Reproducibility in bioinformatics : a key challenge

Nathalie Lehmann
Hub of Bioinformatics and Biostatistics
Institut Pasteur (Paris)

25.10.2024

What happens in theory… vs real-life

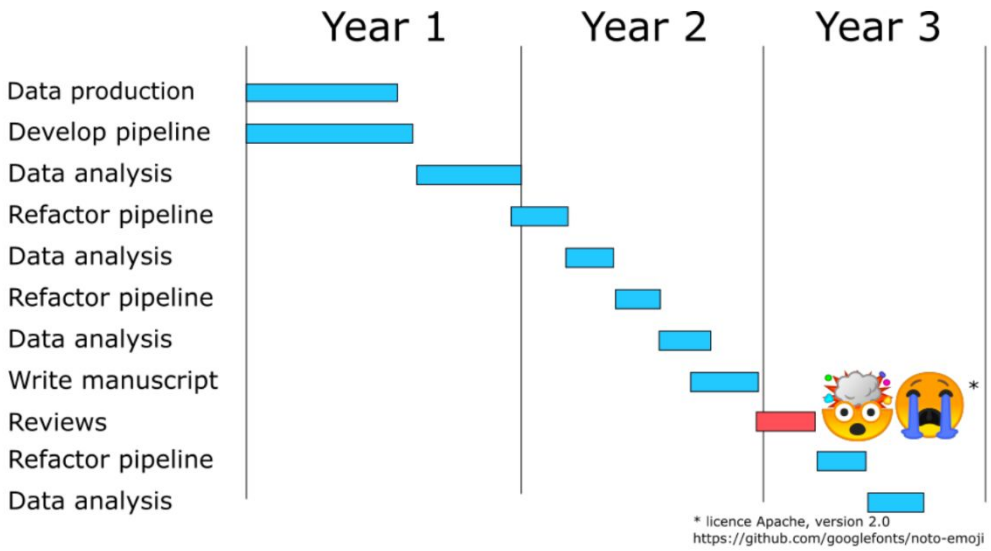
Wet lab experiment
data production

Bioinformatics data analysis

Publish results

Images made by DALL-E
Slide from Frédéric Lemoine

What happens in theory… vs real-life



Reviewers ask for a different version of fig. 1:

- How was this figure generated?
- Where is the right data?
- Where is the right script?
- What version of the libraries?
- How was this file called?

→ **Reproducibility**

Slide from Frédéric Lemoine

What happens in theory… vs real-life

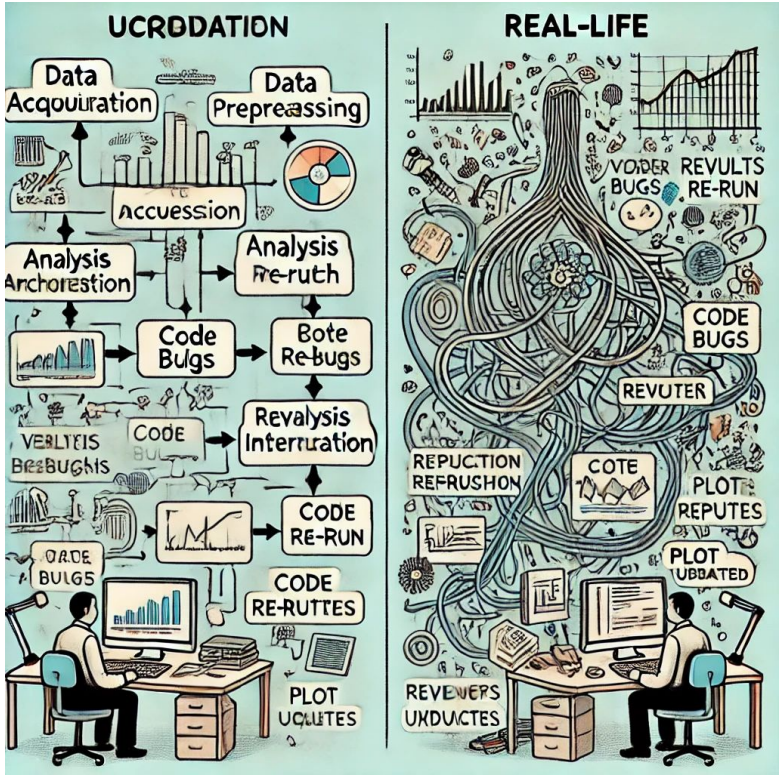


Image made by DALL-E

Importance of reproducibility: concrete examples

1. A collaborator leaves
2. A scientist wants to reproduce your analysis
3. A reviewer asks for new analyses
4. You want to be 100% sure of the results
5. The journal asks for raw data + scripts/workflows (more and more **mandatory**)

- Where is the data?
- Where are the scripts?
- What tools where used, which versions?
- What is the history of the project?
- How to run the analysis?

Slide from Frédéric Lemoine

Different types of reproducibility

As defined by Victoria Stodden, 2013 :

1. **Empirical reproducibility**

2. **Statistical reproducibility**

3. **Computational reproducibility**

Different types of reproducibility (I)

- **Empirical reproducibility** (*Methodological Reproducibility*) :
  - Ability to repeat the same experiment using the same methodology and obtain the same results
  - It focuses on ensuring that enough details are provided so others can replicate the experiment exactly as described
  - *Ex in scRNA-seq : 2 teams working on the same tissue, get similar distribution of cells*

Cell Reports
Commentary

Cell PRESS
Open ACCESS

Sorting Out the FACS: A Devil in the Details

William C. Hines,[1,5,*] Ying Su,[2,3,4,5,*] Irene Kuhn,[1] Kornelia Polyak,[2,3,4,5] and Mina J. Bissell[1,5]
[1]Life Sciences Division, Lawrence Berkeley National Laboratory, Mailstop 977R225A, 1 Cyclotron Road, Berkeley, CA 94720, USA
[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[3]Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[4]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
[5]These authors contributed equally to this work
*Correspondence: chines@lbl.gov (W.C.H.), ying_su@dfci.harvard.edu (Y.S.)
http://dx.doi.org/10.1016/j.celrep.2014.02.021

The reproduction of results is the cornerstone of science; yet, at times, reproducing the results of others can be a difficult challenge. Our two laboratories, one on the East and the other on the West Coast of the United States, decided to collaborate on a problem of mutual interest—namely, the heterogeneity of the human breast. Despite using seemingly identical methods, reagents, and specimens, our two laboratories quite reproducibly were unable to replicate each other's fluorescence-activated cell sorting (FACS) profiles of primary breast cells. Frustration of studying cells close to their context in vivo makes the exercise even more challenging.

Paired with in situ characterizations, FACS has emerged as the technology most suitable for distinguishing diversity among different cell populations in the mammary gland. Flow instruments have evolved from being able to detect only a few parameters to those now capable of measuring up to—and beyond—an astonishing 50 individual markers per cell (Cheung and Utz, 2011). As with any exponential increase in data complexity, breast reduction mammoplasties. Molecular analysis of separated fractions was to be performed in Boston (K.P.'s laboratory, Dana-Farber Cancer Institute, Harvard Medical School), whereas functional analysis of separated cell populations grown in 3D matrices was to take place in Berkeley (M.J.B.'s laboratory, Lawrence Berkeley National Lab, University of California, Berkeley). Both our laboratories have decades of experience and established protocols for isolating cells from primary normal breast tissues as well as the capabilities required for

Different types of reproducibility (II)

- **Statistical reproducibility** :
  - Refers to the reproducibility of statistical results or findings derived from data analysis
  - It ensures that the statistical inferences drawn from the data are consistent when the analysis is repeated, either using the same dataset and methods or slightly different but valid statistical approaches
  - Closely related to robustness or **Inferential Reproducibility**
  - *Ex in scRNA-seq : 2 similar analyses would lead to the same types of results in terms of DEG, p-values and logFC*

P-hacking
False discoveries
Inappropriate models
Model robustness to parameter chang

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

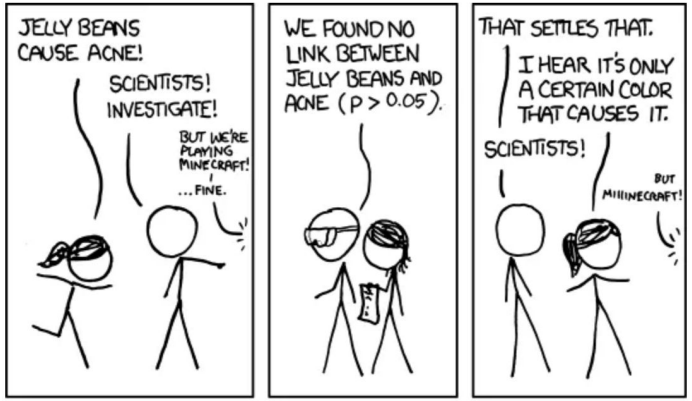
Different types of reproducibility (II)

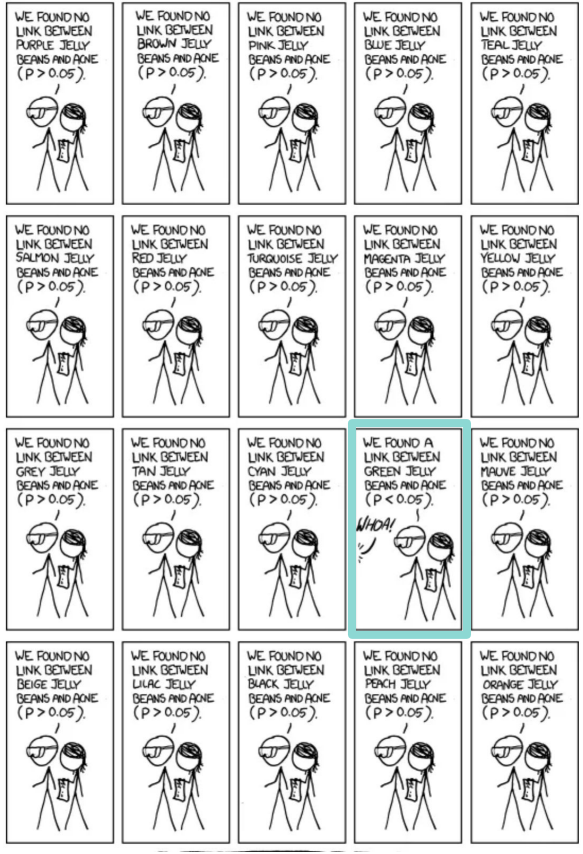
"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

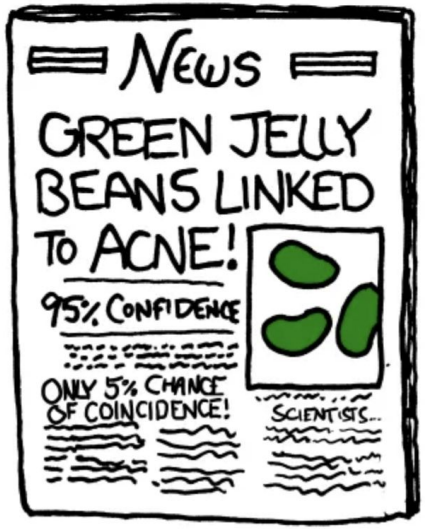
Source: Atoz Markets

Left image from https://labs.getninjas.com.br/p-hacking-eac7186dcd7f
Right image from https://medium.com/nerd-for-tech/p-hacking-explained-45d4980abf11

Different types of reproducibility (II)





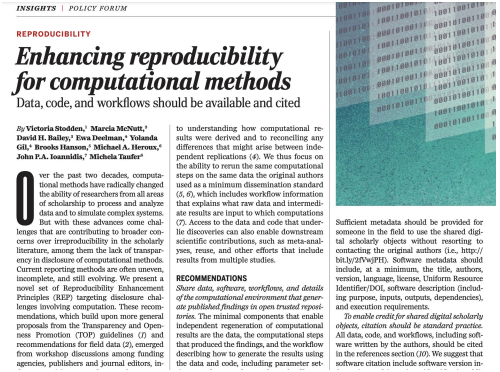


Source: P-Hacking

Image from https://medium.com/nerd-for-tech/p-hacking-explained-45d4980abf11

Different types of reproducibility (III)

- **Computational Reproducibility** (*Result Reproducibility*) **:**
  - Ability to reproduce the computational aspects of a study, which includes reproducing the figures, tables, or other outputs from the data and code provided
  - The goal is to make sure that **given the same code, data, and computing environment, the same results can be obtained**
  - *Ex in scRNA-seq : study publishes UMAP plots for cell clustering and provides the code and processed data, another researcher should be able to generate the same UMAP plots*



INSIGHTS | POLICY FORUM

REPRODUCIBILITY

Enhancing reproducibility for computational methods

Data, code, and workflows should be available and cited

By Victoria Stodden,[1] Marcia McNutt,[2] David H. Bailey,[3] Ewa Deelman,[4] Yolanda Gil,[5] Brooks Hanson,[2] Michael A. Heroux,[6] John P.A. Ioannidis,[7] Michela Taufer[8]

Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transparency in disclosure of computational methods. Current reporting methods are often uneven, incomplete, and still evolving. We present a novel set of Reproducibility Enhancement Principles (REP) targeting disclosure challenges involving computation. These recommendations, which build upon more general proposals from the Transparency and Openness Promotion (TOP) guidelines (*1*) and recommendations for field data (*2*), emerged from workshop discussions among funding agencies, publishers and journal editors, in-

to understanding how computational results were derived and to reconciling any differences that might arise between independent replications (*4*). We thus focus on the ability to rerun the same computational steps on the same data the original authors used as a minimum dissemination standard (*5*, *6*), which includes workflow information that explains what raw data and intermediate results are input to which computations (*7*). Access to the data and code that underlie discoveries can also enable downstream scientific contributions, such as meta-analyses, reuse, and other efforts that include results from multiple studies.

RECOMMENDATIONS

Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories. The minimal components that enable independent regeneration of computational results are the data, the computational steps that produced the findings, and the workflow describing how to generate the results using the data and code, including parameter set-

Sufficient metadata should be provided for someone in the field to use the shared digital scholarly objects without resorting to contacting the original authors (i.e., http://bit.ly/2fVwJFH). Software metadata should include, at a minimum, the title, authors, version, language, license, Uniform Resource Identifier/DOI, software description (including purpose, inputs, outputs, dependencies), and execution requirements.

To enable credit for shared digital scholarly objects, citation should be standard practice. All data, code, and workflows, including software written by the authors, should be cited in the references section (*10*). We suggest that software citation include software version in-

Reproducibility Enhancement Principles (REP) *2016*

Computational reproducibility : 20 years of concern

Nekrutenko & Taylor, *Nature Reviews Genetics*, 2012

Published: 17 August 2012

Next-generation sequencing data interpretation: enhancing reproducibility and accessibility

Anton Nekrutenko & James Taylor

Nature Reviews Genetics **13**, 667–672 (2012) | Cite this article

- 50 papers sampled from 378 published in 2011 using BWA

- 31 : no version, parameters, nor ref. genome version
- 4 : settings
- 8: version
- 7: all details

Alsheikh-Ali et al, *PLoS one* (2011)

Public Availability of Published Research Data in High-Impact Journals

Alawi A. Alsheikh-Ali, Waqas Qureshi, Mouaz H. Al-Mallah, John P. A. Ioannidis

Published: September 7, 2011 • https://doi.org/10.1371/journal.pone.0024357

- 500 papers published in 2009 in highest IF journals

- 149 (30%): no data availability policy → no public data online
- Among the 70% remaning:
  - 208 (59%): did not fully adhere to the policy
  - 143 (41%): deposited only required data + willingness to share

--

- Overall: 47 (9%) full primary raw data online

Slide inspired from Frédéric Lemoine

Many many stories…

Steven Salzberg @StevenSalzberg1 · 1 août

Major, fatal errors found in the data and methods of a 2020 paper in @Nature, including millions of reads mis-identified as bacteria. The "cancer microbiome" in this study was simply not there. @abrahamgihawi @elapertea @YuchenGe1 @JenniferLu717

bioRxiv — THE PREPRINT SERVER FOR BIOLOGY

biorxiv.org
Major data analysis errors invalidate cancer micro...
We re-analyzed the data from a recent large-scale study that reported strong correlations between ...

35 · 582 · 1 502 · 564,8 k

Steven Salzberg @StevenSalzberg1 · 3 août

New story in @statnews by @angRchen and @matthewherper about the big, big problems we discovered in a @Nature paper that reported finding a microbiome associated with 32 cancer types

STAT @statnews · 3 août
Computational biologist @StevenSalzberg1 says the problems with a Nature paper about a microbiome cancer diagnostic are serious. trib.al/N3M72f0

1 · 10 · 46 · 28,6 k

Steven Salzberg @StevenSalzberg1 · 3 août

and once again, in the quotes from Knight et al., they don't address any of the problems in their study, instead just claiming that "other work" supports it. That doesn't fix the problems

2 · 9 · 2 999

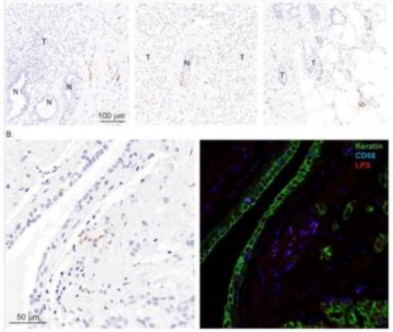
Steven Salzberg @StevenSalzberg1

Yet another major blow to the hypothesis that a microbiome of cancer exists. TLDR: the main results from a 2020 @ScienceMagazine paper claiming to find bacteria in breast cancer simply doesn't hold up. Well done @NFdeMiranda, Jacques Neefjes, et al

Traduire le post

Noel F. de Miranda @NFdeMiranda · 29 août

In a bid to replicate a prior study, we couldn't confirm LPS presence within breast cancer cells. We did spot it around ducts & in macrophages, aligning with its biology.
#ResearchReplication #Cancer #Microbiome
biorxiv.org/content/10.110...



Legend Figure 1 – A - Representative examples of LPS immunodetection with typical granular pattern. LPS expression never co-localized with cancer cells. B – Left: LPS detection of a breast cancer section

2:17 PM · 29 août 2023 · **96,5 k** vues

Human Microbiome | Research Article | 9 October 2023

Major data analysis errors invalidate cancer microbiome findings

Authors: Abraham Gihawi, Yuchen Ge, Jennifer Lu, Daniela Puiu, Amanda Xu, Colin S. Cooper, Daniel S. Brewer, Mihaela Pertea, Steven L. Salzberg | AUTHORS INFO & AFFILIATIONS

DOI: https://doi.org/10.1128/mbio.01607-23 • Check for updates

PDF/EPUB

ABSTRACT

We re-analyzed the data from a recent large-scale study that reported strong correlations between DNA signatures of microbial organisms and 33 different cancer types and that created machine-learning predictors with near-perfect accuracy at distinguishing among cancers. We found at least two fundamental flaws in the reported data and in the methods: (i) errors in the genome database and the associated computational methods led to millions of false-positive findings of bacterial reads across all samples, largely because most of the sequences identified as bacteria were instead human; and (ii) errors in the transformation of the raw data created an artificial signature, even for microbes with no reads detected, tagging each tumor type with a distinct signal that the machine-learning programs then used to create an apparently accurate classifier. Each of these problems invalidates the results, leading to the conclusion that the microbiome-based classifiers for identifying cancer presented in the study are entirely wrong. These flaws have subsequently affected more than a dozen additional published studies that used the same data and whose results are likely invalid as well.

Slide inspired from Frédéric Lemoine

Many many stories…

SCIENTIFIC PUBLISHING

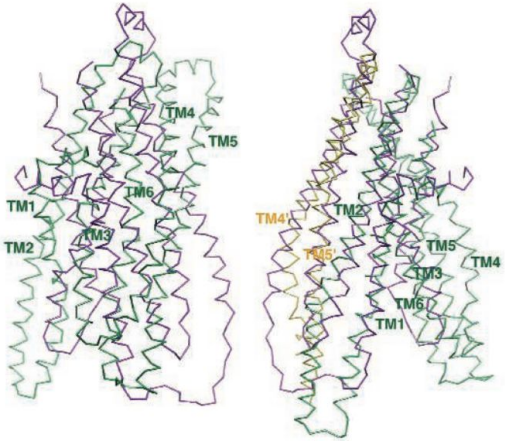
A Scientist's Nightmare: Software Problem Leads to Five Retractions

Until recently, Geoffrey Chang's career was on a trajectory most young scientists only dream about. In 1999, at the age of 28, the protein crystallographer landed a faculty position at the prestigious Scripps Research Institute in San Diego, California. The next year, in a ceremony at the White House, Chang received a Presidential Early Career Award for Scientists and Engineers, the country's highest honor for young researchers. His lab generated a stream of high-profile papers detailing the molecular structures of important proteins embedded in cell membranes.

Then the dream turned into a nightmare. In September, Swiss researchers published a paper in *Nature* that cast serious doubt on a protein structure Chang's group had described in a 2001 *Science* paper. When he investigated, Chang was horrified to discover that a homemade data-analysis program had flipped two columns of data, inverting the electron-density map from which his team had derived the final protein structure. Unfortunately, his group had used the program to analyze data for

2001 *Science* paper, which described the structure of a protein called MsbA, isolated from the bacterium *Escherichia coli*. MsbA belongs to a huge and ancient family of molecules that use energy from adenosine triphosphate to transport molecules across cell membranes. These so-called ABC transporters perform many

Sciences ar
EmrE, a dif
Crystal
five memb
was an inc
postdoc ad
nia Institut
proteins ar
because th
ously diff
needed for
determinat
cess: "He
ethic. He r



Flipping fiasco. The structures of MsbA (purple) and Sav1866 (green) overlap little (*left*) until MsbA is inverted (*right*).

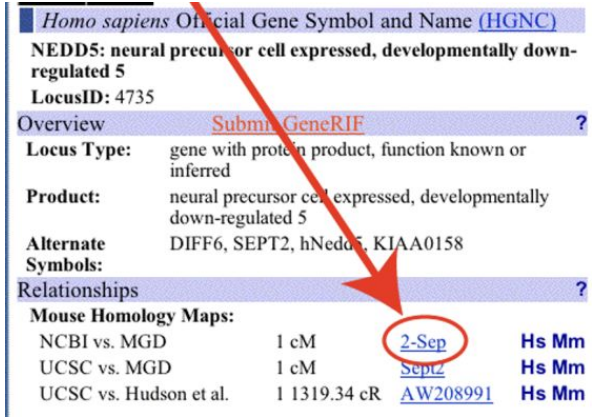
The curse of the Excel spreadsheet (but not only…)

Correspondence | Open access | Published: 23 June 2004

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

Barry R Zeeberg, Joseph Riss, David W Kane, Kimberly J Bussey, Edward Uchio, W Marston Linehan, J Carl Barrett & John N Weinstein

BMC Bioinformatics **5**, Article number: 80 (2004) | Cite this article



The curse of the Excel spreadsheet (but not only…)

Comment | Open access | Published: 23 August 2016

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren & Assam El-Osta

Genome Biology **17**, Article number: 177 (2016) | Cite this article

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

The curse of the Excel spreadsheet (but not only…)

PLOS COMPUTATIONAL BIOLOGY

PLOS Comput Biol. 2021 Jul 30;17(7): e1008984. doi: 10.1371/journal.pcbi.1008984

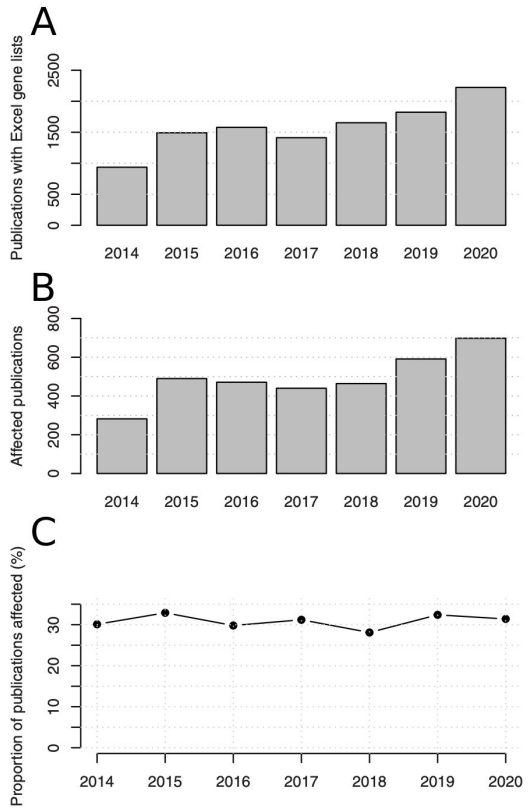
Gene name errors: Lessons not learned

Mandhri Abeysooriya [1], Megan Soria [1], Mary Sravya Kasu [1], Mark Ziemann [1,*]

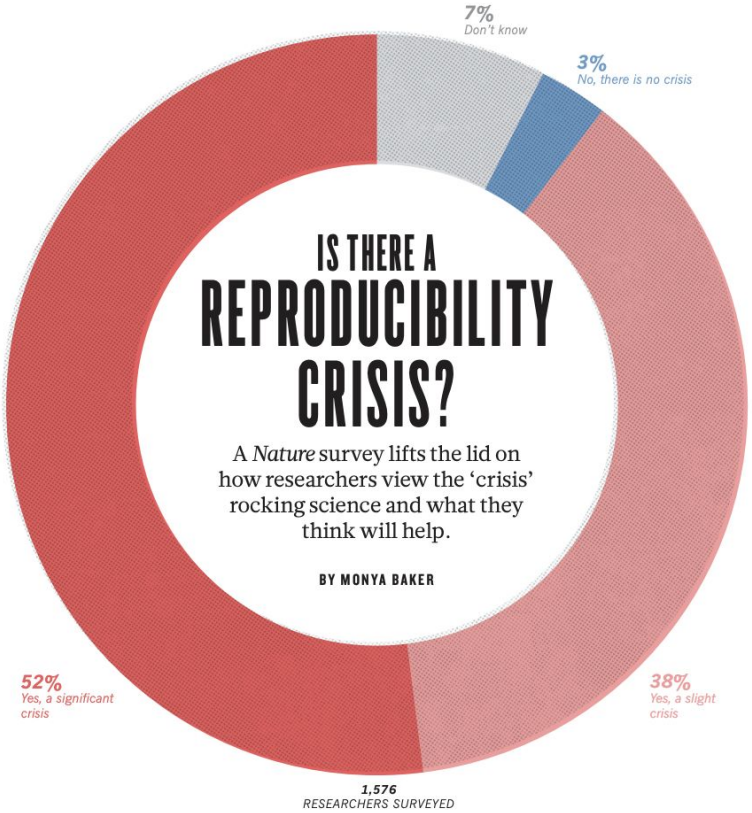
Editor: Christos A Ouzounis[2]

Author information · Article notes · Copyright and License information

PMCID: PMC8357140 PMID: 34329294

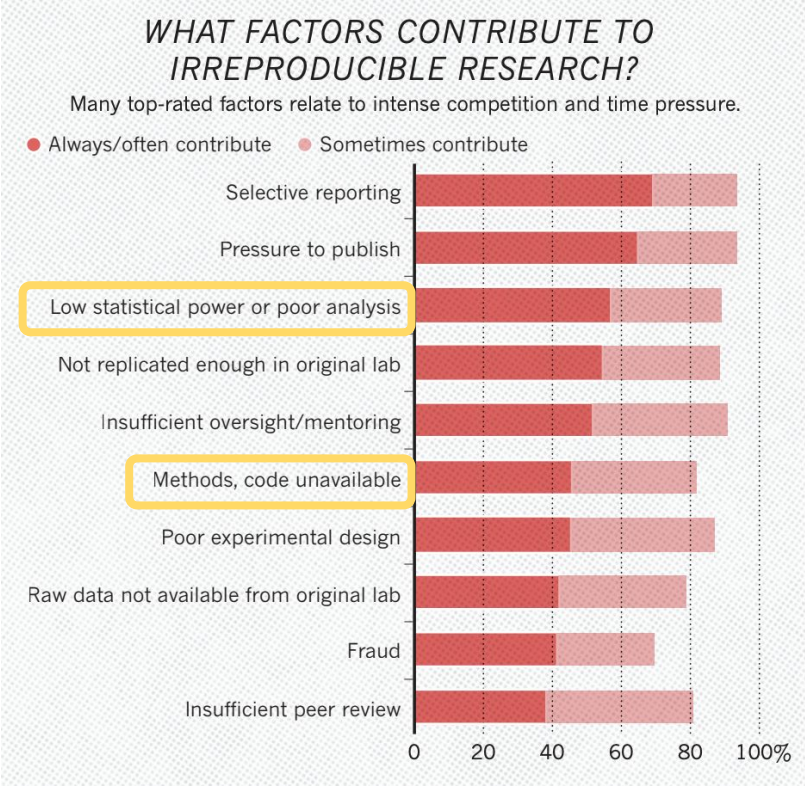


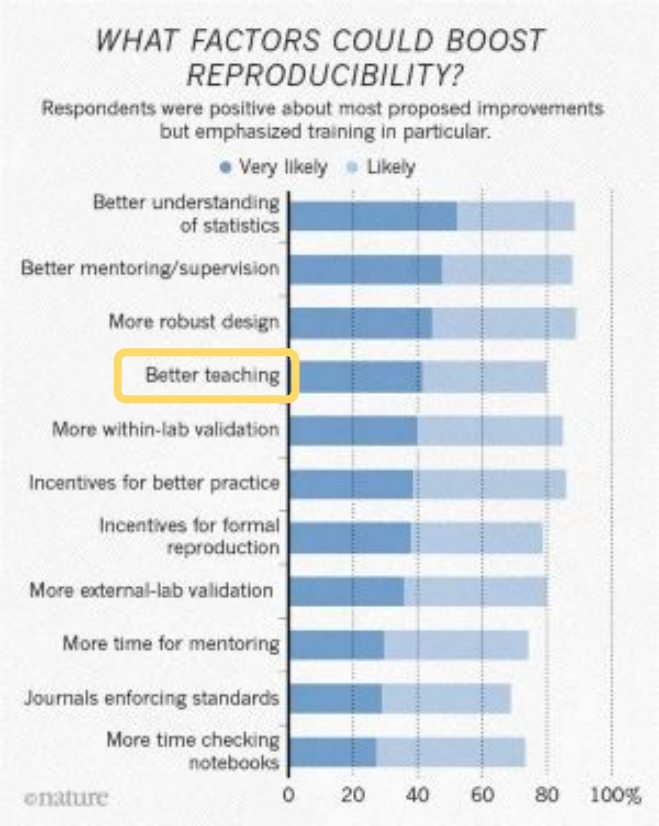
Current awareness



Nature, 2016

What can we do to improve the situation ?





https://www.nature.com/articles/533452a

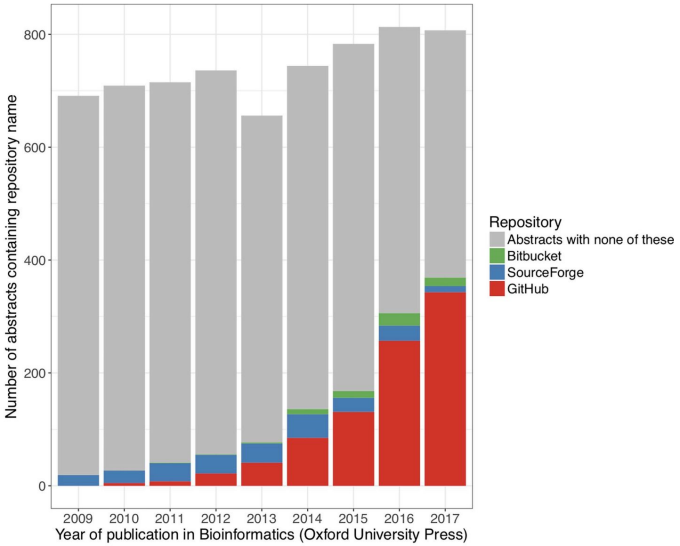
Nature, 2016

Investigation of the state of source code in the bioinformatics community

A large-scale analysis of bioinformatics code on GitHub

Pamela H. Russell, Rachel L. Johnson, Shreyas Ananthan, Benjamin Harnke, Nichole E. Carlson

Published: October 31, 2018 • https://doi.org/10.1371/journal.pone.0205898



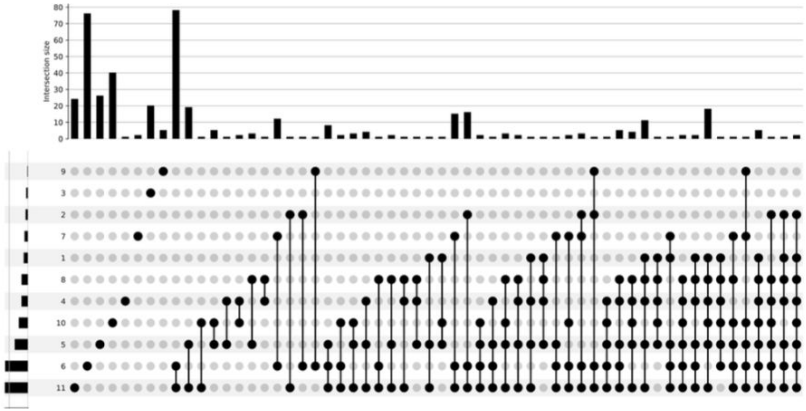
Recent initiatives : reprohackathons

JOURNAL ARTICLE

Reprohackathons: promoting reproducibility in bioinformatics through training

Thomas Cokelaer, Sarah Cohen-Boulakia, Frédéric Lemoine Author Notes

Bioinformatics, Volume 39, Issue Supplement_1, June 2023, Pages i11–i20,
https://doi.org/10.1093/bioinformatics/btad227
Published: 30 June 2023



- High variability in the results
- Very **few genes found in common** between groups
- Revealing the high difficulty to reproduce analyses!

Slide inspired from Frédéric Lemoine

Recent initiatives : scFAIR

scFAIR

Home Data Resources Metadata schema Tools Community About Contact us

There exist many tools that can help capture, store, access, and share single-cell data in a FAIR way.

Cell Type Annotation

Cell Annotation Schema (CAS) and cas-tools: The Cell Annotation Schema is a general, open-standard schema for cell annotations and related metadata. CAS provides a programmatically accessible standard designed that allows users to record additional metadata about individual cell type annotations, including marker genes used as evidence and details of automated annotation transfer. The standard is represented as JSON schema as this allows all metadata to be gathered in a single, compact validatable file - which includes a link to a cell by gene matrix file of annotated data. However, the schema is designed so that it can be decomposed into individual tables suitable for use in dataframes/TSVs and flattened onto obs in AnnData format. CAS-Tools is a comprehensive utility package designed to facilitate the effective use and manipulation of the Cell Annotation Schema (CAS) in single-cell transcriptomics data analysis.

Annotation File Validators

CELLxGENE schema validator: CELLxGENE curation tools includes a schema validator that can validate single-cell annotation h5ad files in accordance with the CELLxGENE metadata schema. The `cellxgene-schema validate` command checks an annotation file and will print validation failure messages or a validation success message.

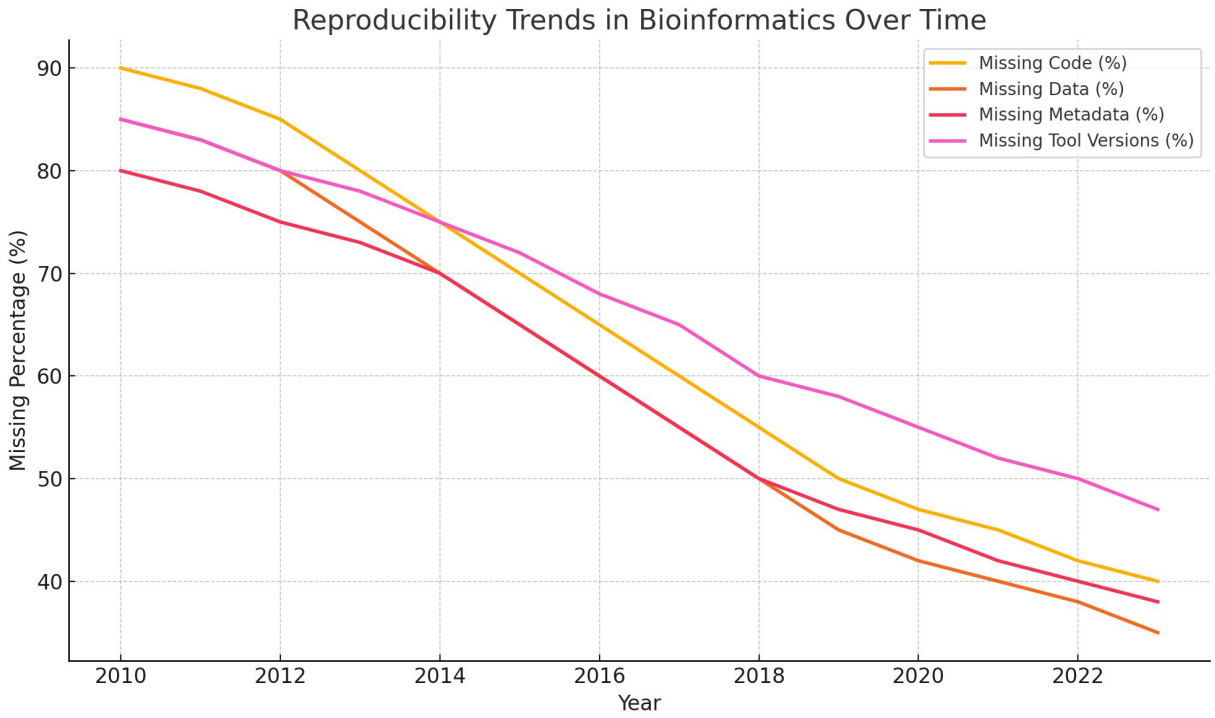
I asked chatGPT about reproducibility in single cell…

In single-cell RNA-seq studies, reproducibility is often compromised due to incomplete reporting of critical details such as code, data, metadata, and tool versions.

1. **Code Availability**: Approximately 40-50% of single-cell studies provide access to the code used for analysis. While many repositories such as GitHub now include projects associated with single-cell papers, a significant proportion of studies still lack the availability of reproducible code PAPERS WITH CODE GITHUB.
2. **Data and Metadata**: Only about 13-25% of studies deposited in public repositories like the Gene Expression Omnibus (GEO) include necessary cell-level metadata. Metadata is crucial for replicating results, but a substantial number of studies fail to provide this, limiting the reproducibility of key analyses like differential expression tests PLOS POLLY.
3. **Tool and Version Reporting**: Tool and version reporting is often inconsistent. While some guidelines are emerging, many studies still do not comprehensively document the versions of software, algorithms, or specific parameters used, which is critical since minor differences in these can significantly affect outcomes PLOS POLLY.

In summary, missing metadata and the lack of standardized reporting on code and software versions pose significant barriers to reproducibility in the field. Efforts to harmonize metadata and standardize reporting practices are ongoing but remain far from universal BIGOMICS ANALYTICS.

I asked chatGPT about reproducibility in bioinformatics over time…

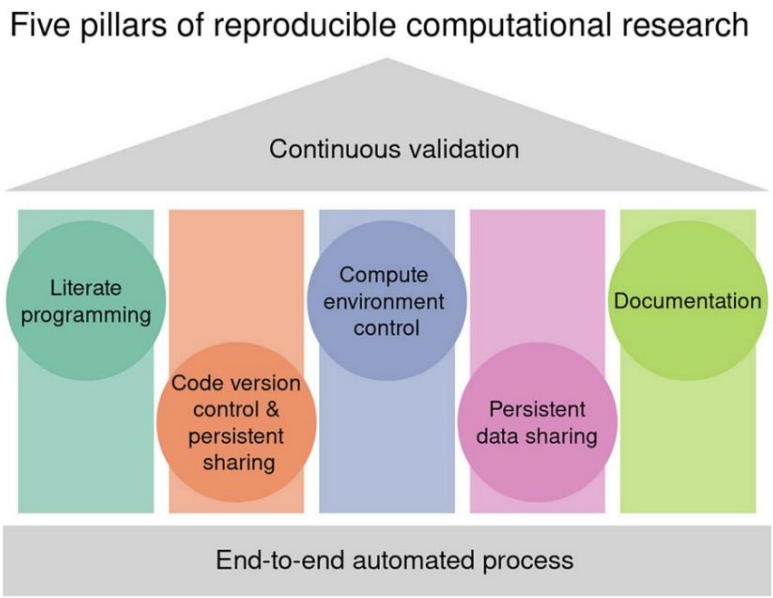


What can we do to improve the situation: best practices

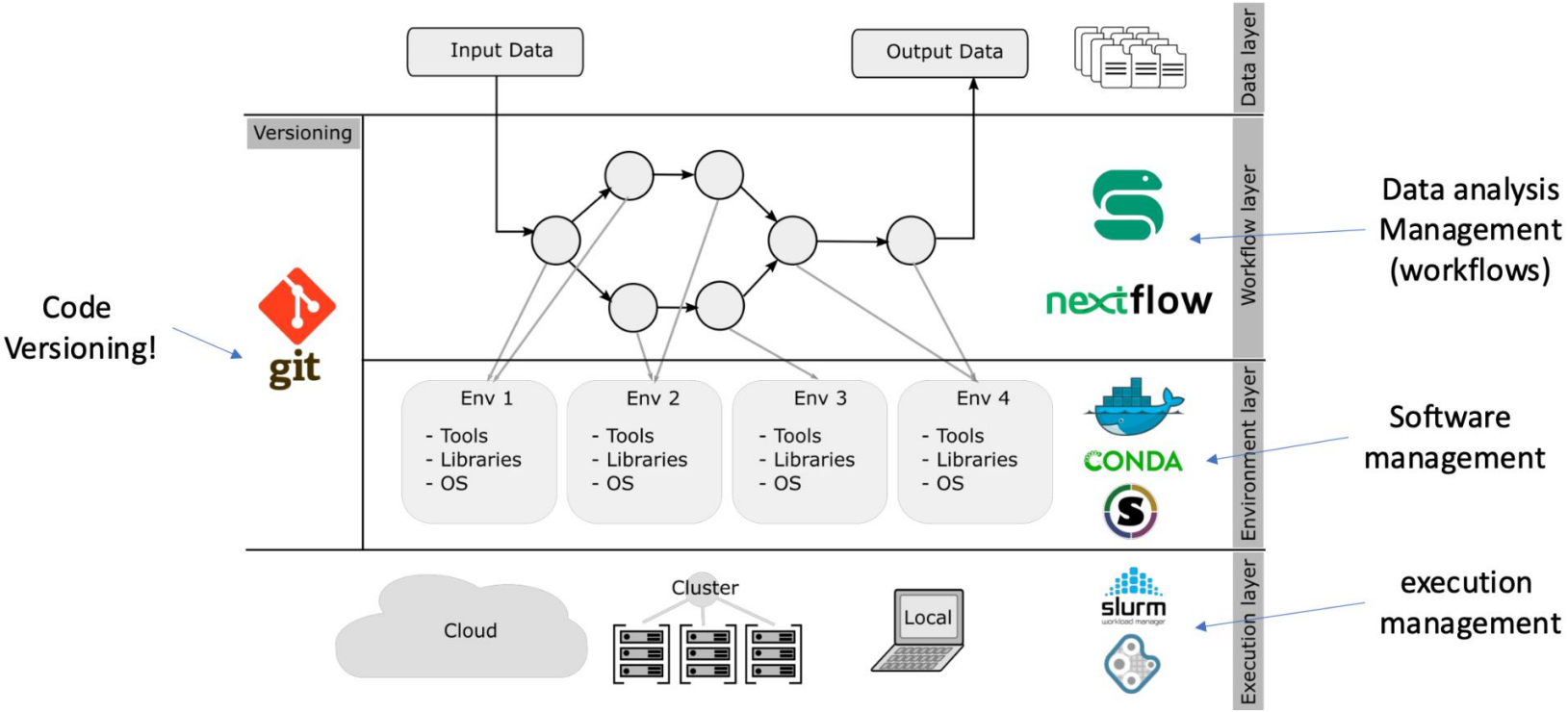
The five pillars of computational reproducibility: bioinformatics and beyond

Mark Ziemann, Pierre Poulain and Anusuiya Bora

Corresponding author: Mark Ziemann, School of Life and Environmental Sciences, Deakin University, 75 Pigdons Rd, Waurn Ponds, VIC 3216, Australia. Tel.: +61 3 522 78965; E-mail: m.ziemann@deakin.edu.au



What can we do to improve the situation: best practices



And last but not least

“If everyone on a research team knows that everything they do is going to someday be published for reproducibility, they’ll behave differently from day one” (Donoho et al, 2009)

Reproducibility practices are for yourself first !

Resources (I)

FUN FRANCE UNIVERSITÉ NUMÉRIQUE
The freedom to study

News | Courses | FormaPro | Organizations | About FUN | FAQ | LOG IN | English

You are here: Home › Courses
› Reproducible research: methodological principles for transparent science

Search for courses

Tools for research | Digital and technology

Reproducible research: methodological principles for transparent science

Ref. 41016

This Mooc proposes methodological principles for open and transparent science. It deals in a practical way with note-taking, computational documentation, replicability of analyses.

Effort: 24 hours | Pace: Self paced

Languages: French

Enrollment
From Jan 10, 2020 to Dec 11, 2024

Outils pour la recherche | Numérique et technologie

Reproducible Research II: Practices and tools for managing computations and data

Réf. 41023

In this MOOC, we will show you how to improve your practices and your ability to manage and process larger amounts of data, complex computations, while controlling your software environment.

Durée : 4 mois | Effort : 35 heures | Rythme: ~8h45/mois

Langues: Anglais

Resources (II)

IFB Cours Français (fr)

Institut Français de Bioinformatique - Les formations

Accueil

La formation à l'IFB

Les ressources pédagogiques de l'Institut Français de Bioinformatique

Cours

- Formations IFB sur le thème du FAIR
  - FAIR-BIOINFO
  - FAIR-DATA
- Formations analyse de données de séquençage haut débit
- Formations Bioinformatique Intégrative
- Single-Cell Workshops
- Omics analysis
- E-formation
- Pratiques pédagogiques

Cours disponibles

WF4bioinfo 2024 : Les langages de workflows pour une analyse bioinformatique reproductible

FAIR Bioinfo 2024 [Strasbourg] : principes FAIR dans un projet de bioinformatique

scFAIR

Home Data Resources Metadata schema Tools Community About Contact us

Community

Workshops

Future workshops:

- March 7 2024: "Enabling FAIR access to single-cell RNA-Seq data for reproducible analyses", International Biocuration Conference, India
- June 24 2024: "Single cell RNA sequencing data analysis: Requirements for reproducibility and meaningful multi-omics integration", The Swiss Bioinformatics Summit

Past workshops:

- September 11 2023: "Standardization of single-cell metadata: an Open Research Data initiative", Basel Computational Biology Conference

Feedback | Privacy policy

Resources (III)

Baykal *et al. Genome Biology* (2024) 25:213
https://doi.org/10.1186/s13059-024-03343-2

Genome Biology

REVIEW **Open Access**

Genomic reproducibility in the bioinformatics era

Pelin Icer Baykal[1,2], Paweł Piotr Łabaj[3,4], Florian Markowetz[5,6], Lynn M. Schriml[7], Daniel J. Stekhoven[2,8], Serghei Mangul[9,10*†] and Niko Beerenwinkel[1,2*†]

The five pillars of computational reproducibility: bioinformatics and beyond

Mark Ziemann, Pierre Poulain and Anusuiya Bora

Corresponding author: Mark Ziemann, School of Life and Environmental Sciences, Deakin University, 75 Pigdons Rd, Waurn Ponds, VIC 3216, Australia. Tel.: +61 3 522 78965; E-mail: m.ziemann@deakin.edu.au

Resources (IV)

Guidelines for reporting single-cell RNA-Seq experiments

31 Oct 2019 · Anja Füllgrabe, Nancy George, Matthew Green, Parisa Nejad, Bruce Aronow, Laura Clarke, Silvie Korena Fexova, Clay Fischer, Mallory Ann Freeberg, Laura Huerta, Norman Morrison, Richard H. Scheuermann, Deanne Taylor, Nicole Vasilevsky, Nils Gehlenborg, John Marioni, Sarah Teichmann, Alvis Brazma, Irene Papatheodorou · Edit social preview

Single-cell RNA-Sequencing (scRNA-Seq) has undergone major technological advances in recent years, enabling the conception of various organism-level cell atlassing projects. With increasing numbers of datasets being deposited in public archives, there is a need to address the challenges of enabling the reproducibility of such data sets. Here, we describe guidelines for a minimum set of metadata to sufficiently describe scRNA-Seq experiments, ensuring reproducibility of data analyses.

5 sacrés travers de la science par p-value

YouTube · Science4All
10 juin 2019

La plus grosse confusion des sciences : la p-value !! 🌶️

YouTube · Science4All
27 mai 2019

Resources (V)

Top-10 Readings in Reproducibility

Early this year, my student Olivier and I were getting started writing a book chapter and later a full-length journal article; the first was about our reproducible-research workflow and the second on our CFD replication study. These represented about three years of work, not exclusively on this project, but taking most of the graduate student's time. As part of our "pre-writing" tasks, we decided to build—collectively as a group—our list of Top 10 papers discussing reproducible research in computational science. Here's our current reading list (modified from our first version of Feb. 2016):

1. Schwab, M., Karrenbach, N., Claerbout, J. (2000) Making scientific computations reproducible, *Comp. Sci. Eng.* 2(6):61–67, doi: 10.1109/5992.881708

2. Donoho, D. et al. (2009), Reproducible research in computational harmonic analysis, *Comp. Sci. Eng.* 11(1):8–18, doi: 10.1109/MCSE.2009.15

3. Reproducible Research, by the Yale Law School Roundtable on Data and Code Sharing, *Comp. Sci. Eng.* 12(5): 8–13 (Sept.-Oct. 2010), doi:10.1109/mcse.2010.113

4. Peng, R. D. (2011), Reproducible research in computational science, *Science* 334(6060): 1226–1227, doi: 10.1126/science.1213847

5. Diethelm, Kai (2012) The limits of reproducibility in numerical simulation, *Comp. Sci. Eng.* 14(1): 64-72, doi: 10.1109/MCSE.2011.21

6. Setting the default to reproducible (2013), ICERM report of the Workshop on Reproducibility in Computational and Experimental Mathematics (Providence, Dec. 10-14, 2012), Stodden et al. (eds.), https://icerm.brown.edu/tw12-5-rcem/ // report PDF

7. Sandve, G. K. et al. (2013), Ten simple rules for reproducible computational research, *PLOS Comp. Bio.* (editorial), Vol. 9(10):1–4, doi: 10.1371/journal.pcbi.1003285

8. Leek, J. and Peng, R (2015), Opinion: Reproducible research can still be wrong: Adopting a prevention approach, *PNAS* 112(6):1645–1646, doi: 10.1073/pnas.1421412111

9. M. Liberman, "Replicability vs. reproducibility — or is it the other way around?," Oct. 2015, http://languagelog.ldc.upenn.edu/nll/?p=21956

10. Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine* 8(341), 341ps12-341ps12, doi: 10.1126/scitranslmed.aaf5027

https://lorenabarba.com/blog/barbagroup-reproducibility-syllabus/