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: Perspective

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

Abstract

Microbiologists have long struggled to make their research reproducible. These struggles are not particular to microbiology and have lead many to speak of a “reproducibility crisis” in science. We need to respect that science and microbiology, in particular are difficult. There are many reasons why a result may not be reproducible or replicable and even if a result is reproducible it may not be correct. Here I lay out a framework that describes the reproducibility, replicability, robustness, and generalizability of a particular result. I then describe factors that can threaten this framework and approaches microbiologists can take to overcome the threats. Finally, I provide several provocative exercises for individuals and research groups to discuss to gain a better appreciation for their own research practices. There is much to be learned from failures to validate previous results that have much to teach us about the scientific process and microbial life itself.

Keywords: Reproducibility, Microbiome, Scientific method, Research ethics, American Academy of Microbiology

Introduction

In 1677 Antonie van Leeuwenhoek submitted a letter to the Royal Society, “Concerning little animals” (1). 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 with others. Adding to these problems he wrote in “low Dutch” and his writing was translated to English and significantly edited. 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* (2). The response to that challenge resulted in a unique analysis that challenged 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.

Both 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* (3, 4). 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. Missing from these reports was 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 forgotten 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) (5). The most widely used definitions are 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 (5). 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 (Table 1) (8). 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 (6). Furthermore, I emphatically agree 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 (3, 4, 9). Finally, just because a result is reproducible or even generalizable does not guarantee that the result is correct (5). 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 (10), have attempted to validate the result that obese individuals were more likely to have lower bacterial diversity and higher abundances of *Bacteroidetes* and

lower abundances of *Firmicutes* in their feces (11, 12). The original observation was published in 2008 using 16S rRNA gene sequencing and engendered much enthusiasm for the role of the microbiome in human health (13). 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 key in Table 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. They initially sought microbiome-based signatures using mouse models (14). They observed stark differences in the microbiota of genetically lean and obese mice and that the microbiota of obese mice could transmit the propensity to gain weight to germ free mice (14). In a human cohort, they generated multiple datasets that each reflected different regions of the 16S rRNA gene. In obese individuals, they observed reduced diversity and relative abundance of *Bacteroidetes* (13). They also used shotgun metagenomic sequencing to postulate the enrichment of carbohydrate processing genes in obese individuals (13). In a smaller cohort study, although the subjects' diversity remained constant, there was the predicted increase in *Bacteroidetes* as subjects lost weight (16). Although each part of their approach had significant weaknesses including methodological biases and underpowered experimental designs, their results supported the hypothesis that there were microbial signatures associated with obesity. Their overall conclusion appears to have been robust within 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, Materials and Methods sections become a chain of citations to previous work that each cite previous work (9). The resulting rabbit holes can largely be addressed by improved documentation in supplementary materials or archives such as protocols.io (https://www.protocols.io) for lab-based methods or through GitHub (https://github.com) for data analysis workflows. For data analysis workflows, software such as GNU Make (https://www.gnu.org/software/make/) and the Common Workflow Language (17) are available that allow one to track data dependencies and automate a workflow. For example, the workflow used in

our obesity meta-analysis was written using GNU Make such that one should be able to download 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.

- Problems with experimental design are often a threat to reproducibility because investigators fail to account for confounding variables. In sequence-based analyses, threats to reproducibility are encountered when samples are not randomized across sequencing runs. These so-called batch effects have been a problem with a large number of analytical techniques beyond sequencing (18). One notable example occurred within the Human Microbiome Project where 150 people were recruited in Houston, TX and 150 from St. Louis, MO (19). DNA extractions for the two sets of subjects were performed at Baylor College of Medicine and Washington University, respectively. The DNA from the Houston subjects were then sequenced at Baylor College of Medicine, the J. Craig Venter Institute, and the Broad Institute and the DNA from the St. Louis subject were sequenced at Washington University. The variable with the largest effect size was the subject's city, although all parties used the same standard operating procedures (19, 20). Because the city of origin and the center that did the extractions were perfectly confounded, it was impossible to quantify the impact of regional differences on the microbiome. Instead of being a single study, this became two replicate studies.

- Access to the raw data behind a result is often not accessible and makes an analysis of a result's reproducibility impossible (21, 22). 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. (13) study only provided the subjects' body mass index (BMI) as categories. The actual heights, weights, and BMIs were not available. Three large datasets from two studies were not included in the analysis because their data were practically inaccessible due to onerous data sharing agreements (23, 24). Two other datasets required at least a month of effort to obtain (19, 25). Beyond sequence data, other data can be archived in databases including FigShare (https://figshare.com) and Dryad (https://datadryad.org).

- 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 (19). 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 (26). The RDP [(27); http://rdp.cme.msu.edu] and SILVA [(28); https://www.arb-silva.de] databases that many use as a reference for aligning and classifying 16S rRNA gene sequences are updated annually and the popular greengenes database files have not been updated since 2013 [(29); http://greengenes.lbl.gov and http://greengenes.secondgenome.com]. With each release, curators 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 (https://www.ncbi.nlm.nih.gov/genbank), greengenes, RDP, and SILVA preclude the ability to analyze new data with older versions of the sites. The problem with web-based workflows is exemplified by the greengenes website, which removed their online tools in April 2017. Their database files, but not tools, are now available through the company, Second Genome.

- 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 (30) or "P-hacking" where large numbers of statistical hypothesis tests are attempted without adequately correcting for performing multiple tests (31). Although it is possible to pre-register data analysis plans (32–34), these are often too stringent for most exploratory research. Alternatives include making research notebooks publicly available using commercial platforms or free tools such as RMarkdown documents (35) and Jupyter notebooks (36). 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 (37). 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 (https://orcid.org) 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 (https://www.zotero.org) 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 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 (15, 38). 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 Farms, but not Jackson Laboratories (39, 40). 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 (41). Similarly, comparing the microbiota of obese and lean individuals from a cohort of twins and their mothers in Missouri (13) may have confounding factors that differ from members of Amish communities (25). 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 (42, 43). 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 (44, 45). 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 (46). These contaminants have led to the identification of contaminants as being important members of the lung and placental microbiota if mock extractions are not sequenced in parallel (47–49).

- A replication may fail because replication is statistical rather than deterministic (6). 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. Stated

more colloquially, they are willing to incorrectly reject a null hypothesis in 5% of the replicates. Scientists also rarely quantify the risk they are willing to accept of falsely accepting a null hypothesis (i.e. Type II errors) (50). 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 (10). 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 (51). Solutions to these problems include authors pre-registering their data analysis plans (32–34), justifying sample sizes based on on power calculations (9, 10, 50), and using Bayesian frameworks that allow the interpretation of new results to be influenced by prior knowledge of the system (52, 53).

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 (54). 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 (55–57). 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 (58). In reality, it affords the second researcher the opportunity to increase the field’s confidence that a result is robust or generalizable (59). 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 and on to human subjects (60, 61). 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 (2) 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 "best practices" documents have emerged in recent years (62–67). 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 (68). 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 at the Johns Hopkins Data Science Lab (http://jhudatascience.org). 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. Have the participants trade instructions and implement the instructions while working away from each other. How closely did the final airplanes resemble that of the person who developed the instructions? 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. Table 2 provides a rubric that someone working within the host-associated microbiome field might use to assess their research. 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 within this rubric and constantly be developing an ethic of fostering 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” (5). 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).

Table 1. Simple grid-based system for defining concepts that can be used to describe the validity of a result. This is a generalization of the approach used by Whitaker (8) who used it to describe computational analyses.

	Same Experimental System	**Different Experimental System**
Same Methods	Reproducibility	Replicability
Different Methods	Robustness	Generalizability

Table 2. An aspirational rubric for evaluating the practices host-associated microbiome researchers might use to increase the reproducibility and replicability of their work. Although many of the questions can be thought of as having a yes or no answer, a better approach would be to see the questions as being open ended with the real question being, “What can I do to improve the status of my project on this point?”. With this in mind, a researcher is unlikely to have a project that satisfies the “Best” column for each line of the table. Researchers are encouraged to adapt the categories to modify the categories to suit their own needs.

Practice	Good	Better	Best
Handling of confounding variables	Prior to generating data, did I identified a list of possible confounding variables - biological and technical - that may obscure the interpretation of my results?	In my manuscript do I indicate the level of randomization and experimental blocking that I performed to minimize the effect of the confounding variables (i.e. batch effects)?	Does the interpretation of my results limit itself to only those variables that are not obviously confounded?
Experimental design considerations	Do I have an active collaboration with a statistician who helps with experimental design and analysis?	Do I indicate the number of hypothesis tests I performed and have I corrected any P-values for multiple comparisons?	For my primary research questions, have I run a power analysis to determine the necessary sample size?
Data analysis plan	Before starting an analysis, have I articulated a set of primary and secondary research questions	Has someone else reviewed my data analysis plan prior to analyzing the data?	Have I registered my data analysis plan with a third party before starting the project?
Provenance of reagents	Does my manuscript have a table of all cell lines, strains, genotypes, and primer sequences used in the study?	Where possible, have I obtained reagents from certified entities like the American Type Culture Collection (ATCC)?	Is there a statement in the manuscript regarding how I know the provenance and purity of each cell line and strain used in my study?
Controlling for initial microbiota	Are mice obtained from a breeding facility that allows me to track their pedigree?	Prior to an experiment, are mice co-housed to control for differences in initial microbiota? Where possible, are mice from different treatment groups co-housed?	Are comparisons between mice with different genotypes made using mice that are the result of matings between animals that are heterozygous for that genotype?
Clarity of methods descriptions	Are all methods, databases, and software tools cited in the manuscript? Do I follow the relevant licensing requirements of each tool?	Do I indicate dates and version numbers of websites that were used to obtain data, code, and other third party resources?	Are detailed methods registered on a website like protocols.io or GitHub?
Sex as confounding variables	Does my manuscript indicate the sex of research animals and participants?	Do I provide a justification for the lack of even representation?	Do I have even representation of male and females in my studies? Do I account for sex as a variable?
DNA contamination	Did I quantify the background DNA concentration in my reagents? Did I sequence an extraction control?	Am I taking steps to minimize reagent contamination?	What methods do I take to confirm a result that a sequencing result may be clouded by contaminating DNA?
Availability of data products	Is all of the raw data used in the manuscript publicly available?	Are intermediate and final data files publicly available?	Are tools like Amazon Machine Images (AMIs) used to make a snapshot of my working directory?
Availability of metadata	Does the manuscript include the metadata necessary to repeat any analyses I performed?	Have I adhered to standards in releasing a minimal amount of metadata about my samples?	Did I go beyond the minimum to incorporate other pieces of metadata that will inform future studies?
Data analysis organization	Are all data, code, results, and documentation housed within a monophyletic folder structure?	Is this project contained within a single directory on my computer and does it separate my raw and processed data, code, documentation, and results?	Is this folder structure under version control? Is the project's repository publicly available? Are there assurances that this repository will remain accessible?
Availability of data analysis tools	Are free tools used in preference to proprietary commercial tools?	Is the computer code required to run analyses available through a service like GitHub?	Are Amazon Machine Images or Docker containers used to allow recreation of my work environment?
Documentation of data analysis workflow	Is my code well documented? Do I use a self-commenting coding practice?	Do each of my scripts have a header indicating the inputs, outputs, and dependencies? Is it documented how files relate to each other?	Are automated workflow tools like GNU Make and CommonWL used to convert raw data into final tables, figures, and summary statistics?
Use of random number generator	Do I know whether any of the steps in my data analysis workflow depend on the use of a random number generator?	For analyses that utilize a random number generator, have I noted the underlying random seed?	Have I repeated my analysis with multiple seeds to show that the results are insensitive to the choice of the seed?
Defensive data analysis	Is my data analysis pipeline flexible enough to add new data?	Does my code include tests to confirm that it does what I think it does?	Do I make use of automated tests and continuous integration tools to insure internal reproducibility?
Insuring short and longterm reproducibility	Did I release the underlying code and new data at the time of submitting a paper with their DOIs and accession numbers?	Did I include a reproducibility statement or declaration at the end of my paper(s)? Are ORCID identifiers provided for all authors?	What mechanisms are in place to insure my project remain accessible and reproducible in 5 years?
Open science to foster reproducibility	Have I released any embargoes on my code repository and raw data prior to submitting the manuscript?	Did I post a preprint version of my manuscript prior to official submission to a journal?	Have I published my manuscript under a Creative Commons license? Is a permissive reuse license posted with my code
Transparency of data analysis	Is it clear where one would go to find the data and processing steps behind any of my figures?	Are electronic notebooks publicly accessible and accompany the manuscript?	Were literate programming tools used so that summary statistics, tables, and figures are generated directly from the data?

References

1. **Lane N**. 2015. The unseen world: Reflections on leeuwenhoek (1677) Concerning little animals. Philosophical Transactions of the Royal Society B: Biological Sciences **370**:20140344–20140344. doi:10.1098/rstb.2014.0344.

2. **Garijo D**, **Kinnings S**, **Xie L**, **Xie L**, **Zhang Y**, **Bourne PE**, **Gil Y**. 2013. Quantifying reproducibility in computational biology: The case of the tuberculosis drugome. PLOS ONE **8**:e80278. doi:10.1371/journal.pone.0080278.

3. **Casadevall A**, **Ellis LM**, **Davies EW**, **McFall-Ngai M**, **Fang FC**. 2016. A framework for improving the quality of research in the biological sciences. mBio **7**:e01256–16. doi:10.1128/mbio.01256-16.

4. **Davies EW**, **Edwards DD**, **Casadevall A**, **Ellis LM**, **Fang FC**, **McFall-Ngai M**. 2016. Promoting responsible scientic research. American Society for Microbiology. http://www.asmscience.org/content/colloquia.54.

5. **Leek JT**, **Peng RD**. 2015. Reproducible research can still be wrong: Adopting a prevention approach. Proceedings of the National Academy of Sciences **112**:1645–1646. doi:10.1073/pnas.1421412111.

6. **Patil P**, **Peng RD**, **Leek JT**. 2016. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. Perspectives on Psychological Science **11**:539–544. doi:10.1177/1745691616646366.

7. **Goodman SN**, **Fanelli D**, **Ioannidis JPA**. 2016. What does research reproducibility mean? Science Translational Medicine **8**:341ps12–341ps12. doi:10.1126/scitranslmed.aaf5027.

8. **Whitaker K**. 2017. Publishing a reproducible paper. doi:10.6084/m9.figshare.5440621.v2.

9. **Collins FS**, **Tabak LA**. 2014. NIH plans to enhance reproducibility. Nature **505**:612–613. doi:10.1038/505612a.

10. **Sze MA**, **Schloss PD**. 2016. Looking for a signal in the noise: Revisiting obesity and the microbiome. mBio **7**:e01018–16. doi:10.1128/mbio.01018-16.

11. **Walters WA**, **Xu Z**, **Knight R**. 2014. Meta-analyses of human gut microbes associated with obesity and IBD. FEBS Letters **588**:4223–4233. doi:10.1016/j.febslet.2014.09.039.

12. **Finucane MM**, **Sharpton TJ**, **Laurent TJ**, **Pollard KS**. 2014. A taxonomic signature of obesity in the

microbiome? Getting to the guts of the matter. PLOS ONE **9**:e84689. doi:10.1371/journal.pone.0084689.

13. **Turnbaugh PJ**, **Hamady M**, **Yatsunenko T**, **Cantarel BL**, **Duncan A**, **Ley RE**, **Sogin ML**, **Jones WJ**, **Roe BA**, **Affourtit JP**, **Egholm M**, **Henrissat B**, **Heath AC**, **Knight R**, **Gordon JI**. 2008. A core gut microbiome in obese and lean twins. Nature **457**:480–484. doi:10.1038/nature07540.

14. **Turnbaugh PJ**, **Ley RE**, **Mahowald MA**, **Magrini V**, **Mardis ER**, **Gordon JI**. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature **444**:1027–131. doi:10.1038/nature05414.

15. **Ley RE**, **Backhed F**, **Turnbaugh P**, **Lozupone CA**, **Knight RD**, **Gordon JI**. 2005. Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences **102**:11070–11075. doi:10.1073/pnas.0504978102.

16. **Ley RE**, **Turnbaugh PJ**, **Klein S**, **Gordon JI**. 2006. Human gut microbes associated with obesity. Nature **444**:1022–1023. doi:10.1038/4441022a.

17. **Amstutz P**, **Crusoe MR**, **Nebojša Tijanić**, **Chapman B**, **Chilton J**, **Heuer M**, **Kartashov A**, **Leehr D**, **Ménager H**, **Nedeljkovich M**, **Scales M**, **Soiland-Reyes S**, **Stojanovic L**. 2016. Common workflow language, v1.0. doi:10.6084/m9.figshare.3115156.v2.

18. **Leek JT**, **Scharpf RB**, **Bravo HC**, **Simcha D**, **Langmead B**, **Johnson WE**, **Geman D**, **Baggerly K**, **Irizarry RA**. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics **11**:733–739. doi:10.1038/nrg2825.

19. **Human Microbiome Project Consortium**. 2012. Structure, function and diversity of the healthy human microbiome. Nature **486**:207–214. doi:10.1038/nature11234.

20. **Ding T**, **Schloss PD**. 2014. Dynamics and associations of microbial community types across the human body. Nature **509**:357–360. doi:10.1038/nature13178.

21. **Langille MGI**, **Ravel J**, **Fricke WF**. 2018. Available upon request: Not good enough for microbiome data! Microbiome **6**. doi:10.1186/s40168-017-0394-z.

22. **Ravel J**, **Wommack K**. 2014. All hail reproducibility in microbiome research. Microbiome **2**:8. doi:10.1186/2049-2618-2-8.

23. **Zhernakova A**, **Kurilshikov A**, **Bonder MJ**, **Tigchelaar EF**, **Schirmer M**, **Vatanen T**, **Mujagic Z**,

Vila AV, **Falony G**, **Vieira-Silva S**, **Wang J**, **Imhann F**, **Brandsma E**, **Jankipersadsing SA**, **Joossens M**, **Cenit MC**, **Deelen P**, **Swertz MA**, **Weersma RK**, **Feskens EJM**, **Netea MG**, **Gevers D**, **Jonkers D**, **Franke L**, **Aulchenko YS**, **Huttenhower C**, **Raes J**, **Hofker MH**, **Xavier RJ**, **Wijmenga C**, **and JF**. 2016. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science **352**:565–569. doi:10.1126/science.aad3369.

24. **Goodrich JK**, **Davenport ER**, **Beaumont M**, **Jackson MA**, **Knight R**, **Ober C**, **Spector TD**, **Bell JT**, **Clark AG**, **Ley RE**. 2016. Genetic determinants of the gut microbiome in UK twins. Cell Host & Microbe **19**:731–743. doi:10.1016/j.chom.2016.04.017.

25. **Zupancic ML**, **Cantarel BL**, **Liu Z**, **Drabek EF**, **Ryan KA**, **Cirimotich S**, **Jones C**, **Knight R**, **Walters WA**, **Knights D**, **Mongodin EF**, **Horenstein RB**, **Mitchell BD**, **Steinle N**, **Snitker S**, **Shuldiner AR**, **Fraser CM**. 2012. Analysis of the gut microbiota in the old order Amish and its relation to the metabolic syndrome. PLOS One **7**:e43052. doi:10.1371/journal.pone.0043052.

26. **Schloss PD**, **Westcott SL**, **Ryabin T**, **Hall JR**, **Hartmann M**, **Hollister EB**, **Lesniewski RA**, **Oakley BB**, **Parks DH**, **Robinson CJ**, **others**. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology **75**:7537–7541.

27. **Cole JR**, **Wang Q**, **Fish JA**, **Chai B**, **McGarrell DM**, **Sun Y**, **Brown CT**, **Porras-Alfaro A**, **Kuske CR**, **Tiedje JM**. 2013. Ribosomal database project: Data and tools for high throughput rRNA analysis. Nucleic Acids Research **42**:D633–D642. doi:10.1093/nar/gkt1244.

28. **Yilmaz P**, **Parfrey LW**, **Yarza P**, **Gerken J**, **Pruesse E**, **Quast C**, **Schweer T**, **Peplies J**, **Ludwig W**, **Glöckner FO**. 2013. The SILVA and All-species living tree project (LTP) taxonomic frameworks. Nucleic Acids Research **42**:D643–D648. doi:10.1093/nar/gkt1209.

29. **DeSantis TZ**, **Hugenholtz P**, **Larsen N**, **Rojas M**, **Brodie EL**, **Keller K**, **Huber T**, **Dalevi D**, **Hu P**, **Andersen GL**. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology **72**:5069–5072. doi:10.1128/aem.03006-05.

30. **Gelman A**, **Loken E**. 2014. The statistical crisis in science. American Scientist **102**:460. doi:10.1511/2014.111.460.

31. **Head ML**, **Holman L**, **Lanfear R**, **Kahn AT**, **Jennions MD**. 2015. The extent and consequences of

p-hacking in science. PLOS Biology **13**:e1002106. doi:10.1371/journal.pbio.1002106.

32. **Errington TM**, **Iorns E**, **Gunn W**, **Tan FE**, **Lomax J**, **Nosek BA**. 2014. An open investigation of the reproducibility of cancer biology research. eLife **3**. doi:10.7554/elife.04333.

33. **Pain E**. 2015. Register your study as a new publication option. Science. doi:10.1126/science.caredit.a1500282.

34. **Nosek BA**, **Ebersole CR**, **DeHaven AC**, **Mellor DT**. 2017. Preprint: The preregistration revolution. OSF Preprints. doi:10.17605/OSF.IO/2DXU5.

35. **Xie Y**. 2015. Dynamic documents with R and knitr, 2nd ed. Chapman; Hall/CRC, Boca Raton, Florida.

36. **Kluyver T**, **Ragan-Kelley B**, **Pérez F**, **Granger B**, **Bussonnier M**, **Frederic J**, **Kelley K**, **Hamrick J**, **Grout J**, **Corlay S**, **Ivanov P**, **Avila D**, **Abdalla S**, **Willing C**. 2016. Jupyter notebooks – a publishing format for reproducible computational workflows. IOS Press.

37. **Klein M**, **Sompel HV de**, **Sanderson R**, **Shankar H**, **Balakireva L**, **Zhou K**, **Tobin R**. 2014. Scholarly context not found: One in five articles suffers from reference rot. PLOS ONE **9**:e115253. doi:10.1371/journal.pone.0115253.

38. **Kim D**, **Hofstaedter CE**, **Zhao C**, **Mattei L**, **Tanes C**, **Clarke E**, **Lauder A**, **Sherrill-Mix S**, **Chehoud C**, **Kelsen J**, **Conrad M**, **Collman RG**, **Baldassano R**, **Bushman FD**, **Bittinger K**. 2017. Optimizing methods and dodging pitfalls in microbiome research. Microbiome **5**. doi:10.1186/s40168-017-0267-5.

39. **Ivanov II**, **Llanos Frutos R de**, **Manel N**, **Yoshinaga K**, **Rifkin DB**, **Sartor RB**, **Finlay BB**, **Littman DR**. 2008. Specific microbiota direct the differentiation of IL-17-producing t-helper cells in the mucosa of the small intestine. Cell Host & Microbe **4**:337–349. doi:10.1016/j.chom.2008.09.009.

40. **Ivanov II**, **Atarashi K**, **Manel N**, **Brodie EL**, **Shima T**, **Karaoz U**, **Wei D**, **Goldfarb KC**, **Santee CA**, **Lynch SV**, **Tanoue T**, **Imaoka A**, **Itoh K**, **Takeda K**, **Umesaki Y**, **Honda K**, **Littman DR**. 2009. Induction of intestinal th17 cells by segmented filamentous bacteria. Cell **139**:485–498. doi:10.1016/j.cell.2009.09.033.

41. **Laukens D**, **Brinkman BM**, **Raes J**, **Vos MD**, **Vandenabeele P**. 2015. Heterogeneity of the gut microbiome in mice: Guidelines for optimizing experimental design. FEMS Microbiology Reviews **40**:117–132. doi:10.1093/femsre/fuv036.

42. **Horbach SPJM**, **Halffman W**. 2017. The ghosts of HeLa: How cell line misidentification contaminates

the scientific literature. PLOS ONE **12**:e0186281. doi:10.1371/journal.pone.0186281.

43. **Huang Y**, **Liu Y**, **Zheng C**, **Shen C**. 2017. Investigation of cross-contamination and misidentification of 278 widely used tumor cell lines. PLOS ONE **12**:e0170384. doi:10.1371/journal.pone.0170384.

44. **Han S-W**, **Sriariyanun M**, **Lee S-W**, **Sharma M**, **Bahar O**, **Bower Z**, **Ronald PC**. 2013. Retraction: Small protein-mediated quorum sensing in a gram-negative bacterium. PLOS ONE **8**. doi:10.1371/annotation/880a72e1-9cf3-45a9-bf1c-c74ccb73fd35.

45. **Lee S-W**, **Han S-W**, **Sririyanum M**, **Park C-J**, **Seo Y-S**, **Ronald PC**. 2013. Retraction. Science **342**:191–191. doi:10.1126/science.342.6155.191-a.

46. **Salter SJ**, **Cox MJ**, **Turek EM**, **Calus ST**, **Cookson WO**, **Moffatt MF**, **Turner P**, **Parkhill J**, **Loman NJ**, **Walker AW**. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology **12**. doi:10.1186/s12915-014-0087-z.

47. **Perez-Muñoz ME**, **Arrieta M-C**, **Ramer-Tait AE**, **Walter J**. 2017. A critical assessment of the sterile womb and in utero colonization hypotheses: Implications for research on the pioneer infant microbiome. Microbiome **5**. doi:10.1186/s40168-017-0268-4.

48. **Lauder AP**, **Roche AM**, **Sherrill-Mix S**, **Bailey A**, **Laughlin AL**, **Bittinger K**, **Leite R**, **Elovitz MA**, **Parry S**, **Bushman FD**. 2016. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. Microbiome **4**. doi:10.1186/s40168-016-0172-3.

49. **Morris A**, **Beck JM**, **Schloss PD**, **Campbell TB**, **Crothers K**, **Curtis JL**, **Flores SC**, **Fontenot AP**, **Ghedin E**, **Huang L**, **Jablonski K**, **Kleerup E**, **Lynch SV**, **Sodergren E**, **Twigg H**, **Young VB**, **Bassis CM**, **Venkataraman A**, **Schmidt TM**, **Weinstock GM**. 2013. Comparison of the respiratory microbiome in healthy nonsmokers and smokers. American Journal of Respiratory and Critical Care Medicine **187**:1067–1075. doi:10.1164/rccm.201210-1913oc.

50. **Guo Q**, **Thabane L**, **Hall G**, **McKinnon M**, **Goeree R**, **Pullenayegum E**. 2014. A systematic review of the reporting of sample size calculations and corresponding data components in observational functional magnetic resonance imaging studies. NeuroImage **86**:172–181. doi:10.1016/j.neuroimage.2013.08.012.

51. **Ioannidis JPA**. 2005. Why most published research findings are false. PLOS Medicine **2**:e124. doi:10.1371/journal.pmed.0020124.

52. **Etz A**, **Vandekerckhove J**. 2016. A bayesian perspective on the reproducibility project: Psychology.

PLOS ONE **11**:e0149794. doi:10.1371/journal.pone.0149794.

53. **Gelman A**, **Hill J**, **Yajima M**. 2012. Why we (usually) dont have to worry about multiple comparisons. Journal of Research on Educational Effectiveness **5**:189–211. doi:10.1080/19345747.2011.618213.

54. **Munafò MR**, **Smith GD**. 2018. Robust research needs many lines of evidence. Nature **553**:399–401. doi:10.1038/d41586-018-01023-3.

55. **Mallick H**, **Ma S**, **Franzosa EA**, **Vatanen T**, **Morgan XC**, **Huttenhower C**. 2017. Experimental design and quantitative analysis of microbial community multiomics. Genome Biology **18**. doi:10.1186/s13059-017-1359-z.

56. **Jenior ML**, **Leslie JL**, **Young VB**, **Schloss PD**. 2017. *Clostridium difficile* colonizes alternative nutrient niches during infection across distinct murine gut microbiomes. mSystems **2**:e00063–17. doi:10.1128/msystems.00063-17.

57. **Califf KJ**, **Schwarzberg-Lipson K**, **Garg N**, **Gibbons SM**, **Caporaso JG**, **Slots J**, **Cohen C**, **Dorrestein PC**, **Kelley ST**. 2017. Multi-omics analysis of periodontal pocket microbial communities pre- and posttreatment. mSystems **2**:e00016–17. doi:10.1128/msystems.00016-17.

58. **Pearson H**. 2003. Competition in biology: Its a scoop! News@Nature. doi:10.1038/news031124-9.

59. **The PLOS Biology Staff Editors**. 2018. The importance of being second. PLOS Biology **16**:e2005203. doi:10.1371/journal.pbio.2005203.

60. **Seok J**, **Warren HS**, **Cuenca AG**, **Mindrinos MN**, **Baker HV**, **Xu W**, **Richards DR**, **McDonald-Smith GP**, **Gao H**, **Hennessy L**, **Finnerty CC**, **López CM**, **Honari S**, **Moore EE**, **Minei JP**, **Cuschieri J**, **Bankey PE**, **Johnson JL**, **Sperry J**, **Nathens AB**, **Billiar TR**, **West MA**, **Jeschke MG**, **Klein MB**, **Gamelli RL**, **Gibran NS**, **Brownstein BH**, **Miller-Graziano C**, **Calvano SE**, **Mason PH**, **Cobb JP**, **Rahme LG**, **Lowry SF**, **Maier RV**, **Moldawer LL**, **Herndon DN**, **Davis RW**, **Xiao W**, **and RGT**. 2013. Genomic responses in mouse models poorly mimic human inflammatory diseases. Proceedings of the National Academy of Sciences **110**:3507–3512. doi:10.1073/pnas.1222878110.

61. **Nguyen TLA**, **Vieira-Silva S**, **Liston A**, **Raes J**. 2015. How informative is the mouse for human gut microbiota research? Disease Models & Mechanisms **8**:1–16. doi:10.1242/dmm.017400.

62. **Wilson G**, **Bryan J**, **Cranston K**, **Kitzes J**, **Nederbragt L**, **Teal TK**. 2017. Good enough practices in

scientific computing. PLOS Computational Biology **13**:e1005510. doi:10.1371/journal.pcbi.1005510.

63. **Noble WS**. 2009. A quick guide to organizing computational biology projects. PLOS Computational Biology **5**:e1000424. doi:10.1371/journal.pcbi.1000424.

64. **Taschuk M**, **Wilson G**. 2017. Ten simple rules for making research software more robust. PLOS Computational Biology **13**:e1005412. doi:10.1371/journal.pcbi.1005412.

65. **Hart EM**, **Barmby P**, **LeBauer D**, **Michonneau F**, **Mount S**, **Mulrooney P**, **Poisot T**, **Woo KH**, **Zimmerman NB**, **Hollister JW**. 2016. Ten simple rules for digital data storage. PLOS Computational Biology **12**:e1005097. doi:10.1371/journal.pcbi.1005097.

66. **Perez-Riverol Y**, **Gatto L**, **Wang R**, **Sachsenberg T**, **Uszkoreit J**, **Veiga Leprevost F da**, **Fufezan C**, **Ternent T**, **Eglen SJ**, **Katz DS**, **Pollard TJ**, **Konovalov A**, **Flight RM**, **Blin K**, **Vizcaíno JA**. 2016. Ten simple rules for taking advantage of git and GitHub. PLOS Computational Biology **12**:e1004947. doi:10.1371/journal.pcbi.1004947.

67. **Sandve GK**, **Nekrutenko A**, **Taylor J**, **Hovig E**. 2013. Ten simple rules for reproducible computational research. PLOS Computational Biology **9**:e1003285. doi:10.1371/journal.pcbi.1003285.

68. **Wilson G**. 2016. Software carpentry: Lessons learned. F1000Research. doi:10.12688/f1000research.3-62.v2.