**Homework 2**

**Statistical Genomics (545)**

**Spring 2023**

Instructor: Zhiwu Zhang

Due on Monday, March 20, 2023, 3:10PM PST

**Data files**: Choose a dataset (please specify the dataset number) from the recommended list (<http://zzlab.net/StaGen/2023/Data/PublicData.pdf>), or a dataset outside the list (please specify source, 5 extra points), or your own data that can be released to the public (please specify source, 5 extra points). You can sample a subset of individuals or markers; however, the final dataset must contain at least 100 individuals and 5,000 markers with known chromosome and base pair positions.

**Hand in:** Email your R RMD and Knit (PDF) with subject of “StaGen545 HW2” to [Zhiwu.Zhang@WSU.edu](mailto:Zhiwu.Zhang@WSU.edu). Name your files as

“Homework2\_ firstname\_lastname”. PDF file is limited to one page per question (5 points off for violation). Your PDF file should focus on design, results, and interpretation by text, tables and figures. Leave the R code in the RMD file.

**Objectives**: 1) Simulate phenotype from genotype; 2) Perform GWAS by correlation; 3) Evaluate true and false positives.

1. Out of the available SNPs, sample 10 SNPs as QTNs. Simulate QTN effects from a standard normal distribution. Assign genetic effects for each of the individuals. Simulate normally distributed residual effects with appropriate variance to have a heritability of 0.75. Add residual effects to genetic effects to create phenotypes. You can either use the G2P R function or code everything by yourself. Describe the distribution of genetic effects, residual effects, and phenotypes, and explore the relationships among them (20 points).
2. Perform GWAS by using the correlation method. You can either use the GWASbyCor R function or code everything by yourself. Create a Manhattan plot and label the positions of the QTNs (20 points).
3. Find the number of QTNs among the top ten associated SNPs (20 points).
4. Count the number of SNPs with P values smaller than the P value of the seventh significant QTN (20 points).
5. Redo (3-4) for 100 replicates. Report the averages and standard deviations (20 points).

**Extra credit**

1. Simulate phenotypes from genotypes so that the phenotypes with skew normal distribution due to genetic effect with a long tail on the right (25 points, report is limited to one extra page).