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

Format: ??????

http://mbio.asm.org/site/misc/authors.xhtml * Commentaries are short invited articles (1,000 words) that discuss mBio papers or issues of special interest. * Perspectives are brief reviews (2,000 words) on a topic in which opinion and synthesis are encouraged. * Minireviews should be approximately 6,000 words maximum (with up to two figures or tables). * Opinions/Hypotheses should be approximately 2,500 words maximum

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

Abstract

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 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 the ability to reproduce another group's work when using the same data and methods (i.e. reproducibility). The investigators demonstrated that the value of reproducibility, the necessary degree to which research can be reproduced, 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. As Garijo and colleagues went on to demonstrate, even when a project led by leaders in the area of reproducible research, reproducibility is a thorny issue.

Their conclusion is in contrast to the report by the American Academy for Microbiology's (AAM's) 2015 colloquium, "Promoting Responsible Scientific Research" and its accompanying editorial in *mBio*. Although the report from this colloquium 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 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 four case studies that accompanied the original report all concern misconduct. In addition, problems related to "sloppy science" are also prominent throughout the reports as well. Missing from these reports is any of the nuance or humility found in the earlier Garijo study. Few would suggest that Bourne's group was sloppy, yet the follow up reproduction analysis estimated that it would take a novice at least 280 hours to reproduce the work with at least 160 hours of that time spent trying to decipher the approach used in the original analysis.

When I ask other microbiologists how difficult it would be for me to reproduce a randomly selected figure from their latest paper, universally, they turn ashen. Does the inability to reproduce that figure indicate that their work is incorrect? No. However, the inability to reproduce a value significantly reduces the confidence that one might have in the work. In contrast, one might also ask, does the ability to reproduce a figure indicate that the work is correct? No. Investigators may release all of their data and the workflow used to generate a figure, but a coding error may result in accidentally switching the colors of the legend. Alas, the ability to interrogate the code, would quickly identify the error and allow the investigators to resolve the issue. Finally,

we might also ask whether research that is not reproducible is replicable. Again, if the methods used to run the study and analyze the data are not transparent, then it would be impossible to confidently understand why the inability to replicate a result was due to the use of a different study population, cell line, or analytical approach. Put simply, most problems with reproducibility and replicability of a study are not due to sloppy science, bias, or misconduct. While those are certainly issues, leaders in microbiology clearly underestimate the difficulty inherent in insuring that one's research design and methods are sufficiently clear.

The goals of this mini-review 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 as well. 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 case studies that can be used within research groups to motivate discussions of problems with reproducibility.

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 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 a 2x2 grid is created when the same or different data are used and when the same or different data analysis workflow is 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 are threats to the reproducibility, replicability, robustness, and generalizability of a result. Whether a result holds up is not just a product of poor scientific practice, but also a product of stochastic and deterministic forces. Furthermore, we must acknowledge 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 measure the replicability of 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 data curation methods that were not well described, we and others were all able to obtain the same results as the original study when using the same dataset, but with different methods. 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. The original result is not replicable to other populations. This failure to replicate the result may be due to methodological differences across the replicating studies, differences in study populations, or statistical variation. Furthermore, this failure to replicate across human populations and studies also indicates the failure to generalize earlier results using murine models of obesity to humans. This example, exemplifies the importance of making datasets available, providing explicit methods descriptions, and asking the same research question in different experimental systems and populations.

Threats. A useful discussion that research groups should have is to think about the threats to reproducibility, replication, robustness, and generalizability within the scientific domain of their research group and what mechanisms are available to overcome these threats.

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.

- It is clear to see that Results sections of papers have grown at a much faster rate than the Methods section. Because of word limits in many journals, the Methods sections become a chain of citations to previous work that each cite previous work. In some cases the previous methods clearly contradict what was done in the study in question. These methods rabbit holes can largely be addressed by improved documentation either 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 that we used to analyze the ten studies in the Sze obesity 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. By tracing the workflow through these tools, it is possible to trace the provenance of a summary statistic in a manuscript back to the raw data.

- Access to the raw data behind a summary statistic is often not publicly available and makes an analysis of a result's reproducibility impossible. Although well-established databases exist for sequence-based data, these data are often missing, lack the necessary metadata to make the data useful, or are only available upon request from the original authors after going through significant and unnecessary bureaucratic layers. As we developed the obesity meta-analysis we found that the sequence data and metadata for X of the 10 studies were publicly available; we were dependent on the original authors to provide the information for the other datasets. For example, the data made available from the Turnbaugh et al study only provides the subjects' body mass index (BMI) as categories of lean, overweight, and obese. The actual heights, weights, and BMIs are not available. In addition, two large datasets were not included in the analysis because their data were practically inaccessible. Beyond sequence data other raw data can be archived in databases including FigShare and Dryad. These databases are free and the data will persist indefinitely.

- Microbiology and microbiome research, in particular, are dependent on rapidly developing methods and information. 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 seminal Human Microbiome Project used Roche's 454 platform to sequence regions of 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, that my lab develops has had 39 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 and with each release they expand the number of sequences in the database and make modifications to the taxonomic outline that is employed. For software and databases, it is critical that authors report the version number that they are using if there is to be any hope of replicating previous work. Unfortunately, the reliance of some on web-based workflows like the greengenes, RDP, and SILVA websites preclude the ability to analyze new data with older versions of the database. In each of these cases the cause for new releases is an improvement of our knowledge of microbiology. Although it would be ideal to regenerate the original results with the exact copy of the analysis workflow, it would probably be better to obtain the original data and re-analyze it with more modern workflows.

- Science is often portrayed as a linear process resembling a pipeline. In reality, it is more like a bush

that has been carefully sculpted. This is an important distinction because depending on how much pruning is done a research risks falling into the trap 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. This can be done through a number of software tools. Two popular tools for data analysis include RMarkdown documents and Jupyter notebooks. Combined with version control software such as git, these literate programming documents can allow a researcher to properly depict their research as a bush-like structure that others can migrate through later. Furthermore, by making the data analysis scripts publicly available in these documents on a website like GitHub, a researcher overcomes the need for others to contact the original authors for the scripts, which is another problem that plagues much data analysis in microbiology.

- When a researcher's best efforts to make their work transparent and reproducible fall short, there is still the opportunity for a later researcher to contact the original. A persistent problem with many research articles is the problem of "link rot" where a web or email address will be deprecated. I have been a corresponding author on papers while at two institutions. Someone trying to contact me regarding work I did at my 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 while I was a postdoctoral researcher in 2005 are no longer functioning (although the software is now available elsewhere online). Both types of electronic rot can be remedied through the use of persistent identifiers. 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.

- It is a mark of reproducibility that numerous laboratories have been able to replicate the Turnbaugh et al. results using the same dataset, but with different data curation methods. Such a result indicates that the observation, for that dataset, was robust to a variety of approaches. The original study sequenced different regions of the 16S rRNA gene from the same subjects and reached similar conclusions. Again, the result was robust to the variations in bias that are imposed by using different PCR primers. All this being said, there is value in benchmarking methods to characterize their strengths, weaknesses, and biases. If a poor method was used to analyze the Turnbaugh dataset and yielded a different result, then that should reflect more on the lack of validity of the method rather than the lack of robustness

of the original study. It is likely that some results are biologically valid, but marginal and require more sensitive methods. Within the microbiome research field, questions surrounding the sequencing error rates, chimera filtering methods, and clustering methods could each vary and impact the ability to yield the same result as the most rigorous methods.

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.

- In microbiome research where animal models are used to study a disease, it is widely appreciated that the microbiota of animals from the same litter and breeding facility are largely clonal. A C57Bl/6 mouse from two breeding facility 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, it would not be surprising to suggest different roles for the microbiota depending on the origin of the mice. Thus, unless one controls for the initial microbiota, the observed differences could be due to the differences in the starting microbiota rather than the specific mutation. This is particularly a problem for genetic models when researchers obtain mutant animals and animals with the wild type background as their control. The observed differences could be due more to the animals environmental exposure than due to the specific differences in their microbiota. 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 between comparing obese and lean members of Amish communities or between the general population of St. Louis, MO or Houston, TX. 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 is interesting biology underlying the differences in the results.

- Uncertain provenance and purity of reagents, organisms, and samples can also be a threat against replicability. Many consider it would be too onerous to resequence strains and cell lines to confirm their identity and purity. Perhaps the best known example is the discovery that HeLa cells contaminate many other cell lines, generally from the same laboratory. There are also known cases of investigators working with bacterial strains that turn out to be a different strain 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 (i.e. the "kitome"). Although this contamination would have a minimal impact on studies of high biomass samples (e.g. soil and feces), these contaminants can lead to the identification of contaminants as being important members of the lung and placental microbiota.

- An undervalued threat to replicability is that there are stochastic factors that give rise to differences in observations. The results of two replicates that attempt to match methods and cohorts may differ because replication is statistical rather than deterministic. Every experiment has a margin of error and in cases where 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. In layman's terms, they are willing to incorrectly reject a null hypothesis in 5% of the replicates. Because of biases 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 inadvertently p-hacking. Conversely, 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 sizes. 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 where the interpretation of new results is impacted by prior knowledge of the system. Of course, each of these suggestions may significantly hamper the rate of discovery in exploratory and basic research.

Robustness. Every method has its own strengths and weaknesses. To counter these biases, it is important to address a research question from multiple and hopefully orthogonal directions. Through a multi-pronged approach, it is possible to combine the strengths of each method to overcome their individual weaknesses.

Returning to the Turnbaugh et al obesity example, it is worth noting that the group pursued multiple approaches to better understand the question of whether the microbiota is important in obesity. In one study they generated multiple datasets from the same cohort that each reflected different regions of the 16S rRNA gene. This allowed them to better understand the relationships between diversity and bacterial taxa with obesity. In that study and then expanded in another study they 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 gain. Although each part of their approach had significant weaknesses including biases in methods and underpowered experimental designs, together they support the hypothesis that within the cohort under study, the microbiota had a role in the participants' obesity status.

Evaluating the robustness of a result from a single cohort is becoming more common as researchers pursue a multi-omics approach whereby, different approaches including 16S rRNA, metagenomics, metatranscriptomics, and metabolomics can be used to strengthen the claims that are reached by the other methods. Of course, if there are biases in the underlying cohort design, sample collection and storage, or the nucleic acid processing those biases will propagate through the analyses. To remedy this, it is important for the methods to be as independent from each other as possible. For example, sequencing the V3, V4, and V6 regions of the 16S rRNA gene separately would not be considered truly independent datasets since the same DNA would be used, PCR would be expected to have similar types of biases across the regions, and the data analysis pipelines would not be meaningfully different. 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. Other methods including 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, but they can be carefully selected to overcome the weaknesses of the other methods.

Generalizability. The gold standard of science is to have a result that is generalizable across populations. Alas, the lack of willingness to replicate previous 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 junior scientist. In reality, it affords the second researcher to increase the field's confidence that a result is generalizable, which provides the field a great service. 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.

As was exemplified in the AAM Report and mBio Editorial, the genre of "Reproducible Research" scholarship focuses on threats to reproducibility, replicability, robustness, and generalizability [@refs]. It is important to see that attempts to guard against these threats is a positive force for doing better science. In fact, just as changing someone's diet to be more healthy will improve their health and hopefully prevent future problems, using reproducible research practices is considered as 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. Furthermore, if there is a problem in how the experiments and analyses were performed, it will be easier to identify the problem and possibly correct them. Although much of Reproducible Research scholarship focuses on the ability of an independent researcher to get the same result as the original researcher, the most important person that needs to get the same result is the original researcher. A motivating concept to improving the reproducibility of one's research is that your most important collaborator is you from six months ago, and old you does not have email access. I would also add that your second most important collaborator is the director of the research group after you have left the lab. The reality is that most research is repeated multiple times within a research group prior to and after publication. 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.

Need for training

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 even get started. 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 that microbiologists can make use of to improve their ability to perform more reproducible data analyses. I have created the Riffomonas project, which contains a 14 module series of lessons targeted to microbiologists on the importance of reproducible research and how to maximize the reproducibility of one's data analysis (http://www.riffomonas.org). In addition, organizations including Software Carpentry and Data Carpentry offer workshops to introduce all researchers to the best practices in reproducible research. Finally, several massively open, online courses (MOOCs) have been developed to teach scientists best practices for data analysis in a reproducible manner. 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 data analysis without learning the basics of data analysis. As all microbiologists begin to analyze larger datasets, it is essential that we seriously consider the training they receive to analyze those data.

Case studies to spawn discussion

As I have outlined in this Mini-Review, the issues addressed by Reproducible Research scholarship are complex. Many of the threats to these aspects of research are practical, but many are also cultural. The following case studies are meant to motivate conversations within a research group on their own research practices and culture of fostering reproducibility.

1. Working away from each other, get 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 my 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 do 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, analysis relies 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

In 1677 Antonie van Leeuwenhoek submitted a letter to the Royal Society, “Concerning little animals” [@Lane2015]. This seminal work in microbiology described microscopic observations of bacteria that had never been previously seen. There was pushback from the scientific community because initially no one could replicate his work and Leeuwenhoek had little interest in sharing his methods with others. Adding to these problems he wrote in “low Dutch” and his writing was translated to English and significantly edited to shorten the letter. Several years later, Robert Hooke developed a compound microscope that was inferior to Leeuwenhoek’s single lens microscope, but was able to replicate the earlier findings. In the process, Hooke popularized the compound microscope. This anecdote is illustrative of many of the current problems in current problems microbiologists face in reproducing and replicating each other’s work. There was a lack of transparency that held microbiologists back. Of course, Leeuwenhoek’s work was rigorous, impactful, and robust. It was not sloppy and there was no fraud. But it was not reproducible. There is much to be learned by striving for transparency in our research in terms of making our original data available, describing our methods, and publishing our results.

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.

References

- https://simplystatistics.org/2017/11/21/rr-sress/
- http://mbio.asm.org/content/7/4/e01256-16.full?sid=e0608d60-2133-4d24-a00c-54579527c544
- https://www.asm.org/images/Colloquia-report/Promoting_Responsible_Scientific_Research.pdf