**Homework 1**

**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 “GS2016HarbinHW1” to:

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

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

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

Name your files as following: Homework1\_ firstname\_lastname.pdf and Homework1\_ 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) validation; 2) invalidate validation; 3) cross validation; 4) MAS.

1. Use GAPIT.Phenotype.Simulation function to simulate phenotypes with heritability of 50% controlled by 20 QTNs having effects with standard normal distribution. Display the distribution of QTN effects, and the correlation between the total genetic effects (breeding values) and phenotypes of individuals (20 points).
2. Perform GWAS on the simulated phenotypes with all the individuals by using FarmCPU and selected the top 20 associated SNPs. Perform random division of all the individuals into two even (roughly) sub populations A and B. Estimate the effects of the 20 associated markers in sub population A. Use the estimated effects of the 20 SNPs to predict the phenotypes and BV in sub population B. Repeat the random division 30 times. Report the means and standard deviations of the prediction accuracy (30 points).
3. Repeat (2) except randomly shuffling the simulated phenotypes before GWAS. Describe the difference from (2) and your expectation (20 points).
4. With the simulated phenotypes from (1), randomly select 80% of the individuals as training population and the rest as testing population. Perform MAS. 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 deviations of the correlations in training and testing population (30 points).