FAIR_bioinfo for bioinformaticians

Introduction to the tools of reproducibility in bioinformatics

C. Hernandez[1] T. Denecker[1] J.Sellier[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

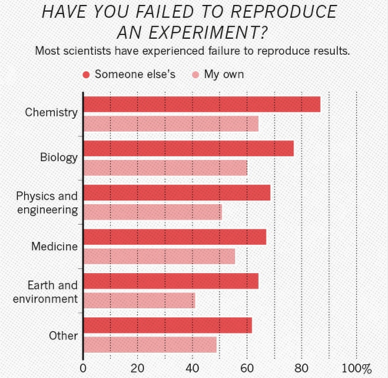
[2] Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC)
CNRS UMR 7104 - Inserm U 1258
67404 - Illkirch cedex, France

Sept. 2020

Introduction to reproducibility

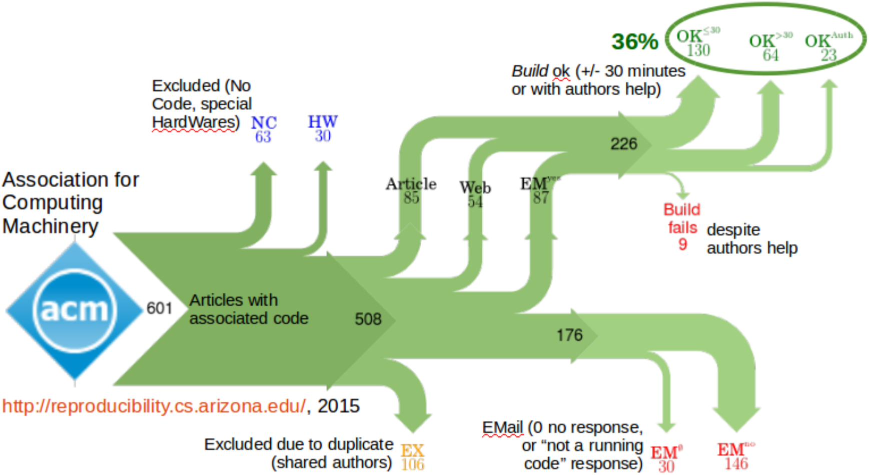
A reproducibility problem, Biology

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
36%
Build ok (+/- 30 minutes or with authors help)
OK$^{\leq 30}$ 130
OK$^{>30}$ 64
OK^{Auth} 23
Build fails 9 despite authors help
176
Excluded due to duplicate (shared authors)
EX 106
EMail (0 no response, or "not a running code" response)
EM^{0} 30
EM^{no} 146
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

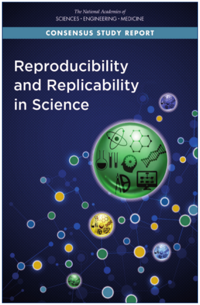
Ten-Year Reproducibility Challenge, Konrad Hinsen
Can your 2009 code still run?
special issue of ReScience and result comments in *Nature*

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]

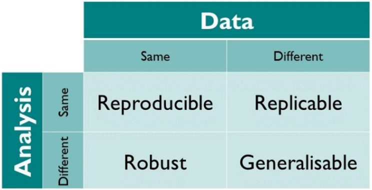

The National Academies of
SCIENCES · ENGINEERING · MEDICINE
CONSENSUS STUDY REPORT
Reproducibility and Replicability in Science

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`

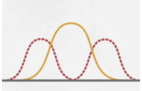
c: `https://doi.org/10.6084/m9.figshare.5443201.v1`, Slide number 7

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 - algorithms - workflows

→

Validation
Publication: thesis, article, report, etc

How to gain in reproductibility?

Focus on codes, algorithms, workflows used throughout the 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

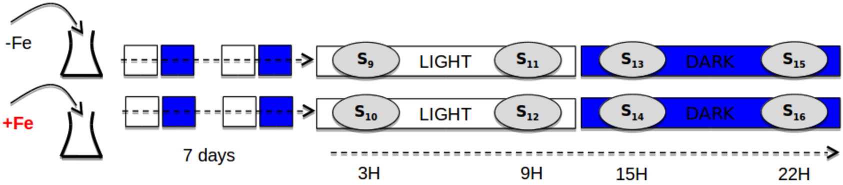
Example based

Classical RNA-seq analysis
(finding genes with differential expression between 2 conditions)
used as an example (not explained)

Working example

The biological study example

- Study of the green alga *Ostreococcus tauri* response to iron deprivation.
- 16 RNAseq samples in triplicate, single-end of 100bp.
- Choice of the 9h point of the long-term adaptative response (s11 and s12 samples):



Lelandais G, Scheiber I, Paz-Yepes J, Lozano JC, Botebol H, Pilátová J, Žárský V, Léger T, Blaiseau PL, Bowler C, Bouget FY, Camadro JM, Sutak R, Lesuisse E.
Ostreococcus tauri is a new model green alga for studying iron metabolism in eukaryotic phytoplankton.
BMC Genomics. 2016 May 3;17:319. doi: 10.1186/s12864-016-2666-6.

Reduced RNAseq Data

Genome

- sequence: GCF_000214015.3_version_140606_genomic.fna (https://www.ncbi.nlm.nih.gov/assembly/GCF_000214015.3/)
- annotation: GCF_000214015.3_version_140606_genomic.gff
- ⇒ 13.0328 Mb, 20 chromosomes, mitochondria, & chloroplast

RNAseq samples

- Project: PRJNA304086
- Selection samples 11 and 12: SRR3099585-87, SRR3105697-99 (fastq.gz ∼360M each x 6 files)
- Reads selection to reduce data volume for the course (mapped on the smalest chromosome, chr18, NC_014443.2 + 100000 first) ⇒ *_chr18.fastq.gz ∼19M each (https://zenodo.org/record/3997237)
- Counts table, complete RNAseq: https://zenodo.org/record/4008452

Data

Access in the IFB ressources

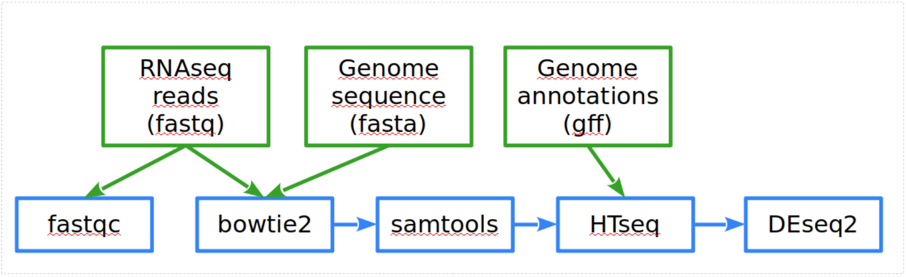
```
/shared/projects/fair_training2020/Data/
```

Or raw download in a local "Data" directory

```
mkdir Data ; cd Data
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/214/015/GCF_000214015.3_version_140606/GCF_000214015.3_version_140606_genomic.fna.gz
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/214/015/GCF_000214015.3_version_140606/GCF_000214015.3_version_140606_genomic.gff.gz
wget https://zenodo.org/record/3997237/files/FAIR_Bioinfo_data.tar.gz
wget https://zenodo.org/record/3997137/files/counts.txt
cd ..
```

RNAseq analysis

Analysis workflow



RNAseq reads (fastq)

Genome sequence (fasta)

Genome annotations (gff)

fastqc

bowtie2

samtools

HTseq

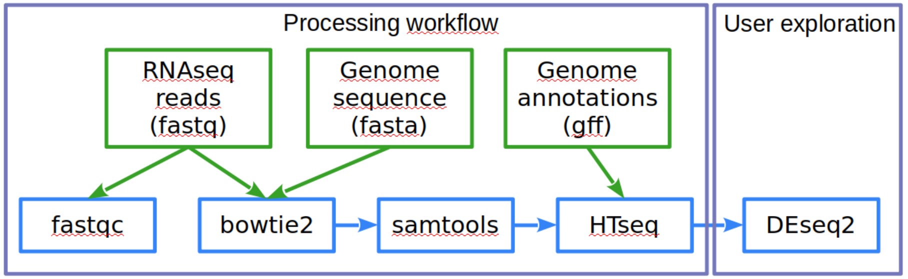
DEseq2

green=input, blue=tool

fastqc	control quality of the input reads
bowtie2	reads mapping on the genome sequence
samtools	mapped reads selection & formatting
HTseq	count table of mapped reads on genes (annotations)
DEseq2	statistical analysis: genes list having differential expression

2 bioinformatician skills

Analysis workflow



green=input, blue=tool, violet=skill

Reproducibility

Processing workflow automatization, scripting

User exploration report choices (or import choices for further analysis)

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