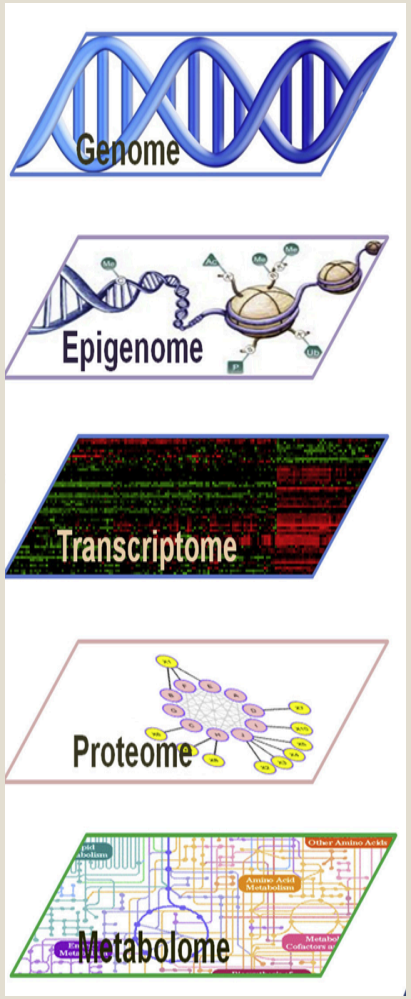
ANALYSIS OF THE USABILITY AND ARCHIVAL STABILITY OF OMICS COMPUTATIONAL TOOLS AND RESOURCES

Serghei Mangul
University of California, Los Angeles

@serghei_mangul

Flood of omics software tools



S. Mangul, Interpreting and integrating big data in the life sciences, preprint -- https://doi.org/10.7287/peerj.preprints.27603v1

Many bioinformatics tools are hard to install

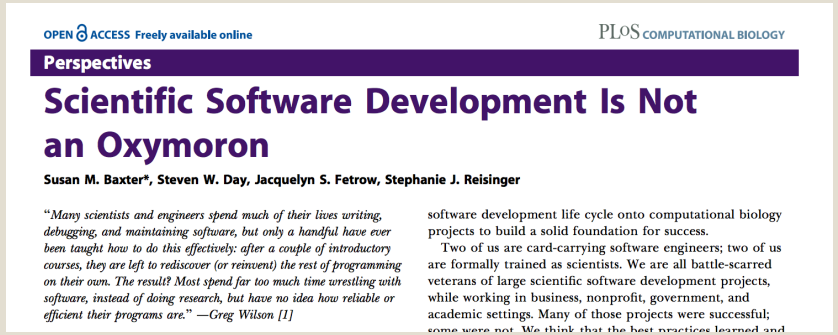
You Retweeted

Emily Madden @_emadden · Oct 27

The tweet I come across as I'm taking a break from trying to properly install and run a bioninformatics tool. #perfecttiming

> **Ran Blekhman** @blekhman
> Say you pick 100 random bioinformatics software tools -- how many will you actually be able to access, install, and run?
>
> Our new paper: biorxiv.org/content/early/… twitter.com/serghei_mangul…
>
> Show this thread

Is user-friendly scientific software an oxymoron?



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Perspectives

Scientific Software Development Is Not an Oxymoron

Susan M. Baxter*, Steven W. Day, Jacquelyn S. Fetrow, Stephanie J. Reisinger

"*Many scientists and engineers spend much of their lives writing, debugging, and maintaining software, but only a handful have ever been taught how to do this effectively: after a couple of introductory courses, they are left to rediscover (or reinvent) the rest of programming on their own. The result? Most spend far too much time wrestling with software, instead of doing research, but have no idea how reliable or efficient their programs are.*" —Greg Wilson [1]

software development life cycle onto computational biology projects to build a solid foundation for success.

Two of us are card-carrying software engineers; two of us are formally trained as scientists. We are all battle-scarred veterans of large scientific software development projects, while working in business, nonprofit, government, and academic settings. Many of those projects were successful;

PLOS COMPUTATIONAL BIOLOGY

BROWSE PUBLISH

OPEN ACCESS

EDITORIAL

Ten Simple Rules for Developing Usable Software in Computational Biology

Markus List, Peter Ebert, Felipe Albrecht

OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for the Open Development of Scientific Software

Andreas Prlić[1]*, James B. Procter[2]

1 San Diego Supercomputer Center, University of California San Diego, La Jolla, California, United States of America, 2 School of Life Sciences Research, College of Life Sciences, University of Dundee, Dundee, Scotland, United Kingdom

OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Perspective

Current Practice in Software Development for Computational Neuroscience and How to Improve It

Marc-Oliver Gewaltig[1]*, Robert Cannon[2]

1 Blue Brain Project, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2 Textensor Ltd., Edinburgh, United Kingdom

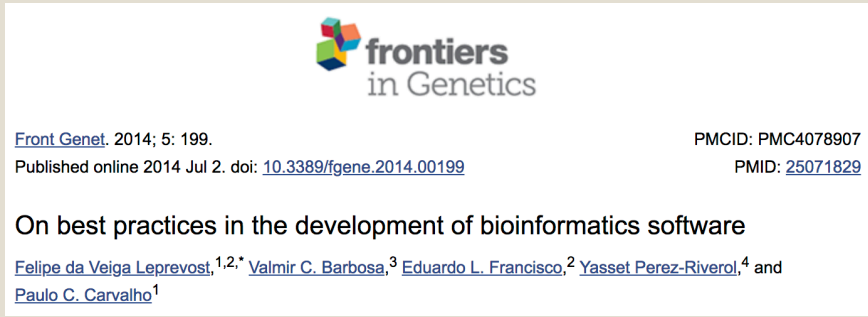
nature biotechnology

Feature | Published: 08 October 2013

The anatomy of successful computational biology software

Stephen Altschul, Barry Demchak, Richard Durbin, Robert Gentleman, Martin Krzywinski, Heng Li, Anton Nekrutenko, James Robinson, Wayne Rasband, James Taylor & Cole Trapnell

Nature Biotechnology **31**, 894–897 (2013) | Download Citation



frontiers in Genetics

Front Genet. 2014; 5: 199.

Published online 2014 Jul 2. doi: 10.3389/fgene.2014.00199

PMCID: PMC4078907

PMID: 25071829

On best practices in the development of bioinformatics software

Felipe da Veiga Leprevost,[1,2,*] Valmir C. Barbosa,[3] Eduardo L. Francisco,[2] Yasset Perez-Riverol,[4] and Paulo C. Carvalho[1]

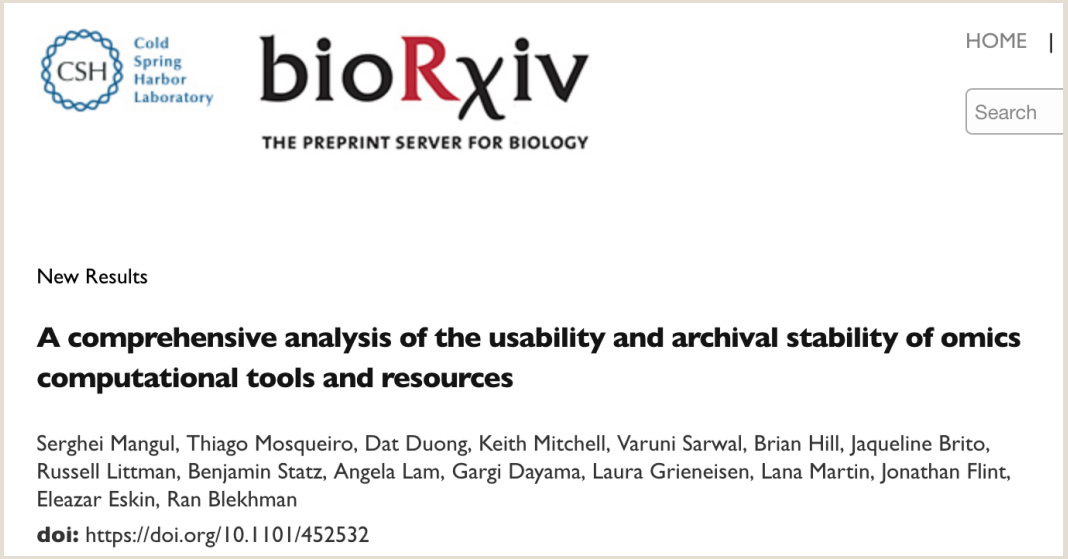
Challenges to effective software development and distribution in academia

- Software written by researchers tends to be written with the idea that all users will know as much about the code as its original authors
- Incentives in academia heavily favor the publication of new software, not the maintenance of existing tools
- There is lack of protocols to check insatiability of published software tools in academia

Software development in academia versus industry

- Industry receive considerably more resources for developing user-friendly tools
- Companies efficiently distribute industry-produced software using dedicated company units or contractors
- There is little reward for continuous, long-term development and maintenance of tools in academia

Poorly implemented tools will ultimately hinder progress in big data-driven fields



Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME |

Search

New Results

A comprehensive analysis of the usability and archival stability of omics computational tools and resources

Serghei Mangul, Thiago Mosqueiro, Dat Duong, Keith Mitchell, Varuni Sarwal, Brian Hill, Jaqueline Brito, Russell Littman, Benjamin Statz, Angela Lam, Gargi Dayama, Laura Grieneisen, Lana Martin, Jonathan Flint, Eleazar Eskin, Ran Blekhman

doi: https://doi.org/10.1101/452532

Mangul *et al. Genome Biology* (2019) 20:47
https://doi.org/10.1186/s13059-019-1649-8

Genome Biology

EDITORIAL Open Access

Check for updates

Improving the usability and archival stability of bioinformatics software

Serghei Mangul[1,2*], Lana S. Martin[2], Eleazar Eskin[1,3] and Ran Blekhman[4,5]

Software crisis is a reproducibility crisis

- Limited software usability and archival stability of computational tools leads to (**computational**) **reproducibility crisis**
- **Computational reproducibility** is the ability to replicate published findings by running the same computational tool on the data generated by the published study

Data: archival stability of omics computational tools and resources

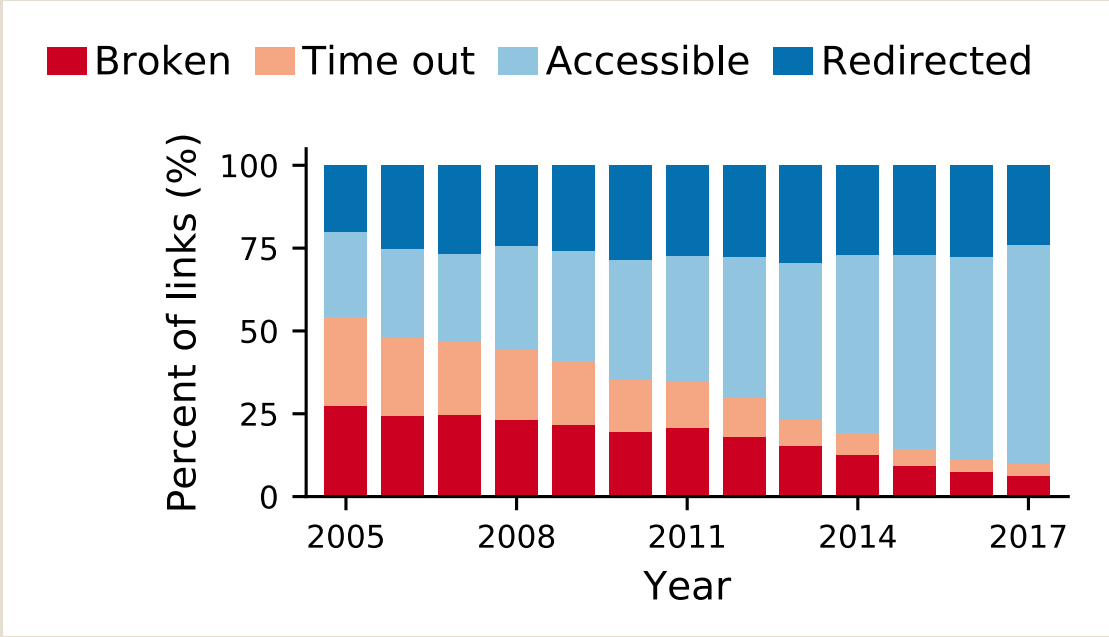
- 51,236 biomedical papers from PubMed
- 99 randomly selected bioinformatics tools

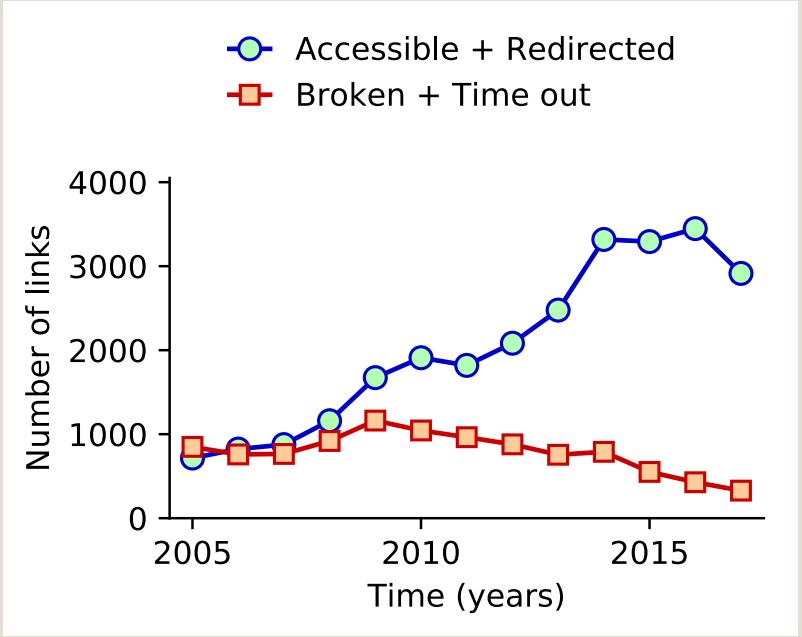
Journal name	Number of URLs
Nature Biotechnol	180
Genome Medicine	352
Nature Methods	403
Genome Biology	904
BMC Systems Biology	912
Bioinformatics	3131
PLoS Comp. Biology	3226
BMC Bioinformatics	6840
BMC Genomics	7651
Nucleic Acids Research	13103

Analysis of archival stability

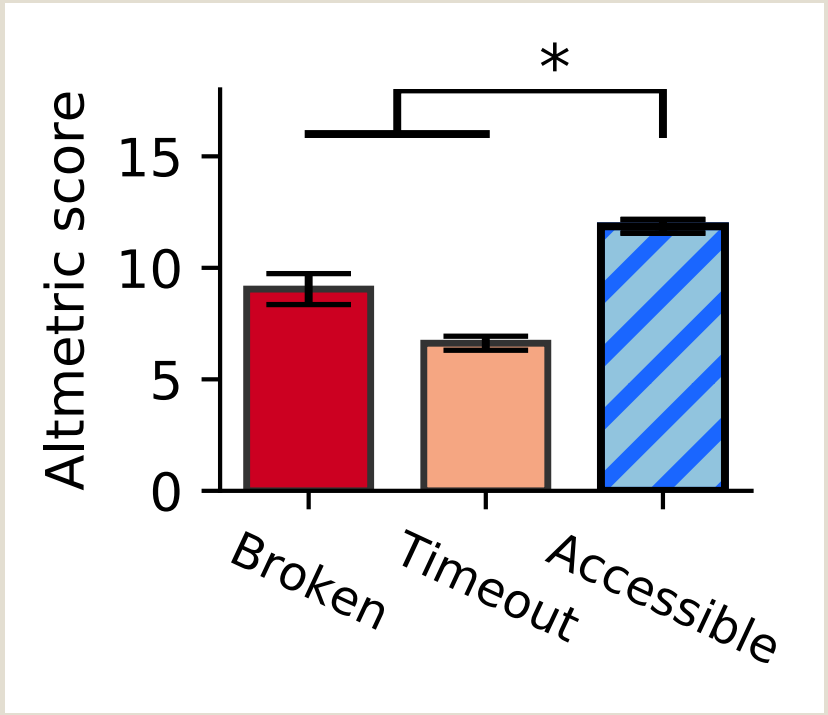
- We downloaded **51,236** open access papers via PubMed from 10 systems and computational biology journals (Raw data in XML format)
- We developed an approach to extract **36,702** software links from the downloaded papers and verify the archival stability of links
- Timeout links were manually verified

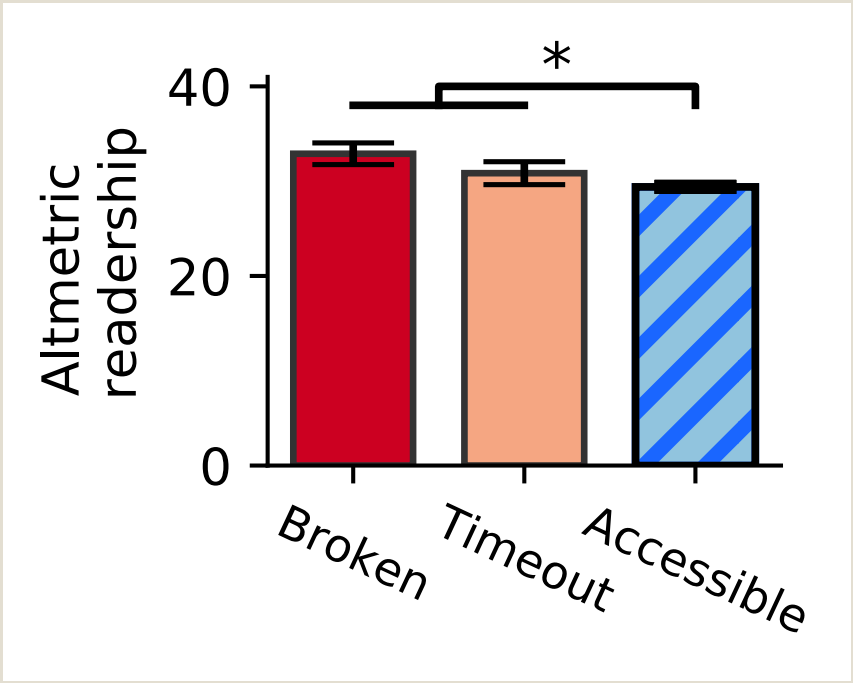
Archival stability of abstract URLs



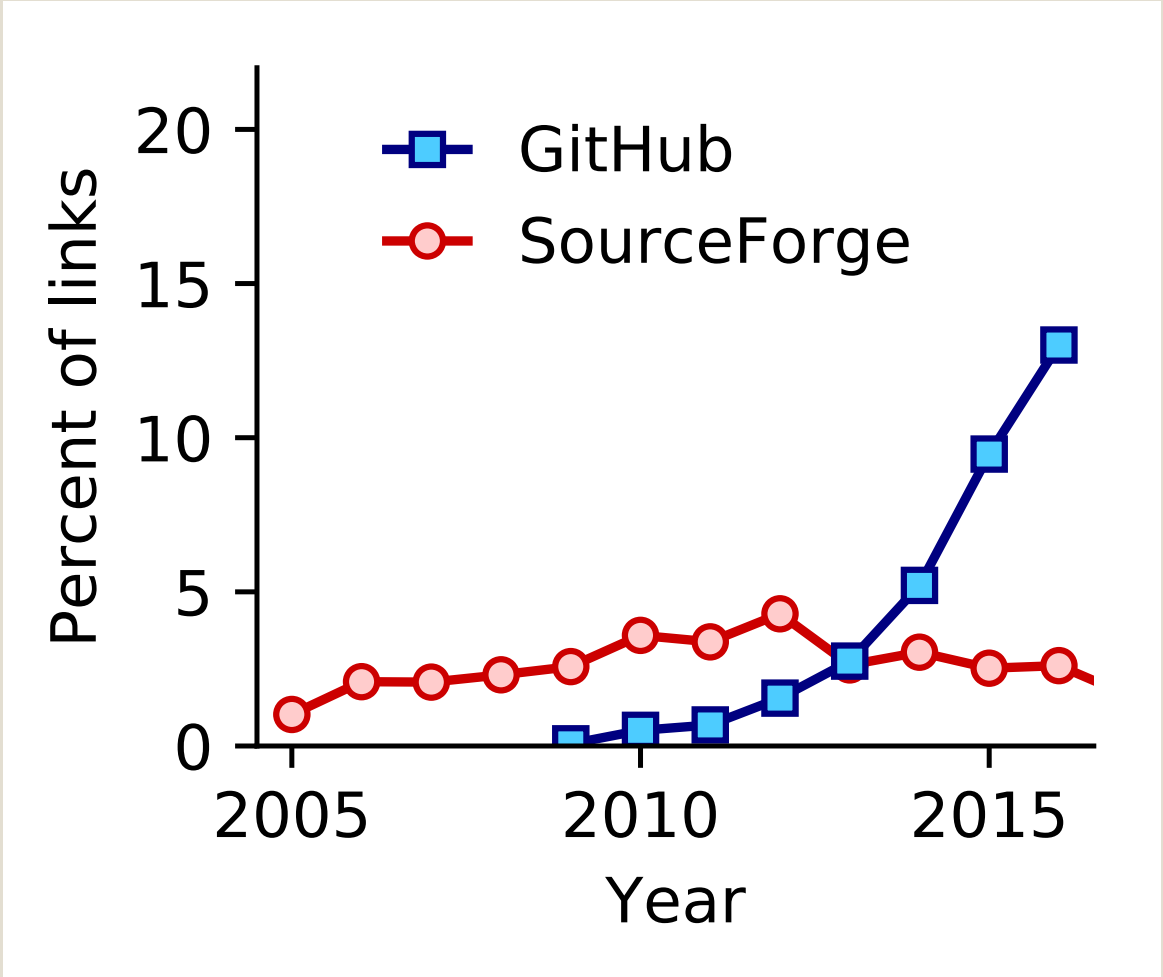


Effect of social media

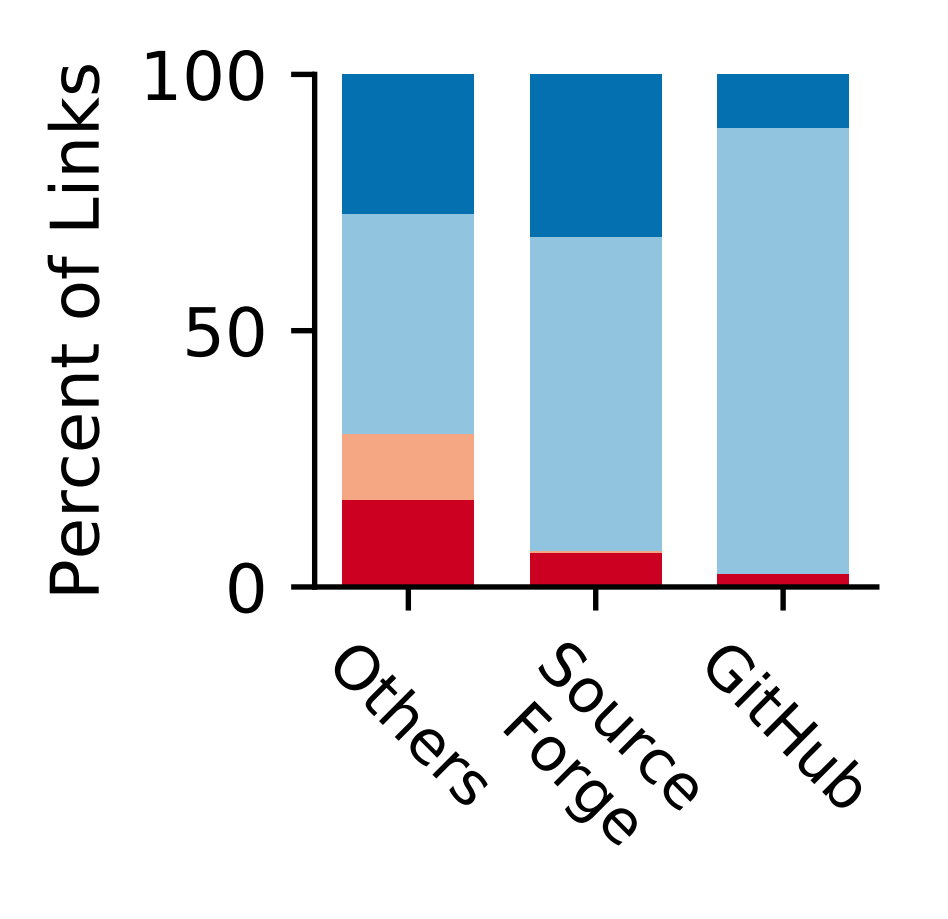




GitHub is the most popular platform to host scientific code



Tools hosted on GitHub have <3% of unreachable links

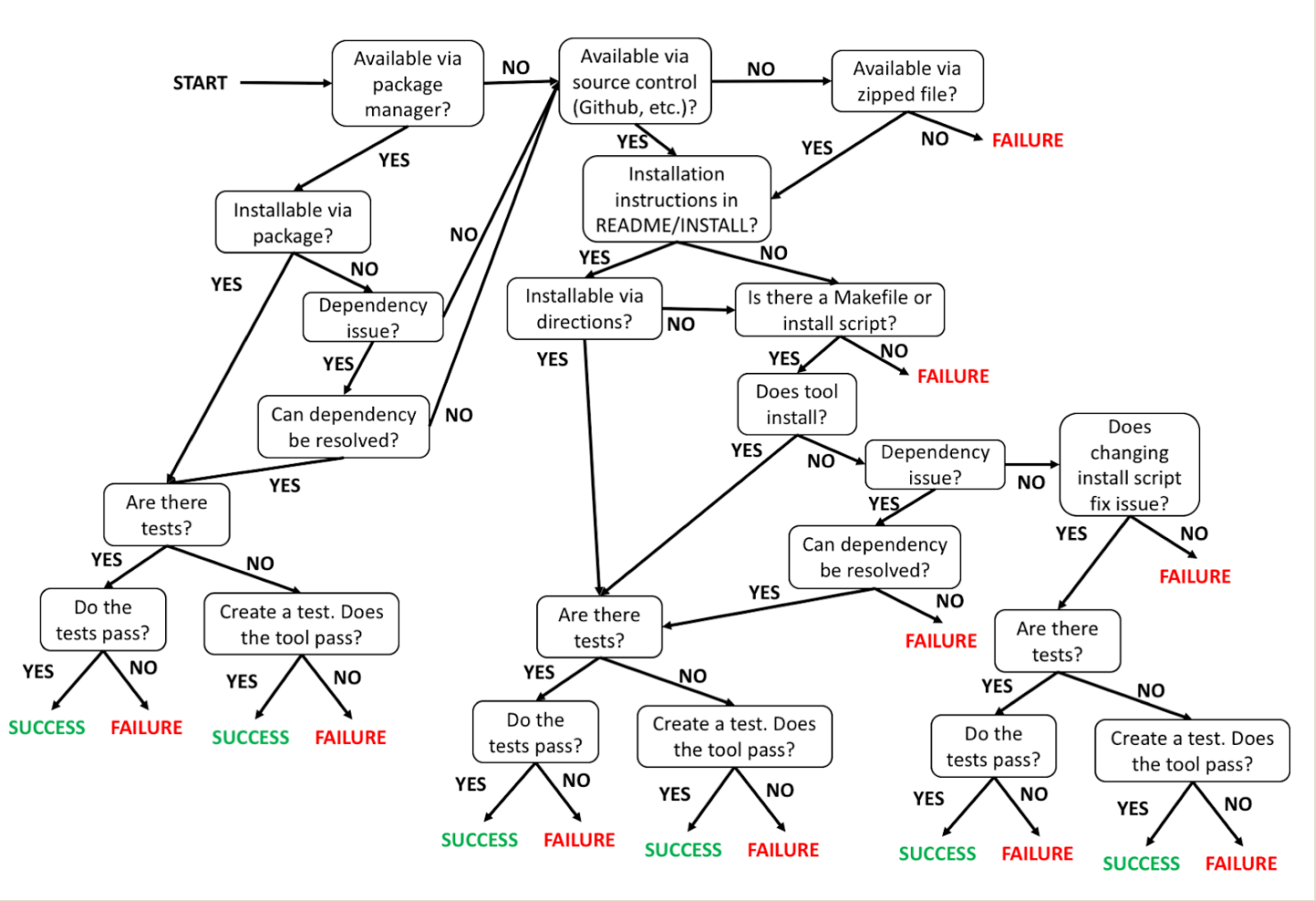


Analysis of the software usability

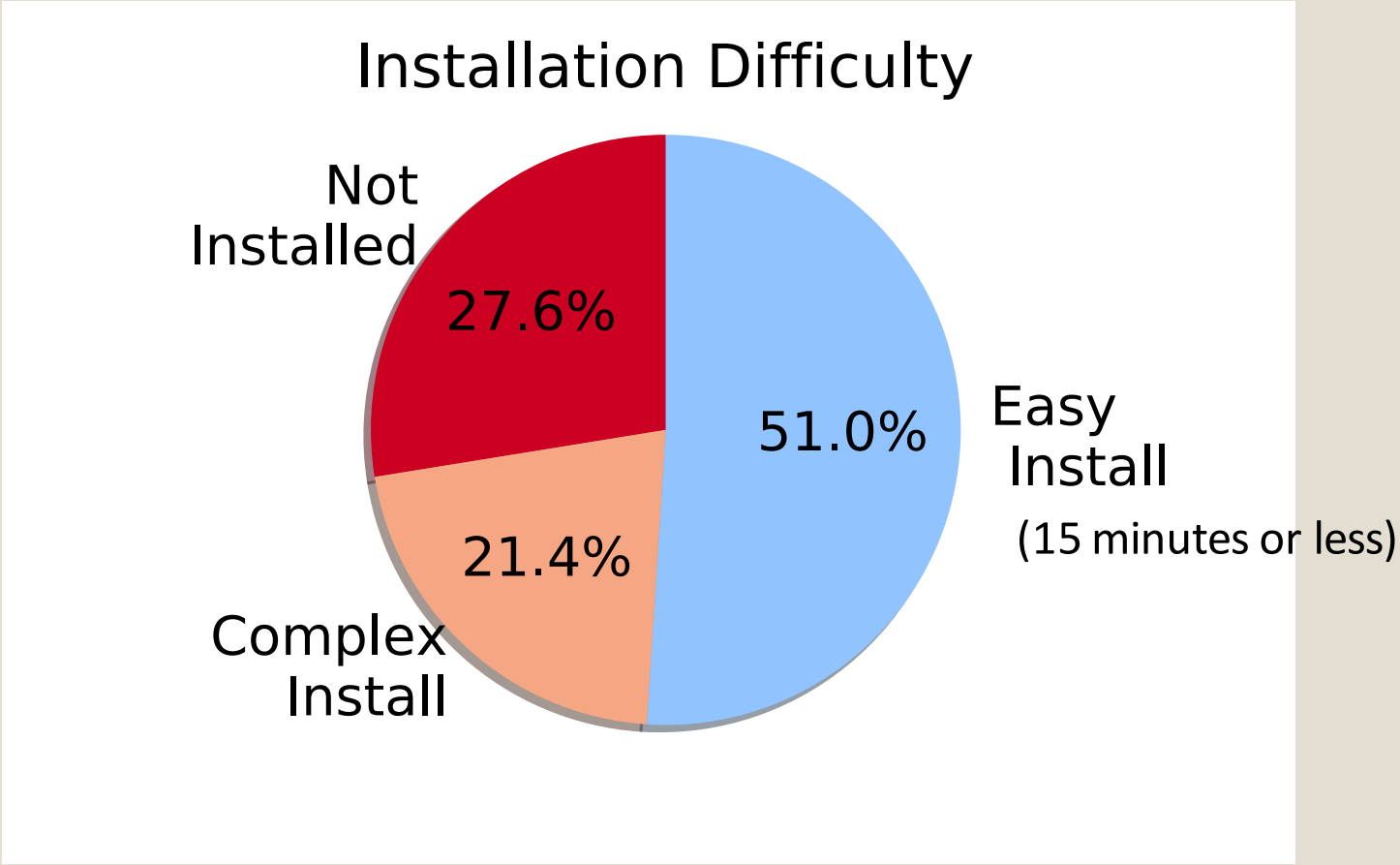
- We have randomly selected **99** tools across various domains of computational biology.
- Assign one undergraduate student to install 10 tools or more
- Record total time, number of commands to install the tools
- Total human time to install 99 tools -- **72 hours**

S. Mangul, et. al , Involving undergraduates in genomics research to narrow the education–research gap, *Nature biotechnology* 36.4 (2018): 369.

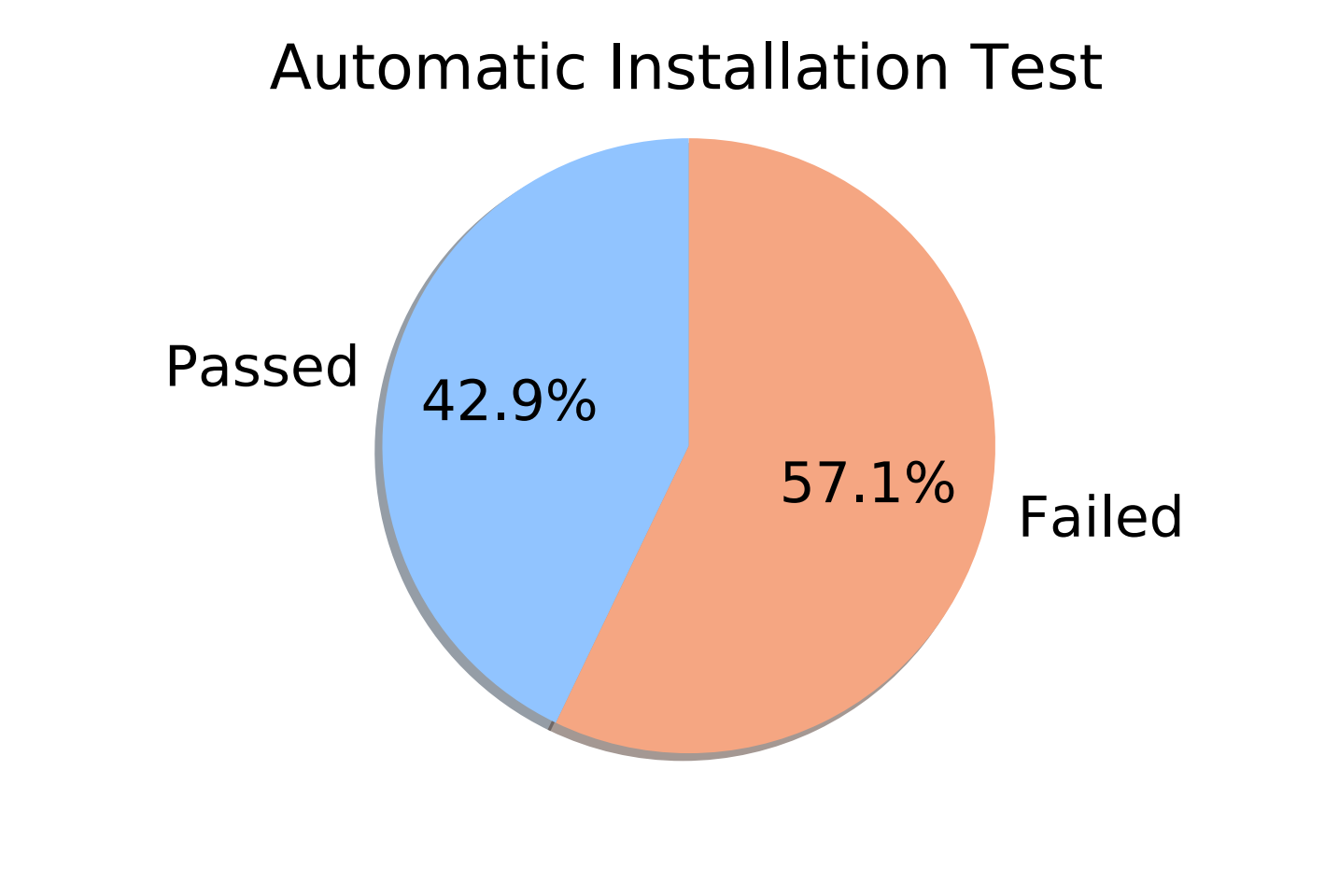
Protocol to check the insatiability of the tool



Many tools are hard or impossible to install

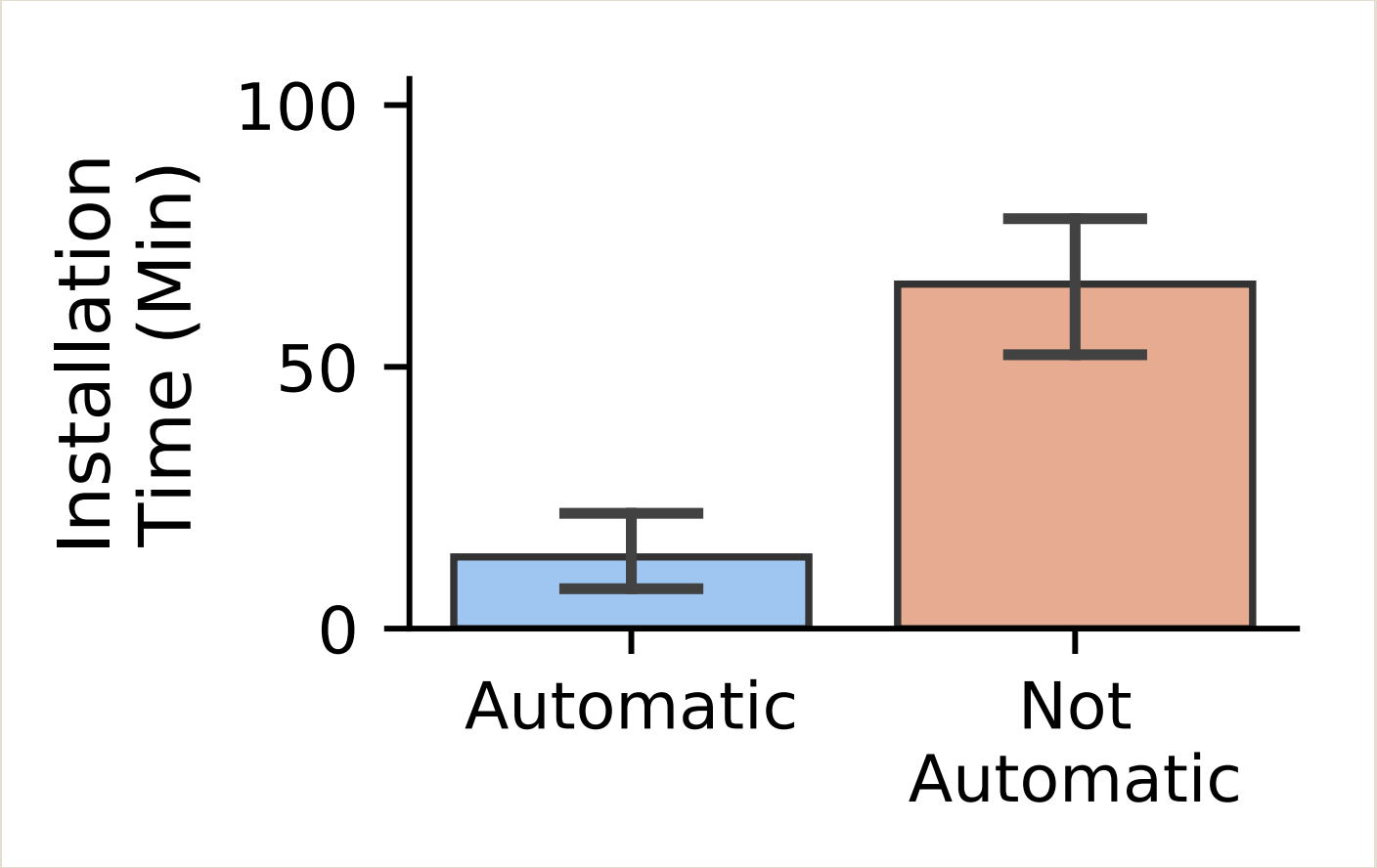


Automatic installation test

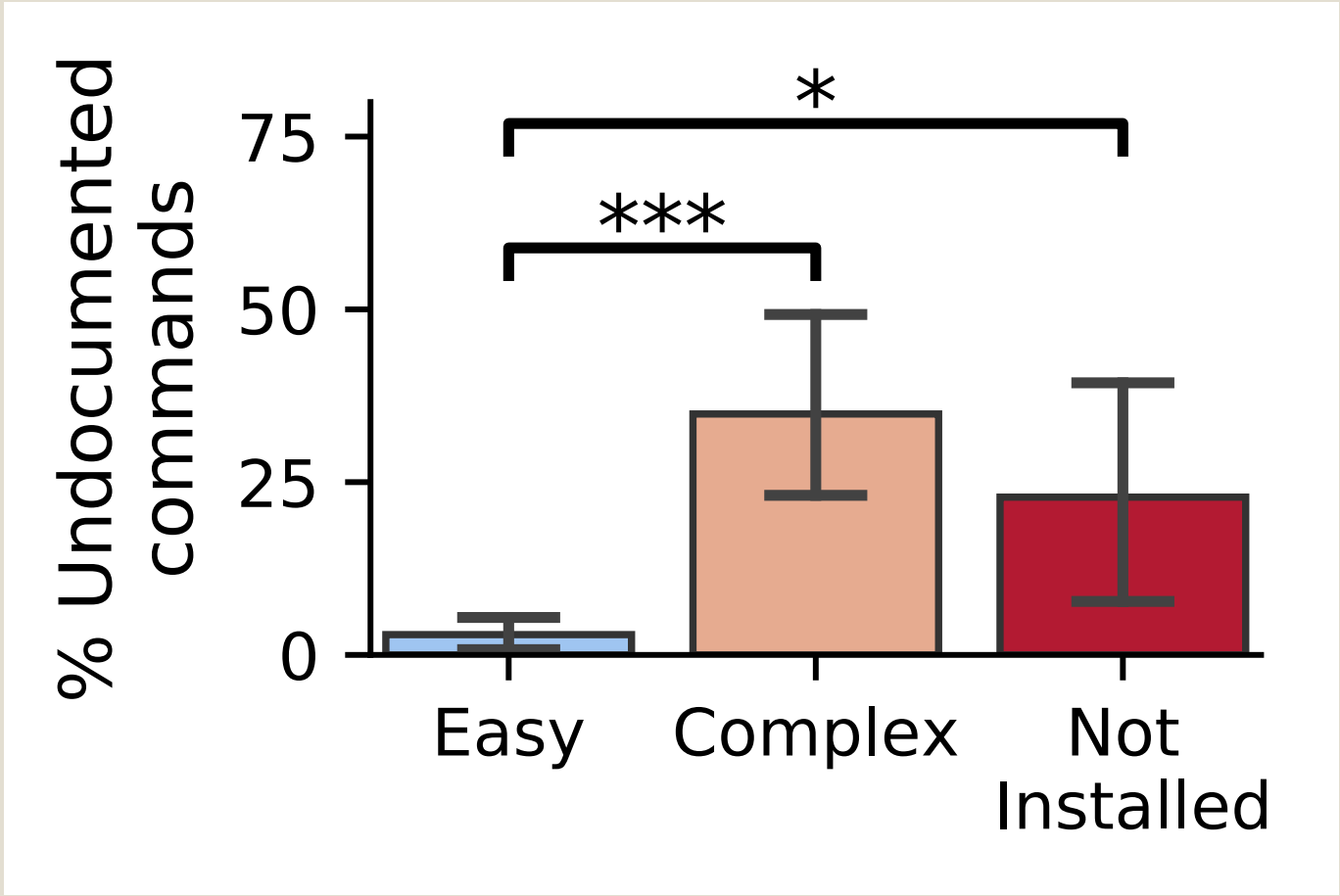


Passed: no manual intervention

Manual interventions are time consuming

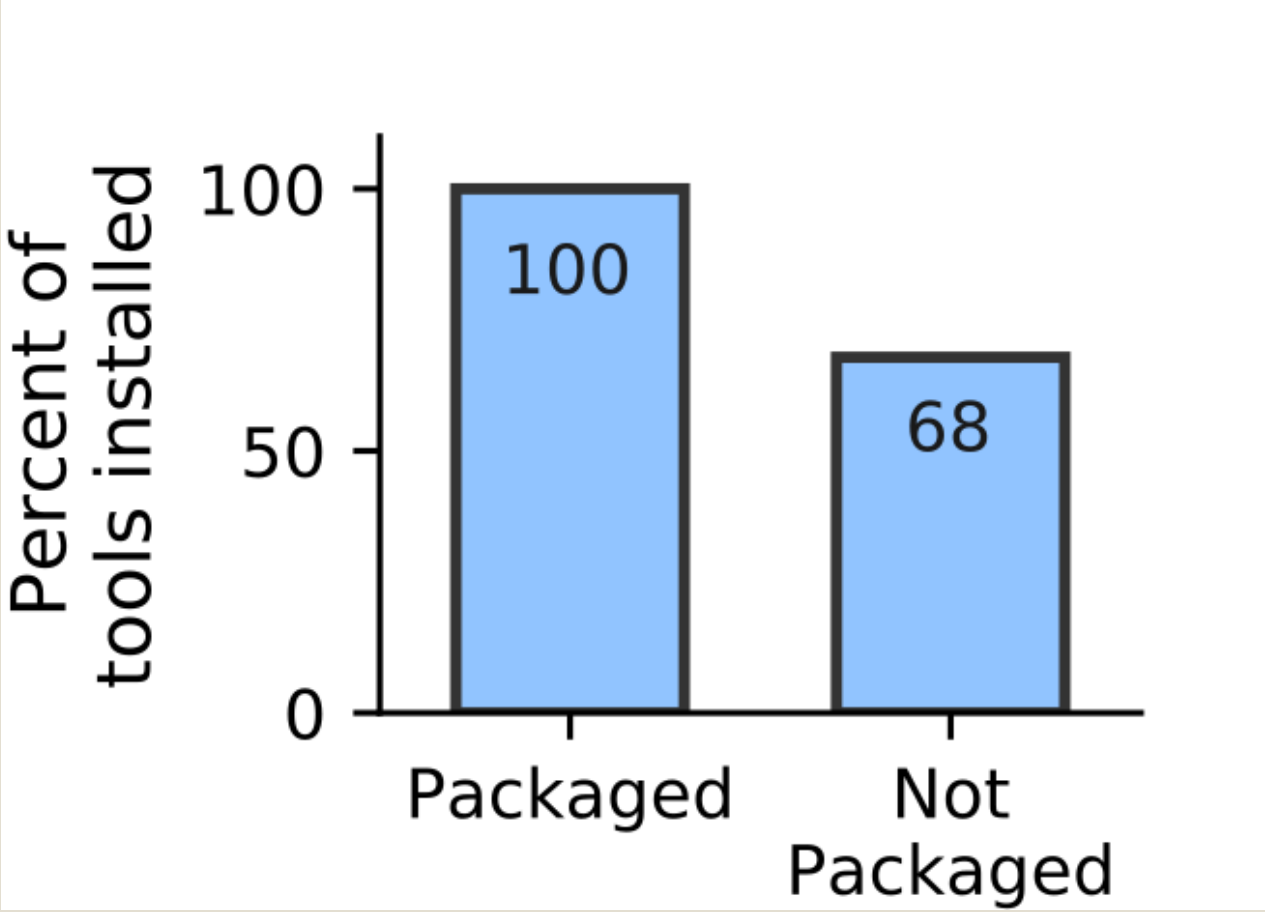


Easy-to-install* tools have fewer undocumented commands

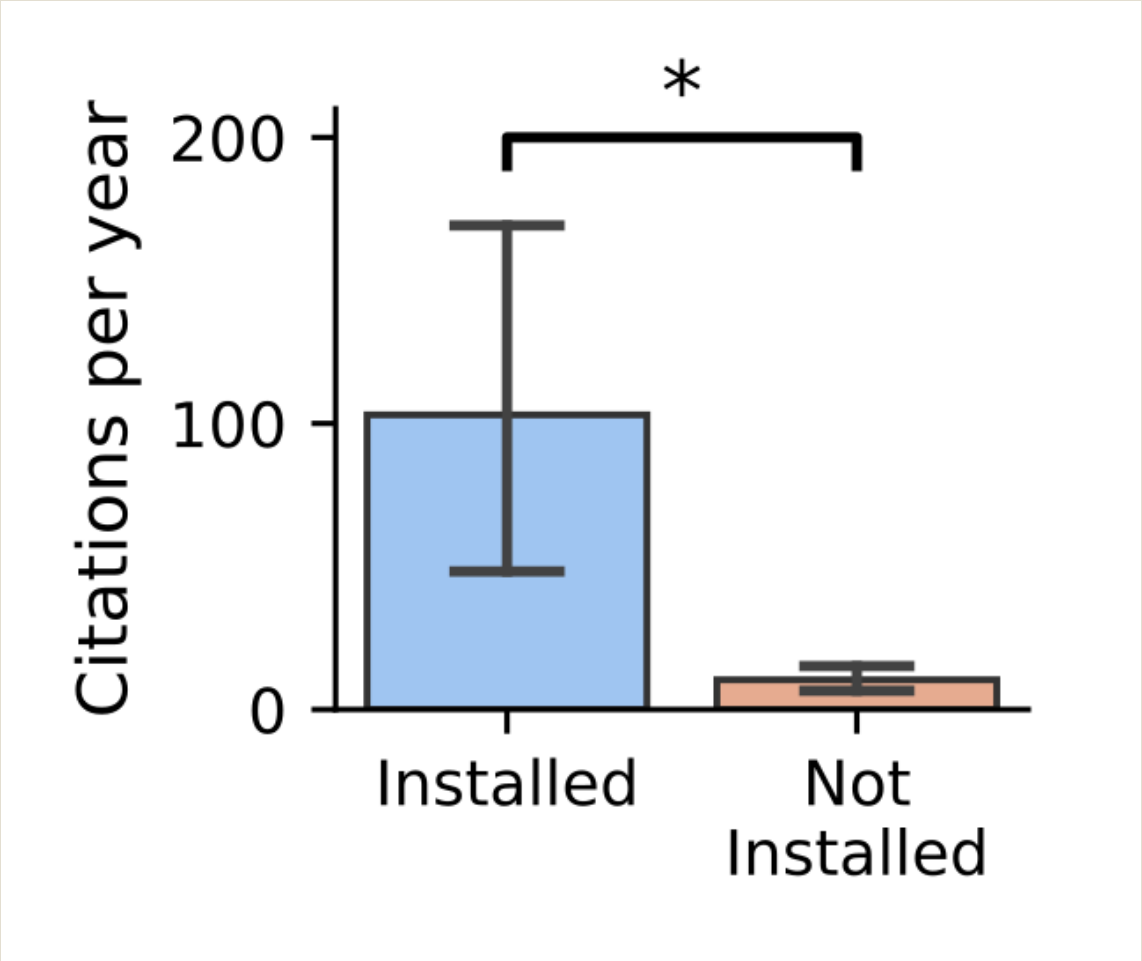


*<15min

Bioconda tools were always installable

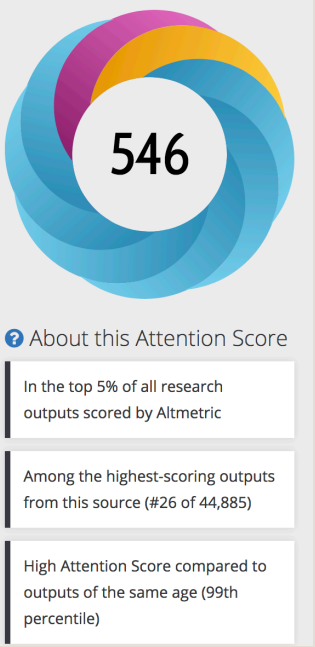


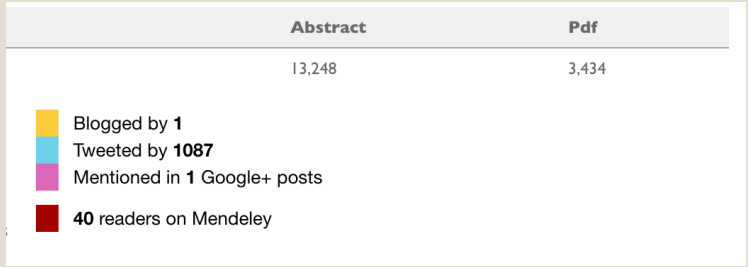
Installable tools have more citations



Conclusions

- Host software on archaically stable services e.g. **GitHub**
- Provide easy-to-use installation interface and get more citations
- Tools wrapped in **package managers (e.g. Bioconda)** are 100% installable!





Editorial (Genome Biology) is available at https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1649-8
Preprint is available at https://www.biorxiv.org/content/early/2018/10/25/452532

Future work


Serghei Mangul @serghei_mangul · Oct 30
1/2 CALL FOR PARTICIPANTS: Given the interest in our study about bioinformatics tools usability, we want to work on the follow up a paper about the package managers for bioinformatics tools. Please reply in this tweet if you are interested to participate! @thmosqueiro @blekhman
CALL FOR PARTICIPANTS
5
5
10

Acknowledgment

- Thiago Mosqueiro, Dat Duong, Keith Mitchell, Varuni Sarwal, Brian Hill, Jaqueline Brito, Russell Littman, Benjamin Statz, Angela Lam, Gargi Dayama, Laura Grieneisen, Lana Martin, Jonathan Flint, Eleazar Eskin, Ran Blekhman

We have prepared Jupyter Notebooks that utilize the raw data described above to reproduce the results and figures presented in our paper

→ https://github.com/smangul1/good.software

smangul@ucla.edu
http://www.sergheimangul.com/

THANK YOU.