Reproductibilité computationnelle en sciences de la vie et *workflows* scientifiques : état-des lieux et retour d'expérience

Sarah Cohen-Boulakia

LISN

Laboratoire Interdisciplinaire des sciences du Numérique

Université Paris-Saclay

université PARIS-SACLAY

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

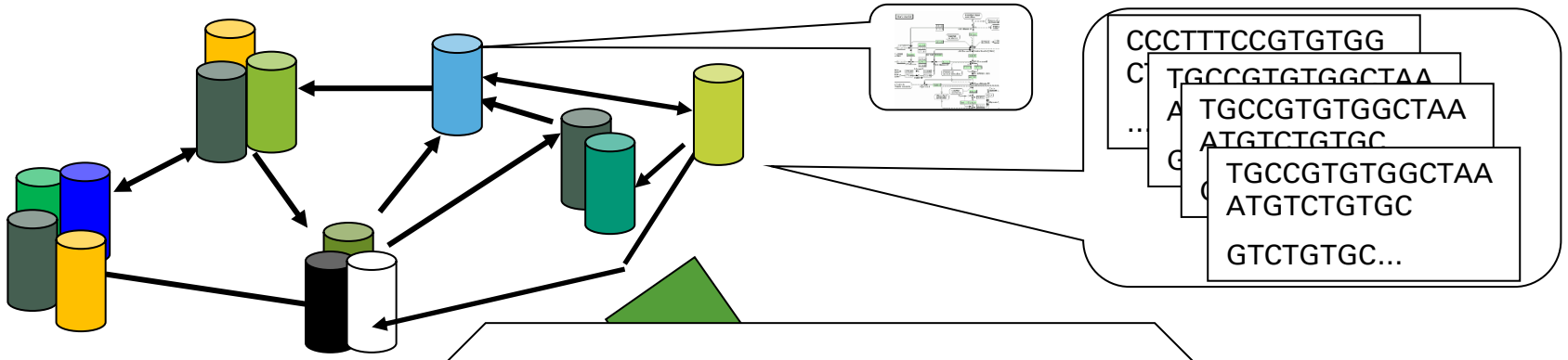
SIF Société Informatique de France

Bioinformatics analysis

Public and private data sources

- Distributed
- Heterogeneous
- **> 1,500**

Nucleic Acids Research



Binarization Water Use Efficiency
Segmentation
Python
Java
Web services

Pipelines
Combi tools
WorkflowHub
nf-core

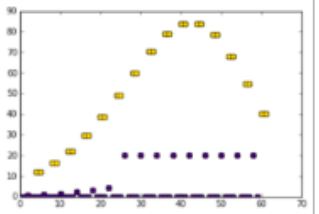
How has this plot been generated?
With which input data?
With which tools? Parameters?

What is the difference between these experiments?

Tools

- Distributed
- Heterogeneous
- **> 21 000 tools**

elixir bio.tools

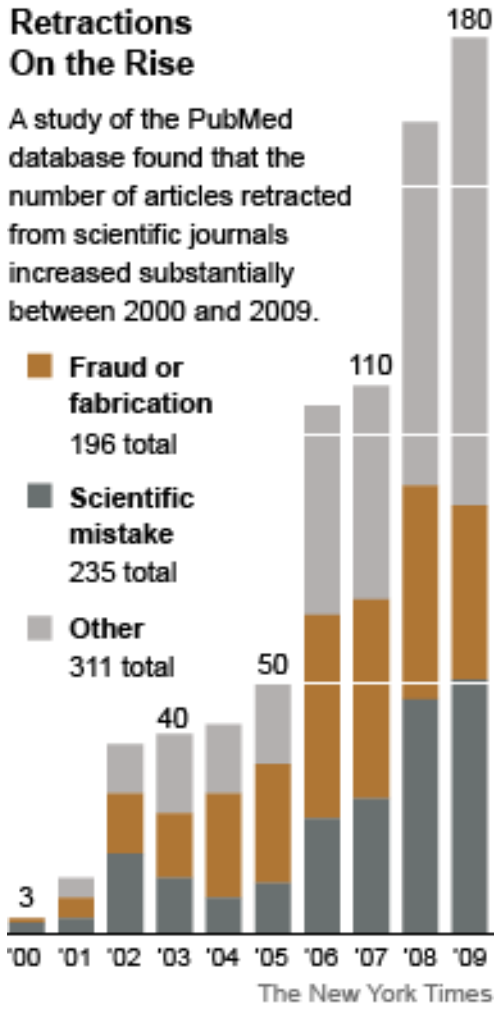


Studies on (lack of) reproducibility

- 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

Reproductibility Crisis…



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

→ *Nature* checklist

→ *Science* requirements for data and code availability

Société Informatique de France

Kinds of 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 code, software, hardware and implementation details

    → Goal: document how data has been produced

V. Stodden
et al.

Société Informatique de France

Scripts and reproducibility?

- Providing your 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

SWFS = “Data analysis pipeline ”

Data flow driven

Encapsulation of scripts, Modularity

WF specification: connected tools steps of the analysis

WF execution: data consumed/produced

Provenance modules data management SWFS scheduling, logging, …

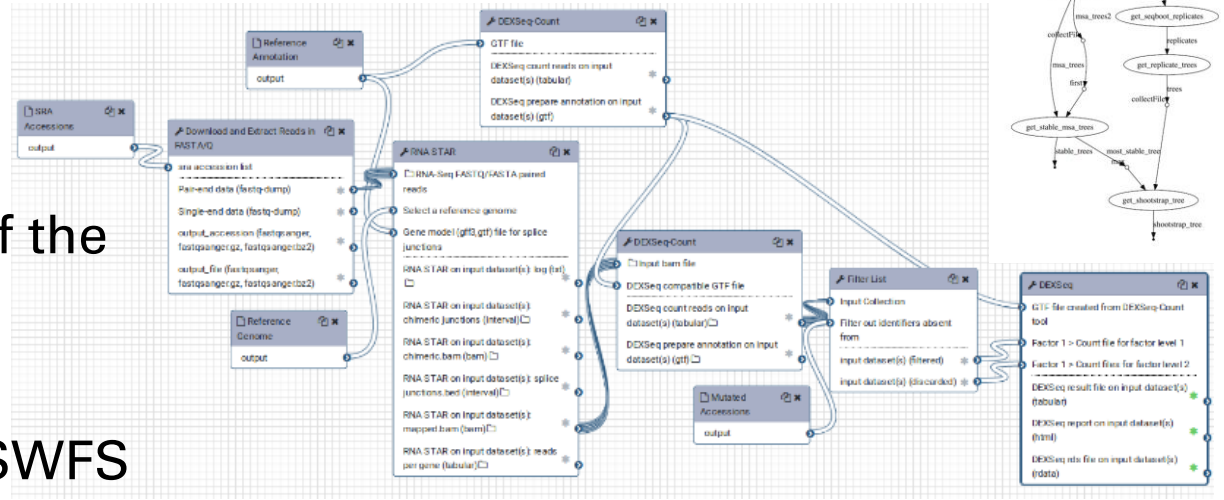
Transparency, optimisation, traceability

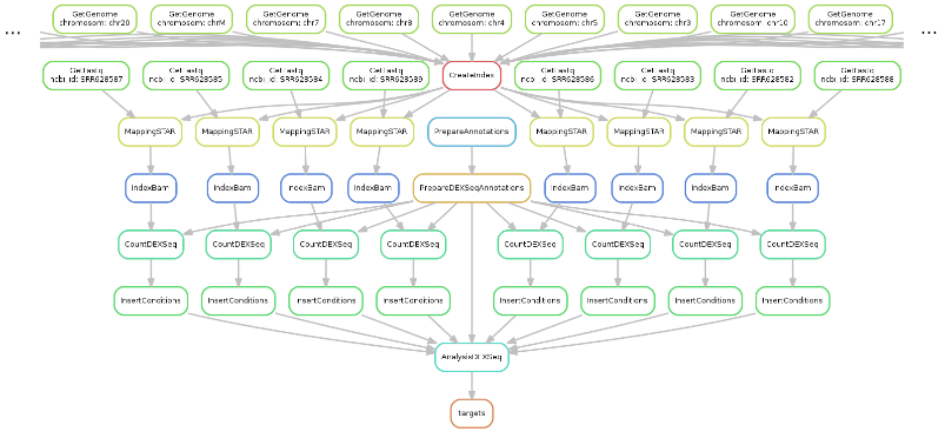
WF environment: companion tools

Virtual machines, containers

Docker, singularity,

Systems: Galaxy, NextFlow, SnakeMake





Société Informatique de France

Members of our Action @MaDICS

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

CNRS UMR & UMS

IRISA Univ. Rennes

Univ. CHU Nantes

Centre de Biophysique Moléculaire, CNRS Orléans

IRD, CIRAD, INRA, Inria, Univ. Montpellier

Univ. Lyon 1 LIRIS

LRI Univ. Paris Sud

CDS, Center for Data Science Saclay

Institut Francais Bioinformatique Gif s/Yvette

Institut Pasteur, Paris

Lamsade Univ. Paris Dauphine

LIG (Grenoble)

GDR MaDICS, GDR Bioinfo, Gpes de travail
IFB, Centres *Data Sciencesinternationaux*

Société Informatique de France

Aims of our Action@MaDICS

Concepts, Needs/solutions

- Which *levels* of reproducibility can we consider?
- Which are the solutions currently available ?

Opportunities, challenges

- What is missing?
- Which are the *research* (vs technical) *open issues*?

Evaluation of solutions on practice and state-of-the-art

- Experience of developers in using solutions in real contexts
- ReproHackathon

  → Real use cases from the Bioinformatics Domain

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

Results of our Action

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

(1) Paper @ FGCS

- Levels of reproducibility
- Criteria of choice
- Open Challenges

Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities

Sarah Cohen-Boulakia[a, b, c], Khalid Belhajjame[d], Olivier Collin[e], Jérôme Chopard[f], Christine Froidevaux[a], Alban Gaignard[g], Konrad Hinsen[h], Pierre Larmande[i, c], Yvan Le Bras[j], Frédéric Lemoine[k], Fabien Mareuil[l, m], Hervé Ménager[l, m], Christophe Pradal[n, b], Christophe Blanchet[o]

(2) ReproHackathon

- New concept designed

(2) 3 hour Webinar : Tutorial + 2 demos (A. Legrand)

Levels of computational reproducibility

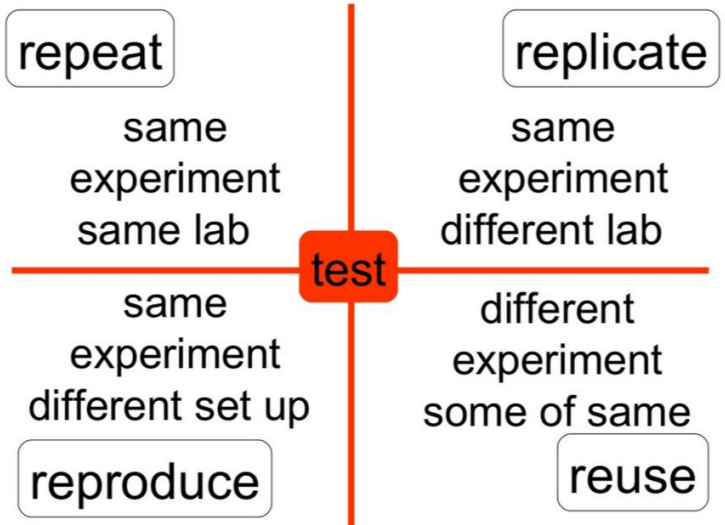
Future Generation Computer Systems
Volume 75, October 2017, Pages 284–298

Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities

Sarah Cohen-Boulakia[a, b, c], Khalid Belhajjame[d], Olivier Collin[e], Jérôme Chopard[f], Christine Froidevaux[a], Alban Gaignard[g], Konrad Hinsen[h], Pierre Larmande[i, c], Yvan Le Bras[j], Frédéric Lemoine[k], Fabien Mareuil[l, m], Hervé Ménager[i, m], Christophe Pradal[n, b], Christophe Blanchet[o]

3 ingredients

- Workflows Specification
  - Chained Tools
- Workflow Execution
  - Input data and parameters
- Environnement
  - 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 workflow, execution setting, environement
  - Identical *output*

  →Aim = proof for reviewers ☺

- Replicate
  - Variation allowed in the workflows, execution setting, environement
  - Similar *output*

  → Aim = robustness

Société Informatique de France

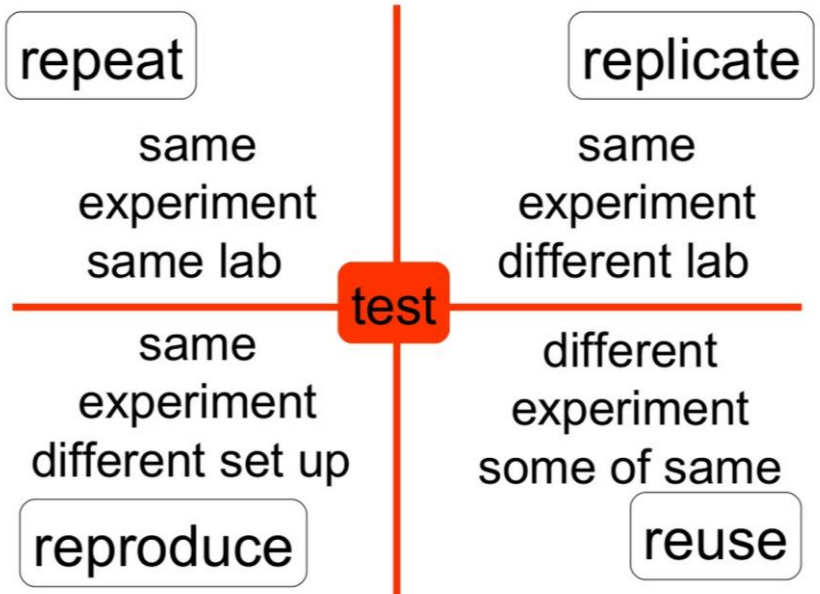
A continuum of possibilities

Future Generation Computer Systems
Volume 75, October 2017, Pages 284–298

Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities

Sarah Cohen-Boulakia[a, b, c], Khalid Belhajjame[d], Olivier Collin[e], Jérôme Chopard[f], Christine Froidevaux[g], Alban Gaignard[g], Konrad Hinsen[h], Pierre Larmande[i, c], Yvan Le Bras[j], Frédéric Lemoine[k], Fabien Mareuil[e, m], Hervé Ménager[l, m], Christophe Pradal[b, n], Christophe Blanchet[o]

- Reproduce
  - Same *scientific result*
  - But the means used may be changed
  - Different workflows, 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-friendly features in scientific workflows

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

researchobject.org ED AM

Execution

Language and standard (PROV…,) → repeat … reuse

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

Annotations → reuse

researchobject.org

Environment (companion tools)

Ability to run workflows in a given environment → repeat (… reuse)

Virtual machines

- Package, *freeze,* and expose the environment
- VMWare, KVM, VirtualBox, Vagran,…

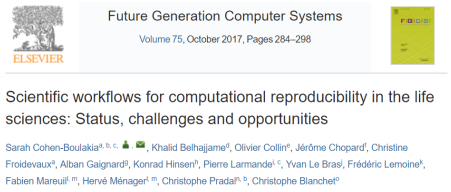
Lighter solutions (containers)

- Software dependencies
- Docker, Singularity, Rocket, OpenVZ, LXC, …

Capturing command-line history input/output, specification

CDE, ReproZip (NewYork University)

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


Future Generation Computer Systems
Volume 75, October 2017, Pages 284–298

Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities

Sarah Cohen-Boulakia, Khalid Belhajjame, Olivier Collin, Jérôme Chopard, Christine Froidevaux, Alban Gaignard, Konrad Hinsen, Pierre Larmande, Yvan Le Bras, Frédéric Lemoine, Fabien Mareuil, Hervé Ménager, Christophe Pradal, Christophe Blanchet

Our new concept: Reprohackathon

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

Société Informatique de France

ReproHackathons

Gif, 06/2017
RNA-Seq data from patients with uveal melanoma
Lyon, 07/2018
Phylogeny data
Montpellier, 11/2019
Plants Image analysis

https://ifb-elixirfr.github.io/ReproHackathon/hackathon_1.html

Systems : SnakeMake, NextFlow, iPython notebooks, Galaxy, scripts…
Executed in the Cloud@IFB

Testing several levels of reproducibility: repeat and replicate

Société Informatique de France

Conclusion

- Too many scientific results are not reproducible
- Reproducibility is necessary to ensure reuse
  - Cumulative science
- Several scientific workflow systems and companion tools are mature solutions
  - Repeat is (almost) always reachable
  - Next levels may be more difficult to reach
- Scientific workflow systems can offer support
  - Automatical Data annotation: FAIR, Data Management Plan…
- Need to teach how to use Scientific workflow systems
  - Now a dedicated class in the M2 AMI2B@Paris-Saclay (F. Lemoine & Th. Cokelaer)!















cnrs GDR Groupement de recherche

MaDICS Masses de données, informations et connaissances en sciences

ifb

[Reproductibilité de la Recherche | 10.05.2021]

SIF Société Informatique de France