Project Title:

Training modules for improving the reproducibility of experimental data recording and pre-processing with examples from microbiology and immunology

Cover Letter:

Please accept the attached grant "Training modules for improving the reproducibility of experimental data recording and pre-processing with examples from microbiology and immunology" by Drs. Brooke Anderson and collaborators in response to to *RFA-GM-18-002: Training Modules to Enhance the Rigor and Reproducibility of Biomedical Research*.

This submission is from Colorado State University. The proposal does not include Human Subjects, Vertebrate Animals, Biohazards, or Select Agents.

This proposal [does what]. Therefore, it would be useful if one of the reviewers had [specific expertise].

Please assign the application to the following:

National Institute of General Medical Sciences (NIGMS). *Rationale:* [Why]

We thank you for the opportunity to submit this proposal.

Regards,

Brooke Anderson (PI)
Mike Lyons (Co-I)
Marcela (Co-I)
Mercedes (Co-I)
Others?

1681 Campus Delivery
Fort Collins, Colorado 80523-1681
Telephone: 203-508-2738
Email: brooke.anderson@colostate.edu

Budget justification

Personnel

Brooke Anderson, *Ph.D., Principal Investigator, x academic person-months, x summer person-months (x% effort) in Years 01 through 02, x academic person-months, x summer person-months (x% effort) in Year 03.* Dr. Anderson is an assistant professor of Epidemiology in the Department of Environmental & Radiological Health Sciences at Colorado State University, with an affiliate position at the Department of Statistics. She is an expert in R programming and has created and published several open-source R packages, in particular to facilitate environmental epidemiological research. She has experience creating R programs to work with large data, including climate model output and large weather datasets, as well as programs that interface with open web-based datasets. She is the co-instructor of a series of Massive Open Online Courses on *Mastering Software Development in R* through Coursera and an associated open online book. Dr. Anderson will lead the development of all training modules developed through this grant, including through supervising the development and integration of training materials from co-investigators. She will also lead user testing and other evaluation of all developed modules to ensure the developed modules are clear, effective, and well-matched to meet the needs of biological researchers from a variety of scientific backgrounds, including those new to programming.

Mike Lyons, *Ph.D., Co-Investigator, x academic person-months, x summer person-months (x% effort) in Years 01 through 02, x academic person-months, x summer person-months (x% effort) in Year 03.*

Marcela Henao-Tamayo, *Ph.D., Co-Investigator, x academic person-months, x summer person-months (x% effort) in Years 01 through 02, x academic person-months, x summer person-months (x% effort) in Year 03.*

Mercedes Gonzalez-Juarrero, *Ph.D., Co-Investigator, x academic person-months, x summer person-months (x% effort) in Years 01 through 02, x academic person-months, x summer person-months (x% effort) in Year 03.*

Travel

Domestic travel. Funds for PI (Dr. Anderson) to travel to one domestic conferences ($1,500 per conference) to learn about cutting edge reproducible research techniques. Potential conferences to be attended include the International R Users' Conference (UseR) on the years it is in the U.S. (typically every other year) or the Annual RStudio Conference. This domestic travel budget also includes funds ($900 per trip) for Dr. Anderson (PI) to travel to up to two Program Meetings over the course of the project. No travel is anticipated in Year 3 of the project, when the focus will be on final evaluation and refining of the training developed and published in Years 01 and 02.

- Year 01: $2,400
- Year 02: $2,400
- Year 03: $0

Materials and Supplies

Annual funds ($1,000 / year) are also requested for recording equipment (e.g., microphone), screen-capture software (e.g., *Camtasia*), other software, and books to facilitate the proposed

training module development.

Conference Registrations. Funds are budgeted for Years 01 and 02 for Dr. Anderson to register for a yearly domestic conference on cutting-edge tools and principals for conducting computationally reproducible research (e.g., UseR or RStudio Conferences). Attending a conference each year will help us include the most up-to-date tools and approaches in the developed training modules.

- Year 01: $500
- Year 02: $500
- Year 03: $0

Hospitality. Funds are budgeted each year to provide breakfast, lunch, coffee, and snacks for two days to 20 people (for the proposed annual extended user testing meetings at Colorado State University; budgeted at $850 / year). Colorado State University provides room reservations free to faculty for similar events, and so funds to rent a space are not required (See Letter of Support, Dr. Jac Nickoloff).

- Year 01: $850
- Year 02: $850
- Year 03: $850

Consulting? For evaluation of the completed modules. Also, for recording demographics, etc., for evaluation purposes. Check the requirements in the grant call for this.

Project Abstract:

We propose to develop training modules for improving the reproducibility of experimental data recording and pre-processing in scientific research. We aim to ensure these training modules are useful to laboratory-based researchers, who may have less prior training in open source software tools for reproducible research than researchers from biostatitics, epidemiology, and other disciplines. To ensure this, we will feature in these training modules examples from microbiology and immunology. We will create two sequences of modules that focus on improving computational reproducibility in the recording and pre-processing of experimental data, each containing approximately 12 modules, each featuring 5–30 minute videos, with supplemental online text, references, and practice exercises. The first sequence will be "Improving the Reproducibility of Experimental Data Recording", and it will include modules on ... . The second sequence will be "Improving the Reproducibility of Experimental Data Pre-Processing", and it will include modules on ... . These two sequences of training modules will be collectively published as an open online book using the *bookdown* technology. Each module will form a chapter of this book, and will feature an embedded YouTube video of 5–30 minutes, with accompanying text in the book to provide trainees with a more detailed written reference they can refer to after completing the video module. Each module's chapter will conclude with practical exercises or open discussion questions to complement the material taught in the video. To ensure this material is completely free and open to researchers in the United States, we will publish this online book and the videos under a Creative Commons license.

Project Narrative:

We will develop training modules for improving the reproducibility of experimental data recording and pre-processing in scientific research. To ensure accessibility and relevance to laboratory-based researchers, we will feature in these training modules examples from microbiology and immunology. We will create approximately 25 short training modules collected in two sequences, one focusing on on reproducible approaches to recording experimental data and one on reproducible approaches to pre-processing experimental data. Each module will feature a video lecture of 5–30 minutes, and the full collection of video modules will be embedded in a free, open online book, with supplemental text and practical exercises to accompany each module. Training modules will be evaluated for researchers at a variety of training levels (undergraduates to faculty), drawing mainly from the Microbiology, Immunology, & Pathology Department of Colorado State University.

Specific Aims

Significance and educational aims of proposed modules.

Proposed content of training modules.

Format of training modules.

Evaluation of training modules.

Project team. This project will bring together experts in R programming (Anderson, Lyons), including its use to improve the computational reproducibility of health-related research, with laboratory-based academic researchers in Microbiology and Immunology (Co-Is Henao-Tamayo, Gonzalez-Juarrero) who are attuned to the needs of and barriers to improving the reproducibility of experimental data collection and pre-processing. Our team will allow us to develop training modules that both present state-of-the-art approaches and tools to reproducibility, but do so in a way that is prioritized to be most useful and accessible to health researchers whose training has focused on laboratory-related, rather than computational, methods, and for whom existing training materials on computational reproducibility might be hard to understand or apply to their own research projects.

A Significance

"Does the proposed program address a key audience and an important aspect or important need in training in rigor and reproducibility? Is there convincing evidence in the application that the proposed program will significantly advance the stated goal of the program?"

B Innovation

"Taking into consideration the nature of the proposed research education program, does the applicant make a strong case for this program effectively reaching an audience in need of the program's offerings? Where appropriate, is the proposed program developing or utilizing innovative approaches and latest best practices to improve the knowledge and/or skills of the intended audience?"

C Approach

"Does the proposed program clearly state its goals and objectives, including the audience to be reached, the content to be conveyed, and the intended outcome? Is there evidence that the program is based on a sound rationale, as well as sound educational concepts and principles? Is the plan for evaluation sound and likely to provide information on the effectiveness of the program?"

C.1 Proposed Research Education Program Plan

"While the proposed research education program may complement ongoing research training and education occurring at the applicant institution, the proposed educational experiences must be distinct from those research training and research education programs currently receiving federal support. When research training programs are on-going in the same department, the applicant organization should clearly distinguish between the activities in the proposed research education program and the research training supported by the training program. The research education proposed must be **targeted to trainees and investigators at any level**. State the **goals for education** and **justify the area of training** selected for module development in terms of its **relevance and potential impact** on improving the development of skills and knowledge important for conducting rigorous and reproducible research. Describe the **subject material** to be covered. Describe the **format** for the training module proposed and **justify it in terms of the education goals**. The **length** of the proposed training module should be explained in terms of **scope and depth of coverage** of the subject matter. In addition, **how the research education will be utilized by trainees or investigators** should be described—for example, a module on how to avoid confirmation bias to be taken by all beginning laboratory workers, or a module on appropriate design of animal studies to be taken immediately prior to beginning such work. Describe the **plans for piloting and evaluating the effectiveness** of the training module. Describe **plans for making the proposed training module section 508 compliant of the Rehabilitation Act** (29 U.S.C. '794 d), as amended by the Workforce Investment Act of 1998 (P.L. 105 220; see http://www.section508.gov/ for additional information). Provide a **timeline for module development, piloting and refinement, dissemination, evaluation, and maintenance**. This timeline must propose **making the training publicly available within two years** of the award date."

C.1.1 Educational goals of the modules

C.1.2 Module subject material

We propose to develop two collections of modules, **Improving the Reproducibility of Experimental Data Recording** and **Improving the Reproducibility of Experimental Data Pre-Processing**.

The **Improving the Reproducibility of Experimental Data Recording** collection will include the following modules:

1. The principals of "tidy" data
2. Creating spreadsheet templates for experimental data collection
3. Example spreadsheet template: A template for collecting CPU data for a tuberculosis study [make more specific]
4. Choosing a spreadshet program for reproducible data collection: Excel, Google Sheets, and RStudio
5. Organizing data recording and meta-data recording through RStudio "Projects"
6. Creating "Project Templates" for consistency across projects
7. Example of creating a "Project Template": A project template for a lab group studying tuberculosis [make more specific]
8. Harnessing version control (git and GitLab) to improve transparency in data recording
9. Using git from RStudio
10. Using GitLab for version controlled collaborations

The **Improving the Reproducibility of Experimental Data Pre-Processing** collection will include the following modules:

1. An introduction to R code scripts
2. The benefits of scripts for data pre-processing
3. Getting started with RMarkdown
4. The relationship between R code scripts and RMarkdown documents
5. Creating reproducible data pre-processing protocols using Rmarkdown
6. Example of a reproducible data pre-processing protocol: Automated gating for flow cytometry data
7. Example of a reproducible data pre-processing protocol: Measuring metabolite feature intensities for metabolomics LC/MS data
8. Complex data types in R and their use in Bioconductor packages
9. Converting from complex data types to "tidy" formats for data analysis and visualization with R's "tidy" data tools

C.1.3 Format for the training modules

- Online book created through the "bookdown" format, with each module as a book chapter. We can use Git Pages to host this (CSU options for web hosting?).
- Training videos embedded for each module, each 5–30 minutes. Videos will be similar to online course lectures and will be hosted using YouTube. Embedding in the book will allow users to watch videos without leaving the book's webpage.
- Each chapter will end with exercise questions (around 10 questions, combination of discussion questions and applied exercises), as well as an embedded video with discussion of the discussion questions and a detailed walk-through of answers to applied exercises.
- Possibly host this through an online course platform like DataCamp?

C.1.4 Piloting and evaluating effectiveness of training modules

C.1.5 Insuring compliance with Rehabilitation Act

C.2 Team

C.2.1 Program Director/Principal Investigator

> "Is the PD/PI capable of providing both administrative and scientific leadership to the development and implementation of the proposed program? Is there evidence that an appropriate level of effort will be devoted by the program leadership to ensure the program's intended goal is accomplished? If the project is collaborative or multi-PD/PI, do the investigators have complementary and integrated expertise; are their leadership approach, governance and organizational structure appropriate for the project?"

> "Describe **arrangements for administration** of the program. Provide evidence that the Program Director/Principal Investigator is actively engaged in research and/or teaching in an area related to the mission of NIH, and can **organize, administer, monitor, and evaluate the research education program**. For programs proposing multiple PDs/PIs, describe the complementary and integrated expertise of the PDs/PIs; their leadership approach, and governance appropriate for the planned project."

C.2.2 Other members of the team

C.3 Institutional Environment and Commitment

> "Describe the institutional environment, reiterating the **availability of facilities and educational resources** (described separately under Facilities & Other Resources), that can contribute to the planned Research Education Program. Evidence of institutional commitment to the research educational program is required. A **letter of institutional commitment** must be attached as part of Letters of Support (see below). Appropriate institutional commitment should include the provision of adequate staff, facilities, and educational resources that can contribute to the planned research education program."

C.4 Evaluation Plan

> "Applications must include a plan for evaluating the activities supported by the award in terms of their **frequency of use** and their **usefulness**. The use of **multiple evaluation approaches** is highly encouraged as is **testing several groups with different characteristics**. The application must specify **baseline metrics (e.g., numbers, educational levels, and demographic characteristics of test group)** in a structured format, as well as **measures to gauge the short and long-term success of the research education award in achieving its objectives**. Applicants are expected to **obtain feedback from test group** to help identify weaknesses and to provide suggestions for improvements, and **make the evaluation and feedback data** available to NIGMS staff."

Pilot / text group evaluation:

- Work with GAUSSI to use some students as pilot testers?
- Recruit researchers / faculty as pilot testers?
- Work with CSU's Research Ethics group to figure out ways to pilot?

Long-term evaluation:

- Google Analytics for online book. How often are people accessing the book? How long are they spending on the book website? Where are the people accessing the book?
- YouTube analytics for the embedded videos. How often are people accessing the book? How long are they spending on the book website? Where are the people accessing the book?
- Quiz for each chapter of the book? Use to evaluate how well they've mastered the material? (Possibly could use embedded Shiny apps for this? Other ways to do this?)
- Rating options for each chapter of the online book? Usefulness? What they learned?
- Survey within each chapter of the online book? Educational level, demographic characteristics.

C.5 Dissemination Plan

> "A specific plan must be provided to disseminate the finished training modules **nationally** and make them **freely accessible**. In addition, links to these modules will be posted and maintained on the NIGMS web site."

C.6 Timeline

> "Provide a timeline for **module development**, **piloting and refinement**, **dissemination**, **evaluation**, and **maintenance**. This timeline must propose making the training publicly available within two years of the award date."

D Works cited

E Environment

"Will the scientific and educational environment of the proposed program contribute to its intended goals? Is there a plan to take advantage of this environment to enhance the educational value of the program? Is there tangible evidence of institutional commitment? Where appropriate, is there evidence of collaboration and buy-in among participating programs, departments, and institutions?"

- Computers. Access to needed software.
- Library. Access to many recent books online.
- Recording equipment / studio?
- Teaching expertise?
- Research Rigor & Ethics center / training?
- Tech Transfer