**AGRO-932 Spring 2022 HW1 (100 points):**

**Due Date: Feb. 15th, 2022 at 11 AM**

Simulate an NGS dataset for two populations with a small reference genome (< 1 Mb) of your choice, each population with 10 diploid individuals, to address a hypothesis about population differentiation.

1. Establish a version-controlled directory system to host the project (paste the link here). Show your work in the GitHub repository in a user-friendly and reproducible manner.

[JIN-HY/agro932-hw1: homework1 of agro932 (github.com)](https://github.com/JIN-HY/agro932-hw1)

1. Describe your simulation strategy and the hypothesis to test (positive, negative, or neutral selection).

The data is from a resequencing experiment of a Sorghum panel which includes 403 inbred lines. The .vcf file of 2,000,000 to 3,000,000 bp of the chromosome 9 is used. Although there are subpopulations in this panel, this information is unknown due to mislabeling and/or contamination.

I divide the whole panel into 2 subpopulations of size 196 and 207 respectively. Thetas are calculated for each population to examine the nucleotide diversity of this 1 Mb region. Fst for each SNP locus between two populations is computed to test if there is population differentiation. The Fsts were averaged by a sliding window strategy, with window size of 10kb and step size of 2kb. Based on Wright 1978, Fst is the value above which there is very great differentiation, the range 0.15 to 0.25 as indicating moderately great differentiation. If high Fst value is observed, I will infer that there is positive selection. If low Fst, for example <0.05, is observed, I will infer that there is negative selection. If the Fst is relative low, the region could be under neutral selection.

1. Calculate thetas for each population and compute Fst between the two populations.

The theta for population 1 is 0.007904414.

The theta for population 2 is 0.007703527

The Fst between two populations is shown as the plot below.

1. Visualize and briefly interpret your results.

We can see from the graph that most loci have low Fst, which means it may be under negative selection and the deleterious mutations are lost. A few regions have very high Fst, which can be larger than 0.5, this indicates there can be strong positive selection. It will be promising to dig into these positions to check what functions the SNP are associated with. I am still trying to troubleshoot the missing data in the Fst calculation, which likely causes the negative values in the plot. Another possible explanation for the negative values is that some loci are not biallelic, which in the .vcf file could be coded differently, but my code does not take these into consideration.

Besides, another interesting finding is that although negative, or wrong values, cannot be seen from the plot generated from sliding window, they do exist, so it is always good to plot the data step by step.



