BIOL 3411: Applied Sequence Analysis Protocols | Summer 2025 | MTWR 3:20 - 5:05 | Instructor: Jamie Henzy
Location: Snell Library 009

Molecular biology has become a big-data science. The wealth of sequencing data currently available holds keys to pressing issues in the life sciences such as the relationship between genotype and phenotype in complex traits; evolutionary dynamics of pathogens and their hosts; and rare variants involved in disease. Introduces students to concepts and skills needed to productively ask questions of genomic data and move toward participating in research projects that involve large sequencing data sets. Students learn good data practices while working with various online tools, databases, Unix commands, and basic R packages to analyze and visualize genomic data. While knowledge of coding is not a prerequisite, a fearless attitude toward computers is an asset!

Prerequisites: BIOL 2301

Learning objectives. Upon completion of the course, students will be able to:

- Access and download various types of sequencing data from a variety of databases
- Understand file structure and organize files and data in a Unix environment
- Use basic Unix commands to parse and perform operations on high-throughput data files
- Work with large amounts of data on Northeastern's high-performance computing cluster
- Use RStudio to perform differential gene expression analysis on RNA-seq data
- Perform a basic phylogenetic analysis
- Generate various types of plots to display genomic data
- Understand and implement "best practices" in computational research

Format: In-person. Each session will contain some lecture/discussion to introduce concepts, followed by lots of hands-on exercises working with various tools, as in a workshop.

Material:

- GitHub: readings and other material for the course is on my GitHub site
- Canvas: a schedule of the class sessions is in Modules on the Canvas site
- Laptop: please bring one to each class

Please note: The Mac operating system is strongly preferred for bioinformatics and this course is set up for Macs, with very limited support for Windows; **if you use a computer that runs on Windows, expect a few extra headaches!**

Submission of work: Each student in the course has an account on NU's "Explorer cluster". You'll submit assignments by placing them in specific folders that I can view in your directory on Explorer. A few assignments require pdf files that you'll submit to Canvas.

Grading: 80% of grade is based on timely and successful completion of six **modules**
20% of grade is based on two **quizzes**

Due dates for modules and quizzes:

Module 1 (UNIX)	Sun May 11
Module 2 (HPC)	Mon May 19
Module 3 (RStudio)	Tue May 27
Quiz 1, in-class	**Thu May 29**
Module 4 (DGE)	Thu Jun 5
Module 5 (Plots)	Wed Jun 11
Module 6 (Phylogen.)	Thu Jun 18
Quiz 2, take-home	**Tue Jun 24 – due**

Modules are each worth a maximum of 50 points. Each quiz is worth a maximum of 30 points. Total points available: 360. Grades are determined according to this chart showing the **absolute minimum percentages necessary** for each grade:

A	93%	B-	80%	D+	67%
A-	90%	C+	77%	D	63%
B+	87%	C	73%	D-	60%
B	83%	C-	70%		

Late policy for modules: Every student gets one free 2-day grace period. After that, each assignment submitted within 48 hours after the due date results in a drop of 5 points. **Please note: no work will be accepted beyond 48 hours after the due date.**

Attendance: Please note that this is not an online course. Much of the information you'll need to successfully work through the modules will be presented during class, where you'll also get the benefit of troubleshooting with your classmates and getting help from me.

If you miss a class: No need to let me know. It is your responsibility to get notes from a classmate on what you missed**.**

Assignments

UNIX/HPC: Upon completion you'll be very comfortable working on the command line and navigating the structure of your file system and that of the **h**igh-**p**erformance **c**omputing cluster, “Explorer”.

Exploring: **H**igh-**P**erformance **C**omputing refers to a cluster of interconnected computers that supply memory and processing power far beyond what your laptop or desktop provides. Anyone who works with genomic datasets needs to know how to interact with such a system to run analyses on large datasets. You'll learn to navigate the system and use various tools to explore and analyze sequences. You’ll align sequencing reads to a reference genome and find variants -- the spice of life-- involved in disease and evolution!

RStudio: Whole suites of R packages have sprung up to allow researchers to do bioinformatics in R. In this assignment, you'll learn how to use the various features of RStudio efficiently to prepare yourself for the fun to come in the remaining three assignments.

DGE: Differential gene expression analysis is a common tool in the transcriptomics tool kit. You'll perform this type of analysis start-to-finish on a couple of datasets, including actual data generated by a Northeastern U researcher. This assignment is the crown jewel of the set!

Plots: Visualization of complex data is tricky. This assignment will introduce you to common types of plots used in visualizing genomic data and outcomes of analyses.

Phylogenomics: Many insights into evolution and disease can be gained from comparing sequences of homologs, so much so that a whole subfield of "phylogenomics" (or phylogenetics) began developing on the coattails of sequencing technology. You'll learn how to perform simple phylogenetic analyses in R.