The *Riffomonas* YouTube Channel: An Educational Resource to Foster Reproducible Research Practices

Running title: *Riffomonas* YouTube Channel

Patrick D. Schloss[1†]

† To whom correspondence should be addressed: pschloss@umich.edu

1 Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109

Educational resource

Abstract

Methods for analyzing data in a reproducible manner are often viewed as impenetrable to scientists more familiar with laboratory research. The *Riffomonas* YouTube channel is committed to teaching these scientists and others how to engage in reproducible research using modern data science tools.

As high throughput data generation becomes more common in microbiology and other disciplines there is a significant need for laboratory scientists to develop data science skills (1). Unfortunately, traditional undergraduate and graduate biology training programs are often deficient in opportunities for scientists to develop the skills necessary to analyze large datasets in a reproducible and robust manner (2, 3). Numerous organizations seek to fill this void including the Carpentries, Codeacademy, and DataCamp (4). There are also numerous video tutorials available on YouTube. Although the content available through these platforms are popular, there has been a gap in content that emphasizes project-based learning.

The *Riffomonas* YouTube channel (https://www.youtube.com/c/RiffomonasProject) seeks to fill this gap. I started consistently posting videos at the beginning of the COVID-19 pandemic in ~~the Spring of~~ April 2020. As of the end of November 2022, the channel included 285 videos that had been viewed 635,947 times; the channel had 11,327 subscribers. The majority of these are 264 videos in the "Code Club" playlist (5) (Table 1). Other videos are related to a previously described tutorial series on reproducible research (6) and series where reproducible reseach practices are used to address topical questions. Code Club videos are typically between 20 and 30 minutes long. The code that is developed in the videos is available through a website (https://riffomonas.org/code_club/) and the channel's GitHub-hosted account (https://github.com/riffomonas).

The name, *Riffomonas*, comes from the concept of "riffing" where musical themes are adapted to achieve a similar sound, albeit perhaps in a different context (6). This is to emphasize the value of reproducibility not only to recreate a set of results but to apply a method with a different dataset (7). The channel covers topics related to reproducible data analysis practices including R programming, data visualization, project organization, version control, command line programming, workflow tools, and scientific publishing (Table 1). Each video includes a brief introduction followed by me live coding to achieve a goal. I emphasize the use of live coding to modulate the rate of instruction and to show viewers my own coding practices. Observing a experienced analyst make mistakes normalizes some level of failure and demonstrates the strategies they can use to resolve their own mistakes. Viewers are encouraged to follow along with each video and to apply the new information to their own project.

Each video emphasizes a specific topic, but includes other content that is selected to review topics covered in recent videos. Although videos can be watched individually, they often form a project arc (Table 1). For example, between July 2020 and July 2021, I formulated a research question, obtained and analyzed data to answer the question, and wrote a paper that was published in *mSphere* (8). This series of 67 videos covered every topic from creating the initial directory on my computer to house the project files through reviewing the proofs of the published manuscript. Other project arcs have included visualizing

microbiome data, modeling microbiome data using machine learning tools, analyzing the impacts of rarefying microbiome data, and other topics. Going forward, the *Riffomonas* channel will continue to post project-based content to help researchers develop their reproducible research skills.

Acknowledgements

I am grateful to the audience of the *Riffomonas* channel for their feedback on topics that I should cover in future episodes.

References

1. **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.

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

3. **Williams JJ**, **Drew JC**, **Galindo-Gonzalez S**, **Robic S**, **Dinsdale E**, **Morgan WR**, **Triplett EW**, **Burnette JM**, **Donovan SS**, **Fowlks ER**, **Goodman AL**, **Grandgenett NF**, **Goller CC**, **Hauser C**, **Jungck JR**, **Newman JD**, **Pearson WR**, **Ryder EF**, **Sierk M**, **Smith TM**, **Tosado-Acevedo R**, **Tapprich W**, **Tobin TC**, **Toro-Martínez A**, **Welch LR**, **Wilson MA**, **Ebenbach D**, **McWilliams M**, **Rosenwald AG**, **Pauley MA**. 2019. Barriers to integration of bioinformatics into undergraduate life sciences education: A national study of US life sciences faculty uncover significant barriers to integrating bioinformatics into undergraduate instruction. PLOS ONE **14**:e0224288. doi:10.1371/journal.pone.0224288.

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

5. **Hagan AK**, **Lesniak NA**, **Balunas MJ**, **Bishop L**, **Close WL**, **Doherty MD**, **Elmore AG**, **Flynn KJ**, **Hannigan GD**, **Koumpouras CC**, **Jenior ML**, **Kozik AJ**, **McBride K**, **Rifkin SB**, **Stough JMA**, **Sovacool KL**, **Sze MA**, **Tomkovich S**, **Topcuoglu BD**, **Schloss PD**. 2020. Ten simple rules to increase computational skills among biologists with code clubs. PLOS Computational Biology **16**:e1008119. doi:10.1371/journal.pcbi.1008119.

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

7. **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.

8. **Schloss PD**. 2021. Amplicon sequence variants artificially split bacterial genomes into separate clusters. mSphere **6**. doi:10.1128/msphere.00191-21.

Table 1. Description of the data science topics and project based series covered in the playlists found on the *Riffomonas* YouTube Channel. Because most videos cover more than one topic they are found in mutiple playlists. Playlists and counts were current as of December 1, 2022. Playlists can be found under the Playlist tab at https://www.youtube.com/c/RiffomonasProject.

	Number of videos
Videos covering data science topics	
Data visualization with R's tidyverse and allied packages	146
Data manipulation within R's tidyverse and other packages	116
Data analysis with base R	39
Tools for reproducible data analysis	33
Working at the command line	26
Literate programming with R markdown	18
Machine learning with mikropml R package	16
Version control with git and GitHub	15
Scientific writing	15
Project organization	3
Project based series	
All Code Club videos since April 2, 2020	265
Microbiome data analysis and visualization	86
ASV/OTU senstitivity and specificity analyses	67
Visualizing COVID-19 vaccination attitudes	31
Climate change data visualization	29
Evaluating rarefaction and its alternatives	18
Drought index visualization	17
Reproducible reserach tutorial series	14
Commemorating Juneteenth 2022 with a visualization	5
2018 MLB All Star Break data analysis sprint	4