## **Getting Started**

## Download “markers.csv”, “pheno.csv”, “geno.csv”, “modelSelFunc.r” and “GS Practical 1.r” from the provided Google Drive link and save them to your working directory.

## 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.

Follow the script “Practical R.r” script to upload, format, filter, and impute the genotypic data. We will go over each of these tasks in class. Next, starting at line 81 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 examine the effect of proportion of missing genotype data, marker number, and population size on the fit of the model. For changing the population size, change the number of RILs being randomly sampled at line 76. 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.

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 to line 67 and work with object geno\_num5. To add additional missing data to this marker matrix randomly, use the following code:

totElem <- dim(geno\_num5)[1]\*dim(geno\_num5)[2]

addNa <- sample.int(n=totElem, size=10000)

geno\_num5[addNa] <- NA

To look at the correlation between predictions and observed phenotypes, you can use

cor(rrGebv, pheno2\_sub$Yield)

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**

1. What is the correlation between GEBVs calculated using RR-BLUP and the observed phenotypes? How about for G-BLUP?
2. Make a plot to show how training population size affects the correlation between GEBVs and observed phenotypes.
3. Make a plot to show how marker number affects the correlation between GEBVs and observed phenotypes.
4. How does the proportion of missing data affect the correlation between GEBVs and observed phenotypes?