**Homework 2**

**Computational Genomics**

**Workshop of Genomic Prediction**

**Harbin, China, December 26-30, 2016**

Professor: Zhiwu Zhang

Due on January 15, 2017, 8:00PM

**Data files**: mdp\_numeric.txt from GAPIT demo data. The data file can be download from <http://www.zzlab.net/GAPIT/GAPIT_Tutorial_Data.zip>. The data contains 281 individuals (row wise) and 3093 SNPs (column wise) coded as 0/1/2. The SNP ID, chromosome and position is indicated by a file named mdp\_SNP\_information.txt

**Hand in:** Email your report (PDF, limited to five page) and R source code (text file) with subject of “GS2016HarbinHW2” to

NEAU participants: Jiabo Wang (wangjiaboyifeng@163.com).

Onsite participants: Zhao Li (yslizhao@163.com).

Online participants: You Tang (tangyou9000@163.com).

Name your files as following:

Homework2\_ firstname\_lastname.pdf and Homework2\_ firstname\_lastname.R

**Grade components**: 1) Hypothesis or statement; 2) Results; 3) Methods; 4 presentation; 5) R source code (clarity, simplicity and documenting comments)

**Objectives**: 1) gBLUP; and 2) ridge regression.

1. Use GAPIT.Phenotype.Simulation function to simulate phenotypes with heritability of 50% controlled by 20 QTNs having effects with standard normal distribution. Calculate correlation coefficient between the total genetic effects (breeding values) and phenotypes of individuals. Vary the parameter of heritability and explore the relationship between the heritability and the correlation coefficient between breeding values and phenotypes. Your observation should be based on average and standard error of multiple replicates (20 points).
2. With the simulated phenotypes from (1), randomly select 80% of the individuals as training population and the rest as testing population. Perform gBLUP with GAPIT. Calculate the correlations between the predictions and phenotypes, and the correlation between predictions and breeding values in training and testing populations separately. Repeat the random selection and prediction 30 times. Compare the means and standard error of the correlations in training and testing population (40 points).
3. With the simulated phenotypes from (1), perform random division of the population into five even (roughly) sub populations (5-folds). Perform 5-folds cross validation to evaluate accuracy in testing population by using Ridge regression with rrBLUP. Calculate accuracy as the average of the correlations between predicted and observed phenotypes, and the correlations between predicted and observed breeding values in the testing populations. Repeat the random division and prediction 30 times. Compare the mean and standard error of the accuracy with the result from (2) (40 points).