~~Reproducible Research Is Really F$%ing Hard~~Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research

Running title: Reproducible Research Is Really F#$%ing Hard

Patrick D. Schloss†

† pschloss@umich.edu; Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI

Format: Perspective

Counts: ~~4400 words plus 69~~ 5500 words plus 71 references, 2 tables, 0 figures, and a ~~144~~ 148 word abstract

Abstract

The “reproducibility crisis” in science affects microbiology as much as any other area of inquiry, and microbiologists have long struggled to make their research reproducible. We need to respect that ~~science and microbiology, in particular, are difficult.~~ ensuring that our methods and results are sufficiently transparent is difficult. This difficulty is compounded in interdisciplinary fields such as microbiome research. There are many reasons why a ~~result may not be reproducible or replicable. Even~~ researcher is unable to reproduce a previous result and even if a result is reproducible, it may not be correct. ~~This Perspective lays out~~ Furthermore, failure to reproduce previous results have much to teach us about the scientific process and microbial life itself. This Perspective delineates a framework for ~~improving the~~ identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability of ~~a particular result. It then describes the factors that can threaten this framework and approaches microbiologists can take to overcome the threats. Finally, it provides several exercises for individuals and research groups who wish to gain a better appreciation of how own research practices facilitate reproducibility and replication by others . Failure to validate previous results have much to teach us about the scientific process and microbial life itself.~~ microbiome research. Instead of seeing signs of a crisis in others’ work, we need to appreciate the technical and social difficulties that limit reproducibility in the work of others as well as our own.

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

Introduction

On first blush, one might argue that any scientist should be able to reproduce another scientist's research with no friction. Yet two anecdotes suffice to describe why this is not the case. The first goes to the roots of microbiology when Antonie van Leeuwenhoek submitted a letter to the Royal Society in 1677, "Concerning little animals" (1). This seminal work and several of his prior investigations described novel observations of microorganisms, but the scientific community rejected his observations for several reasons. First, because Leeuwenhoek had little interest in sharing his methods with others, they could not be ~~replicated~~reproduced. Second, he wrote in "low Dutch" and his writing was translated to English and edited to half their original length. This likely removed a significant amount of information regarding his methods. After several failures, Robert Hooke refined his own compound microscope and was able to reproduce Leeuwenhoek's observations. The precision of Hooke's observations was hindered by his use of a compound microscope, which had inferior optics to that of Leeuwenhoek's single lens microscope. In the process, Hooke popularized the compound microscope. This succession of events is illustrative of many of the current problems microbiologists face in ~~reproducing and replicating~~ validating each other's work. ~~Of course,~~ Time has proven that Leeuwenhoek's work was rigorous, impactful, and robust. It was not sloppy and there was no fraud. But, it required multiple efforts by one of the greatest minds in science to ~~replicate~~ reproduce the results and even then it was a poor ~~replication~~ reproduction of the original.

The second anecdote took place more recently. In 2011 Philip Bourne challenged those attending the *Beyond the PDF* workshop (https://sites.google.com/site/beyondthepdf/) 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 collaborative analysis involving the original authors and scientists from Spain, China, and the United States that challenged concepts critical to understanding reproducible research (3). The reanalysis 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 questions without simple answers. ~~Few would suggest that~~ Bourne's ~~group was sloppy or that they failed to be transparent~~track record in science and as a leader in the field of bioinformatics suggest that his group was not sloppy and his challenge indicated a level of transparency that is rare in science. Yet the investigators who sought to reproduce the findings found that someone with basic bioinformatics skills would require at least 160 hours to decipher the approaches used in the original analysis and an additional 120 hours to implement them to complete the reproduction.

Both of 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* (4, 5). 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 enveloped in Leeuwenhoek's case or Bourne's challenge: insuring that one's research design and methods are sufficiently clear is enormously difficult. Researchers are frequently frustrated with their own lack of documentation 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. I contend that many of the difficulties we face in ensuring the reproducibility of our research is social and driven by cultural forces within science.

Although the issues identified by the AAM colloquium participants are important, this Perspective argues that they are not the main reason for a reproducibility crisis in microbiology. It is scientifically valuable to consider what other factors threaten our ability to ~~validate~~ reproduce a result. Although these factors highlight the technical limitations and cultural forces we face, our inability to validate a result may also indicate that we still have much to learn about biology. Furthermore, we must remember that whether we can validate a result is not just a product of rigorous scientific practice, but also a product of stochastic forces (6, 7). We must also be on guard against assuming that just because a result is reproducible that it is correct (~~7~~8). With these general points in mind, the goals of this Perspective are three-fold. First, I present a framework for thinking about how science is conducted within the microbial sciences. Second, I provide an overview of various factors that threaten the field's ability to validate prior results and the tools that we can use to overcome these problems. Third, based on these issues, I provide five exercises that research groups can use to motivate important discussions of their practices and how ~~they~~ their practices foster or impede efforts to validate ~~their~~ the researchers' results. Although I will primarily focus on examples from microbiome research, the principles are generalizable to other areas of microbiology as all scientists struggle to ensure the reproducibility of their research.

Threats to reproducibility

Developing a framework. One of the struggles in discussing reproducibility, replicability, and the factors that can limit them, is agreeing upon how they should be defined (7). Reproducibility is used as a vague

term for being able to repeat another researchers' work whether that is with the same protocols or the same populations. ~~The AAM report used the term reproducibility where others would use replicability.~~ This Perspective will use ~~the most widely used definitions, which describe reproducibility as the~~ definitions that have greater precision and that are based on definitions that are widely used in the statistics literature. Reproducibility is the ability to regenerate a result with the same dataset and data analysis workflow and replicability ~~as~~ is the ability to produce a consistent result with an independent experiment asking the same scientific question (~~7~~8). I propose a similar framework that accounts for the practice of applying multiple methods to the same samples to improve the robustness and generalizability of a result (Table 1) (~~9~~10). It is critical for scientists to give attention to the right hand column of the framework. Most research is exploratory and 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 (4, 5, ~~10~~7, 11). Results must be reproducible and robust, but they ~~must also~~ also need to be replicable and generalizable.

An example. The question of whether there are microbiome-based signatures of obesity is a useful illustration to demonstrate the factors that affect each of the quadrants of the grid in Table 1 and it can be used to underscore the difficulty of ensuring the reproducibility, replicability, robustness, and generalizability of results. Several research groups, including mine (~~11~~12), have attempted to validate the result that obese individuals were more likely to have lower bacterial diversity and relative abundances of *Bacteroidetes* (~~12,~~ 13, 14). The original observation was published in 2008 using 16S rRNA gene sequence data and continues to engender much enthusiasm for the role of the microbiome in human health (~~14). Although~~ 15). It is important to note that the original study was one of the first to use high throughput amplicon sequencing and so there was minimal infrastructure to deposit and store such sequences in public databases. Furthermore, many of the software tools that we now rely on for facilitating reproducible workflows were not available. Regardless, although the original study was performed using poorly described data curation methods, we 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 (Table 1). However, when we used the same methods with data from nine other cohorts, we and others have failed to replicate the result (~~11–13~~12–14). These failures to replicate the original result may be due to methodological differences across the replicating studies, differences in study populations, or statistical variation. ~~It~~ Our study demonstrated that each of ten cohorts were significantly underpowered to identify a 10% difference in Shannon diversity (12). Therefore, the lack of statistical power may have been responsible for an inability to detect a difference. Each of these studies were rather large for the time that they were published within the development of the microbiome

research field and so the original researchers likely thought they had obtained the best statistical power that was feasible. Identifying what a biologically meaningful difference in any parameter within the microbiome literature to complete a meaningful power analysis has been a challenge. Each of these factors still make it nearly impossible to perform a meaningful *a priori* power analysis to aid in the design of any cohort. Next, it is worth noting that those involved in the original 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 (~~15~~16). 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 (~~15~~16). In a human cohort, they generated multiple datasets that each reflected different regions of the 16S rRNA gene. In obese individuals, they observed lower diversity and relative abundance of *Bacteroidetes* (~~14~~15). They also used shotgun metagenomic sequencing to postulate the enrichment of carbohydrate processing genes in obese individuals (~~14~~15). In a smaller cohort study, although the subjects' diversity remained constant, as the authors predicted, the relative abundance of *Bacteroidetes* increased as the subjects lost weight (~~17~~18). Although each part of their approach had significant weaknesses including methodological biases and underpowered experimental designs, their results supported the hypothesis that there are microbial signatures associated with obesity. This conclusion was robust within the cohort they studied, but it was not generalizable to other cohorts. Within this example it is apparent that scientists acted in good faith given the technological and cultural conditions that they were working in. These conditions underscore the difficulty of replicating and generalizing results.

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 many journals impose word limits on manuscripts, Materials and Methods sections become a chain of citations to previous work that each cite previous work (10). 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 would make it easier for researchers to avoid these rabbit holes. For data analysis workflows, software such as GNU Make (https://www.gnu.org/software/make/) and the Common Workflow Language (18) make it possible to track data dependencies and automate a workflow. For example, we used GNU Make to write a workflow in our meta-analysis of the obesity data , such that downloading a copy of the scripts from the project's GitHub repository and writing "make write.paper" in the command line will reproduce our analysis. Workflow tools make it possible to trace the provenance of a summary statistic from the manuscript back to the raw data .~~

~~Unfortunately, the raw data behind a result is~~ Too often the underlying raw sequencing data and associated data that contextualizes the sequencing data are often not accessible~~, which~~. Clearly, this makes reproducing a prior analysis impossible (19, 20). Well-established databases for storing a variety of "omics" data exist and other data should be archived in third-party databases such as FigShare (https://figshare.com) and Dryad (https://datadryad.org). However, some researchers still fail to post their sequencing data to public databases or do not provide the necessary metadata with the sequencing data. 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 original study only provided the subjects' body mass index (BMI) as categories (~~14~~15). We were unable to access the actual heights, weights, and BMIs. We did not include three large datasets from two studies because their data were inaccessible due to onerous data sharing agreements (21, 22). Two other datasets required at least a month of effort to obtain (23, 24). More broadly, Stodden et al. (25) recently showed that although *Science* magazine has had clear guidelines requiring authors to make the data and code for their studies available, only 44% of the authors who published papers in 2011 and 2012 were willing to provide the resources. Lack of access to the data and underlying code for an analysis clearly limits the ability of others to reproduce and build upon that analysis.

~~Rapid~~ "Link rot" - the fact that web or email addresses become deprecated - is a significant problem for those attempting to access the data and methods needed to reproduce a result (26). Changes in institutional affiliation frequently render email addresses invalid. ORCID (https://orcid.org) has emerged as a technology to solve the email rot problem and many journals use it to provide a persistent link to an individual's many scientific identities over their career. The fraction of manuscripts including web resources continues to grow and yet at least 70% of those manuscripts include URLs that are inaccessible (26). To prevent 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. Unfortunately, the developer of the web resources must ensure that the resource remains active. The inevitability of link rot further emphasizes the importance of using public and stable servers that are likely to persist.

Related to link rot, rapid advances in sequencing technology, data curation, databases, and statistical techniques present an additional threat to reproducibility ~~.~~ because resources and what are considered best practices are constantly evolving. This evolution is not always well documented. For example, the ~~Human Microbiome Project used Roche's 454 platform to sequence the 16S rRNA gene (23). 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 (~~25~~27). The RDP [(~~26~~28); http://rdp.cme.msu.edu] and SILVA [(~~27~~29); 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 [(~~28~~30); 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~~ resources and workflows at sites such as GenBank (https://www.ncbi.nlm.nih.gov/genbank), greengenes, RDP, and SILVA preclude analyzing new data with older versions of the sites. The greengenes website removed their online tools in April 2017, exemplifying the problem with web-based workflows. Their database files are now available through the company, Second Genome, but their tools are not. Combined with the development of new sequencing platforms and deprecation of old platforms, these changes in technology, references, and software underscore the importance of adequately documenting workflows and enabling users to recreate the conditions that the original researchers worked within.

~~"Link rot" - the fact that web or email addresses become deprecated - is a significant problem for those attempting to access the data and methods needed to reproduce a result (29). Changes in institutional affiliation frequently render email addresses invalid.ORCID~~ Because many journals impose word limits on manuscripts, Materials and Methods sections become a chain of citations to previous work that each cite previous work (11). Improved documentation in supplementary materials or archives such as protocols.io (https://~~orcid.org) has emerged as a technology to solve the email rot problem and many journals use it to provide a persistent link to an individual's many scientific identities over their career. The fraction of manuscripts including web resources continues to grow and yet at least 70% of those manuscripts include URLs that are inaccessible (29). To prevent link rot, services like Zotero~~ www.protocols.io) for lab-based methods or through GitHub (https://github.com) for data analysis workflows would make it easier for researchers to avoid these rabbit holes. For data analysis workflows, software such as GNU Make (https://www.~~zotero.org) can provide a digital object identifier (DOI) that persists even if the link that it points to changes. Unfortunately, the developer of the web resources must insure that the resource remains active. The inevitability of link rot further emphasizes the importance of using public and stable servers that are likely to persist for at least a decade~~gnu.org/software/make/) and the Common Workflow Language (31) make it possible to track data dependencies and automate a workflow. For example, we used GNU Make to write a workflow in our meta-analysis of the obesity data, such that downloading a copy of the scripts from the project's GitHub repository and writing "make write.paper" in the command line will reproduce our

analysis. Although considerable effort is required to make them work, workflow tools make it possible to trace the provenance of a summary statistic from the manuscript back to the raw data.

~~Other problems with reproducibility~~ The use of workflow tools, literate programming tools (e.g. RMarkdown (32) and Jupyter (33)), and version control software provide researchers with mechanisms to track the development of their analyses. Furthermore, these tools can help researchers reflect the fact that ~~science is~~ their analysis was not a linear process resembling a pipeline. In reality, questions change and scientists can fall into the traps of the "Garden of Many Forking Paths" where they go looking for a desired result (~~30~~34) or "P-hacking" where large numbers of statistical hypothesis tests are attempted without adequately correcting for performing multiple tests (~~31~~35). Although it is possible to pre-register data analysis plans (~~32–34~~36–38), these plans 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 programmingdocuments can allow researchers to document and share the evolution of their analyses~~An increasing number of microbiome researchers are using workflow, literate programming, and version control tools to document their analyses. I have yet to observe widespread exploration of the history of projects' repositories or the adoption of pre-registration of data analysis plans among microbiome researchers. Although these have their technical and cultural limitations, they offer greater transparency to improved reproducibility.

Replicability. A number of threats similar to those for reproducibility could explain why a previous result cannot be replicated. In addition to those detailed previously, there are threats related to differences in systems or populations and the ability to control for those differences.

~~Problems with~~ Forgotten in discussions of replication failures by many microbiologists is that 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 with which 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. Further, they rarely quantify the risk of falsely accepting a null hypothesis (i.e. Type II errors) (39). In some cases, an insufficient sample size in the replicate study may explain the failure to replicate a study. In other cases, the original study may have been underpowered, rendering it susceptible to an inflated risk of Type I errors (40). Solutions to these problems include pre-registering data analysis plans (36–38), justifying sample sizes based on power calculations

(11, 12, 39), and using Bayesian frameworks that allow prior knowledge of the system to influence the interpretation of new results (41, 42). It needs to be underscored, however, that to measure statistical power and use that information to inform sample size selections one must know what a biologically relevant difference is. The microbiome field has yet to make that determination. Our previous power analysis used varying differences in Shannon diversity (12). As we indicated, those levels were picked because they seemed reasonable, not because of a biological foundation. Furthermore, there was no reason to think that diversity metrics are the most biologically meaningful parameters to base the calculations on.

Beyond problems of sample size and statistical power calculations, problems with experimental design are also often a threat to replicability because investigators fail to account for confounding variables in the original study. A subsequent study may fail to find the same result because their design is not impacted by the confounding variable. In sequence-based analyses, threats to replicability 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 (~~37~~43). One notable example occurred within the Human Microbiome Project where 150 people were recruited in Houston, TX and 150 in St. Louis, MO (23). Researchers at the Baylor College of Medicine and Washington University performed the DNA extractions for the two sets of subjects, respectively. Researchers at the Baylor College of Medicine, the J. Craig Venter Institute, and the Broad Institute sequenced the DNA from the Houston subjects and researchers from Washington University sequenced the DNA from the St. Louis ~~subject~~subjects. The subject's city was the variable with the largest effect size, although all parties used the same standard operating procedures to sample the subjects and extract and sequence the DNA (23, ~~38~~44). Because the city of origin and the center that did the extractions were perfectly confounded, it was impossible to quantify the impact of geographic differences on the microbiome. Instead of being a single study that intended to address associations between geographical and microbiome variation, this became two replicate studies that were unable address the influence that geography has on the microbiome. It is easy to blame the those that designed the study for this confounding, but it is important to acknowledge the social conditions that were resolved via negotiations that may have impacted the design and the need to garner buy in from different centers.

In addition to variation between human cohorts, variation between bacterial and model organism strains 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 (~~16, 39~~17, 45). 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 (~~40, 41~~46, 47). Thus, the origin of the

mice and not the experimental treatment may explain the roles ascribed to the microbiota. 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 (~~42~~48). Similarly, comparing the microbiota of obese and lean individuals from a cohort of twins and their mothers in Missouri (~~14~~15) may have confounding factors that differ from members of Amish communities (24). In these cases, the problem with replicability is not due to the quality of the investigator's experimental practices, but to the differences that may be biological, demographic, or anthropological. Thus failure to replicate a study across different strains or cohorts could suggest that other interesting factors play a role in the phenomenon under study.

~~Uncertain~~ Just as uncertainty over the variation in mouse and human populations can impact the replicability of results, uncertain provenance and purity of reagents, organisms, and samples can also threaten replicability. Perhaps the best-known example is the discovery that HeLa cells contaminate many other cell lines, especially those in the same laboratory (~~43, 44~~49, 50). 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 (~~45, 46~~51, 52). 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~~ However, it is part of our scientific culture that if a colleague sends a strain to another researcher, the recipient generally trusts that they get the correct strain. There is also a growing awareness that DNA extraction kits can be contaminated with low levels of bacterial DNA (~~47~~53). These contaminants have led to the identification of contaminants as being important members of the lung and placental microbiota when mock extractions are not sequenced in parallel (~~48–50).~~

~~Forgotten in discussions of replication failures is that 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 with which 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. Further, they rarely quantify the risk of falsely accepting a null hypothesis (i.e. Type II errors) (51). 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 (11). In some cases, an insufficient sample size in the replicate study may explain the failure to replicate a study. In other cases, the original study may have been underpowered, rendering it susceptible to an inflated~~

risk of Type I errors (52). Solutions to these problems include pre-registering data analysis plans (32–34), justifying sample sizes based on power calculations (1054–56). For each of these threats to replication, 11, 51), and using Bayesian frameworks that allow prior knowledge of the system to influence the interpretation of new results(53, 54)we would be well served by following the proverb to "trust, but verify" by testing the robustness of the results.

Robustness. Every method has its own strengths and weaknesses. Therefore, it is important to address a research question from multiple and hopefully orthogonal directions. This strategy combines the strengths of different methods to overcome their individual weaknesses (5557). Evaluating the robustness of a result from a single cohort is becoming more common as researchers pursue multiple approaches including 16S rRNA gene sequencing, metagenomics, metatranscriptomics, and metabolomics (56–5858–60). Of course, biases in the underlying cohort design, sample collection and storage, or the nucleic acid processing will propagate through the analyses. The way to remedy this is to select methods that are as independent from each other as possible. For example, data collected from multiple regions of the 16S rRNA gene would not be considered truly independent datasets since amplicon sequencing would have been applied to the same samples. The results would be marginally more independent if one were to layer shotgun metagenomic data onto the 16S rRNA gene sequence data because although the same DNA would be used for sequencing, metagenomics 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 it requires completely different sample handling and processing steps. Quantitative PCR, cultivation, and microscopy could be similarly layered on these data. Ultimately, it is impossible for the results of each set of methods to be fully independent. If the underlying design of the study is flawed by insufficient statistical power or failure to account for confounding variables, then any attempts to test the robustness of a result will also be flawed.

Generalizability. A motivating goal in science is to have a result that is generalizable across populations or systems. Within a scientific culture that does not place value on publishing negative results, it is difficult to assess whether scientists' bias to support their prior results affects the ability to claim that a result is robust or generalizable.

Generalizability. The gold standard of science is to have a result that is generalizable across populations. Failing to Similarly, failing to attempt replication studies hinders the ability of researchers to test the generalizability of most results. Scientists often fear being "scooped" (5961). In reality, it is the second researcher who examines the same question that has the opportunity to increase the field's confidence that

a result is valid (~~60~~62). Generalizability is an important and broad question. 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 (~~61, 62~~63, 64). Like a failure to reproduce, replicate, or demonstrate the robustness of a result, a failure to generalize a result is not a failure of science. Rather, it is an opportunity to better understand the complex biology of bacteria and how they interact with their environments.

Fostering a culture of greater reproducibility and replicability

Training. Throughout my discussion of the threats to reproducibility, replicability, robustness, and generalizability failures on the part of scientists to be more transparent, provide greater documentation, or design better experiments have been balanced by an appreciation that we work within a scientific culture. This culture is limited by our ignorance of biology, rapid expansion in technology, misaligned rewards, and a lack of necessary training. A key observation from the work of Garijo and colleagues (3) 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. Yet, many microbiology training programs focus on laboratory skills while ignoring data analysis skills. A number of excellent "best practices" documents have emerged in recent years (~~63–68). I have created the Riffomonas project, which expounds on the threats to reproducibility and tools that microbiologists can use to maximize the computational reproducibility of their analyses (http://www.riffomonas.org).~~ 65–70). In addition, organizations including Software Carpentry and Data Carpentry offer workshops to introduce researchers to the best practices in reproducible research (~~69)~~ 71) (https://carpentries.org). Massively open online courses have been developed 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 important as learning the fundamentals of how to implement reproducible research methods is honing those skills in one's research. A novice could not reproduce Beethoven's "Für Elise" from sheet music without prior experience playing the piano. Similarly, novices cannot expect to reproduce a result without learning the methods of their discipline. With this analogy in mind, I have created the Riffomonas project, which expounds on the threats to reproducibility and tools that microbiome researchers can use to maximize the computational reproducibility of their analyses (http://www.riffomonas.org). The Riffomonas materials

use microbiome-related examples to illustrate the importance of transparency, documentation, automated workflows, version control, and literate programming to improving the computational reproducibility of an analysis. The goal is that once scientists have been trained in these practices they can apply them to their own work and use them to "riff" or adapt and build on the work of others.

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

1. Working away from each other, have two or more people to write instructions on how to fold a paper airplane. Have the participants trade instructions, separate, and implement the instructions. After the participants come back together ask: 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 other person's prior knowledge of paper airplanes, resources, and abilities were assumed? What challenges would length limitations place on this exercise? 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. Imagine a graduate student is really excited about an analysis that you performed in your most recent paper and would like to replicate the analysis with their own data. But first, they want to make sure that they reproduce your results. What steps are likely to cause the student problems? ~~Find~~ If it is not clear to you what problems they might face, find your favorite figure from ~~your favorite paper from~~ a paper by a different research group than your own. Can you reproduce the figure? What is standing in your way?

3. Take a figure from your recent paper and improve the likelihood that another researcher would be able to reproduce it. Where are the data and how would the researcher access them? What calculations were performed to summarize the data? What software was used to generate the figure? Is that software freely available? What steps would the researcher need to take to generate the figure? When you write your methods, what experience level are you writing for? Who should you be writing for? When you are confident that you have made the figure as reproducible as you can, give the instructions to several colleagues and ask for their feedback.

4. Complete an audit of the reproducibility practices in your research group. Table 2 provides a rubric that someone working within the host-associated microbiome field might use to assess their research. Within your research group, modify this rubric to suit your needs. For your next paper, work to improve one element from the rubric and constantly be developing an ethic of fostering greater reproducibility.

5. 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. Some might give into despair thinking that one person or research group can only have a minor impact. Have a discussion within your group about why things are this way, whether your group's practices should change, and what would be the easiest and most impactful thing to change.

Conclusion

A motivating concept ~~to~~ that has been attributed to many people to improve the reproducibility of one's research is that ~~the~~ they should think of themselves from a month ago as their most important collaborator~~is themselves six months ago, and old them does not have email access. Their~~ . They are not available to answer questions to things that they have forgotten in the intervening period. This is a common occurrence for many researchers who put projects to the side for a time to prepare for examinations, go on vacations, or work on other projects. Trying to piece together what they did previously is often a frustrating process. If instead they had been using tools to improve reproducibility, then they will be doing themselves a favor when they return to the project. Similarly, I consider their supervisor or co-authors to be their second most important ~~collaborator is the director of their research group.~~ collaborators. It is likely that the corresponding author was not the person that implemented the details of the analysis plan. Thus, it is important that they have access and the ability to navigate the project when they receive a query about how the analysis was done. Anyone that has done research can attest to how difficult it can be to ~~satisfying their two most important~~ satisfy these two sets of "collaborators". And yet, if can satisfy these collaborators, then we should be able to satisfy the third collaborator, the reader who hopes to build upon our work to generalize it or go in a new direction.

It is important to see that attempts to guard against threats to reproducibility, replicability, robustness, and generalizability are positive forces that will improve science. They have been considered a form of scientific "preventative medicine" (~~7~~8). 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. Beyond ensuring "correctness" the goal of these efforts, and I would argue their primary goal, should be to enable future scientists to build upon the work to go further. Before attributing difficulties with reproducibility, replicability, robustness, and generalizability to a dim view of our fellow scientists as being sloppy, biased, or untrustworthy, it is worth seriously considering the many factors - biological, statistical, and sociological - that pose a threat.

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). I am grateful to Ada Hagan for providing comments on an early version of the manuscript and Kate Epstein for assisting me with language editing.

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 (~~9~~10) 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~~ we do to improve the status of ~~my~~ our 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~~ we identify a list of possible confounding variables - biological and technical - that may obscure the interpretation of ~~my~~ our results?	Do ~~I~~ we indicate the level of randomization and experimental blocking that ~~I~~ we performed to minimize the effect of the confounding variables (i.e. batch effects)?	Does the interpretation of ~~my~~ our results limit itself to only those variables that are not obviously confounded?
Sex/gender as confounding variables	Do ~~I~~ we indicate the sex/gender of research animals~~and~~ /participants?	Do ~~I~~ we provide a justification for the lack of even representation?	Do ~~I have even representation of male and females in my~~ we have an equitable representation of sexes/genders in our studies? Do ~~I account for sex~~ we account for them as a variable?
Experimental design considerations	Do ~~I~~ we have an active collaboration with a statistician who helps with experimental design and analysis?	Do ~~I~~ we indicate the number of hypothesis tests ~~I~~ we performed and have ~~I~~ we corrected any P-values for multiple comparisons?	For ~~my~~ our primary research questions, have ~~I~~ we run a power analysis to determine the necessary sample size?
Data analysis plan	Before starting an analysis, have ~~I~~ we articulated a set of primary and secondary research questions	Has someone else reviewed ~~my~~ our data analysis plan prior to analyzing the data?	Have ~~I registered my~~ we registered our data analysis plan with a third party before starting the project?
Provenance of reagents	~~Do I include~~ Is there a table of ~~the~~ reagents such as cell lines, strains, ~~genotypes,~~ and primer sequences that ~~I~~ were used?	Where possible, have ~~I~~ we obtained reagents from certified entities like the American Type Culture Collection (ATCC)?	Is there a statement indicating how ~~I~~ we know the provenance and purity of each cell line and strain?
Controlling for initial microbiota	Are mice obtained from a breeding facility that allows me to track their pedigree?	Where possible, are mice from different treatment groups co-housed to control for differences in initial microbiota?	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 software descriptions	Are all methods, databases, and software tools cited? Do ~~I~~ we follow the relevant licensing requirements of each tool?	Do ~~I~~ we 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?
DNA contamination	Did ~~I~~ we quantify the background DNA concentration in ~~my~~ our reagents? Did ~~I~~ we sequence an extraction control?	~~Am I~~ Are we taking steps to minimize reagent contamination?	What methods do ~~I~~ we 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 publicly available?	Are intermediate and final data files publicly available?	Are tools like Amazon Machine Images (AMIs) used to make a snapshot of ~~my~~ our working directory?
Availability of metadata	Are all of the metadata necessary to repeat any analyses ~~I~~ we performed publicly available?	Have ~~I~~ we adhered to standards in releasing the minimal amount of metadata about ~~my~~ our samples?	Did ~~I~~ we 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 on ~~my~~ our computer?	Is this project contained within a single directory on ~~my~~ our computer and does it separate ~~my~~ our 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 and open 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~~ our work environment?
Documentation of data analysis workflow	Is ~~my~~ our code well documented? Do ~~I~~ we use a self-commenting coding practice?	Do each of ~~my~~ our 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~~ we know whether any of the steps in ~~my~~ our data analysis workflow depend on the use of a random number generator?	For analyses that utilize a random number generator, have ~~I~~ we noted the underlying random seed?	Have ~~I repeated my~~ we repeated our analysis with multiple seeds to show that the results are insensitive to the choice of the seed?
Defensive data analysis	Is ~~my~~ our data analysis pipeline flexible enough to add new data?	Does ~~my~~ our code include tests to confirm that it does what ~~I~~ we think it does?	Do ~~I make~~ we made use of automated tests and continuous integration tools to ~~insure~~ ensure internal reproducibility?
Insuring short and longterm reproducibility	Did ~~I~~ we release the underlying code and new data at the time of submitting a paper with their DOIs and accession numbers?	Did ~~I~~ we include a reproducibility statement or declaration at the end of the manuscript? Are ORCID identifiers provided for all authors?	What mechanisms are in place to ~~insure my~~ ensure our analysis remains accessible and reproducible in 5 years?
Open science to foster reproducibility	Have ~~I~~ we released any embargoes on ~~my~~ our code repository and raw data prior to submitting the manuscript?	Did ~~I~~ we post a preprint version of ~~my~~ our manuscript prior to submission?	Have ~~I~~ we published under a Creative Commons license? Is a permissive reuse license posted with ~~my~~ our code

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. **Kinnings SL**, **Xie L**, **Fung KH**, **Jackson RM**, **Xie L**, **Bourne PE**. 2010. The mycobacterium tuberculosis drugome and its polypharmacological implications. PLoS Computational Biology **6**:e1000976. doi:10.1371/journal.pcbi.1000976.

3. **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.

4. **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.

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

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. **Casadevall A**, **Fang FC**. 2010. Reproducible science. Infection and Immunity **78**:4972–4975. doi:10.1128/iai.00908-10.

8. **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.

~~8.~~ 9. **Goodman SN**, **Fanelli D**, **Ioannidis JPA**. 2016. What does research reproducibility mean? Science Translational Medicine **8**:341ps12–341ps12. doi:10.1126/scitranslmed.aaf5027.

~~9.~~ 10. **Whitaker K**. 2017. Publishing a reproducible paper. doi:10.6084/m9.figshare.5440621.v2.

~~10.~~ 11. **Collins FS**, **Tabak LA**. 2014. NIH plans to enhance reproducibility. Nature **505**:612–613. doi:10.1038/505612a.

~~11.~~ 12. **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.

~~12.~~ 13. **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.

~~13.~~ 14. **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.

~~14.~~ 15. **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.

~~15.~~ 16. **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.

~~16.~~ 17. **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.

~~17.~~ 18. **Ley RE**, **Turnbaugh PJ**, **Klein S**, **Gordon JI**. 2006. Human gut microbes associated with obesity. Nature **444**:1022–1023. doi:10.1038/4441022a.

~~18. **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:.~~

19. **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.

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

21. **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.

22. **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.

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

24. **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.

25. **Stodden V**, **Seiler J**, **Ma Z**. 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. Proceedings of the National Academy of Sciences **115**:2584–2589. doi:10.1073/pnas.1708290115.

26. **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.

27. **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.

~~26.~~ 28. **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.

~~27.~~ 29. **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.

~~28.~~ 30. **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.

~~29.~~ 31. ~~**Klein M**~~**Amstutz P**, ~~**Sompel HV de**~~**Crusoe MR**, ~~**Sanderson R**~~**Nebojša Tijanić**, ~~**Shankar H**~~**Chapman B**, ~~**Balakireva L**~~**Chilton J**, ~~**Zhou K**~~**Heuer M**, ~~**Tobin R. 2014. Scholarly context not found: One in five articles suffers from reference rot. PLOS ONE 9**~~**Kartashov A, Leehr D, Ménager H, Nedeljkovich M, Scales M, Soiland-Reyes S, Stojanovic L**. 2016. Common workflow language, v1.0. doi:~~e115253. doi:~~10.6084/m9.figshare.3115156.v2.

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

33. **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.

~~30.~~ 34. **Gelman A**, **Loken E**. 2014. The statistical crisis in science. American Scientist **102**:460. doi:10.1511/2014.111.460.

~~31.~~ 35. **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.~~ 36. **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.~~ 37. **Pain E**. 2015. Register your study as a new publication option. Science. doi:10.1126/science.caredit.a1500282.

~~34.~~ 38. **Nosek BA**, **Ebersole CR**, **DeHaven AC**, **Mellor DT**. 2017. Preprint: The preregistration revolution. OSF Preprints. doi:10.17605/OSF.IO/2DXU5.

~~35.~~ 39.

~~36. Kluyver T, Ragan-Kelley B, Pérez F~~Guo Q~~**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**~~Guo Q~~**, Granger B**~~**Thabane L**~~**, Bussonnier M**~~**Hall G**~~**, Frederic J**~~**McKinnon M**~~**, Kelley K**~~**Goeree R**~~**, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C. 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.**

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

41. Etz A, Vandekerckhove J. 2016. ~~Jupyter notebooks – a publishing format for reproducible computational workflows. IOS Press.~~ A bayesian perspective on the reproducibility project: Psychology. PLOS ONE 11:e0149794. doi:10.1371/journal.pone.0149794.

42. 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.

~~37.~~ 43. 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.

~~38.~~ 44. Ding T, Schloss PD. 2014. Dynamics and associations of microbial community types across the human body. Nature 509:357–360. doi:10.1038/nature13178.

~~39.~~ 45. 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.

~~40.~~ 46. 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.

~~41.~~ 47. 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.

~~42.~~ 48. 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.

~~43.~~ 49. 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.

~~44.~~ 50. 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.

~~45.~~ 51. 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.

~~46.~~ 52. 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.

~~47.~~ 53. 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.

~~48.~~ 54. 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.

~~49.~~ 55. 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.

~~50.~~ 56. 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.

~~51. 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:.~~

~~52. Ioannidis JPA. 2005. Why most published research findings are false. PLOS Medicine 2:e124. doi:.~~

~~53. Etz A, Vandekerckhove J. 2016. A bayesian perspective on the reproducibility project:~~

Psychology. PLOS ONE 11:e0149794. doi:.

54. 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:.

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

56. 58. 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.

57. 59. 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.

58. 60. 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.

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

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

61. 63. 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.

62. 64. 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.

63. 65. 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.

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

65. 67. 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.

66. 68. 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.

67. 69. 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.

68. 70. 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.

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