**#R Assignment**

Your R assignment will consist of two components:

1. Replicating your UNIX assignment in R

2. Additional analysis and visualization

Do your R assignment in your "homework" repository. Explain all your steps and include your code in an "R Notebook" file. In addition, update the `README.md` file to indicate the extra work you have done.

**## Part I - Data Inspection**

Load the two data files you used for your UNIX assignment in R and inspect their context. Use as many functions as you can to describe their structure and their dimensions (file size, number of columns, number of lines, ect...). You don't have to limit yourselves to the functions we learned in class.

As a reminder, the files are:

* `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
* `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**

Now we will manipulate the two files in R in order to format them for a downstream analysis. 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 may need to transpose your genotype data so the columns become rows. You just have to know one letter to do this in R: `t()`. However, check the results carefully, as there will be surprises ;)
* As in the UNIX assignment, it might help to write out the entire workflow that will be necessary to produce the files described above before doing the actual analysis.
* If you get stuck or confused, first, use the help() function; second, search the Internet; and, finally, post to the "scripting\_help" channel on Slack and we will provide hints that may be helpful for the whole class.

## Part II

I will add this part after the class on Tuesday, 2/21.