**Homework 2**

**Statistical Genomics**

**CROPS 545, Spring 2016**

Professor: Zhiwu Zhang

Due on February 17, 2016, Wednesday, 3:10PM, PST

**Objectives**: To exam the impact of 1) missing rate; 2) sample size; and 3) method on imputation accuracy evaluated as correlation coefficient, match proportion across genotypes, and match proportions on major and minor allele homozygous separately.

**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 “CROPS545 HW2” to Zhiwu.Zhang@WSU.edu. 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)

1. Randomly select 5%, 10%, 25%, 50% and 75% of data points and set them as missing values. Impute these missing values with the stochastic imputation method. Calculate imputation accuracy as correlation coefficient or match proportion. Repeat this process at least 30 times, report average, standard deviation and number of replicates. Describe the relationship between the missing rate and imputation accuracy (20 points).
2. Fix the missing rate at 25% and randomly sample 75%, 50%, and 25% of individuals to perform imputation with the stochastic imputation method. Calculate imputation accuracy as correlation coefficient or match proportion. Repeat this process at least 30 times, report average, standard deviation and number of replicates. Describe the relationship between the sample size and imputation accuracy (20 points).
3. Fix the missing rate at 25% and use all the individuals to perform imputation with KNN method with K=2, 5, 10 and 20. Calculate imputation accuracy as correlation coefficient or match proportion. Repeat this process at least 20 times, report average, standard deviation and number of replicates. Describe the relationship between the number of nearest neighbors and imputation accuracy (20 points).
4. Fix the missing rate at 25% and use all the individuals to perform imputation with the stochastic method, KNN and BEAGLE. Calculate imputation accuracy as correlation coefficient or match proportion. Repeat this process at least 10 times, report average, standard deviation and number of replicates. Describe the relationship between imputation method and imputation accuracy (20 points).
5. The neighbors in KNN refer to individuals and attributes refer to genetic markers for imputation of missing genotypes. Redo (3) by switching neighbors to genetic markers and attribute to individuals. Describe the differences (20 points).

**Extra credit**:

1. Randomly select 5%, 10%, 25%, 50% and 75% of data points and set them as missing values. Use all the individuals to perform imputation by using KNN method with K=10. Calculate imputation accuracy as correlation coefficient, match proportion for all genotypes, and match proportions for major allele homozygous and minor allele homozygous separately. Repeat this process at least 30 times, report average, standard deviation and number of replicates. Describe the relationship among different types of imputation accuracy (25 points, report is limited to one extra page).
2. Find another method and demonstrate that it has better imputation accuracy than both KNN and BEAGLE (30 points, report is limited to one extra page).