UNIX & File Format Exercises Solutions

1. If you have not already completed the exercises on the PFB2017 website, please do so first. If you have, proceed to the next item.

   The solutions to these may be found at:
   https://github.com/srobb1/pfb2017/blob/master/problemsets/answers/JVB_Unix_01_problemset.md
   Ignore the solutions to #1 and #2; they are not specific to your machine or our course.

2. Log into the AWS server

   Instructions are here:
   https://github.com/bredeson/HandsOnDArT#logging-into-the-aws-server

3. Make a directory called “exercises” and navigate into it

   From your home directory:

   ```
   mkdir exercises
   cd exercises
   ```

UNIX & File Format Exercises Solutions

4. Copy all files starting with “data” from the /files directory into your exercises directory.

```
cp /files/data* .
```

5. Examine the contents of data1 (hint: use less).

```
less data1
```

   a. What format of file is it? Rename the file with the appropriate file extension.

   FASTQ

```
mv data1 data1.fastq
```

   b. Is the contents of the file single-end or paired-end?

   paired-end

UNIX & File Format Exercises Solutions

6. Examine the contents of data2.

```
less data2
od -c data2
```

   a. What format of file is it? Rename the file with the appropriate file extension.

      FASTA

      ```
      mv data2 data2.fasta
      ```

   b. What is incorrect about its formatting?

      Extra line break, between Sequence ID and Sequence description, putting the description on its own line. The sequence description should be on the same line as the sequence ID, with the sequence line immediately following on the next line.

UNIX & File Format Exercises Solutions

7. Examine the contents of data3 (hint: use man to look at the -S option).

```
less -S data3
```

   a. What format of file is it? Rename the file with the appropriate file extension.

      VCF

```
mv data3 data3.vcf
```

   b. How many loci are there? How many SNPs? How many Indels?

```
grep -v -e# data3.vcf | wc -l
```

      41 loci
      All are SNPs
      No Indels

UNIX & File Format Exercises Solutions

8. Examine the contents of data4.

```
less -S data4
```

   a. What format of file is it?

      CSV

   b. What is incorrect about this file (hint: see od)?

      ```
      od -c data4
      ```

      Carriage returns (\r) instead of line feeds (\n)

   c. From which operating system might it have come from?

      e.g. Legacy Mac OS

   d. Correct the file using UNIX commands (hint: see tr)

      ```
      cat data4 | tr '\r' '\n' > data4.fixed
      ```

UNIX & File Format Exercises Solutions

9. Examine the contents of data5.

```
od -c data5
od -ta data5
```

   a. The file is tab-delimited, but what is incorrect about this file?
   The first line has spaces instead of a tab

   b. Correct the file using UNIX commands.

```
cat data5 | tr -s ' ' '\t' > data5.fixed
```