AUTOMATIC REPRODUCTION OF WORKFLOWS IN SNAKEMAKE WORKFLOW CATALOG AND NF-CORE

Samuel Grayson, Darko Marinov, Daniel S. Katz, Reed Milewicz

Automatic Reproduction of Workflows in Registries (Snakemake Workflow Catalog and Nf-core)

WHY AUTOMATIC REPRODUCIBILITY?

- Test the current state of reproducibility in practice
- Reproducible $:=$ different team/machine + same code $\Rightarrow$ consistent measurement [4]
- True “research” reproducibility would need knowledge of the specific experiment
- Instead, look at crash-free reproducibility
- Crash-free reproducibility is necessary for resarch reproducibility

WHY WORKFLOWS?

- Workflow ≔ script that generates directed acyclic graph (DAG) of tasks
  - DAG edges specify data dependencies to other nodes
  - Usually each node runs in a container
- Workflow engine ≔ interpreter that runs the workflow and executes the DAG
- Workflow registry ≔ archive of workflows (e.g. GitHub)
- Workflows are easier to code than programs
  - Especially parallelism

PRIOR WORK

- “Why workflows break — understanding and combating decay in Taverna workflows” Zhao et al. 2012 [6]
- “Repeatability in Computer Systems Research” Collberg and Proebsting 2016 [1]
- “A large-scale study on research code quality and execution” Trisovic et al. 2022 [5]

RESEARCH QUESTIONS

0. Characterize the registries
1. How many workflows were crash-free reproducible?
2. Causes of crashes?

RQ0: WORKFLOW REGISTRIES

- Drawn from https://workflows.community/

REGISTRY: NF-CORE [2]

- Nextflow engine only
- Mostly multiomics users (genomics, proteomics, …)
- Community-curated workflows for common tasks
- Nf-core workflows follow certain conventions
  - Have `./main.nf`
  - Define profile for Singularity, Docker
- 48 workflows
- All less than 4.5 years
- Hosted in GitHub, can be viewed at https://nf-co.re

REGISTRY: SNAKEMAKE WORKFLOW CATALOG

- Snakemake engine only
- Mostly multiomics users (genomics, proteomics, …)
- All GitHub repositories that follow certain standards
  - Have a snakefile in a specific place
- 2,045 workflows but only 53 workflows with GitHub releases
- Developers can customize the usage command with their run file `.snakemake-workflow-catalog.yml`
- Almost all less than 2.5 years
- Hosted in GitHub, can be viewed at https://snakemake.github.io/snakemake-workflow-catalog/

REGISTRY: WORKFLOWHUB AND DOCKSTORE

- WorkflowHub.eu [3] and Dockstore
  - Multiple workflow engines ⇒ no automatic run commands
  - Future work

RESULTS

Quantity	All	SWC	nf-core
# workflows	101	53	48
# releases	584	333	251

RQ1: % OF AUTOMATIC REPRODUCIBILITY

- The registries advertise a command which runs these repositories.
- We want to know how often this command works without manual input.

RESULTS

Quantity	All	SWC	nf-core
# workflows	101	53	48
% of workflows with ≥ 1 non-crashing release	53%	23%	88%
# releases	584	333	251
% of releases with no crash	28%	11%	51%

RQ2: WHAT ARE COMMON ERROR CAUSES

1. Look at one unclassified crashing execution by hand.
    - Describe the high-level reason.
2. Write a regular expression to catch this kind of error (stderr, stdout, logs).
    - Make sure it is exclusive with the other regular expressions.
3. Mark these as classified.
4. Repeat to 1
5. Spotcheck the results

- Workflow task error
  - Timeout
  - Network resource changed
  - Missing software dependency
  - Other
- Workflow script error
  - Missing data/config input
  - Other
- Workflow engine error
  - Singularity error
  - Conda environment unsolvable
- Unclassified

RESULTS

Kind of crash	All	SWC	nf-core
Missing data/config input	32.2%	43.8%	16.7%
Conda environment unsolvable	10.8%	18.9%	0.0%
Unclassified reason	7.9%	12.0%	2.4%
Timeout reached	7.0%	5.7%	8.8%
Singularity error	6.0%	6.6%	5.2%
Other (workflow script)	5.7%	1.5%	11.2%
Other (workflow task)	1.2%	0.0%	2.8%
Network resource changed	0.7%	0.0%	1.6%
Missing software dependency	0.5%	0.9%	0.0%
No crash	28.1%	10.5%	51.4%
Total	100%	100%	100%

OBSERVATIONS

- Misisng example data/config is prominent
  - SWC YAML run file does not have a place for example data
  - Sometimes the nf-core `test` profile is insufficient!
  - Workflows should default to example data (downloaded or generated)
- Conda environment solve is also a common factor
  - Conda, by default, does not generate lockfiles
  - Difficult to debug
  - Packages can get yoinked
  - Source-level package managers could be more robust (Spack, Nix, Guix)
    - But Conda is specially supported by Snakemake

DISCUSSION

CONTAINERS REQUIRE SUPERUSER TO INSTALL

- Users won’t necessarily have root on shared systems
- Rootless user-namespaces exist, but may be too new or not enabled
- Why do we need root to reproduce someone else’s code?
    - Linux filesystems are kernel modules, requires root to modify
- Future work could look at CharlieCloud

TOWARDS AN EXECUTION DESCRIPTION LANGUAGE

- SWC supports a YAML file which says how to run the workflow
- Develop workflow-agnostic way of saying how to run a workflow
  - Stored with the workflow code
  - Could be done by workflow authors or by reproducers
  - Simplify artifact evaluation, CI, reusability, reproducibility

REFERENCES

[1] Collberg, C. and Proebsting, T.A. 2016. Repeatability in computer systems research. *Communications of the ACM*. 59, 3 (Feb. 2016), 62–69. DOI:https://doi.org/10.1145/2812803.

[2] Ewels, P.A. et al. 2020. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*. 38, 3 (Mar. 2020), 276–278. DOI:https://doi.org/10.1038/s41587-020-0439-x.

[3] Ferreira da Silva, R. et al. 2020. WorkflowHub: Community Framework for Enabling Scientific Workflow Research and Development. *2020 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS)* (Georgia, USA, Nov. 2020), 49–56.

[4] staff, A.Inc. 2020. Artifact Review and Badging.

[5] Trisovic, A. et al. 2022. A large-scale study on research code quality and execution. *Scientific Data*. 9, 1 (Feb. 2022), 60. DOI:https://doi.org/10.1038/s41597-022-01143-6.

[6] Zhao, J. et al. 2012. Why workflows break — understanding and combating decay in Taverna workflows. *2012 IEEE 8th International Conference on E-Science (e-Science)* (Chicago, IL, Oct. 2012), 9.

BACKUP SLIDES

LINKS

- This presentation
- Paper preprint
- Code (GitHub) (archived)
- Raw data (rendered) (archived)

  `./data/results.html`

CONTINUOUS INTEGRATION

- Use current CI scripts
  - If the script multiple targets, which one to use?
  - None of them may actually run the experiment, if it is too expensive to run in CI.
- Use CI script format but write new scripts
  - No way to say **what** the command does
  - Need linked data for that