BIOL 2406: Computational Tools for Genome Analysis

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, a high-performance computing cluster, 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: Can be taken concurrently with 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
- Design basic A to Z workflows for common genomic analyses
- 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

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.

Attendance: This course is formatted as a workshop and is similar to language courses in that you'll be learning the new languages UNIX and R and using them to write scripts, requiring frequent practice and use of your new "words". Additionally, computational work involves a lot of troubleshooting which is best done in the company of others who may be facing the same annoying snags and can help work through them. Therefore, attendance is an important component of the course and counts for 15% of the grade. Attendance is scored through a series of daily in-class quizzes that allow you to test your under-standing of course topics. There will be a quiz at the start of every class beginning with our 2nd meeting, totaling 23 quizzes. Everyone is generously allowed three absences, no questions asked (exam day is an exception). **The three absences include wellness days and any other personal absences, whether due to illness, accident, or bad hair.** Any absence beyond the allotted, however, will result in a zero on a quiz.

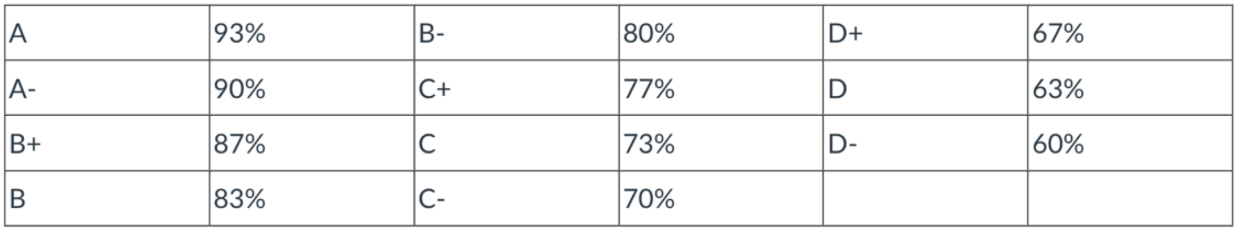
AI use: Personally, I'm tremendously excited about the possibilities offered by AI, while also wary of the downsides, and I look forward to experimenting with its use in this class, with you! We will have frequent conversations during the term about ways in which AI is useful for our work, and ways in which it can lead us astray. For purposes of this experiment, please document your uses of AI and share them in class. The objective is to use AI to make us smarter and not dumber!

Grading:	Timely and successful completion of six **modules**	60%
	In-class demo (2)	05%
	In-class daily quizzes (23, drop lowest 3)	15%
	Exams	20%

Due dates for assignments and quizzes:

1 UNIX_HPC	Sun Jan 25
2 Explore!	Mon Feb 9
3 RStudio	Tue Feb 24
4 Plots	Tue Mar 10
Quiz 1, in-class	**Thu Mar 12**
5 DGE workflow	Sun Apr 5
6 Phylogenomics	Sun Apr 12
Quiz 2, take-home	**Tue Apr 21 – due**

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%		

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”.

Explore!: **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.

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.

DGE: Differential gene expression analysis is a required 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!

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.

Below is a rough idea of the topics we'll cover from session to session. However, I'll post details for each session in Canvas Modules as we progress through the semester.

#	Date	Topics	Due dates
1	Thu Jan 8	Course introduction; Command line setup, best practices	
2	Mon Jan 12	UNIX basic commands	1st daily quiz
3	Thu Jan 15	Writing bash scripts; file permissions	
	Mon Jan 19	NO CLASS (Holiday)	
4	Thu Jan 22	Regular expressions; FASTQ files	UNIX_HPC tutorial
5	Mon Jan 26	Intro to Explorer: nodes and directories	
6	Thu Jan 29	Using modules on explorer	
7	Mon Feb 2	Running jobs on cluster: SLURM	
8	Thu Feb 5	Indexing a genome, mapping reads	
9	Mon Feb 9	Calling variants: genotype likelihoods	Explorer tutorial
10	Thu Feb 12	RStudio introduction	
	Mon Feb 16	NO CLASS (Holiday)	
11	Thu Feb 19	Wrangling data	
12	Mon Feb 23	Reading in and modifying GEO datasets	RStudio tutorial
13	Thu Feb 26	Plotting: basic and ggplot2	
	Mon Mar 2	NO CLASS (Spring Break)	

#	**Date**	**Topics**	**Due dates**
	Thu Mar 5	NO CLASS (Spring Break)	
14	Mon Mar 9	Generating a heatmap from real data	Plotting tutorial
15	**Thu Mar 12**	**Quiz 1**	
16	Mon Mar 16	Intro to differential gene expression (DGE) analysis	
17	Thu Mar 19	Statistics used in DGE analysis	
18	Mon Mar 23	DGE workflow in RStudio; PCA plots	
19	Thu Mar 26	Visualizing DGE results: heatmaps, scatterplots	
20	Mon Mar 30	Start-to-finish DGE with new data	
21	Thu Apr 2	Analyzing DGE results	DGE tutorial
22	Mon Apr 6	Pairwise and multiple-sequence alignment	
23	Thu Apr 9	Tree-building methods; bootstrapping	
24	Mon Apr 13	Phylogenetic analysis with "ape" package	
25	Thu Apr 16	Develop multi-platform workflow	Phylo. tutorial