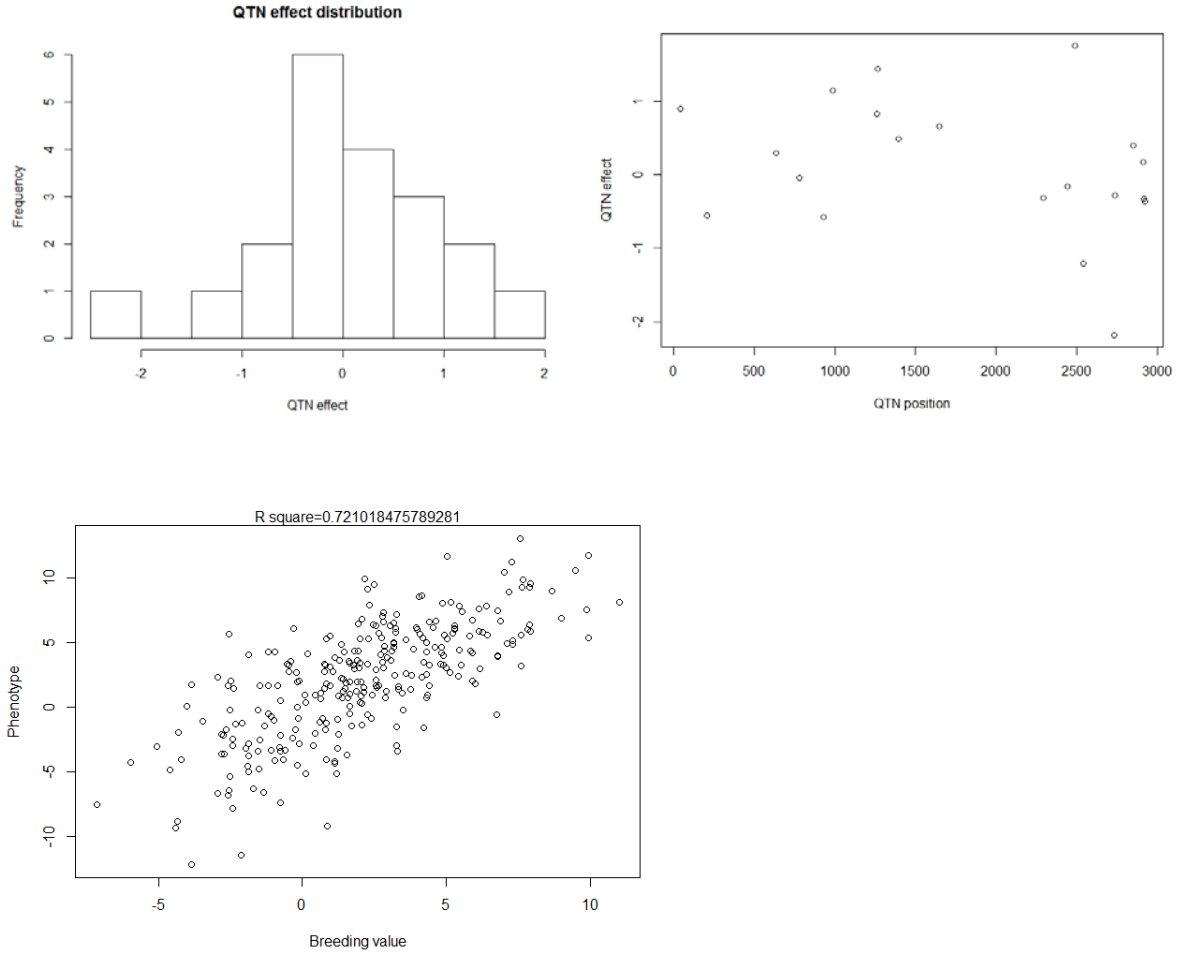
Ryan Oliveira

Statistical Genomics (CROPS 545), Spring 2017 Homework #6

(1) Using the GAPIT demo data, the GAPIT.Phenotype.Simulation function is run to simulate phenotypes on the dataset with a heritability of 50% and 20 QTNs with a standard normal distribution. The graphs showing the distribution of the QTNs is below, confirming (top left) that they approximately follow a normal distribution by histogram and (top right) that they are fairly randomly distributed across the genome. On the bottom left, a decently strong correlation is seen between breeding value (estimated genetic effect) and actual phenotype ($r^2 = 0.721$), confirming at least a moderate degree of heritability.



(2) Using FarmCPU and the simulation from question 1, GWAS is done on the simulated phenotypes using the FarmCPU program from ZZLab. The top 20 SNPs are selected as the QTN list for purposes of prediction. Using GAPIT, prediction is tested via cross-validation of random division of the population into roughly equal testing and training sets. The process is repeated 30 times, and the average and standard deviation of the prediction accuracy for phenotype and breeding value is reported.

Assessment	Mean	Standard deviation
Phenotype prediction accuracy	0.614	0.0618
Breeding value prediction accuracy	0.611	0.0556

As expected, using the top 20 SNPs yields a strong accuracy for both phenotype and prediction, each with $r^2 > 0.60$ and a small standard deviation. These values can be used as a basis for comparison in question 3.

(3) The same procedure is repeated as in question 2. However, this time the "sample()" function in R is used to randomly shuffle phenotypes after the GAPIT phenotype simulation and before GWAS, breaking the link between genetic effects and phenotype (effectively destroying any QTN effects).

Assessment	Mean	Standard deviation
Phenotype prediction accuracy	0.219	0.0393
Breeding value prediction accuracy	0.0294	0.0233

Despite the data set effectively having no effects from QTNs, prediction using the effects of the top 20 SNPs from GWAS is still able to yield $r^2 = 0.219$ for phenotype prediction. The breeding value accuracy is effectively null, with 0% being contained within two standard deviations, as would be expected given the complete lack of a real genetic effect.

(4) Using the simulated phenotypes, 80% of the population is selected as a training population, and gBLUP is performed is to predict phenotype and breeding value accuracy in the testing and training population separately. The entire process is repeated 30 times. The accuracy in both sets is reported below.

Assessment	Set	Mean	Standard deviation
Phenotype prediction accuracy	Training	0.802	0.0351
	Testing	0.085	0.0758
Breeding value prediction accuracy	Training	0.438	0.0258
	Testing	0.188	0.0941

The accuracy is extremely high within the training population ($r^2 = 0.802$ for phenotype, $r^2 = 0.438$ for breeding value); however, the accuracy is significantly lower (again, r^2 for phenotypic accuracy encompasses 0% within two standard deviations) in the testing population. We may speculate the reason(s) for this. The gBLUP algorithm aims to minimize the least squared error and, like any measure using this method, is prone to fitting the shape of the data set. As the testing population consists of different individuals, and the training population is four times the size of the testing population, gBLUP's failure may be attributed to a degree of "overfitting."

(5) A similar procedure is repeated as in question 4 with two important modifications. First, ridge regression BLUP (rrBLUP) is used instead of gBLUP to make predictions. Second, cross-validation is performed using the "k-fold" method with k = 5, in which the data set is divided randomly into five groups with the testing population iteratively selected, and the predictions in each iteration are averaged. Again, the process is repeated thirty times. Only the accuracy in predicting the testing population is reported below.

Assessment	Mean	Standard deviation
Phenotype prediction accuracy	0.280	0.0624
Breeding value prediction accuracy	0.668	0.0735

The improvement in accuracy is immediately apparent. The reasons for this improvement are perhaps identical to those behind the failure of gBLUP. First, the five-fold cross validation mitigates the drop in

accuracy from any potential outliers in the testing population by effectively iterating through the entire population for the testing group. Second, ridge regression does not utilize a least squares model that is prone to tightly fitting the original data but instead treats SNPs as a random effect and attempts to maximize the likelihood. As a result, ridge regression and five-five cross validation better predict the phenotype and breeding value in a separate testing population.