R Assignment

Your R assignment will consist of three parts:

1. Replicating your UNIX assignment in R
2. Additional analysis and visualization
3. Reviewing two assignments from your peers

Create an "R Markdown" file, which includes both the code and your description of the workflow in a version-controlled repository using git. The repository should also include the files you will create as described below. You will be given emails of two randomly selected participants of the class. Please send them an url linking to the GitHub (public) repository you have created **by 1pm on Friday, October 12**. In turn, you will receive links to two repositories to review. When you receive a link, first fork the repository, than clone the forked repository on your computer and write a review inside it named [your lastname]_review.Rmd. Push your review to the forked repository and submit a Pull request **by 9am on Tuesday, October 16**. Accept the pull requests of your reviewers. Finally, submit your assignment in Canvas **by 1pm on Wednesday, October 17**.

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:

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

Manipulate the two files in R in order to format them for a downstream analysis. During this process, we will need to join 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 that 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

We will use ggplot to visualize our data in this part. Note, that it may be easier to reshape the original data (make it tidy) using the `melt` command in the `reshape2` package before attempting this part.

SNPs per chromosome

Plot the total number of SNPs in our dataset on each chromosome. What groups contribute most of these SNPs?

Missing data and amount of heterozygosity

Create a new column to indicate whether a particular site is homozygous (has the same nucleotide on both chromosomes (i.e., A/A, C/C, G/G, T/T) or heterozygous (otherwise)). Recode the missing data as NA. Sort your dataframe using Group and Species_ID values. Make a graph that shows the proportion of homozygous and heterozygous sites as well as missing data in each species (you won't be able to see species names) and each group. For groups normalize the height of individual bars using one of the ggplot "position adjustments" options.

Your own visualization

Visualize one other feature of the dataset. The choice is up to you!