Exploring Reproducibility in Bioinformatics: Lessons Learned from an NLM Workshop

1. Background & Motivation

In recent years, alarms have been raised over the lack of reproducibility in various research fields. According to a 2016 *Nature* survey of over 1,500 researchers from various fields, more than 70% had tried and failed to reproduce another researcher's study. Researchers in a number of disciplines—from psychology to cancer biology to social science—have systematically worked to replicate studies in their respective fields with varying degrees of success.

The scientific community has made several efforts to improve the rigor of research and facilitate reproducibility. Standards, recommendations, and guidelines like the Transparency and Openness Promotion (TOP) Guidelines have been proposed to promote and recognize reproducible research practices. A number of journals and funders, including the NIH, have adopted policies requiring researchers to make their research data publicly available, in hopes of improving both transparency and reproducibility. Consequently, reproducibility has garnered a great deal of attention from the research community, creating a need to develop training and best practices around reproducible research practices.

To begin to explore how a curriculum around reproducibility might take shape, we piloted a three-day workshop on Reproducibility in Bioinformatics for NIH intramural researchers in September 2018. We then incorporated lessons learned, hosting a second workshop in May 2019. In both iterations, participants were tasked with reproducing a paper from the bioinformatics literature, while learning tools to facilitate reproducible research. The selected papers had data publicly available in either NCBI's Gene Expression Omnibus (GEO) repository or in GenBank. The objective was to provide participants with:

1. a working knowledge of tools for reproducible research—specifically, executable notebooks, version control, and containerization—and NLM's data resources for bioinformatics;
2. an understanding how to incorporate these tools into their research practices; and
3. a path towards a deliverable, in the form of an executable notebook and/or publication.

The workshops also served as an exploratory study to investigate the experience of researchers attempting to reproduce bioinformatics papers. These findings would better inform the development of training materials on reproducible research practices, as well as to understand problems that may prevent researchers from reproducing published research results.

2. Workshop Structure

A. Participants

In the first iteration of the workshop, 51 interested researchers submitted applications for participation. Of these, 25 participants from ten NIH Institutes and Centers were selected, based on level of interest and basic programming experience, and 19 of the selected participants ultimately participated in the workshop. In the second iteration of the workshop, 42 researchers applied and 25 were selected from 11 NIH Institutes and Centers; 23 ultimately participated in the workshop.

Participants were required to have command line knowledge and to self-identify as having at least an “intermediate” level of knowledge of either Python, R, Java, C/C++, JavaScript, or Perl. Level of interest was determined by response to the question: "Please tell us in 2-3 sentences why you want to participate in this workshop?” Applicants who did not answer the question were disqualified. In addition to being used to select participants, responses to the question informed the topics covered over the course of the workshop.

Applicants were given links to ten possible papers, all with underlying data hosted in one of NLM’s repositories, from which to select their top three choices for studies to reproduce. In advance of the workshop, participants were split into five teams and each group was assigned a paper from their top three choices to reproduce. Each team had a range of programming expertise and contained members from varying career stages and positions across NIH, to encourage mentorship within the groups.

In both workshops, three instructors were identified to give brief workshops related to reproducibility including specific tools, such as Docker and Github, as well as more general lectures on topics like open science and publishing reproducibility studies. In addition, between four and five mentors were available throughout the three days to provide guidance and feedback to the teams. These mentors had expertise in both bioinformatics and various coding languages.

B. Format

The workshops took place over the course of three days and were run in a codeathon-style format, with participants working in teams to reproduce one of five papers from the bioinformatics literature. The schedules for both workshops were largely the same and are outlined in Appendix A. Papers were selected by the workshop organizers based on availability of the underlying data in one of NLM’s repositories, providing either a Gene Expression Omnibus (GEO) or GenBank accession codes. In the second iteration of the workshop, participants could opt into a pre-workshop “download-a-thon” to download the underlying datasets in advance to give themselves more time during the workshop to work on the actual reproduction of results.

The workshop began with an hour-long primer on open science and reproducibility to establish common definitions and the greater open science framework in which reproducibility fits. Because there is no one standardized definition for reproducibility—and standards for reproducible research may vary by discipline—definitions were given for a range of contexts, for example, distinguishing computational, empirical, and statistical reproducibility. The lecture also provided an overview of tools to make research more reproducible, including those covered over the course of the workshop.

Participants spent the rest of the morning working to launch cloud instances on Amazon Web Services, download the underlying datasets, and develop a plan for reproducing their assigned papers. In the afternoon, instructors led participants through sessions to introduce three tools to facilitate reproducible research:

1. the computable notebook, Jupyter notebook;
2. version control using Git and GitHub; and
3. containerization of computational environments using Docker.

Participants spent the remainder of the workshop carrying out their attempts at reproduction with occasional assistance from mentors. On the final day, groups presented their progress to each other and discussed challenges they faced. The workshop closed with a presentation on publication options and award opportunities for efforts in reproducibility and open science.

3. Themes

Over the course of the workshop, a number of themes emerged as teams worked to reproduce their assigned papers. While no team was able to fully reproduce a paper, each failed to do so in a different way, which provided practical insights into the challenges around reproducibility, as well as opportunities to improve dissemination of research findings to enhance reproducibility and openness.

A. Reproducibility is not trivial

Journal publications are not written with the idea that researchers may use them to independently validate the conclusions. During the workshops, participants ran into a variety of obstacles, which included:

- **Missing underlying data**. While GEO or GenBank accession numbers were listed for each of the selected papers, in some cases, participants were unable to locate the underlying raw, unprocessed data. One of the teams was unable to locate data for ten papers before finally locating a study with available raw data. Another team found a note in their assigned paper that additional data are listed in supplementary information files, but no such files existed; the paper did, however, detail the NCBI GEO accession codes for microarray and RNA-seq data in the body of the text.

- **Missing software and tools**. In several papers, the software and tools used in the analysis were not made openly available and were instead described as “custom in-house scripts.” One paper included a link to the tool the authors had developed, but when the team tried to access it, the URL was no longer valid.
- **Inadequate descriptions of software and tools**. Several teams had difficulties installing and running the required software and tools used to analyze the data for a variety of reasons. For instance, tool versions were not listed in the methods section of some of the papers, making it difficult to set up the proper computing environment in the case of tools with software dependencies.
- **Workflows inadequately described or difficult to follow**. While some papers clearly detailed the workflow analysis pipeline, others left the details open for interpretation. As a result, some teams experienced difficulty trying to understand what exactly was done to process and analyze the data. Even where specific tools were referenced, the parameters used to run them were not indicated, which could significantly affect how data were process and the ultimate results of a given workflow. Participants also noted that they were unsure of which tools or workflows were used to analyze which datasets. This lack of clarity might have contributed to results inconsistent with those detailed in the original study.
- **Data did not map to the conclusions or described workflows**. While the underlying data for each paper were shared in NLM repositories, it was not immediately clear how the data supported a given set of results or conclusions detailed in the paper. For instance, participants found that certain datasets did not map to any figures or conclusions, though they were still referenced in the paper. Others could not determine which datasets were used in particular analysis workflows, making a reproduction of those workflows and their results difficult.

B. Need better minimum standards for peer review

Given the logistical challenges to reproduction detailed in the section above, teams discussed a minimum set of standards for peer review to facilitate reproducibility, which would significantly alleviate some of the challenges identified during these workshops but would not be onerous for reviewers. These are:

- **Underlying raw data are made readily available**. While many journals require that underlying data are made available within a reasonable period of time following publication, there appears to be little enforcement to ensure compliance. In some cases, the accession codes for datasets pointed to processed, rather than raw data, while in other cases the accession codes or links to datasets led to dead ends. One paper referred to data in the supplemental information, but no supplemental information was provided. Peer reviewers can play a role in ensuring that the supplements and data are indeed available in the location the authors indicate.
- **Underlying data are well-organized and clearly described**. While data may be readily available, they may be disorganized or named in a manner that is unclear. As a result, data users may have difficulty understanding what exactly the datasets contain and which

datasets were used for what part of the analysis. Reviewers can ensure that data are named in a consistent and intuitive manner. This review measure will enable readers to better understand which data underly a specific set of results and conclusions detailed in the research paper, while further promoting reproducibility.

- **All software and tools must detail the appropriate version**. Reviewers should ensure that versions are listed for all software and tools listed in the Methods section. Doing so will enable independent researchers to more easily create the right computing environment for reproduction. Reviewers may also ensure that custom-built software or tools reference their stated purpose in data analysis (e.g., cleaning the data) and the corresponding datasets they were used to analyze.
- **Underlying analysis tools are made readily available**. Several of the teams ran into some variation of the phrase "analysis performed with custom in-house scripts." Peer reviewers can request that those scripts are made available as a condition for publication.

C. Still many different ways to interpret reproducibility

While there is no one standard definition of reproducibility, workshop participants were introduced to the following working definition for reproducible computational research, outlined by Stodden et al:

> Open or reproducible research [is] auditable research made openly available. This comprised well-documented and fully open code and data that are publicly available that would allow one to (a) fully audit the computational procedure, (b) replicate and also independently reproduce the results of the research, and (c) extend the results or apply the method to new problems.

For a variety of reasons—ranging from data and tool availability to simple time restrictions—teams were unable to reproduce their respective papers in a way that satisfies the above definition. As a result, each team lowered their standards for reproduction during the constrained time period of the workshop; however, the lowered bars for reproduction varied from team to team. One group aimed to recreate just the figures using the available processed data, while another aimed to organize the raw data so that they could at least make sense of which datasets underlie what conclusions. While some teams worked to re-engineer scripts and workflows according to their respective paper's methods section, others used tools provided by the author to reproduce their results. One team struggled to recreate the original computing environment needed for the author-provided tool to run, spending the entirety of the workshop trying to install the appropriate versions of the software tools so they were compatible with one another. Given these challenges, some teams concluded that even if they could make progress toward reproducing some aspect of the paper, a paper cannot truly be reproduced without the raw, unprocessed data.

D. Communication for open science

Some groups made an effort to contact the corresponding authors to seek information that was missing from the paper, such as raw data or information about custom in-house scripts. Some authors responded with clarifications on the methods or datasets used in their respective papers. One author agreed to share the data the team was missing, with the caveat that she was still working on another publication and would only share the data if the group agreed not to publish any results before she did. She also pointed out that the group would need more than just the dataset they had requested in order to reproduce the results; had the group not contacted the author, they likely would have unknowingly proceeded without the relevant data and been unable to reproduce the result.

The quick response by authors, within hours of receiving the email, suggests that lack of reproducibility is not the result of bad faith or unwillingness to share on behalf of the authors. Rather, the authors' positive responses suggest that they were willing to share code and data, but that ensuring reproducibility is challenging. However, it should be noted that participants indicated in their emails that they were requesting the information as part of an NIH workshop on reproducibility. Had the emails come from another source in another context, authors may not have responded so quickly or positively.

4. Conclusions

Overall, the workshops reinforced the common perception that reproducibility is not trivial. While teams had varying degrees of success, none were able to fully reproduce their assigned papers, starting from raw underlying data to conclusions. Nevertheless, participants noted the exercise of attempting reproduction was an informative experience. As participants ran into roadblocks, they reflected that similar errors and gaps in documentation or methods communication likely existed in their own publications. During the discussion periods, participants commented on how they plan to apply the tools covered over the course of the workshops to facilitate reproducibility of their own work, even before the publication process.

While the will to publish results in a manner that facilitates reproducibility may exist, there is still a great deal of work to be done developing, disseminating, and, perhaps most importantly, implementing guidelines and best practices for reproducible publication. These workshops uncovered several opportunities to make research results more reproducible. These insights can both inform the peer review process, particularly as more articles rely on computational and data science methods, as well as training curricula for reproducible methods for bioinformatics or computational research.

Appendix A. Workshop Schedules

The schedules for both iterations of the workshop were the same with minor adjustments based on the needs of the workshop participants.

Day 1

Time	Topic
9:00 – 10:00 am	**Introduction to Open Science** *Overview of tools to facilitate best practices and discussion of what open science and reproducibility means*
10:00 – 12:00 pm	**Team introductions and preparations for reproducing** • Discuss a data analysis plan and begin execution • Download relevant datasets • Launch instances on AWS or Biowulf • Install software required for analysis
12:00 – 12:30 pm	Break for lunch (not provided)
12:30 – 1:00 pm	**Breakout Session 1: Jupyter Notebooks** (and working lunch) Instructor: Burke Squires
1:00 – 1:30 pm	**Breakout Session 2: Git and Version Control** Instructor: Keith Hughitt
2:00 – 2:30 pm	**Breakout Session 3: Containerization and Docker** Instructor: Ryan Dale or Steve Tsang
4:00 – 4:30 pm	**Share Out** Short 3-5 minute reports from each team, citing progress made and challenges encountered.

Day 2

Time	Topic
9:00 am – 12:00 pm	Continue working in teams to reproduce papers
12:00 – 1:00 pm	Working Lunch (lunch not provided)
1:00 – 4:00 pm	Continue working in teams to reproduce papers
4:00 – 4:30 pm	**Share Out** Short 3-5 minute reports from each team, citing progress made and challenges encountered.

Day 3

Time	Topic
9:00 am – 12:00 pm	Continue working in teams to reproduce papers
12:00 – 1:00 pm	Working Lunch (lunch not provided)
1:00 – 3:00 pm	Continue working in teams to reproduce papers
2:00 – 2:30 pm	**Discussion of Next Steps** *Opportunities for publication, awards to recognize open science* Instructor: Adam Thomas
2:30 – 5:00 pm	**Share Out and Discussion** Short 3-5 minute reports from each team, citing progress made and challenges encountered.