High Throughput Biomedical Data Analysis

Spring (April-June), 2024

COURSE DESCRIPTION:

This class is a an introduction to methods and tools for biomedical data analysis. The course will introduce basic and intermediate bioinformatics and statistical analyses for high-throughput biology, including sequencing data preprocessing and quality control, genome-wide association studies, gene expression and proteomics, epigenetics, metagenomics, and imaging data analysis. We will describe relevant methods and procedures for each context, and practice and apply code in each case. Examples will come from a variety of applications from molecular biology and biomedical research. Students be expected to use R and GitHub throughout this course.

COURSE OBJECTIVES:

Students who take this course will:

1. Become familiar with bioinformatics workflows for several biomedical data types
2. Learn to effectively plot and visualize data generated from multiple technologies
3. Gain advanced experience in writing code and using existing bioinformatics tools
4. Practice principles and concepts of data sharing, reproducibility, and open science

PREREQUISITES

A working knowledge and experience with R is required. We recommend that students complete MBGC 5002Q and prior to registering for this course, or have equivalent experience. An introductory course in statistics, biostatistics, or equivalent experience in statistical analysis is also recommended (e.g., GSND 5135Q or MSBS N5030). Students without sufficient coding or statistics experience will need to obtain these skills before the course. Please contact Dr. Johnson to obtain a list of the required proficiencies and asynchronous resources available to obtain the needed programming or statistics skills.

COURSE FORMAT:

This class will be taught virtually using a synchronous remote modality, although students will be provided a classroom to gather for each lecture. A co-instructor will be present in the classroom for each lecture. Class will occur Mondays and Wednesdays from 10:00am-11:50am. Courses may also be recorded and made available for students who need to miss classes due to personal reasons, illness, or research related needs.

ZOOM LINK AND CLASSROOM:

Zoom Meeting ID for all sessions is 912 9274 4505, with the password: 961614, or use the following direct link (the link is also available though the course GitHub page): https://rutgers.zoom.us/j/91292744505?pwd=L1RCWHBkQXdZTkQ5T0g4SWpIYXJIdz09.

Room MSB B619 will also be available for the students to congregate for each lecture, with a co-instructor present.

FACULTY AND STAFF:

W. Evan Johnson, Ph.D.
Email: w.evan.johnson@rutgers.edu
Cell Phone: (801) 472-6951

Co-instructor Samantha Bell, Ph.D.
Email: samantha.l.bell@rutgers.edu

Co-instructor Arthur VanValkenburg
Email: ajv120@njms.rutgers.edu

OFFICE HOURS:

Instructor: Dr. Johnson will be available virtually by appointment only. Email or text him any time to set up a time to meet!

Teaching Assistant: There is not TA for this course. However, Dr. VanValkenburg will be available for in-person help and support by appointment on a limited basis.

GitHub REPOSITORY:

The course GitHub repository is: https://github.com/wevanjohnson/2024_Spring_BDMA. This page will contain all information in this syllabus plus more. Homework assignments and other information pertinent to this course will be posted on this web site, which will be updated frequently, so you should visit it regularly.

CANVAS:

There will also be a Canvas course page for this course. This is where you will be able to access links to past lectures, and also turn in your homework (and track your HW grades). The rest of the course materials will only be posted on GitHub.

COURSE TEXTBOOKS:

We will use multiple text resources in this class. None are required, all are freely available online or can be purchased in hard-copy. Many of my materials are adapted from these resources (thanks to the authors for these):

1. *Modern Data Science with R*, 2nd edition, By Benjamin S. Baumer, Daniel T. Kaplan, Nicholas J. Horton, Chapman and Hall/CRC, 2021. https://mdsr-book.github.io/mdsr2e/
2. *Introduction to Data Science: Data Analysis and Prediction Algorithms with R*, 1st edition, By Rafael A. Irizarry, Chapman and Hall/CRC, 2020. https://rafalab.github.io/dsbook/
3. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, 2nd edition, By Hadley Wickham, Garrett Grolemund, O'Reilly, 2017 https://r4ds.had.co.nz
4. *Mathematical Foundations for Data Analysis*, By Jeff M. Phillips: https://mathfordata.github.io.
5. *The Elements of Statistical Learning*, 2nd Edition, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman: https://hastie.su.domains/ElemStatLearn/

EVALUATION METHODS & COURSE GRADING

Assessment/Evaluation: This course is a hands-on, project-based course. You will be graded based on homework assignments/mini projects (70%) and your final project (30%; there will be no final exam). Homework assignments and mini projects will be usually assigned at the beginning of each week and will be due by Wednesday of the week after the material is covered. However, please plan to be flexible on due dates based on the material covered in class.

Course Grading: Grade Scale:

$\geq$	90%	85%	80%	75%	70%	<70%
Grade	A	B+	B	C+	C	F

ATTENDANCE:

This course is being taught through a synchronous remote modality through Zoom. Attendance is mandatory; lecture recordings will only be available to students with university approved absences or pre-approved special circumstances. If you are sick or have any other justified reason to miss a lecture, please reach out to Dr. Johnson in advance and you will be reasonably accommodated.

WORKLOAD:

This is an 8-week, 2.0 credit class near the end of Spring 2024. In general, you should expect four hours of in class each week, and two hours outside of class for every hour in class.

OTHER HELP:

I **strongly** encourage you to contact early me if you have difficulty with the material. This course builds on material from prior lectures, so do not fall behind! My job is to help you understand the material as well as possible, and I am flexible with meeting times.

ACADEMIC INTEGRITY:

You are expected to have read and follow the guidelines at the university's academic integrity website (http://academicintegrity.rutgers.edu). These principles forbid plagiarism and require that every Rutgers University student to:

- Properly acknowledge and cite all use of the ideas, results, or words of others
- Properly acknowledge all contributors to a given piece of work
- Make sure that all work submitted as his or her own in a course or other academic activity isproduced without the aid of unsanctioned materials or unsanctioned collaboration
- Treat all other students in an ethical manner, respecting their integrity and right to pursue their educational goals without interference. This requires that a student neither facilitate academic dishonesty by others nor obstruct their academic progress (reproduced from: ttp://academicintegrity.rutgers.edu/academic-integrity-at-rutgers/).

Violations of academic integrity will be treated in accordance with university policy, and sanctions for violations may range from no credit for the assignment, to a failing course grade to (for the most severe violations) dismissal from the university.

COURSE TOPICS AND OUTLINE (BY WEEK; MAY BE MODIFIED AS NEEDED)

Sequencing data preprocessing and quality control (Week 1)

- 4/22/24: Course introduction; Installation; Introduction to R
- 4/24/24: More Introduction to R, R Markdown, and Unix

Sequencing data preprocessing and quality control (Week 2)

- 4/29/24: Introduction to Amarel and OARC
- 5/1/24: Introduction to Github; Introduction to FASTA and Other Sequencing Formats

Genomics data analysis: mutation, linkage, and phylogenetics (Week 3)

- 5/6/24: Genomic sequence variation analysis; Plots and summaries
- 5/8/24: Linkage analysis; Phylogenetic trees

Genome-wide association study (GWAS) analysis (Week 4)

- 5/13/24: Genome-wide association testing
- 5/15/24: Quantitative trait analysis

Gene expression and proteomics analysis (Week 5)

- 5/20/24: Clustering/heatmaps; Dimension reduction
- 5/22/24: Batch correction; Differential expression analysis

Network analysis (Week 6)

- 5/27/24: Memorial Day (No Class)
- 5/29/24: Pathway/network analysis

Single cell and epigenetic data analysis (Week 7)

- 6/3/24: Single cell RNA-seq analysis
- 6/5/24: Bump hunting/peak finding; Regulatory network analysis

Microbiome data analysis (Week 8)

- 6/10/24: Diversity analysis; Dimension reduction
- 6/12/24: Differential abundance; Co-occurrence and functional analyses; Final Projects Due

Imaging and spatial data analysis (????)

- Types of imaging data; Tools for image analysis
- Pattern recognition and machine learning