Data Bootcamp for Genomic Prediction in Plant Breeding
University of Minnesota Plant Breeding Center
July 5 - 7, 2023
Practical 2: Cross validation and assessing model accuracy

Getting Started

This script introduces the use of the BGLR package, which is a powerful and flexible package useful for implementing many genomic prediction models, mostly Bayesian models.

The Practical 2 script implements a cross-validation analysis to assess the predictive ability of the model. Assessing model accuracy is best accomplished by removing observations, fitting the model, and then predicting values of individuals not included in the model. This provides an independent assessment of model performance by avoiding the effects of model overfitting.

For starters, the script removes one whole family within the SoyNAM population. Train a model with a dataset that excludes phenotypic data for this family, then predict the performance of the lines in that family. Although the analysis will take longer, it may be worthwhile to increase the population size from 1000 to 5000.

Tasks to perform and questions to consider

1. Using family 64 as a validation set, choose a trait and compare the predictive ability of at least three different models. Feel free to add rrBLUP code (can borrow from Practical 1) to run an RR-BLUP or G-BLUP model.

2. Compare the marker effects against one another for two contrasting models, such as Bayesian ridge regression and Bayes B.

3. Execute a Reproducing Kernel Hilbert Spaces model using an additive kernel. Hint, you can use the A.mat function in the rrBLUP package to calculate the genomic additive relationship matrix.

4. Try fitting a non-linear kernel, such as the Gaussian kernel as described near the bottom of page 15 in Perez and de los Campos BGLR documentation posted to the literature folder on Google Drive.