

A decade after the reproducibility crisis: Toward reusable data analysis workflows



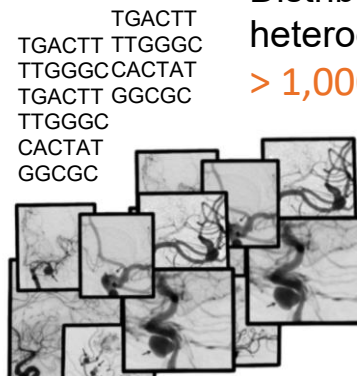
Sarah Cohen-Boulakia

LISN – Laboratoire Interdisciplinaire des Sciences du Numerique

Université Paris-Saclay

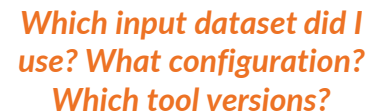


institut
universitaire
de France



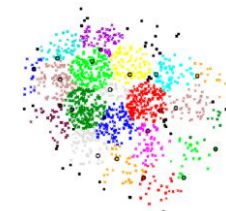
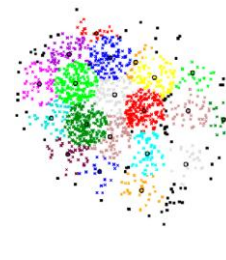
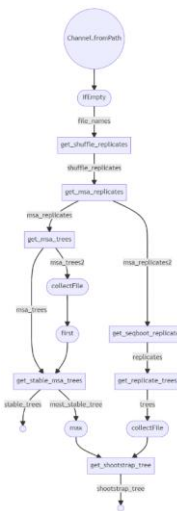
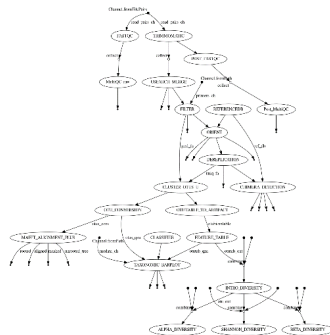
Distributed
heterogeneous network
> 1,000 (NAR)

Distributed
Heterogeneous
> 30,000
(bio.tools)



Combination of tools various environments & platforms

> 7,000 workflows
(GitHub, GitLab,)



The Reproducibility Crisis

Each community has its reference papers

Nekrutenko & Taylor - Nature Genetics (2012)

50 papers using the Burrows-Wheeler Aligner

31/50 (62%) provide no information

no version of the tool no parameters

no genomic reference sequence

7/50 (14%) provide all the necessary details

Alsheikh-Ali et al, PLoS one (2011)

10 papers in the top-50 IF journals → 500 papers

149 (30%) were not subject to any data availability policy

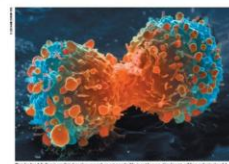
(0% data available)

Of the remaining 351 papers

208 papers (59%) did not adhere to the data availability instructions

143 make a statement of willingness to share

47 papers (9%) deposited full primary raw data online



Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

47/53 "landmark" publications could not be replicated
(Begley, Ellis Nature, 483, 2012)

Must try harder

Too many sloppy initiatives are creeping into scientific papers, at the data – and at themselves.

Error prone

Biologists must realize the pitfalls massive amounts of data.

If a job is worth doing, it is worth doing twice

Researchers need leading agencies to lead in publication or ensuring that results are reproducible, argues Jonathan F. Russell.

The case for open computer programs

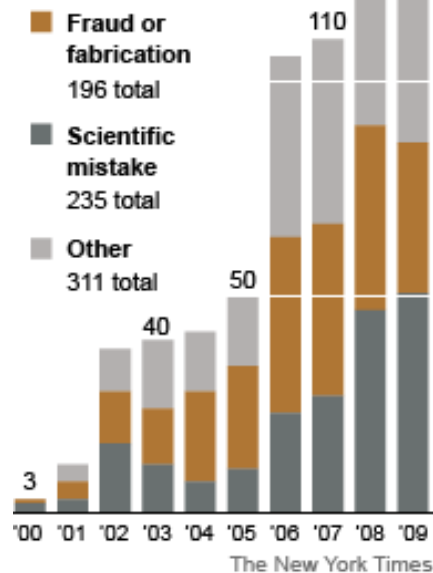
Six red flags for suspect work

C. Glenn Begley explains how to recognize the preclinical papers in which the data won't stand up.

Know when your numbers are significant

Retractions On the Rise

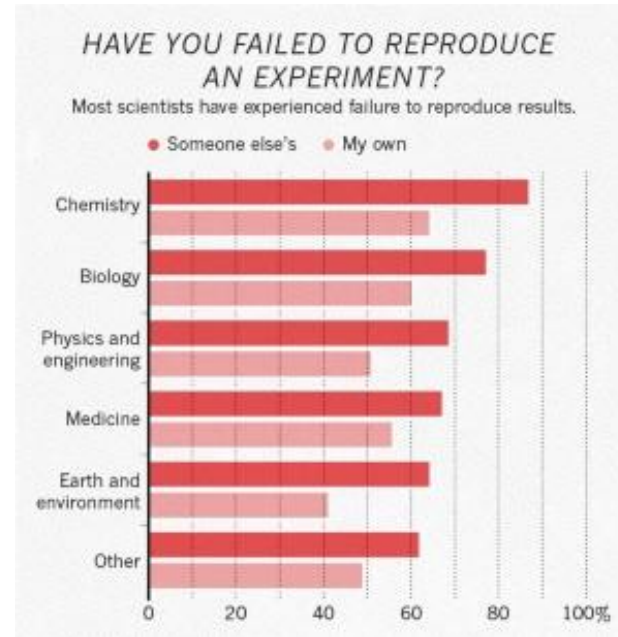
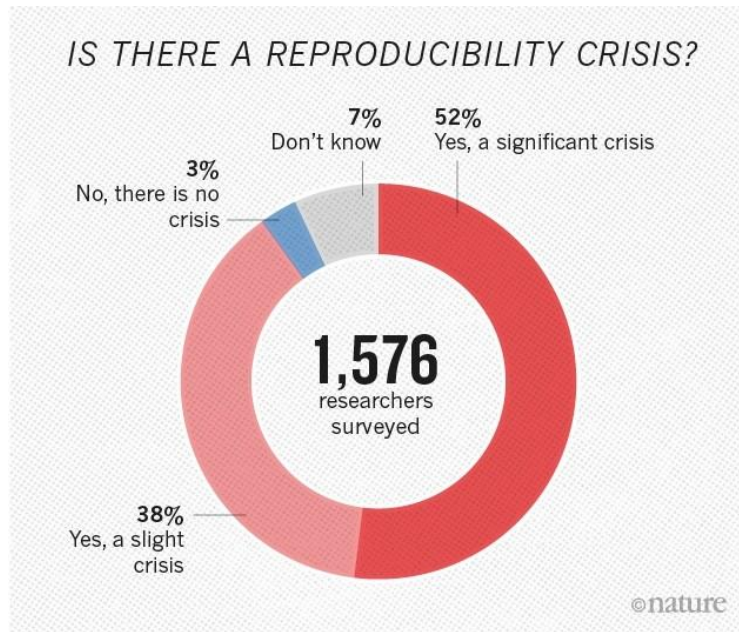
A study of the PubMed database found that the number of articles retracted from scientific journals increased substantially between 2000 and 2009.



New York Times 2011

Survey Nature - 1,500 scientists lift the lid on reproducibility (2016)

The crisis has affected all disciplines



nature

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [news feature](#) > [article](#)

News Feature | Published: 25 May 2016

1,500 scientists lift the lid on reproducibility

[Monya Baker](#)

[Nature](#) 533, 452–454 (2016) | [Cite this article](#)

209k Accesses | 2372 Citations | 5151 Altmetric | [Metrics](#)

Types of Reproducibility

Empirical Reproducibility

Detailed information on experiments

Note: The researcher controls the experimental setting

Their expertise may play a role

Record of data and data collection methods

Observational Reproducibility

Detailed information on observations

Note: The researcher does not control the setting – they observe

Record of data and data collection methods

Statistical Reproducibility

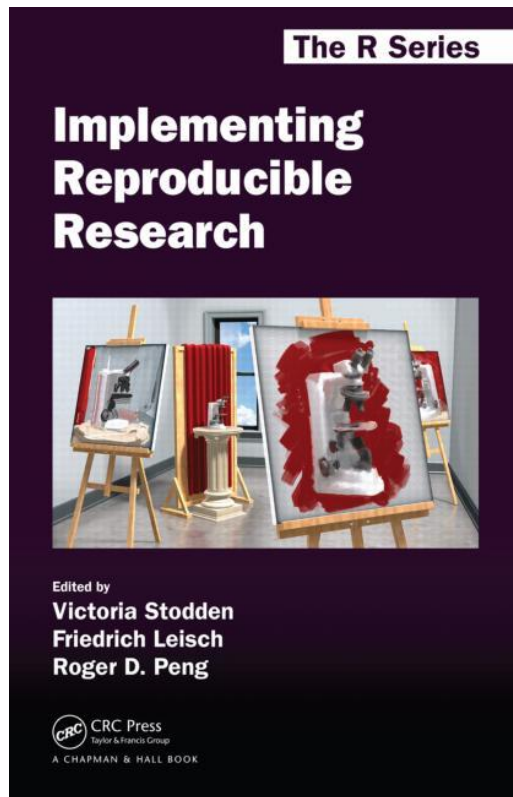
Detailed information on the choice of statistical tests, model parameters, decision thresholds, etc.

Pre-registration of the study design to prevent p-value manipulation and other manipulations

Computational Reproducibility

Detailed information on code, software, hardware, and implementation details.

Documenting *how* the data was produced



Levels of Reproducibility: Framework (proposal)

Level 1

Repeat

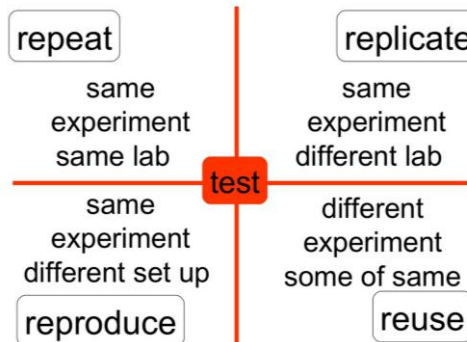
Complete traceability

Exact repetition

Same data

Redo – Reexecute

Goal: capture as much information as possible to explain a result



Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science. Science 2 Dec 2011; 1226-1227.

Level 2

Replicate

Some variations are allowed

Same results – similar data

Goal: test the limits of an approach

Levels of Reproducibility: Framework (proposal)

Level 1

Repeat

Complete traceability

Exact repetition

Same data

Redo – Reexecute

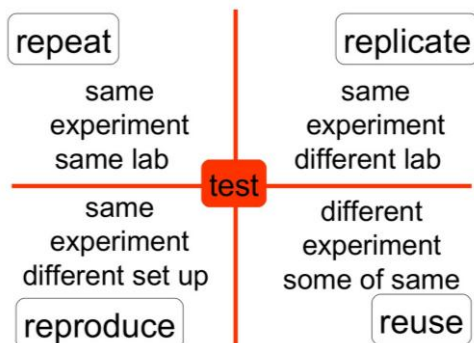
Goal: capture as much information as possible to explain a result

Level 3

Reproduce

Same result – same inference

But the means/procedures/methods/data may have changed



Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science. Science 2 Dec 2011; 1226-1227.

Level 2

Replicate

Some variations are allowed

Same results – similar data

Goal: test the limits of an approach

Reuse

Adapted to new needs

Partial reuse – in another context

A different result may be obtained

→ Cumulative Science

Context of Reproducibility

**Zoom into computational reproducibility in
bioinformatics**

Reproducibility networks

Scripts and computational reproducibility?

Providing scripts is an excellent first step

Using **git/github** for **versioning, collaborative** development

But

No clear distinction between **steps of the analysis**

- piece of codes, methods/functions

- ... and execution of the analysis

- data sets used as inputs and then produced

Major steps of the analysis may be difficult to get

No solution for **data management**

- Naming convention for produced files, storage...

→ Difficult to share, exchange and reuse (repurpose)

Scientific Workflow Management Systems

“Data analysis pipeline ”

Data flow driven

Encapsulation of scripts in **processors**

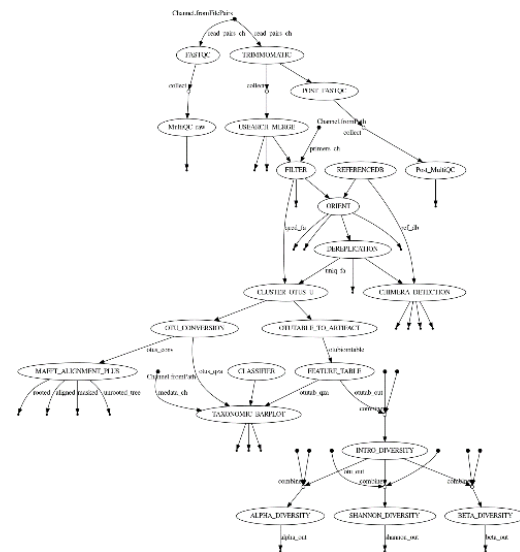
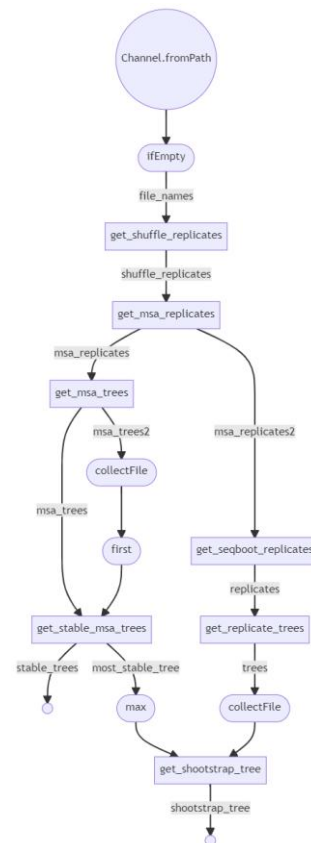
Algebraic transformation of data in **operations**

Modularity

WF specification: connected tools
steps of the analysis

Possible subworkflow - encapsulation

Systems:



Many changes in the last 15 years in repro solutions...

In terms of workflow systems

~2010s



Graphical
user
interface

~2020s



Many changes in the last 15 years...

~2010s

In terms of workflow systems

~2020s



Graphical
user
interface

Code-based



nextflow



snakemake



In terms of users

End users & Developers
Biologists & Computer scientists

Bioinformaticians trained to develop
More Python less Java
Reprohackathons



Bioinformatics, 2023, 36, 1-12
<https://doi.org/10.1093/bioinformatics/btad227>
ISMB/ECCB 2023



Reprohackathons: promoting reproducibility in
bioinformatics through training

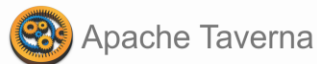
Thomas Cokelaer ^{1,2}, Sarah Cohen-Boulakia ^{1,2,3}, Frédéric Lemoine ^{1,2,4,5}

Many changes in the last 15 years...

~2010s

In terms of workflow systems

~2020s



Graphical
user
interface

End users & Developers
Biologists & Computer scientists

Code-based



In terms of users

Bioinformaticians trained to develop
More Python less Java
Reprohackathons



Bioinformatics, 2023, 36, 1-12
<https://doi.org/10.1093/bioinformatics/btad227>
ISMB/ECCB 2023



Reprohackathons: promoting reproducibility in
bioinformatics through training

Thomas Cokelaer ^{1,2}, Sarah Cohen-Boulakia ^{1,2,3}, Frédéric Lemoine ^{1,2,4,5}

In terms of means to share workflows



S. Cohen-Boulakia, U. Paris-Saclay



Challenges encountered when reusing

~2010s

Workflows were dependent of
web services

Workflow decay



~2020s

Workflows are dependent of
libraries

Challenges encountered when reusing

~2010s

Workflows were dependent of
web services



Workflow decay



Workflow rerun
(as a black box)



~2020s

Workflows are dependent of
libraries

Containers-based solutions help a lot!



docker



Challenges encountered when reusing

~2010s

Workflows were dependent of web services



Workflow decay



Workflow rerun
(as a black box)



Workflow replicate & reproduce



~2020s

Workflows are dependent of libraries

Containers-based solutions help a lot!



With a lot of efforts,
it is now possible!
(data availability
remains problematic)



Challenges encountered when reusing

~2010s

Workflows were dependent of web services



Workflow decay

Workflow rerun
(as a black box)

Workflow replicate & reproduce

Workflow reuse
(in part)

Very difficult in practice
Plumbing workflows

~2020s

Workflows are dependent of libraries

Containers-based solutions help a lot!



docker



With a lot of efforts,
it is now possible!
(data availability
remains problematic)



In code-based workflow systems, it is
as easy as reusing the code of
somebody else 😊

Workflow reuse status



2014

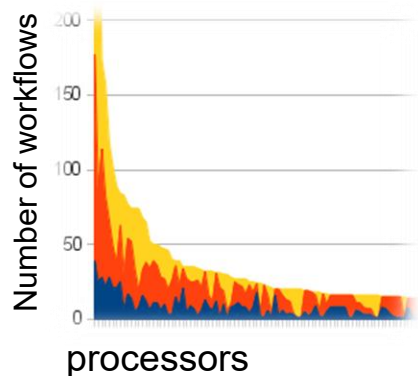
1,700 workflows

Taverna

10 242 processors (analysis steps)

Centralized in myExperiment

Re-use rates have a **Zipf-like distribution**



The top ten authors published 62% of all workflows

2023

2,443 workflows

Nextflow & Snakemake

15 540 processors (analysis steps)

Distributed in github

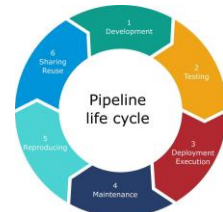
[Developing and reusing bioinformatics data analysis pipelines using scientific workflow systems](#)

M Djaffardjy, G Marchment, C Sebe, R Blanchet et al
Computational and Structural Biotechnology Journal
21, 2075-2085

Still low reuse

The top ten authors published 15% of all workflows

Higher reuse (as a black box) of processors from nf-core (repo of high quality)



Example of a project: ShareFAIR

ShareFAIR : Sharing reliable workflows to transform datasets into gold standards: Application to Neuro-Vascular Pathologies

PEPR Digital Health

Coordinator

- Sarah Cohen-Boulakia

Management structure

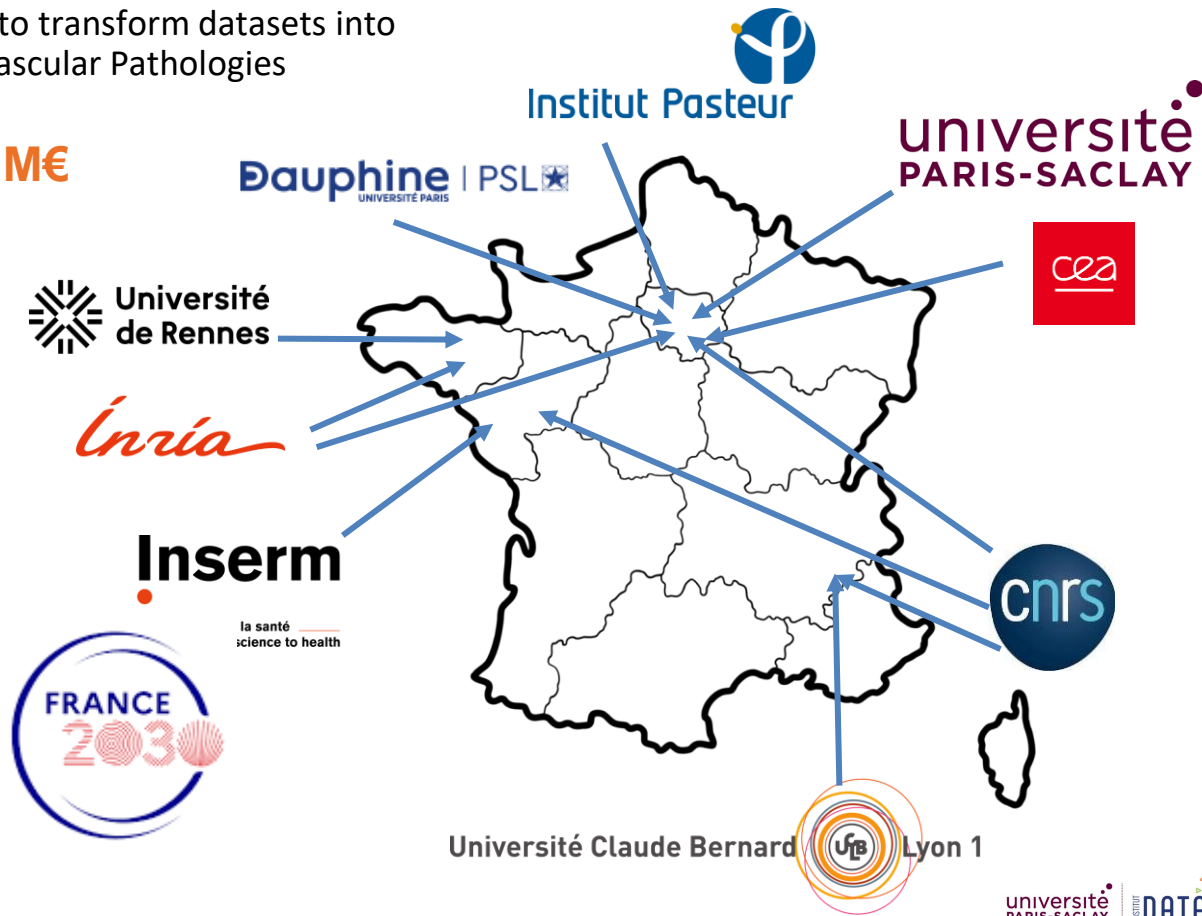
- Université Paris-Saclay

Partners

- Université Paris-Saclay
- Université Paris Dauphine PSL
- Institut Pasteur
- Université Lyon
- Université Rennes
- Inria
- CEA
- CNRS, INSERM (ITX)

Duration: 48 mois

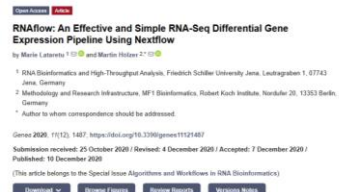
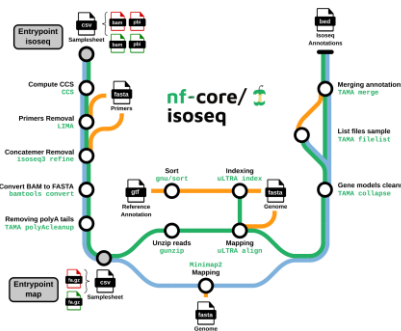
2 M€



Exploiting the 3 forms of workflow

```
#####
# RNAseq annotation for LIGO 2017 and 2018
#####
# Workflow: nf-core/seqr (v1.0.0)
#
# Parameters:
#   - input: path to the FASTQ files (required)
#   - output: path to the output directory (required)
#   - config: path to the configuration file (optional)
#   - debug: boolean flag to enable debug mode (optional)
#
# Description:
#   This workflow performs RNA-seq analysis using nf-core/seqr.
#   It includes steps for data ingestion, quality control, alignment,
#   and differential gene expression analysis.
#
# Usage:
#   nf-core/seqr --input /path/to/FASTQ --output /path/to/output
#
# License:
#   This workflow is licensed under the Creative Commons Attribution
#   license (CC BY 4.0).
#####
```

Originality of ShareFAIR 3 views of workflows



Code: GitHub

Graph: workflow structure

Text: scientific paper

- Automatic **generation of metro map** view based on the code of any Nextflow workflow (with Nextflow / nf-core groups) – G. Marchment
- **Alignment** of commands names extracted from code and tool names extracted for paper with **NLP technics** – C. Sebe

Stay tuned ☺

Outline

Context of Reproducibility

Zoom into computational reproducibility in
bioinformatics

Reproducibility networks

NETWORKS

Global Networks

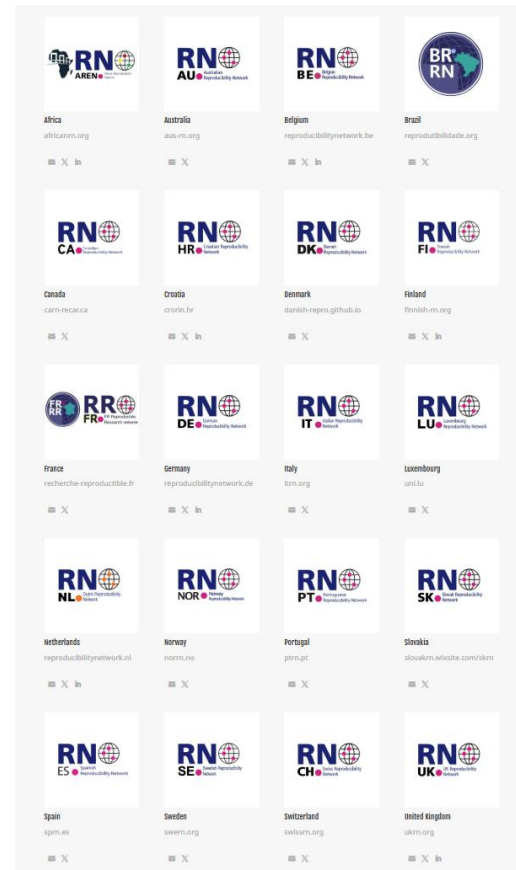
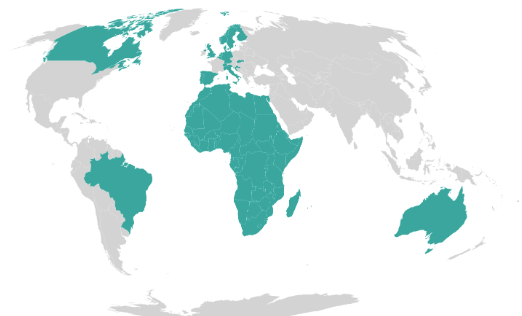
Outside the UK? Find a Reproducibility Network in your area

[See full Global Networks Statement](#)

Global Reproducibility Networks

A Reproducibility Network (RN) is a national, peer-led consortium of researchers that aims to promote and ensure rigorous research practices by establishing appropriate training activities, designing and evaluating research improvement efforts, disseminating best practice and working with stakeholders to coordinate efforts across the sector. RNs aim for broad disciplinary representation and an intensive interdisciplinary dialogue (e.g., with funding agencies, publishers, learned societies and other sectoral organisations, as well as researchers from all disciplines and across all career stages).

To reach as many researchers as possible, and to operate as efficiently as possible, we are keen to support other countries interested in creating similar networks. If you are interested in setting up a national RN, or finding out who in your country is working towards this, please email: contact@ukrn.org.



The Reproducible Research Network (France)

March 2022 : First days of the network

Brings together 260+ members (individuals)

Connected with existing networks, including French ones (LORIER, INSERM)



Ongoing structuring

European relations committee

Exchanges: Summer schools – meetings – joint workshops

Animation committee

Webinars – seminars – events...

Bibliographic monitoring committee and Institutional relations committee

Several working groups: Training WG – Notebook WG...

Newsletter – Monitoring – link with the network – event organization

30 disciplines: Computer Science 20% –
Bioinformatics 10% – Physics 8% –
Neurosciences 8%...

Comité de pilotage



Céline Anany-Robert – Laboratoire Jean Kuntzmann / GRICAD



Sarah Cohen-Boulakia – Laboratoire Interdisciplinaire des Sciences du Numérique



Raphaëlle Krummeich – Laboratoire IDEES UMR6266



Arnaud Legrand – LIG



Frédéric Lemaire – Institut Pasteur



Dominique Müller – Laboratoire Interuniversitaire de Psychologie



Sébastien Rey/Coynehourcq – Laboratoire IDEES UMR6266



François Ric – Laboratoire de Psychologie UR 4139



Nicolas P. Rougier – Institute of Neurodegenerative Diseases

5 Computer scientists
(multidisciplinary)

2 psychologists

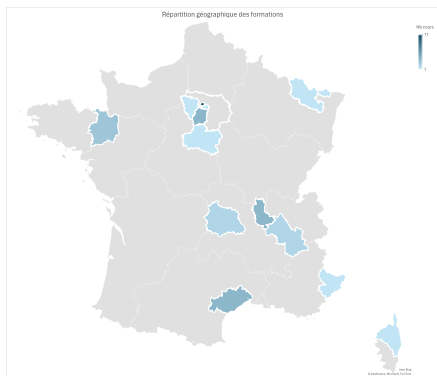
2 geographers

Example Action – Mapping Trainings (ongoing)

Questionnaire sent to members of the network

16 training modules + 2 MOOCs

Taught in pluridisciplinarity



Objectives

Map existing training programs

Content sharing

Create a network for training the trainers

Dedicated website with available resources




**MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE**
*Liberté
Égalité
Fraternité*



Network Life – Ongoing and Upcoming Actions

<https://www.recherche-reproductible.fr>

Conférences

- | | |
|--------------|---|
| Oct 3, 2025 | Replication Games - Paris |
| Apr 3, 2025 | Journées du Réseau Français de Recherche Reproductible |
| Dec 17, 2024 | Journées Ardoise : ouverture des codes sources et logiciels de la recherche |



Programme des séminaires à venir pour l'année 2024 - 2025

- **6 Mai 2025, 14h** : Présentation du réseau anglais (UKRN), organisation, fonctionnement et initiatives, *Etienne Roesch*
- **23 Mai 2025, 14h** : Reproductibilité et environnements de développement, *Pol Dellaiera*
- **13 Juin 2025, 16h** : {rix}, un paquet R qui s'appuie sur Nix, *Bruno Rodriguez*.

Formations

ANF "Reproductibilité computationnelle des résultats de publications scientifiques, pratiques et outils", 1-3 Juillet 2025.

Le mouvement pour la science ouverte a depuis quelques années pris une ampleur inédite. Les questions autour de la reproductibilité des résultats de publications scientifiques s'inscrivent naturellement dans cette dynamique. Elles sont d'une importance capitale pour assurer la transparence de la science, la confiance de la société, et sont en lien fort avec les problématiques d'éthique de la science. Dans ce cadre, le service formation du CNRS et CNRS Mathématique (INSMI) proposent propose pour la deuxième année consécutive une formation autour du logiciel libre. Cette formation s'inscrit dans cette évolution, afin de donner aux participantes et aux participants tous les éléments pour aller dans le sens de plus de reproductibilité. Plus d'informations et inscriptions sur le [site de la conférence](#). (Une priorité sera donnée aux membres des laboratoires de mathématiques.)



Comment nous contacter ?

Pour entrer en contact avec des membres du réseau : envoyez nous un courriel à contact@recherche-reproductible.fr

Pour intégrer le réseau : vous pouvez vous abonner à la liste de diffusion sur la page : <https://groupes.renater.fr/sympa/info/recherche-reproductible>

Conclusion

No cumulative science without reproducibility

Compared to 15 years ago: **many technical solutions exist** to the reproducibility crisis

Challenges now lie in **shifting from redo to reuse**

- reusing code

- understanding code

- documenting (at various levels of granularity) code

Such problems need to be treated in a **multidisciplinary** way within computer science: data science – graphs – NLP – visualization – software engineering...

nature

[View all journals](#) [Search](#) [Log in](#)

[Explore content](#) [About the journal](#) [Publish with us](#)

[nature](#) > [comment](#) > [article](#)

COMMENT | 07 October 2025

Stop treating code like an afterthought: record, share and value it

Scientists, research institutions, funders, libraries and publishers must all improve software practices.

By [Roberto Di Cosmo](#), [Sabrina Granger](#), [Konrad Hinsén](#), [Nicolas Jullien](#), [Daniel Le Berre](#), [Violaine Louvet](#), [Camille Maumet](#), [Clémentine Maurice](#), [Raphaël Monat](#) & [Nicolas P. Rougier](#) 



Thanks!



References

Extracting Information in a Low-Resource Setting: Case Study on Bioinformatics Workflows

C Sebe, S Cohen-Boulakia, O Ferret, A Névéol

IDA - International Symposium on Intelligent Data Analysis, 274-287, 2025

Reproducibility in Named Entity Recognition: A Case Study Analysis

CC Villarmin, S Cohen-Boulakia, N Naderi

2024 IEEE 20th International Conference on e-Science (e-Science), 1-10

BioFlow-Insight: facilitating reuse of Nextflow workflows with structure reconstruction and visualization

G Marchment, B Brancotte, M Schmit, F Lemoine, S Cohen-Boulakia

NAR Genomics and Bioinformatics 6 (3), Iqae092

Representing bioinformatics Nextflow workflows in RO-Crate: challenges and opportunities

G Marchment, M Schmit, C Sebe, F Lemoine, H Ménager, ...

Semantic Web Applications and Tools for Health Care and Life Sciences

Towards improving workflows reproducibility: Extracting information on workflows from text and code repositories

C Sebe, F Lemoine, A Gaignard, O Ferret, S Cohen-Boulakia, A Névéol

31st Annual Intelligent Systems For Molecular Biology (ISMB Workshop)

Developing and reusing bioinformatics data analysis pipelines using scientific workflow systems

M Djaffardjy, G Marchment, C Sebe, R Blanchet, K Belhajjame, ...

Computational and Structural Biotechnology Journal 21, 2075-2085

Reprohackathons: promoting reproducibility in bioinformatics through training

T Cokelaer, S Cohen-Boulakia, F Lemoine

Bioinformatics 39 (Supplement_1), i11-i20