**BIOL 5153 Project Proposal: Performing Genomic Selection in Wheat Through the Use of the University of Arkansas High Performance Computing Center**

**Introduction**

Genomic selection is an important tool for plant breeders because it allows for breeders to select experimental breeding lines while using less resources for field phenotyping. A genomic selection program typically consists of a prediction set and a training set. The prediction set contains early generation breeding lines that have been genotyped but not phenotyped. The training set is a small set of breeding lines that have been both genotyped and phenotyped. The data from the training set are used to predict the phenotypic performance of the prediction set. Predictions are made through the use of a prediction model. There are many models that can be used, however more complex models, such as Bayesian models require more computing power and struggle with large genotypic data sets. As a result, most breeding programs rely on a model called ridge-regression best linear unbiased prediction (RR-BLUP), which assumes that all markers have equal variances with a small but non-zero effect and requires less computing power (Desta and Ortiz, 2014).

**Objectives**

The objective of this project is to use a training set of soft red winter wheat genotypes from the Southern United States to predict the performance of five traits, including yield, test weight, heading date, plant height, and maturity date for two panels of soft red winter wheat breeding lines from the SunGrains 2017 and 2018 Genotyping by Sequencing populations. Forward predictions will be performed using the rrBLUP package in R (Endelman, 2011). As genomic selection requires a large amount of computing power, the analysis will be performed at the University of Arkansas High Performance Computing Center (HPCC) through the use of remote computing. Using the HPCC for this analysis will help to produce forward prediction results at a faster pace compared to a standard desktop computer. These forward prediction results will help the University of Arkansas wheat breeding program in making early generation selections for the 2019 growing season.

**Data Sources**

*Training Set*

The training set consists of 354 soft red winter wheat breeding lines from the University of Arkansas, University of Georgia, Louisiana State University, and North Carolina State University wheat breeding programs. The training set was grown at two locations over four years for a total of eight location-years. The training set was phenotyped at each location year for five traits including grain yield, test weight, heading date, plant height, and maturity date. Phenotypic means were evaluated for all five traits at each individual location-year using a mixed linear model fitted with a spatial covariance structure using the mixed procedure in SAS 9.4 (SAS Analytics, Cary, NC). Best linear unbiased predictions (BLUP) of all five traits for each line were then calculated by combining the phenotypic means from each of the location-years. Each of the 354 breeding lines were also genotyped by the North Carolina State University Genotyping Lab using genotyping by sequencing (GBS) for a total number of 45,749 single nucleotide polymorphisms (SNP). Missing data were imputed using the LD-kNNi imputation method in TASSEL 5.0 (Bradbury et al., 2007). The genotypic data was then converted from a HapMap format to a numerical format, where 0 and 2 represent homogeneous alleles and 1 represents the heterogeneous allele, using GAPIT Version 2 in R (Tang et al., 2016).

*2017 Prediction Set*

The 2017 prediction set consists of 3,852 soft red winter wheat breeding lines from the University of Arkansas, University of Georgia, Louisiana State University, Texas A&M University, Clemson University, North Carolina State University, Virginia Polytechnic Institute, University of Kentucky, University of Florida, and USDA-ARS wheat breeding programs. Each of the breeding lines were genotyped using the same method as the training set and they include the same 45,749 SNP as the training set. Missing SNP data was imputed using the LD-kNNi imputation method in TASSEL 5.0 (Buckler Lab, Ithaca, NY). Phenotype data was not collected for the 2017 data set. The genotypic data was then converted from a HapMap format to a numerical format, where 0 and 2 represent homogeneous alleles and 1 represents the heterogeneous allele, using GAPIT Version 2 in R (Tang et al., 2016).

*2018 Prediction Set*

The 2018 prediction set consists of 4,765 soft red winter wheat breeding lines from the University of Arkansas, University of Georgia, Louisiana State University, Texas A&M University, Clemson University, North Carolina State University, Virginia Polytechnic Institute, University of Kentucky, University of Florida, University of Maryland, and USDA-ARS wheat breeding programs. Each of the breeding lines were genotyped using the same method as the training set and 2017 prediction set. The genotypic data consists of 45,749 SNP like the training set and the 2017 prediction set. Missing SNP data was imputed using the LD-kNNi imputation method in TASSEL 5.0 (Bradbury et al., 2007). The genotypic data was then converted from a HapMap format to a numerical format, where 0 and 2 represent homogeneous alleles and 1 represents the heterogeneous allele, using GAPIT Version 2 in R (Tang et al., 2016).

**Methods**

The genomic selection analysis will be performed using the rrBLUP package in R version 3.4.4 (Endelman, 2011; R Core Team, 2018). Two forward prediction analyses will be performed, the first for the 2017 prediction set and the second for the 2018 prediction set. Both analyses will use the same training set. Genotypic data and phenotypic data for all five traits will be used from the training set to predict the performance of the 3,852 and 4,765 breeding lines from the 2017 and 2018 prediction sets, respectively. The output from each analysis will include a comma separated file containing predictions for all five traits for each of the predicted breeding lines.

A subset of 283 randomly selected lines from the training set will then be used to predict the performance of the rest of the training set for the purpose of cross-validation. The cross-validation analysis will be run 500 times in order to test the accuracy of the prediction model. Output from this analysis will include a comma separated file containing the prediction accuracies from each run and a boxplot of the results from cross-validation analyses for all five traits.

**Conclusions**

This project should test my ability to successfully utilize the University of Arkansas High Speed Computing Center, as taught in class. The results from the aforementioned analyses will also be helpful for my doctoral research and the University of Arkansas wheat breeding program.

**Citations**

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, E.S. Buckler. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.

Desta, Z.A and R. Ortiz. 2014. Genomic selection: genome-wide prediction in plant improvement. Trends in Plant Science 19(9):592-601.

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250-255.

R Core Team. 2018. R: A language and environment for statistical computing [Online]. Available at <https://www.R-project.org/> (accessed 27 Mar. 2018). R Foundation for Statistical Computing, Vienna, Austria.

Tang, Y., X. Liu, J. Wang, M. Li, Q. Wang, F. Tian, et al. 2016. GAPIT version 2: an enhanced integrated tool for genomic association and prediction. The Plant Genome 9(2):1-9.