UNIX Assignment

Your UNIX assignment will consist of two components:

1. Inspection of data files
2. Processing of these same data files to produce output files that are formatted for a subsequent analysis

Please document your work in a version-controlled repository using `git`. Your repository should include a `README.md` file in Markdown format that describes your workflow for both data inspection and processing and the files you will create as described below. When you are ready to turn in your workflow and files, please send a url linking to the GitHub or Bitbucket repository you have created via a Slack direct message to all three instructors.

Data Inspection

I will upload on both Slack and GitHub (look for the "UNIX_Assignment" folder) two data files for you to inspect and describe with the various UNIX programs we have learned about over the last few weeks. Use these programs to become familiar with the files and to describe their structure and their dimensions (file size, number of columns, number of lines, ect...)

The files are:

1. `fang_et_al_genotypes.txt` : a published SNP data set including maize, teosinte (i.e., wild maize), and Tripsacum (a close outgroup to the genus *Zea*) individuals

2. `snp_position.txt` : an additional data file that includes the SNP id (first column), chromosome location (third column), nucleotide location (fourth column) and other information for the SNPs genotyped in the `fang_et_al_genotypes.txt` file

Data Processing

Our goal is to process these files with UNIX tools in order to format them for a downstream analysis (we won't actually be carrying out this analysis, just preparing the input files...something you'll likely need to do time and again). During this process, we will need to `join` (hint, hint) these data sets so that we have both genotypes and positions in a series of input files. All our files will be formatted such that the first column is "SNP_ID", the second column is "Chromosome", the third column is "Position", and subsequent columns are genotype data from either maize or teosinte individuals.

For maize (Group = ZMMIL, ZMMLR, and ZMMMR in the third column of the `fang_et_al_genotypes.txt` file) we want 20 files in total:

- 10 files (1 for each chromosome) with SNPs ordered based on increasing position values and with missing data encoded by this symbol: ?
- 10 files (1 for each chromosome) with SNPs ordered based on decreasing position values and with missing data encoded by this symbol: -

For teosinte (Group = ZMPBA, ZMPIL, and ZMPJA in the third column of the `fang_et_al_genotypes.txt` file) we want 20 files in total:

- 10 files (1 for each chromosome) with SNPs ordered based on increasing position values and with missing data encoded by this symbol: ?
- 10 files (1 for each chromosome) with SNPs ordered based on decreasing position values and with missing data encoded by this symbol: -

A total of 40 files will therefore be produced.

A few notes and hints:

- In order to join these files, you will first need to transpose your genotype data so the columns become rows. This is beyond the scope of what we will learn in our UNIX section so I have uploaded an `awk` script (`transpose.awk`) that you can use to transpose these data. It can be run as:

```
$ awk -f transpose.awk fang_et_al_genotypes.txt > transposed_genotypes.txt
```

- You may wish to transpose the data after extracting the teosinte and maize data in which case your command line might looks something like this:

```
$ awk -f transpose.awk teosinte_genotypes.txt > transposed_teosinte_genotypes.txt
```

- `awk` will likely come in handy for separating SNP data based on chromosome and we will cover this in more detail on Thursday
- It might help to write out the entire workflow that will be necessary to produce the files described above before diving in with individual UNIX programs
- If you get stuck or confused, post to the "scripting_help" channel on Slack and we will provide hints that may be helpful for the whole class