Data Bootcamp for Genomic Prediction in Plant Breeding
University of Minnesota Plant Breeding Center
July 5 - 7, 2023
Practical 1: Effect of missing data, training population size, and marker number on prediction accuracy

Getting Started

These data are from Soybean Nested Association Mapping Population (SoyNAM) project. More information on this population and associated genotypic and phenotypic data can be found in the course slides and Diers et al. (2018) provided in the course literature. The whole SoyNAM population consists of over 5000 RILs, but I subset the dataset down to 500 RILs to the data manipulations and computations do not take too long for this demonstration.

Follow the "practical_1.R" script to upload, format, filter, and impute the genotypic data. We will go over each of these tasks in class. Next, we will fit two models using the rrBLUP package. This is a simple and easy to use package that implements the RR-BLUP and G-BLUP models. Fit the models to the data and compare the GEBVs from each model.

Next, let's implement a cross-validation analysis. A cross-validation analysis is a useful way to assess the predictive ability of a statistical model while protecting against model overfitting. Basically, a *k*-fold cross-validation leaves one k^{th} part of the data out of the model training process, then predicts these "unobserved" genotypes, and compares the predictions to the observations. This way, the model is not influenced by noise in the observed data, testing the models ability to predict unobserved samples (genotypes) in this case. In this example, a 10-fold cross validation is implemented, but there are many flavors of this, including a leave-one-out cross-validation (leave one sample out of model training, then predict its value).

You can use the cross-validation analysis to assess the effect of various factors on model predictive ability such as training population size, marker number, and proportion of missing data. For changing the population size, change the number of RILs being randomly sampled around line 92 in the script (line indicated in script). For changing the marker number by randomly selecting a subset of markers, you can use similar code but now index the columns instead of the rows. In this example, 500 markers are being randomly sampled.

```
mrkNdx <- sample.int(n=dim(pheno2)[2], size=500)
geno_imp_sub <- geno_imp_sub[, mrkNdx]
```

To vary the amount of missing data, go above and work with object geno_num5. To add additional missing data to this marker matrix randomly, use the following code. For this example, we are randomly setting 10,000 marker datapoints as NA.

```
totElem <- dim(geno_num5)[1]*dim(geno_num5)[2]
addNa <- sample.int(n=totElem, size=10000)
geno_num5[addNa] <- NA
```

Note: You will need to make subtle changes based on any changes you make to names of the objects.

Tasks to perform and questions to consider

Data Bootcamp for Genomic Prediction in Plant Breeding
University of Minnesota Plant Breeding Center
July 5 - 7, 2023
Practical 1: Effect of missing data, training population size, and marker number on prediction accuracy

1. Examine the effect of imputation on model predictive ability by varying the imputation method between naïve imputation and the Markov chain method.

2. Choose a trait and look at the effect of training population size, marker number, or proportion of missing data on prediction accuracy. Feel free to choose multiple variables and traits if you have time.

3. Determine the extent of overfitting, and examine how the level of overfitting is influenced by training population size.