**Homework 2**

**Statistical Genomics**

**CROPS 545, Spring 2020**

Professor: Zhiwu Zhang

Due on February 24, 2020, Monday, 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.

**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 and interpretation; 3) Methods; 4) R source code (clarity, simplicity and documenting comments); and 5) presentations.

1. Choose a dataset (please specify the dataset number) from the recommended list (<http://zzlab.net/StaGen/2018/Data/PublicData.pdf>), or a dataset outside the list (please specify source, 5 extra points), or your own data that can be released to public (please specify source, 5 extra points). You can sample partial of individuals or makers, however, the final data set must contain over 100 individuals and 5,000 markers with known chromosome and base pair positions. Display marker locations on chromosomes, distribution of missing rate (both maker wise and individual wise), and minor allele frequency (20 points).
2. Randomly select 5%, 25% and 50% 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).
3. Redo (2) by replacing stochastic method with KNN method. Describe the differences of results from (2) (20 points).
4. 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).
5. Fix the missing rate at 25% and perform imputation with 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 advantage over KNN (20 points).

**Extra credit**:

1. Redo (3) with imputation accuracies for major allele homozygous and minor allele homozygous separately. (20 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 (20 points, report is limited to one extra page).