**Genomics Week Practical 1:**

**Bioinformatics, Data, and Databases**

Silwood Park, November 2017

Dr. Jason Hodgson

**Introduction**

Most people doing genomics research and bioinformatics use a Unix-based computer operating system.

In this practical you will:

* Learn to interact with the computer using a terminal shell.
* Learn basic Unix commands.
* Learn to manipulate text files
* Learn to manipulate directories
* Learn to execute programs
* Learn the most important genomics databases

**UNIX**

For these practicals you will need access to a computer that runs Unix. If you are using Mac OS X or some flavor of Linux, congratulations, your computer is Unix based. If you are using a Windows computer you will need to either install a Unix emulator or install a Linux operating system in a virtual machine. Also, the Linux operating system Ubuntu, is installed in virtual machines on all of the computers in the Hamilton teaching laboratory. For our purposes, the important thing is that we can interact with the operating system using a command line terminal shell.

Although it may seem that working on the command line is an old fashioned way to interface with a computer, when you become proficient it is actually much faster than clicking through windows. Also, by typing your commands into a script, you have a record of every analysis or action taken. This is imperative for performing reproducible science!

**Mac Users:**  You will need a good text editor. Download and install “Text Wrangler” for free from the apple store. This is an excellent text editor for writing programs and viewing text files.

**Windows people try:**

<https://stefanoallesina.github.io/BSD-QBio3/>(scroll down to "UNIX Emulator")

or

<https://www.cygwin.com/>

**Warning for those using the Hamilton computers:** Imperial IT has set up the Hamilton virtual machines to wipe clean every night. It is imperative that you move any files that you want to keep from the virtual machine to the Windows environment. This can be done simply by dragging and dropping from the Ubuntu virtual machine to any normal folder on the computer.

**Useful Unix commands:**

**Command Action**

pwd print working directory

ls lists file in working directory

cd change directory

rm remove a file or directory (careful! This is permanent!)

cp copy a file

mv move a file or directory

more read a file (type q to quit out of more)

head read the first ten lines of a file (use –n to specify # of lines)

tail read the last ten lines of a file (use –n to specify # of lines)

wc count the number of words, lines, and characters in a file

man read the manual for a specific unix command (e.g. man cd)

unzip unzip a .zip file

zip zip a directory into a .zip file

history Gives the chronological history of commands entered

**Useful Unix tutorials online:**

Dr. Christina Bergey has made some very useful beginning Unix youtube tutorials. Find these by searching youtube for “Christina Bergey”, or use the following link to begin:

<https://www.youtube.com/watch?v=Ms5sNYyejEw>

**Exercise 1, Unix tutorial:**

1. Direct your web browser to: <http://www.ee.surrey.ac.uk/Teaching/Unix/unix1.html>
2. Launch the “terminal” application on your Unix machine. (**Mac users** can use spotlight to search for “terminal”; **Ubuntu users** can click the uppermost icon in the toolbar and search “terminal”, **Virtual machine users**  will need to click the Ubuntu icon on the Windows desktop, click the play button to start the virtual machine, and then sign in using your university account)
3. Complete the **Unix Tutorial One** through **Unix Tutorial Five** from the website.

**Useful tips for speedy work:**

1. **Rapidly complete typing tasks using the tab key**. Begin typing the first few characters of one of the files in your directory then hit the tab key. The terminal will try to complete your typing for you. Try it!
2. **Scroll through your command history using the up and down arrows.** Try hitting the up arrow. The last command you typed should appear on the prompt. Hit the up arrow again and the command previous to that one will appear.
3. **Use wildcards to specify groups of files.** The \* symbol says to match groups of files or directories. For example

mv plink\* plinkdir/

moves all files in the directory with names beginning with plink to the directory called plinkdir.

**Exercise 2, Setting up your machine:**

Bioinformatics involves downloading and running numerous programs that are freely distributed by researchers for researchers. It is imperative that you can download and implement applications into your workflow.

An easy way to accomplish this is by setting up a local “bin”. In Unix a bin is a directory of binaries (compiled executable programs). You will need to create a bin, and instruct your terminal that it should always look for applications in your bin.

1. On the terminal type:

echo $PATH

$PATH is a Unix environment variable. These are the directories that your computer looks in when you invoke some command in the terminal. For example when you type “ls” the terminal looks into each of the directories listed in $PATH for an executable program called “ls”.

We first need to create a directory called “bin” in your home directory:

2. First direct your terminal to your home directory. Type:

cd

3. Make a directory called “bin”:

mkdir bin

4. Check that bin exists:

ls

5. Now we need to add this directory to the $PATH variable. The easiest way to do this is by creating or editing the hidden file called “.bash\_profile”. In Unix there are many hidden files that exist that the operating system reads but the user generally does not (and often should not!). Hidden Unix files begin with a fullstop. Check if you have a “.bash\_profile” file in your home directory. Type:

ls -a

The –a flag tells the “ls” program to list all files in the directory including hidden files. Do you see .bash\_profile?

6. If you do not see .bash\_profile you will need to create one. You can do this simply using the terminal. Type:

echo PATH=$PATH:~/bin >.bash\_profile

echo export PATH >>.bash\_profile

7. If you do have a .bash\_profile file you will need to edit it. Open .bash\_profile using your system’s text editor (“text” in Ubuntu or Text Wrangler in Mac). The first line of the file will be a list of the directories in $PATH each separated by a colon. Add your directory to the end of this line by adding the following text:

:~/bin

Once you have done so, save the file and close it.

8. Close your terminal application and relaunch it. Again type:

echo $PATH

You should now see the bin you created added to the end of the $PATH environment variable. Now, any program you place into your bin will be available to the computer no matter what working directory you are in. This is where you should keep Unix programs that you download for your workflow.

**Exercise 3, adding programs to your bin.**

We will be using a few programs throughout the week. You will need to have these available in your bin. Download the appropriate versions of the following:

ADMIXTURE: <https://www.genetics.ucla.edu/software/admixture/download.html>

PLINK: <https://www.cog-genomics.org/plink2>

You should now be comfortable enough with Unix to unpack each download using the unzip command, and mv the executable to your bin, e.g.

unzip plink\_mac.zip

mv plink ~/bin/plink

1. Make sure you have executable permissions for each of your programs:

chmod +x ~/bin/plink

chmod +x ~/bin/admixture

2. Once you have added both PLINK and ADMIXTURE to your bin and made them executable, check if you can execute the programs. Type:

plink

If everything is set up correctly you should see something like:

PLINK v1.90b4.10 64-bit (3 Nov 2017) www.cog-genomics.org/plink/1.9/

(C) 2005-2017 Shaun Purcell, Christopher Chang GNU General Public License v3

If things are not set up correctly you might see:

-bash: plink: command not found

If you are getting this error your computer cannot see the program. Check that you have named your program correctly and that you can see it in your bin.

3. Now check that ADMIXTURE is running correctly:

admixture

You should see something like:

\*\*\*\* ADMIXTURE Version 1.21 \*\*\*\*\*

\*\*\*\* Copyright 2008-2011 \*\*\*\*\*

\*\*\*\* David Alexander, John Novembre, Ken Lange \*\*\*\*\*

\*\*\*\* Please cite our paper! \*\*\*\*\*

\*\*\*\* Information at www.genetics.ucla.edu/software/admixture \*\*\*\*\*

Usage: admixture <input file> <K>

See manual for more advanced usage.

If you see this, congratulations, your computer is ready to go!

**Data and Databases:**

**NCBI GenBank and BLAST**

NCBI GenBank is the primary database and repository for DNA sequence data. Most journals require DNA sequence data to be deposited to GenBank as a condition of publication. GenBank often provides the comparative data necessary to understand your DNA sequences of interest.

BLAST stands for Basic Local Alignment Search Tool. BLAST is one of the most important tools available for bioinformatics. The two primary papers describing BLAST have each been cited more than 50k times! BLAST is used to search for similar nucleotide sequences to a query sequence. The idea is to identify homologous sequences.

**Exercise 4**

1. navigate to: <http://digitalworldbiology.com/dwb/BLAST>

2. Perform the BLAST tutorial. This will teach you how to perform and interpret a BLAST search

3. Perform BLAST searches on the set of unknown sequences: http://www.digitalworldbiology.com/BLAST/62000sequences.html

**Question: What organisms and genes are the unknown sequences from?**

**DRYAD**

Dryad is a repository for any type of scientific data. It is becoming increasingly common for genomic data to be deposited in Dryad. This is particularly true for genomic data that does not easily fit into GenBank (e.g. SNP data rather than sequence data).

**Exercise 5**

First, navigate to the Dryad website: <http://datadryad.org/>

1. Pick three taxa that interest you and perform a search by entering the common names. Now try the scientific names.

**Question: How many data entries are there for each taxa? What types of data are available?**

2. Pick three geographic locations.

**Question: How many data entries are there for each location? What types of data are available?**

3. Now pick a few researchers and answer the same questions (e.g. Hodgson JA, Savolainen V, etc.).

**Note that Dryad often contains additional data to the genomics, such as phenotype data. The really interesting biology couples genomic data with other lines of inquiry!**

**Exercise 6**

Construct a database of primate mitochondrial genomes for phylogenetic analysis. Be sure to save the files you create to your computer (copy it outside the virtual machine). We will use these files in a later practical.

1. Navigate to the NCBI Nucleotide web portal: <http://www.ncbi.nlm.nih.gov/nuccore>
2. Choose an mtDNA genome accession number from Hodgson et al. 2009 Table 1 and search for the accession on NCBI Nucleotide: <http://www.pnas.org/content/106/14/5534.full>
3. BLAST this entry (note you can select “run BLAST” from the right hand side of the page).
4. Choose as many similar sequences as possible, however avoid choosing multiple individuals from the same species. Check the box for all appropriate sequences.
5. Export the sequences to a Fasta format text file by clicking the “Download” dropdown menu, and choosing “FASTA complete sequence”.
6. Also export an aligned Fasta file by choosing “FASTA aligned sequences”.
7. Using the terminal move these rename these files “primate\_raw.fasta”, and “primate\_aligned.fasta”, and move them from the Downloads directory to your desktop.
8. Drag the files out of the virtual desktop onto the Windows desktop for safekeeping.

**Other databases to be aware of:**

**HapMap.** The International HapMap project was designed to characterise human genetic variation worldwide. High density (2.4 million SNPs) SNP data exists for 13 global populations. The SNPs were characterized primarily through SNP chips.

**1000 Genomes.** The 1000 Genomes project is another attempt to characterize global genetic variation in humans. These data are produced through whole genome sequencing, and do not have the same biases as the HapMap data.