JOURNAL OF MICROBIOLOGY & BIOLOGY EDUCATION, March 2016, p. 46-50
DOI: http://dx.doi.org/10.1128/jmbe.v17i1.1034

Turning Participatory Microbiome Research into Usable Data: Lessons from the American Gut Project

Justine W. Debelius[1], Yoshiki Vázquez-Baeza[2], Daniel McDonald[1], Zhenjiang Xu[1], Elaine Wolfe[1], and Rob Knight[1,2]*

[1]*Department of Pediatrics, University of California, San Diego, CA 92110,*
[2]*Department of Computer Science and Engineering, University of California, San Diego, CA 92093-0404*

The role of the human microbiome is the subject of continued investigation resulting in increased understanding. However, current microbiome research has only scratched the surface of the variety of healthy microbiomes. Public participation in science through crowdsourcing and crowdfunding microbiome research provides a novel opportunity for both participants and investigators. However, turning participatory science into publishable data can be challenging. Clear communication with the participant base and among researchers can ameliorate some challenges. Three major aspects need to be considered: recruitment and ongoing interaction, sample collection, and data analysis. Usable data can be maximized through diligent participant interaction, careful survey design, and maintaining an open source pipeline. While participatory science will complement rather than replace traditional avenues, it presents new opportunities for studies in the microbiome and beyond.

INTRODUCTION

The human microbiome is a poorly understood, but critical, component of health. Community structure is influenced by many factors, including genetics, diet, and xenobiotic and antibiotic use (4, 5, 9, 15). The gut microbiome, in particular, plays an important role in metabolism, immune development, and endocrine and neurological signaling (10, 16). Dysbiotic gut communities have been associated with a host of human diseases including obesity, inflammatory bowel disease, type I and type II diabetes, autism, multiple sclerosis, and malnutrition (3, 16). The gut microbiome can predict risk for conditions like Crohn's disease (8). Fecal material transplant may also transmit clinical phenotypes in some cases: one report suggested a donor transmitted a risk for obesity to her human recipient along with her stool, while trans-species transmission of obesity is well established (1, 17).

Human microbiome work has primarily focused on case-control studies of a few dozen to a few hundred individuals. Budget restrictions and strict disease focus by funding agencies often limit the size and scope of investigation. Although studies supported by traditional mechanisms have led to considerable advances, there are also major pitfalls with the traditional approach. Small cohorts create inconsistent observations among studies. Trends in community structure are often shared between studies, but the individual taxa driving these trends often are not. For example, studies have found correlations between obesity and both an increase and a decrease in *Methanobrevibacter smithii* (19). Meta-analysis can ameliorate inconsistencies due to data analysis, although it cannot correct for differences due to sample handling or the characteristics of the control and clinical groups (14). The problem is compounded by the absence of effective, mathematically justified ways to quantify effect size or the signal-to-noise landscape in the microbiome.

Even previous efforts to define "healthy" microbiomes have been small compared with cohorts used for other types of studies, such as genome-wide association studies. The Human Microbiome Project (HMP) focused on 252 healthy professional students in their twenties and thirties living in two regions of the United States (11). The HMP contributed valuable information about the microbiome, including the variation in taxonomic abundance in healthy adults and the lack of a core healthy microbiome. However, the HMP did not answer all the open questions about the healthy microbiome. For instance, the cohort was not well suited to describe how the microbial communities change between age groups or what a healthy microbiome looks like relative to dietary or lifestyle choices.

Public participation in microbiome research, through crowdsourcing and crowdfunding, may provide some

*Corresponding author. Mailing address: 9500 Gilman Drive MC 0763, La Jolla, CA 92093-0763. Phone: 858-822-2379. Fax: 858-246-1981. E-mail: robknight@ucsd.edu.

potential solutions to these problems. Both models transform science into a public, participatory area, rather than a practice for experts in semi-isolation. Crowdfunding involves the public in science by asking for a monetary investment in a project. Lay people can determine what they consider worthy or unworthy of funding, whether it be comparative studies of the cat microbiome (https://fundrazr.com/campaigns/410aC4/ab/f4vYF9?) or a qualitative survey of the best burritos in San Francisco (https://experiment.com/projects/qualitative-survey-of-burritos-in-san-francisco). Crowdsourcing engages the public in collecting the data to be analyzed. Individuals participate by contributing data or samples. This typically involves the contribution of observational data, such as bird sightings or flu symptoms, in projects like Flu Near You (https://flunearyou.org), but may also involve crowdsourced data collection, modeled by the Personal Genome Project (www.personalgenomes.org), or even crowdsourcing data analysis, through platforms like the online games Foldit (https://fold.it/portal/) and EteRNA (http://eterna.cmu.edu/web/). Crowdsourcing may open opportunities to access populations, areas, or information that is difficult for a finite group of researchers to access. It also offers opportunities in exploratory science: the wealth of data allows for a degree of exploration that can be more difficult in traditionally sourced studies, where participant recruitment is more focused.

The American Gut Project (www.americangut.org) is a crowdfunded, crowdsourced microbiome project run through the University of California at San Diego, which was initiated as a collaboration between the Earth Microbiome Project and the Human Food Project. Participants provide a physical sample (fecal, oral, skin, pet, or environmental), answer a survey about their health, lifestyle, and diet, and a make a monetary contribution that covers the cost of microbial DNA sequencing. Individual participants receive a report describing their results. De-identified data are also deposited in a public repository. We have used American Gut to draw conclusions about factors that affect participant health in the human microbiome (Debelius, McDonald, et al., in preparation). Here, we present three stages that have been important for aggregating the American Gut results and presenting usable data.

CRITICAL CONSIDERATIONS FOR GETTING USEFUL DATA FROM CROWDSOURCING

Communication is central to successful science, especially crowdsourced science. There is an added complexity in disseminating research to the general public, because complex concepts must be translated into messages that can be readily digested by individuals without specific domain knowledge. The inherent difficulty is magnified in participatory science, as there is a continual interaction with members of the general public. The challenge of communication can be broken down into three major areas critical to crowdfunding: participant recruitment and retention, data collection (both sample and metadata collection and quality), and data dissemination.

Participant recruitment and retention

The first area, recruitment and retention, is a crowdsourced project's initial and primary interaction. At the outset, members of the public are unlikely to be interested in your project if they are unable to understand why you are doing the project in the first place. They also want to know how they benefit by participating. In the case of the American Gut Project, one of our goals was to provide an avenue through which members of the general public could engage in cutting-edge research and, in turn, learn about the organisms that inhabit their bodies.

Participatory science can be self-selecting, and this may create a biased cohort, rather than a true representation of the population. Gut microbiome research, for example, may be more likely to attract individuals with diagnosed gastrointestinal conditions, such as inflammatory bowel disease (IBD). In the American Gut, we see a six-fold enrichment in participants with IBD compared with the US population as a whole (Debelius, McDonald et al., in preparation). Sponsors have also contributed funds to provide kits for participants in other populations of interest, including children with autism spectrum disorder. These sub-studies may lead to an understanding of compositional patterns associated with these specific populations. Other, less explicit biases may also appear in the data. The role of the Internet in participatory science cannot be discounted, meaning that participation is likely linked to Internet access (6). Coupling crowdfunding to crowdsourcing may limit the participant population to those able afford the cost. The financial burden may also create self-selection, even for those with the available disposable income. Many of the early American Gut participants were individuals who emphasized the importance of diet in health, and therefore tended toward more extreme dietary choices. These implicit biases in the population may be hard to identify, and harder to correct (although they are less important for the original goal of the project in terms of identifying the diversity of types of microbiome "out there in the wild"). Decoupling crowdsourcing and crowdfunding, at least for some cohorts, may help ameliorate some biases in data.

A second important interaction with participants arises when, inevitably, participants have questions. Mismanagement of the participant base can be a major reason projects fail (7). The help burden stems not just from the number of questions coming in, but the number of personnel hours necessary to answer these questions. Given the nature of crowdfunding, the rate at which a project will grow is not known in advance, which makes scoping personnel effort difficult and risky (e.g., if the project "fails"). In microbiome research, the potential for health discoveries adds a new level of complexity. Participants and backers may choose to engage in a project in which the research personally benefits

them. For the American Gut Project, despite all our efforts at dispelling the notion that the data generated have current medical value, we still frequently receive questions along the lines of "I have condition X. Given my microbiome, what do you recommend I do?"

Sample collection and quality

Once participants are recruited, the next major hurdle is collecting their data. In microbiome studies, this typically involves a physical sample, or set of physical samples, and information about the participant and sample. Physical sample collection poses a challenge for biologically-based projects. The sampling protocol needs to be simple and safe. However, even simple protocols can be complicated for novices without clear instructions. The unfortunate reality is that people are bad at following instructions (for example, we anticipate that few of the readers of this article read their cell phone manual cover to cover). Explicit, succinct, and engaging instructions are vital to minimize variability in how instructions are followed. To this end, the American Gut Project took two approaches. The first is an eye-catching "quick instructions" sheet that gives a rundown of the necessary steps. In addition, detailed instructions are provided, including video examples on the website. During the course of the project so far, it has been necessary to revise the instructions, based on feedback from participants, and address obvious issues with sample collection. Notably, we discovered that the amount of fecal matter to send in was ambiguous, leading us to provide graphic examples of good and bad samples. As we refined the instructions, we encountered fewer questions, and higher quality samples were returned.

Metadata collection and quality

The human microbiome is contextually dependent, making it impossible to understand a microbiome community without information about its host (12, 18). Therefore, participant and sample metadata (i.e., contextual information) are also an important consideration in participatory microbiome research. The goal of metadata collection is to maximize the amount of accurate, usable data that can be collected for every sample. Survey design and implementation can support or impede this end. Although it is possible to analyze a few dozen free response fields for a small number of samples, it is prohibitive to analyze large numbers of free-response fields for large numbers of samples. Free response fields are also more likely to contain human error: in the American Gut dataset, individuals have reported chicken as their most common carbohydrate, which would be surprising if true (standard nutritional data for chicken breast report zero carbohydrates). Questions with controlled vocabulary, such as multiple-choice questions or fields limited to accept bounded numeric responses, can help improve accuracy. It may also be important to consider the level of detail that is possible to record in a survey. Controlled vocabulary represents one of these trade-offs. Another is the decision of whether or not to pursue information about a specific medical condition. The American Gut has addressed these issues with triggered response questions, condition-specific surveys, and the option to follow up with participants.

Metadata errors are inevitable—whether in self-reported data or well-funded clinical studies (2). There are two major considerations with error reporting: how the errors are identified and the way the errors are corrected or removed. Identifying obvious errors can be easy. In the American Gut, participants who reported birth dates prior to the start of the twentieth century were identified as obvious errors. There are also profound differences between adult microbial communities based on body site, which can help when participants forget which sample was collected on which swab (11). However, other errors can be more difficult to identify. In certain American Gut analyses, we noticed that alcohol had a larger effect than antibiotic use, and that infants (birth to three years of age) had microbiomes that were more diverse than older children; a contrast with previous publications (20). When we examined the infant data further, we identified several individuals with age listed as less than three years of age but self-reported height over four feet and reported drinking more than once a week, leading us to question the age data. In a large dataset, it can be useful to remove clearly erroneous information, especially if the correct answer is difficult to determine. Age values that are likely incorrect, given the rest of the contextual information, are therefore removed from analysis within the American Gut data. Mislabeled body sites can be corrected, even against a high background mislabeling rate, using a supervised learning technique, due to the strength of the association between body site and community structure (13). The same associations may be true for other parameters as we continue to collect data.

Data analysis and dissemination

Data dissemination and communication is a final step in the scientific process. In a traditional scientific model, this has taken the form of publication in grant reports, scientific journals, and the deposition of data to repositories. Participatory science opens questions about data ownership, dissemination, and communication. Rather than delivering results to a grant committee of peers, scientists instead must communicate results to a wider community. In crowdsourced projects, individualized results may be offered as an incentive for participation. When the project focuses on characterizing human biology, it may be challenging to balance providing novel results with avoiding presenting information that could be interpreted as a medical diagnosis. In crowdfunded projects, regular updates showing progress are important to continued investment and re-investment (7); for a scientific project, this can mean everything from

a blog with regular updates to a public release of data and analyses techniques.

Providing aggregated crowdsourced data to the general public can also crowdsource the analysis. It sends a clear message that the data are owned by the public. Large datasets present opportunities for exploration, new technique development, and technique refinement. Providing the dataset to a collaborator network early on fosters opportunities for new analyses and directions. Collaborations that play on the strengths and expertise of each group can accelerate the rate of discovery. Making the full dataset available through open access mechanisms early in the analysis process is one of the simplest ways to disseminate data to multiple collaborators at a variety of institutions.

However, data release can raise privacy concerns. Institutional Review Board (IRB) protocols must make it clear how participants' de-identified data can and will be used. Participants' de-identified microbial DNA sequence data and per-sample and per-individual metadata will be made publicly available if that is a goal of the project. Releasing data into repositories without monitoring may make dissemination easier, but it can also mean that after participants withdraw, their data cannot be retracted. Additionally, extensive care has to be taken to avoid compromising the anonymity of the participants. Such steps include separating clearly identifying participant data from survey information; limiting access to raw survey answers; and removing identifying information from publicly available survey results, even inadvertently identifying information. To this end, the surveyed data must be validated against possible identification threats; for example, a combination of date of birth and zip code could provide an attacker with the identified personal information of a participant.

PROSPECTS

Crowdfunding and crowdsourcing, while powerful ways to fund projects, recruit participants, and raise public awareness and interest, are novel approaches and have their own pitfalls. The nature of a crowdfunded project requires different approaches from traditional study designs and considerations, especially with respect to public relations and communication. Defining the intention and standing of the project is vital when individuals have a personal and financial stake. Communication of the project expectations, what participants can expect to receive, and progress of the project and of the participants' specific samples, especially if there is a waiting period between financial contribution and tangible results, cannot be overlooked. The participants themselves also must be considered. The topic of the crowdfunded research project is almost certainly expected to draw in a specific subset of the population, leading to potentially biased sampling. The financial aspect of participation may exclude an additional subset due to inability to afford participation (although this can be ameliorated by supplementing crowdfunding by philanthropic contributions and/or foundation support). Additionally, considerations of how to reduce and respond to errors in the data must be considered. Data dissemination, in the form of individualized results, and sharing analysis tasks can also benefit or hinder projects.

In summary, citizen science provides a new opportunity for microbiome research. While it is unlikely to replace grant funding from government and private agencies, it may act as an additional mechanism for answering questions that are difficult to explore through traditional means.

ACKNOWLEDGMENTS

Thanks to Dr. Embriette Hyde and Joshua Shorenstein for their help during manuscript preparation. The American Gut Project acknowledges financial support from Second Genome, MoBio, and thousands of members of the public. The authors declare that there are no conflicts of interest.

REFERENCES

1. **Alang, N., and C. R. Kelly.** 2015. Weight gain after fecal microbiota transplantation. Open Forum Infect. Dis. **2**:ofv004.
2. **Archer, E., G. Pavela, and C. J. Lavie.** 2015. The inadmissibility of what we eat in America and NHANES dietary data in nutrition and obesity research and the scientific formulation of national dietary guidelines. Mayo Clin. Proc. **90**:911–926.
3. **Brestof, J. R., and D. Artis.** 2013. Commensal bacteria at the interface of host metabolism and the immune system. Nat. Rev. Immunol. **14**:676–684.
4. **David, L. A., et al.** 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature **505**:559–563.
5. **Dethlefsen, L., and D. A. Relman.** 2011. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. Proc. Natl. Acad. Sci. U. S. A. **108**(Suppl):4554–4561.
6. **Gerber, E. M., J. S. Hui, and P.-Y. Kuo.** 2012. Crowdfunding: why people are motivated to post and fund projects on crowdfunding platforms. Proc. Int. Work. **10**.
7. **Gerber, E. M., and J. Hui.** 2013. Crowdfunding: motivations and deterrents for participation. ACM Trans. Comput. Interact. **20**:32.
8. **Gevers, D., et al**. 2014. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe **15**:382–392.
9. **Goodrich, J. K., et al.** 2014. Human genetics shape the gut microbiome. Cell **159**:789–799.
10. **Hooper, L. V., D. R. Littman, and A. J. Macpherson.** 2012. Interactions between the microbiota and the immune system. Science **336**:1268–1273.
11. **Human Microbiome Project Consortium (The).** 2012. Structure, function and diversity of the healthy human microbiome. Nature **486**:207–214.
12. **Knight, R., et al.** 2012. Unlocking the potential of metagenomics through replicated experimental design. Nat. Biotechnol. **30**:513–520.

13. **Knights, D., E. K. Costello, and R. Knight.** 2011. Supervised classification of human microbiota. FEMS Microbiol. Rev. **35**:343–359.
14. **Lozupone, C. A., et al.** 2013. Meta-analyses of studies of the human microbiota. Genome Res. **23**:1704–1714.
15. **Maurice, C. F., H. J. Haiser, and P. J. Turnbaugh.** 2013. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. Cell **152**:39–50.
16. **Neuman, H., J. W. Debelius, R. Knight, and O. Koren.** 2015. Microbial endocrinology: the interplay between the microbiota and the endocrine system. FEMS Microbiol. Rev. **39**:509–521.
17. **Ridaura, V. K., et al.** 2013. Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science **341**:1241214.
18. **Subramanian, S., et al.** 2014. Persistent gut microbiota immaturity in malnourished Bangladeshi children. Nature **510**:417–421.
19. **Walters, W. A., Z. Xu, and R. Knight.** 2014. Meta-analyses of human gut microbes associated with obesity and IBD. FEBS Lett. **588**:4223–4233.
20. **Yatsunenko, T., et al.** 2012. Human gut microbiome viewed across age and geography. Nature **486**:222–227.