High Throughput Biomedical Data Analysis (GSND 5340Q)

Spring 2025, April-June

## COURSE DESCRIPTION:

This course introduces methods and tools for biomedical data analysis, focusing on basic and intermediate bioinformatics and statistical techniques used in high-throughput biology. Topics include sequencing data preprocessing and quality control, genome-wide association studies, gene expression and proteomics, epigenetics, metagenomics, and imaging data analysis. For each topic, we will review relevant analytical methods and procedures, with hands-on coding exercises to reinforce concepts. Examples will be drawn from a variety of molecular biology and biomedical research applications. Students are expected to use R and GitHub throughout the course.

## COURSE OBJECTIVES:

By the end of this course, students will:

1. Become familiar with bioinformatics workflows for a variety of biomedical data types
2. Learn to effectively visualize and interpret data generated from multiple technologies
3. Develop advanced skills in coding and in using established bioinformatics tools
4. Apply core principles of data sharing, reproducibility, and open science

## PREREQUISITES

A working knowledge and experience with R, GitHub, and Unix is required. Students are required to complete GSND 5345Q (Fundamentals of Data Science) prior to taking this course, or have equivalent experience. An introductory course in statistics or biostatistics is also recommended (e.g., GSND 5135Q or MSBS N5030), or comparable experience in statistical analysis. Students without sufficient background in coding or statistics are expected to acquire these skills before the course begins. Please contact Dr. Johnson for a list of required proficiencies and available asynchronous resources to help build the necessary programming and statistical foundations.

## 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-12:00pm. 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 Meeting ID: 924 8657 8277, with the password: 063331, or use the following direct link (the link is also available though the course GitHub page): <https://rutgers.zoom.us/j/92486578277?pwd=NAG3fMvBlLbrOv6e3zg1xuZjREu9Pm.1>

There will also be available for the students to congregate for each lecture, with a co-intructor present. This room is still TBA.

## FACULTY AND STAFF:

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

Co-instructor Arthur VanValkenburg  
Email: [ajv120@njms.rutgers.edu](mailto: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!

## GitHub REPOSITORY:

The course GitHub repository is located at: <https://github.com/wevanjohnson/2025_Spring_BMDA>. 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:

|  | 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 2025. 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)

Introduction: Sequencing data formats and quality control (Week 1)

* 4/23/25: Course introduction; Essential tools for Data Science
* 4/23/25: Data formats (.fasta, .fastq, .sam/.bam); FASTQC and multiQC **(asynchronous lecture)**

Genomics data analysis: Data wrangling and variation analysis (Week 2)

* 4/28/25: Sequence trimming and QC; alignment and wrangling
* 4/30/25: Genomic sequence variation analysis **(asynchronous lecture)**

Genomics data analysis: Linkage and genome-wide association (GWAS) (Week 3)

* 5/5/24: Linkage analysis; Phylogenetic trees
* 5/7/24: Genome-wide association testing

Gene expression and proteomics analysis (Week 4)

* 5/12/24: Quantitative trait analysis
* 5/14/24: Clustering/heatmaps; Dimension reduction

Gene expression and proteomics analysis; Network analysis (Week 5)

* 5/19/24: Batch correction; Differential expression analysis
* 5/21/24: Pathway/network analysis

Single cell and epigenetic data analysis (Week 6)

* 5/26/25: Single cell RNA-seq analysis **(Memorial Day; asynchronous lecture)**
* 5/28/25: Bump hunting/peak finding; Regulatory network analysis **(asynchronous lecture)**

Microbiome data analysis (Week 7)

* 6/2/25: Diversity analysis; Dimension reduction
* 6/4/25: Differential abundance; Co-occurrence and functional analyses; Final Projects Due

Imaging and spatial data analysis (Week 8)

* 6/9/25: Types of imaging data; Tools for image analysis
* 6/11/25: Pattern recognition and machine learning