University of Pennsylvania
BIOL4536 Fall 2023

HW#4

(UNIX)

Assigned September 20th
Due September 27th, 3:30pm

You should have received an email with information about your UNIX account, that you access it through a web browser.

In this HW you will be asked to construct some unix commands to perform certain tasks. There usually is no one right way to do things. But points can be deducted for doing them in an unreasonably long or inefficient way.

In your home directory, create a subdirectory called “hw4”.

Move into that directory. Execute `pwd` to make sure you are in the right place.

Copy the file `biol4536_course_files.tar.gz` from the `/data` directory to the `hw4` dir. This is a `.tar` archive that has been gzipped. The class slides explain how to unzip and unpack such a file. Do that now.

Do an `ls -ltr` to make sure the files are there.

The file `reads.fa` has $5,000$ sequence reads in fastA format.

(1) (1/2 pt) Use `grep` to count the number of sequences that have five consecutive A’s and five consecutive T’s in them. Show your grep command and the answer. Would your command work if there were read IDs with five A’s and five T’s in them? Even though there aren’t, make sure your command would still work if there were.

(2) (1/2 pt) Use `grep` to count the number of sequences that have consecutive five A’s and five consecutive T’s in them, but the A’s occur before the T’s. Show your grep command and the answer.

(3) (1/2 pt) Use `grep` to count the number of sequences that have five consecutive A’s and five consecutive T’s in them, but the T’s occur before the A’s. Show your grep command and the answer.

(4) (1 pt) The answer for #2 is the number of sequences with five consecutive A’s upstream of five consecutive T’s, and the answer to #3 is the number of sequences with five consecutive A’s downstream of five consecutive T’s. While #1 is the number of sequences with five consecutive A’s upstream *or downstream* of five consecutive T’s. Why isn’t the answer to #2 equal to the answer to #2 plus the answer to #3?

Advance to next page

Do the following to prepare for the next problem. We will make a simple "shell script" (as you've seen in the UNIX slides last topic).

Create a sub-directory of `hw4` called `temp`. Create a file called `makefiles.sh` with the following code in it:

```
for i in {1..100}
do
    echo "This is the contents of file $i" > temp/file_$i.txt
done
```

Now execute the command '`source makefiles.sh`' from the `hw4` directory.
The directory `temp` should now have 100 files in it. Do an `ls -l temp` to make sure.

(5) (2 pts) Write a python script to execute the unix commands to delete all files `file_N.txt` in the temp directory where *N* is divisible by 3 but not by 5.

Copy the python code to your write-up, or include as an attachment. It should not be too long.

(6) (1/2 pt) Use a Unix command to count how many rows the file `arraydata1.txt` has. Give the answer and the command.

(7) (1/2 pt) You can use the following Unix command to count the number of columns in a tab-delimited file. Note, if you copy/paste this you might need to retype the quotes and some other characters, since they may vary.

```
> head -1 <filename> | awk '{print gsub(/\t/,"")'}
```

This counts the number of tabs in the first row of the file. So add one to get the number of columns. Run this on `arraydata1.txt` to find out how many columns there are. Give your answer and the command.

(8) (1/2 pt) Use a Unix command to count how many rows have their 15th column equal to 2. Give the answer and the command.

(9) (2 pts) Write a python program to run the unix command from the previous exercise for each column, and report the column with the greatest number of 2's. Copy your code into your write-up, or include as an attachment. *Note: we know how many columns there are from problem 7, you can hard code this into your script.*

(10) (1 pt) Column three of the file `UCSC_mouse_knowngene_id_mapping` has the so-called "refseq" gene identifier. Use `cut` and `sort` to figure out how many uniquely different ones there are in that column. Report the answer and your unix command. *Note that the first row is the header row.*

(11) (1 pt) The file `reads.fa` has high-throughput sequencing reads of RNA. The ones with ID's that end in an "`a`" are the so-called "forward" reads and the ones that end in a "`b`" are the "reverse". We'll learn more about what that means later. But for now, all you need to know is `a`'s are foward and `b`'s are reverse. Suppose we are looking for polyadenylated genes, so we want to find polyA tails. Suppose to this end we want to find all sequences with 20 `A`'s in a row, or more. Out of the 5,000 reads in the file, how many foward reads are there with 20 `A`'s and how many reverses? Do this with `grep` commands, report the answer and the command.