Specific Aims

Significance and educational aims of proposed modules. Among the steps in conducting experimental research, the collection and pre-processing of experimental data are steps where it is critical to ensure good practices to ensure research reproducibility. However, much of the training available for computational reproducibility focuses on later steps of the research process, such as the analysis of processed experimental data. Here, we aim to develop training modules that focus on principals and techniques of improving the reproducibility of research at these stages of the research process. We will focus on making these modules accessible and useful to laboratory-based researchers by including examples of improving reproducibility at these stages of microbiology and immunology research projects and will evaluate the training modules through user testing among laboratory-based researchers.

Proposed content of training modules. We will develop two sequences of modules to train researchers in how and why to improve reproducibility in the collection and pre-processing of experimental data, including training on implementing these approaches using open source R software tools. Working with laboratory-based co-investigators on our team, we will ensure that these modules and the examples used in them are approachable and useful to researchers without extensive computational training, helping the modules address an audience beyond those of existing training materials on R programming tools for reproducible research. Further, the training modules will focus on reproducibility in the data collection and pre-processing stages of projects, rather than in experimental design or data analysis.

The first sequence will be "Improving the Reproducibility of Experimental Data Recording", and it will include modules on the principals of the tidy data format, creating and using spreadsheet templates for data collection, developing project templates for organized and consistent data and meta-data collection, harnessing version control to improve transparency in data collection, and using GitLab for version-controlled collaborations. The second sequence will be "Improving the Reproducibility of Experimental Data Pre-Processing", with modules on how and why to use code scripts for data pre-processing, creating reproducible data pre-processing protocols using *rmarkdown*, examples of developing reproducible data pre-processing protocols, and converting from complex Bioconductor data types to tidy data formats to allow use of R's "tidy" data tools.

Format of training modules. Each module within each of these sequences will focus on a video lecture of 5–30 minutes. These videos will be collected together in an online book, where each module will form a chapter with an embedded video, and text and applied exercises or discussion questions will accompany each video. This book will be freely and openly published online under the Creative Commons license, to ensure it is available to any U.S. researcher. To ensure compliance with x, we will include in the online book transcripts for each training video.

Evaluation of training modules. We will conduct two-day user testing sessions each year of the grant. These will be focused on scientists at a variety of levels (undergraduate to faculty) and will determine the usefulness, clarity, and relevance of the developed modules to these researchers. These evaluation sessions will include live presentations of the lectures to be taped as modules, as well as directed work-throughs of the practical exercises included in the online book for each module. We will focus these testings on members of Colorado State University's Department of Microbiology, Immunology, & Pathology.

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 (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.