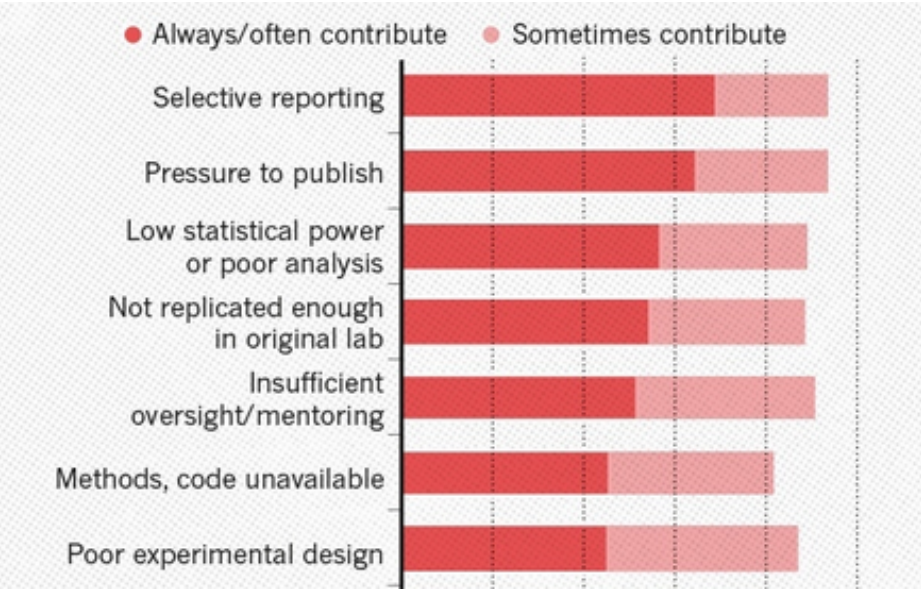
Redefinition of a workflow

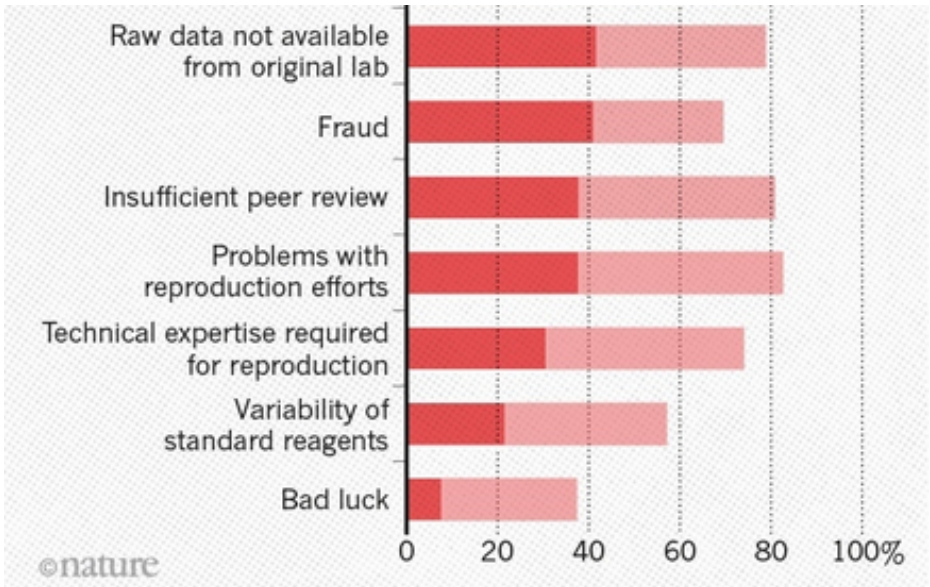
ANF Workflow et reproductibilité en Bioinformatique

Claire Toffano-Nioche, Laurent Jourdren

november 2025

Reproducibility crisis





survey on 1,500 scientists, Baker et al., Nature, 2016

What are the difficulties (scientific uses)?

Publication bias and/or pressure

- we publish “only” what is new
- we publish “only” positive results
- the “publish or perish” culture encourages quantity over quality

Scientific practices

- HARKing (Hypothesizing After the Results are Known)
  reformulating hypotheses after obtaining results
- p-hacking: manipulating data to achieve the desired statistical threshold

Problems accessing data

- *data available upon request*
- data available, but non-existent or insufficient metadata (not FAIR)

What are the difficulties (numerical aspects)?

Problems accessing tools

- installation difficulties
- old or obsolete tools
- configuring the analysis:
  tools (version + settings + sequence)

Problems accessing the necessary resources

- calculation
- storage

How to improves?

For this part of the training devoted to analysis:

Working assumptions:

- we want to produce our analyses ethically (far from the problems of p-hacking and HARKing)
- without publication pressure (because is not always within our control)
- access to data is operational (e.g., data specific to the institute)

issues related to data will be addressed later in the training

$\Longrightarrow$ the remaining numerical difficulties define our needs/solutions

How to improves?

Problems accessing tools

- installation difficulties ⟹ use a package manager
- old or obsolete tools ⟹ define a container
- configuring the analysis:

  tools (version + settings + sequence) ⟹ workflow

Problems accessing the necessary resources

- calculation ⟹ HPC generic (with a workflow configuration file)
- storage ⟹ fix the deposit access (with versions)

How to improves?

Problems accessing tools

- installation difficulties ⟹ use a package manager
- old or obsolete tools ⟹ define a container
- configuring the analysis:

  tools (version + settings + sequence) ⟹ **workflow**

Problems accessing the necessary resources

- calculation ⟹ HPC generic (with a **workflow** configuration file)
- storage ⟹ fix the deposit access (with versions)

Definitions

Workflow: sequence of the (bioinformatics analysis) steps

e.g. successive interactive launches, script (bash, python)

- rely on both our own code or some third-party tools
- each step can have input(s), output(s) and parameter(s)
- derived today toward the workflow file (e.g. snakefile, nextflow, galaxy)

Workflow management system: software device designed to help scientists develop, execute, and monitor chains of data analysis programs (wikipedia)

Advantages of Workflow management systems

▸ With a workflow management system, developers focus solely on workflow logic: defining steps and their sequence.

The workflow management system handle:

- the support ressources (local or HPC)
- the access to the third-party tools (software environment or container)
- free parallelization

Parallelisation

Workflow management systems cannot parallelize the tools to be executed, they can only schedule data processing on available resources (cpus and cluster nodes)

This can be achived because a workflow can be described as a **Directed Acyclic Graph** (DAG)

→ As soon as all the inputs of a step are available, a step can be started

Abstraction over infractruture

Workflow management systems provide an **abstraction layer** across multiple **infrastructures**

A workflow developed on a laptop can be **run on any other infrastructure** (workstation, computing cluster, cloud…), once the workflow system is installed and configured

History of workflow management system

	mode	composition	date	comment
Static WF	local	hard code	1975-1985	hard to maintain
Kepler	web-service	graphic	2002-	BSD licence
Taverna	web-service	graphic	2003-2020	ApacheSF
KNIME	local, HPC	graphic	2004-	GPLv3
Galaxy	local, HPC	graphic	2005-	MIT (from 2021-04-07; Academic Free License v3 before)
Snakemake	local, HPC	textual	2012-	MIT
Nextflow	local, HPC, Cloud	textual	2013-	Apache License 2.0

Share

WorkflowHub	
Galaxy	713
NextFlow	167
CWL	111
COMPSs	67
Snakemake	60
Python	59
Jyputer	35
Docker	21
Scipion	17
Shell script	16
KNIME	14

Taverna:

myexperiment

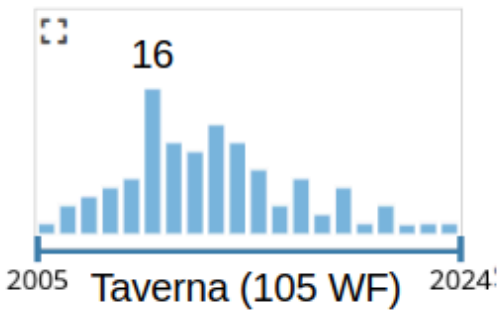
Nextflow:

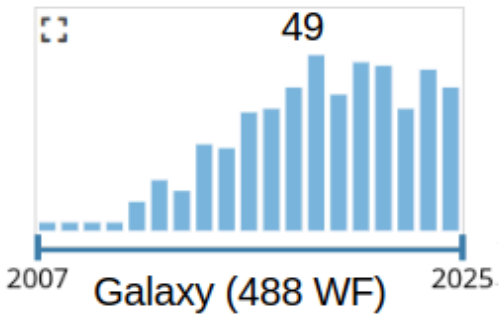
nf-core

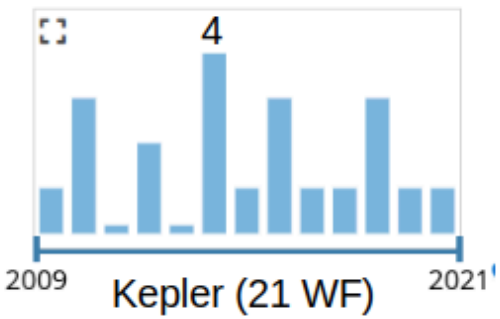
Snakemake: workflows wrappers

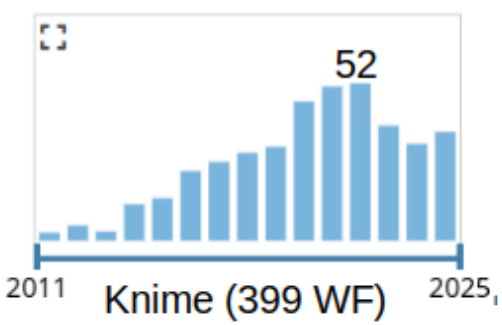
nov. 2025, more than 10 elements

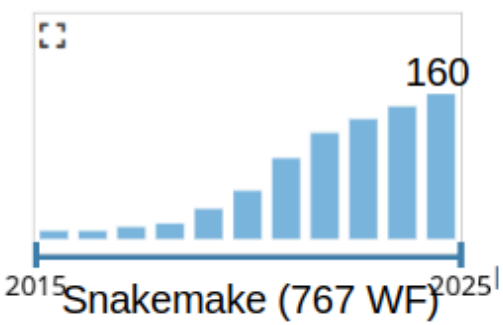
Conclusion

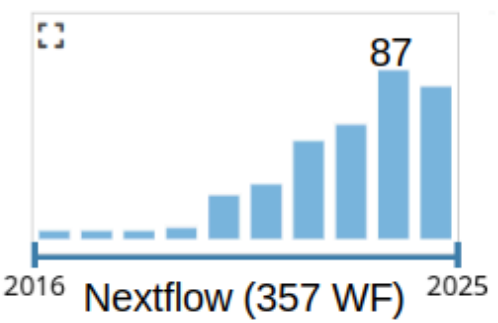












Number of citations in PMC europe (nov.2025):

Workflow management systems and their workflow files are a best practice for improving reproducibility.

Bibliography

- Scientific Workflows for Computational Reproducibility in the Life Sciences: Status, Challenges and Opportunities S. Cohen-Boulakia et.al 2017
- Workflows and e-Science: An overview of workflow system features and capabilities E. Deelman et.al 2009
- Kepler: an extensible system for design and execution of scientific workflows I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, S. Mock, Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004
- myExperiment: social networking for workflow-using e-scientists. C.Goble, D.Charles De Roure:WORKS_HPDC 2007: 1-2
- Taverna: a tool for the composition and enactment of bioinformatics workflows. T.Oinn, M.Addis, J.Ferris, D.Marvin, M.Senger, R.Greenwood, T.Carver, K.Glover, M.Pocock, A.Wipat, P.Li Bioinform. 20(17): 3045-3054 (2004)