Joshua Hallas

Data Science for Biology I: Unix and Python

Proposed project: Splitting Fastqs

I will create a python script to split fastq files of double digest restriction-site associated DNA sequences (ddRADseq). In the Parchman Lab, we have established scripts for processing ddRAD data, but all are in Perl. This project will strength my ability to write in Python, which is vital to processing genomic data files.

My project is focused on characterizing population structure in garter snakes and newts across western North America. I have ddRAD data for >700 individuals and the only way to process this data is through the terminal using Unix and a combination of custom scripts. This script will take a cleaned Illumina data fastq file and generate individual fastq files for each specimen sequenced. This script will read in 2 files (barcode ID; clean fastq), create a directory, create individual specimen fastqs, pull out the appropriate sequences, and place new files in the new directory. These fastq files must be in a very specific file structure to be used in genetic filtering pipelines.