Assessing the reproducibility of published microbiome analyses by teaching scientists methods to improve their own reproducibility

Patrick D. Schloss, PhD

April 29, 2015

Background. 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". 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. 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. In 2010, 78 papers were published that included the keywords microbiota or microbiome. In 2014, there were over 4,000. 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 (1).

This explosion in interest has created a unique scientific environment. Microbiologists who have been classically trained as reductionist molecular biologists are now expected to be holistic molecular ecologists integrating the tools of statistics, computer programming and bioinformatics, ecology, and clinical science. As developers of the most widely used software package for analyzing the data generated by these projects we have first hand experience of the difficulties novices experience. Our experience suggests that these difficulties have a significant impact on the ability of others to reproduce their research. In addition, because novices do not have the ability to discriminate between competing methods, they may not perform the ideal methods. Instead they engage in a bioinformatics game of "telephone" where they do what previous researchers in their lab told them to do. 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 (1).

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. Quantify the factors that affect the ability of researchers to reproduce each other's analyses.***

- ***Objective 2. Assess the reproducibility of***

Both of these objectives will improve the reproducibility of research within the microbiome community. Perhaps most importantly, they will be achieved through outreach and training activities across the microbiome research community. Given the significant role of the microbiome in human health improving the reliability of the results form these studies will have a meaningful positive impact.

budgets: 2-3 yrs proposal: 2-5 page executive summary public policy - broadly defined

Objective 1 The question wouldn't be to identify fraud, rather to determine whether another person could follow the published methods to get what the initial investigators get. This would allow us to quantify the scope of any reproducibility problems we have in the field and the covariates that effect reproducibility. On a positive note, by enlisting the help of the community we would further the training goals of the first aim.

Objective 2 In other words, can I take data from a researcher that used their favorite analysis pipeline and get their results if I use my pipeline? This would assess the sensitivity of results to variation in analytical methods. This of course underlies the robustness of the science. This aim would also underscore the importance of reproducibility and the goals of the other aims.

Significance

The NIH-funded Human Microbiome Project (HMP) Roadmap Initiative engendered great enthusiasm in understanding how the structure and function of the microbiome relates to human health {The Human Microbiome Consortium, 2012 #2616; The Human Microbiome Consortium, 2012 #2617}. This initiative has resulted in an expansion of support at the NIH, who now supports these research efforts across 16 of the 27 institutes, centers, and offices. The funding for microbiome research more than doubled between fiscal years 2010 and 2012 to a total between $120 and 150 million for nearly 300 grants and contracts (Lita Proctor, NHGRI, personal communication). This financial support, combined with the development of next generation sequencing platforms resulted in a meteoric rise in microbiome-related publications (see figure at right). This massive expansion in microbiome research has relied on biomedical researchers to do their own bioinformatics leading to a common complaint that researchers are either unaware of or underserved by existing tools {Mardis, 2010 #3452;Gevers, 2012 #2711}. The experience across the microbiome research domain has largely paralleled that of other biomedical research areas were, as the RFA for this competition notes, "graduate students were often significantly dependent on the mentor or the mentor's lab for the training received, and postdoctoral fellows were primarily dependent on the mentor or mentor's lab at all institutions. Rather than being learned in prescribed curricula, training in good laboratory practices that influence data reproducibility appears to be largely passed down from generation to generation of working scientists, with substantial variation from laboratory to laboratory." It is clear that microbiome research is an important part of the NIH portfolio and that there is a dearth of training for its practitioners to insure reproducibility of methods. The proposed project will provide this growing community of researchers the tools they need to improve the reproducibility of their research. ***This will yield a significant vertical step in the field because we will have greater confidence in the results and we will be better enabled to use previous studies as a launch point for further investigations.*** This project will yield multiple benefits to NIH-funded projects and beyond. First, although the current proposal focuses on microbiome research, it is reasonable to expect that with some customization, the materials could be easily tailored to other disciplines where novice practitioners are implementing their own bioinformatic analyses. Second, microbiome research is a specialized form of microbial ecology. Other microbial ecologists who study environments as diverse as hydrothermal vents and soil {Lesniewski, 2012

#2732;Schloss, 2006 #992} will benefit from the training they can receive through the proposed instructional materials.

Innovation

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 2,200 times, making it the most widely cited software package for analyzing microbiome data {Schloss, 2009 #1816}. The mothur project has allowed us to engage thousands of researchers through the wiki, forum, and mailing list and through face-to-face workshops held at conferences, universities, and as part of larger training initiatives. We recently conducted a survey that was advertised through our newsletter and received 170 responses. First, 41% of users were graduate students, 41% were PhD-level scientists, and 12% were faculty. Second, 48% of the respondents have used mothur within Windows, 32% within Mac OS X, and 45% within Linux. Finally and perhaps most surprising, 41% had no programming experience. The others reported knowing R (35%), Perl (32%), Python (27%), C/C++ (16%), Java (10%), or some other language (10%). These results and our experiences emphasize that most individuals carrying out microbiome research have limited experience in performing bioinformatics research and are largely self-taught. In fact, from these experiences, we have met the graduate students, postdocs, and faculty that the quote from the RFA above describes and know that we are in a unique position to engage this community. Although there are some exceptions, the *status quo* has generally involved 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. ***Therefore, the proposed research is innovative, in our opinion, because it will fulfill and unmet need to develop instructional materials to train the growing number of microbiome researchers best practices and insure that their analyses are reproducible by others.*** Once it is possible to confidently reproduce analyses it will be possible to move on to determining whether the analyses were done well and to build off of previous analyses to expand our knowledge of how the microbiome affects transitions between health and disease.

budgets: 2-3 yrs proposal: 2-5 page executive summary public policy - broadly defined

1. **Abrams PA**, **Ruokolainen L**, **Shuter BJ**, **McCann KS**. 2012. Harvesting creates ecological traps: Consequences of invisible mortality risks in predatorprey metacommunities. Ecology **93**:281–293.