CROPS_545
Homework 6
Evan Craine

(1) Use GAPIT.Phenotype.Simulation function to simulate phenotypes with heritability of 50% controlled by 20 QTNs having effects with standard normal distribution. Display the distribution of QTN effects, and the correlation between the total genetic effects (breeding values) and phenotypes of individuals (5 points).

Hypothesis: Based on the designated parameters for the function to simulate phenotype, I predict that the QTN effects will follow a normal distribution and that the r^2 value will be close to .5.

Methods: The *GAPIT.Phenotype.Simulation* function was used to simulate phenotypes according to the designated parameters. Genetic effects were set to *mySim$u*, phenotypes were set to *mySim$Y*, and r^2 was set to *cor((effects, myY[,2])^2* rounded to the third decimal place.

Results and Discussion: Genetic effect values ranged from -14.237 to 7.223, and phenotype values ranged from -18.376 to 12.275. A somewhat strong correlation of 0.511 was found between the genetic effects and phenotypes when the covariates we removed (Figure 1b). The distribution for QTN effects, follows a normal distribution (Figure 1b). These results agree with the predictions made according to the parameters set for the phenotype simulation.

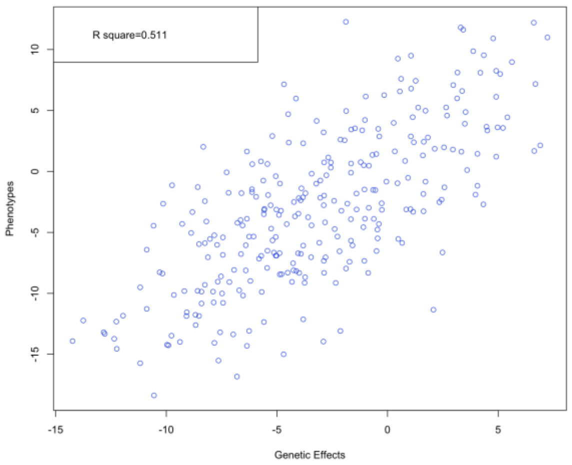

Figure 1a: Phenotypic values from phenotype simulation using *GAPIT* plotted against genetic effects from the same simulation. R squared value is given in the top left of the figure.

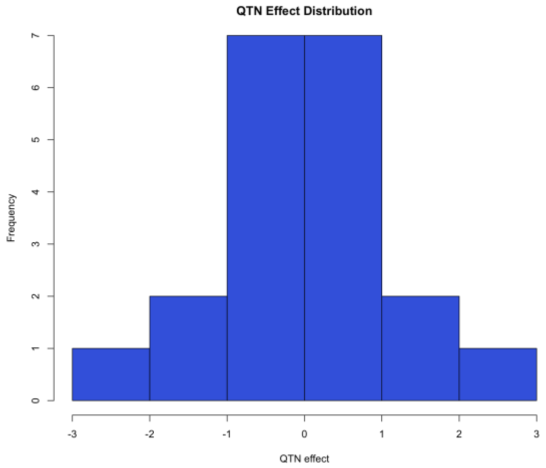

Figure 1b: Histogram of QTN effects plotted for phenotype simulation by *GAPIT*.

CROPS_545
Homework 6
Evan Craine

(2) Perform GWAS on the simulated phenotypes with all the individuals by using *FarmCPU* and selected the top 20 associated SNPs. Perform random division of all the individuals into two even (roughly) sub populations A and B. Estimate the effects of the 20 associated markers in sub population A. Use the estimated effects of the 20 SNPs to predict the phenotypes and BV in sub population B. Repeat the random division 30 times. Report the means and standard deviations of the prediction accuracy (20 points).

Hypothesis: Since we will be using the top 20 SNPs from GWAS using *FarmCPU* in sub population A to predict the phenotypes and BV in sub population B, I predict that a relatively strong accuracy will be calculated. Specifically, I predict that the process will produce a mean accuracy of 50%, with a standard deviation of one.

Methods: We first simulate the phenotypes and perform GWAS to calculate PCA using *GAPIT*. We then perform GWAS using *FarmCPU* and save the genotypic data from the top 20 associated SNPs. Using a loop, we divide the individuals and use one half as the testing population in GWAS using the 20 SNPs as QTN with *GAPIT* to produce effect values, which are then used to calculate the phenotypes for the training population. We replicate this process thirty times and calculate means for accuracy using the *cor* function for phenotypes and GEBVs.

Results and Discussion: Mean accuracy for predicted phenotype (mean = 0.642) and predicted breeding value (mean = 0.598) was higher than expected (Table 1). This is likely because we used a GLM with QTN included to perform GWAS, which is a relatively strong method. Therefore, the effect values are similar between predicted values in the testing population, and the actual values produced through phenotype simulation using *GAPIT*.

Test	Mean	SD
Accuracy of Predicted Phenotype	0.642	0.045
Accuracy of Predicted Breeding Value	0.598	0.056

Table 1: Means and standard deviations reported for accuracy of predicted phenotype and predicted breeding value as calculated by *cor* function in R.

(3) Repeat (2) except randomly shuffling the simulated phenotypes before GWAS. Describe the difference from (2) and your expectation (15 points).

Hypothesis: I predict that shuffling the simulated phenotypes after simulation with *GAPIT*, but before GWAS, will decrease the accuracy of predicting phenotype and predicted breeding value. I think that this will occur because by shuffling the phenotypes, we disrupt the association between genotypic data and phenotypic data for each taxon. When we use *FarmCPU* to generate the top 20 QTN, the quality of these QTN will be diminished. Thus, the prediction will suffer as a result of the shuffled phenotypes.

CROPS_545
Homework 6
Evan Craine

Methods: The methods will be the same as question two, expect for adding a section of code to shuffle the phenotype between the simulation and GWAS using FarmCPU to generate the top 20 QTN. We use the *sample* function to randomly grab individuals by row name, and then bind this to the phenotype data to destroy the connection.

Results and Discussion: As predicted, mean accuracy for predicting phenotype (mean =0.013) and breeding value (mean = 0.011) via GWAS analysis with *GAPIT* using the top 20 QTN identified with FarmCPU was much lower when phenotypes were shuffled, compared to the results from (2) where the phenotypes were not shuffled. This is likely because the top 20 QTN are not actually QTN, and were incorrectly identified as a result of destroying the association between genotype and phenotype via shuffling. Thus, the accuracy of prediction is zero within one standard deviation.

Test	Mean	SD
Accuracy of Predicted Phenotype	0.013	0.013
Accuracy of Predicted Breeding Value	0.011	0.014

Table 2: Means and standard deviations reported for accuracy of predicted phenotype and predicted breeding value as calculated by *cor* function in R.

(4) With the simulated phenotypes from (1), randomly select 80% of the individuals as training population and the rest as testing population. Perform gBLUP with GAPIT. Calculate the correlations between the predictions and phenotypes, and the correlation between predictions and breeding values in training and testing populations separately. Repeat the random selection and prediction 30 times. Compare the means and standard deviations of the correlations in training and testing population (20 points).

Hypothesis: I predict that correlations will be higher in the training population, due to the large percentage of individuals that make up this population, compared to the testing population. Since *gBLUP* does not rely solely on QTN to make predictions, and instead incorporates genetic effects and environmental variance, I predict that *gGLUP* will deliver a more accurate prediction overall compared to *GAPIT.*

Methods: Phenotypes were simulated using *GAPIT* and stored. The *replicate* function in R was used to first randomly select 80% of the individuals for the training population, and then the rest of the individuals were assigned to the testing population. With *GAPIT* function, gBLUP was used to calculate genetic effects and phenotypes for both the training and testing population using only the training population inside the *replicate* function. The *cor* function was used to calculate r^2 values to assess the accuracy via correlation between predicted phenotype and breeding values from gBLUP and the actual phenotype and breeding values for both the training and testing populations.

Results and Discussion: As predicted, accuracy of predicting the phenotype and breeding value was higher in the training population (mean accuracy of predicted phenotype: r^2 = 0.785; mean accuracy of predicted breeding value: r^2 = 0.504) than in the testing population (mean accuracy of predicted phenotype: r^2 = 0.088; mean accuracy of predicted breeding value: r^2 = 0.140) (Table 3). We see the accuracy of the breeding value close to 50%, which happens to be the heritability value we set during the phenotype simulation. The cause of this decreased accuracy is likely a result of the size of the testing population (i.e. 20% of individuals). When the number of markers (i.e. large *p*) is much larger than the number of individuals (i.e. small *n*), we suffer from the curse of dimensionality, or multicollinearity, which causes a model that is over fitted to the data due to a lack of sufficient degrees of freedom and analysis using ordinary least squares linear regression. Alternatively, we can consider the fact that gBLUP uses all of the markers to estimate kinship, and then kinship is used to generate the breeding values. Therefore, we see lower accuracy for the predicted breeding values compared to the predicted phenotypes, which may be an artifact of the quality of the kinship calculation. The much lower accuracy for the predicted breeding value in the testing population compared to the training population is likely a result of the smaller population size limiting the quality of the kinship matrix and the accuracy of the prediction.

Test	Population	Mean	SD
Accuracy of Predicted Phenotype	Training	0.785	0.033
	Testing	0.088	0.043
Accuracy of Predicted Breeding Value	Training	0.504	0.020
	Testing	0.140	0.081

Table 3: Means and standard deviations reported for accuracy of predicted phenotype and predicted breeding value for training and testing populations from gBLUP as calculated by *cor* function in R.

(5) With the simulated phenotypes from (1), perform random division of the population into five even (roughly) sub populations (5-folds). Perform 5-folds cross validation to evaluate accuracy in testing population by using Ridge regression with rrBLUP. Calculate accuracy as the average of the correlations between predicted and observed phenotypes, and the correlations between predicted and observed breeding values in the testing populations. Repeat the random division and prediction 30 times. Compare the mean and standard deviation of the accuracy with the result from (4) (20 points).

Hypothesis: I predict that with 5-fold cross validation the accuracy of predicting the phenotype and breeding value will be higher compared to the results in (4) because we will be more thoroughly dividing up the individuals into training and testing during each cross validation, and will therefore be able to average correlations which will strengthen our genomic selection model. If we were performing one iteration, I would predict that gBLUP and rrBLUP would produce similar results, because the two methods are mathematically identical albeit employing different approaches. For instance,

CROPS_545
Homework 6
Evan Craine

Methods: In (4), we only used one "learning event" to train and test for genomic selection. We chose 80% of the individuals to serve as the training population, and 20% of the individuals to serve as the testing population. In (5), we will instead be performing 5-fold cross validation (i.e. 5 "learning" events), where the individuals are divided into "folds" five times to represent the training (80%) and testing (20%) populations. This effectively allows us to divide up the total population of individuals to create a better model to predict the phenotypes and breeding values. We use the *replicate* function in R to accomplish the task. We first create null objects to fill later in the loop, and then use the *seq* function to divide the individuals into five folds by setting the *break* parameter equal to five, and then randomly assigning the groupings. Please see Figure 2 for a visual representation of this methodology.

Results and Discussion: Using rrBLUP, we see an improvement on the accuracy the mean accuracy of predicted phenotype and breeding value compared to the results from (4) (Table 4). This is most likely as result of the 5-cross validation process employed in (5), and the advantages conferred by replacing gBLUP with rrBLUP. As we have learned in lecture, when the top SNP as determined through GWAS can be used as cofactors along with the PCA and covariates, the accuracy of predicting the phenotype and GEBV improves greatly from these fixed effects. This was not the case in (4), since we do not define the parameter CV in *GAPIT*. In rrBLUP, we do not define any fixed effects, and we assume a multivariate normal distribution with a covariance structure according to "G" matrix (elemets are realized genomic relations according to markers). To address the 'curse of dimensionality' and correct the problem of overfitting, rrBLUP takes the approach of placing a penalty (i.e. dividing unit variance by total number of markers) on the model to make sure it does not explain more than 100% of the variance. Lastly, I was surprised to see a higher accuracy in predicting the phenotype compared the GEBV in (5). The GEBV represents the additive genetic effect that makes up the phenotype along with environmental noise, and thus one might expect a better correlation when there isn't an attempt to predict this environmental noise. However, we do define *pcEnv* as the PCA from *GAPIT* and the covariates from *myCV*.

Test	Mean	SD
Accuracy of Predicted Phenotype	0.784	0.005
Accuracy of Predicted Breeding Value	0.434	0.019

Table 4: Means and standard deviations reported for accuracy of predicted phenotype and predicted breeding value from rrBLUP as calculated by *cor* function in R.