BIOLOGY 4559: Computational Evolutionary Biology
SYLLABUS AND SCHEDULE
Fall 2023

Instructor: Alan Bergland
Contact: aob2x@virginia.edu

TA: Connor Murray
Contact: csm6hg@virginia.edu

Class time & location: Thursday 9AM-12PM
Class location: Gilmer XYZ (in-person)
Professor Office hours:

TA office hours:

COURSE DESCRIPTION AND OBJECTIVES:

The evolutionary history of a population can be studied by examining patterns of genetic variation among individuals from a species. Using information about genetic variation among individuals, one can infer demographic events such as migration or population size changes, as well as identify molecular signatures and genetic targets of recent adaptation. Characterizing these basis evolutionary features of a population or species is important for making accurate inferences of the genetic basis of disease, predicting future evolutionary change, and explaining levels of genetic diversity. These days, evolutionary biologists utilize genome data from samples collected in the wild or in laboratory experiments to infer this evolutionary history.

In this lab course, you will learn how to utilize genomic data to make evolutionary inferences. Together, we will learn…

1) … fundamentals of population genetics.
2) … how to conduct bio-informatic research on high-performance computers.
3) … how to take raw sequence data and turn it into biological insight.
4) … how to present the results of your research.

FUNDAMENTALS

In this course, we will focus on understanding the ways that variation in genomes is measured, and how levels and patterns of genetic variation among individuals informs us about evolutionary history. The evolutionary history of a population or species can be studied across different time scales – from speciation events deep in the past to contemporary evolution such as response to climate change or artificial selection. This course will focus on contemporary evolution. Therefore, our discussion about fundamentals will largely focus on ways to interpret patterns of *allele frequencies* (AKA mutation frequencies, SNP frequencies) across the genome. We will delve into the basic theoretical models of allele frequency dynamics, and gain intuition about how allele frequencies can reflect the evolutionary history of a species.

CORE SKILLS

Research in this course relies on computational analysis of large genomic datasets. Most of this computational work will be conducted on UVA's research mainframe, Rivanna. In order to conduct this research, we will learn:

1) Navigating Rivanna: Basics of unix command-line, basics of text parsing, job management 'slurm', version control on git-hub.
2) Mapping pipelines: from raw data to workable data.
3) Data analysis in R-studio: data-structures, functions, graphs, statistical tests using allele frequencies.

THE RESEARCH PROJECT

Drosophila melanogaster, the common vinegar fly, has been a workhorse for genetics since the early 20th century. Very likely, you learned about *D. melanogaster's* central role in the important discoveries made about mechanisms of genetic inheritance, the nature of a gene, and the organization of the chromosome. Drosophila also became a model system for other aspects of genetic research, such as quantitative genetics (the science that grew out of animal and plant breeding) and population genetics (the science of allele frequencies). From the quantitative genetic perspective, *D. melanogaster* were useful because of their fast generation time (~14 days) and because relatively large populations (n~500-5000) could be maintained in the lab and subjected to all sorts of selective pressures. Hundreds of artificial selection experiments have been conducted on flies, and the general conclusion from these studies is that virtually any trait, behavior, phenotype can be selected for increased or decreased value. The ease by which artificial selection can shift phenotypes in flies implies that there is a large amount of genetic variation for traits in flies, and that are specific mutations that cause increased or decreased trait values. The development of inexpensive next-generation sequencing technologies allows us to identify those genetic mutations (in principle) and reach the holy-grail of genetics: understanding the genotype to phenotype map.

The primary research project you will undertake in this class is the re-analysis of an *Evolve-and-Resequence* experiment conducted on Drosophila. E&R experiments are artificial selection followed by high-throughput sequencing of evolved and control lines at various points in time during the selection experiment. You will learn to map the raw data and then how to interpret the output.

Although the flies you are studying lived in the lab, the conclusions you reach can be put into a global context. For instance, fly populations around the world are known to genetically vary in phenotype. For instance, flies from Maine are larger than flies from Florida, when reared in a common environment. Therefore, it is reasonable to hypothesize that an allele that is associated with larger body size from an E&R study might be at higher frequency in Maine than Florida. We will test this hypothesis by incorporating your E&R data into an existing repository of allele frequencies from flies sampled around the world for nearly a decade (https://dest.bio).

PARTICIPATION & GROUP WORK

Because of the nature of this course, participation is essential. Participation will take place in many ways, including solo-work, group-work, in-class discussions, engagement outside of class.

A major component of participation is group-work. For your final project, you will be working in a small group (groups of 2, mostly) that forms early in the semester. In addition, we will have opportunities to work with others in the class for in-class groups.

Everyone will be entering this course with a different computational skill set, and different experiences with research. Everyone is expected to contribute to group work in an equitable manner, and to be responsible for contributing to all aspects of the research project.

You are expected to attend class in-person (no Zoom will be conducted). If you need to miss a class, please contact Alan ahead of time.

EVALUATION

Evaluation will be based on completion of small assignments, lab-reports, as well as completion of the final project. Final grading scheme TBD. Rubrics will be provided for the final project evaluation.

PREREQUISITES:

Genetics & Molecular Biology - BIOL 3010
Evolution & Ecology - BIOL 3020
Stats - Stats 2120/2020

READINGS

All assigned readings will be available as PDFs on Canvas. You will also be reading the primary literature that you find yourself.

COMPUTATIONAL WORK ENVIRONMENT

Mainframe computers, such as Rivanna, can be accessed in different ways using different types of programs. The simplest way is some sort of terminal application ("Terminal" on Macs; PUTTY or MobaXTerm on Windows). The more advanced way is to use software designed for writing scripts/code that also has access to the terminal (e.g., Atom, Visual Studio, etc, etc). Rstudio has very nice capabilities for the later category, and will be using this program in many ways. To facilitate using Rivanna, we will take advantage of the excellent "OnDemand" tool. This service lets you use terminal programs and Rstudio via web-browser with direct access to data on Rivanna. This is great because it means that you do not need to install anything. If you are already using another script/code editor or terminal application, I ask that you please use the OnDemand version that is describes in the activities.

COURSE SCHEDULE

Class number	Class date	Topics	Presentation
1	24-Aug	1) Intoductions 2) pop-gen basics 3) Rivanna, bash, Github	1) Presentation: The unit and measure of genetic diversty
2	31-Aug	1) Artificial selection concepts 2) NGS, FASTQ, SRA 3) slurm & modules	1) Presentation: Additive variation, P&G response to selection 2) paper discussion 2) Presentation: NGS sequence data
3	7-Sep	1) group discussions about papers 2) mapping raw data	1) student presentations 2) mapping pipelines
4	14-Sep	1) post-mapping quality control 2) variant calling	1) group reports on FASTQ quality 2) presentation: estimating allele frequencies & precision
5	21-Sep	1) VCFs, GDS, looking at the data 2) Allele frequency spectrum, pop-gen stats	1) Presentation: VCF data structures, fields, annotation 2) bash: grep, awk 3) R: GDS, SFS
6	28-Sep	1) Data filtering 2) Imputation	
7	5-Oct	Where in the world did these samples come from? PCA	
8	12-Oct	Inversions	
9	19-Oct	Field day	Field trip to Carter's Mountain Orchard to collect flies
10	26-Oct	What are the targets of selection? Fst	
11	2-Nov	Enrichment tests (NS/Syn)	
12	9-Nov	Local adaptation & geographic concordance	
13	16-Nov	Catch-up & in-class workday	Poster presenation guidelines
	23-Nov	No class - Thanksgiving	
14	30-Nov	Poster presentations	