2015 Český Krumlov Genomics Workshop - UNIX Homework 2

This homework was built from a real task we needed to complete as part of an experimental analysis we have been working on in the Cresko Lab for the past few weeks. The basic problem is this: we have a vast number of SNPs generated from a large RADseq[1] analysis and we want to load this data into a program called GenePop. Unfortunately, our data set is too large for this program which was written before data sets of this size became common. So, our goal is to select a random subset of our data to load into GenePop. We don't know exactly how much data will fit into the program so we want to make three different sized, randomly-selected subsets.

1. Copy the following file into your working directory:

   ```
   ~/workshop_data/unix/batch_2.genepop.gz
   ```

   This file contains the genotypes from 40,928 loci in nine populations of Stickleback fish. Line one contains a definition of the program versions used and the date it was generated. Line two contains all the loci in the dataset, separated by commas. Each locus is listed multiple times, once for each SNP at the RAD locus. Here is an example of a RAD locus, with two detected SNPs listed in **RED**:

   Locus 138586:

```
TGCAGGAGATTAAAGGAAAACCACACGTGGAGATGAAAGGATACGAGCTCGGCAAGTGACACAG
```

   This locus would appear in line two as: `138586_15,138586_40`. (So, its the locus ID, followed by a "_" character, followed by the column where the SNP was found.)After line two, each individual in the dataset is listed, followed by their genotype at every locus in the dataset. These genotypes are encoded with two, two digit numbers:

   ```
   01 = A
   02 = C
   03 = G
   04 = T
   ```

   So, `0103` represents an A/G SNP at that position in that individual.

2. Your goal is to extract line two from the file, *translate* the commas into newlines ("\n"). This will give you a list of loci, one per line.
3. Next, you will want to extract the locus number off each ID, so given `138586_40, locus 138586` with a SNP at position 40, we only want to keep the locus part of the number. This can be done with by using a regular expression to match the pattern and then replacing the part we don't care about ("_40" in this case), with nothing.
4. We then have a duplicated list of loci that we want to make *unique* to get our final list of loci.

Steps 2-4 can be completed in a single command, with several pipes.

5. Finally, I want you to use a new command, `shuf (man shuf)`, to randomly order the list of loci and select 500 of them into a file. Then repeat again and this time select 1000 loci, and finally do it one more time and select 10,000 loci. These three data sets should be *captured* into three different files.
6. So, in the end, you should have three files, one with 500 lines, with each line containing a randomly-selected locus. Another files with 1000 lines and a different random set of loci, and then the third file with 10,000 random loci. The 10,000 locus file should have different loci then the 500 locus file (although one or two might be in common because they were randomly selected).

Given these lists of loci, I can separately generate new GenePop files containing the data just for those loci (you don't have to do this part).

[1] https://www.wiki.ed.ac.uk/display/RADSequencing/Home