SSI & Jisc Software Deposit and Preservation Workshop, Oxford, July 11 2018

Research software preservation: a publisher's perspective

Naomi Penfold, Innovation Officer

Email: n.penfold@elifesciences.org

Twitter: @eLifeInnovation

Maria Guerreiro

Journal Development Editor

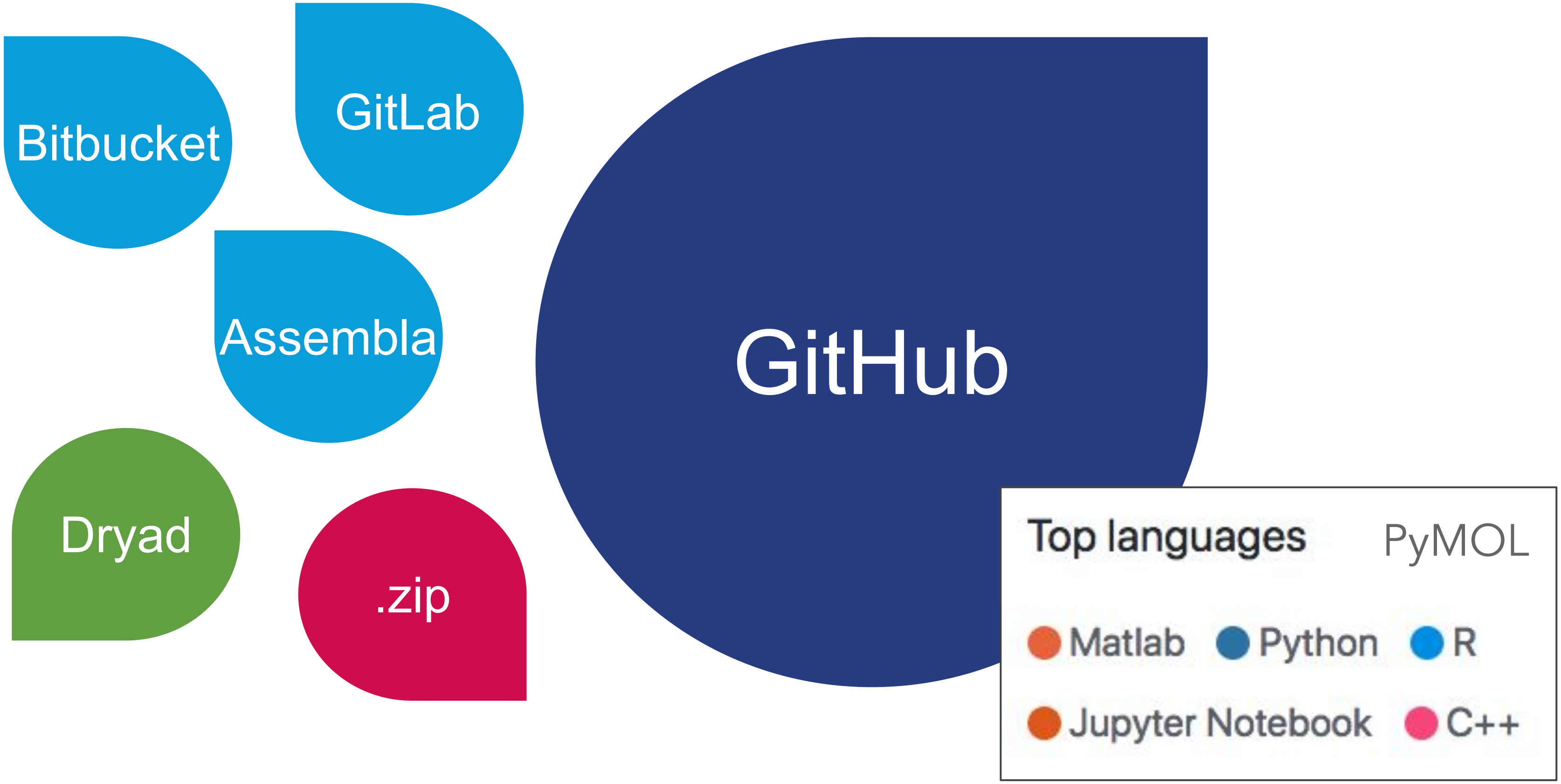
Summary and slides available at elifesci.org/software-preservation

Helping scientists accelerate discovery by operating a platform for research communication that encourages and recognises the most responsible behaviours in science

Today…

- Software shared with eLife
- How we cite and preserve software
- Cost-benefit of developing?
- Key requirements
- Opportunity to encourage best practice?

Research software shared with eLife



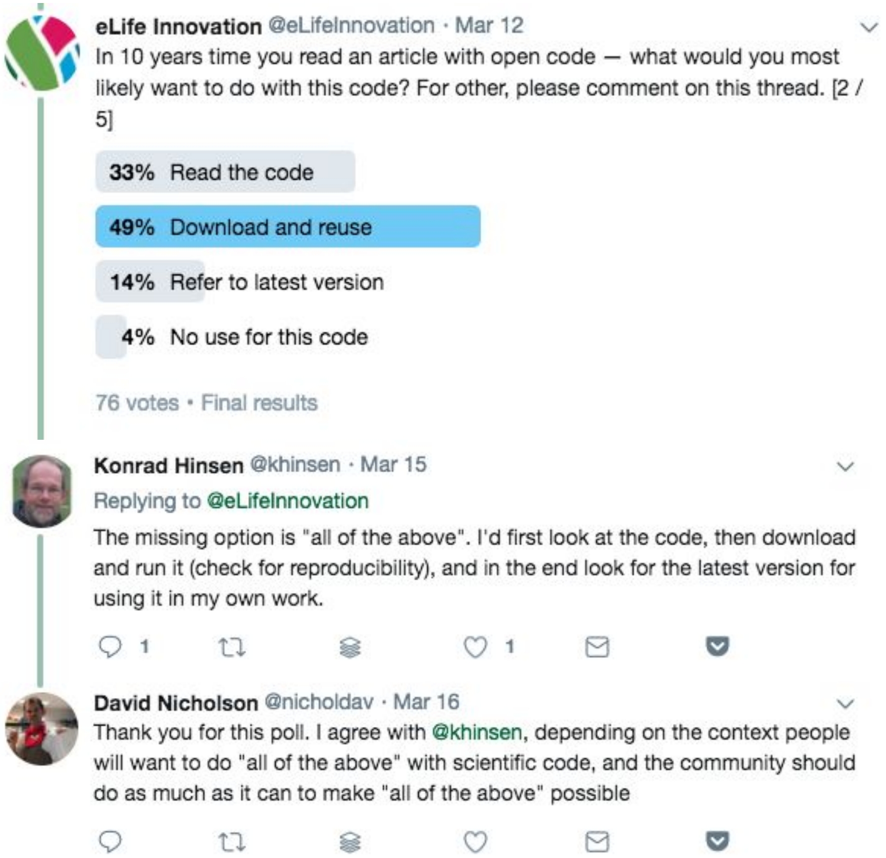
Top languages

PyMOL

● Matlab ● Python ● R

● Jupyter Notebook ● C++

Researchers want to keep options open for reuse



eLife Innovation @eLifeInnovation · Mar 12

In 10 years time you read an article with open code — what would you most likely want to do with this code? For other, please comment on this thread. [2 / 5]

33% Read the code

49% Download and reuse

14% Refer to latest version

4% No use for this code

76 votes • Final results

Konrad Hinsen @khinsen · Mar 15

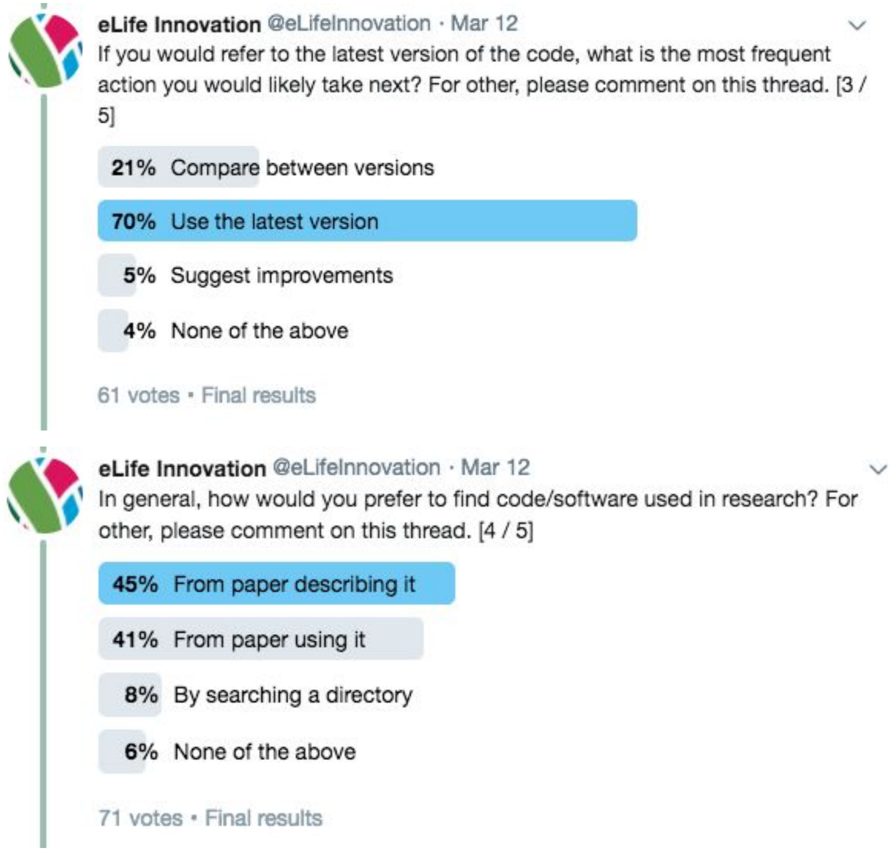
Replying to @eLifeInnovation

The missing option is "all of the above". I'd first look at the code, then download and run it (check for reproducibility), and in the end look for the latest version for using it in my own work.

💬 1 ♡ 1

David Nicholson @nicholdav · Mar 16

Thank you for this poll. I agree with @khinsen, depending on the context people will want to do "all of the above" with scientific code, and the community should do as much as it can to make "all of the above" possible



eLife Innovation @eLifeInnovation · Mar 12

If you would refer to the latest version of the code, what is the most frequent action you would likely take next? For other, please comment on this thread. [3 / 5]

21% Compare between versions

70% Use the latest version

5% Suggest improvements

4% None of the above

61 votes • Final results

eLife Innovation @eLifeInnovation · Mar 12

In general, how would you prefer to find code/software used in research? For other, please comment on this thread. [4 / 5]

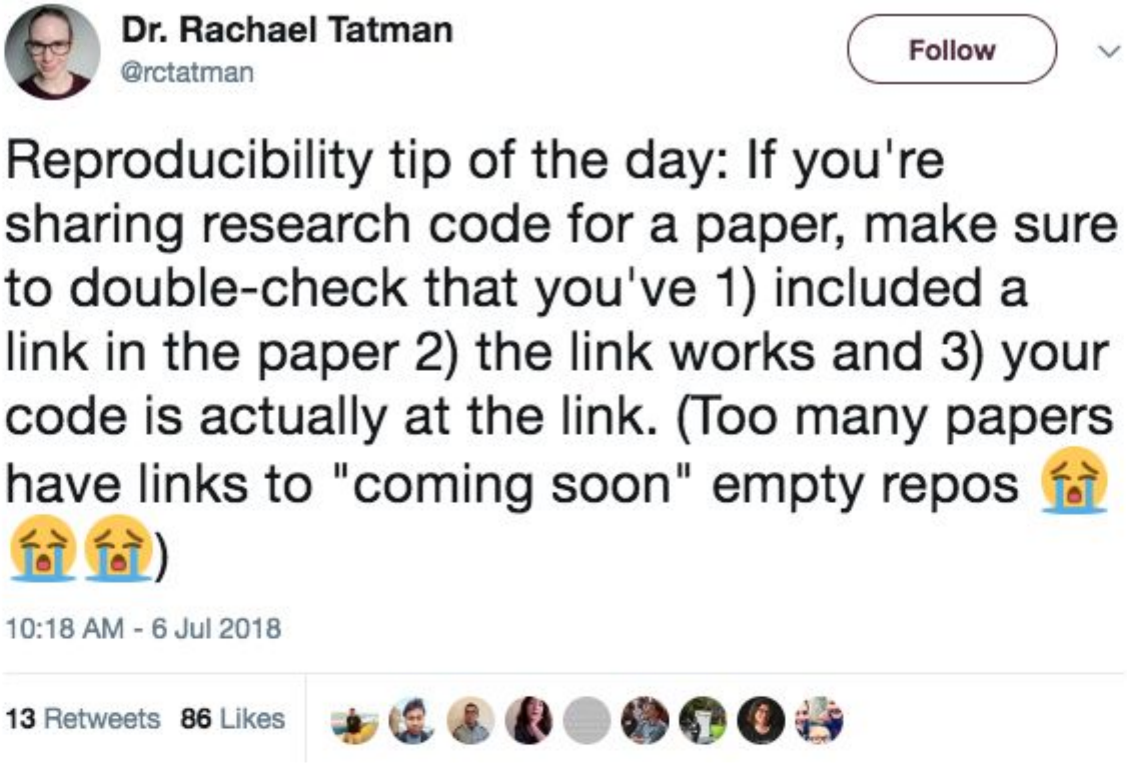
45% From paper describing it

41% From paper using it

8% By searching a directory

6% None of the above

71 votes • Final results



Dr. Rachael Tatman
@rctatman

Reproducibility tip of the day: If you're sharing research code for a paper, make sure to double-check that you've 1) included a link in the paper 2) the link works and 3) your code is actually at the link. (Too many papers have links to "coming soon" empty repos 😭😭😭)

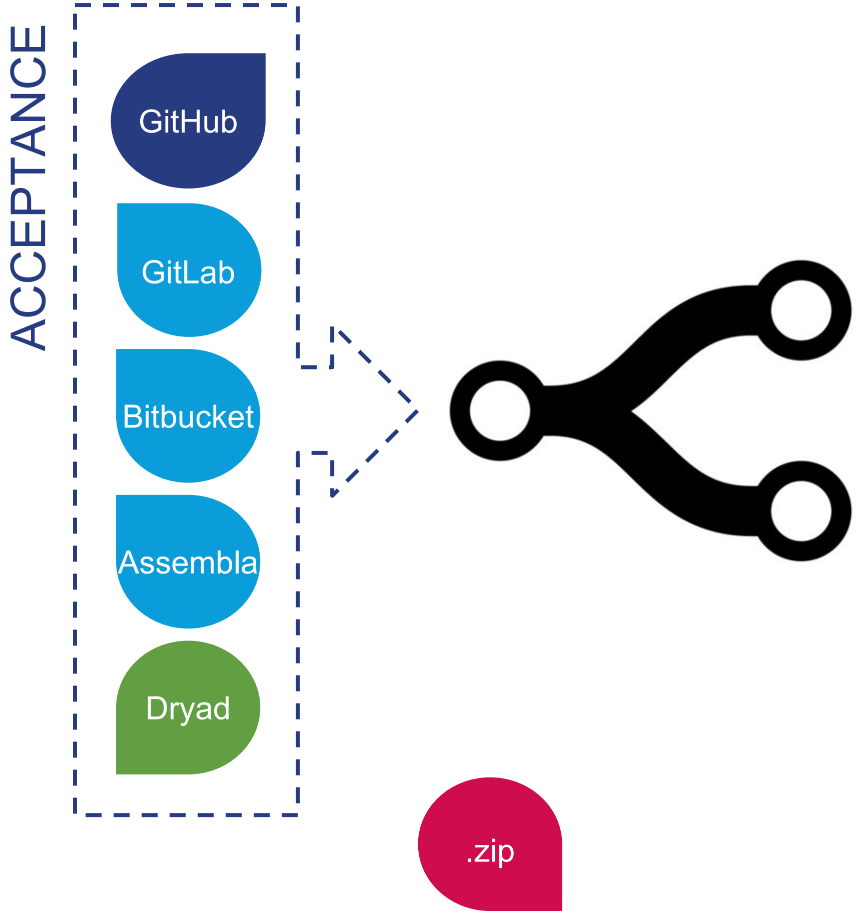
10:18 AM - 6 Jul 2018

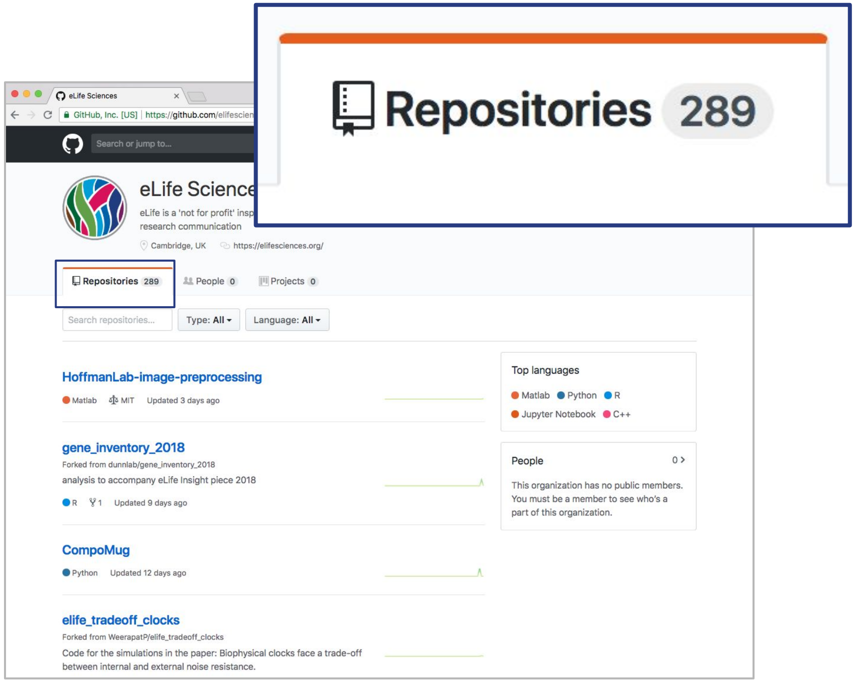
13 Retweets **86** Likes

https://twitter.com/rctatman/status/1015283853131304960

Learn more: https://elifesciences.org/inside-elife/dbcb6949/forking-software-used-in-elife-papers-to-github

We fork to our own GitHub repository




Repositories 289

eLife Sciences

GitHub, Inc. [US] | https://github.com/elifescien

Search or jump to...

eLife Science

eLife is a 'not for profit' insp research communication

Cambridge, UK https://elifesciences.org/

Repositories 289 People 0 Projects 0

Search repositories... Type: All Language: All

HoffmanLab-image-preprocessing

● Matlab MIT Updated 3 days ago

gene_inventory_2018

Forked from dunnlab/gene_inventory_2018

analysis to accompany eLife Insight piece 2018

● R 1 Updated 9 days ago

CompoMug

● Python Updated 12 days ago

elife_tradeoff_clocks

Forked from WeerapatP/elife_tradeoff_clocks

Code for the simulations in the paper: Biophysical clocks face a trade-off between internal and external noise resistance.

Top languages

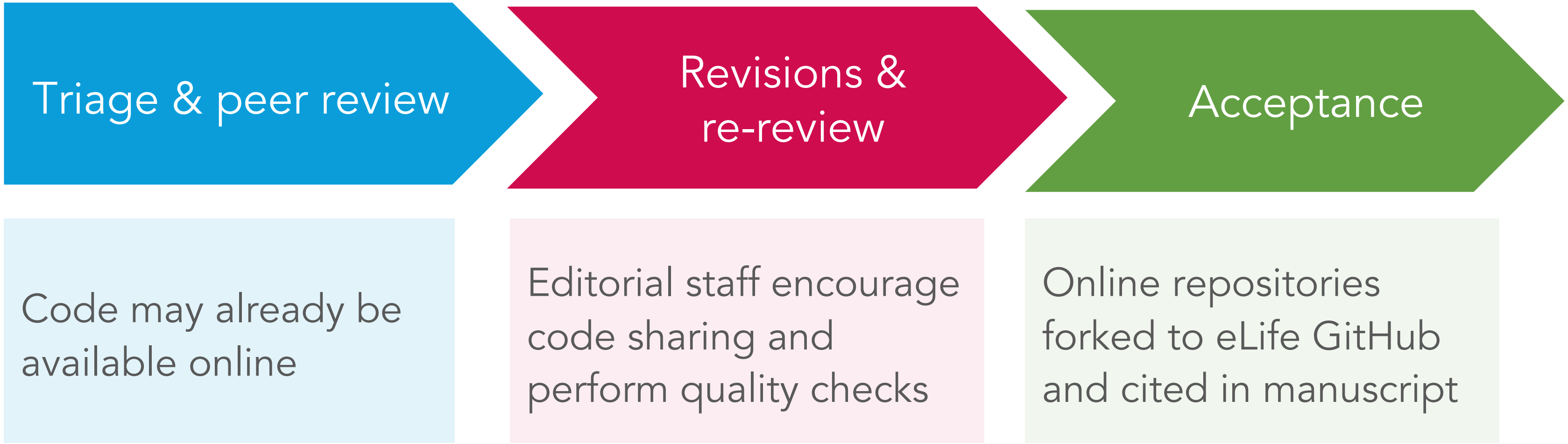
● Matlab ● Python ● R
● Jupyter Notebook ● C++

People 0 >

This organization has no public members. You must be a member to see who's a part of this organization.

https://github.com/elifesciences-publications/

The workflow in detail



Triage & peer review

Code may already be available online

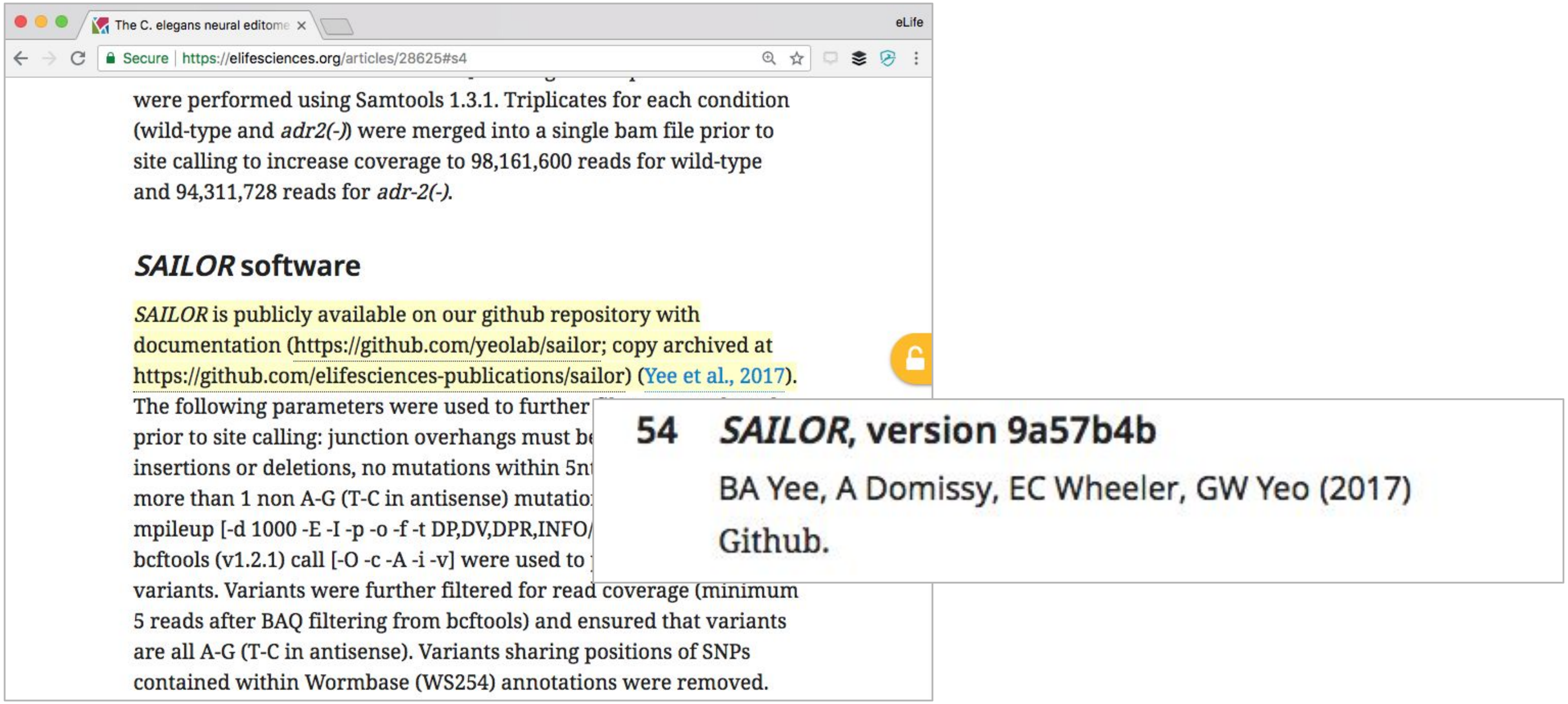
Revisions & re-review

Editorial staff encourage code sharing and perform quality checks

Acceptance

Online repositories forked to eLife GitHub and cited in manuscript

Citation in the text and references


The C. elegans neural editome
eLife
Secure
https://elifesciences.org/articles/28625#s4

were performed using Samtools 1.3.1. Triplicates for each condition (wild-type and *adr2(-)*) were merged into a single bam file prior to site calling to increase coverage to 98,161,600 reads for wild-type and 94,311,728 reads for *adr-2(-)*.

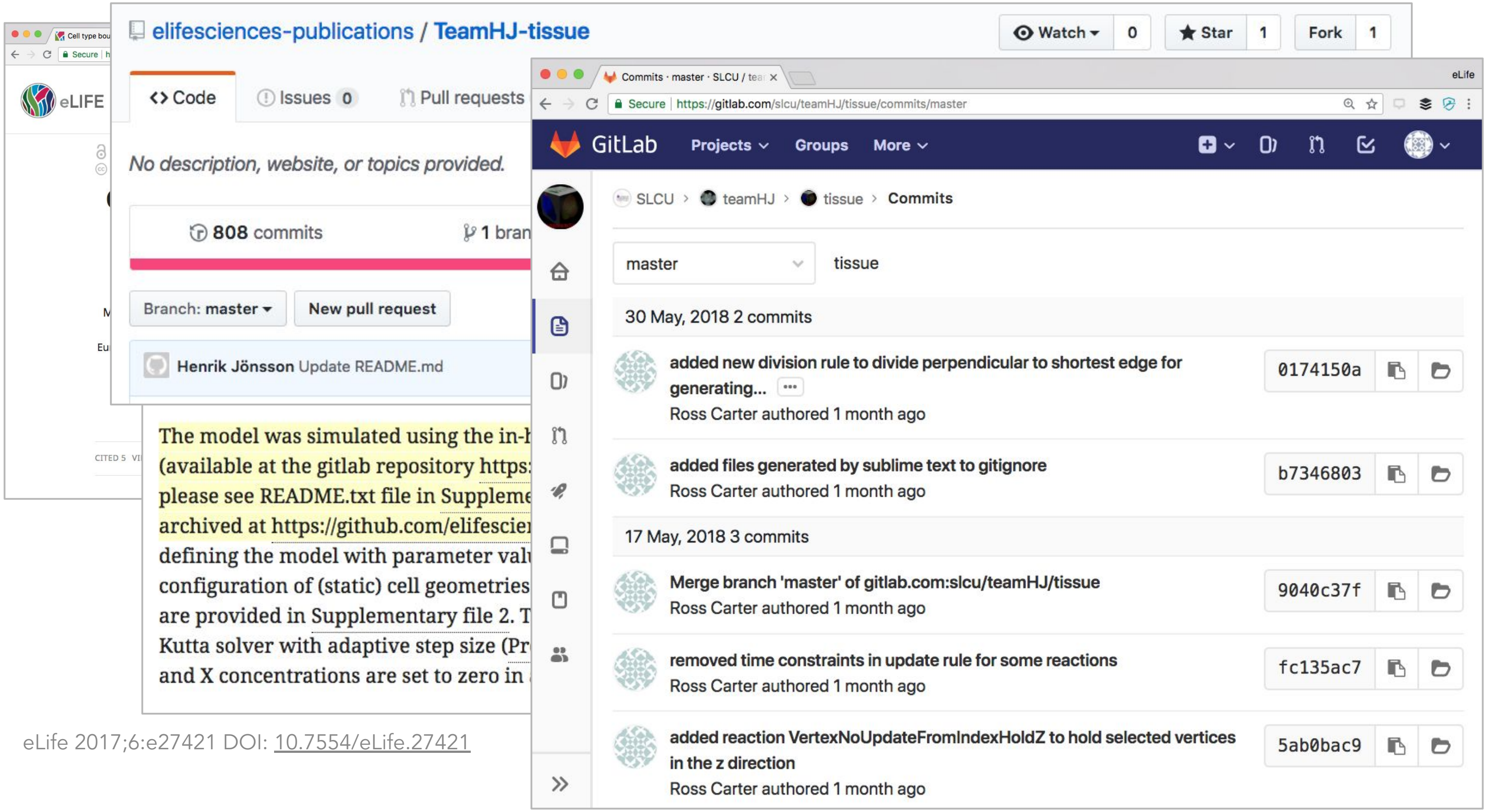
SAILOR software

SAILOR is publicly available on our github repository with documentation (https://github.com/yeolab/sailor; copy archived at https://github.com/elifesciences-publications/sailor) (Yee et al., 2017). The following parameters were used to further
prior to site calling: junction overhangs must be
insertions or deletions, no mutations within 5n
more than 1 non A-G (T-C in antisense) mutatio
mpileup [-d 1000 -E -I -p -o -f -t DP,DV,DPR,INFO/
bcftools (v1.2.1) call [-O -c -A -i -v] were used to
variants. Variants were further filtered for read coverage (minimum 5 reads after BAQ filtering from bcftools) and ensured that variants are all A-G (T-C in antisense). Variants sharing positions of SNPs contained within Wormbase (WS254) annotations were removed.

54 ***SAILOR*, version 9a57b4b**
BA Yee, A Domissy, EC Wheeler, GW Yeo (2017)
Github.

eLife 2017;6:e28625 DOI: 10.7554/eLife.28625

Research software continues to develop



No description, website, or topics provided.

Henrik Jönsson Update README.md

The model was simulated using the in-h
(available at the gitlab repository https:
please see README.txt file in Suppleme
archived at https://github.com/elifescien
defining the model with parameter valu
configuration of (static) cell geometries
are provided in Supplementary file 2. T
Kutta solver with adaptive step size (Pr
and X concentrations are set to zero in

Commits · master · SLCU / tea
Secure | https://gitlab.com/slcu/teamHJ/tissue/commits/master
GitLab
Projects
Groups
More
SLCU › teamHJ › tissue › Commits
master
tissue

30 May, 2018 2 commits

- added new division rule to divide perpendicular to shortest edge for generating... — Ross Carter authored 1 month ago — 0174150a
- added files generated by sublime text to gitignore — Ross Carter authored 1 month ago — b7346803

17 May, 2018 3 commits

- Merge branch 'master' of gitlab.com:slcu/teamHJ/tissue — Ross Carter authored 1 month ago — 9040c37f
- removed time constraints in update rule for some reactions — Ross Carter authored 1 month ago — fc135ac7
- added reaction VertexNoUpdateFromIndexHoldZ to hold selected vertices in the z direction — Ross Carter authored 1 month ago — 5ab0bac9

eLife 2017;6:e27421 DOI: 10.7554/eLife.27421

Preserving research software at eLife

Benefits

- Scientist-driven
- Less work for authors
- Reuse is facilitated
- **Encourages best practice: powerful when combined with data and other resources**

Limitations

- No DOI
- Reliant on Github
- Requires staff time
- Not for source code files shared directly with the journal; what are the advantages/disadvantages of hosting code on the journal website?

Going further: is it worth it?

- How much are we prepared to invest as a community in process innovation or development? Or to support an archive?
- For how much added value? To whom?

How do we evaluate cost-benefit for software preservation?

Image credits, all from the Noun Project: (left to right, top to bottom) Work by BGBOXXX Design; Scale by Amelia; Reuse by Desbenoit, FR (Public domain, CC0); tags by Thomas Helbig; archive by Yuri Mazursky; Download by Landan Lloyd

Requirements as a publisher

Minimise work for authors, make it sustainable for publishers

Scalable and agnostic to platform, format or tool

For reusability wherever possible

Metadata collected at source, compliant with citation guidelines

Persistent and retrievable for as long as is reasonable

Image credits: Check by Icon Depot from the Noun Project; Workflow by Kirby Wu from the Noun Project

Can we encourage best practice?

☑ LICENSE.md

☑ CITATION.cff or codemeta.json

Include event-driven process in open source publishing platform

Together with:

- Text
- Data
- Key Resources Table

→ open and reproducible research package

What if the software is **not** open?

How do we do the best for all cases?

Further investigations

- Software citations, interactions, and activity at eLife
- Other publisher workflows

Collaboration welcome

In summary…

- Software shared with eLife is mainly on Github but we need to cater for any source
- We fork online repositories to our Github upon acceptance but have no process to preserve source code files
- We would like to minimise burden and cost, encourage best practice, and support researchers to reuse where reasonable

Today we ask:

- What do other publishers do?
- Can we encourage best practice?
- Can we help you test new process(es)?
- **Is it worth it?**

Questions?

Slides: elifesci.org/software-preservation

Email: n.penfold@elifesciences.org

Twitter: @eLifeInnovation