CONTENT SUBJECT TO MINOR CHANGES UP UNTIL THE FIRST DAY OF CLASS

BIOL 3411: *Applied Sequence Analysis Protocols* | Summer 1 | MTWR 3:20 - 5:00 pm | Instructor: Jamie Henzy
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. This course introduces students to concepts and skills needed to productively "interrogate" genomic data and move toward participating in this exciting and rapidly evolving field. Working with various online tools, databases, Unix commands, and basic R packages to analyze and visualize genomic data, students will learn good data practices while building a portfolio of their work on GitHub, if desired. While knowledge of coding is not a prerequisite, a fearless attitude toward computers is an asset.

Prerequisites: BIOL 2301 (Currently the course requires "instructor's permission".

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
- 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.

Grading: Problem sets and quizzes will accompany each unit. One or two group mini-projects will involve completing a list of "deliverables". The course grade will be based on the student's overall performance with respect to completing tasks satisfactorily and participating in class.

Material: The Canvas site will be used only as a document repository, where you can access the Syllabus and Deliverables lists. All other material will be available online. Instead of submitting assignments to Canvas, you'll post them in your Discovery Cluster directory or on your GitHub site.

I. Genomes and their contents
Concepts:

- Survey of the genomes of model organisms
- Genomic databases and sequence files
- Best practices in computational work

Toolbox:

- Package managers, installation, and troubleshooting
- Basic sequence analysis tools: reverse complement; ExPASy translate; file formats and conversions
- UCSC Genome Browser
- Getting started on the Discovery Cluster and Unix commands

II. Variation
Concepts:

- Basics of sequencing, assembly, and annotation
- Types of sequence variation
- Indexing, aligning reads, and calling variants

Toolbox:

- Unix commands continued
- Accessing and downloading genomic datasets
- Bowtie2 and STAR

III. Expression

Concepts:

- Transcripts as proxies for expression
- Normalization
- Heatmaps and scatterplots

Toolbox:

- Wrangling data in RStudio
- Performing DESeq2 in R
- Generating heatmaps and scatterplots in R

IV. *Comparative genomics

Concepts:

- Homolog searches and pairwise alignments
- Inferring divergence times from sequence comparison
- Phylogenetic trees

Toolbox:

- BLASTing on the cluster
- Clustal omega and GBlocks
- R packages for phylogenomics

*Students have the option of choosing **Protein structure analysis** for this module, in which case they will focus on learning to use the program Alphafold.