Specific Aims

Many excellent free training resources exist to improve the computational reproducibility of biomedical research. However, most of these materials target researchers at the stage of *data analysis*, and provide much less guidance principals and techniques to improve the reproducibility of the earlier steps of **experimental data recording** and **experimental data pre-processing**. In this project, we will create training modules to fill this gap. A key aim is to make these modules **accessible and useful to laboratory-based researchers** by including examples from real microbiology and immunology research projects and by piloting the training modules among laboratory-based biomedical researchers.

Content of training modules. We will develop two sequences of modules. The first sequence, **"Improving the Reproducibility of Experimental Data Recording"**, will explore the pitfalls of combining experimental data recording and analysis through the use of macro-enabled spreadsheets, explain the power of recording data in structured data formats, present the 'tidy' data format as one implementation of structured data, explain how reproducibility can be improved by using consistently-structured 'Project' directories to store all research project files, and demonstrate the use of *git* and *GitLab* to maintain single, current versions of all files while tracking the evolution of those files. The second sequence, **"Improving the Reproducibility of Experimental Data Pre-Processing"**, will focus on improving the reproducibility of experimental data pre-processing steps, like gating for flow cytometry data and peak finding / quantifying for mass spectrometry data. Training materials will explain how the use of code scripts for these steps dramatically improve reproducibility compared to using vendor-supplied point-and-click software and will introduce trainees to some of the popular R software extensions for this pre-processing. This sequence will also include advice on how to use literate programming tools (*Rmarkdown*) to create well-documented data pre-processing protocols that a research group can reuse to consistently and reproducibly pre-process the experimental data they collect. Each module will fall into one of three categories for teaching reproducibility: (1) principals; (2) implementation; and (3) case study examples. Implementation modules will focus on tools available through the popular open source R software and its RStudio interface. 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.

Format and dissemination of training modules. All training modules will be collected together in an online book, with each chapter covering one module. The chapter will center on an embedded YouTube video with a recorded lecture of 10–25 minutes, recorded in Colorado State University's professional-grade video recording facilities. The chapter will include written text to supplement the video lecture and to be used as a later reference by trainees. Each chapter will end with additional educational materials crafted to reinforce the video lecture, including discussion questions, applied exercises, and multiple choice quizzes. We will create this book using R's *bookdown* framework and will publish it freely and openly online—under the Creative Commons 3.0 license—using Git Pages, with Google Analytics enabled to aid in evaluation.

Evaluation of training modules. Our evaluations of the training modules developed under this grant will be assisted by an expert in program evaluation from Colorado State University's Science, Technology, Education, and Mathematics (STEM) Center (Maertens). **These evaluations 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.** We will conduct project evaluations of: (1) on-campus pilot testers; (2) off-campus pilot testers; (3) workshop participants at a national microbiology meeting; and (4) online users of the final online book. We will collect evaluation resultsthrough website analytics, quantitative survey questions, open-ended survey questions, and focus-group-style feedback generated through biannual full-day pilot testing sessions at Colorado State University and at a workshop at the American Association for Microbiology's annual meeting. Results on the long-term benefits of the training modules will be collected by one-year follow-up surveys to the pilot testers and workshop participants.

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 among laboratory-based biomedical researchers**. Our team will allow us to develop training modules that 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.