## **Getting Started**

## Use the same files “markers.csv”, “pheno.csv”, “geno.vcf”, “modelSelFunc.r” and “GS Practical 2.r” from Practical 1. The start of the script will be the same, only the end differs.

This script also 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, starting at line 105, the script removes one whole family within the SoyNAM. 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 by increasing the size ssNdx at line 76.

Next, starting at line 141, the script performs a 10-fold cross validation. The dataset is shuffled, then phenotypic data is set as missing in 10 folds, and predictions are made and stored in the predStor object.

**Tasks to perform and questions to consider**

1. For Family 09, how do the correlations between GEBVs and phenotypes compare between when RILs of this family are left in the model fitting, versus when they are removed for validation?
2. Compare the predictive ability assessed when Family 09 was left out, versus the predictive ability estimated using the 10-fold cross validation.
3. Using the BGLR documentation and instructional manuscript, try fitting at least four other models. You can fit a RR-BLUP model using the rrBLUP package if you want. Use the cross-validation to compare the models for their performance.