# 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

***The reproducibility crisis.*** At the height of the “reproducible research crisis” there was concern that most biomedical research was not reproducible [REFS]. Ironically, the reports that heralded this crisis did not provide the level of rigor and transparency that they decried in the studies they claimed could not be reproduced. Less thoughtful commentators might see the lack of reproducibility as a sign of academic misconduct or that a result was incorrect [REFS]. Others may call such work “sloppy” [REFS]. The reality is that academic misconduct is rare, reproducible research can be wrong, and *everyone* struggles to ensure that their work is reproducible. Performing reproducible research practices is hard. The descriptions of a reproducibility crisis put a much needed spotlight on well known difficulties within biomedical research [REFS]. These led to a renewed effort to improve the rigor and reproducibility across the lifecycle of a research project [REFS]. As highlighted at the XXXX Clearinghouse XXXXX, NIH has supported the development of instructional modules including one that we generated for the field of microbiome research [REFS].

***Defining reproducibility.*** There is a general understanding that a result is reproducible if others can obtain the same result that we did. Yet reproducibility and replicability are often used interchangeably or with different definitions. It is impossible to discuss improving “reproducibility” if the term is poorly defined. We previously described a framework for outlining how to think about reproducibility within the field of microbiome research that is easily generalized to other disciplines (Table 1). We can also think of this framework in terms of the data analysis step of a research project.

|  |  |  |
| --- | --- | --- |
| Methods | Same dataset | Different datasets |
| Same methods | Reproducibility | Replicability |
| Different methods | Robustness | Generalizability |

Briefly, if someone were to take my data and methods and generated the same results, then the results would be **reproducible**. While reproducibility should always be achieved, failure to generate a replicable, robust, or generalizable result is not necessarily a failure. If they used the same methods to test a hypothesis using data collected from different populations and got the same results, as one would do in a meta analysis, it would be **replicable**. Failure to replicate a result could indicate that there is some underlying variable that distinguishes the different datasets that needs to be better understood and could point to important biology. If different methods were applied to the same experimental system and the same results were produced, the result would be **robust**. Since not all methods are equally valid and may make different assumptions, failure to achieve a robust result is not necessarily a failure. In fact, as will be discussed next, subjective decisions that are made in an analysis can have a large impact on the conclusions of an analysis. Finally, if different methods were applied to multiple datasets, the results would be considered **generalizable**. Such results are relatively rare and point to overarching theories that drive science (e.g. climate change, evolution). Failure to generalize a result usually indicates that the underlying hypothesis is incorrect.

As an example, Silberzhan et al (10.1177/2515245917747646) performed a study that tested the reproducibility and robustness of the hypothesis that soccer referees are biased against players with darker skin tones. They recruited 29 teams of data scientists to analyze the same dataset to measure the bias and determine whether it was statistically significant. Although the methods varied by team, all of the methods were peer reviewed at multiple stages of the study. This level of oversight and transparency is not typical or practical for most studies, but points to the importance the team placed on reproducibility. In spite of their reproducibility, the variation in study design reflected the teams’ subjectivity and the peer review. The teams calculated odds ratios ranging from 0.89 to 2.93. That 69% of the teams found a significant odds ratio suggests that the relationship between bias and skin tone was robust to differences in methodology. As the authors pointed out, the decisions made at each step are subjective and may have oversized impact on the conclusions of the studies. The key factor in this analysis is that because the individual analyses were reproducible and the methods were transparent, it was possible to understand how subjective decisions affected the robustness of the results.

***Reproducibility.*** As we move forward through this proposal, ***reproducibility*** will be used in the sense of Table 1. Just as lacking reproducibility should not invalidate a result, if an analysis is reproducible, it is not necessarily correct. However, if an incorrect result is obtained from a reproducible analysis, it is much easier to find and correct the problem and test the effect of the error on the final result. The past 10 years has seen a significant growth in the availability of tools to help improve the reproducibility of data analyses. We have found that assuming the analyst will need to reproduce the work in six months is a strong motivator to think more diligently about reproducibility. A six month gap is a good motivator because it is likely that any scientist will have one such gap in their analysis such as between when they finalized their analysis plans and when reviews come back from referees who ask for additional or different analyses to be done. There may be another such gap between when the author is done with the study and when readers start to ask questions about the analysis. If an analyst considers themselves six months from now as a muse, they will be more likely to ensure that the analysis is reproducible by their collaborators and third parties.

To highlight the challenges of reproducibility, Philip Bourne challenged researchers to reproduce his group’s 2010 study “The *Mycobacterium tuberculosis* drugome and its polypharmacological implications”. It is important to note that this study was performed before many of the tools used today to ensure reproducibility were popularized. The team of researchers attempting to reproduce the original work interacted with Bourne and his team to resolve questions. Bourne and his team have an excellent reputation as being concerned with reproducibility and conducting rigorous research. By the time the re-analysis was completed, the team estimated that it would take someone with basic bioinformatics skills 160 hrs to decipher the analysis and another 120 hrs to implement and execute the re-analysis. This would represent 13% of someone’s annual effort to reproduce an analysis. This is a considerable cost and does not even consider the cost of data storage and processing, much less how much longer it would have taken if Bourne and colleagues were not as helpful. Although the re-analysis of the original study was ultimately reproducible, it was only with great expense. In hindsight, the re-analysis effort may have been easier had the original authors had started with the assumption that their analysis would be reproduced by someone in the future.

***Why reproducibility?*** There are three reasons that scientists should be concerned about the reproducibility of their research. The first is a “negative” reason. If a result is not reproducible, then other researchers will doubt the rigor of the original study and there will be concerns regarding misconduct. Furthermore, failure to reproduce an earlier finding is expensive and wastes scarse resources. The second is a neutral reason. One of the more intriguing perspectives on the reproducibility crisis in data science was a call to see efforts to improve reproducibility as “preventative medicine” [REFS]. The analogy suggests that using reproducible data analysis practices will help researchers better identify any problems that occur in their analysis. The third is a positive reason and what we consider the most important. Reproducibility is important because we want other to build upon our work. Another researcher cannot extend a result if they cannot reproduce the initial result. They also cannot apply my innovative methods if they cannot reproduce the work. They cannot build off of the initial result and methods if they do not have access to the relevant data and code. If scientists want their research to have the maximum impact, it must be reproducible. **This philosophy is central to the proposed Research Education Program: analysts need to ensure reproducibility at every stage of their analysis so that they or anyone else can reproduce and then build upon the work.**

***Developing data analysis skills.*** Preparing a data analysis to ensure that it is reproducible by yourself or others six months from now is not trivial. Software packages, databases, and operating systems change over time and may cause results to change or cause code from the original analysis to break. For researchers analyzing their data using a graphical user interface (GUI; e.g. Microsoft Excel, GraphPad Prism), ensuring reproducibility with such tools requires painstakingly documenting every operation. For those who use a scripting language (e.g. R or Python), the code can become the documentation, but it can also be written in a manner that is impossible to read and reproduce. Of course, accessibility to the raw data, code, and documentation is critical. Unfortunately, in some subfields this level of openness is not widely practiced. It may also be impractical to make protected, proprietary, or large datasets available. Regardless of the challenges, we have found that as traditionally-trained bench scientists are expected to do more of their own analysis using ever growing datasets, are ill-equipped to employ modern approaches to maximize reproducibility.

The past 10 years has seen an explosion in the availability of materials to help people wanting to learn how to analyze data.

Much of these materials have been targeted to autodidacts. This is evident in the myriad books teaching people programming and data analysis skills using languages such as R and Python, the popularity of websites such as Stack Overflow (https://stackoverflow.com) that answer programming questions, the accessibility of online tutorials through for profit companies including DataCamp (https://www.datacamp.com) and Codecademy (https://www.codecademy.com), and in the number of YouTube tutorial videos. There are three significant challenges for the autodidact. The first is learning how to sift through the various opinions to learn the best practices for performing reproducible analyses. The second is how to piece together tools from different ares of data science to perform a complete analysis since most tutorial focus on teaching a single concept rather than on how it fits in with other concepts. The third is that by definition, an autodidact lacks a community in which to develop, correct, and strengthen their new skills.

Simultaneously, the traditional training environment has been converted into a concentrated format giving rise to workshops and bootcamps. Organizations such as The Carpentries (also known as Software Carpentry and Data Carpentry) have popularized the use of workshops to introduce data analysis skills to learners in a concentrated format. These workshops are popular because they are free, require a minimal time commitment from the learner (i.e. 2 to 3 days), employ excellent pedagogy, and offer a welcoming and inclusive environment. Other, for profit, bootcamps and university-based certificate programs, require a more extensive time commitment and often assume a foundation in math, statistics, and programming.

Learning how to become a data scientists requires more than learning to program. Someone wanting to become a data scientist needs to know how to program, but they also need to understand project organization, data visualization, statistics, version control, and automation. This “stack” of skills is overwhelming to most people setting out to learn to to engage in reproducible data science practices.

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

### Produce Code Club sessions that highlight concepts important for performing rigorous and reproducible data science

## 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.