LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

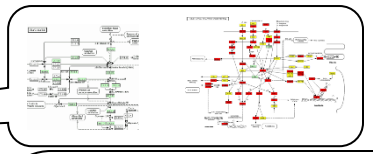
Reproducibility of Bioinformatics pipelines Some lessons learnt

Sarah Cohen-Boulakia

Université Paris-Saclay, Laboratoire Interdisciplinaire des Sciences du Numérique, UMR 9015, Université Paris-Saclay Orsay, France

cnrs
dépasser les frontières

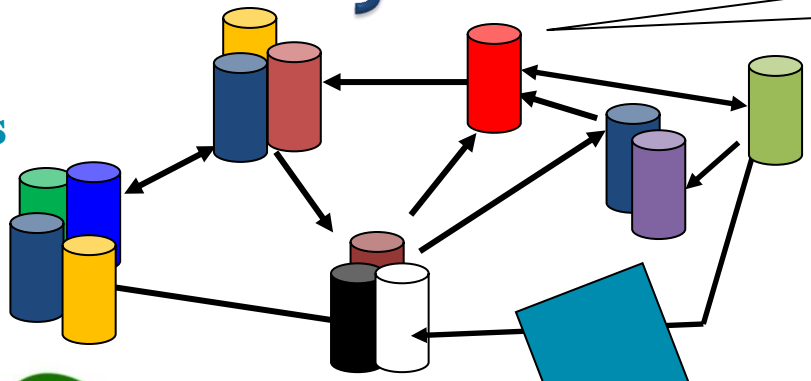
Biological analysis

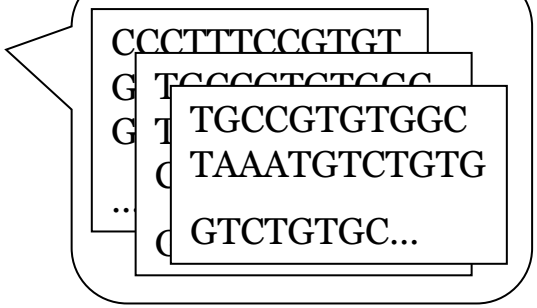


Public sources

- Distributed
- Heterogeneous
- Network

> 1,500 (NAR)





Tools Python

Scripts JAVA

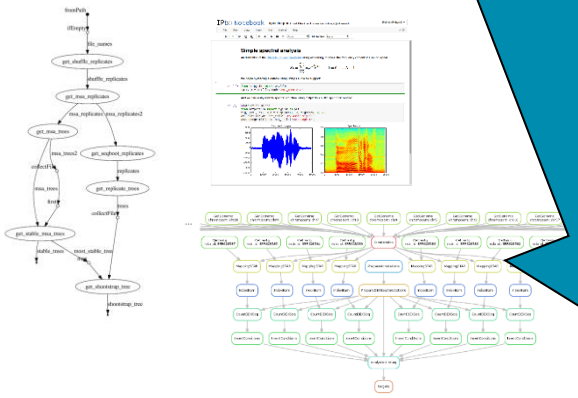
Web services

Tools - Scripts

- Distributed
- Heterogeneous

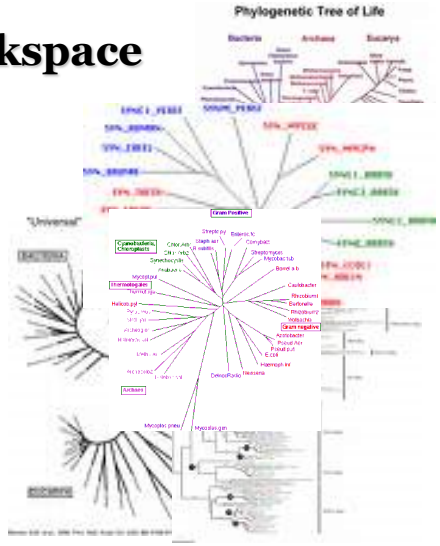
> 23,000 (bio.tools)

How the data (result) have been obtained? Which exact input data? Which tools? Which parameter settings?

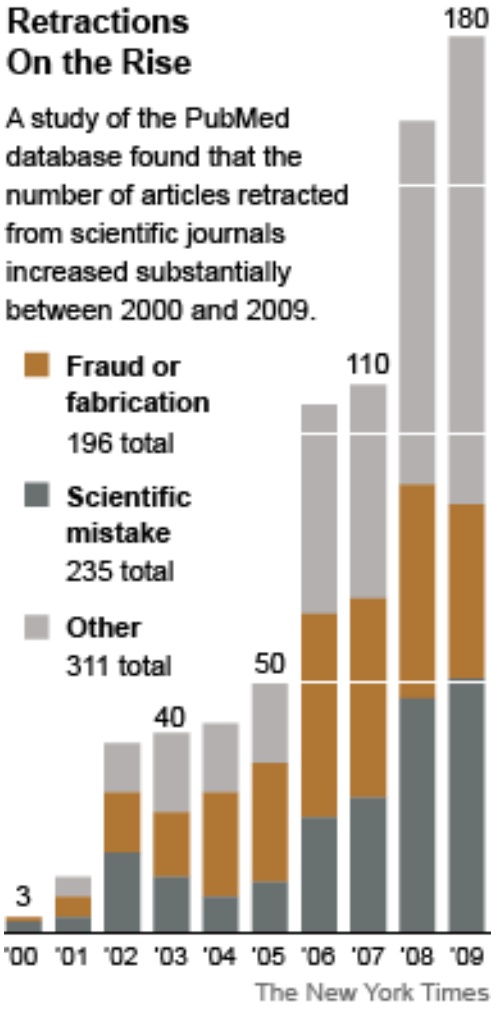


Analysis pipelines

Workspace



Reproducibility Crisis…



→ *Nature* checklist

→ *Science* requirements for data and code availability

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 mistakes 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 and funding agencies need to put a premium on 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

http://www.nature.com/nature/focus/reproducibility/index.html

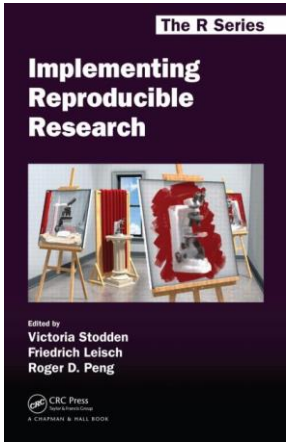
Reproducibility

- *Empirical reproducibility*
  - detailed information about non-computational empirical scientific experiments and observations
  - In practice this is enabled by making data freely available, as well as details of how the data was collected.

- *Statistical reproducibility*
  - detailed information about the choice of statistical tests, model parameters, threshold values, etc.
  - This relates to pre-registration of study design to prevent p-value hacking and other manipulations.

- *Computational reproducibility*
  - detailed information about data sets, code, software, hardware and implementation details

    → Goal: document how data has been produced



V. Stodden *et al.*

Outline

The Reproducibility Crisis

Levels of Reproducibility

Elements of solutions for reproducibility & reuse

Current Actions

Reproducibility in Bioinfo pipelines

3 ingredients to track

- ***Tools used / code*** — pipeline specification
  - Version of black/grey boxes
  - Version of (possibly open source) code
- ***Computational environment*** — pipeline environment
  - Many dependencies - libraries used ++
  - Different behavior on different OS
- ***Data and parameters used at run time*** — pipeline execution
  - Biological data change a lot
  - Tools have an increasing number of parameters

Levels of computational reproducibility

3 ingredients

Pipeline (Specification)

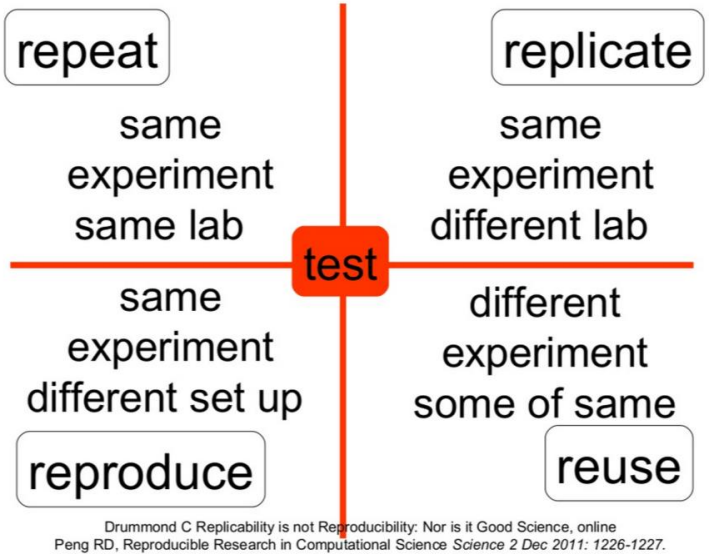
Chained Tools

Pipeline Execution

Input data and parameters

Environment

OS/librairies …



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.*

Repeat

- *Redo*: exact same context
- Same pipeline, execution setting, environment
- Identical *output*

→Aim = proof for reviewers ☺

Replicate

- Variation allowed in the pipeline, execution setting, environment
- Similar *output*

→ Aim = robustness

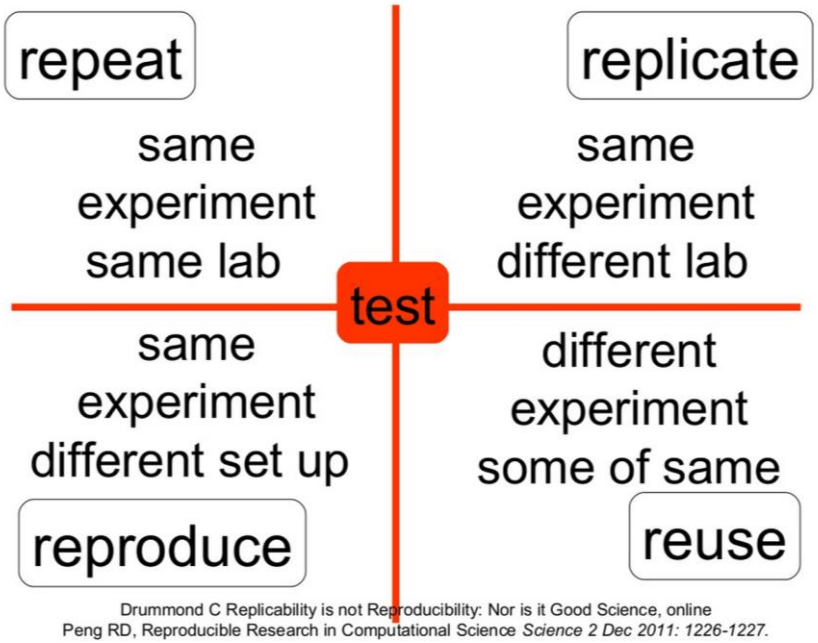
A continuum of possibilities

Reproduce

- Same *scientific result*
- But the means used may be changed
- Different pipelines, execution setting, environment
- Different output but in accordance with the result

Reuse

- Different scientific result
- Use of tools/… designed in another context



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.*

Reproducibility as a necessary condition for Reuse

Outline

The Reproducibility Crisis

Levels of Reproducibility

Elements of solutions for reproducibility & reuse

Current Actions

Open Science and reproducibility

A large part of the solution lies in

- Providing (Opening) source code
- Using versioning code, collaborative development
- Using solutions for tracking the dev environment (popular in bioinformatics: Singularity, Docker – containers) – **GUIX?** ☺

Some problems still occur…

- 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 (1/2)

SWFS = “Data analysis pipeline ”

Data flow driven

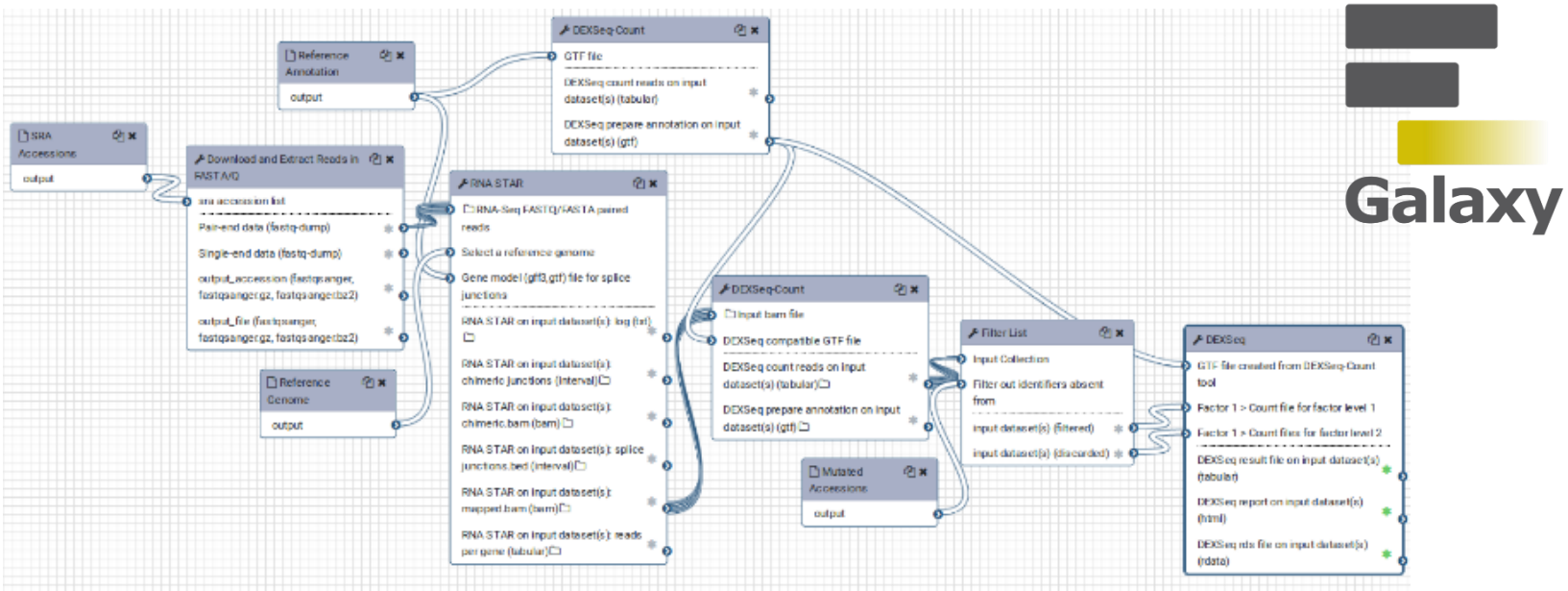
Encapsulation of scripts

Modularity

Snakemake,
Nextflow,
Galaxy…

nextflow

WF specification: connected tools steps of the analysis



Galaxy

Scientific Workflow Management Systems (2/2)

WF execution: data consumed/produced

Provenance modules data management SWFS scheduling, logging, …

Transparency, optimisation, traceability

Bio.tools

https://bio.tools/

- Registry of Tools for the Life Sciences
  - find, understand, compare and select resources == **discovery**
  - use and connect them in workflows == **(inter)operability**
- Leaded by ELIXIR (European network of Excellence)

- Each tool must be described using biotoolsSchema
  - a formalized XML schema (XSD) which defines a description model for bioinformatics software (inputs, outputs and operations)
  - EDAM Ontology Terms are used

- EDAM Ontology
  - bioinformatics types of data including identifiers, data formats, operations and topics

Outline

The Reproducibility Crisis

Levels of Reproducibility

Elements of solutions for reproducibility & reuse

Current Actions

ReproHackathon: reproducing with workflows

Hackathon

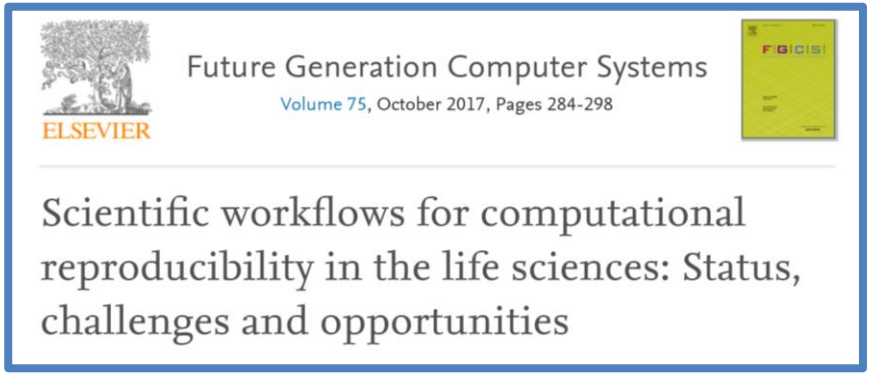
- Several developers in the same room
- Same goal to achieve (e.g., predicting growth from plants images)
- Create useable software in a short amount of time
- Aim: Demonstrating feasibility

ReproHackathon

Testing workflow systems in practice

- A hackathon where
  - Given a scientific publication + input data (+ possibly contacts with authors)
  - Several (groups of) developers reimplement the methods to try to get the same result
- Aim: Ability of current tools to reproduce a scientific result

Reproducibility-friendly features in workflows



5 Systems: Galaxy, VisTrails, Taverna, OpenAlea, NextFlow

Workflow specification

Language (XML, Python…) → repeat … reuse

Interoperability (CWL…) → replicate … reuse

Description of steps

- Remote services → repeat
- Command line → repeat … reuse
- Access to source code → replicate

Modularity (nested workflows?) → reuse

Annotation (tags, ontologies…) → reuse

Execution

Language & standard (PROV) → repeat … reuse

Presentation (interactivity with the results/ provenance, notebooks) → replicate … reuse

Annotations → reuse

Environment

Ability to run workflows in a given environment

→ repeat … reuse

VM: VMWare, KVM, VirtualBox, Vagran,…

Lighter solutions (containers): Docker, Rocket, OpenVZ, LXC, Conda

Command-line history: CDE, ReproZip

Conclusion

Bioinformatics pipelines need reproducibility & reuse

Several elements of solution exist

GUIX deserves to be more used by the bioinfo community!

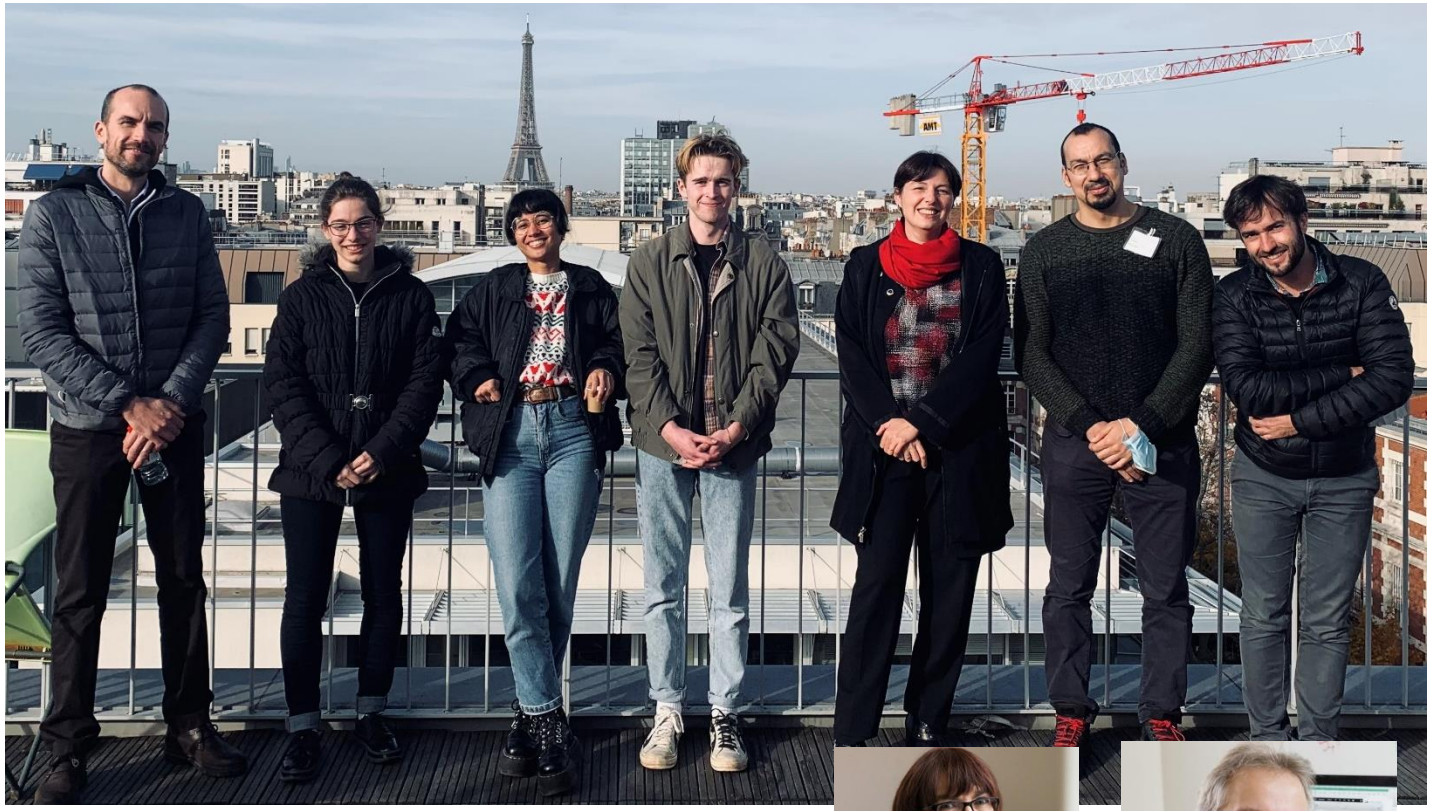
Eager to know more!

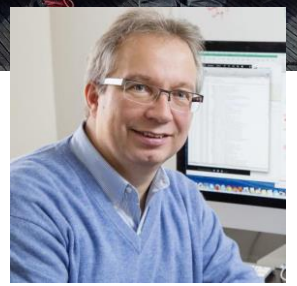
Today & over the weekend











THANKS!




cnrs
GDR
Groupement
de recherche
MaDICS Masses de données, informations
et connaissances en sciences

université
PARIS-SACLAY

Studies on reproducibility

- Nekrutenko & Taylor, Nature Genetics (2012)
  - 50 papers published in 2011 using the Burrows-Wheeler Aligner for Mapping Illumina reads.
  - 31/50 (62%) provide no information
    - no version of the tool + no parameters used + no exact 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 (publishers)
    - 149 (30%) were not subject to any data availability policy
      (0% made their 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