OMB Number: 4040-0010
Expiration Date: 12/31/2022

APPLICATION FOR FEDERAL ASSISTANCE
SF 424 (R&R)

3. DATE RECEIVED BY STATE	**State Application Identifier**

1. TYPE OF SUBMISSION*

❍ Pre-application ● Application ❍ Changed/Corrected Application

4.a. Federal Identifier

b. Agency Routing Number

c. Previous Grants.gov Tracking Number

2. DATE SUBMITTED

Application Identifier 20-PAF08584

5. APPLICANT INFORMATION **Organizational DUNS*:** 073133571

Legal Name*: Regents of the University of Michigan
Department:
Division:
Street1*: 3003 S. State St
Street2:
City*: Ann Arbor
County: Washtenaw
State*: MI: Michigan
Province:
Country*: USA: UNITED STATES
ZIP / Postal Code*: 481091274

Person to be contacted on matters involving this application
Prefix: First Name*: Kellie Middle Name: Last Name*: Buss Suffix:
Position/Title: Project Representative
Street1*: Research & Sponsored Projects, Wolverine Tower
Street2: 3003 S. State St
City*: Ann Arbor
County:
State*: MI: Michigan
Province:
Country*: USA: UNITED STATES
ZIP / Postal Code*: 481091274
Phone Number*: 734-936-1361 Fax Number: Email: klbuss@umich.edu

6. EMPLOYER IDENTIFICATION NUMBER *(EIN) or (TIN)** 38-6006309

7. TYPE OF APPLICANT* H: Public/State Controlled Institution of Higher Education

Other (Specify):
Small Business Organization Type ❍ Women Owned ❍ Socially and Economically Disadvantaged

8. TYPE OF APPLICATION*

● New ❍ Resubmission
❍ Renewal ❍ Continuation ❍ Revision

If Revision, mark appropriate box(es).
❍ A. Increase Award ❍ B. Decrease Award ❍ C. Increase Duration
❍ D. Decrease Duration ❍ E. Other *(specify)* :

Is this application being submitted to other agencies?* ❍Yes ●No What other Agencies?

9. NAME OF FEDERAL AGENCY*
National Institutes of Health

10. CATALOG OF FEDERAL DOMESTIC ASSISTANCE NUMBER
TITLE:

11. DESCRIPTIVE TITLE OF APPLICANT'S PROJECT*
Code Clubs: Repeated practice opportunities to develop reproducible data analysis skills

12. PROPOSED PROJECT
Start Date* Ending Date*
04/01/2021 03/31/2024

13. CONGRESSIONAL DISTRICTS OF APPLICANT
MI-012

SF 424 (R&R) APPLICATION FOR FEDERAL ASSISTANCE

14. PROJECT DIRECTOR/PRINCIPAL INVESTIGATOR CONTACT INFORMATION

Prefix: First Name*: Patrick Middle Name: Last Name*: Schloss Suffix:
Position/Title: Professor
Organization Name*: Regents of the University of Michigan
Department: Microbiology & Immunology
Division: Medical School
Street1*: 1520A MSRB I
Street2:
City*: Ann Arbor
County:
State*: MI: Michigan
Province:
Country*: USA: UNITED STATES
ZIP / Postal Code*: 481095620
Phone Number*: 734-647-5801 Fax Number: Email*: pschloss@umich.edu

15. ESTIMATED PROJECT FUNDING

a. Total Federal Funds Requested*	$248,103.00
b. Total Non-Federal Funds*	$0.00
c. Total Federal & Non-Federal Funds*	$248,103.00
d. Estimated Program Income*	$0.00

16.IS APPLICATION SUBJECT TO REVIEW BY STATE EXECUTIVE ORDER 12372 PROCESS?*

a. YES ○ THIS PREAPPLICATION/APPLICATION WAS MADE AVAILABLE TO THE STATE EXECUTIVE ORDER 12372 PROCESS FOR REVIEW ON:
DATE:

b. NO ● PROGRAM IS NOT COVERED BY E.O. 12372; OR
○ PROGRAM HAS NOT BEEN SELECTED BY STATE FOR REVIEW

17. By signing this application, I certify (1) to the statements contained in the list of certifications* and (2) that the statements herein are true, complete and accurate to the best of my knowledge. I also provide the required assurances * and agree to comply with any resulting terms if I accept an award. I am aware that any false, fictitious, or fraudulent statements or claims may subject me to criminal, civil, or administrative penalties. (U.S. Code, Title 18, Section 1001)

● I agree*

** The list of certifications and assurances, or an Internet site where you may obtain this list, is contained in the announcement or agency specific instructions.*

18. SFLLL or OTHER EXPLANATORY DOCUMENTATION File Name:

19. AUTHORIZED REPRESENTATIVE

Prefix: First Name*: Craig Middle Name: Last Name*: Reynolds Suffix:
Position/Title*: Asst. VP Research and Sponsored Projects
Organization Name*: Regents of the University of Michigan
Department: Research & Sponsored Projects
Division:
Street1*: 3003 S. State St
Street2:
City*: Ann Arbor
County: Washtenaw
State*: MI: Michigan
Province:
Country*: USA: UNITED STATES
ZIP / Postal Code*: 481091274
Phone Number*: 734-936-1361 Fax Number: Email*: msgrants@umich.edu

Signature of Authorized Representative*
Craig.Reynolds

Date Signed*
06/16/2020

20. PRE-APPLICATION File Name:

21. COVER LETTER ATTACHMENT File Name:

424 R&R and PHS-398 Specific
Table Of Contents

OMB Number: 4040-0010
Expiration Date: 12/31/2022

Project/Performance Site Location(s)

Project/Performance Site Primary Location

❍ I am submitting an application as an individual, and not on behalf of a company, state, local or tribal government, academia, or other type of organization.

Organization Name:	Regents of the University of Michigan
Duns Number:	073133571
Street1*:	3003 S. State St
Street2:	
City*:	Ann Arbor
County:	
State*:	MI: Michigan
Province:	
Country*:	USA: UNITED STATES
Zip / Postal Code*:	481091274
Project/Performance Site Congressional District*:	MI-012

Additional Location(s) File Name:

Funding Opportunity Number: RFA-GM-20-001. Received Date: 2020-06-16T12:11:57.000-04:00

OMB Number: 4040-0010
Expiration Date: 12/31/2022

RESEARCH & RELATED Other Project Information

1. Are Human Subjects Involved?* ❍ **Yes** ● **No**

1.a. If YES to Human Subjects

Is the Project Exempt from Federal regulations? ❍ Yes ❍ No

If YES, check appropriate exemption number: _ 1 _ 2 _ 3 _ 4 _ 5 _ 6 _ 7 _ 8

If NO, is the IRB review Pending? ❍ **Yes** ❍ **No**

IRB Approval Date:

Human Subject Assurance Number

2. Are Vertebrate Animals Used?* ❍ **Yes** ● **No**

2.a. If YES to Vertebrate Animals

Is the IACUC review Pending? ❍ Yes ❍ No

IACUC Approval Date:

Animal Welfare Assurance Number

3. Is proprietary/privileged information included in the application?* ❍ **Yes** ● **No**

4.a. Does this project have an actual or potential impact - positive or negative - on the environment?* ❍ **Yes** ● **No**

4.b. If yes, please explain:

4.c. If this project has an actual or potential impact on the environment, has an exemption been authorized or an environmental assessment (EA) or environmental impact statement (EIS) been performed? ❍ Yes ❍ No

4.d. If yes, please explain:

5. Is the research performance site designated, or eligible to be designated, as a historic place?* ❍ **Yes** ● **No**

5.a. If yes, please explain:

6. Does this project involve activities outside the United States or partnership with international collaborators?* ❍ **Yes** ● **No**

6.a. If yes, identify countries:

6.b. Optional Explanation:

	Filename
7. Project Summary/Abstract*	Abstract.pdf
8. Project Narrative*	Narrative.pdf
9. Bibliography & References Cited	References.pdf
10.Facilities & Other Resources	Resources.pdf
11.Equipment	equipment.pdf

Funding Opportunity Number: RFA-GM-20-001. Received Date: 2020-06-16T12:11:57.000-04:00

Project Summary

The development of high throughput data generation tools used across the biomedical sciences has led to a situation where researchers with excellent bench skills struggle to appropriately and reproducibly analyze their data. With the increased size of the datasets, the complexity of the analyses has also grown. Although many institutions provide bioinformatic and statistical consulting services, the reality is that these services are overburdened and ultimately require the researcher who generated the data to also analyze the data. Researchers that once used paper notebooks to record data and spreadsheet-based tools to analyze their data now struggle to use command line tools. The ***long-term goal*** of this work is to enable bench scientists to analyze their biomedical data with robust, rigorous, and reproducible approaches. Traditional training programs have not been able to meet the needs of these researchers. Although very popular and well rated, workshops and bootcamps have proven ineffective at establishing lasting competency. *The lack of repeated reinforcement of the content over time is the most likely explanation for the poor outcomes of these workshops and resources.* Code Clubs have proven critical for providing this repeated reinforcement. Code Clubs are weekly activities that are analogous to a traditional Journal Club, but that focus on developing data analysis skills. The ***overall objective*** of this proposal is to develop a collection of virtual Code Club sessions that researchers can use on their own or with colleagues at their own institutions to improve reproducible data analysis skills. These sessions will cover concepts important for performing rigorous and reproducible data science, will be intentionally designed to develop local communities of practice, and will implement robust pedagogical approaches to teaching. These efforts are aligned with the overall goal of this RFA to create “exportable training modules designed to enhance the rigor, reproducibility, and responsible conduct of biomedical and behavioral data science research.” The ***central hypothesis*** is that completing Code Club sessions will improve the retention of concepts covered in prior workshops and allow learners to more quickly develop their skills expand beyond those covered in a workshop. This hypothesis based on 20 years of experience helping bench scientists learn to do their own data analysis and the excitement of colleagues who have run their own Code Clubs. The ***rationale*** for developing additional Code Club sessions is that by increasing the diversity and number of videos available, researchers will make quicker and deeper gains in their knowledge of reproducible research practices. This project will yield a ***significant vertical step*** in the field because it will put tools into the hands of researchers performing the analyses, empowering them to perform sophisticated and reproducible analyses. The approach taken in the proposed research is ***innovative*** because it represents the first concentrated effort to develop materials that use on repeated engagement of the same content in different contexts to help researchers develop data analysis skills.

Project Narrative

The proposed research is relevant to public health because it supports researchers within the domain of biomedical research who need to develop and strengthen their data science skills. Thus the research is relevant to the part of NIH's mission that pertains to the development, maintenance, and renewal of scientific resources that will assure our ability to perform robust and reproducible research in order to prevent disease.

Facilities and Other Resources

Environment
After serving for three years at the University of Massachusetts in Amherst as an Assistant Professor, he joined the faculty at the University of Michigan in September 2009. Schloss was hired as part of a faculty-led initiative to hire four junior faculty positions as part of a Microbial Ecology and Health cluster hire to the departments of Microbiology & Immunology, Molecular & Cell Biology, Ecology & Evolutionary Biology, and Epidemiology. Schloss and the three other faculty members hired as part of this effort join an already strong cadre of scientists interested in relating microbial ecology to health. He has received extensive support at the University of Michigan to insure his success as an academic researcher. In September 2013 he was promoted with tenure to the rank of Associate Professor and in September 2017 he was promoted to Full Professor. In 2016 he was named the Frederick Novy Collegiate Professor of Microbiome Research. His 12-month tenure track appointment includes a total of 12 calendar months dedicated to research and 20 hours per year of formal classroom teaching. Beyond this, Schloss is the Director of the University of Michigan's chapter of The Carpentries where he works with trainees and staff to develop their skills as instructors and helps deploy more than 10 on-campus workshops per year; he directly helps teach 2 of these each year. In addition, he teaches 6 in person and online workshops per year to international scientists that focus on reproducible data analysis material. These intellectual resources are complemented by the strong program and computational resources provided through his membership as an affiliate faculty member in the Center for Computational Medicine and Bioinformatics and the Michigan Institute for Data Science at the University of Michigan provide collective support for the proposed research to be successful. The facilities and resources available to the PI and his research team at the University of Michigan include everything that they will need to successfully implement the proposed research.

Institutional Commitment to Research Education Program Plan
In the Letters of Support, Dr. Bethan Moore, PhD, the interim chair of the Department of Microbiology & Immunology asserts, "*You continue to have the full support of the Department of Microbiology & Immunology and the School of Medicine in this endeavor including assistance in the provision of adequate staff, facilities, and educational resources that to contribute to your planned research education program*." In addition, the letter from the University of Michigan Center for Research on Learning and Teaching (CRLT), states that this project has the support of CRLT. *These statements of support from members of the University of Michigan and the other letters from colleagues in the are of microbiome research emphatically demonstrate an institutional commitment to the proposed research education program plan.*

Facilities

Laboratory:
Not applicable

Office:
The Schloss lab has 500 sq ft of office space, which includes Dr. Schloss' office as well as a dedicated office for bioinformatics work. This office space consists of telephones, desks, chairs, shelving, and cabinet space. The offices are hardwired for high-speed Internet access and there is access to the Internet through the university's wireless network. Within the Department of Microbiology & Immunology there are a number of conference rooms available for large-group meetings. *These facilities assure that Schloss' research team will have sufficient space and opportunities to collaborate and successfully carry out the proposed research.*

Computer:
Personal computer resources. The bioinformatics component of the Schloss lab consists of Unix, MacPro and iMac computers, which are attached to the university network. The network is maintained by a staff of technicians that are experienced in Windows XP, UNIX, and MacOSX.

HPC resources. Schloss has access to the resources at the Advanced Research Computing (ARC) center, which supports a computing infrastructure that consists of a number of computing clusters and

independent computing and application servers all connected to a high-speed campus network. These resources are described in the Equipment page.

Between the computer resources within Schloss's research program and that of the ARC he will not be limited in their ability to successfully implement the proposed research.

Animal:
Not applicable

Clinical:
Not applicable

Major equipment at the University of Michigan that is relevant to this proposal

Great Lakes HPC cluster. The primary computing resource that we will utilize for this project is the new, Great Lakes HPC cluster. Great Lakes is an HPC Linux-based cluster intended to support parallel and other applications that are not suitable for departmental or individual computers. Each Great Lakes compute node comprises multiple CPU cores with at least 4 GB of RAM per core; Great Lakes has approximately 14,000 cores. All compute nodes are interconnected with InfiniBand networking. The larger memory Great Lakes hardware comprises 3 compute nodes, each configured with 1.5 TB RAM. Great Lakes contains 20 GPU nodes, with a total of 40 NVIDIA Tesla V100 CUDA-capable GPUs. There are also 4 visualization nodes available, each equipped with a single NVIDIA Tesla P40. Computing jobs on Great Lakes are managed through the SLURM Scheduler. The high-speed scratch file system provides 2 petabytes of storage at approximately 80 GB/s performance (compared to 7 GB/s on Flux). All Great Lakes nodes are interconnected with InfiniBand HDR100 networking, capable of 100 Gb/s throughput. In addition to the InfiniBand networking, there is 25 Gb/s ethernet for the login and transfer nodes and a gigabit Ethernet network that connects the remaining nodes. This is used for node management and NFS file system access. Great Lakes is connected to the University of Michigan's campus backbone to provide access to student and researcher desktops as well as other campus computing and storage systems. The campus backbone provides 100 Gbps connectivity to the commodity internet and the research networks Internet2 and MiLR. The Great Lakes cluster includes a comprehensive software suite of commercial and open source research software, including major software compilers, and many of the common research-specific applications. Great Lakes computing services are provided through a collaboration of University of Michigan units: Advanced Research Computing (in the Office of the VP of Research and the Provost's Office), and computing groups in schools and colleges at the university.

OMB Number: 4040-0010
Expiration Date: 12/31/2022

RESEARCH & RELATED Senior/Key Person Profile (Expanded)

PROFILE - Project Director/Principal Investigator

Prefix: First Name*: Patrick Middle Name Last Name*: Schloss Suffix:

Position/Title*: Professor
Organization Name*: Regents of the University of Michigan
Department: Microbiology & Immunology
Division: Medical School
Street1*: 1520A MSRB I
Street2:
City*: Ann Arbor
County:
State*: MI: Michigan
Province:
Country*: USA: UNITED STATES
Zip / Postal Code*: 481095620

Phone Number*: 734-647-5801 Fax Number:

E-Mail*: pschloss@umich.edu

Credential, e.g., agency login: PSCHLOSS

Project Role*: PD/PI Other Project Role Category:

Degree Type: Degree Year:

Attach Biographical Sketch*: File Name: Schloss_biosketch.pdf

Attach Current & Pending Support: File Name:

OMB No. 0925-0001 and 0925-0002 (Rev. 03/2020 Approved Through 02/28/2023)

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Schloss, Patrick

eRA COMMONS USER NAME (credential, e.g., agency login): PSCHLOSS

POSITION TITLE: Professor

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
Cornell University, Ithaca, NY	BS	05/1997	Agricultural & Biological Engineering
Cornell University, Ithaca, NY	PHD	12/2001	Biological & Environmental Eng
University of Wisconsin, Madison, WI	Postdoctoral Fellow	05/2006	Microbial ecology

A. Personal Statement

My research group is broadly interested in beneficial and pathogenic host-microbiome interactions with the goal of improving our understanding of how the microbiome can be used to reach translational outcomes in the prevention, detection, and treatment of colorectal cancer and *Clostridium difficile* infection. To support these efforts, we develop and apply bioinformatic tools to facilitate our analysis. This has made us leaders in the field of host-microbiome research. Leveraging this expertise, we have become highly involved in training others to use the same types of tools that we use to ensure that our research is computationally reproducible. Toward this goal, I direct the University of Michigan's local chapter of The Carpentries for whom I teach several workshops each year. I also teach reproducible research practices as a class at the University of Michigan and in off campus workshops. Through these activities, I have taught more than 1200 individuals. Beyond these traditional forms of teaching, I have also been a leader within the biomedical sciences for developing instructional materials for improving the reproducibility of data science. The rich history of practice and teaching reproducible data science in the environment of an active research laboratory and my past oversight of numerous federally-funded projects makes it an ideal environment to conduct the proposed research. My h-index is 51 and our microbiome-focused research spanning more than 100 peer reviewed publications has been cited more than 29,000 times (Web of Science; accessed 6/4/2020).

B. Positions and Honors

Positions and Employment

1997 - 2002	Graduate Research Assistant, Dept of Biological and Environmental Engineering, Cornell University, Ithaca, NY
2002 - 2006	Associate Researcher, Dept of Plant Pathology, U of Wisconsin, Madison
2006 - 2009	Assistant Professor, Dept of Microbiology, U of Massachusetts, Amherst
2009 - 2016	Associate Faculty, Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor
2009 - 2013	Assistant Professor, Dept of Microbiology & Immunology, U of Michigan, Ann Arbor
2012 - 2013	Assistant Professor, Dept of Civil & Environmental Engineering, U of Michigan, Ann Arbor
2013 - 2017	Associate Professor, Dept of Microbiology & Immunology, U of Michigan, Ann Arbor
2013 - 2015	Associate Professor, Dept of Civil and Environmental Engineering, U of Michigan, Ann Arbor

2014 - 2017 Editor, Applied & Environmental Microbiology, Washington, DC
2017 - Chair of American Society for Microbiology Journals Board, Washington, DC
2017 - Professor, Dept of Microbiology & Immunology, U of Michigan, Ann Arbor

Other Experience and Professional Memberships

2004 - Member, American Society for Microbiology
2005 - Member, International Society for Microbial Ecology
2017- Member, American Association for the Advancement of Science

Honors

2003 Soil Biology Postdoctoral Fellow, United States Department of Agriculture
2003 University of Wisconsin Teaching Fellowship, Howard Hughes Medical Institute
2008 Chancellor's Junior Faculty Fellow, University of Massachusetts
2013 Distinguished Alumnus, University of Wisconsin Department of Bacteriology
2014 League of Research Excellence, University of Michigan Medical School
2016 Frederick Novy Collegiate Professorship in Microbiome Research
2016 Elected to American Academy for Microbiology

C. Contributions to Science

1. A critical aspect of the scientific method is the ability to reproduce the research performed by others so that the field can correct itself as well as build upon previous work and methods to move forward. My lab's efforts in this area have included implementing these materials in our own research, carrying out meta-analyses to validate and synthesize the work of others, and developing instructional materials to disseminate best practices for ensuring that microbiome research is reproducible. Believing that the best way to lead is through our own example, in each of the manuscripts published by the Schloss lab since 2014, our lab has posted the code and literate programming documents for each of our papers to a GitHub repository to insure transparency to better demonstrate the methods behind each of the numbers and figures in our papers. This has led to numerous other research groups being able to build off of our own research. As the Chair of the American Society for Microbiology Journals Board, I am committed to developing protocols to improve the reproducibility of the research reported in our society's journals. To make it easier for others to develop the skills to implement these methods that focus on transparency, automation, version control, and literate programming, we have leveraged funding from an NIH grant to develop instructional materials that others can use to implement the practices used in our lab in their own research. This effort builds upon a tradition in other areas of our laboratory known for producing an open source software package, mothur, which has formalized much microbiome research making it more reproducible. These resources have been posted at https://www.riffomonas.org and their accompanying videos have been posted on YouTube at https://www.youtube.com/channel/UCGuktEl5InrcxPfCjmPWxsA.

    a. **Schloss PD**. The Riffomonas Reproducible Research Tutorial Series. Journal of Open Source Education. 2018. 1(3), 13. https://doi.org/10.21105/jose.00013. [Not indexed in PubMed]
    b. **Schloss PD**. Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. mBio. 2018 Jun 5;9(3). doi: 10.1128/mBio.00525-18. PubMed PMID: 29871915; PubMed Central PMCID: PMC5989067.
    c. **Schloss PD**. Preprinting Microbiology. MBio. 2017 May 23;8(3). e00438-17. PMID: 28536284; PMCID: PMC5442452

2. Sequencing 16S rRNA genes and clustering those sequences into operational taxonomic units (OTUs) is the primary analysis method that underlies most microbiome research projects. When I began developing software to analyzing 16S rRNA gene sequences, researchers either assigned sequences to OTUs manually or they used private scripts. Our tool, DOTUR, automated and standardized the process and made the source code publicly available. DOTUR has gone on to be cited 1,900 times since it was published in 2005 (Web of Science, 1/23/2020). Noticing that a growing number of tools were being published without providing their source code, we resolved to create a fully open source software package

that any researcher could use to perform a broader set of analyses. The result was mothur. In the ten years since it was published, mothur has been cited more than 9,300 times. We are able to keep mothur relevant through regular feature releases and by publishing articles that describe and test new algorithms. The long list of co-authors attests to our mission of serving the research community. While the first three authors wrote the source code, the rest provided documentation and a diverse array of use cases. We are frequently commended for supporting the diverse community of researchers who frequently have limited bioinformatics skills. We are proud of the broad adoption of mothur by users across the microbiome field and around the world. The citations alone are a measure of the significance of this paper. More importantly, mothur has resulted in the standardization of methods and increased the bioinformatics literacy within the field. Sequencing of 16S rRNA genes will continue to be part of microbiome research and mothur will remain a significant part of that effort.

a. **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, Van Horn DJ, Weber CF. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009 Dec;75(23):7537-41. PubMed PMID: 19801464; PubMed Central PMCID: PMC2786419.
b. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, **Schloss PD**. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol. 2013 Sep;79(17):5112-20. PubMed PMID: 23793624; PubMed Central PMCID: PMC3753973.
c. Westcott SL, **Schloss PD**. OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units. mSphere. 2017 Mar-Apr;2(2): e00073-17.. PubMed PMID: 28289728; PubMed Central PMCID: PMC5343174.
d. **Schloss PD**. Reintroducing mothur: 10 Years Later. Appl Environ Microbiol. 2020. Jan 7;86(2):e02343-19. PubMed PMID: 31704678; PubMed Central PMCID: PMC6952234.

3. The standard microbiome analysis will determine whether the communities from healthy and diseased individuals have the same diversity or composition. By analogy, these studies are similar to genome-wide association studies that seek to identify single alleles that can be associated with the disease. Just as geneticists sought out the gene responsible for Huntington's Disease, there are microbiome researchers searching for the "obesity bug". It is far more likely that the microbiome involvement for many diseases is analogous to polygenic traits. Geneticists are also looking for the combination of alleles that result in diabetes and so microbiome researchers need to seek out the consortia within the broader microbiome that is responsible for colon cancer. Another difficulty with the standard microbiome study is that they rarely incorporate clinical data; even if the clinical data are reported it only serves a descriptive role and is not incorporated into the overall analysis. In this study we overcame these limitations to identify the subsets of microbiome found in patients' microbiomes that were associated with *Clostridium difficil*e colonization. *C. difficil*e infections have emerged as the leading nosocomial infection in the US. Through animal models and epidemiological studies, it has been determined that antibiotic perturbations alter the composition of the gut microbiome to allow colonization by *C. difficile*. We sequenced the microbiome of individuals with and without diarrhea and used their microbiome and clinical data to identify collections of bacteria and clinical data that were associated with *C. difficil*e infection. This was a significant result demonstrating that incorporating the microbiome into diagnostic and risk models improve models based on clinical data alone. In subsequent work in my laboratory, we are using multi-omics approaches in animal models of *C. difficile* infection.

a. Schubert AM, Rogers MA, Ring C, Mogle J, Petrosino JP, Young VB, Aronoff DM, **Schloss PD**. Microbiome data distinguish patients with Clostridium difficile infection and non-*C. difficile*-associated diarrhea from healthy controls. MBio. 2014 May 6;5(3):e01021-14. PubMed PMID: 24803517; PubMed Central PMCID: PMC4010826.
b. Schubert AM, Sinani H, **Schloss PD**. 2015. Antibiotic-induced alterations of the murine gut microbiota and subsequent effects on colonization resistance against *Clostridium difficil*e. mBio. 6: e00974-15. PubMed PMID: 26173701; PubMed Central PMCID: PMC4502226.

c. Jenior ML, Leslie JL, Young VB, **Schloss PD**. *Clostridium difficile* Colonizes Alternative Nutrient Niches during Infection across Distinct Murine Gut Microbiomes. mSystems. 2017. 2(4). PubMed PMID: 28761936; PubMed Central PMCID: PMC5527303.

d. Jenior ML, Leslie JL, Young VB, **Schloss PD**. *Clostridium difficile* Alters the Structure and Metabolism of Distinct Cecal Microbiomes during Initial Infection to Promote Sustained Colonization. mSphere. 2018. 3(3). PubMed PMID: 29950381; PubMed Central PMCID: PMC6021602.

4. Whether changes in the microbiome induce tumorigenesis or does the microbiome change as a result of tumorigenesis is the heart of our research into the role of the microbiome in colorectal cancer. Our studies in this area have been significant because they demonstrated an experimental framework for establishing causation in microbiome research and the use of machine learning algorithms to identify biomarkers that are diagnostic of tumors. Through a series of studies using a mouse model of colorectal cancer, we used 16S rRNA gene sequencing to identify changes in the microbiome in a murine model of colon cancer, demonstrated that altering the gut community with antibiotics suppressed tumor formation, and showed that transferring the original tumor-associated microbiome into germ free mice and applied the tumorigenesis-inducing treatment increased the number and size of tumors. Overall, our results point to the microbiome as a necessary component to the process of tumorigenesis. With these results, we then proceeded to apply sequencing and metabolomics techniques to identify biomarkers of disease in human populations. The first wave of microbiome research has been limited to the characterization of the composition of the microbiome under a variety of conditions. Our works is significant because it moves beyond this threshold and delves into designing experiments that manipulate the communities to test predictions about the role of the microbiome and translate those results into results into a potential diagnostic tool for the clinic. Furthermore, it demonstrates an extreme level of rigor that my lab is working at to understand the role of the microbiome in health and disease.

   a. Zackular JP, Baxter NT, Iverson KD, Sadler WD, Petrosino JF, Chen GY, **Schloss PD**. The gut microbiome modulates colon tumorigenesis. MBio. 2013 Nov 5;4(6):e00692-13. PubMed PMID: 24194538; PubMed Central PMCID: PMC3892781.

   b. Zackular JP, Baxter NT, Chen GH, **Schloss PD**. Manipulation of the gut microbiota reveals role in colon tumorigenesis. *mSphere*. 2015 Nov;1(1):e00001-15. PubMed PMID: 27303681; PubMed Central PMCID: PMC4863627.

   c. Baxter NT, Ruffin MT IV, Rogers MAM, **Schloss PD**. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine*. 2016: Apr;8:37. PubMed PMID: 27056827; PubMed Central PMCID: PMC4823848.

   d. Sze MA, Topçuoğlu BD, Lesniak NA, Ruffin MT 4th, **Schloss PD**. Fecal Short-Chain Fatty Acids Are Not Predictive of Colonic Tumor Status and Cannot Be Predicted Based on Bacterial Community Structure. MBio. 2019.10(4). PubMed PMID: 31266879; PubMed Central PMCID: PMC6606814.

5. My research group participated in the first phase of the Human Microbiome Project (HMP) as a member of the Data Analysis Working Group. We developed the data curation pipeline that was used to process the data that was ultimately used in publications from the first phase of the project. This series of papers has symbolic significance indicating my overall service to the community of microbiome researchers. Several dozen papers were published from the HMP in the initial phase that were based on a single time point, including several involving my lab. The HMP has since released data collected from additional time points. The final dataset included sampling 300 individuals at up to 18 body sites on two or three occasions. In Ding & Schloss (2014), we asked whether there were enterotypes, or more broadly, community types, that could be identified at the 18 body sites across the body. Previous analyses were limited to single time points and were unable to quantify the stability of community types through time. To address these questions, we characterized the stability of the community types at each body site, identified associations between the community types found at each body site, and quantified the association between each community type and the subjects' metadata. Overall, we showed that the interpersonal variation of the microbiome sampled from healthy individuals is considerable and that we still do not understand which factors drive differences in the structure of the microbiome. This study was significant because it demonstrated that "healthy" does not represent matching some ideal microbiome composition. Furthermore, it established a framework to connect clinical data with community types that will prove useful in developing diagnostics and assessing risks for developing diseases.

a. A framework for human microbiome research. Nature. 2012 Jun 13;486(7402):215-21. PubMed PMID: 22699610; PubMed Central PMCID: PMC3377744.
b. Structure, function and diversity of the healthy human microbiome. Nature. 2012 Jun 13;486(7402):207-14. PubMed PMID: 22699609; PubMed Central PMCID: PMC3564958.
c. Ding T, **Schloss PD**. Dynamics and associations of microbial community types across the human body. Nature. 2014 May 15;509(7500):357-60. PubMed PMID: 24739969; PubMed Central PMCID: PMC4139711.

Additional publications:
https://www.ncbi.nlm.nih.gov/myncbi/patrick.schloss.1/bibliography/public/

D. Additional Information: Research Support and/or Scholastic Performance

Ongoing Research Support

2018/01/15-2020/12/31
R01 CA215574-01, National Cancer Institute (NCI)
Schloss, Patrick David (contact-PI) & Ruffin IV, Mack (multi-PI)
Identification of Microbiome Based Markers to Improve Colorectal Cancer Detection
Identify and validate microbiome-based biomarkers for a non-invasive diagnostic of colorectal cancer
Role: PI

2016/03/01-2021/02/28
U01 AI124255-01, National Institute of Allergy and Infectious Diseases (NIAID)
Young, Vincent B (contact-PI) & Schloss, Patrick David (multi-PI)
Systems biology of *Clostridium difficile* infection
Model the infection and severity of Clostridium difficile in hospital and long-term care facilities
Role: PI

Completed Research Support

2015/09/01-2017/08/31
R25GM116149-01, National Institute of General Medical Sciences (NIGMS)
Schloss, Patrick David (PI)
Development of reproducible informatics skills among microbiome researchers
Developing instructional materials to improve reproducible informatics skills among microbiome researchers
Role: PI

OMB Number: 4040-0010
Expiration Date: 12/31/2022

RESEARCH & RELATED BUDGET - SECTION A & B, Budget Period 1

ORGANIZATIONAL DUNS*: 073133571

Budget Type*: ● Project ❍ Subaward/Consortium

Enter name of Organization: Regents of the University of Michigan

Start Date*: 04-01-2021 **End Date*:** 03-31-2022 **Budget Period: 1**

A. Senior/Key Person

	Prefix	First Name*	Middle Name	Last Name*	Suffix	Project Role*	Base Salary ($)	Calendar Months	Academic Months	Summer Months	Requested Salary ($)*	Fringe Benefits ($)*	Funds Requested ($)*
1 .		Patrick		Schloss		PD/PI	0.00	2.4	0	0	40,112.00	12,034.00	52,146.00

Total Funds Requested for all Senior Key Persons in the attached file **0.00**

Additional Senior Key Persons: File Name: **Total Senior/Key Person** **52,146.00**

B. Other Personnel

Number of Personnel*	Project Role*	Calendar Months	Academic Months	Summer Months	Requested Salary ($)*	Fringe Benefits*	Funds Requested ($)*
1	Post Doctoral Associates	3	0	0	13,407.00	4,022.00	17,429.00
	Graduate Students						
	Undergraduate Students						
	Secretarial/Clerical						
1	**Total Number Other Personnel**					**Total Other Personnel**	**17,429.00**
						Total Salary, Wages and Fringe Benefits (A+B)	**69,575.00**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

Funding Opportunity Number: RFA-GM-20-001 . Received Date: 2020-06-16T12:11:57.000-04:00

RESEARCH & RELATED BUDGET - SECTION C, D, & E, Budget Period 1

ORGANIZATIONAL DUNS*: 073133571

Budget Type*: ● Project ❍ Subaward/Consortium

Organization: Regents of the University of Michigan

Start Date*: 04-01-2021 **End Date*:** 03-31-2022 **Budget Period: 1**

C. Equipment Description	
List items and dollar amount for each item exceeding $5,000	
Equipment Item	**Funds Requested ($)***
Total funds requested for all equipment listed in the attached file	**0.00**
Total Equipment	**0.00**

Additional Equipment: File Name:

D. Travel	**Funds Requested ($)***
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	3,000.00
2. Foreign Travel Costs	0.00
Total Travel Cost	**3,000.00**

E. Participant/Trainee Support Costs		**Funds Requested ($)***
1. Tuition/Fees/Health Insurance		
2. Stipends		
3. Travel		
4. Subsistence		
5. Other:		
Number of Participants/Trainees	**Total Participant Trainee Support Costs**	

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K, Budget Period 1

ORGANIZATIONAL DUNS*: 073133571

Budget Type*: ● Project ❍ Subaward/Consortium

Organization: Regents of the University of Michigan

Start Date*: 04-01-2021 **End Date*:** 03-31-2022 **Budget Period: 1**

F. Other Direct Costs	Funds Requested ($)*
1. Materials and Supplies	0.00
2. Publication Costs	3,000.00
3. Consultant Services	0.00
4. ADP/Computer Services	0.00
5. Subawards/Consortium/Contractual Costs	0.00
6. Equipment or Facility Rental/User Fees	0.00
7. Alterations and Renovations	0.00
8 . Program evaluation costs	500.00
Total Other Direct Costs	**3,500.00**

G. Direct Costs	Funds Requested ($)*
Total Direct Costs (A thru F)	**76,075.00**

H. Indirect Costs

Indirect Cost Type	Indirect Cost Rate (%)	Indirect Cost Base ($)	Funds Requested ($)*
1 . MTDC	8	76,075.00	6,086.00
		Total Indirect Costs	**6,086.00**

Cognizant Federal Agency (Agency Name, POC Name, and POC Phone Number)	Department of Health and Human Services, Matthew Dito (214) 767-3764

I. Total Direct and Indirect Costs	Funds Requested ($)*
Total Direct and Indirect Institutional Costs (G + H)	**82,161.00**

J. Fee	Funds Requested ($)*
	0.00

K. Total Costs and Fee	Funds Requested ($)*
	82,161.00

L. Budget Justification*	File Name: budget_justification.pdf (Only attach one file.)

RESEARCH & RELATED Budget {F-K} (Funds Requested)

OMB Number: 4040-0010
Expiration Date: 12/31/2022

RESEARCH & RELATED BUDGET - SECTION A & B, Budget Period 2

ORGANIZATIONAL DUNS*: 073133571

Budget Type*: ● Project ❍ Subaward/Consortium

Enter name of Organization: Regents of the University of Michigan

Start Date*: 04-01-2022 **End Date*:** 03-31-2023 **Budget Period: 2**

A. Senior/Key Person

	Prefix	First Name*	Middle Name	Last Name*	Suffix	Project Role*	Base Salary ($)	Calendar Months	Academic Months	Summer Months	Requested Salary ($)*	Fringe Benefits ($)*	Funds Requested ($)*
1 .		Patrick		Schloss		PD/PI	0.00	2.4	0	0	40,112.00	12,034.00	52,146.00

Total Funds Requested for all Senior Key Persons in the attached file **0.00**

Additional Senior Key Persons: File Name: **Total Senior/Key Person** **52,146.00**

B. Other Personnel

Number of Personnel*	Project Role*	Calendar Months	Academic Months	Summer Months	Requested Salary ($)*	Fringe Benefits*	Funds Requested ($)*
1	Post Doctoral Associates	3	0	0	13,407.00	4,022.00	17,429.00
	Graduate Students						
	Undergraduate Students						
	Secretarial/Clerical						
1	**Total Number Other Personnel**					**Total Other Personnel**	**17,429.00**
						Total Salary, Wages and Fringe Benefits (A+B)	**69,575.00**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

Funding Opportunity Number: RFA-GM-20-001 . Received Date: 2020-06-16T12:11:57.000-04:00

RESEARCH & RELATED BUDGET - SECTION C, D, & E, Budget Period 2

ORGANIZATIONAL DUNS*: 073133571

Budget Type*: ● Project ❍ Subaward/Consortium

Organization: Regents of the University of Michigan

Start Date*: 04-01-2022 **End Date*:** 03-31-2023 **Budget Period: 2**

C. Equipment Description	
List items and dollar amount for each item exceeding $5,000	
Equipment Item	**Funds Requested ($)***
Total funds requested for all equipment listed in the attached file	**0.00**
Total Equipment	**0.00**

Additional Equipment: File Name:

D. Travel	**Funds Requested ($)***
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	3,000.00
2. Foreign Travel Costs	0.00
Total Travel Cost	**3,000.00**

E. Participant/Trainee Support Costs		**Funds Requested ($)***
1. Tuition/Fees/Health Insurance		
2. Stipends		
3. Travel		
4. Subsistence		
5. Other:		
Number of Participants/Trainees	**Total Participant Trainee Support Costs**	

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K, Budget Period 2

ORGANIZATIONAL DUNS*: 073133571

Budget Type*: ● Project ❍ Subaward/Consortium

Organization: Regents of the University of Michigan

Start Date*: 04-01-2022 **End Date*:** 03-31-2023 **Budget Period: 2**

F. Other Direct Costs	Funds Requested ($)*
1. Materials and Supplies	0.00
2. Publication Costs	3,000.00
3. Consultant Services	0.00
4. ADP/Computer Services	0.00
5. Subawards/Consortium/Contractual Costs	0.00
6. Equipment or Facility Rental/User Fees	0.00
7. Alterations and Renovations	0.00
8 . Program evaluation costs	500.00
Total Other Direct Costs	**3,500.00**

G. Direct Costs	Funds Requested ($)*
Total Direct Costs (A thru F)	**76,075.00**

H. Indirect Costs

Indirect Cost Type	Indirect Cost Rate (%)	Indirect Cost Base ($)	Funds Requested ($)*
1 . MTDC	8	76,075.00	6,086.00
		Total Indirect Costs	**6,086.00**

Cognizant Federal Agency (Agency Name, POC Name, and POC Phone Number)	Department of Health and Human Services, Matthew Dito (214) 767-3764

I. Total Direct and Indirect Costs	Funds Requested ($)*
Total Direct and Indirect Institutional Costs (G + H)	**82,161.00**

J. Fee	Funds Requested ($)*
	0.00

K. Total Costs and Fee	Funds Requested ($)*
	82,161.00

L. Budget Justification*	File Name: budget_justification.pdf (Only attach one file.)

RESEARCH & RELATED Budget {F-K} (Funds Requested)

OMB Number: 4040-0010
Expiration Date: 12/31/2022

RESEARCH & RELATED BUDGET - SECTION A & B, Budget Period 3

ORGANIZATIONAL DUNS*: 073133571

Budget Type*: ● Project ❍ Subaward/Consortium

Enter name of Organization: Regents of the University of Michigan

Start Date*: 04-01-2023 **End Date*:** 03-31-2024 **Budget Period: 3**

A. Senior/Key Person

	Prefix	First Name*	Middle Name	Last Name*	Suffix	Project Role*	Base Salary ($)	Calendar Months	Academic Months	Summer Months	Requested Salary ($)*	Fringe Benefits ($)*	Funds Requested ($)*
1 .		Patrick		Schloss		PD/PI	0.00	2.4	0	0	40,112.00	12,034.00	52,146.00

Total Funds Requested for all Senior Key Persons in the attached file **0.00**

Additional Senior Key Persons: File Name: **Total Senior/Key Person** **52,146.00**

B. Other Personnel

Number of Personnel*	Project Role*	Calendar Months	Academic Months	Summer Months	Requested Salary ($)*	Fringe Benefits*	Funds Requested ($)*
1	Post Doctoral Associates	3	0	0	13,407.00	4,022.00	17,429.00
	Graduate Students						
	Undergraduate Students						
	Secretarial/Clerical						
1	**Total Number Other Personnel**					**Total Other Personnel**	**17,429.00**
						Total Salary, Wages and Fringe Benefits (A+B)	**69,575.00**

RESEARCH & RELATED Budget {A-B} (Funds Requested)

Funding Opportunity Number: RFA-GM-20-001 . Received Date: 2020-06-16T12:11:57.000-04:00

RESEARCH & RELATED BUDGET - SECTION C, D, & E, Budget Period 3

ORGANIZATIONAL DUNS*: 073133571

Budget Type*: ● Project ❍ Subaward/Consortium

Organization: Regents of the University of Michigan

Start Date*: 04-01-2023 **End Date*:** 03-31-2024 **Budget Period: 3**

C. Equipment Description	
List items and dollar amount for each item exceeding $5,000	
Equipment Item	**Funds Requested ($)***
Total funds requested for all equipment listed in the attached file	**0.00**
Total Equipment	**0.00**

Additional Equipment: File Name:

D. Travel	**Funds Requested ($)***
1. Domestic Travel Costs (Incl. Canada, Mexico, and U.S. Possessions)	3,000.00
2. Foreign Travel Costs	0.00
Total Travel Cost	**3,000.00**

E. Participant/Trainee Support Costs		**Funds Requested ($)***
1. Tuition/Fees/Health Insurance		
2. Stipends		
3. Travel		
4. Subsistence		
5. Other:		
Number of Participants/Trainees	**Total Participant Trainee Support Costs**	

RESEARCH & RELATED Budget {C-E} (Funds Requested)

RESEARCH & RELATED BUDGET - SECTIONS F-K, Budget Period 3

ORGANIZATIONAL DUNS*: 073133571

Budget Type*: ● Project ❍ Subaward/Consortium

Organization: Regents of the University of Michigan

Start Date*: 04-01-2023 **End Date*:** 03-31-2024 **Budget Period: 3**

F. Other Direct Costs	Funds Requested ($)*
1. Materials and Supplies	0.00
2. Publication Costs	3,000.00
3. Consultant Services	0.00
4. ADP/Computer Services	0.00
5. Subawards/Consortium/Contractual Costs	0.00
6. Equipment or Facility Rental/User Fees	0.00
7. Alterations and Renovations	0.00
8 . Program evaluation costs	2,000.00
Total Other Direct Costs	**5,000.00**

G. Direct Costs	Funds Requested ($)*
Total Direct Costs (A thru F)	**77,575.00**

H. Indirect Costs

Indirect Cost Type	Indirect Cost Rate (%)	Indirect Cost Base ($)	Funds Requested ($)*
1 . MTDC	8	77,575.00	6,206.00
		Total Indirect Costs	**6,206.00**

Cognizant Federal Agency (Agency Name, POC Name, and POC Phone Number)	Department of Health and Human Services, Matthew Dito (214) 767-3764

I. Total Direct and Indirect Costs	Funds Requested ($)*
Total Direct and Indirect Institutional Costs (G + H)	**83,781.00**

J. Fee	Funds Requested ($)*
	0.00

K. Total Costs and Fee	Funds Requested ($)*
	83,781.00

L. Budget Justification*	File Name: budget_justification.pdf (Only attach one file.)

RESEARCH & RELATED Budget {F-K} (Funds Requested)

Tracking Number: GRANT13145169 Funding Opportunity Number: RFA-GM-20-001 . Received Date: 2020-06-16T12:11:57.000-04:00

BUDGET JUSTIFICATION

A. Senior/Key Personnel

Patrick Schloss, PhD. – Principal Investigator (2.4 calendar months). Schloss is the Federick G. Novy Collegiate Professor in the Department of Microbiology & Immunology at the University of Michigan. He will direct the design and evaluation of the proposed module and will be responsible for the dissemination of the results from the project. Fringe Benefits are estimated to be 30% of total salary.

B. Other Personnel

To be named, PhD – Postdoctoral Research Associate (3.0 calendar months). This individual will aid in the development of the instructional materials and the dissemination of the materials. Schloss will recruit a postdoctoral research associates for this position who has an interest in teaching as a career path. Historically, all members of the Schloss lab have become trained as instructors for The Carpentries organization and take an active role in promoting reproducible research practices to other scientists. They will also participate in activities hosted by the University of Michigan's Center for Research on Learning and Teaching (CRLT). The researcher will split their effort between this and other projects in the Schloss lab. Fringe Benefits are estimated to be 30% of total salary.

C. Equipment

NA

D. Travel

Support is requested for the PI and Postdoctoral Research Associate to both attend the two annual meetings in Bethesda, MD that are described in the RFA. Funds are also requested for the Postdoctoral Research Associate to travel to give onsite workshops at conferences (e.g. The American Society for Microbiology General Meeting and the meeting of the International Society for Microbial Ecology) to facilitate the development of the instructional materials.

E. Participant/Trainee Support

NA

F. Other direct costs

Publication costs: Funds are requested to facilitate the publication of two publications to disseminate the modules and broadly discuss the need for reproducible research within the microbiome research community.

Program evaluation costs: Funds are requested to help develop, implement, and analyze tools for assessing the efficacy of the instructional materials. Because year 3 is when most of the assessment activities will be performed, the anticipated costs are weighted accordingly.

RESEARCH & RELATED BUDGET - Cumulative Budget

	Totals ($)	
Section A, Senior/Key Person		156,438.00
Section B, Other Personnel		52,287.00
Total Number Other Personnel	3	
Total Salary, Wages and Fringe Benefits (A+B)		208,725.00
Section C, Equipment		0.00
Section D, Travel		9,000.00
1. Domestic	9,000.00	
2. Foreign	0.00	
Section E, Participant/Trainee Support Costs		0.00
1. Tuition/Fees/Health Insurance	0.00	
2. Stipends	0.00	
3. Travel	0.00	
4. Subsistence	0.00	
5. Other	0.00	
6. Number of Participants/Trainees	0	
Section F, Other Direct Costs		12,000.00
1. Materials and Supplies	0.00	
2. Publication Costs	9,000.00	
3. Consultant Services	0.00	
4. ADP/Computer Services	0.00	
5. Subawards/Consortium/Contractual Costs	0.00	
6. Equipment or Facility Rental/User Fees	0.00	
7. Alterations and Renovations	0.00	
8. Other 1	3,000.00	
9. Other 2	0.00	
10. Other 3	0.00	
Section G, Direct Costs (A thru F)		229,725.00
Section H, Indirect Costs		18,378.00
Section I, Total Direct and Indirect Costs (G + H)		248,103.00
Section J, Fee		0.00
Section K, Total Costs and Fee (I + J)		248,103.00

PHS 398 Cover Page Supplement

OMB Number: 0925-0001

Expiration Date: 02/28/2023

1. Vertebrate Animals Section

Are vertebrate animals euthanized? ❍ Yes ❍ No

If "Yes" to euthanasia

Is the method consistent with American Veterinary Medical Association (AVMA) guidelines?

❍ Yes ❍ No

If "No" to AVMA guidelines, describe method and provide scientific justification

2. *Program Income Section

*Is program income anticipated during the periods for which the grant support is requested?

❍ Yes ● No

If you checked "yes" above (indicating that program income is anticipated), then use the format below to reflect the amount and source(s). Otherwise, leave this section blank.

*Budget Period *Anticipated Amount ($) *Source(s)

Tracking Number: GRANT13145169 Funding Opportunity Number: RFA-GM-20-001 . Received Date: 2020-06-16T12:11:57.000-04:00

PHS 398 Cover Page Supplement

3. Human Embryonic Stem Cells Section

*Does the proposed project involve human embryonic stem cells? ❍ Yes ● No

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://grants.nih.gov/stem_cells/registry/current.htm. Or, if a specific stem cell line cannot be referenced at this time, check the box indicating that one from the registry will be used:

Specific stem cell line cannot be referenced at this time. One from the registry will be used.

Cell Line(s) (Example: 0004):

4. Human Fetal Tissue Section

*Does the proposed project involve human fetal tissue obtained from elective abortions? ❍ Yes ● No

If "yes" then provide the HFT Compliance Assurance

If "yes" then provide the HFT Sample IRB Consent Form

5. Inventions and Patents Section (Renewal applications)

*Inventions and Patents: ❍ Yes ❍ No

If the answer is "Yes" then please answer the following:

*Previously Reported: ❍ Yes ❍ No

6. Change of Investigator/Change of Institution Section

❑ Change of Project Director/Principal Investigator

Name of former Project Director/Principal Investigator

Prefix:

*First Name:

Middle Name:

*Last Name:

Suffix:

❑ Change of Grantee Institution

*Name of former institution:

Tracking Number: GRANT13145169 Funding Opportunity Number: RFA-GM-20-001 . Received Date: 2020-06-16T12:11:57.000-04:00

PHS 398 Research Plan

OMB Number: 0925-0001
Expiration Date: 02/28/2023

Introduction	
1. Introduction to Application (for Resubmission and Revision applications)	
Research Plan Section	
2. Specific Aims	Aims.pdf
3. Research Strategy*	Research_Education_Program_Plan.pdf
4. Progress Report Publication List	
Other Research Plan Section	
5. Vertebrate Animals	
6. Select Agent Research	
7. Multiple PD/PI Leadership Plan	
8. Consortium/Contractual Arrangements	
9. Letters of Support	LOS.pdf
10. Resource Sharing Plan(s)	Resource_sharing_plan.pdf
11. Authentication of Key Biological and/or Chemical Resources	
Appendix	
12. Appendix	

Tracking Number: GRANT13145169
Funding Opportunity Number: RFA-GM-20-001. Received Date: 2020-06-16T12:11:57.000-04:00

The expansion of high-throughput laboratory techniques and availability of large public databases has made it clear that the ability to generate data has far outpaced most biomedical scientists' ability to analyze those data. Although many institutions have core facilities that provide statistical and bioinformatic consulting services, these facilities are overrun with clients and are typically a cost sink for their institution. As a solution, researchers attempt to develop their data analysis skills through workshops and online tutorials. The demand and number of learning resources that have become available through organizations like The Carpentries, Data Camp, and CodeAcademy are a testament to their popularity. Unfortunately, empirical analysis of outcomes from workshops has shown that although learners universally love the format and content, they have minimal long term retention of the material. **The lack of repeated reinforcement of the content over time is the most likely explanation for the poor outcomes of these workshops and resources.** Consequently, there is a need to create a library of tutorials that present concepts in different contexts that are relevant that allow learners to assess their retention and confidence with employing the concepts. Surprisingly, most online tutorials focus on how to implement individual concepts to answer abstract questions rather than integrating different concepts to answer interesting questions. If these tutorials are designed with the intent of also building local and more distributed communities around data analysis (i.e. communities of practice), then it will be possible to significantly improve the retention of material covered in data analysis workshops. As a solution to these problems, our research group has developed the concept of a weekly **Code Club**, which is analogous to a traditional Journal Club, but focused around programming. These interactive sessions have successfully helped bench scientists develop skills in data analysis that are strong enough to go on to careers as data scientists at leading universities and pharmaceutical companies.

Our ***long-term*** goal is to enable bench scientists to analyze biomedical data with robust, rigorous, and reproducible approaches. The ***overall objective*** of this proposal is to develop a a collection of virtual Code Club sessions that researchers at any career stage can use on their own or with colleagues. These sessions will cover concepts important for performing rigorous and reproducible data science, will be intentionally designed to develop communities of practice, and use robust pedagogical approaches to teaching. This is aligned with the overall goal of this RFA to create "exportable training modules designed to enhance the rigor, reproducibility, and responsible conduct of biomedical and behavioral data science research." The ***central hypothesis*** is that completing Code Club sessions will improve the retention of concepts covered in prior workshops and allow learners to more quickly develop their skills expand beyond those covered in a workshop. We arrived at this hypothesis based on 20 years of experience helping bench scientists learn to do their own data analysis and the excitement of colleagues who have run their own Code Clubs. Furthermore, during the COVID-19 pandemic we have posted weekly Code Club sessions to help bench scientists developing their R programming skills. The ***rationale*** for developing additional Code Club sessions is that by increasing the diversity and number of videos available, researchers will make quicker and deeper gains in their knowledge of reproducible research practices. We are uniquely poised to achieve the overall objective by executing the following ***Proposed Research Education Program***:

Produce Code Club sessions that highlight concepts important for performing rigorous and reproducible data science. Each session will be composed of a brief web tutorial with exercises for participants to complete along with a video version of the tutorial along with solutions to the exercises. The selection of content and datasets will be motivated by questions relevant to diverse areas across biomedical research that cover a range of concepts related to rigor and reproducibility in data science. To help build a community around the videos, we will solicit questions from viewers. We ***hypothesize*** that researchers who take a workshop and then go on to complete the Code Club sessions will have longer retention of the concepts covered in the workshop than researchers who only take the workshop. We further expect that researchers who complete the Code Club tutorials will more quickly pursue concepts beyond the scope of the workshop. To evaluate the success of this Education Program, we will leverage a network of programming workshops to track retention of content with and without the supplemental tutorials.

Successful completion of this Proposed Research Education Program will significantly enhance the availability and efficacy of materials for improving reproducible data science. In addition to being available through the NIH Clearing House, the tutorials developed through this project will be added to the Riffomonas Project website and YouTube channel. The Riffomonas Project was initiated in 2015 to develop a workshop-style instructional series to help microbiome researchers develop skills that foster reproducibility. The material developed through this proposal will complement earlier microbiome content and help to broaden its audience to include other areas of biomedical research.

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 (1–4). 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 nicely into traditional laboratory training. Laboratory training has not incorporated data analysis skills into existing courses and training programs have been slow to develop or adopt stand alone data science courses (5). Although many programs may require a course in statistics, these typically focus on experimental design and choosing the appropriate statistical tests. They rarely discuss data management, data curation, data visualization, or data dissemination (6, 7). Furthermore, faculty who appreciate that they need to develop these skills themselves 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 neglected. 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 (8). Learners have positive experiences in these workshops and rate the value of the material highly (9–12). Yet, it is necessary to ask whether these intensive training activities are effective. A 2017 analysis measured learning outcomes and found that such activities did not have a statistically significant effect on learning (8). **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 (13, 14). 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. Others called such work "sloppy" (15). 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 (4, 15–18). These led to a renewed effort to improve the rigor and reproducibility across the lifecycle of a research project (1). As highlighted at the NIGMS Clearinghouse for Training Modules to Enhance Data Reproducibility, NIH has supported the development of instructional modules. This includes one that we generated for the field of microbiome research (19). **To improve reproducibility in science, the training must improve.**

Defining reproducibility. There is a general understanding that a result is reproducible if others can obtain the same result as the original researchers. Yet reproducibility and replicability are often used interchangeably or with different definitions (1, 15, 20–23). 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) (24). We can also think of this framework in terms of the data analysis step of a research project (23).

Table 1. System for defining concepts related to reproducibility

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

Briefly, if someone were to take data and methods and generated the same results as the original researchers, 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 biological phenomena. 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 (25–28). 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. (25) 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 teams placed on reproducibility. In spite of their individual reproducibility, the variation in study design reflected the teams' subjectivity and the peer review. The teams produced 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. **Insuring that a result is reproducible is challenging; however, by following best practices we can understand the technical and biological reasons why a result fails to be reproducible, replicable, robust, or generalizable.**

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, a reproducible analysis is not necessarily correct because there may be limitations in the data and methods (22). 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 and resources to help improve the reproducibility of data analyses (6, 7, 9, 11, 19, 29–33). We have found that assuming the analyst will need to reproduce their 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 co-authors' 2010 study "The *Mycobacterium tuberculosis* drugome and its polypharmacological implications" (34). 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 (35). 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, how much longer it would have taken if Bourne and colleagues were not as helpful, or the cost of independent researchers repeating the same process. 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. **Through the development and use of improved tools, data scientists are now in a better position to insure the reproducibility of their data analyses than 10 years ago.**

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 cynical concerns of misconduct. Furthermore, failure to reproduce an earlier finding is expensive and consumes scarce 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" (22). 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 salient. Reproducibility is important because scientists should want others to build upon their work. Another researcher cannot extend a result if they cannot reproduce the initial result. They also cannot apply innovative methods if they cannot reproduce the work. 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 openness and reproducibility at every stage of their analysis so that they or anyone else can reproduce and then build upon the work.**

Necessary data analysis skills. As described above, 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 part of 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. **The challenges of reproducibility are both cultural and technical.**

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 engage in reproducible data science practices. In earlier work, we developed the Riffomonas project to lead microbiome scientists through the development of these skills (19). The proposed project will build off of the earlier Riffomonas platform to generalize the concepts to other biomedical and general science fields. In a Commentary that came from that work, Schloss proposed an aspirational rubric for how researchers could grade the reproducibility of a study (24). The practices outlined in the rubric highlight the skills that a scientist needs to develop to perform reproducible analyses:

- Handling of confounding variables
- Sex/gender as confounding variables
- Experimental design considerations
- Data analysis plan
- Clarity of software descriptions
- Availability of data products
- Availability of metadata
- Data analysis organization
- Availability of data analysis tools
- Documentation of data analysis workflow
- Use of random number generator seed
- Defensive data analysis
- Insuring short and longterm reproducibility
- Open science to foster reproducibility
- Transparency of data analysis

Each of these practices were associated with a grade of “good”, “better”, or “best” depending on how the scientist answered specific questions. For example, under the “Documentation of data analysis workflow” practice:

- **Good: Is our code well documented? Do we use a self-commenting coding practice?** To get this grade, one would need to be able to program and use its commenting system along with using descriptive function and variable names.
- **Better: Do each of our scripts have a header indicating the inputs, outputs, and dependencies? Is it documented how files relate to each other?** This grade requires a more extensive use of commenting and organization.
- **Best: Are automated workflow tools like GNU Make and CommonWL used to convert raw data into final tables, figures, and summary statistics?** A grade of best requires the use of an automation tool and a high level of organization across the project in addition to well documented code.

These questions demonstrate that the ability to answer “yes” is often dependent on using multiple tools. Considering it is an *aspirational* rubric, the expectation is not that every practice be rated “best”. Rather, the goal should always be at least “good” and the scientist should be striving to move to the “better” and “best” grades. **This rubric requires both the basic knowledge of how to use the tools, but also the skill to integrate tools to achieve a goal.**

Teaching data analysis skills. 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, they are ill-equipped to employ modern approaches to maximize reproducibility.

Teaching the self-learner. The past 10 years has seen an explosion in the availability of materials to help people learn to analyze data. Much of these materials have been targeted to self-learners. 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. We have created our own online content as text and videos covering topics related to reproducible research practices (https://www.riffomonas.org). Although these resources are generally excellent, there are multiple challenges for the self-learner (36). The first is deciphering what they need to learn and how sift through the various opinions to learn best practices. The second is how to piece together tools from different areas of data science to perform a complete analysis since most tutorials focus on teaching a single concept rather than on how the concept fits in with other concepts. The third is that by definition, an self-learner lacks a community in which to develop, correct, and strengthen their new skills. These challenges likely limit the progress of bench scientists trying to develop data analysis skills leading them to persist in their skills that limit reproducibility.

Teaching in workshops. Simultaneously, the traditional spread out format of a traditional semester-long course 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), and offer a welcoming and inclusive environment (9, 12). Software and Data Carpentry workshops cover command line tools, programming, and version control. 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. For the past 12 years, we have taught our own 3-day data analysis workshops covering reproducible research topics for scientists studying the host-associated and other microbiomes. An important study from Feldon et al. (8) acknowledged the popularity of workshops, but wanted to know whether the workshop model was effective for long-term retention of the material. Using a cohort of 294 life sciences PhD students they assessed skill development, productivity, and socialization among students 1 and 2 years after joining their graduate program. Among those students who participated in a boot camp or bridge program designed to enhance data analysis and writing skills and acclimate students to academia prior to starting graduate school, there were no significant benefits of the training relative to those that did not participate in a

program. This result was jarring, but aligns with anecdotal evidence of past workshop learners who comment that they are taking a workshop covering the same content for the second or third time. It also aligns with the education literature, as discussed below.

Teaching via live coding online. Data science workshops can be taught in a variety of formats ranging from lectures where code is discussed and shown to live coding demonstrations where the instructor teaches as they interact with a computer and make time for learners to parallel their activities. The latter approach is far more active and produces a better experience and learning outcomes (37, 38). Beyond slowing down the delivery of content, this process of live coding has a few benefits. First, learners can see the instructor make mistakes and watch as the instructor diagnoses problems and works through the solutions. This also normalizes mistakes so that learners realize that even experienced programmers make mistakes. Second, in a live coding environment there are opportunities for learners to ask the instructor questions that pull the instructor off their script. This customizes the training the learner receives. The popularization of online video tutorials available on YouTube, Vimeo, and Twitch offer a virtualized version of the live coding instruction delivered in a workshop. Surprisingly the number of videos posted to these sites on topics related to reproducible research is relatively low. These videos generally fall into two categories. The first category is represented by short tutorials that cover the syntax of a command. Such videos are demonstrations of materials commonly found in reference materials. There is little integration with other tools or concepts. There is no opportunity for viewers to engage and practice the material. The second category is represented by longer videos that are recordings of someone working on a project. These videos are primarily demonstrations and performances rather than instruction. In rare cases, these are live streamed and viewers can comment and ask questions of the presenter. Future viewers can watch a recording of the exchanges. Similar to the short tutorials in the first category, there is no opportunity for the viewer to practice the concepts with new material. **Live coding online is an approach that has yet to be used to effectively teach reproducible data analysis skills.**

There is nothing inherently wrong with the materials that are available to scientists wanting to learn how to analyze data. The challenge is that too much is asked of these approaches. A self-learner struggles to advance because they do not know enough to know what to study - a grounding provided by workshops. Conversely, the workshop-based approach suffers because the learner does not continue to practice the material that was introduced during the workshop. **These approaches likely fail because they do not incorporate the extensive lessons from the cognition literature, which emphasize the value repeated practice over mass learning (39–42).**

Code Club. A common strategy for keeping up with the literature is participating in journal clubs, which involve group discussion of a pre-selected paper. In addition to staying current on the literature, journal clubs help strengthen skills in critical thinking, communication, and integrating the literature (43). Over the past 4 years, our research group has leveraged the similar problems of integrating the overwhelming scientific literature and learning data analysis skills. We have experimented with what we call a ***Code Club*** to improve data analysis skills in a community environment. Code Club sessions generally include a brief tutorial, a set of exercises related to the tutorial for learners to work through in a small group, and an opportunity to debrief and report back each group's solutions. Table 2 includes several examples of successful Code Club topics that we have done within a separate 1-hour long Code Club session.

Table 2. Examples of Code Club sessions facilitated within the Schloss lab's weekly group meetings

Title	Description
base vs. ggplot2	Given input data and a figure, recreate the figure using R's base graphics or ggplot2 syntax
Snakemake	Given a bash script that contains an analysis pipeline, convert it to a Snakemake workflow (can also be done with GNU Make)
DRYing code	Given script with repeated code, create functions to remove repetition
mothur and Vegan	Given a pairwise community dissimilarity matrix, compare communities using the adonis function in the Vegan R package
tidy data	Given a wide-formatted data table, convert it to a long, tidy-formatted data table using tools from R's tidyverse
GitFlow	Learners file and claim an issue to add their name to a README file in a GitHub-hosted repository and file a pull request to complete the issue
R with Google docs	Scrape a Google docs workbook and clean the data to identify previous Code Club presenters

As described in the letters of support from Drs. Lauring, Balunas, and Snitkin, this model has been used by other research groups with great success and enthusiasm. Because of the collaborative nature of the Code Club format, there is significant peer-to-peer instruction and customization of concepts, data sets, and questions that are relevant to the research group. However, often researchers who want to develop their data analysis skills feel isolated in a lab of traditional bench scientists or are in a small research group that lacks the critical mass to implement their own Code Club.

With these factors in mind and the isolation many have felt due to the shutdown of research laboratories due to the COVID-19 pandemic, we have experimented with creating virtual Code Club sessions. Initially, the model included live learners on a Zoom call. The synchronous nature of that format made the model unsustainable. More recently, we have created an asynchronous model where a motivating question, tutorial, and set of exercises and their solutions are provided as a blog post with an accompanying video or "vlog" posted to the Riffomonas project YouTube channel (https://www.youtube.com/riffomonasproject). These Code Club sessions are released once a week and have received a positive response from its growing community. As the sessions gain wider reception, it is hoped that the community will make greater use of the commenting features to ask questions and propose topics for future sessions. **The development of additional Code Club sessions will significantly enhance the opportunity for learners to strengthen their reproducible data analysis skills through repeated and deliberate practice.**

The pedagogical benefits of Code Club over previous methods of instruction. As highlighted above, self-learners and those who participate in workshops struggle because the available resources fail to facilitate learners' ability to engage in the best metacognition practices. The basis of this proposal is that introducing Code Club activities will overcome those limitations. The primary benefit of the Code Club format is that because the sessions incorporate exercises for the learner to practice with, they are engaging in repeated practice, which is a more effective means of learning new material and developing automaticity than massed practice (39–42). By developing a large collection of sessions that learners can chose from, they will have many opportunities to cover similar content helping to make their practice deliberate rather than superficial. The design of the sessions requires that the amount of material be kept to at most 2 concepts. This minimalist design reduces the cognitive load for the learners, which can be significant in workshops (44). In fact, a key part of the design of each session is stripping out extraneous material to minimize the cognitive load for learners. Furthermore, in addition to the limited number of concepts that can be covered in a single session, it is not possible to completely cover those concepts. Therefore, it is unlikely that the learner will attain complete mastery the first time they see the concept. Interleaving concepts over multiple sessions without first attaining complete mastery is an effective way to help learners develop complete mastery because each time they see the concept, they need to practice their retrieval skills to apply what they already know (45, 46). Each time they do this they approach mastery and automaticity. The exercises that are provided with each session increase in difficulty. The first exercise asks the learner to do a simple modification of what was done in the tutorial to answer a related question. By the third question, the learners are asked to apply the concepts to a completely new question with the same or different data. The increasing difficulty of the exercises provides a self assessment to the learner (47). Even if the learner is only able to solve one or two of three exercises on their own, by trying the other exercises they are creating a mental model of how they think the exercise should be solved, which they can adjust once they see the solution (48). **Each Code Club session is intentionally designed to incorporate the cognition literature to enhance the development of reproducible data science skills.**

Teaching portfolio. Over the past 12 years Schloss has taught between 4 and 8 courses per year related to reproducible data analysis skills. Although most of these have been 3 day workshops, Schloss has also taught workshops lasting 2 hours to traditional courses that were a full semester. Schloss's teaching has largely been devoted to general reproducible research practices including R programming and the use of the mothur software package for analyzing microbial ecology data. More than 1,200 scientists have participated in these workshops. In addition, Schloss is a trained Carpentries Instructor and co-teaches one or two Software or Data Carpentry workshops per year. At the University of Michigan, Schloss directs the local chapter of the Carpentries organization where he helps coordinate 10 workshops that are taught each year on campus, pedagogical topics, and the development of additional workshop. Prior to 2020, these workshops were taught in person. Prior to the COVID-19 pandemic Schloss had already transitioned to teaching his workshops via Zoom and was well prepared to start teaching the material virtually. In April 2020, Schloss taught a three day,

R-based workshop to over 100 learners with the help of four teaching assistants. Between this large workshop and two other virtual workshops Schloss has already taught in 2020, he has significantly improved his ability to teach remotely. These experiences demonstrate that Schloss is connected to a large network of scientists who have participated in workshop style learning environments. Schloss will draw upon this network to recruit learners to continue their learning through virtual Code Club sessions. Furthermore, the depth of his experience shows that he has a unique familiarity with the strengths and weaknesses of different teaching approaches. **Schloss's network and deep experience will be significant assets for the proposed Research Eduction Program.**

Teaching philosophy. Schloss believes that anyone can learn to analyze their own data. A data analysis is strongest when the person who designed the experiments and generated the data analyzes with the advice of experts in statistics and their sub-discipline. The best way to motivate learners to learn the concepts they are taught is by answering real world questions rather than using generic questions derived from simulated or overly abstract datasets (e.g. the mtcars or diamonds datasets, which are popular in R teaching materials, e.g. (49)). It would be dishonest to teach one set of methods and use a different set for our own professional work. Therefore, to answer these questions, Schloss teaches the approaches that his research group uses for their research. Schloss teaches learners in his classes as though they were trainees working in his lab that need skills to create the reproducible papers that his lab strives to publish. Just as Philip Bourne demonstrated in the anecdote above, the best data scientists acknowledge that they still have room to grow. With this in mind, Schloss uses an encouraging and growth-minded outlook that asks learners to do better with each new analysis they perform. This incremental approach may feel slow or incomplete. But this approach is far more effective than expecting scientists to take on a large set of skills at once.

Schloss's style of teaching can be seen in the current proposal and in the module he created for work funded under an RFA similar to the current RFA, RFA-GM-15-006. In that module, Schloss developed a series of 14 modules related to reproducible research practices for microbiome research (19). The nearly 14 hours of content was motivated by real world scientific questions, uses live-coding to demonstrate practices, and includes activities for learners to engage in to develop their own skills. The current proposal goes beyond the materials developed for the initial phase of the Riffomonas project and more fully integrates Schloss's teaching philosophy and the pedagogical goals outlined above. These materials are hosted as part of the Riffomonas Project (https://www.riffomonas.org). This name encapsulates how we have seen others and ourselves make the greatest gains in learning reproducible methods. "Riffing" involves taking a musical theme and either repeating it or adapting it to a new setting. The Riffomonas Project, seeks to help people learn concepts by showing how the concept can be employed to answer one question and then encouraging them to adapt the solution to a new question. By starting with solutions that they know work, they can dissect the solution to understand why it works and expand upon it to derive solutions to new problems. **This cycle helps a learner work through the levels of Bloom's Taxonomy as they interact with the concepts (50).**

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 learners 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, they 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 likely 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 our 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. Those activities build off of prior 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. Learners will then have the opportunity to answer related questions using the concepts they just learned. With a high volume of resources, learners will see the same concepts multiple times over many sessions and in different contexts. This will build off of an initial workshop experience 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.

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

Introduction

Most bench scientists struggle to apply modern tools that enable them to insure the reproducibility of their data analyses. The ***overall objective*** of this Research Education Program is to develop a collection of virtual Code Club sessions that researchers can use on their own or with colleagues to strengthen their reproducible data analysis skills. The materials will be targeted to biomedical scientists at any career stage. The ***central hypothesis*** is that completing Code Club sessions will improve the retention of concepts introduced in prior workshops and allow learners to more quickly develop their skills beyond those covered in a workshop. The ***rationale*** for this hypothesis is that although the workshop format is popular, its effectiveness is limited because a workshop asks learners to remember a large number of concepts and only provides superficial opportunities to practice the material. A workshop forces learners to engage in massed practice; an approach that is known to be ineffective (39–42). In contrast, Code Club sessions provide brief, regular opportunities to engage in repeated deliberate practice. By limiting each session to one or two concepts, we will lessen the cognitive load for learners and encourage them to practice their retrieval and application skills (44–46). The framework of the Code Club is based on Schloss's considerable experience helping bench scientists learn to do their own data analyses and observing the experiences of colleagues who have run their own Code Clubs. The ***outcome*** of this Research Education Program will be a validated collection of materials for more than 100 Code Club sessions that cover multiple areas of reproducible data analysis.

Design

Format. Each Code Club session will have the same structure and will be motivated by a question. For example, "what is the half-saturation constant for β-galactosidase?". After a brief introduction stating the question, the host will provide a 20-25 minute tutorial on that concept in which they will answer the question. Depending on the relative balance of concepts that have been covered in other Code Club sessions (see next paragraph), we will select one or two concepts to demonstrate for this question. For example, the host could focus on linear regression for this session. The tutorial would cover the assumptions that must be true to fit data to a linear regression, how to evaluate the quality of the fit, and how to perform the fit in R. To answer the question, the host would discuss the strengths and weaknesses of using a double-reciprocal version of the Michaelis–Menten equation (i.e. the Lineweaver-Burk plot) to determine the relevant constants. Next, viewers will be encouraged to pause the video to complete three exercises. The first question will ask the viewer to adapt their code to a related question (e.g. determine the half-saturation constant for β-glucoronidase). The second question will ask the viewer to adapt their code to answer a related, but more distant question (e.g. fit data using the Hanes-Woolf or Eadie–Hofstee transformations). The third question will have the viewer apply

the concept in a different context (e.g. construct and apply a linear calibration curve for a Bradford assay). To conclude the Code Club session, the host will spend 15 minutes sharing their solutions to the exercises. The goal for each Code Club is not to achieve mastery, but to supplement prior knowledge, provide an opportunity to practice concepts in diverse settings, and to give an opportunity for self assessment. To emphasize this point, a subsequent session might use linear regression to solve a different problem or might build off of this session by discussing logistic regression. Another session might cover the same question, but demonstrate how to fit the Michaelis-Menten equation using non-linear regression. Schloss will be the primary host, but will periodically invite other hosts and co-hosts to diversify the delivery of the message and provide more context for biological questions.

Each Code Club session will consist of a blog post hosted as webpages on the Riffomonas project website (https://www.riffomonas.org/code_club) and as a video hosted on the Riffomonas project YouTube channel (https://www.youtube.org/riffomonasproject). All materials will be released under the Creative Commons By Attribution (CC-BY v4.0) license. The materials will be developed and disseminated keeping in mind the best practices to comply with revised Section 508 Standards. We will regularly review and adopt the best practices described in the standards described in the Guide to Accessible Web Design & Development (https://www.section508.gov/content/guide-accessible-web-design-development) and the W3C, WCAG, and ARIA authoring best practices. The videos will be captioned using the Google speech-to-text algorithm and if the quality is poor, we will use a transcription service and upload our own captions. This is the approach we used previously for the microbiome-based reproducible research tutorial series. We are committed to facilitating the learning of all scientists.

Topic areas. The concepts that will be covered in each Code Club session will be selected from topics that are relevant to insuring reproducible and robust data analyses. The topic areas will include (the estimated number of sessions per topic area are shown in parentheses):

- **Scripting of analyses (n=60 sessions).** Scripting is a critical topic area because it allows the analyst to show the code that was used to transform raw data into the final results. Because this generally involves applying programming skills, this topic area has the most concepts that need to be covered. Sessions will teach learners best practices for using R and bash scripts to clean, process, and validate raw data, visualize data, and statistically analyze and model data. In addition, best practices for building and working with spreadsheets to facilitate scripting of analyses will be covered.
- **Automation (n=10 sessions).** Related to scripting analyses, automation is an important consideration since an automated pipeline details how scripts are integrated with each other to complete a full analysis and how to track the data and code dependencies across a project. Concepts will be demonstrated using bash scripts, GNU Make, and snakemake.
- **Project organization and documentation (n=10 sessions).** A project that is automated using elegant code is worthless if the project is poorly documented or organized. Concepts related to this topic area will detail the value of project structure, self-documenting file and directory names, and providing a guide to the reader to navigate the project.
- **Version control (n=10 sessions).** The ability to see how a data analysis has evolved over its life is possible if good version control practices are used. Concepts related to both creating that historical thread for a project and looking back over the historical thread will be covered. git and GitHub will be used to demonstrate the concepts for this topic area.
- **Literate programming (n=10 sessions).** A literate programming document embeds code within a written narrative and insures that any results in a document are directly linked to the code responsible for producing that result (32, 51). Concepts and applications using R Markdown will be used to cover this topic area.
- **FAIR (Findable, Accessible, Interoperable, and Reusable) principles (n=4 sessions).** The ability to reproduce and build off of an existing analysis is central to the philosophy of the Riffomonas project. To achieve that goal, we will highlight the FAIR principles throughout the series of Code Club sessions. We will also emphasize the value of open science, data and code accessibility, and the creation of containers and machine images for facilitating these principles (https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf)

Although we will limit the number of learning objectives to one or two concepts per session, they will be demonstrated in the context of other concepts to emphasize how the concept of the session can be integrated with other concepts. For example, a session discussing licensing might involve using version control to put a copy of a license in a git repository hosted on GitHub. There are many tools available to demonstrate the concepts in each of these topic areas. We have selected a set of tools that are widely used. As we develop the Code Club sessions we will reassess that the tool we are demonstrating is still preferred. In each case we will also present the strengths and weaknesses of different tools.

Motivating questions. Using real questions and situations faced by scientists is an important component of the design of the Code Club sessions. Some questions will be related to the preparation of reports, presentations, and manuscripts while others will be driven by a biological question. For these latter questions, we will initially rely on our own interests and the suggestions of colleagues to insure that we have a broad range of topics that cover the biomedical sciences (see letters of support). At a minimum, we foresee using questions inspired from genomics, microbiome science, physiology, bacteriology, viral evolution, biochemistry, immunology, epidemiology, psychology, and neurobiology. We will always use data that are freely available and if the question is from a domain outside of our area of expertise (i.e. microbiology), we will consult with an expert to review our materials.

Dissemination Plan

To disseminate the materials generated as part of the proposed research, we will pursue several avenues beyond linking the materials to the NIGMS clearinghouse web site:

Recruitment through workshops. The primary direct approach that we will use to recruit people to participate in Code Clubs is through workshops that Schloss helps facilitate. Each year, Schloss teaches 6 3-day workshops that are attended by researchers from around the world. In addition, at the University of Michigan, he teaches his Microbial Informatics class that is taught as a 3 day workshop and he co-teaches 2 2-day Carpentries workshops. Schloss is the director of the University of Michigan's Carpentries Partner Organization. Including the workshops that Schloss co-teaches, the local Carpentries organization teaches 10 workshops per year. Through these diverse teaching venues, Schloss has the ability to annually draw from a population of more than 120 scientists from outside the University of Michigan and more than 200 scientists from the University of Michigan. We will make every effort to recruit scientists that participate in these activities. Our surveys of past workshop participants indicates that our learners are primarily graduate students (~40%) and postdocs (~40%), but research staff (~15%), faculty (~5%), and occasionally undergraduates (~1%) also participate. An equal number of women and men participate in these workshops.

Advertising of materials. We will use social media (e.g. YouTube and Twitter) to promote the Code Club materials. Schloss currently has over 8,000 followers on Twitter, where he has a reputation for discussing data analysis issues. We will advertise new Code Club sessions through his Twitter account. We will also use search engine optimization (SEO) strategies targeting YouTube's recommendation algorithm to grow a new viewership base. Both avenues will create enthusiasm in the biomedical research community and beyond to foster their interest in the sessions.

Publications and presentations. We anticipate publishing at least two manuscripts for this project. The first will announce the availability of the resources, similar to what we previously did for the Riffomonas project's reproducible research tutorial series (19). The second will report the results of our study into the benefits of engaging in Code Club materials over workshop-based learning alone. Finally, it is likely that there will also be opportunities to give seminar, conference, and webinar presentations describing the project to interested groups.

Availability of materials. All instructional materials will be made freely available through the Riffomonas project website at (https://www.riffomonas.org/code_club) and all videos will be hosted on YouTube under the Riffomonas project channel (https://www.youtube.com/riffomonasproject). All materials related to the project will be maintained as a public GitHub project repository (https://www.github.com/riffomonas/code_club). In fact, the development of this proposal is available at www.github.com/riffomonas/2020_RR_R25. All content will be released under a Creative Commons by Attribution (CC-BY-4.0) license.

Overall, we have a structure in place to disseminate the Code Club materials developed in the proposed plan to a large number of researchers and a plan to expand their reach beyond our current network.

Evaluation Plan

The central hypothesis of the proposed Research Education plan ***is that completing Code Club sessions will improve the retention of concepts covered in prior workshops and allow learners to more quickly develop their skills expand beyond those covered in a workshop.*** To test this hypothesis, we will establish three groups of learners in our study. The first are those that participated in a programming workshop and watched at least one Code Club session. The second are those who only participated in a workshop. The third are those that only participate in Code Club sessions. We will follow up with learners during the week after participating in the workshop to establish a baseline and 2 and 6 months after the workshop. For those in the third group who did not participate in a workshop, we will assess their baseline as soon as they are recruited into the study. We will partner with the Center for Research on Learning and Teaching at the University of Michigan to create survey and assessment instruments (see attached letter from Malinda Matney, PhD). We will finalize the survey and assessment tools in the first year while we are refining the style and building the collection of Code Club materials. The survey and assessment tools will then be deployed in the second and third years of the project (see timeline below).

Surveying participation. Learners will take a survey to record demographic information. It will be important to determine whether factors like gender, race, or career stage impact whether one is more likely to participate in Code Club sessions. We will also survey the learners to measure covariates that we expect to be important to account for retention and growth including the type of workshop they participated in, how many Code Club sessions they participated in, the number of hours they engaged in practice and application of the content, and their career stage.

Evaluating efficacy of Code Club Materials. Learners will complete an evaluation tool that asks the learners to answer a series of questions and perform a series of tasks. These tasks will be brief to minimize the time and effort required by the learner. This evaluation will be a mixture of problem types that ask the learner to:

- Modify a series of steps to achieve a solution to a new question
- Debug a series of steps to achieve a solution with an actual pipeline, using blanked out commands and/or arguments, or rearranging steps to achieve a solution
- Generate the code to convert data to the specified output

Evaluating the materials. Beyond assessing the learners, we need to assess the reception of the learning materials. We will assess the materials with several tools. First, every Code Club session will include an anonymous survey asking learners to evaluate the materials for the clarity of their presentation, relevance, and clarity. The data will be aggregated using Google Forms. Second, we will use the built in analytic tools for YouTube to track the number of views, time spent on each video, viewer demographics, how viewers found the video, likes vs dislikes, and comments. Third, we will use Google's analytics tools to track how learners find each Code Club session's blog post, how long they spend on the site, and where they go after visiting the site. We will track all of these metrics and adjust accordingly throughout the funding period. We anticipate that we will make the most significant changes to the style and strategy in the first year.

This evaluation plan demonstrates that we have a comprehensive plan to evaluate our materials, the effectiveness of the materials, and the types of people engaging in the material.

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 who is anxious to utilize innovative teaching methods to communicate complex materials. Over the past 12 years, Schloss has been the PI on 9 research grants funded by NIH and other agencies including 4 R01 and U01 projects related to the microbiome and an R25 related to developing instructional modules for engaging in reproducible research practices. He has served as a co-Investigator on 16 additional projects during that time. Over the course of his career he has published 103 manuscripts, which covered topics including microbial ecology and the microbiome, science policy and communication, and reproducible practices. The R25 that Schloss was awarded under RFA-GM-15-006, "Development of reproducible informatics skills among microbiome researchers (R25GM116149)"

successfully yielded two peer-reviewed publications (19, 24). The most watched video from that series has been viewed more than 1,300 times. Beyond the funding period of that project, Schloss has continued to develop and post educational content to the Riffomonas project website at https://www.riffomonas.org. 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. Initially designed as a semester-long, 3 credit course, Schloss revamped the course to a 3 full-day workshop to better serve more diverse researchers who could not commit to a semester-long course. Over the 5 years that he has offered the course in this format, the class has grown and he has diversified its content from focusing on microbiome-related data sets to data sets that appeal to a broader audience. Although this course touches on the content of the proposed teaching materials, it has focused on developing R programming skills and not broader 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 12 years Schloss has offered numerous other workshops each year describing how microbiologists can use mothur and R to analyze data from their research projects. This experience has given him a unique perspective into the needs and competencies of the biomedical 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

We have 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 we are utilizing the best practices to evaluate the 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.**

Expected Outcomes

By the end of the proposed project, we expect to find that scientists who engage in a workshop retain and grow their skills better if they also participate in a weekly Code Club session. This expectation derives from the well-validated benefits of repeated practice for improving automaticity. To achieve this outcome, we will create a library of more than 100 recorded Code Club sessions that scientists and the general public can use to improve their reproducible data analysis skills. By consistently releasing a session each week and following our broad dissemination plan, we will build a large subscriber base of more than 1,000 individuals; the Riffomonas project YouTube channel currently has 193 subscribers and has increased the number of subscribers by 56 individuals since Schloss started posting the recent Code Club materials (as of June 15, 2020). Subscribers represent the core group of individuals who will engage with the materials, participate by leaving comments where they ask questions and make suggestions for future concepts they would like to see us cover. **Ultimately, we expect to significantly enhance the reproducible data science skills of a diverse range of scientists at every career level in ever sub-discipline of biomedical research.**

Potential Problems & Alternative Strategies

We are confident that the proposed Research Education Program Plan will be successful; however, there several potential problems that we may face. First, we may find that video is not an attractive medium for scientists to engage with the Code Club sessions and that they prefer the accompanying blog posts. We believe this problem is unlikely because of the popularity of live coding. However, if this is the case, we would pivot to text-based content and cease producing the video. Second, we may find that the number of people engaging the Code Club materials grows at a slower than expected pace. We are optimistic that this will be unlikely because there is a large population of scientists interested in learning data science skills. Furthermore, we will be consistently producing content, which has been shown to be key to building viewership. If this does become a problem, we will conduct focus groups and interviews to assess what is causing slow adoption. We

will also ask scientists with a strong social media presence to help evaluate the materials and promote our content. Finally, it is possible that we will be unable to keep up with the pace of producing one video per week. Considering each Code Club session takes about 8 hours for one person to develop, record, edit, and release, this is unlikely since Schloss is devoting 20% of his effort (i.e. 8 hours per week) and will be working with a postdoctoral research fellow (25% effort) to develop the material. This should be a sufficient amount of effort to cover the proposed pacing, especially since we anticipate finding efficiencies as we go forward. One solution might be to produce multiple weeks' worth of Code Club sessions simultaneously and deploy them once per week. Although less desirable, we could also shift to releasing the sessions every two weeks. **We have developed a robust frameworks for developing, disseminating, and evaluating the Code Club sessions, which will yield a successful outcome.**

Timeline and Benchmarks for Success

The proposed Research Education Program Plan will consist of developing 52 Code Club sessions per year that are released weekly to the Riffomonas Project website and YouTube channel (https://www.youtube.com/riffomonasproject) for the first two years of the program. If we continue to develop videos beyond this time, it will be supported by independent sources of funding. In the first year we will also develop and refine our survey and evaluation tools. In the second and third years we will recruit learners to the study from our workshop attendees, social media network, and those who find the materials through search engines. In the third year we will continue to recruit learners and report the results of our surveys and evaluations.

Task	2021			2022				2023				2024
	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1
Deploy sessions												
Maintenance and refinement												
Develop assessment tools												
Recruit learners												
Evaluate learners												
Summarize and report results												

Future directions

Similar to how Journal Clubs are used to improve research skills and develop new research directions, Code Clubs have the potential to improve reproducible data analysis skills and inspire scientists to push their sub-disciplines further. If the proposed Research Education Program Plan is successful, then we will be likely to have inspired "copy cats" to produce a similar type of content within the data science arena. Hopefully, we can build enthusiasm around the Code Club concept and build a community of people helping each other to learn to engage in more reproducible practices. In the future, the need for us to produce all of the content ourselves would lessen and we would receive Code Club session proposals from the community who could produce the sessions under the Riffomonas Project. We expect that we will be able to adapt the Code Club concept to other components of research including writing and bench skills to complement programs that align with the goals of other educational programs relevant to NIGMS's training portfolio.

PHS Human Subjects and Clinical Trials Information

OMB Number: 0925-0001

Expiration Date: 02/28/2023

Use of Human Specimens and/or Data

Does any of the proposed research in the application involve human specimens and/or data *	❍ Yes ● No
Provide an explanation for any use of human specimens and/or data not considered to be human subjects research.	
Are Human Subjects Involved	❍ Yes ● No
Is the Project Exempt from Federal regulations?	❍ Yes ❍ No
Exemption Number	❑ 1 ❑ 2 ❑ 3 ❑ 4 ❑ 5 ❑ 6 ❑ 7 ❑ 8
Other Requested Information	

References

1. **Collins FS**, **Tabak LA**. 2014. NIH plans to enhance reproducibility. Nature **505**:612–613. doi:10.1038/505612a.

2. **Huang Y**, **Liu Y**, **Zheng C**, **Shen C**. 2017. Investigation of cross-contamination and misidentification of 278 widely used tumor cell lines. PLOS ONE **12**:e0170384. doi:10.1371/journal.pone.0170384.

3. **Horbach SPJM**, **Halffman W**. 2017. The ghosts of HeLa: How cell line misidentification contaminates the scientific literature. PLOS ONE **12**:e0186281. doi:10.1371/journal.pone.0186281.

4. **Head ML**, **Holman L**, **Lanfear R**, **Kahn AT**, **Jennions MD**. 2015. The extent and consequences of p-hacking in science. PLOS Biology **13**:e1002106. doi:10.1371/journal.pbio.1002106.

5. **Barone L**, **Williams J**, **Micklos D**. 2017. Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. PLOS Computational Biology **13**:e1005755. doi:10.1371/journal.pcbi.1005755.

6. **Wilson G**, **Aruliah DA**, **Brown CT**, **Hong NPC**, **Davis M**, **Guy RT**, **Haddock SHD**, **Huff KD**, **Mitchell IM**, **Plumbley MD**, **Waugh B**, **White EP**, **Wilson P**. 2014. Best practices for scientific computing. PLoS Biology **12**:e1001745. doi:10.1371/journal.pbio.1001745.

7. **Wilson G**, **Bryan J**, **Cranston K**, **Kitzes J**, **Nederbragt L**, **Teal TK**. 2017. Good enough practices in scientific computing. PLOS Computational Biology **13**:e1005510. doi:10.1371/journal.pcbi.1005510.

8. **Feldon DF**, **Jeong S**, **Peugh J**, **Roksa J**, **Maahs-Fladung C**, **Shenoy A**, **Oliva M**. 2017. Null effects of boot camps and short-format training for PhD students in life sciences. Proceedings of the National Academy of Sciences **114**:9854–9858. doi:10.1073/pnas.1705783114.

9. **Stefan MI**, **Gutlerner JL**, **Born RT**, **Springer M**. 2015. The quantitative methods boot camp: Teaching quantitative thinking and computing skills to graduate students in the life sciences. PLOS Computational Biology **11**:e1004208. doi:10.1371/journal.pcbi.1004208.

10. **Bentley AM**, **Artavanis-Tsakonas S**, **Stanford JS**. 2008. Nanocourses: A short course format as an educational tool in a biological sciences graduate curriculum. CBELife Sciences Education **7**:175–183. doi:10.1187/cbe.07-07-0049.

11. **Wilson G**. 2016. Software carpentry: Lessons learned. F1000Research. doi:10.12688/f1000research.3-62.v2.

12. **Huppenkothen D**, **Arendt A**, **Hogg DW**, **Ram K**, **VanderPlas JT**, **Rokem A**. 2018. Hack weeks as a model for data science education and collaboration. Proceedings of the National Academy of Sciences **115**:8872–8877. doi:10.1073/pnas.1717196115.

13. **Begley CG**, **Ellis LM**. 2012. Raise standards for preclinical cancer research. Nature **483**:531–533. doi:10.1038/483531a.

14. **Prinz F**, **Schlange T**, **Asadullah K**. 2011. Believe it or not: How much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery **10**:712–712. doi:10.1038/nrd3439-c1.

15. **Casadevall A**, **Ellis LM**, **Davies EW**, **McFall-Ngai M**, **Fang FC**. 2016. A framework for improving the quality of research in the biological sciences. mBio **7**:e01256–16. doi:10.1128/mbio.01256-16.

16. **Ioannidis JPA**. 2005. Why most published research findings are false. PLOS Medicine **2**:e124. doi:10.1371/journal.pmed.0020124.

17. **Langille MGI**, **Ravel J**, **Fricke WF**. 2018. “Available upon request”: Not good enough for microbiome data! Microbiome **6**. doi:10.1186/s40168-017-0394-z.

18. **Stodden V**, **Seiler J**, **Ma Z**. 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. Proceedings of the National Academy of Sciences **115**:2584–2589. doi:10.1073/pnas.1708290115.

19. **Schloss PD**. 2018. The riffomonas reproducible research tutorial series. Journal of Open Source Education **1**:13. doi:10.21105/jose.00013.

20. **Casadevall A**, **Fang FC**. 2010. Reproducible science. Infection and Immunity **78**:4972–4975. doi:10.1128/iai.00908-10.

21. **Goodman SN**, **Fanelli D**, **Ioannidis JPA**. 2016. What does research reproducibility mean? Science Translational Medicine **8**:341ps12–341ps12. doi:10.1126/scitranslmed.aaf5027.

22. **Leek JT**, **Peng RD**. 2015. Reproducible research can still be wrong: Adopting a prevention approach. Proceedings of the National Academy of Sciences **112**:1645–1646. doi:10.1073/pnas.1421412111.

23. **Whitaker K**. 2017. Publishing a reproducible paper. doi:10.6084/m9.figshare.5440621.v2.

24. **Schloss PD**. 2018. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. mBio **9**. doi:10.1128/mbio.00525-18. PMCID: PMC5989067

25. **Silberzahn R**, **Uhlmann EL**, **Martin DP**, **Anselmi P**, **Aust F**, **Awtrey E**, **Bahník**, **Bai F**, **Bannard C**, **Bonnier E**, **Carlsson R**, **Cheung F**, **Christensen G**, **Clay R**, **Craig MA**, **Rosa AD**, **Dam L**, **Evans MH**, **Cervantes IF**, **Fong N**, **Gamez-Djokic M**, **Glenz A**, **Gordon-McKeon S**, **Heaton TJ**, **Hederos K**, **Heene M**, **Mohr AJH**, **Högden F**, **Hui K**, **Johannesson M**, **Kalodimos J**, **Kaszubowski E**, **Kennedy DM**, **Lei R**, **Lindsay TA**, **Liverani S**, **Madan CR**, **Molden D**, **Molleman E**, **Morey RD**, **Mulder LB**, **Nijstad BR**, **Pope NG**, **Pope B**, **Prenoveau JM**, **Rink F**, **Robusto E**, **Roderique H**, **Sandberg A**, **Schlüter E**, **Schönbrodt FD**, **Sherman MF**, **Sommer SA**, **Sotak K**, **Spain S**, **Spörlein C**, **Stafford T**, **Stefanutti L**, **Tauber S**, **Ullrich J**, **Vianello M**, **Wagenmakers E-J**, **Witkowiak M**, **Yoon S**, **Nosek BA**. 2018. Many analysts, one data set: Making transparent how variations in analytic choices affect results. Advances in Methods and Practices in Psychological Science **1**:337–356. doi:10.1177/2515245917747646.

26. **Patil P**, **Peng RD**, **Leek JT**. 2016. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. Perspectives on Psychological Science **11**:539–544. doi:10.1177/1745691616646366.

27. **Etz A**, **Vandekerckhove J**. 2016. A Bayesian perspective on the reproducibility project: Psychology. PLOS ONE **11**:e0149794. doi:10.1371/journal.pone.0149794.

28. **Errington TM**, **Iorns E**, **Gunn W**, **Tan FE**, **Lomax J**, **Nosek BA**. 2014. An open investigation of the reproducibility of cancer biology research. eLife **3**. doi:10.7554/elife.04333.

29. **Noble WS**. 2009. A quick guide to organizing computational biology projects. PLOS Computational Biology **5**:e1000424. doi:10.1371/journal.pcbi.1000424.

30. **Sandve GK**, **Nekrutenko A**, **Taylor J**, **Hovig E**. 2013. Ten simple rules for reproducible computational research. PLOS Computational Biology **9**:e1003285. doi:10.1371/journal.pcbi.1003285.

31. **Taschuk M**, **Wilson G**. 2017. Ten simple rules for making research software more robust. PLOS Computational Biology **13**:e1005412. doi:10.1371/journal.pcbi.1005412.

32. **Xie Y**. 2015. Dynamic documents with R and knitr, 2nd ed. Chapman; Hall/CRC, Boca Raton, Florida.

33. **Perez F**, **Granger BE**. 2007. IPython: A system for interactive scientific computing. Computing in Science & Engineering **9**:21–29. doi:10.1109/mcse.2007.53.

34. **Kinnings SL**, **Xie L**, **Fung KH**, **Jackson RM**, **Xie L**, **Bourne PE**. 2010. The mycobacterium tuberculosis drugome and its polypharmacological implications. PLoS Computational Biology **6**:e1000976. doi:10.1371/journal.pcbi.1000976.

35. **Garijo D**, **Kinnings S**, **Xie L**, **Xie L**, **Zhang Y**, **Bourne PE**, **Gil Y**. 2013. Quantifying reproducibility in computational biology: The case of the tuberculosis drugome. PLOS ONE **8**:e80278. doi:10.1371/journal.pone.0080278.

36. **Kirschner PA**, **Merriënboer JJG van**. 2013. Do learners really know best? Urban legends in education. Educational Psychologist **48**:169–183. doi:10.1080/00461520.2013.804395.

37. **Rubin MJ**. 2013. The effectiveness of live-coding to teach introductory programming. *In* Proceeding of the 44th ACM technical symposium on computer science education - SIGCSE 13. ACM Press.

38. **Haaranen L**. 2017. Programming as a performance. *In* Proceedings of the 2017 ACM conference on innovation and technology in computer science education. ACM.

39. **Carpenter SK**, **Cepeda NJ**, **Rohrer D**, **Kang SHK**, **Pashler H**. 2012. Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. Educational Psychology Review **24**:369–378. doi:10.1007/s10648-012-9205-z.

40. **Rohrer D**. 2015. Student instruction should be distributed over long time periods. Educational Psychology Review **27**:635–643. doi:10.1007/s10648-015-9332-4.

41. **Budé L**, **Imbos T**, **Wiel MW van de**, **Berger MP**. 2010. The effect of distributed practice on students' conceptual understanding of statistics. Higher Education **62**:69–79. doi:10.1007/s10734-010-9366-y.

42. **Karpicke JD**, **Blunt JR**. 2011. Retrieval practice produces more learning than elaborative studying with concept mapping. Science **331**:772–775. doi:10.1126/science.1199327.

43. **Lonsdale A**, **Penington JS**, **Rice T**, **Walker M**, **Dashnow H**. 2016. Ten simple rules for a bioinformatics journal club. PLOS Computational Biology **12**:e1004526. doi:10.1371/journal.pcbi.1004526.

44. **Sweller J**. 1988. Cognitive load during problem solving: Effects on learning. Cognitive Science **12**:257–285. doi:10.1207/s15516709cog1202_4.

45. **Rohrer D**, **Dedrick RF**, **Stershic S**. 2015. Interleaved practice improves mathematics learning. Journal of Educational Psychology **107**:900–908. doi:10.1037/edu0000001.

46. **Karpicke JD**, **Roediger HL**. 2008. The critical importance of retrieval for learning. Science **319**:966–968. doi:10.1126/science.1152408.

47. **Andrade H**, **Valtcheva A**. 2009. Promoting learning and achievement through self-assessment. Theory Into Practice **48**:12–19. doi:10.1080/00405840802577544.

48. **Kornell N**, **Hays MJ**, **Bjork RA**. 2009. Unsuccessful retrieval attempts enhance subsequent learning. Journal of Experimental Psychology: Learning, Memory, and Cognition **35**:989–998. doi:10.1037/a0015729.

49. **Wickham H**, **Grolemund G**. 2016. R for Data Science. O'Reilly Media.

50. **Anderson LW**, **Krathwohl DR**, **Bloom BS**. 2001. A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives. Longman.

51. **Knuth DE**. 1984. Literate programming. Comput J **27**:97–111. doi:10.1093/comjnl/27.2.97.

Letters of Support

The following list indicates the authors of the letters of support that follow this page, their affiliation and the content of their letter.

- **Bethany Moore, PhD:** Interim Chair of Department of Microbiology & Immunology at the University of Michigan; Statement of University Support
- **Malinda Matney, PhD:** Instructional consultant within the Center for Research on Learning & Teaching at the University of Michigan; Support of educational assessment and instructional material development
- **Evan Snitkin, PhD:** Assistant Professor in the Department of Microbiology & Immunology; co-instructor with Schloss (PI) in MICRBIOL 612: Microbial Informatics and adopter of in person Code Club model with his research group
- **Adam Lauring, MD/PhD:** Associate Professor in the Department of Microbiology & Immunology and Internal Medicine; adopter of in person Code Club model with his research group
- **Marcy Balunas, PhD:** Associate Professor in the Department of Department of Pharmaceutical Sciences at the University of Connecticut; Schloss Lab Code Club participant during her sabbatical at the University of Michigan and adopter of in person Code Club model with her research group

University of Michigan
Medical School

Bethany Moore, Ph.D.
Professor & Interim Chair, Department of Microbiology & Immunology
Galen B. Toews, M.D. Collegiate Professor of Internal Medicine
University of Michigan Medical School
5641 Medical Science Building II
1150 W. Medical Center Drive
Ann Arbor, MI 48109-5620
Phone: (734) 764-1466
bmoore@umich.edu

June 4, 2020

Dear Pat,

I am writing in support of your proposal, “Code Clubs: Repeated practice opportunities to develop reproducible data analysis skills”, which you are submitting to the “Training Modules to Enhance the Rigor, Reproducibility and Responsible Conduct of Biomedical Data Science Research (R25)” program at NIGMS. ***You continue to have the full support of the Department of Microbiology & Immunology and the School of Medicine in this endeavor including assistance in the provision of adequate staff, facilities, and educational resources to contribute to your planned research education program***. Since you joined the department in 2009, you have successfully leveraged extramural support to continuously fund your research program, train exceptional graduate students and postdocs, and continue your excellent research career.

You have also made a significant impact on the teaching mission within the department by developing popular courses in *Symbiosis* and *Microbial Informatics*. I was excited to hear that you have been using *Microbial Informatics* as a seedbed to develop the ideas you are proposing in this proposal. The enrollment of your courses steadily grow and frequently attract the attention of senior graduate students and postdocs who take the class not because they need it for graduation, but because they need it for their research. This is a high compliment that busy researchers stop to take your class because they find it useful and well taught. In fact, one of my pulmonary clinical fellows took your course when he was just learning to analyze microbiome data and he is now a K99/R00 funded researcher on the tenure-track here at Michigan. Within these courses you do a great job of attracting a diverse array of students from various departments and strong representation of women. This was highlighted by you opening your recent offering of *Microbial Informatics* to bench scientists working from home because of the COVID-19 pandemic. I was impressed that you had more than 100 people participating across the 3 full day workshop. I am confident that you will continue to leverage these types of experiences as you prepare, evaluate, and disseminate the proposed materials to biomedical researchers at other institutions.

My research group is increasingly using high throughput tools to study the role of the immune system in lung fibrosis and stem cell transplant. The challenges you describe in your proposal of helping bench scientists to develop data analysis skills that are robust and reproducible is significant. As your project goes forward, we would be happy to share questions and example datasets with you that you could use to motivate your teaching modules. For instance, we have microbiome datasets from the lung and gut in two different disease models in human and mice, several RNAseq datasets and a large plasma proteomics dataset with longitudinal collections over 12 months. I would be pleased to help provide relevant research questions for the students to practice using these datasets.

I am excited about this proposal and wish you the best of success in this new project!

Sincerely,

Bethany Moore

Interim Chair, Department of Microbiology and Immunology
Galen B. Toews, MD Professor of Internal Medicine

CRLT CENTER FOR RESEARCH ON LEARNING & TEACHING
UNIVERSITY OF MICHIGAN

1071 Palmer Commons
100 Washtenaw Ave.
Ann Arbor, MI 48109-2218
Phone: (734) 764-0505 • Fax: (734) 647-3600

11 June 2020

Dear Pat:

I was excited to hear about your R25 proposal, “Code Clubs: Repeated practice opportunities to develop reproducible data analysis skills” that you are submitting in response to the recently announced “Training Modules to Enhance the Rigor, Reproducibility and Responsible Conduct of Biomedical Data Science Research” RFA at NIGMS. It is clear that you are passionate about the proposed instructional materials and have innovative ideas for implementing the materials and assessing the modules themselves as well as the researchers that take the modules.

The mission of the Center for Research on Learning and Teaching (CRLT) at the University of Michigan is “dedicated to the support and advancement of evidence-based learning and teaching practices and the professional development of all members of the campus teaching community. CRLT partners with faculty, graduate students, postdocs, and administrators to develop and sustain a University culture that values and rewards teaching, respects and supports individual differences among learners, and creates learning environments in which diverse students and instructors can excel.” As you proceed with this project, other CRLT staff and myself would be happy to provide you with fee-based services such as performing focus groups and surveys to assess the materials you develop. We are also able to provide complimentary consultations for developing IRB protocols, should those be needed.

It is exciting to see your project taking shape and we are anxious to see how these modules develop. Please be in touch as you move forward.

Sincerely,

Malinda M. Matney, Ph.D.
Managing Director

University of Michigan
Medical School

Evan S. Snitkin, Ph.D.
Assistant Professor
DEPARTMENT OF MICROBIOLOGY & IMMUNOLOGY
University of Michigan Medical School
1510E Medical Science Research Building I
Ann Arbor, MI 48109-5620
Phone: (734) 763-3531
Fax (734) 764-3562
esnitkin@umich.edu

June 10th, 2020

Dear Pat,

I was excited to hear that you are submitting your proposal, “Code Clubs: Repeated practice opportunities to develop reproducible data analysis skills”, to further develop your concept of Code Clubs. As co-instructors of MICRBIOL 612: Microbial Informatics, we have often discussed the challenges of teaching data science concepts in a compressed workshop-style framework. I know that you have thinking deeply about how to help students to continue to develop their skills when our class is over.

I share your concerns and like you, I struggle to recruit and train scientists to engage in computational research. I first became aware of your lab’s Code Clubs when one of my students told me about her participation in yours. She thought they were worthwhile and has encouraged me to incorporate them into our group’s regular lab meetings. I am confident that continuing to develop materials for running Code Clubs will have a significant impact on the development of people’s data science skills.

As you know, my research group is interested in bacterial genomics and using genome sequences to track outbreaks of antibiotic resistant bacteria. I would be happy to share research questions and existing data that you could use to motivate topics for your Code Club sessions. In addition, I will provide iterative feedback as I employ these Code Clubs in my research group.

Sincerely,

Evan Snitkin, PhD
Assistant Professor
Department of Microbiology and Immunology
Department of Medicine, Division of Infectious Diseases
University of Michigan Medical School

The University of Michigan
Medical School
Ann Arbor, Michigan 48109-5640

ADAM LAURING, M.D., Ph.D
Department of Internal Medicine
Department of Microbiology & Immunology
5510B MSRB I
1150 W. Medical Center Dr.
PHONE: (734) 764-7731
FAX: (734) 764-0101
E-mail: alauring@umich.edu

June 8, 2020

RE: Code Club R25 application

Dear Pat,

I am happy to write in support of your R25 application to develop Code Clubs as a widely distributed teaching tool for scientists. Programming has become an essential skill for biomedical scientists, and "good coding practices" are fundamental to ensuring rigor and reproducibility in research.

During the time that we have both been at the University of Michigan, I have always enjoyed our discussions around implementing data science practices and developing approaches to help members of our labs to develop their skills. Your innovative teaching in microbial bioinformatics (e.g. your local Software Carpentry workshops and your online Riffomonas tutorials) has been a tremendous help to my trainees and provided them with the skills necessary to do their work.

I think the concept of a Code Club, which your group developed, is innovative and has proven itself an effective training tool. As you outline in your proposal, students often have difficulty retaining knowledge or maintaining their coding skills after completing a course or workshop. I have found this to be a frequent problem in my own work, and Code Club effectively addresses this issue. In fact, borrowing your idea of a Code Club, my group has used your original reproducible research modules to seed discussions and tutorials at our lab meetings. It is exciting that your proposal will develop materials to help individuals and labs to strengthen their data analysis skills through the Code Club format. These materials will have a significant impact on training the next generation of scientists.

As you know, my research group studies the biology, evolution, and epidemiology of influenza and polio. We would be thrilled to help you find interesting questions and datasets to as you develop the materials for your Code Club sessions. Good luck on your exciting proposal!

Sincerely,

Adam Lauring, M.D., Ph.D.
Associate Professor

UCONN
UNIVERSITY OF CONNECTICUT

School of Pharmacy
Department of Pharmaceutical Sciences
Marcy J. Balunas, Ph.D.
Associate Professor of Medicinal Chemistry

June 6, 2020

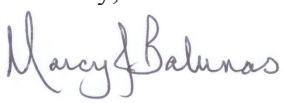
Patrick Schloss, Ph.D.
Frederick G. Novy Collegiate Professor of Microbiome Research
Department of Microbiology and Immunology
University of Michigan
1520A Medical Science Research Buildng I
1150 W. Medical Center Drive
Ann Arbor, MI 48109

Dear Pat,

I am writing in enthusiastic support for your National Institutes of Health R25 proposal, entitled "Code Clubs: Repeated practice opportunities to develop reproducible data analysis skills". In 2018, I had the opportunity to spend my sabbatical working in your research group. As a natural products chemist, it was eye opening to see how you interacted with your group and helped them develop the tools they needed to analyze large microbiome datasets. I face a similar struggle with training students to analyze large metabolomics datasets. While at the University of Michigan, I took your data analysis with R class and participated in your group's Code Club sessions. I was impressed that trainees with very different interests and level of comfort with R could improve their skills in such a safe, friendly, and nurturing environment. When I returned to the University of Connecticut, I started doing similar Code Club activities with my research group motivated by questions related to what we are trying to do with mass spectrometry data. Although I still do not feel proficient in my programming skills, the format works because complimentary skills from members of my lab allow us to continually teach each other. I have enjoyed seeing your initial attempts at virtual Code Clubs as YouTube videos and know that those you make for the proposed project will give my group more material to strengthen our skills with reproducible research practices.

As we have discussed, my research group generates mass spectrometry data from a variety of symbiotic populations. These are large, messy datasets that require us to move data between different proprietary software packages. I would love to be able to help you find interesting questions and datasets that you could use to develop Code Club sessions that incorporate metabolomics data. I look forward to working with you and supporting the proposed project.

Sincerely,



Marcy J. Balunas, Ph.D.
Associate Professor of Medicinal Chemistry

69 NORTH EAGLEVILLE ROAD, UNIT 3092
STORRS, CT 06269-3092
PHONE 860.486.3051
FAX 860.486.6857
EMAIL marcy.balunas@uconn.edu

Resource Sharing Plan

As outlined in the project description, all materials will be made publicly open to any individual through a website (www.riffomonas.org) and its GitHub repository (www.github.com/riffomonas). All materials will be available under a Creative Commons Attribution 2.0 Generic (CC BY) license.