# Research Education Program Plan

## Significance

### Importance of the problem to be addressed

Insuring that biomedical research is performed in a rigorous and reproducible manner is critical to the advancement of science and improvement of human health. Significant emphasis has been placed on improving the rigor and reproducibility of laboratory science by improving the description of protocols, confirming the authenticity of strains and reagents, and improving experimental design [REF]. Implementing rigorous and reproducible practices in the analysis of the resulting experimental data has not received the same level of attention. This is perhaps because bench scientists receive extensive training in how to do laboratory techniques and the discussion of improving rigor and reproducibility fits into traditional laboratory training. This training has not incorporated data analysis skills into existing courses or by creating stand alone data science courses. Although many programs may require a course in statistics, these courses focus on experimental design and choosing the appropriate statistical tests. They rarely discuss data management, data curation, data visualization, or data dissemination. Furthermore, faculty who appreciate that they need to develop these skills have limited availability to do so. Given the heightened emphasis on rigor and reproducibility and the broad adoption of technologies that generate massive datasets, there is a great need for these skills. Because trainees’ time is limited for activities outside of developing laboratory skills or for what is relevant to learning the background literature of their sub-discipline, training in data science skills has been limited. To overcome this problem, workshops (also referred to as boot camps, short courses, or short form training) have been grown in popularity because they can provide a lot of information in a short period of time. For 2016, it was estimated that NIH and NSF provided such programs $27.8 million [REF]. Participants have positive experiences in these workshops and rate the value of the material highly. Yet, it is necessary to ask whether these intensive training activities are effective [REF]. A 2017 analysis measured learning outcomes and found that such activities did not have a statistically significant effect on learning [REF]. **Given the considerable time and financial investments made in these activities, it is important that we find effective means of training scientists the best practices in performing reproducible data science.** Furthermore, if poor data analysis practices persist, they will continue to undercut the rigor and reproducibility of biomedical research. This problem is central to the RFA that this proposal is in response to, which calls for the development of “exportable training modules with the potential to enhance the scientific rigor, reproducibility, and responsible conduct of biomedical data science research, and to provide for communication and coordination of the development and deployment of such modules.”

### Rigor of the Prior Research Supporting the Proposed Research Education Program

Examples of workshops

Workshop study

Pedagogy \* Repeated practice vs drill/kill \* Chunked learning \* Value of testing \* Cycle - incomplete mastery before moving on \* Deliberate practice

Short form offerings \* online resources - cost - long - lack of community \* youtube videos - short demos of how an individual concept works - no question - live coding of people doing tasks - no question - they are performances and not didactic

**preliminary results**

Workshops with personal connection \* MICRBIOL 612 \* Carpentries workshops \* minimalR/generalR workshops

Riffomonas \* Video series \* Stats about views \* Publication \* This proposal is an extension of the ideas in those videos

Code Club \* Lay out format \* Examples of topics \* Videos during COVID-19

### Significance of the Proposed Research Education Program

Successful completion of the proposed Research Education Program will result in a library of resources that individuals or groups of researchers can use to engage in repeated practice of concepts important in conducting data analysis. ***This contribution is expected to be significant because it will address the problem of wasting the significant resources that are extended to participate in workshops only to be ineffective because participants do not have the additional resources for deliberate practice.*** It is likely that similar types of repeated practice activities would improve learning in areas where researchers also use workshops to engage in intensive learning activities including laboratory skills and safety training. Central to the proposed research is the problem that researchers participate in workshops with every intention of learning to program. The leave the workshop enthusiastic and feeling like they have learned a lot. Then they struggle to find opportunities to apply their skills. Because they fail to practice the material in the weeks following the workshop, they lose those skills. When another workshop is offered, the dutifully sign up again hoping that the outcome will be different. The materials developed for the proposed Research Education Program will provide opportunities to practice what was covered in the workshop, breaking the cycle of learning and forgetting.

## Innovation

The *status quo* as it pertains to bench scientists developing data analysis skills is for them to take short and intensive workshops. This approach works well if they have an immediate need for these skills; however, this is rarely the case and the learner hopes to retain enough information from the workshop to apply it when they reach the data analysis portion of their project. The reality is that the bench scientist typically forgets the information by the time they are ready to use it. They have effectively crammed as much information as they could during the workshop hoping to retain it for later application. A consistent message from educational research is that cramming is ineffective, but that repeated and deliberate practice is essential to long term learning. ***The proposed Research Education Program is innovative, in my opinion, because it represents a substantive departure from the status quo by providing bench scientists with a library of resources to engage in repeated and deliberate practice of reproducible data analysis concepts.*** The Code Club concept is drawn from traditional Journal Clubs where a paper is presented, critiqued, and used to think of additional research questions. The Journal Club activities teach scientists best practices in experimental design, methods, and interpretation. They build off of coursework to reinforce the concepts covered in the classroom. Similarly, the Code Club format seeks similar goals but with data analysis concepts. Analogous to a Journal Club presentation, Code Club resources will include a motivating research question and the data and data analysis concepts needed to answer that question. Participants will then have the opportunity to answer related questions using the concepts they just learned. With a high volume of resources, participants will see the same concepts multiple times over many sessions and in different contexts. This will serve to deepen their understanding of the concepts and ability to integrate different concepts to answer their own questions. The result will be a better-trained scientific workforce that is able to ask better research questions of their data and answer the questions in a robust and reproducible manner.

## Proposed Research Education Program

## Evaluation Plan

**As described above, participants will register with the module through a website where they will provide demographic data. The website will track the amount of time spent on each item and their responses. The autotutorials will request the participants’ account name so that it is possible to look for their GitHub account to track their performance. It is important to note that aside from the account name, the evaluation will be completely anonymous. In general, it is not possible to ascertain a person’s identity from their account name. Each autotutorial will have a variety of quantitative and qualitative assessments that will allow us to track a participant’s progress through the module. We will create a badge-based commendation system where participants gain varying points. In addition, participants will receive a certificate indicating completion of the program. As we track the participants’ activities within each module we will also be able to evaluate the usefulness of each activity and modify the content to overcome misconceptions or confusion in the material. Finally, we will assess the overall success of the module by sending an automated email to each participant at 1, 2, and 3-year anniversaries of the date they completed the module. These emails will invite the participant to complete a survey that describes their current reproducible research practices. We will monitor how they have changed relative to when they started the module. The evaluation plan that we have developed blends quantitative and qualitative measures of the participants’ learning and behavior as well as the quality of the instructional materials.**

## Dissemination Plan

\*\*We hope that this project will be a seed to engage other developers and the microbiome community to improve the reproducibility of microbiome data analysis. All instructional materials will be made freely available through a website at www.riffomonas.org. To disseminate these materials we will pursue several avenues including:

* Social media (e.g. Twitter and blogs) to promote a reproducible microbiome research initiative. This will create enthusiasm in the microbiome community and foster their interest in developing and improving the modules.
* At least two manuscripts that will address the issue of reproducibility in microbiome research and the development of our teaching modules.
* Microbiome-based e-mailing lists that Schloss (PI) is part of including the extensive mothur mailing list, which currently has more than 3,000 registered users and the Human Microbiome Project Data Analysis and Coordination Center website and mailing list (http://www.hmpdacc.org).
* Networking with other microbiome researchers (see letters of support) to facilitate workshops at 3 universities each of the two years of the project (see the letters of support). This will serve to help us develop the modules as well as disseminate the modules.
* Utilize our connections within the American Society for Microbiology to advertise the module through *Microbe* magazine and in workshop offerings at the society’s General Meetings in 2016 and 2017 (see letter of support from ASM).

In addition to www.riffomonas.org, which will host information about the instructional materials, we intend to continue to develop instructional materials and best practices beyond the funded period of this project to foster greater reproducibility in microbiome research (see Future Directions). Once funding for this project has been exhausted, the project should be able to continue on as a democratic, open source, community-supported endeavor. To support this all materials related to the project will be maintained as a public GitHub project repository (www.github.com/riffomonas). In fact, the development of this proposal is available at www.github.com/riffomonas/2020\_RR\_R25.\*\*

## Principal Investigator

**As indicated by his biosketch and the numerous letters of support, Schloss is a respected member of the microbiome research community and is an excellent teacher that is anxious to utilize innovative teaching methods to communicate complex materials. Over the past 6 years, Schloss has been the PI on 9 research grants funded by NIH and other agencies including 2 R01 projects related to the microbiome. He has served as a co-Investigator on 7 additional projects during that time. From this funding he has published 32 peer-reviewed publications and he was the senior author on 16 of these studies; all of these studies have involved research questions related to the microbiome. At the University of Michigan, Schloss has developed two courses: *Symbiosis* and *Microbial informatics*. The latter is a course that is designed to teach microbiologists in MS and PhD programs and postdocs how to use R. This semester, in preparation for this proposal, Schloss successfully experimented with using knitr-based documents and GitHub repositories to submit assignments. In addition, all of his course materials have been made available through the course’s GitHub site including the lecture slides that were prepared using the R slidify package that is based on the knitr package (microbialinformatics.github.io). Although this course touches on the content of the proposed teaching materials, it focused on developing R programming skills rather than data analysis practices. This course and Schloss’s willingness to experiment with the content is indicative of his innovative approach to teaching. Finally, over the past 7 years, Schloss has offered 4-5 workshops describing how microbiologists can use mothur and R to analyze data from microbiome research projects. This experience has given him a unique perspective into the needs and competencies of the microbiome research community. Together, these data and experiences indicate Schloss is “actively engaged in research in an area related to the mission of NIH, and can organize, administer, monitor, and evaluate the research education program.”**

## Institutional Environment and Commitment

Schloss has secured institutional support for this project on multiple levels. First, as indicated by the letter of support from Dr. Bethany Moore, Interim Chair of the Department of Microbiology & Immunology at the University of Michigan School of Medicine, Schloss has the support of the university to gain access to adequate staff, facilities, and educational resources to make the planned research education program successful. Second, Schloss has interacted with the Center for Research on Learning and Teaching (CRLT) at the University of Michigan to plan the assessment program for this project (see letter of support from CRLT). The CRLT provides a mixture of complimentary and fee-based services, but does not participate in projects as personnel on grant proposals. The support provided by CRLT will insure that Schloss it utilizing the latest in pedagogical theory and assessment to develop the proposed teaching modules. Third, as indicated by the letters of support from other researchers at the University of Michigan and across the United States, Schloss has the support and commitment of other investigators to implement this project. They all see the value of developing instructional materials such as those described in this proposal. The multiple levels of commitment and broad support that this proposal enjoys speaks to its importance and the unique qualifications of Schloss to lead the project.