Homework Assignment

This assignment and exercises at the bioinformatics bootcamp will require the use of command line interfaces (CLI) with a Unix shell. On UNIX-like operatings systems such as macOS and Linux operating systems, this is as easy as launching the Terminal application. If you use a Windows operating system, you can use a terminal emulator (e.g. PuTTY) to connect remotely to a Linux machine, or install Linux subsystem for Windows 10 (see Windows user guide to accessing the Alabama Super Computer).

Although certainly not required, we recommend that you purchase a copy of Practical Computing for Biologists by Steven Haddock and Casey Dunn. It is a well written book that introduces a number of Unix shell topics and will serve as an excellent reference for scientific computing in general.

- Alternatively, there are numerous, excellent Unix learning resources online. A few very good, brief introduction to Unix commands and the shell environment are listed below:
    - https://kb.iu.edu/d/afsk
    - https://www.codecademy.com/learn/learnthecommandline
    - http://linuxcommand.org/
- You'll likely be able to find many more out there (Google is your friend!) in addition to resources on Linux mainstays such as `grep`, `awk`, `sed` and shell scripting, all of which we will be using during the Bootcamp.

Tip: many UNIX commands have incredibly complicated options, don't let that stop you from learning and understanding their simplest use case.

- If available, read through Chapters 4 and 5 of the Haddock and Dunn textbook. Following that, go through the online tutorial at this link: http://www.ee.surrey.ac.uk/Teaching/Unix/unix0.html. Complete sections 1-6 and 8. Skip over the "quota" command at the start of section 6 as well as sections 8.5 and 8.6 as they are not that relevant at this point.
- As a supplemental activity, please read over this page: www.auburn.edu/bioinformatics/documents/file_transfers.pdf (additionally, Chapter 20 of Haddock and Dunn)

Part A

1. Before starting the assignment, clear the contents of your command history by typing the following command in terminal: `history -c` Create a directory titled with your first and last name: `<NAME>_Bootcamp_Assignment` in the Desktop folder of your user account. Navigate into this directory and work on the project from there.
2. Direct the output of the following command: `ifconfig` to a file named in the following fashion: `<NAME>_Bootcamp_Assignment.ifconfig` (for example:Smith_Bootcamp_Assignment.ifconfig).
3. Do the same as above for the following commands: `ps aux`, `df .`, `du -sh *` and `whoami`. For `du -sh *`, you will want to: 1) change back to your home directory to get the total sizes of the various folders in your filesystem but 2) write the output to the directory that you are working in for your project. Be sure to name each file as above, changing the extension of the file to match the name of the command that generated that output e.g. Smith_Bootcamp_Assignment.ps, Smith_Bootcamp_Assignment.df.
4. Combine the five files from above into a single file using cat and wildcards. Name the combined file: `<NAME>_Bootcamp_Assignment.system`
5. Create a new directory in `<NAME>_Bootcamp_Assignment` directory named `NAME_Sysinfo` (for example, Smith_Sysinfo). Move the five smaller files and the combined .system file into this new directory using `mv` and wildcards.
6. Send the output of history to a file named: `<NAME>_partA.history` (for example: Smith_partA.history). Place this file in the same directory as the six files from above.

Part B

1. Before starting this section, clear the contents of your command history again by running the command: `history -c`
2. Create a directory in `<NAME>_Bootcamp_Assignment` called `<NAME>_GenbankData` (for example, Smith_GenbankData). Move into this directory to start Part B.
3. Download the file `DinoPro.fasta` from the "student" account at the address the-santos-lab.dynu.net using `scp`. Specifically, the `DinoPro.fasta` file is located in the `homework` directory of the “student” account’s home directory. For a refresher on scp, see:auburn.edu/bioinformatics/documents/file_transfers.pdf
4. Use `grep` to extract lines with the pattern `>gi` in the file `DinoPro.fasta` and direct this output to a file titled `AllEntries.output`. Examine the contents of this file with the utility less.
5. Use `grep` to search for the term `Symbiodinium` in the file `AllEntries.output` and send this output to a file titled `SymEntries.output`. Also examine the contents of this file with less.
6. Use `grep` to exclude entries with the term `Symbiodinium` in the `AllEntries.output` file and send this output to a file titled `NonSymEntries.output`. Also examine the contents of this file with `less`.
7. Use `wc` to obtain the number of lines in `AllEntries.output`, `SymEntries.output` and `NonSymEntries.output`. Note if the line numbers in `SymEntries.output` +

`NonSymEntries.output` = `AllEntries.output`. Send the output from the three separate `wc` commands to a file titled `Values_DinoPro_Fasta.output` in the following order: `SymEntries.output`, `NonSymEntries.output`, `AllEntries.output`

8. In a text editor (e.g. Atom, Visual Studio Code, Vim), write 8-10 sentences describing what you did above and the reasoning behind what was done. Start this description with what type of information was contained in the `DinoPro.fasta` file, followed by what information was extracted and parsed among the various result files. Add this text file to the `<NAME>_GenbankData` directory (if it's not already there) and name it `<NAME>_Methodology.txt`.
9. When you are done, send the output of history to a file named in the following fashion: `<NAME>_partB.history` (for example, Smith_partB.history). Save this file in the same directory as the files you just worked with i.e., `<NAME>_GenbankData`.

Part C

1. Archive the contents of the entire `<NAME>_Bootcamp_Assignment` folder for submission by using tar. While in this directory, execute the following: `tar -czf <NAME>_Bootcamp_Assignment.tar <NAME>_SysInfo <NAME>_GenbankData` This will create a compressed file of the directories containing your work (for example, Smith_Bootcamp_Assignment.tar).
2. Use `scp` to send your `<NAME>_Bootcamp_Assignment.tar` file to the "student" account at the address the-santos-lab.dynu.net and place it in the `completed_assignments` directory. Any concern your last name might be the same as another participant? Tack on your initials to your last name in the command you use above during the homework submission.

Part D

For the sessions relating to Data Visualization, we will be working with software installed directly on your own computer. The three software are all free to you and compatible with all platforms (Mac, Windows, and Linux). In order to participate in the exercises, you will need to download and install each software to your own computer using the instructions below. Please open each once installed to make sure you do not encounter any errors as this will delay your ability to follow along.

1. The base language R can be downloaded at https://cran.rstudio.com/. Choose the appropriate download for your OS.
2. Once you've downloaded and installed R, download RStudio: https://www.rstudio.com/products/rstudio/download/. Although R can be run and loaded from the command line, RStudio offers a graphic user interface for working in R. It includes a set of integrated tools designed to help you be more productive with R, such as a console, syntax-highlighting editor that supports direct code execution, tools for plotting, history, debugging and workspace management.

3. Finally, you will download and install IGV, available from the Broad Institute (which also host GATK). IGV stands for Integrative Genome Viewer and allows you to upload a variety of genome file formats to view on your desktop machine. It runs on java, so you may need to update your version of java to run it. Link to download (make sure to scroll to the appropriate distribution based on your OS): https://software.broadinstitute.org/software/igv/download. Linux users will download the binary distribution, which can be launched directly from terminal using java. Make sure after you install it that it loads properly to avoid delays during Friday's session.