TidyBiology: An Introduction to Biological Data Science in R

CMB710: Cell and Molecular Biology Core Course

Instructor: Matthew Hirschey, Ph.D. (matthew.hirschey@duke.edu)

Teaching Assistant: Allie Mills, Ph.D., Akshay Bareja, D.Phil.

Class Dates: 10/9/2019 - 10/21/2019
Class time: 10:20am – 11:40am
Classroom: MSRB3, Room 1125

Office hours: Slack is a versatile communications and collaboration tool used in this course in place of formal office hours. Students are discouraged from using email for questions and discussions related to content of the course; emails about personal matters are allowed. Hence, course communication will happen using Slack. Both public (for announcements and general questions) and private (for team communication) channels are used. Feel free to post anytime for help.

Overview: This workshop-style module provides an introduction to the emerging field of Data Science in R, including data analysis and visualization, with a particular focus on its utility for biological insight. Students will be provided with biological datasets, and introduced to R packages and code used to examine data. In the first half of each class, students will be lectured on methods and shown demonstrations; in the second half of each class, students will use tools to analyze real data. Methods for filtering, sorting, and transforming data will be discussed along with visualization tools and options. Particular attention will be paid to code interpretation and data provenance methods by learning to generate reproducible data output files. For a final project, students will be given a new dataset to analyze using the tools learned during the course, and will share findings with the class in a short oral presentation. Although specific datasets will be used for analysis in class, this workshop will provide broadly applicable tools to reproducibly analyze and visualize data across the biological sciences.

Objectives: By the end of this course, students will be familar with R, the RStudio Integrated Developer Environment, exploratory data analysis, and the tools and concepts of the tidyverse. The overall goal is to provide students with enough initial training that they can continue to learn these tools beyond the class.

Evaluation: A final grade is an important metric summarizing your fluency and aptitude with the concepts presented in class. In this course, I will ask each student to be the judge of these criteria and submit their proposed grade and a justification for it after the conclusion of the class. The following metrics should be considered when proposing a grade:

40% Code, R markdown file, and summary interpretation of findings
30% In-class presentation of above findings
20% Help vignette
5% Attendance
5% Participation

While the instructors will ultimately determine the final grade for each student, thoughtful reflection and justification for a letter grade provided by each student will be heavily weighted. Self-critique is an important skill for scientists, and this activity presents and opportunity for it.

Attendance & Participation: This course is hands-on and experientially oriented. Students must bring a laptop computer to class. Activities in class will require you to:

1. Complete the class assignments
2. Be punctual
3. Be an active participant
4. Prepare a help vignette
5. Present a final project

You are expected to attend and participate in every class meeting. If you miss any class for any reason, you must let the instructors know and are responsible for finding out about any assignments or information from class.

Academic Integrity: Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and non-academic endeavors, and to protect and promote a culture of integrity. To uphold the Duke Community Standard:

- I will not lie, cheat, or steal in my academic endeavors;
- I will conduct myself honorably in all my endeavors;
- I will speak-up if the standard is compromised.

Commitment to Diversity and Inclusion: Duke aspires to create a community built on collaboration, innovation, creativity, and belonging. Our collective success depends on the robust exchange of ideas—an exchange that is best when the rich diversity of our perspectives, backgrounds, and experiences flourishes. To achieve this exchange, it is essential that all members of the community feel secure and welcome, that the contributions of all individuals are respected, and that all voices are heard. All members of our community have a responsibility to uphold these values. If you have any concerns about diversity and inclusion in our class, please do not hesitate to reach out to your instructors.

Projects

Assignments
Each class will have a small home work assignment. In most instances, these assignments can be accomplished in class. In the rare case we run out of time, these assignments should not take more than 5-10 minutes to complete. The overall goal of including them is to provide tangible examples that reinforce concepts presented in class. Should you struggle with any of these assignments or the concepts presented within, please reach out to your instructors or fellow students on Slack.

Help Vignette
In order to learn how to read help files and vignettes, students will work throughout the module to generate a vignette for a single operation. The help vignette will be due prior to the beginning of the final class on10/21/2019. The series of help vignettes will be available to students after completion of the class, and will be a useful resource for future reference.

Final Project
This class will culminate with a final data science project. The overall objective will be to analyze a data set, provided by the instructors, and generate an image that suggests a testable hypothesis. The final products will be the R markdown code file, a knitted R markdown report, and a final saved image. Each student will give a short 2-3 minute presentation on the final day (10/21/2019) describing their analysis method, accompanying code, and their findings. Students will receive feedback from peers and instructors on presented content, shown visuals, and mechanics of the presentation.

Class Schedule

Date	Topic	Homework
Pre-class	'Data science for the scientific life cycle | eLife'	Sign-up for RStudio.cloud, install Slack and post, install R and RStudio
Oct 9	Introduction to Data Science, the R programming language, the Tidyverse packages	01_homework.Rmd, and begin help file
Oct 11	Dplyr package and data wrangling	02_homework.Rmd, continue help file
Oct 14	ggplot package and making graphs	03_homework.Rmd, continue help file
Oct 16	Joining and shaping data, strings, and miscellaneous	04_homework.Rmd, continue help file
Oct 18	Rmarkdown, knitr,and presentation skills	05_exercises, finalize report, help file, and presentations
Oct 21	Final Presentations	Submit peer reviews and individual grading summary