**Homework 1**

**Statistical Genomics (545)**

**Spring 2025**

Professor: Zhiwu Zhang

TA: Meijing Liang

Due on Monday, February 24, 2025, 3:10PM PST

**Data files**: You can use your own data (five extra points) or the following files that can be download at <http://zzlab.net/GAPIT/data/>:

1. mdp\_numeric.txt from GAPIT demo data. The data file contains 281 individuals (row-wise) and 3093 SNPs (column-wise) coded as 0/1/2.
2. mdp\_SNP\_information.txt. The file contains SNP ID, chromosome, and position.
3. CROP545\_Phenotype.txt. The file contains taxa name and phenotype.
4. CROP545\_Covariates.txt. The file contains taxa name and two covariates.

**Hand in:** Upload your package to GitHub and Email the link with subject of “StaGen545 HW1” to [Zhiwu.Zhang@WSU.edu](mailto:Zhiwu.Zhang@WSU.edu) and copy to TA ([Meijing.Liang@WSU.edu](mailto:Meijing.Liang@WSU.edu)).

**Objective**: Develop your own R package to perform GLM GWAS.

1. The package should contain at least three inputs (y, X, and C) that are R numeric data frame objects. Their dimensions are n by 1, n by m, and n by t, corresponding to phenotype, genotype, and covariate data, where n is number of individuals, m is number of markers, and t is number of covariates. The function should return probability values with dimensions of 1 by m for the association tests between phenotypes and markers. Markers are tested one at a time with covariates in C included as covariates (15 points).
2. The package should perform PCA and incorporate PCs as cofactors for GWAS. Your package should also automatically exclude the PCs that are in linear dependence with the covariates provided by users. (25 points).
3. Develop a user manual and tutorials. Name your package and create a logo. (20 points).
4. Perform GWAS on the data provided or your own data, which must contain cofactors (15 points).
5. Demonstrate that your method is superior to the competing method (GWASbyCor) through simulation with at least 30 replicates (25 points).

**Extra credit**

1. Use your own data other than provided by the course (5 points).
2. Implementation of other methods beyond correlation and GLM (10 points).
3. Demonstrate that your package is better than BLINK C version (<http://zzlab.net/blink>) in either statistical power or speed (25 points).