Reproducible Research Is Really F#$@ing Hard

Patrick D. Schloss†

† To whom correspondence should be addressed: pschloss@umich.edu; Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI

\textbf{Format: Perspective

Counts: ~XXXX words plus XX references, X figures, and a XXX word abstract

Abstract

(150 word limit) Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Introduction

In 1677 Antonie van Leeuwenhoek submitted a letter to the Royal Society, "Concerning little animals" [@Lane2015]. This seminal work described novel observations of microorganisms. The scientific community rejected his observations because they could not be replicated. Leeuwenhoek had little interest in sharing his methods for others. Adding to these problems he wrote in "low Dutch" and his writing was translated to English and significantly edited to shorten the letter. Robert Hooke later developed a compound microscope that was inferior to Leeuwenhoek's single lens microscope, but replicated the earlier findings. In the process, Hooke popularized the compound microscope. Leeuwenhoek and Hooke's experiences are illustrative of many of the current problems microbiologists face in reproducing and replicating each other's work. Of course, Leeuwenhoek's work was rigorous, impactful, and robust. It was not sloppy and there was no fraud. But, it was not reproducible or replicable.

In 2011 Philip Bourne challenged those attending the *Beyond the PDF* workshop to reproduce the analysis performed in his group's 2010 study *The Mycobacterium tuberculosis Drugome and Its Polypharmacological Implications* [@Garijo2013]. The response to that challenge resulted in a unique analysis, which would challenge concepts critical to understanding reproducible research. The investigators demonstrated that the value of reproducibility, the degree to which research should be reproducible, the amount of effort required to reproduce the research, and who should be able to reproduce the research are not simple questions. On first blush, one might argue that any scientist should be able to reproduce another scientist's research with no friction. Few would suggest that Bourne's group was sloppy or that they failed to be transparent. Yet the subsequent attempt to reproduce their study estimated that it would take a novice at least 160 hours to decipher the approaches used in the original analysis and an additional 120 hours to implement them.

These anecdotes are at odds with the tone of a recent report by the American Academy for Microbiology's (AAM's) 2015 colloquium, "Promoting Responsible Scientific Research" and its accompanying editorial in *mBio*. The report is a useful lens into how microbiologists view the reliability of research in their field. The colloquium identified "(i) sloppy science, (ii) selection and experimental bias, and (iii) misconduct" as the primary contributors to the ongoing problems with insuring the reliability of microbiology research. Although the participants were quick to point out that misconduct was a relatively minor contributor to the problem, the XXXX case studies that accompanied the original report all concern misconduct. Missing from these reports is any of the nuance or humility encountered by Bourne and the countless researchers who go out of their way to do good research only to be frustrated when they are contacted about a detail years after a paper is published. Put simply, most problems with reproducibility and replicability are not due to sloppy science, bias,

or misconduct. Although those are certainly issues, the colloquium participants underestimated the difficulty inherent in insuring that one's research design and methods are sufficiently clear.

The goals of this Perspective are three-fold. First, I hope to give a better framework for thinking about how science is conducted within the microbial sciences. Although I will primarily focus on examples from microbiome research, the principles are generalizable to other areas of microbiology. Second, I provide an overview of various factors that threaten the ability to validate prior results and the tools that can be used to overcome these problems. Third, based on these issues, I suggest several exercises that can be used within research groups to motivate discussions of these factors.

Threats to reproducibility

Definitions. One of the struggles in discussing reproducibility, replicability, and the factors that can limit them, is agreeing upon how they should be defined. The AAM report used the term reproducibility where others would use replicability (i.e. the ability to generate the same results after repeating the experiment independently of the first). The most widely differentiation between reproducibility and replicability is that reproducibility is the ability to regenerate a result when given the same dataset and data analysis workflow whereas replicability is the ability to produce a consistent result with an independent experiment asking the same scientific question [@Leek2015]. A similar framework has been proposed in which the same or different system or cohort are used and when the same or different methods are used (Figure 1) [@Whittaker2017]. This second framework highlights attempts to determine whether a result is robust to differences in methods or generalizable to different datasets that may have been collected under different conditions. Aside from issues of sloppiness, bias, and fraud, it is scientifically valuable to consider what factors threaten each of the quadrants in this framework. Whether a result holds up is not just a product of rigorous scientific practice, but also a product of stochastic forces. Furthermore, I emphatically agree with the the AAM report that most research is exploratory and that scientists, editors, and funding agencies generally lack the will or ability to confirm previous studies via independent replications or attempts to generalize results in other model systems or human populations. Finally, just because a result is reproducible or even generalizable does not guarantee that the result is correct. Science is hard and failure to support an earlier observation does not indicate a failure, but a success of the scientific method.

An example. Several research groups, including mine [@Sze2016], have attempted to validate the result that individuals with a lower bacterial diversity and higher ratio of *Bacteroidetes* to *Firmicutes* in their feces

were more likely to be obese [@Knight; @Sharpton]. The original observation was published in 20XX using 16S rRNA gene sequencing and engendered much enthusiasm for the role of the microbiome in human health [@Turnbaugh20XX]. Although the original study was performed using poorly reported data curation methods, we and others were able to independently obtain the same results as the original study when using the same dataset. The original result can thus be considered reproducible by the rubric in Figure 1. However, when we used the same methods with 9 other datasets, we failed to replicate the result. Similarly, other groups have failed to replicate the original result with their own data analysis workflows. This failure to replicate the original result may be due to methodological differences across the replicating studies, differences in study populations, or statistical variation. It is worth noting that those involved in the original Turnbaugh study pursued multiple approaches to better understand the question of whether the microbiota is important in obesity. In the original study they generated multiple datasets from the same cohort that each reflected different regions of the 16S rRNA gene. They also used shotgun metagenomic sequencing to postulate the enrichment of carbohydrate processing genes in obese individuals. Finally, they transferred the feces from members of their human cohort to germ free mice and observed variation in weight. Although each part of their approach had significant weaknesses including methodological biases and underpowered experimental designs, their results support the hypothesis that within their cohort, there were microbial signatures associated with obesity. Their overall conclusion appeared to be robust for this cohort. The inability to replicate these results in other cohorts indicated the conclusions were not generalizable.

Reproducibility. Threats to reproducibility are some of the most fundamental and easiest to lay fault on the original investigators. If a result cannot be reproduced, then it is difficult to have confidence that it can be replicated or generalized. Thus the ability to reproduce a result is critical.

- Because of word limits in many journals, the Methods sections become a chain of citations to previous work that each cite previous work. The resulting rabbit holes can largely be addressed by improved documentation in supplementary materials or archives such as protocols.io for lab-based methods or through GitHub for data analysis workflows. For data analysis workflows, software such as GNU Make and the Common Workflow Language are available that allow one to track data dependencies and automate a workflow. For example, the workflow used in our meta-analysis was written using GNU Make such that one should be able to get a copy of the scripts from the project's GitHub repository and write "make write.paper" from the command line to reproduce our analysis. These tools make it possible to trace the provenance of a summary statistic from the manuscript back to the raw data.

- Access to the raw data behind a result is often not accessible and makes an analysis of a result's

reproducibility impossible. Although well-established databases exist for sequence data, these data are still often missing, lack the necessary metadata, or are only available upon request from the original authors. As we developed the obesity meta-analysis we were dependent on the original authors to provide the information for two of the ten datasets. Furthermore, the data made available from the Turnbaugh et al study only provided the subjects' body mass index (BMI) as categories. The actual heights, weights, and BMIs were not available. Two large datasets were not included in the analysis because their data were practically inaccessible. Two other datasets required at least a month of effort to obtain. Beyond sequence data, other data can be archived in databases including FigShare and Dryad.

- Changes in sequencing technology, data curation, databases, and statistical techniques are quickly rendering the methods used in studies from a few years ago obsolete. For example, the Human Microbiome Project used Roche's 454 platform to sequence the 16S rRNA gene. This sequencing platform is no longer commercially available. Data analysis software and databases are also rapidly changing. The mothur software package has had 40 major updates since it was originally released in 2009. The RDP and SILVA databases that many use as a reference for aligning and classifying 16S rRNA gene sequences are updated annually. With each release they expand the number of sequences in the database and make modifications to their taxonomic outline. For software and databases, it is critical that authors report version numbers if there is to be any hope of replicating previous work. Unfortunately, the reliance on web-based workflows like GenBank, greengenes, RDP, and SILVA preclude the ability to analyze new data with older versions of the sites.

- Science is often falsely portrayed as a linear process resembling a pipeline. In reality, questions change and scientists fall into the traps of the "Garden of Many Forking Paths" where they go looking for a desired result or "P-hacking" where large numbers of statistical hypothesis tests are attempted without adequately correcting for performing multiple tests. Although it is possible to pre-register data analysis plans, these are often too stringent for most exploratory research. Alternatives include making research notebooks publicly available using commercial platforms or tools such as RMarkdown documents and Jupyter notebooks. Combined with version control software such as git, these literate programming documents can allow a researcher to document the evolution of their analysis.

- A persistent problem with many research articles is the problem of "link rot" where a web or email address will be deprecated. Someone trying to contact me regarding work I did while at a prior institution would receive an error message if they used the email address associated with those

manuscripts. Furthermore, the URLs in papers describing software written in 2005 are no longer functioning. To solve the email rot problem, ORCID has emerged as a technology used by many journals to provide a persistent link between an individual's many scientific identities over their career. For link rot, services like Zotero can provide a digital object identifier (DOI) that persists even if the link that it points to changes.

Replicability. Failure to replicate a previous result could be due to an extensive number of factors that are due to threats similar to those for reproducibility. In addition there are threats related to differences in systems or populations and the ability to control for those differences.

- There is tremendous inter-strain and -population variation that can hinder efforts to replicate results. In microbiome research, it is widely appreciated that the microbiota of research animals from the same litter and breeding facility are largely clonal and distinct from other facilities. Mice from two breeding facilities at the same institution may have completely different microbiota. The best example of this phenomenon is the presence of segmented filamentous bacteria in mice purchased from Taconic, but not Jackson Laboratories. Thus, the roles ascribed to the microbiota may be confounded by the origin of the mice and not the experimental treatment. This is particularly a problem for genetic models when researchers obtain mutant animals and animals with the wild type background as their control. In such cases using the offspring of heterozygous matings is critical. Similarly, comparing the microbiota of obese and lean individuals from a cohort of twins and their mothers in Missouri may have confounding factors that differ from members of Amish communities. In these cases, the problem with replicability is not due to the quality of the investigator's experimental practices, but because of possible biological, demographic, or anthropological differences. Instead of being cause for a crisis, failures to replicate a study across different cohorts could suggest that there are other interesting factors that underly the failure to replicate.

- Uncertain provenance and purity of reagents, organisms, and samples also threaten replicability. Perhaps the best known example is the discovery that HeLa cells contaminate many other cell lines, generally from the same laboratory. Similarly, investigators frequently realize that they are working with bacterial strains that were incorrectly typed or that have evolved during serial passages from the freezer stock. Short of resequencing the cells, experimental controls, limiting the number of passages from freezer stocks, and periodic phenotyping of the strains can help to overcome these problems. In the microbiome literature, there is a growing awareness that DNA extraction kits can be contaminated with low levels of bacterial DNA. These contaminants can lead to the identification of contaminants as

being important members of the lung and placental microbiota if mock extractions are not sequenced in parallel.

- In biology, a replication may fail because replication is statistical rather than deterministic. Every experiment has a margin of error and when the effect size is near that margin of error, it is likely that a statistically significant result in one replicate will not be significant in another. Most researchers use a frequentist null model hypothesis testing approach where they are willing to accept a Type I error of 0.05. More colloquially, they are willing to incorrectly reject a null hypothesis in 5% of the replicates. Because of biases correctly described in the AAM report, including the "file drawer effect" and a general reluctance to attempt replications, it is difficult to know the degree to which scientists are succumbing to the problem of the Garden of Many Forking Paths or P-Hacking. Equally troubling, scientists rarely quantify the risk they are willing to accept of falsely accepting a null hypothesis (i.e. Type II errors). In our analysis of the microbiota associated with human obesity, we observed that nearly all studies were underpowered to detect 5 or 10% differences in diversity. In some cases, failure to replicate a study may be because the replicate study did not have a sufficient sample size. In other cases, it may be that the original study was underpowered rendering it susceptible to an inflated risk of Type I errors. Solutions to these problems include authors pre-registering their data analysis plans, justifying sample sizes based on on power calculations, and using Bayesian frameworks that allow the interpretation of new results to be influenced by prior knowledge of the system.

Robustness. Every method has its own strengths and weaknesses. Therefore, it is important to address a research question from multiple and hopefully orthogonal directions. With this strategy the strengths of different methods combine to overcome their individual weaknesses. Evaluating the robustness of a result from a single cohort is becoming more common as researchers pursue multiple approaches where different approaches including 16S rRNA, metagenomics, metatranscriptomics, and metabolomics. Of course, biases in the underlying cohort design, sample collection and storage, or the nucleic acid processing will propagate through the analyses. To remedy this, the methods need to be as independent from each other as possible. For example, sequencing multiple regions of the 16S rRNA gene would not be considered truly independent datasets since the same general method would be applied to the same samples. Layering shotgun metagenomic data onto the 16S rRNA gene sequence results would be marginally more independent because although the same DNA would be used for sequencing, the method provides information about the genetic diversity and functional potential of a community rather than the taxonomic diversity of a community. Metabolomic data would be even more independent from the DNA-based methods since completely different sample handling and processing steps would be needed. Quantitative PCR, cultivation, and microscopy

could be similarly layered on these data. As this discussion illustrates, it is impossible for the results of each set of methods to be fully independent.

Generalizability. The gold standard of science is to have a result that is generalizable across populations. Failing to attempt replication studies hinders the ability of researchers to test the generalizability of most results. Being "scooped" is often seen as the worst thing that can happen to a scientist. In reality, it affords the second researcher the opportunity to increase the field's confidence that a result is robust or generalizable. In addition, model organisms (e.g. *E. coli*) and strains of those organisms (e.g. K-12) have taught us a great deal about the biology of those organisms. However, it is not always trivial to generalize that knowledge to related species and strains or from *in vitro* to *in vivo* conditions. Again, rather than seeing the failure to generalize a result as a failure of science, it should instead be seen as an opportunity to better understand the complex biology of bacteria and how they interact with their environments.

Need for training

Motivation. A key observation from the work of Garijo and colleagues was that the level of detail needed to reproduce an analysis varies depending on the researcher's level of training. An expert in the field understands the nuances and standards of the field whereas a novice may not know how to install the software. This highlights the need for training. Many microbiology training programs focus on laboratory skills while ignoring the skills needed for data analysis. A number of excellent training programs have emerged in recent years. I have created the Riffomonas project, which expounds on the threats to reproducibility and tools that microbiologists can use to maximize the reproducibility of their analyses (http://www.riffomonas.org). In addition, organizations including Software Carpentry and Data Carpentry offer workshops to introduce researchers to the best practices in reproducible research. Massively open online courses (MOOCs) are also available that teach scientists best practices for performing reproducible analyses. The most popular of these is a training program from faculty in the Department of Biostatistics at the Johns Hopkins University. Just as a novice could not reproduce Beethoven's "Für Elise" from sheet music without prior experience playing the piano, a novice cannot expect to reproduce a complex experiment and analysis without learning the methods of their discipline.

Exercises. The following exercises are meant to motivate conversations within a research group to foster a culture improving reproducibility and replicability and underscore the threats outlined above.

1. Working away from each other, have two or more people to write instructions on how to fold a piece of

paper into an airplane. Once the instructions have been written, have the participants trade instructions and implement the instructions while working away from each other. How closely did the airplanes folded from the instructions resemble the first? What would have helped to make the reproductions more faithful? How much did the author of the instructions assume about the second person's prior knowledge of paper airplanes? What resources or abilities were assumed? What challenges would one face if they were limited by the length of the instructions? How does this exercise resemble the descriptions in the Materials and Methods section of papers for standard methods (e.g. PCR) and for novel methods (e.g. bioinformatic workflows)?

2. A graduate student was really excited to see an analysis that you performed in your most recent paper because the would like to reproduce it with their data. Before using their data, however, the want to make sure that they get the same results as you. What steps are likely to cause them problems? Take a figure from your recent paper and improve the likelihood that a third party would be able to reproduce it. Where are the data and how would they get them? What calculations were performed to summarize the data? What software was used to generate the figure? Is that software freely available? What steps need to be taken to generate the figure? When you are confident that you have made the figure as reproducible as you can, give the instructions to a colleague and ask for their feedback. Find your favorite figure from your favorite paper from a different research group. Can you reproduce the figure? What is standing in your way?

3. Many of the threats to reproducibility and replicability are a product of scientific culture: methods sections are terse or vague, original data are not available, analyses rely on expensive and proprietary software, analysis scripts are available "upon request from the authors", papers are published behind pay-walls. Complete an audit of the reproducibility practices in your own research group. Have a discussion within your group about why you do things this way, whether your practices should change, and what would be the easiest to change. For your next paper, work improving one element of reproducibility. Develop an ethic of striving towards greater reproducibility.

Conclusion

A motivating concept to improving the reproducibility of one's research is that the most important collaborator is themselves six months ago, and old them does not have email access. Their second most important collaborator is the director of their research group. The reality is that most research is repeated multiple

times within a research group prior to and after publication. Anyone that has done research can attest to how difficult it can be to satisfying their two most important “collaborators”. If a scientist does not provide sufficient transparency that they and their lab can reproduce a result, then it is unlikely that any one else can. It is important to see that attempts to guard against threats to reproducibility, replicability, robustness, and generalizability are positive forces for improving science. Such attempts have been considered a form of “preventative medicine”. Although guarding against these threats is not a guarantee that the correct conclusion will be reached, the likelihood that the result is correct will be increased. Before slashing at our fellow scientists as being sloppy, biased, or untrustworthy, it is worth seriously considering the many factors - biological, statistical, and sociological - that lead to the failure to yield a similar result. Although there is much room for improvement, we must acknowledge that science is a process of learning and that it is really f#$@ing hard.

Acknowledgements

This work was supported in part by funding from the National Institutes of Health (5R25GM116149).

References