Assessing the reproducibility of microbiome data analysis

Today's most significant health care concerns include the treatment and prevention of obesity, diabetes, autism, antibacterial resistance, cancer, aging, and a growing list of other "Diseases of affluence" [refs]. There is growing sentiment that the bacteria that live in and on the human body (i.e. the human *microbiota*) and the environment they inhabit (i.e. the human *microbiome*) are at least partially involved in their etiology. A commonly expressed hypothesis is that changes in diet and hygiene have altered our microbiota leading to the increased prevalence of these diseases [refs]. Clearly, the human microbiome and the conditions and diseases it affects have profound implications for public policy and health care. Fueled by the decreasing costs of DNA sequencing and significant investments by the US National Institutes of Health and private foundations, the study of the human microbiome has exploded over the past 15 years [refs]. In 2010, 78 papers were published that included the keywords microbiota or microbiome. In 2014, there were over 4,000. Over the past 10 years the median annual increase in number of publications in the microbiome research space has been 36% and it has been only 5% for cancer. Although there is considerable excitement about this new area of research, there is also healthy skepticism that many of the claims made by its proponents are exaggerated [refs].

This explosion has created a unique scientific environment. Microbiologists who have been classically trained as reductionist molecular biologists are now expected to be molecular ecologists integrating the tools of statistics, bioinformatics, ecology, and clinical science. For the past 13 years we have developed software that is now the most widely used software package in the field and worked with numerous researchers. Our experience suggests that many new researchers face considerable difficulties implementing analysis plans in a manner that is reproducible. Furthermore, because novices do not have the ability to discriminate between competing methods, they may not perform the ideal methods. Even when analyses are done well, it is common for a series of complex data manipulation steps distilled to a single sentence in the methods section of a paper. Obviously, the field needs to assess the overall reproducibility and robustness of published data analyses and it needs to create a mechanism to better train novice researchers.

The Arnold Foundation has a special interest in "Research Integrity" and understanding how problems with research integrity affect public health policy. We are confident that this project fits within the mission of the foundation. As outlined above, the human microbiome is central to many issues in current public health discussions. The objective of the proposed project would be to assess the reproducibility of microbiome research by training researchers in the best current practices that improve reproducibility. To achieve our objective, we propose the following objectives:

- **Objective 1. Assess the level of methodological transparency of microbiome studies**
- **Objective 2. Quantify the ability of researchers to reproduce reported analyses**
- **Objective 3. Determine whether similar microbiome studies validate each other**

Together, these objectives will allow us to assess whether microbiome research has a reproducibility problem. As outlined below, we will achieve these objectives by engaging the microbiome research community. Given the significant role of the microbiome in human health improving the reliability of the results form the proposed research will have a meaningful positive impact.

Problem definition

Our research group has been at the forefront of microbiome research through our development of the popular software package, mothur, which has now been cited more than 2,600 times, making it the most widely cited software package for analyzing microbiome data (1). This has allowed us to have a significant role in helping to train literally thousands of microbiome researchers. Overall, these experiences have allowed us to identify three critical problems in the field that relate to the problem of reproducibility.

Lack of training in data analysis. We recently conducted a survey of mothur users to better understand their background. Of respondents, 41% were graduate students, 41% were PhD-level scientists, and 12% were faculty. Microbiome research is clearly being performed across training levels. Perhaps most surprising, 41% of the respondents had no programming experience. These results and our experiences emphasize that most individuals carrying out microbiome research have limited experience and are self-taught. There is a growing set of resources to overcome some of these problems including workshops (e.g. Software and Data Carpentry), literate programming tools (e.g. knitr and IPython), and online instruction (e.g. Codecademy). However, these efforts do not include capstone projects where learners can apply their training to a large, domain-specfic project. Our group is beginning to develop a microbiome-specific curriculum to improve research reproducibility and it is clear that there is considerable enthusiasm for this type of training.

Lack of access to original data and code from published studies. Even if a researcher was trained in the best practices of microbiome research, the fact that many investigators do not make their raw data publicly accessible, even though it is required by the journals and funding agencies, is a significant problem. As a small example, we are interested in understanding the differences between the microbiota found in people living in a Western culture and that of people in these different indigenous communities. Three of the four recent papers describing the microbiota of indigenous peoples have yet to make their raw data publicly accessible [refs]. After much effort, we have only been able to acquire one of the three datasets from the authors. Clearly, our ability to address this question has been stalled. The lack of transparency extends to the publishing of incomplete methods descriptions that are frequently at odds with the papers being referenced and the use of “in house” scripts and pipelines that are not disseminated with the final manuscript. The field would be well served by an effort to quantify the accessibility of raw data and the code and methods used to analyze those data.

Inter-study reproducibility. A larger and perhaps more important issue is the lack of explicit validation of observations using different populations of people. For example, one of the keystone results of microbiome research has been the association between the composition of the gut microbiota with obesity [refs]. It was initially stated that obesity correlated with microbiota diversity and ratios of bacteria in the gut. Two independent groups have since used additional published datasets to validate these predictions [refs]. Both failed to reproduce the original results. Given the great fanfare that much microbiome research has been greeted with, we need to expand the scope of this effort to validate studies related to the definition of health, inflammatory bowel diseases, cystic fibrosis, and others.

Considering the importance of microbiome research in understanding many diseases with major public policy implications, there is a great need to directly address the problems related to inadequate training, limited access to data and methods, and an absence of a culture that attempts to validate results in different populations of subjects. The two innovative objectives outlined below will address these issues.

Objective 1. Assess the level of methodological transparency of microbiome studies

Two challenges affect one's ability to reproduce the analysis of others. First, it is well-established that restrictions on the length of papers and the overall format of research papers results in abbreviated descriptions of analytical methods [refs]. Frequently, the methods describing the data curation steps in microbiome analyses are distilled to at most two or three sentences that cite other studies. In reality, the process of manipulating raw sequence data into a format that can be used to address specific biological questions are complex and depend on myriad options that frequently go undocumented. In some cases papers are cited that actually describe multiple methods and so it becomes unclear what was actually done. Second, journals and funding agencies have made great strides to require that researchers make their raw data publicly accessible [ref]. Unfortunately, in practice, these requirements involve researchers self-reporting and rarely involve independent confirmation that database accession numbers are provided or that the links are live. Clearly, if data are not publicly available, it will be impossible for other researchers to reproduce and build upon the initial work. This limits the checks and balances in science and limits the progress of the field. Together, these two problems limit the transparency of the scientific process. To solve these problems it is important to first quantify the actual practices among microbiome researchers.

To quantify these practices, we will perform an audit of microbiome papers. Although it is likely that the specific questions will evolve as we begin to screen papers, our preliminary list of questions will include:

- Is there a description of where the raw data are deposited?
- Are the data actually available at the stated location?
- How many sentences are devoted to describing data processing?
- Are the actual commands available as supplementary material or in a repository?

This audit will be performed by creating a paper crawler written in the R programming language. It will query PubMed for microbiome papers that generate sequence data, retrieve html versions of those papers, and then parse the html to answer these questions. Human oversight will confirm that the results retrieved by the paper crawler are correct. With the results of this screen, we will assess whether the results vary by the scope of the journal, the journal's impact factor, the number of microbiome papers previously published by the authors, and the scope of the research question. Our *a priori* hypothesis is that there is a general lack of transparency in describing methods and making data available. Disseminating these results will provide a benchmark for where the microbiome literature is currently that can be used to quantify the evolution of practices over time. Furthermore, by developing a rubric to assess reproducibility we will create a badging system that we can assign to microbiome studies that describe their reproducibility practices.

Objective 2. Quantify the ability of researchers to reproduce reported analyses

As described above, the limited space allotted to authors for describing their methods is portrayed as one of the factors that limits the ability of others to implement new methods and reproduce the work of others. Yet, the foundation of science is the ability to reproduce the work of others and then take the next step. To circumvent these limitations, a small, but growing number of researchers, including ourselves, have taken to releasing the code that was used to convert raw sequencing

data to the final results. Using tools such as R-based knitr documents and IPython notebooks it is possible deposit these materials on public repositories such as GitHub. Because GitHub is based on the version control software git, it is even possible to see the complete history of the data analysis process. This represents the ultimate in transparency and openness. Although anecdotal, we have received several emails from individuals that have explored our repositories (https://github.com/SchlossLab) to learn how to adapt our analysis to their research questions as well as others that intend to use our data to address their specific questions. By presenting a fully reproducible workflow, we have made this process easier for them and we are able to present a positive spin to the problem of reproducibility. Currently, we do not know how reproducible the analyses reported in the current literature are. Although the lack of reproducibility is clearly a negative, it is important to quantify the level of reproducibility in the field and the covariates that effect reproducibility.

The goal of this objective is to host in-person and virtual reproducibility parties where participants will reimplement the methods described in previously published papers to determine whether they can reproduce their results as described. The results of this objective will further the training goals of improving the reporting of methods used in the microbiome literature and quantify the level of a perceived reproducibility crisis within the microbiome literature. Through these reproducibility parties, we seek to address three specific questions:

- What percentage of papers can be reproduced by other researchers?
- What factors covary with reproducibility (e.g. journal, bioinformatics training of authors, previous microbiome publications, etc.)?
- Does the ability of an individual to reproduce the work depend on characteristics of that individual (e.g. training history, previous microbiome publications, etc.)?

We will broadly recruit an international and diverse cohort of microbiome researchers to participate in several in-person and virtual reproducibility parties. Upon recruitment, the participants will be given a survey to characterize their background and knowledge of the microbiome literature. All cohort members will be asked to complete autotutorial materials that are being developed with NIH funding to train researchers in the best practices of performing reproducible microbiome data analysis. We will identify a collection of 25-50 microbiome papers that provided their raw data. Then members of the cohort will be assigned to one of the datasets and asked to address one specific claim from the paper while doing their best to follow the methods described in the paper. We will randomly assign three individuals to each paper so we can assess variation in the ability of cohort members to implement the methods and individuals will be able to work on multiple datasets. Cohort members will be blinded to the identity of the authors of each paper and to the identity of the other cohort members participating analyzing the specific claim. Finally, all participants will be asked to use the methods described in the autotutorial to document the actual data analysis steps they used and to submit the documentation to the project's GitHub account. Three reproducibility parties will be hosted at the University of Michigan. Those unable to travel to Ann Arbor, MI will be able to participate at their home institution and interact with the project using google Hangouts and other networking tools. At the end of the study, all of the repositories within the project's GitHub account will be made public and the result published with all cohort members as co-authors. Based on our experience with a large number of microbiome investigators, our a priori hypothesis is that at least 25% of published research cannot be reproduced independently from the original researchers. Beyond assessing the reproducibility of the microbiome literature we anticipate that this objective will improve the training and practices of the cohort members.

Objective 3. Determine whether similar microbiome studies validate each other

Problems related to reproducibility in science have come to the forefront because of studies such as those carried out by Bayer, Amgen, and others who attempt to repeat the entire experimental design of previous biomedical research studies. There have also been concentrated efforts to do the same in the psychology literature. In general, these efforts indicate a high level of poor reproducibility. Such an organized approach has yet to be proposed for microbiome studies. Due to the explosive growth of the microbiome field, it appears that replication studies are already being performed informally. In general, authors will acknowledge the previous results of a study, but will not dig into the actual data to see whether the actual observations of their study replicate what was seen previously. The beauty of providing others with one's data and explicit methods is that others can explore the prior work to answer their own questions. Unfortunately, there are only isolated cases where the results of these studies are being synthesized to determine whether the claims of each study are supported by the others. Both cases of meta-analyses have concerned the role of the microbiota in obesity and both confirm that the initial reports out of the Gordon lab, which implicated differences in diversity and the ratio of Bacteroidetes to Firmicutes were not been replicated in other cohorts [refs]. As critics of replication studies correctly point out there are many reasons for why a study might fail to replicate [refs]. These include use of different protocols across studies, non-standard definitions, and differences in the underlying populations being studied. Regardless, if a signal is biologically relevant, one would expect it to transcend these issues.

The goal of this objective is to apply a common set of analytical methods across datasets that have considered a common question to determine whether they replicate each other. Examples include validation of microbiome-derived biomarkers associated with:

- Colorectal cancer [refs]
- Inflammatory bowel diseases including Crohn's disease and colitis [refs]
- *Clostridium difficile* infection [refs]
- Differences between Western and "primitive" or indigenous populations [refs]

These areas of interest have been explored by multiple studies but they have not been analyzed to address a common question. For example, our research group has not found a difference in the diversity of the gut microbiome of healthy individuals and those with colorectal cancer. By analyzing multiple datasets that have characterized similar cohorts we will be able to determine whether the difference in diversity is real or an artifact. In addition to testing observations across multiple studies, this objective will give us the opportunity to analyze a large number of datasets using a common data analysis approach. This will allow us to ascertain how sensitive observations are to variation in approach. While the goal of Objective 2 is to determine whether the stated methods are reproducible, the goal of this objective will be to determine whether results are reproducible across studies when a common, well-justified approach is applied to the raw data.

Summary

The development of new sequencing technologies has fueled interest in the human microbiome as a key component in transitions between health and disease. The explosion of this research area has resulted in the generation of many large datasets, provocative results, and an unease that the role of the microbiome may be overstated. By completing the proposed objectives, we will be

able to quantify the extent of a reproducibility problem in the field. Perhaps most promising is the prospect of using these objectives to educate novices and lead the field in the best practices for insuring that future analyses are more reproducible.

References

1. **Schloss PD**, **Westcott SL**, **Ryabin T**, **Hall JR**, **Hartmann M**, **Hollister EB**, **Lesniewski RA**, **Oakley BB**, **Parks DH**, **Robinson CJ**, **Sahl JW**, **Stres B**, **Thallinger GG**, **Horn DJV**, **Weber CF**. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology **75**:7537–7541.