

Recommendations to enhance rigor and reproducibility in biomedical research

Jacqueline J. Brito, PhD
Mangul Lab - USC



Cornell University

arXiv.org > q-bio > arXiv:2001.05127

the S

Search.

Help | A

Quantitative Biology > Other Quantitative Biology

Enhancing rigor and reproducibility by improving software availability, usability, and archival stability

Jaqueleine J. Brito, Jun Li, Jason H. Moore, Casey S. Greene, Nicole A. Nogoy, Lana X. Carmire, Serghei Mangul

(Submitted on 15 Jan 2020)

Computational methods have reshaped the landscape of modern biology. While the biomedical community is increasingly dependent on computational tools, the mechanisms ensuring open data, open software, and reproducibility are variably enforced. Publications may describe the software for which source code is unavailable, documentation is incomplete or unmaintained, and analytical source code is missing. Publications that lack this information compromise the role of peer review in evaluating technical strength and scientific contribution. Such flaws also limit any subsequent work that intends to use the described software. We herein provide recommendations to improve reproducibility, transparency, and rigor in computational biology -- precisely the values which should be emphasized in foundational life and medical science curricula. Our recommendations for improving software availability, usability, and archival stability aim to foster a sustainable data science ecosystem in biomedicine and life science research.

Subjects: **Other Quantitative Biology (q-bio.OT)**

Cite as: [arXiv:2001.05127](#) [q-bio.OT]
(or [arXiv:2001.05127v1](#) [q-bio.OT] for this version)

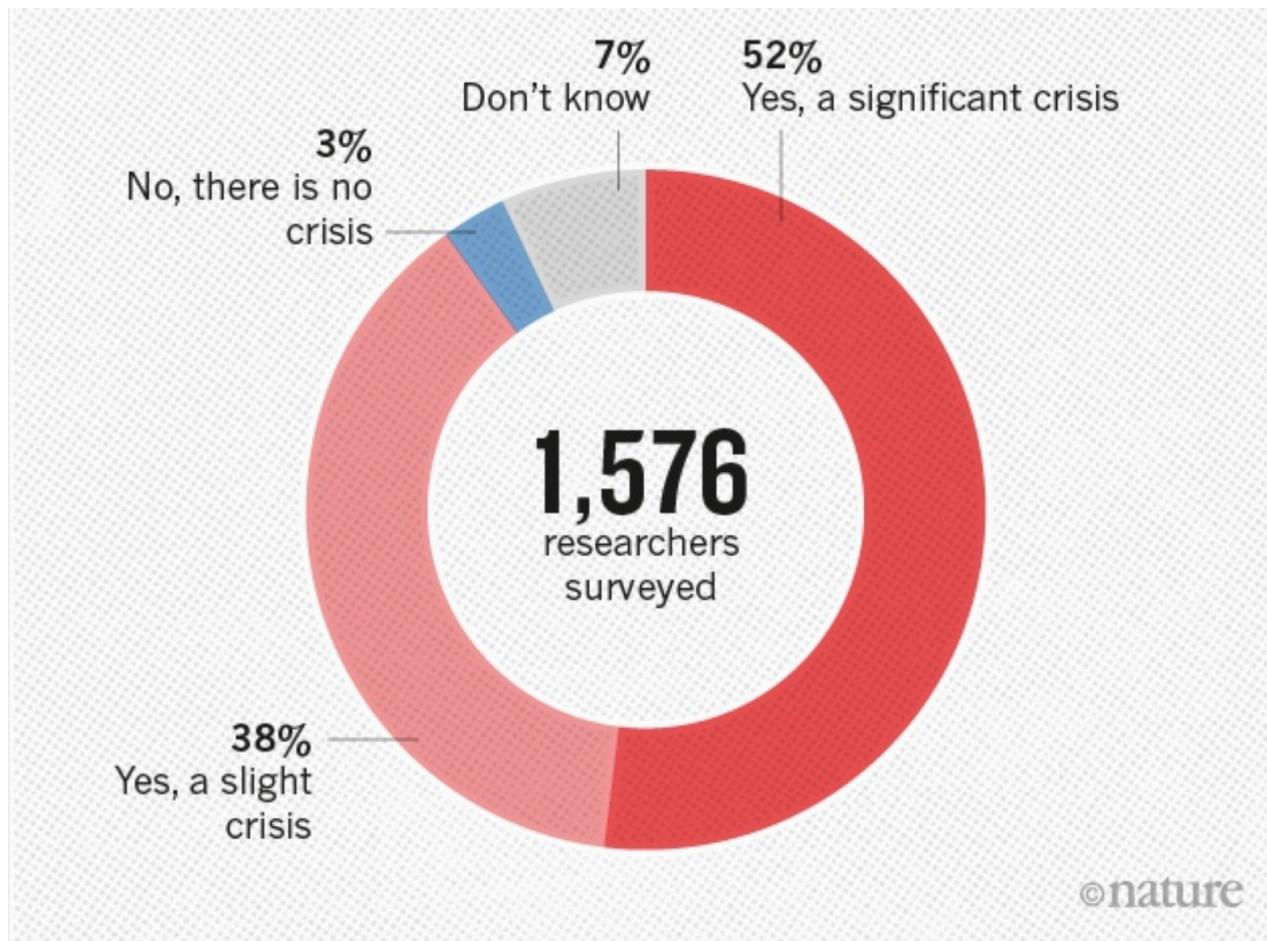
Bibliographic data

[[Enable Bibex](#) ([What is Bibex?](#))]

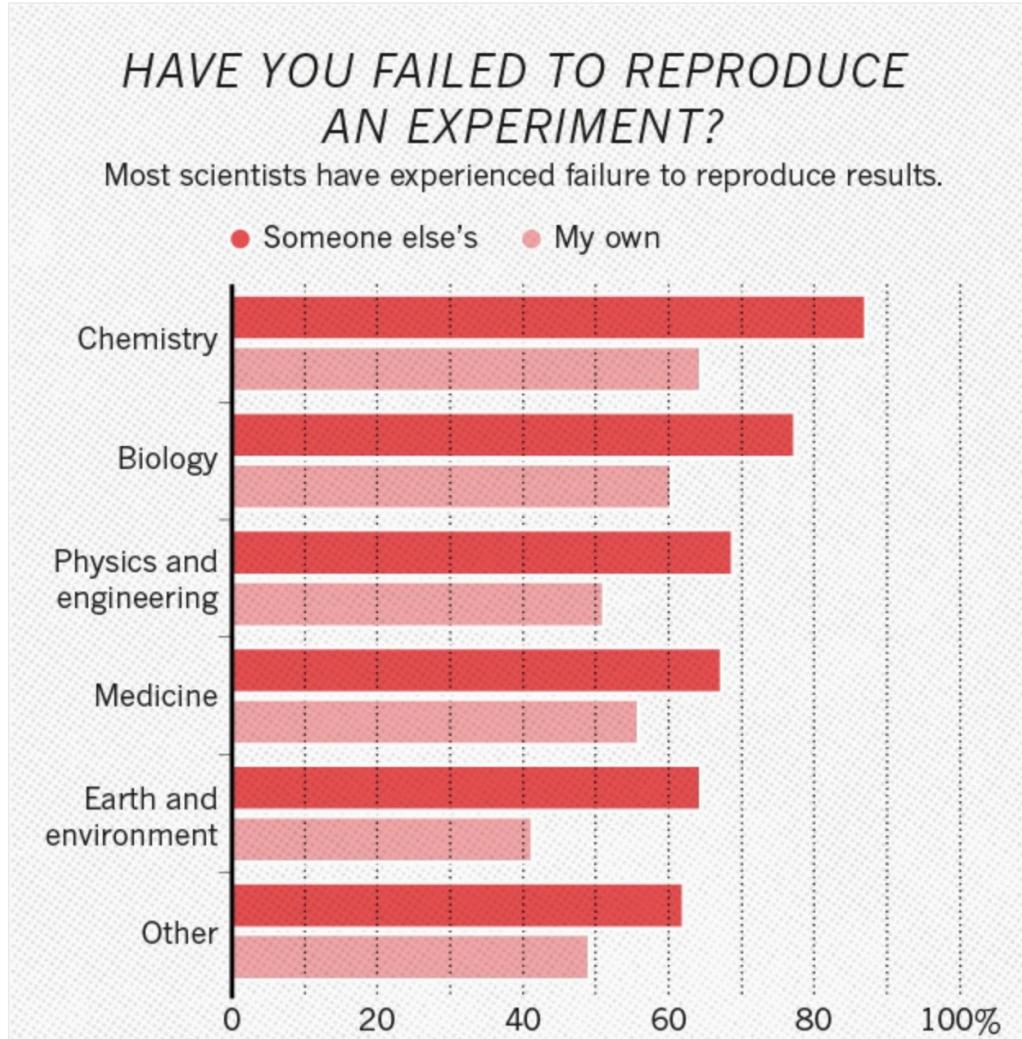
Submission history

From: Jaqueleine Brito [[view email](#)]
[v1] Wed, 15 Jan 2020 04:22:15 UTC (241 KB)

Reproducibility crisis?

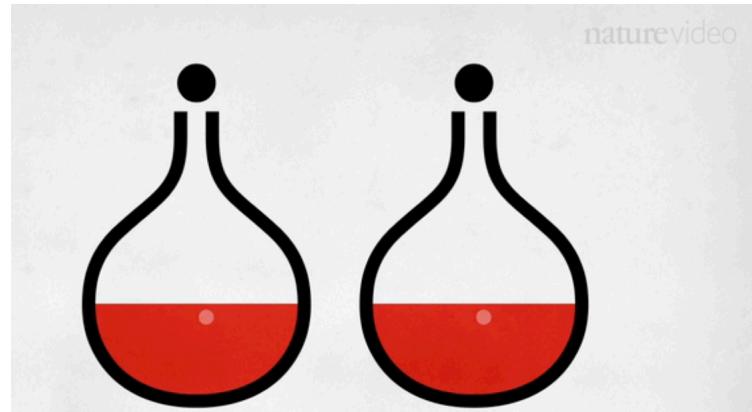


Reproducibility crisis?



Ideal scenario with all information

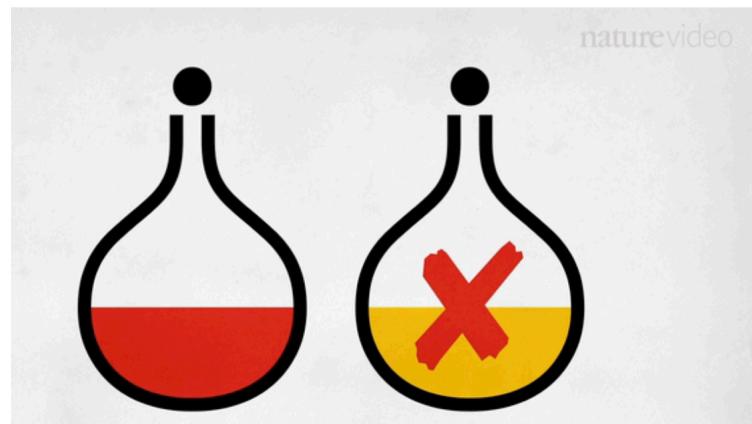
Researchers obtain different results even when they follow detailed procedures



Several factors can affect results

- Human factor
- Environment settings

No guarantee of same results



The reality is even worse

Many researchers do not share the information necessary to reproduce their results



Will it ever be possible to reproduce biomedical research routinely?

Many aspects are involved

- What is the level of details necessary to be provided to ensure reproducibility?
- Can the human factors be leveraged?

Is it feasible to develop procedures to ensure reproducibility?

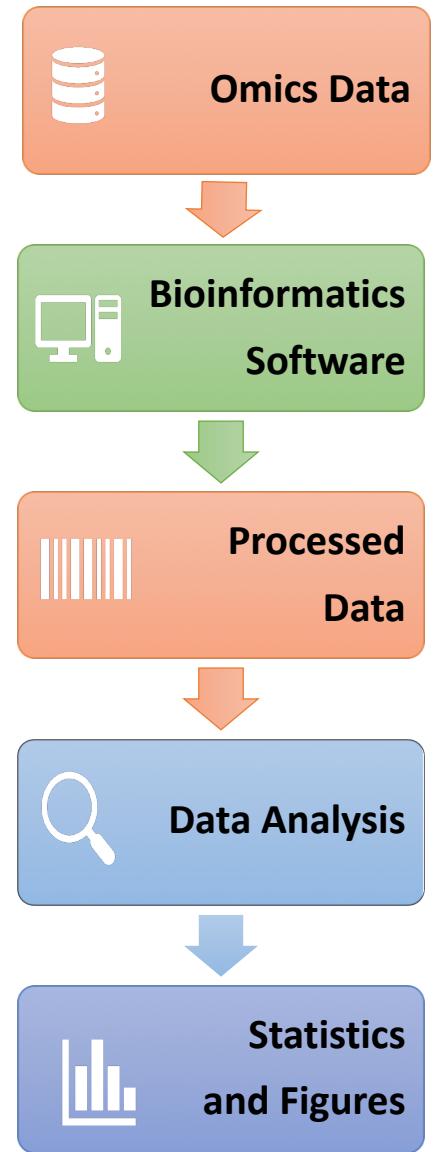
Unique opportunity in computational biomedical research!

Research cycle of automatic processes that can be shared

Still many studies are still not reproducible

- No data
- No source code

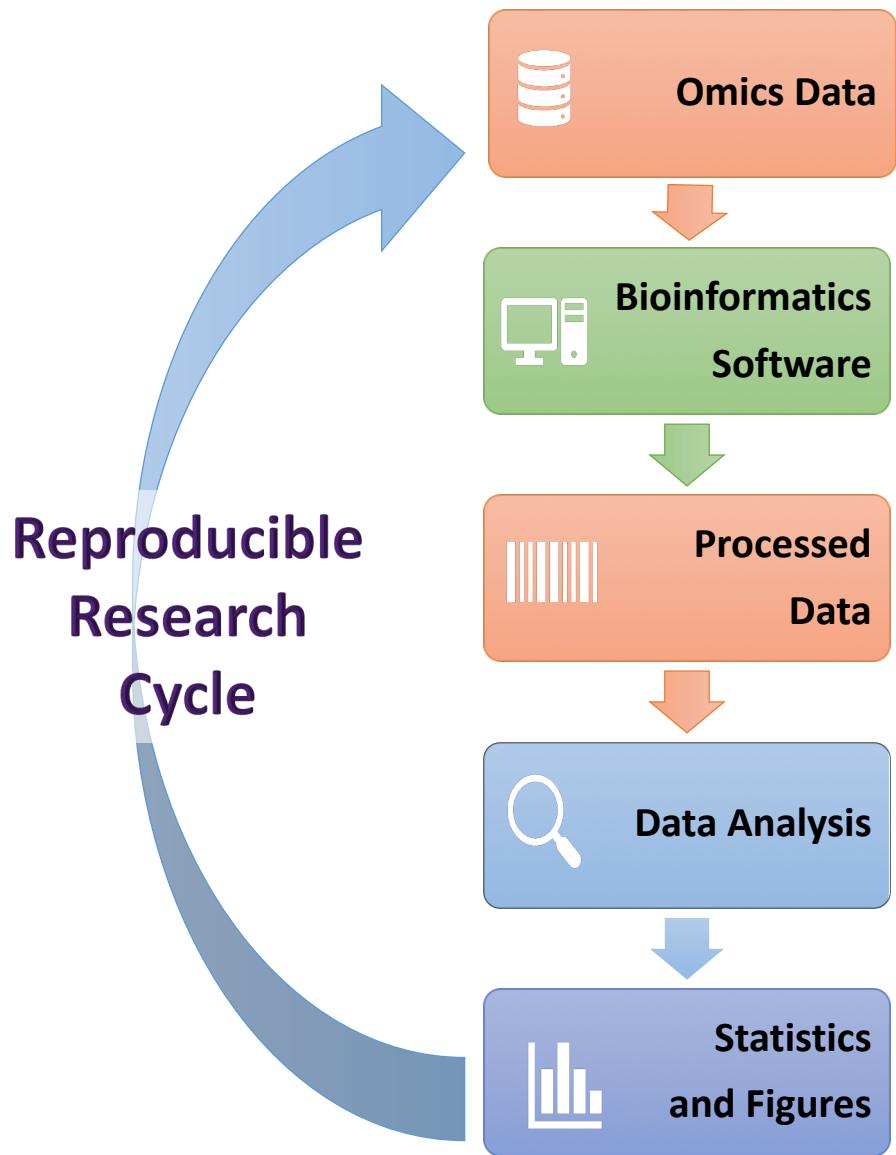
Reproducible Research Cycle



How do we enable reproducible results?

The problem can be solved by sharing all data and code across the entire cycle

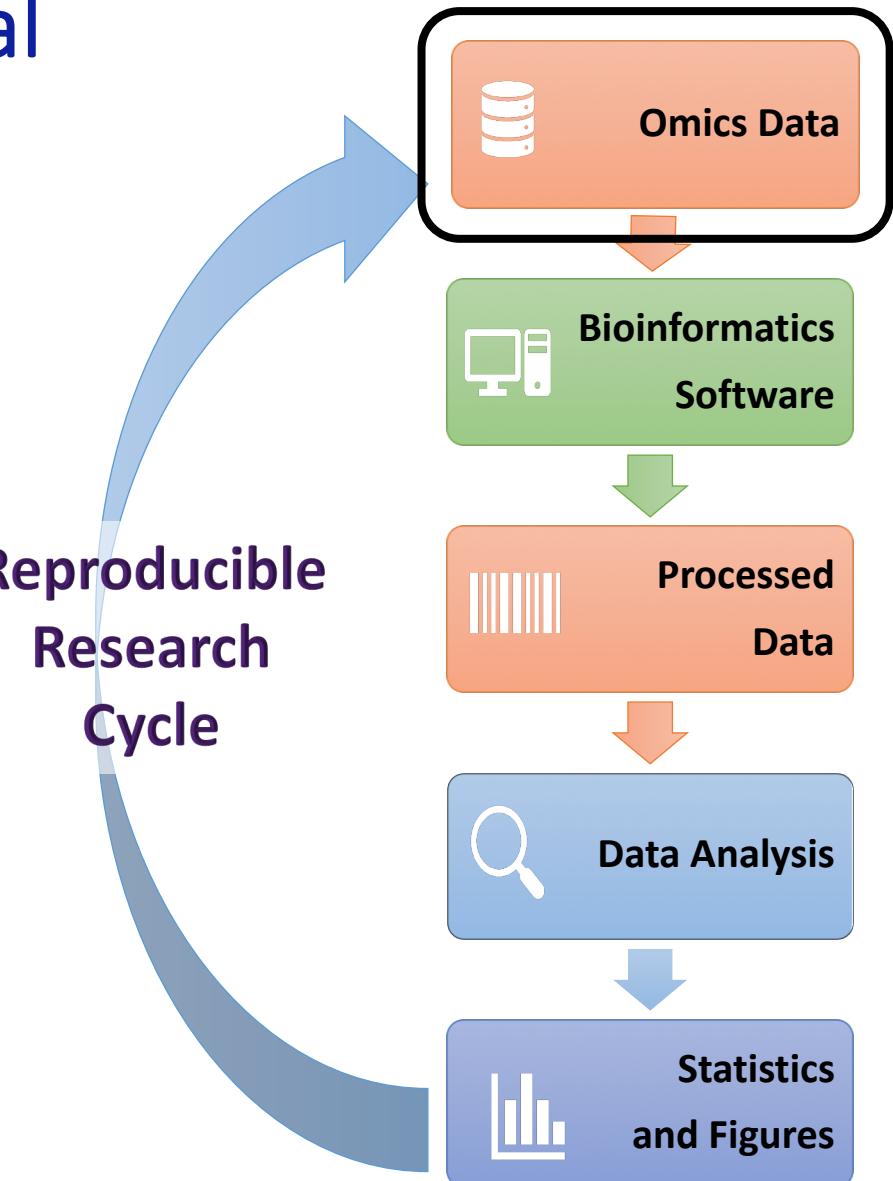
We need to follow certain practices to enable and facilitate reproducibility



Data access is essential

Data is essential for replicating and auditing results

Reproducible Research Cycle

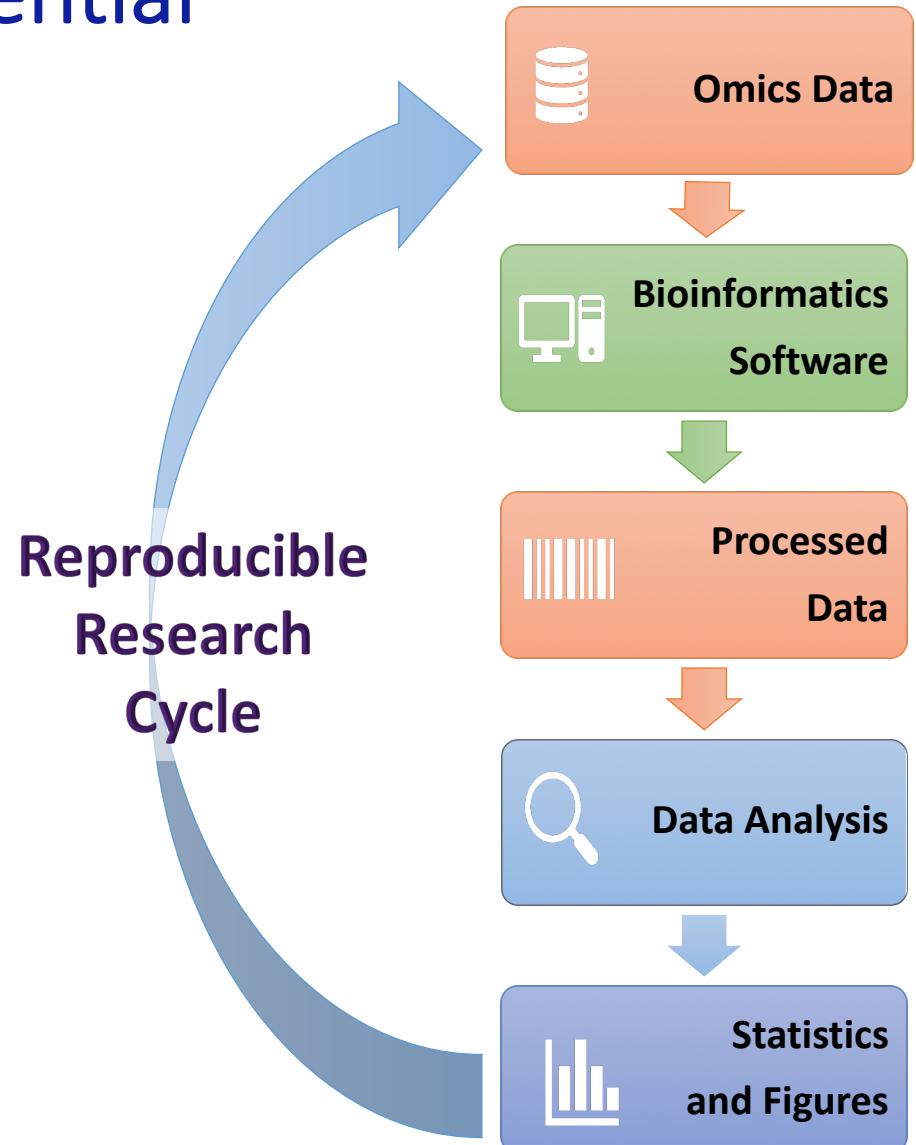


Metadata is also essential

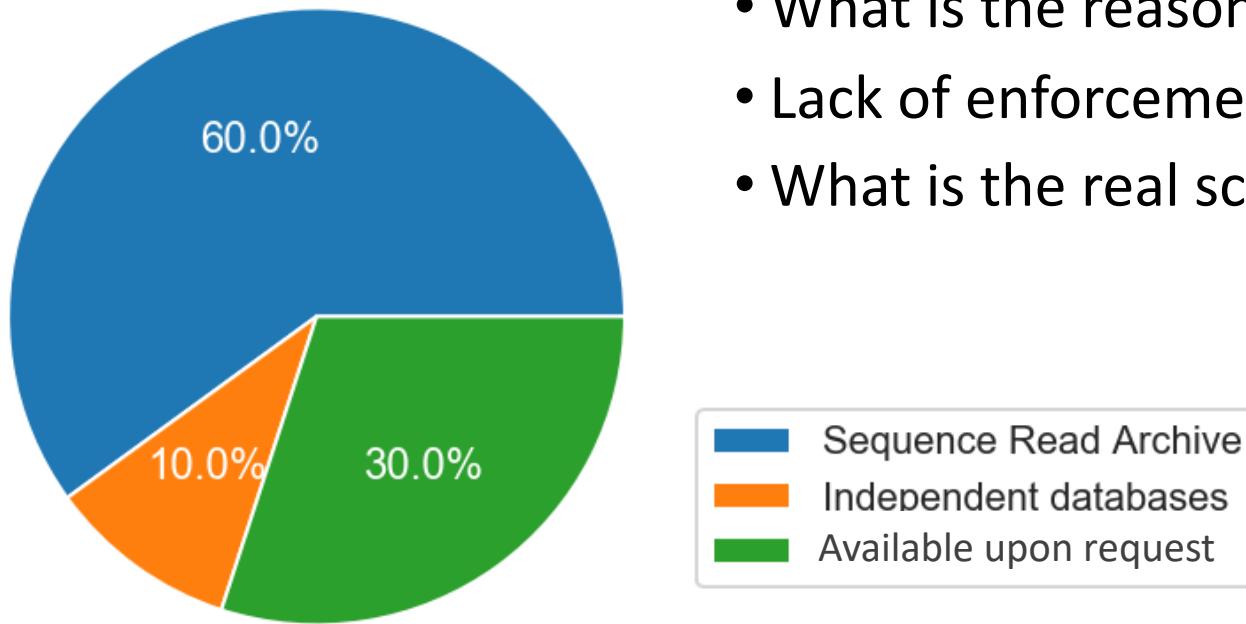
Omics raw data is mostly shared, but metadata is usually not properly formatted and shared

- Often incomplete
- Lack of agreement on the standards

Prohibits or makes it extremely challenging to search for specific phenotypes



Many researchers still don't share data



- What is the reason for not sharing?
- Lack of enforcement from editors?
- What is the real scale of the problem?

Data sharing of clinical metagenomics
studies between 2015-2019

The “Upon request model” is not sustainable

An empirical analysis of journal policy effectiveness for computational reproducibility

 Victoria Stodden, Jennifer Seiler, and Zhaokun Ma

PNAS March 13, 2018 115 (11) 2584-2589; first published March 12, 2018 <https://doi.org/10.1073/pnas.1708290115>

Edited by David B. Allison, Indiana University Bloomington, Bloomington, IN, and accepted by Editorial Board Member Susan T. Fiske January 9, 2018 (received for review July 11, 2017)

Authors response to data access requests

When you approach a PI for the source codes and raw data, you better explain who you are, whom you work for, why you need the data and what you are going to do with it.

I have to say that this is a very unusual request without any explanation! Please ask your supervisor to send me an email with a detailed, and I mean detailed, explanation.

Shared does not mean discoverable

Archival: via random pages that may never be found

Discovery: via centralized repositories with complete metadata allowing discovery via search tools

Meningitis and Epididymitis caused by Toscana Virus Infection Imported to Switzerland Diagnosed by Metagenomic Sequencing

Tschumi Fabian; Schmutz Stefan; Kufner Verena; Heider Maike; Pigny Fiona; Schreiner Bettina; Capaul Riccarda; Achermann Yvonne; Huber Michael

Raw Illumina MiSeq sequencing read in zipped FASTQ format for the RNA and DNA workflow of liquor sample 1000414117.

DNA: 1000414117-LI-DNA_S5_viral_reads.fastq.gz

RNA: 1000414117-LI-RNA_S10_viral_reads.fastq.gz

Files (27.8 kB)		
Name	Size	
1000414117-LI-DNA_S5_viral_reads.fastq.gz	9.9 kB	Download
md5:625524c6a2a1d9299170e106660e6ed5 ?		
1000414117-LI-RNA_S10_viral_reads.fastq.gz	17.9 kB	Download
md5:0ee5d0c96ed3ec10879afb4e27b12a3a ?		

Archived

The screenshot shows the NCBI SRA search interface. The search bar contains 'SRA' and the query 'meningitis'. Below the search bar, there are buttons for 'Advanced' and 'Full' search modes, with 'Full' currently selected. The results page displays a single study entry: SRX7047115.

SRX7047115: Cryptococcus neoformans var. grubii serotype A (H99) gpp2delta strain
1 ILLUMINA (Illumina HiSeq 2500) run: 15.9M spots, 3.2G bases, 1Gb downloads

Design: Deletion of GPP2 causes several phenotypic features related to stress response in C. are hypovirulent in invertebrate animal model of infection

Submitted by: LNCC

Study: Cryptococcus neoformans var. grubii serotype A (H99) gpp2delta strain
[PRJNA578980](#) • [SRP226736](#) • All experiments • All runs
[show Abstract](#)

Sample: WT biological replicate 3
[SAMN13089055](#) • [SRS5565031](#) • All experiments • All runs

Discoverable

Sharing processed data (summary statistics)

Usually small files

Example: GitHub

The screenshot shows a GitHub repository page for 'Mangul-Lab-USC / benchmarking_error_correction'. The repository has 2 stars and 1 fork. The 'Code' tab is selected, showing a branch dropdown set to 'master'. The path 'benchmarking_error_correction / summary_data /' is shown. A commit by 'jaquejbrito' dated Jan 2 is listed, with the message 'cleaned notebooks and figures'. Below the commit, there is a list of nine CSV files, all of which were committed on Jan 2 and updated 2 months ago. The files are:

File	Description	Committed	Last Updated
D1_WGS_E.coli_summary.csv	cleaned notebooks and figures	Jan 2	2 months ago
D1_WGS_human_complexity_summary.csv	cleaned notebooks and figures	Jan 2	2 months ago
D1_WGS_human_cpu_memory.csv	cleaned notebooks and figures	Jan 2	2 months ago
D1_WGS_human_summary.csv	cleaned notebooks and figures	Jan 2	2 months ago
D2_TCRA_real.csv	cleaned notebooks and figures	Jan 2	2 months ago
D3_TCRA_simulated.csv	cleaned notebooks and figures	Jan 2	2 months ago
D5_HIV_diversity_summary.csv	cleaned notebooks and figures	Jan 2	2 months ago
D5_HIV_mixture_summary.csv	cleaned notebooks and figures	Jan 2	2 months ago

Sharing processed data (summary statistics)

Branch: master [▼](#) [benchmarking_error_correction / summary_data / D1_WGS_E.coli_summary.csv](#) [Find file](#) [Copy path](#)

 jaquejbrito cleaned notebooks and figures f1db470 on Jan 2

1 contributor

301 lines (301 sloc) | 62.9 KB [Raw](#) [Blame](#) [History](#)   

Search this file...

1	EC Filename	Wrapper Name	Kmer Size	Read - TP	Read - TN	Read - FN	Read - FN WRONG	Read - FP	Read - FP IN
2	0 bfc_wgsim_rl_100_cov_1_20.corrected.fastq	bfc	20	265	28574	18139	3	209	0
3	1 bfc_wgsim_rl_100_cov_1_22.corrected.fastq	bfc	22	249	28596	18180	3	162	0
4	2 bfc_wgsim_rl_100_cov_1_24.corrected.fastq	bfc	24	236	28605	18186	3	160	0
5	3 bfc_wgsim_rl_100_cov_1_26.corrected.fastq	bfc	26	245	28630	18192	2	121	0
6	4 bfc_wgsim_rl_100_cov_1_28.corrected.fastq	bfc	28	223	28631	18219	2	115	0
7	5 bfc_wgsim_rl_100_cov_1_30.corrected.fastq	bfc	30	225	28634	18220	1	110	0
8	6 bfc_wgsim_rl_100_cov_16_20.corrected.fastq	bfc	20	265904	453943	27986	123	7086	0
9	7 bfc_wgsim_rl_100_cov_16_22.corrected.fastq	bfc	22	262422	453846	31428	113	7233	0
10	8 bfc_wgsim_rl_100_cov_16_24.corrected.fastq	bfc	24	258535	453754	35259	106	7388	0

Sharing raw omics data

Sharing real data

- SRA (Short Sequence Archive)
- GEO (Gene Expression Omnibus)

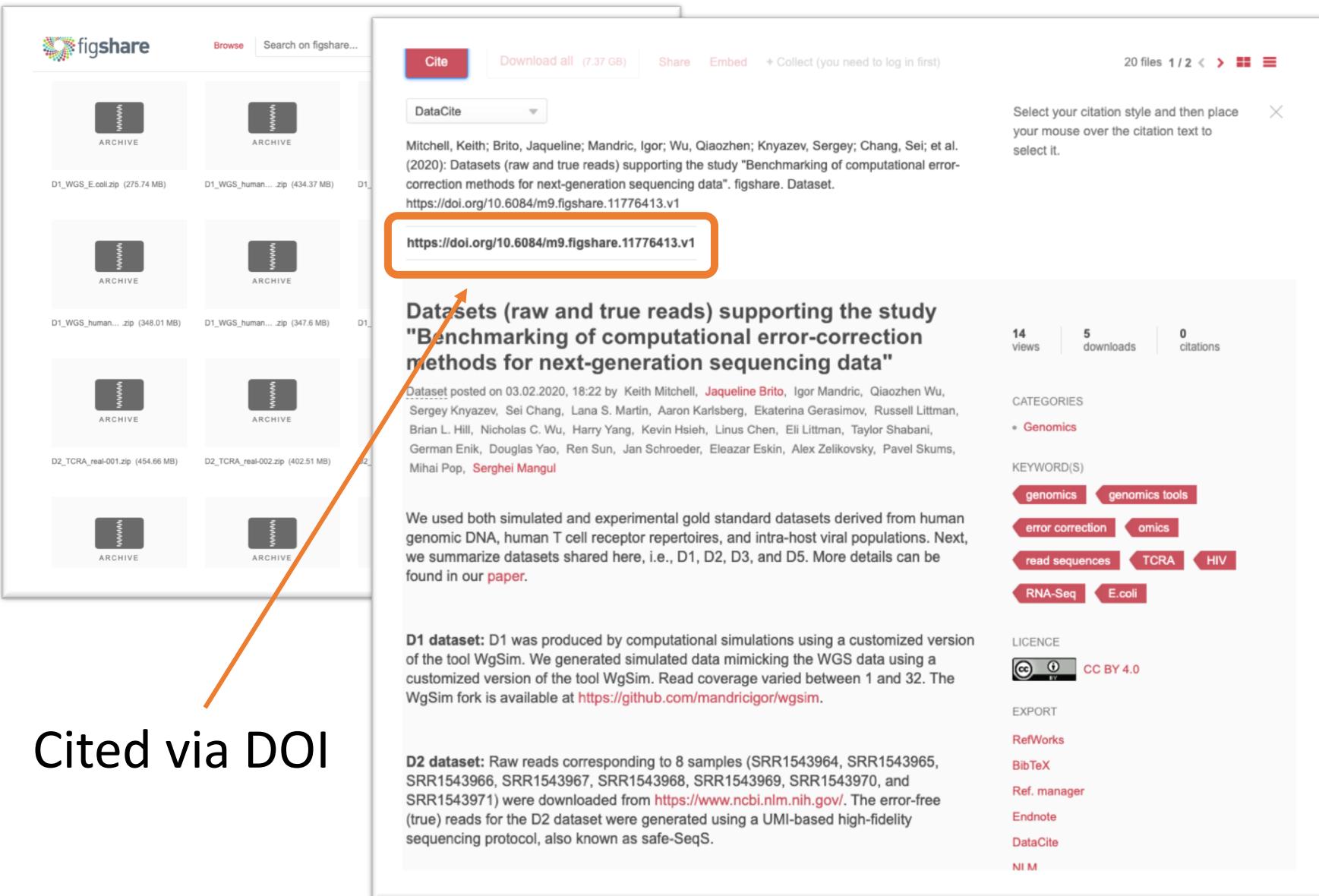
How do we share simulated/manipulated data?

- Platforms such as GitHub do not accept large files

Examples of platforms to store large data files

- Zenodo
- Figshare

Sharing raw omics data



The screenshot shows a figshare dataset page. At the top, there's a navigation bar with 'Browse' and a search bar. Below the navigation, there are two sections of datasets labeled D1 and D2. A red arrow points from the text 'Cited via DOI' at the bottom left towards the DOI link in the dataset details section.

Cite Download all (7.37 GB) Share Embed + Collect (you need to log in first) 20 files 1 / 2 < >

Select your citation style and then place your mouse over the citation text to select it.

DataCite

Mitchell, Keith; Brito, Jaqueline; Mandric, Igor; Wu, Qiaozhen; Knyazev, Sergey; Chang, Sei; et al. (2020): Datasets (raw and true reads) supporting the study "Benchmarking of computational error-correction methods for next-generation sequencing data". figshare. Dataset. <https://doi.org/10.6084/m9.figshare.11776413.v1>

Datasets (raw and true reads) supporting the study "Benchmarking of computational error-correction methods for next-generation sequencing data"

Dataset posted on 03.02.2020, 18:22 by Keith Mitchell, **Jaqueline Brito**, Igor Mandric, Qiaozhen Wu, Sergey Knyazev, Sei Chang, Lana S. Martin, Aaron Karlsberg, Ekaterina Gerasimov, Russell Littman, Brian L. Hill, Nicholas C. Wu, Harry Yang, Kevin Hsieh, Linus Chen, Eli Littman, Taylor Shabani, German Enik, Douglas Yao, Ren Sun, Jan Schroeder, Eleazar Eskin, Alex Zelikovsky, Pavel Skums, Mihai Pop, **Serghei Mangul**

We used both simulated and experimental gold standard datasets derived from human genomic DNA, human T cell receptor repertoires, and intra-host viral populations. Next, we summarize datasets shared here, i.e., D1, D2, D3, and D5. More details can be found in our [paper](#).

D1 dataset: D1 was produced by computational simulations using a customized version of the tool WgSim. We generated simulated data mimicking the WGS data using a customized version of the tool WgSim. Read coverage varied between 1 and 32. The WgSim fork is available at <https://github.com/mandrigor/wgsm>.

D2 dataset: Raw reads corresponding to 8 samples (SRR1543964, SRR1543965, SRR1543966, SRR1543967, SRR1543968, SRR1543969, SRR1543970, and SRR1543971) were downloaded from <https://www.ncbi.nlm.nih.gov/>. The error-free (true) reads for the D2 dataset were generated using a UMI-based high-fidelity sequencing protocol, also known as safe-SeQs.

14 views | 5 downloads | 0 citations

CATEGORIES • Genomics

KEYWORD(S) genomics genomics tools error correction omics read sequences TCRA HIV RNA-Seq E.coli

LICENCE CC BY 4.0

EXPORT RefWorks BibTeX Ref. manager Endnote DataCite NLM

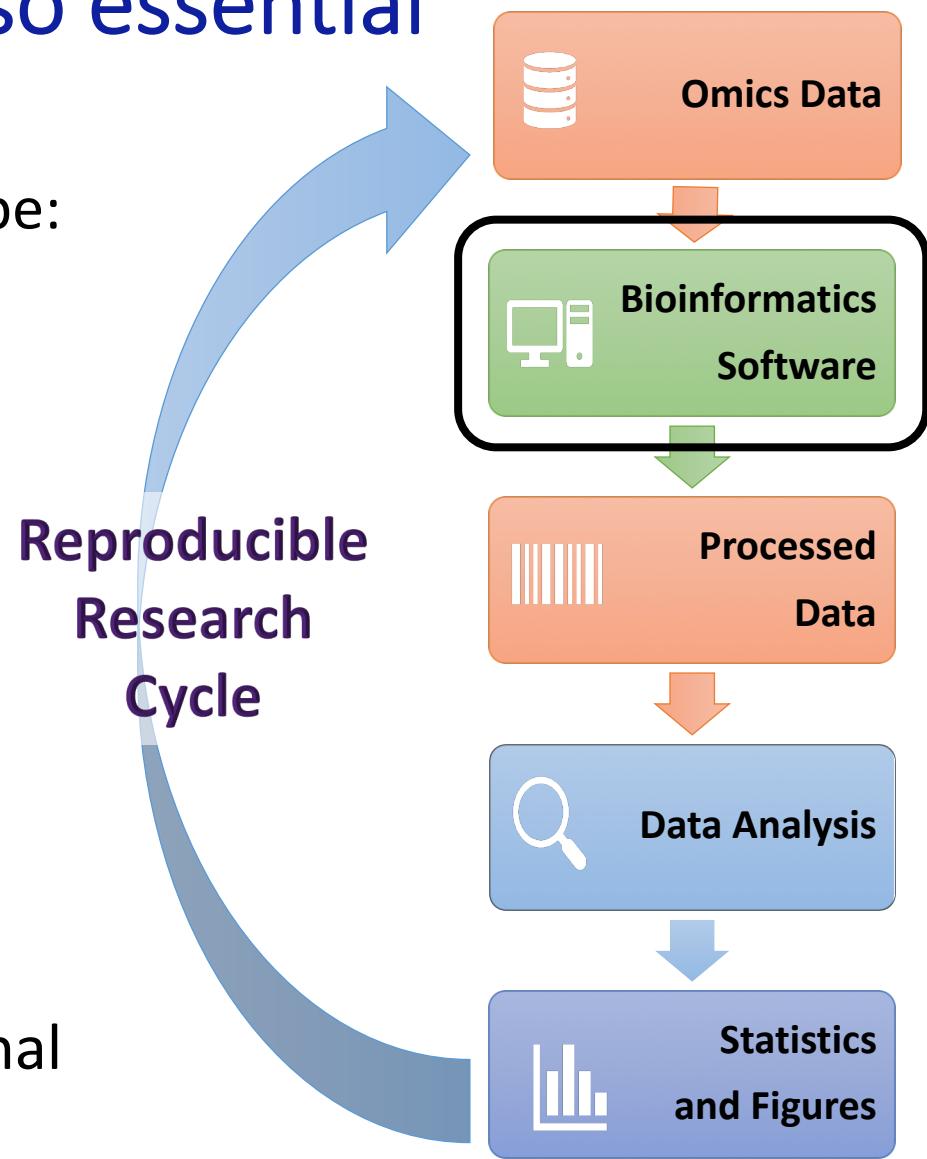
Cited via DOI

Software access is also essential

Ideally, software sharing should be:

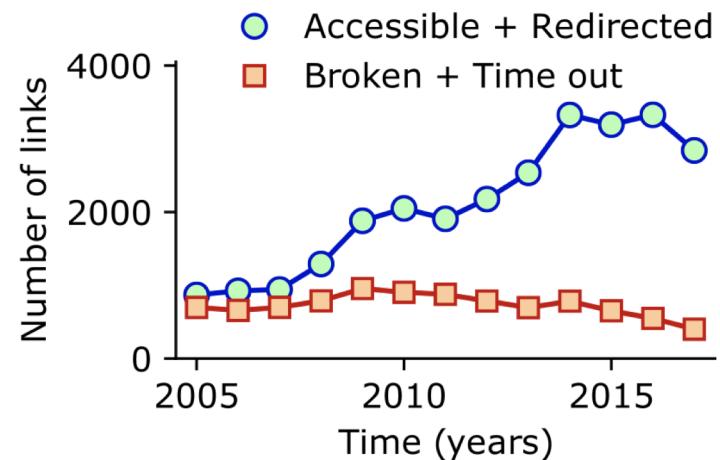
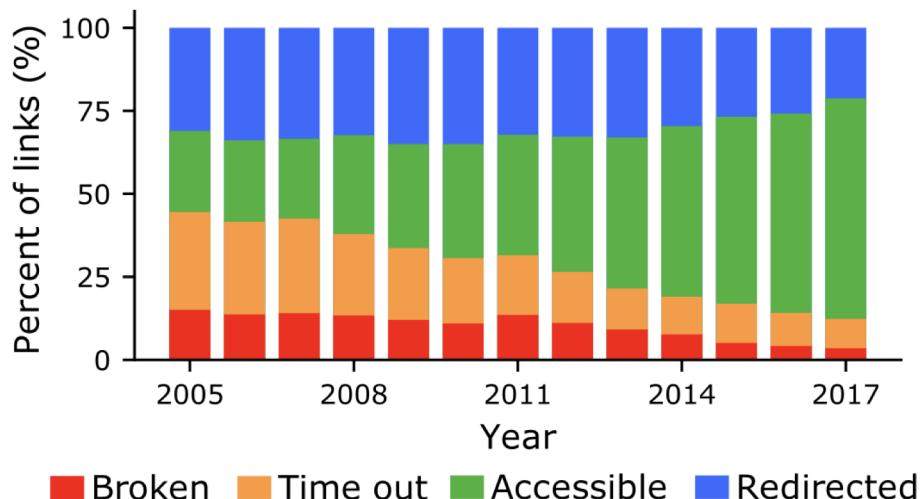
- Open-source
- Well documented
- Easy to install
- Maintained via centralized repositories
- Stored on archivally stable resources

Limited software usability and archival stability of computational tools leads to computational reproducibility crisis



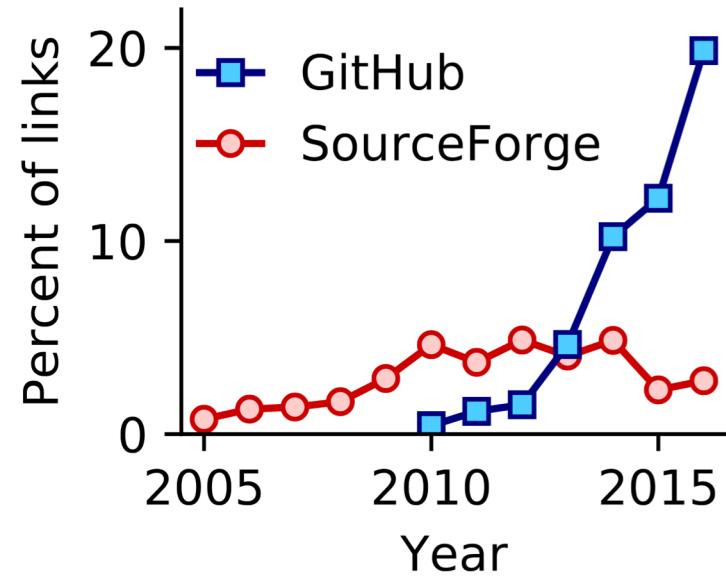
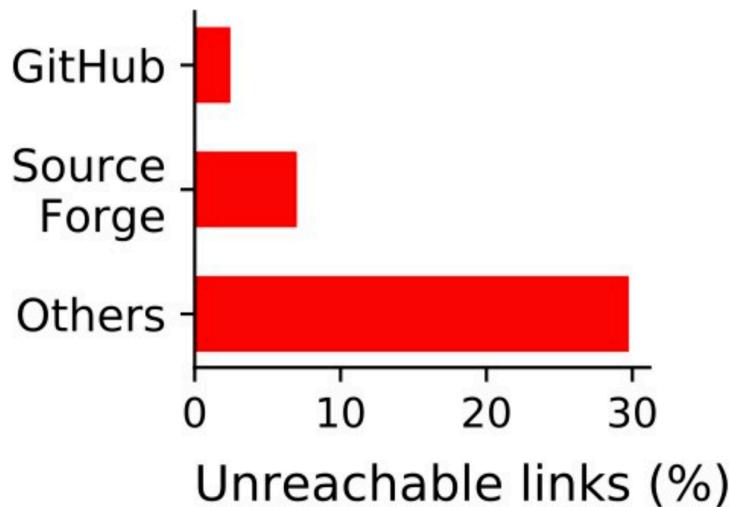
Limited archival stability

Archival stability of 36,702 published URLs across 10 systems and computational biology journals over the span of 13 years



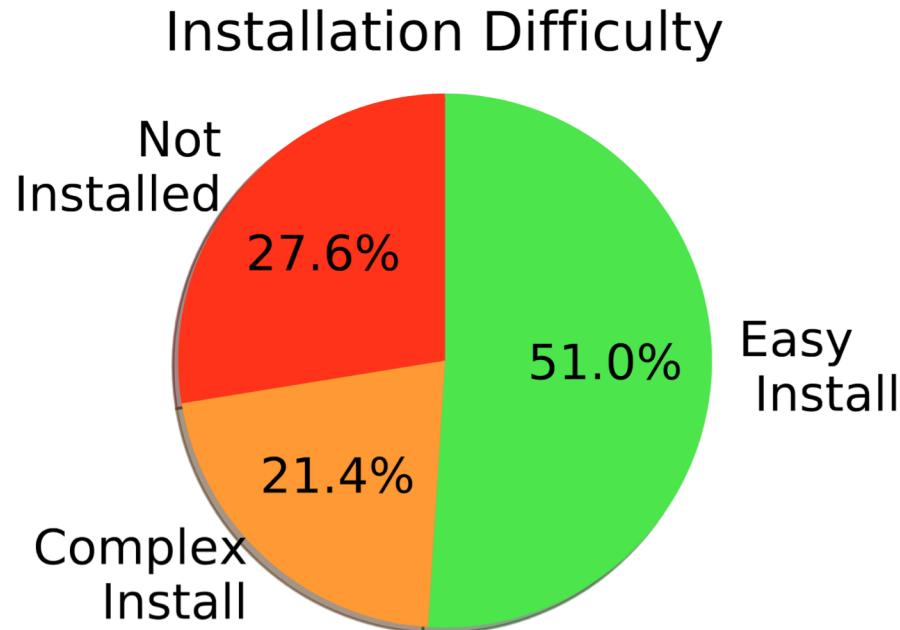
Limited archival stability

Vast majority of URLs hosted on GitHub or SourceForge are archivally stable



Limited Usability

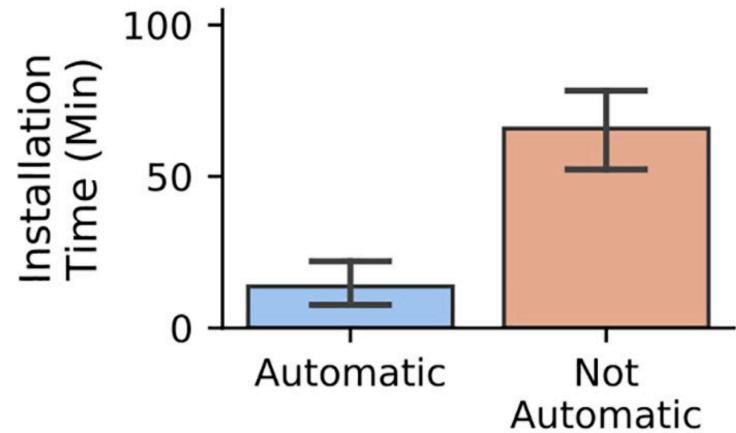
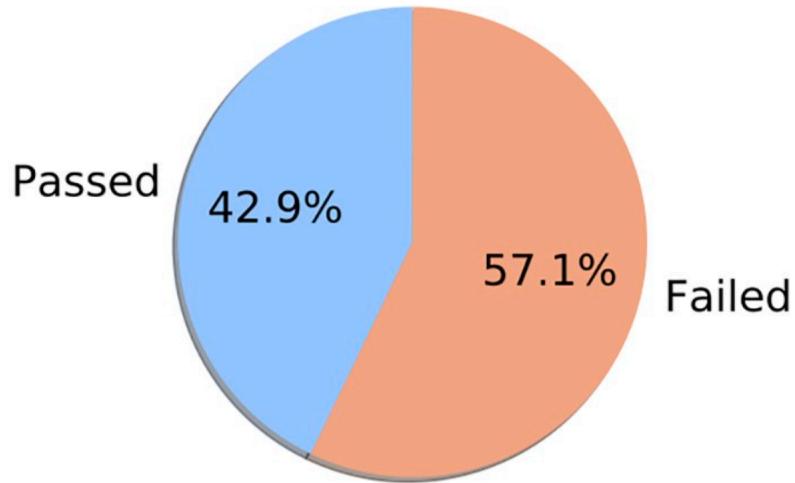
Many tools are hard or impossible to install



Source: Mangul, Serghei, et al. "Challenges and recommendations to improve the installability and archival stability of omics computational tools." *PLoS biology* 17.6 (2019): e3000333.

Limited Usability

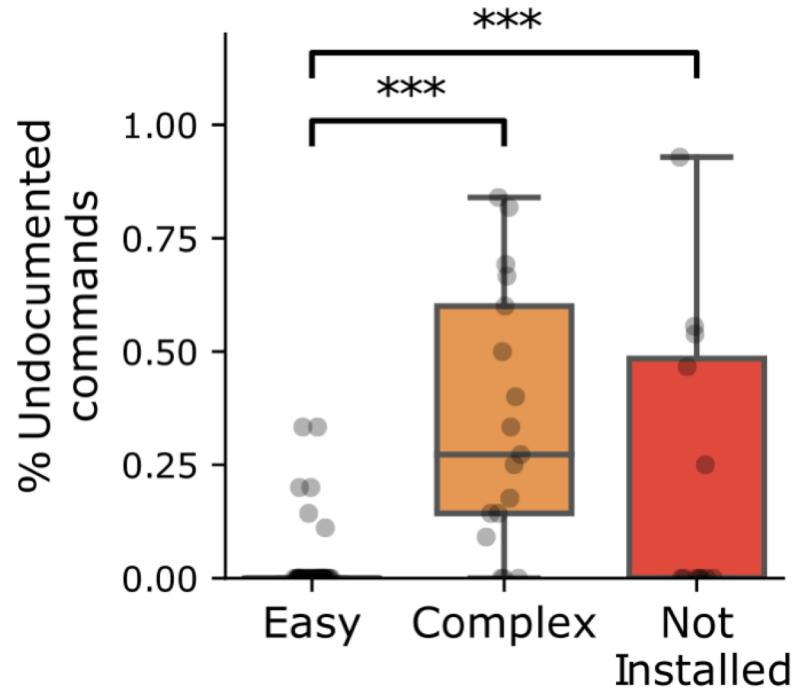
Automatic installation test



Source: Mangul, Serghei, et al. "Challenges and recommendations to improve the installability and archival stability of omics computational tools." *PLoS biology* 17.6 (2019): e3000333.

Limited Usability

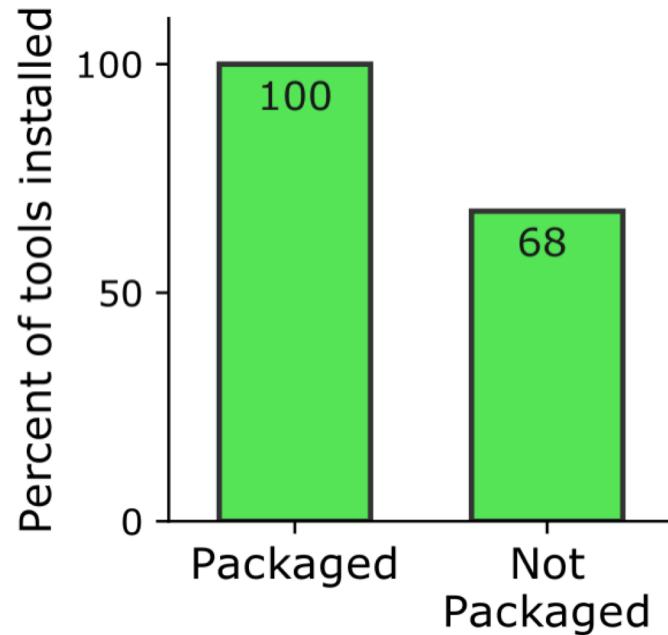
Tools easy to install have few undocumented commands



Source: Mangul, Serghei, et al. "Challenges and recommendations to improve the installability and archival stability of omics computational tools." *PLoS biology* 17.6 (2019): e3000333.

Limited Usability

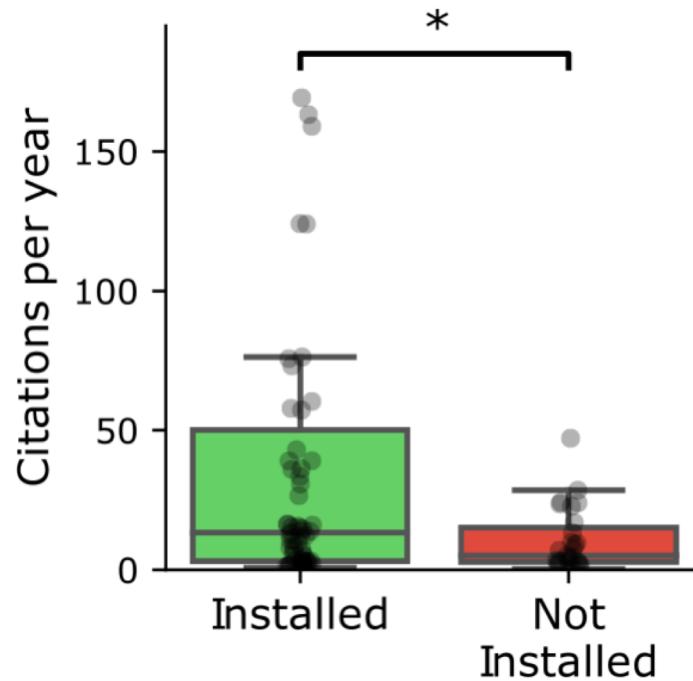
Bioconda tools were always installable



Source: Mangul, Serghei, et al. "Challenges and recommendations to improve the installability and archival stability of omics computational tools." *PLoS biology* 17.6 (2019): e3000333.

Limited Usability

Installable tools are more cited and probably more likely to be reused

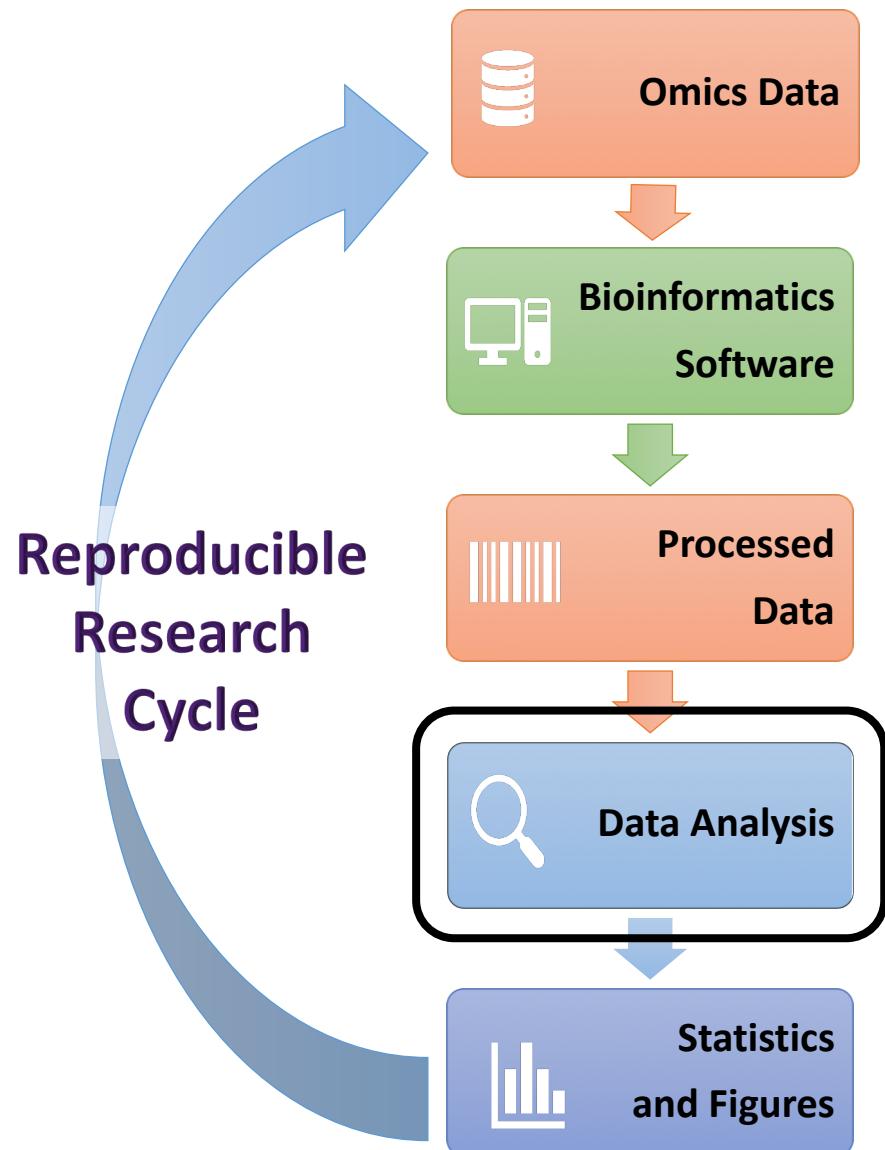


Source: Mangul, Serghei, et al. "Challenges and recommendations to improve the installability and archival stability of omics computational tools." *PLoS biology* 17.6 (2019): e3000333.

Data Analysis

To reproduce the results one need to use the same data analysis procedures from the original publication

Interactive notebooks (e.g., Jupyter) allows straightforward way to reproduce the results



Interactive notebooks allow us to go beyond reproducibility

- Easy change of parameters
- Check results' robustness
- Majority provide support to R and Python

Jupyter Notebooks

 [Mangul-Lab-USC / benchmarking_error_correction](#)

[Unwatch](#) 2 [Star](#) 1 [Fork](#) 1

[Code](#) [Issues 0](#) [Pull requests 0](#) [Actions](#) [Projects 0](#) [Wiki](#) [Security](#) [Insights](#)

This is the GitHub repository for our benchmarking study "Benchmarking of computational error-correction methods for next-generation sequencing".

 35 commits  1 branch  0 packages  0 releases  2 contributors

Branch: master ▾ [New pull request](#) [Create new file](#) [Upload files](#) [Find file](#) [Clone or download](#) ▾

Commit History		
 jaquejbrito	Update of study description	Latest commit df51057 on Jan 6
 figures	cleaned notebooks and figures	2 months ago
 figures_multi_panel	cleaned notebooks and figures	2 months ago
 notebooks	cleaned notebooks and figures	2 months ago
 summary_data	cleaned notebooks and figures	2 months ago
 .gitignore	cleaned notebooks and figures	2 months ago
 README.md	Update of study description	last month

 [README.md](#) 

Benchmarking of computational error-correction methods for next-generation sequencing data

[Preprint online](#) [licence](#) [MIT](#)

This project contains the links to the datasets and the code that was used for our study : "[Benchmarking of computational error-correction methods for next-generation sequencing data](#)"

Jupyter Notebooks

Mangul-Lab-USC / [benchmarking_error_correction](#)

Unwatch 2 Star 1 Fork 1

Code Issues 0 Pull requests 0 Actions Projects 0 Wiki Security Insights

Branch: master [benchmarking_error_correction / notebooks /](#)

Create new file Upload files Find file History

 jaquejbrito	cleaned notebooks and figures	Latest commit f1db470 on Jan 2
..		
D1_WGS_E.coli.ipynb	cleaned notebooks and figures	2 months ago
D1_WGS_human.ipynb	cleaned notebooks and figures	2 months ago
D2_TCRA_real.ipynb	cleaned notebooks and figures	2 months ago
D3_TCRA_simulated.ipynb	cleaned notebooks and figures	2 months ago
D5_HIV.ipynb	cleaned notebooks and figures	2 months ago

Jupyter Notebooks

422 lines (422 sloc) | 341 KB

Raw Blame History

Dataset D1 - WGS E.coli

Importing libraries

```
In [1]: from __future__ import division
import pylab as pl
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
import math
import seaborn as sns
import os
```

Data reading and cleaning

```
In [2]: data = pd.read_csv('../summary_data/D1_WGS_E.coli_summary.csv')
data["Trim Efficiency"] = data["Base - TP TRIM"]/(data["Base - TP TRIM"] + data["Base - FP TRIM"])
data['total_corrections'] = data['Base - TP']+ data['Base - FP']
```

Defining color dictionary

```
In [3]: color_dict=dict({'Bfc':'purple','Bless':'orange','Coral':'brown','Fiona':'gray','Lighter':'pink','Musket':'blue','Pollux':'yellow','Racer':'green','Reckoner':'red','Sga':'black'})
```

Selecting best kmer for each tool

```
In [4]: data_best = data.loc[data.groupby(["Tool","Coverage"])["Base Gain"].idxmax()]
data_best = data_best.fillna(0)
```

Jupyter Notebooks

Figure 2g

Heatmap depicting the gain across various coverage settings.

Each row corresponds to an error correction tool, and each column corresponds to a dataset with a given coverage.

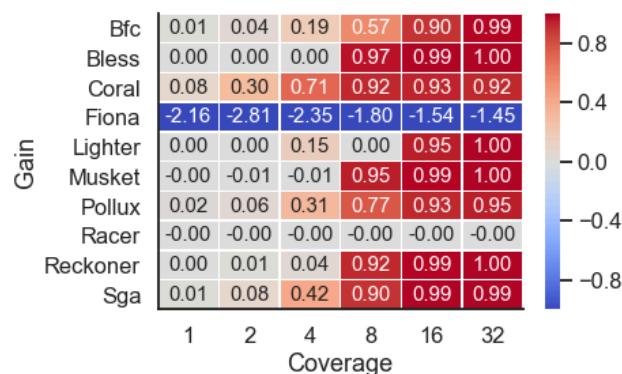
For each tool, the best k-mer size was selected.

```
In [5]: result = data_best.pivot(index='Tool', columns='Coverage', values='Base Gain') \
    .sort_values(by="Tool", ascending=False)

import matplotlib.pyplot as plt
import seaborn as sns
g=sns.set_style("white")
g=sns.set_context("talk")
g=sns.heatmap(result,
               annot=True,
               cmap='coolwarm',
               center=0,
               linewidths=.5,
               annot_kws={'size':15},
               fmt=".2f",
               vmin=-1,
               vmax=1)

g.set(xlabel='Coverage', ylabel='Gain')
g.set_yticks([0, 10])
g.set_yticklabels(['0', '10'])
g.set_xticks([1, 2, 4, 8, 16, 32])
g.set_xticklabels(['1', '2', '4', '8', '16', '32'])

plt.savefig("../figures/D1_WGS_E.coli/Fig2g_heatmap_ecoli_gain.png", bbox_inches="tight")
plt.savefig("../figures/D1_WGS_E.coli/Fig2g_heatmap_ecoli_gain.pdf", bbox_inches="tight")
```



What if we do reproducible research properly?

Enable secondary analysis

Potential for powerful discoveries in science in the dry lab settings

Accessible for labs in countries with limited resources

nature
biotechnology

Career Feature | Published: 04 March 2019

CAREER FEATURE

How bioinformatics and open data can boost basic science in countries and universities with limited resources

Serghei Mangul , Lana S. Martin, Ben Langmead, Javier E. Sanchez-Galan, Ian Toma, Fereydoun Hormozdiari, Pavel Pevzner & Eleazar Eskin

Nature Biotechnology 37, 324–326(2019) | [Cite this article](#)

2160 Accesses | 4 Citations | 107 Altmetric | [Metrics](#)

Providing training and access to standard computing hardware and cloud-based resources can enable scientists in lower-resource institutions and countries to reanalyze published ‘-omics’ data and produce career-enhancing STEM research.

Journal policies for ensuring and promoting reproducibility

Reproducibility can be assured upon publication

Various journals require sharing of materials

- Guidelines based on FAIR principles
- Badges to acknowledge reproducible studies



Journal policies for ensuring and promoting reproducibility



Revolutionizing data dissemination, organization, and use

Home

About ▾

Help ▾

Terms of use

GIGADB DATASETS

GigaDB contains 1862 discoverable, trackable, and citable datasets that have been assigned DOIs and are available for public download and use.

e.g. Chicken, brain etc...

Search

Dataset types

+ more



Genomic (1169)



Software (217)



Transcriptomic (169)



Imaging (1174)



Neuroscience (36)



Epigenomic (26)



Metagenomic (48)



Genome mapping (17)



Workflow (53)



Proteomic (21)



Metabarcoding (6)



Metadata (8)

RSS

New dataset added on 2020-02-13: [10.5524/100714](#)

Supporting data for "/*-DCC: A platform to collect, annotate and explore a large variety of sequencing experiments"

New dataset added on 2020-02-13: [10.5524/100715](#)

Supporting data for "PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S rRNA, ITS and COI marker genes"

New dataset added on 2020-02-09: [10.5524/100707](#)



Supporting data for "Telescope: an interactive tool for managing large scale analysis from mobile devices"

Dataset type: Software

Data released on December 19, 2019

Brutto JJ; Mosqueiro T; Rotman J; Xue V; Chapski DJ; la Hoz JD; Matias P; Martin LS; Zelikovsky A; Pellegrini M; Mangul S (2019):

Supporting data for "Telescope: an interactive tool for managing large scale analysis from mobile devices" GigaScience Database.

<http://dx.doi.org/10.5524/100686>

DOI 10.5524/100686

In today's world of big data, computational analysis has become a key driver of biomedical research. High-performance computational facilities are capable of processing considerable volumes of data, yet often lack an easy-to-use interface to guide the user in supervising and adjusting bioinformatics analysis via a tablet or smartphone. Telescope is a novel tool that interfaces with high-performance computational clusters to deliver an intuitive user interface for controlling and monitoring bioinformatics analyses in real-time. Telescope provides a user-friendly method for integrating computational analyses with experimental biomedical research.

Keywords:

[bioinformatics](#) [job scheduler](#) [high throughput computing](#) [bioinformatics analysis](#)

Additional details

Read the peer-reviewed publication(s):

Brutto, J. J., Mosqueiro, T., Rotman, J., Xue, V., Chapski, D. J., la Hoz, J. D., ... Mangul, S. (2020). Telescope: an interactive tool for managing large-scale analysis from mobile devices. *GigaScience*, 9(1). [doi:10.1093/gigascience/giz163](https://doi.org/10.1093/gigascience/giz163)

Additional information:

https://scicrunch.org/resolver/RRID:SCR_017626

<https://bio.tools/Telescope>

Github links:

<https://github.com/Mangul-Lab-USC/telescope>

<https://github.com/QCB-Collaboratory/telescope>

Files

Funding

History

FTP site

Table Settings

File Name	Sample ID	Data Type	File Format	Size	Release Date	
readme_100686.txt		readme.txt	TEXT	1.76 KB	2019-12-19	
telescope-master.zip		GitHub archive	archive	1.27 MB	2019-12-12	

Improve support and acknowledgement

Producing reproducible research requires a lot of effort from researchers

Reality of most research groups

- Lack of experience on software development
- No knowledge of better practices for long-term maintenance

Improve support and acknowledgement

There is a need of better mechanisms to support reproducibility

- Funding for building proper infrastructure
- Acknowledgement for producing reproducible research

Example: Research Parasite and Symbiotic Awards

- Recognizes rigorous secondary data analyses
- <https://researchparasite.com/>

Mechanisms for ensuring reproducibility

Variably enforced by academic institutions, funders and publishers

- Published software source code is unavailable
- Documentation is incomplete or unmaintained
- Analytical source code is missing

Who is responsible?

How can we measure reproducibility?



Mechanisms for sustaining reproducibility

Where do you store the data and code from the studies?

- GitHub? SourceForge?

Is it sustainable to rely on commercial solutions?

Are these solutions sustainable in the long run?

What are the alternatives solutions?

Beyond reproducibility

README.md 

Automated scholarly manuscripts on GitHub

[manuscript](#) [HTML](#) [manuscript](#) [PDF](#)

Manuscript description

This repository is a dynamic version of the paper "Recommendations to enhance rigor and reproducibility in biomedical research" with an extended list of references.

Manubot

Manubot is a system for writing scholarly manuscripts via GitHub. Manubot automates citations and references, versions manuscripts using git, and enables collaborative writing via GitHub. An [overview manuscript](#) presents the benefits of collaborative writing with Manubot and its unique features. The [rootstock repository](#) is a general purpose template for creating new Manubot instances, as detailed in [SETUP.md](#). See [USAGE.md](#) for documentation how to write a manuscript.

Please open [an issue](#) for questions related to Manubot usage, bug reports, or general inquiries.

Beyond reproducibility

The screenshot shows a GitHub repository interface. At the top, there's a header with the repository name "Mangul-Lab-USC / enhancing_reproducibility" and various stats: Unwatch (2), Star (1), Fork (0). Below the header are navigation links: Code, Issues (0), Pull requests (0), Actions, Projects (0), Wiki, Security, Insights, and Settings. The main content area shows the file "01.abstract.md" with the title "enhancing_reproducibility / content / 01.abstract.md". There are buttons for "Cancel" and "Edit file". To the right of the file content are settings for "Spaces" (set to 2), "Soft wrap", and "Preview changes". The file content itself is a text document with numbered lines from 1 to 21, discussing computational biology, reproducibility, and research practices.

```
1 ## Abstract {.page_break_before}
2 Computational methods have reshaped the landscape of modern biology.
3 While the biomedical community is increasingly dependent on
4 computational tools, the mechanisms ensuring open data, open software,
5 and reproducibility are variably enforced by academic institutions,
6 funders and publishers. Publications may describe the software for which
7 source code is unavailable, documentation is incomplete or unmaintained,
8 and analytical source code is missing. Publications that lack this
9 information compromise the role of peer review in evaluating technical
10 strength and scientific contribution. Such flaws also limit any
11 subsequent work that intends to use the described software. We herein
12 provide recommendations to improve reproducibility, transparency, and
13 rigor in computational biology—precisely the values which should be
14 emphasized in foundational life and medical science curricula. Our
15 recommendations for improving software availability, usability, and
16 archival stability aim to foster a sustainable data science ecosystem in
17 biomedicine and life science research.
18
19 **Keywords**: Rigor, reproducible research, installability, archival
20 stability, big data, open science
21
```

Beyond reproducibility

Recommendations to enhance rigor and reproducibility in biomedical research

This manuscript was automatically generated on February 23, 2020.

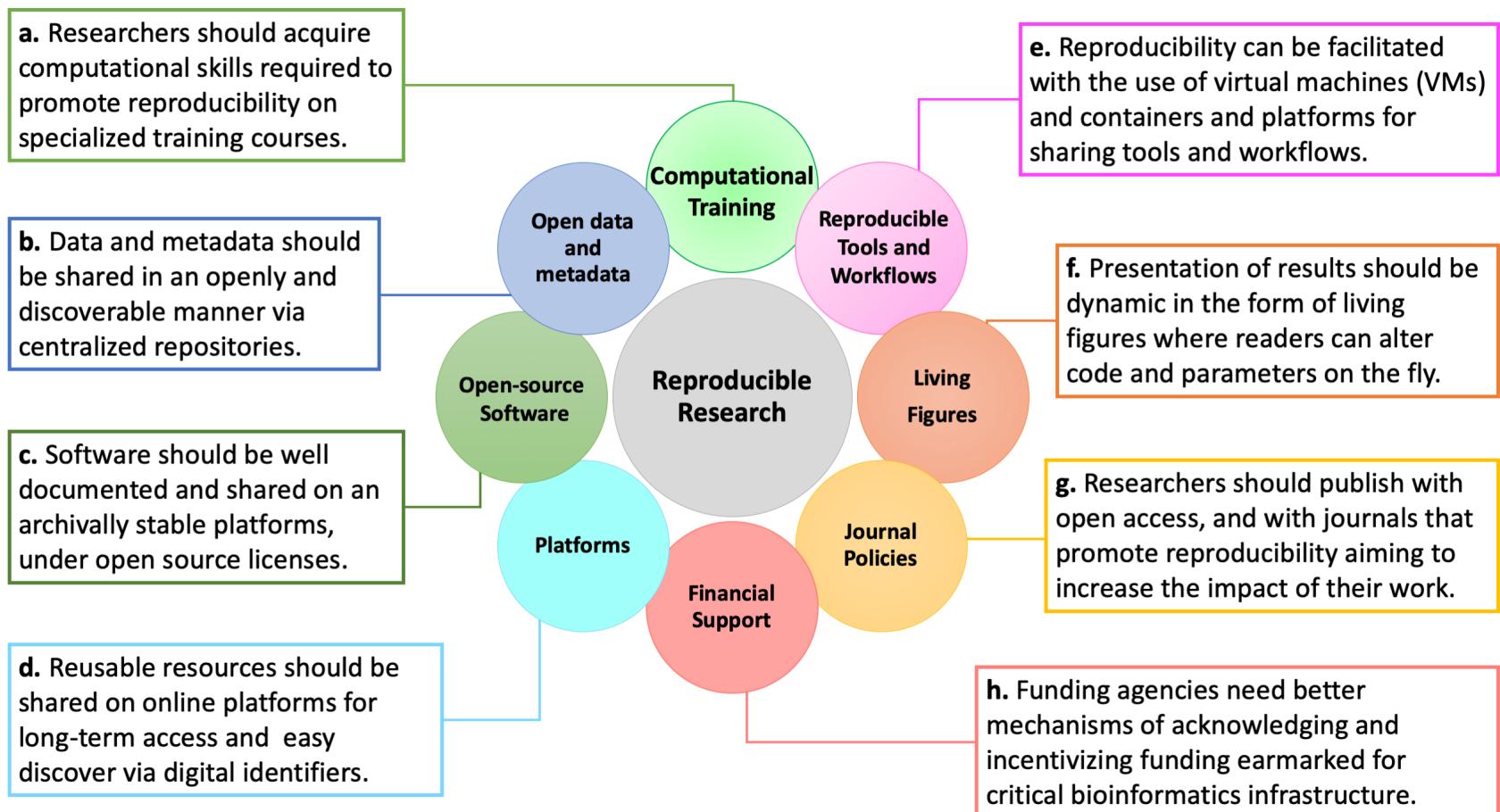
Authors

▼ Abstract

Computational methods have reshaped the landscape of modern biology. While the biomedical community is increasingly dependent on computational tools, the mechanisms ensuring open data, open software, and reproducibility are variably enforced by academic institutions, funders and publishers. Publications may describe the software for which source code is unavailable, documentation is incomplete or unmaintained, and analytical source code is missing. Publications that lack this information compromise the role of peer review in evaluating technical strength and scientific contribution. Such flaws also limit any subsequent work that intends to use the described software. We herein provide recommendations to improve reproducibility, transparency, and rigor in computational biology—precisely the values which should be emphasized in foundational life

Our recommendations

Main components we believe to be necessary for enhancing reproducibility



Conclusions

Our recommendations aim to improve the rigor of biomedical studies and foster reproducibility in computational biology

The infrastructure required to systematically adopt best practices for reproducibility of biomedical research is largely in place, but incentives are not currently aligned to support good practices

Current efforts rely on effort of individual researchers

Successful systematic adoption of best practices will require the buy-in of multiple stakeholders from publishers, academic institutions, funding agencies

Community-wide adoption of best practices is essential to produce reproducible research

Acknowledgments

Jun Li - University of Michigan

Jason H. Moore - University of Pennsylvania

Casey S. Greene - University of Pennsylvania

Nicole A. Nogoy - GigaScience

Lana X. Garmire - University of Michigan

Serghei Mangul - University of Southern California