BIOL 2405: Methods in Genome Analysis | Spring 2025 | MR 11:45 - 1:25 | Instructor: Jamie Henzy
Location: Ryder Hall 266

Introduces students to concepts and computational skills required to productively mine, evaluate, and interpret genomic data. Emphasizes high-quality and reliable data practices by using a variety of online tools, databases, Unix commands, and basic R packages. Topics include navigating a Unix environment, accessing databases and working with sequence files, using a variety of bioinformatics tools to explore sequencing data, organizing data and running jobs on Northeastern's high-performance computing cluster, and implementing RStudio for gene expression analysis and phylogenetic analysis.

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: Please bring a **laptop** to every class. The Mac operating system is strongly preferred for bioinformatics; if you use a computer that runs on Windows, expect a few extra headaches! All readings and assignments can be found on the Canvas site.

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: 70% of grade is based on the number of **modules** completed
30% of grade is based on two **in-class quizzes**, each worth 150 pts

Points for modules completed on time:
7 modules: 700 pts
6 modules: 623 pts
5 modules: 553 pts
4 modules: 483 pts
Fewer than 4 modules: 0 pts

Due dates for modules and quizzes:

Module 1 (UNIX)	Thu Jan 23
Module 2 (HPC)	Thu Feb 6
Module 3 (Sequence Explorer)	Thu Feb 20
Module 4 (RStudio)	Thu Mar 13
Module 5 (DGE)	Mon Mar 24
Module 6 (Plots)	Thu Apr 3
Module 7 (Phylogen.)	Mon Apr 14

Quiz 1: **Mon Mar 17**
Quiz 2: **Mon Apr 21**

Module points plus quiz points enable a maximum of 1000 pts, and grades are determined according to this chart showing the **absolute minimum points 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 50 module 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 work successfully work through the modules will be presented during class. In class you'll also get the benefit of troubleshooting with your classmates and getting help from me. **If you miss a class, be sure to get notes from a classmate.**

Modules

UNIX: Upon completion you'll be very comfortable working on the command line and navigating the structure of your directory, all of which will be extremely useful for working with genomic datasets and the tools for analyzing them on a HPC.

HPC: **H**igh-**P**erformance **C**omputing refers to the use of 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.

Sequence explorer: You'll learn to use various tools to examine and analyze sequences, including tools to align sequencing reads to a reference genome. Uber-handy!

Variant analysis: Variants are the spice of life, involved in disease and evolution, so genomicists are committed to characterizing these differences and sorting the true variants from sequencing mistakes.

R Studio: Whole suites of R packages have sprung up to allow researchers to do bioinformatics in R. You'll use a version of RStudio available through the Explorer cluster to perform a differential gene expression analysis and create plots. With this module, you'll prepare yourself for the fun to come in the remaining three modules.

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 actual data generated by a Northeastern U researcher.

Plots: Visualization of complex data is tricky. This module 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.