BIOL 5153, Practical Programming for Biologists
Assignment #1
Due Wednesday, 05 February 2019 by 9:55 AM

TURN IN A SINGLE BASH SCRIPT WITH ALL THE COMMANDS. DO NOT INCLUDE OUTPUT UNLESS DIRECTED TO DO SO. FOLLOW THE INSTRUCTIONS. USE GOOGLE. SEND ME A SINGLE TEXT FILE BY EMAIL.

1) We can't cover every Unix command in class. You're in your home directory, and you know there are amino acid and nucleotide sequences for the *nad* genes somewhere in $HOME. Which Unix command could you use to search for files with 'nad' in the file name? Note: I am not looking for 'ls *'. Put the command in a Bash script with the comment 'assn01-1' above it.

2) We can't cover every Unix command in class. Using Unix, show all the processes running on your computer. What command did you use? What percent of your CPU is being used by that command? Copy and paste the line of output showing how much total memory you have on your machine, how much is currently being used, and how much is currently available. Show me how you got these numbers. Put the command in a Bash script with the comment 'assn01-2' above it. Put the answers to the questions as comments below the command.

3) Extract all the lines describing chloroplast IR features (annotated as 'misc_feature') in watermelon.gff, sort them by fragment size in descending order, and write them to a file called 'IR_regions.gff'. Put the command or commands you used to answer this in the Bash script with the comment 'assn01-3' above it.

   Refer to the following for information about GFF format:
   http://gmod.org/wiki/GFF3

4) The 'watermelon.gff' file contains all of the features in the watermelon mitochondrial genome. Plant mitochondrial genomes are interesting in that they regularly absorb DNA from the chloroplast genome. If chloroplast DNA is raining it at random, one might predict that the number of fragments from the chloroplast's large inverted repeat region (the 'IR') will outnumber sequences from outside the IR. Is this the case? Do more chloroplast fragments come from inside the or outside the IR? You'll need to look at the file and the annotations to figure out how to identify IR vs. non-IR sequences. Put the command or commands you used to answer this in the Bash script with the comment 'assn01-4' above it. Put the answers to the questions as comments below the command.

5) Are there any watermelon genes that HAVE a BamHI site and LACK an EcoRI site? Print them to the screen, making sure you print both the sequence and the header line. Put the command or commands you used to answer this in the Bash script with the comment 'assn01-5' above it.

6) The dataset in shaver_et_al.csv contains 1708 records. Extract records 500–1000. This is possible with a single command (with pipes) using commands covered in class. Look through all the

commands we've learned so far and be creative. Use Google. Divide and conquer. Put the command you used to answer this in the Bash script with the comment 'assn01-6' above it.

7) Use the sort command to sort fruit.txt. Sort first in descending order by fruit type (column 2), then in ascending order within fruit type by state. You should get this result (after piping through 'column'):

```
5  pear   FL
3  pear   IL
2  pear   OH
4  apple  FL
6  apple  MI
1  apple  OH
```

Put the command or commands you used to answer this in the Bash script with the comment 'assn01-7' above it.