

Specific Aims

Many excellent and 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 on principles 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. **Our primary educational goal is to introduce the language, concepts, and tools for reproducible data recording and pre-processing to laboratory-based scientists whose attention is rather to their experimental technique and collection of accurate data, and who may have little or no background in the use of general purpose software tools.** To meet this goal, we will:

1. Develop training modules for reproducible data recording and data pre-processing. Many common existing practices—including use of spreadsheets with embedded formulas that concurrently record and analyze experimental data, problematic management of project files, and reliance on proprietary, vendor-supplied point-and-click software for data pre-processing—interfere with the transparency, reproducibility, and efficiency of laboratory-based biomedical research projects. **We will create training modules that will teach laboratory-based biomedical researchers how simple computational reproducibility principles can improve reproducibility at the stages of data recording and pre-processing.** We propose to develop two sequences of modules, **Improving the Reproducibility of Experimental Data Recording** and **Improving the Reproducibility of Experimental Data Pre-Processing**. While this proposal uses examples from microbiology and immunology experiments, the collection of primary data in a flexible, open source, transparent, and reproducible format or data structure, which can be kept in its primary state without the need for additional modification, is a solution that extends throughout large- and small-scale biomedical science projects.

2. Format and disseminate training modules as an online book. All training modules will be collected in an online book, with one chapter per module. The chapters will begin with an embedded video lecture of 10–25 minutes, recorded in Colorado State University's professional video recording facilities, and will include written text to supplement the video lecture, which can 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 a Creative Commons license using *GitHub Pages*, with *Google Analytics* enabled to aid in evaluation.

3. Pilot test and evaluate training modules with a broad sample of laboratory-based biomedical scientists. We will pilot test the training modules among scientists at a variety of levels (undergraduate to faculty) to determine the usefulness, clarity, and relevance of the developed modules to our target audience of laboratory-based biomedical researchers. We will conduct evaluations among: (1) on-campus pilot testers at Colorado State University; (2) workshop participants at the American Society for Microbiology's annual conference; and (3) early users of the online book. We will collect feedback through website analytics, quantitative survey questions, open-ended survey questions, and focus-group-style feedback for in-person pilot testing. Results on the long-term benefits of the training modules will be collected by six-month follow-up surveys to pilot testers and workshop participants. Our evaluation of the training modules will be assisted by an expert in program evaluation from Colorado State University's Science, Technology, Engineering, and Math (STEM) Center.

Our project's **primary goal** is to develop training modules that address the needs of laboratory-based biomedical researchers seeking to improve reproducibility, especially of experimental data recording and pre-processing, in their research projects. The **expected result** of this project is an online book that contains twenty short training modules as separate chapters, with video lectures, written text, and additional educational materials collected within each module's chapter. We consider it critical that these training materials be clear, relevant, and useful to a key audience of biomedical scientists whose primary research activities focus on laboratory research, rather than data analysts or statisticians.

Project team. Our project team brings 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, Robertson) 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 for 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.