power(2,2) == 4
```

Functional test: test all the code

```
1 # Functional test
2 power(sum(2,2),2) == 16
```

Continuous integration

Verification at each source code modification that the result of the modifications does not produce:

- no regression in the developed application
- nor any change in the results obtained



Travis CI



GitHub Actions

Some FAIR_bioinfo 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?
- ✓ ...?

Reproducibility checklist¹

- **Code** Enshrine computations and data manipulation in code, avoid workflows based on point-and-click interfaces (eg. Excel)
- **Document** Explain how code works, define parameters and computational environment required: comments, notebooks and README
- **Record** Note key parameters (eg. the 'seed' values of a random-number generator)
- **Test** with test functions using positive and negative control data sets, and run those tests throughout development
- **Guide** with master script (eg. 'run.sh') that downloads data sets and executes workflow
- **Archive** with long-term stability services such as Zenodo, Figshare and Software Heritage (GitHub is impermanent online repository).

¹[Nature](#)

Reproducibility checklist²

- **Track** the project's history with a version-control tools (eg. Git).
Note which version you used to create each result
- **Package** with ready-to-use computational environments using containerization tools (eg. Docker, Singularity), web services (Code Ocean, Gigantum, Binder) or virtual-environment managers (Conda)
- **Automate** the test of your code with continuous-integration services (eg. Travis CI)
- **Simplify** Avoid niche or hard-to-install third-party code libraries
- **Verify** your code's portability by running it in a range of computing environments

²[Nature](#)

Thanks

- Organizational comity (our guardian angels): Yousra, Jacques, Hélène
- IFB Core Cluster taskforce: Julien, Gildas, and those who provide in the shadows
- Helper: Paulette
- beta-testers: Pauline
- Organisations: CNRS, INRAE, IFB, I2BC, Paris Saclay University