Stephanie Sjoberg
CROP_SCI 545 HW3

Objective: 1) Simulate phenotype from genotype; 2) GWAS by correlation; 3) Evaluate true and false positives.

Question 1: Sample 10 SNPs as QTNs out of the 3093 SNPs. Simulate QTN effects from a standard normal distribution. Assign genetic effects for each of the 281 individuals. Simulate normal distributed residual effects with appropriate variance to have a heritability of 0.75. Add residual effects to genetic effect to create phenotypes. You can either use the G2P R function or code everything by yourself. Describe the distribution of genetic effect, residual effects and phenotypes and explore the relationship among them (20 points).

Statement: Ten SNPs will be assigned as QTNs randomly in the dataset. Variance components will be simulated based on an assigned heritability of 0.75.

Methods: Genotype data made up of 3093 SNPs across 10 chromosomes for 281 individuals was used for this part of the assignment. Manual coding in R was used to execute the commands necessary to produce the 10 QTNs and variance components.

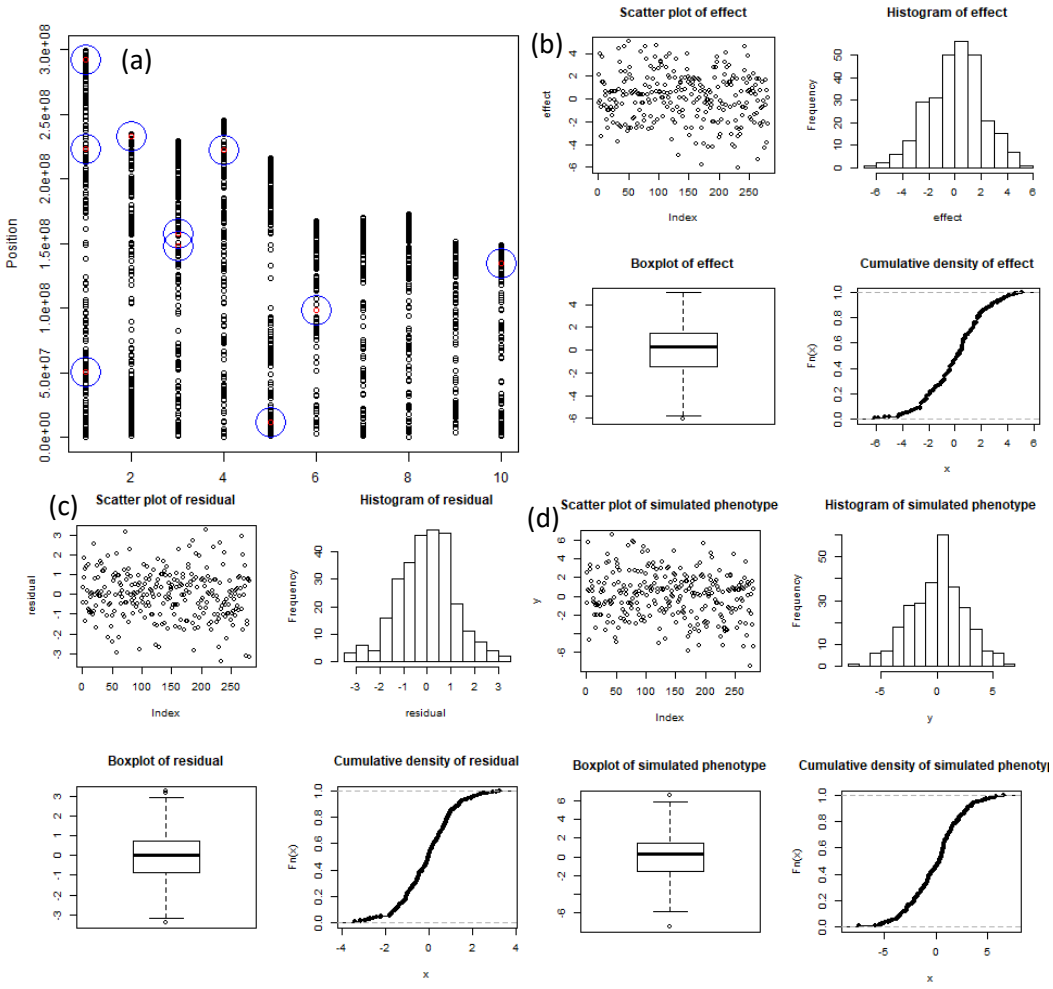


Figure 1. (a) Simulated QTNs mapped across the genome, noted in red and circled in blue, (b) the distribution of the genetic effects shown in scatterplot, histogram, boxplot, and density curve, (c) The distribution of the residual effects, and (d) the distribution of the phenotypes shown in scatterplot, histogram, boxplot, and density curve.

Results and Interpretation: The ten QTN randomly sampled from this dataset are shown in Figure 1a. These QTN covered seven out of the ten chromosomes. The distribution of variance components in including genetic effects, residual effects, and phenotype are shown in Figure1b-d. All components are normally distributed around a mean of zero. The phenotype and genetic effects have a larger variance than the residual effects. Figure 2a shows the density plots for all three effects. The phenotype and genetic effects nearly overlap, which makes sense since the heritability is 0.75, or 75% of the phenotypic variation is due to genetic effects. The phenotypic variance is the sum of the residual and genetic effects, as shown in Figure 2b. The correlation between the phenotypic variance and genetic effects is 0.89, between the phenotypic variance and residual effects is 0.70, and between the genetic and residual effects is 0.01. These values make sense because the phenotypic variance is made up of the genetic and residual effects, while the genetic and residual effects are not tied to each other.

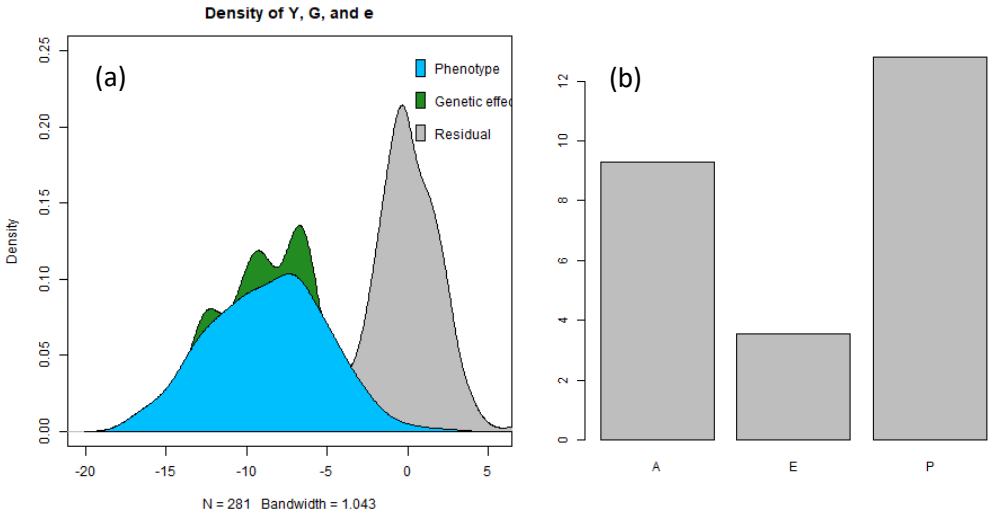


Figure 2. (a) The density plot of the phenotypic variance in black, the additive variance in blue, and the residual variance in red. (b) A bar graph displaying the distribution of variance between additive A, residual or environment E, and phenotype P.

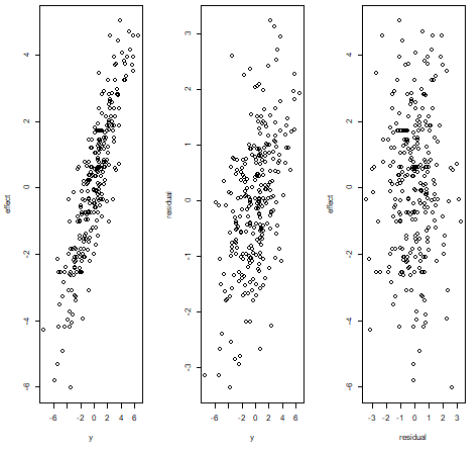


Figure 3. Complete correlations comparisons between genetic effect, residual effect, and phenotype.

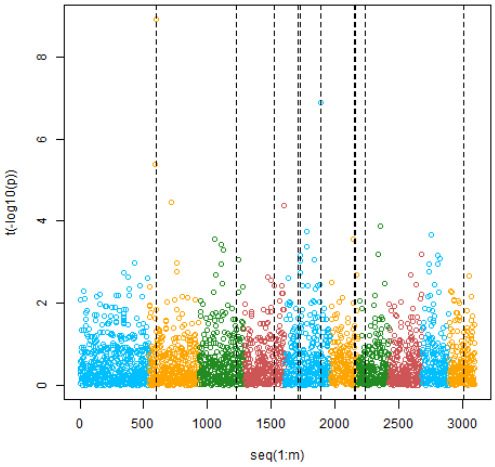


Figure 4. Manhattan plot with LOD plotted across the genome. Vertical lines note where the QTNs are located.

Question 2: Perform GWAS by using the correlation method. You can either use the GWASbyCor R function or code everything by yourself. Create Manhattan plot and label the positions of the QTNs (20 points).

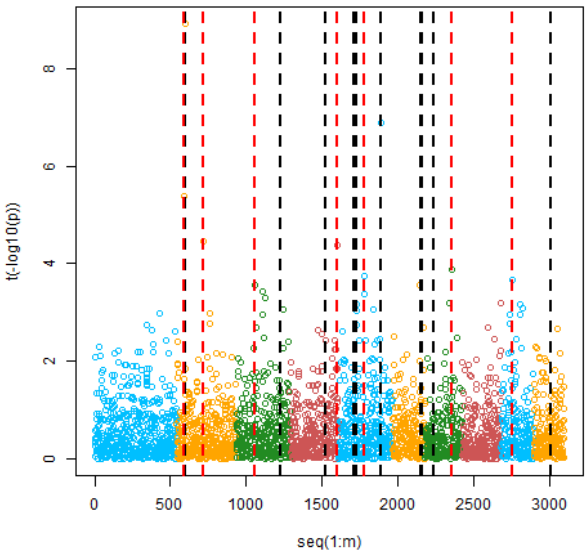
Hypothesis: The correlation method of GWAS will detect significant SNPs.

Methods: The function *GWASbyCor* was used to find correlations between SNPs and the simulated phenotype.

Results and Interpretation: The Manhattan plot shows the LOD value for all SNPs across the genome in Figure 4. The ten QTN are denoted by the dotted vertical lines. Not all the QTN are showing up as significant SNPs and that there are many false positive SNPs.

Question 3: Find number of QTNs among top ten associated SNPs (20 points).

Hypothesis: Some proportion of significant SNPs will with the ten QTN shown in Figure 1a.



Methods: The SNPs were ordered by p-value from low to high. From most ten most significant SNPs, the number of QTNs included were counted.

Results and Interpretation: Of the top ten most significant SNPs, there were only 3 QTNs present (Figure 5). That means 7 of the 10 most associated SNPs were false positives.

Figure 5. Manhattan plot with LOD plotted across the genome. Vertical lines note where the top 10 SNPs are located; red lines indicate QTNs.

Question 4: Count number of SNPs with P values smaller than the P value of the seventh significant QTN (20 points).

Hypothesis: Many SNPs will be detected to be more associated with the phenotype than the seventh most significant QTN. This hypothesis is based on the results from Question 3.

Methods: The QTN were ordered by p-value from low to high. The p-value of the seventh most associated QTN was identified, then that threshold was set on all the SNPs and every SNP with a p-value lower than it was counted.

Results and Interpretation: The seventh most associated QTN had a p-value of 0.061. Of the remaining SNPs, 517 of them had a p-value of less than 0.061.

Question 5: Redo (3-4) for 100 replicates. Report the averages and standard deviations (20 points).

Hypothesis: The average number of QTN amongst the top ten most associated SNPs will be near 3, and the average number of SNPs with P-values smaller than the P-value of the seventh most significant QTN will be near 500.

Methods: The *GWASbyCor* function was utilized in this code, but it was looped to run 100 times, each time sampling the number of QTN amongst the top ten most associated SNPs and the number of SNPs with P-values smaller than the P-value of the seventh most significant QTN. The average and standard deviation was taken for each of these values.

Results and Interpretation: The average number of QTN amongst the top ten most associated SNPs was 5.16 with a standard deviation of 1.25. This is not consistent with my hypothesis that the average would be near 3, even if you expand the range using a 95% confidence interval to 5.035 to 5.285. The average number of SNPs with P-values smaller than the P-value of the seventh most significant QTN was 211.59 with a standard deviation of 103.86. This is also not consistent with my hypothesis that the average would be near 500, even expanding the range using a 95% confidence interval to 107.73 to 315.45. These numbers being so different from the first time the GWAS was run could have to do with the structure of the population influencing the number of false positives detected. Using something like a PCA to correct for population structure should reduce the difference in results between each GWAS.

Question 6 (Extra credit): Simulated phenotypes from genotypes so that the phenotypes skewed normal distribution due to genetic effect with a long tail on the right (25 points, report is limited to one extra page).

Hypothesis: By skewing the genetic effect by adding a long tail on the right of the distribution, most genotypes will have a small genetic effect and only a small number of QTNs will be highly-associated in the GWAS.

Methods: The *G2P* function was used to residual and genetic effects, and the phenotype for a trait with a heritability of 0.75. To skew the phenotypic distribution to have a long tail, the genetic effect, alpha, was reduced to 0.2 compared to 1 in Question #1. This causes the next QTN to have dramatically less genetic effect, as it is the square of the previous QTN genetic effect.

Results and Interpretation: Figure 6a shows the density plots for all three effects. The phenotype and genetic effects nearly overlap, and both have a longer tail on the right, as hypothesized. The phenotypic variance is the sum of the residual and genetic effects, as shown in Figure 6a. Only a small number of individuals have genetic effects greater than 0.1, as shown in Figure 6b. The correlation between the phenotypic variance and genetic effects is 0.87, between the phenotypic variance and residual effects is 0.53, and between the genetic and residual effects is 0.04. These variances are like what we saw without skewing the data, in Question #1.

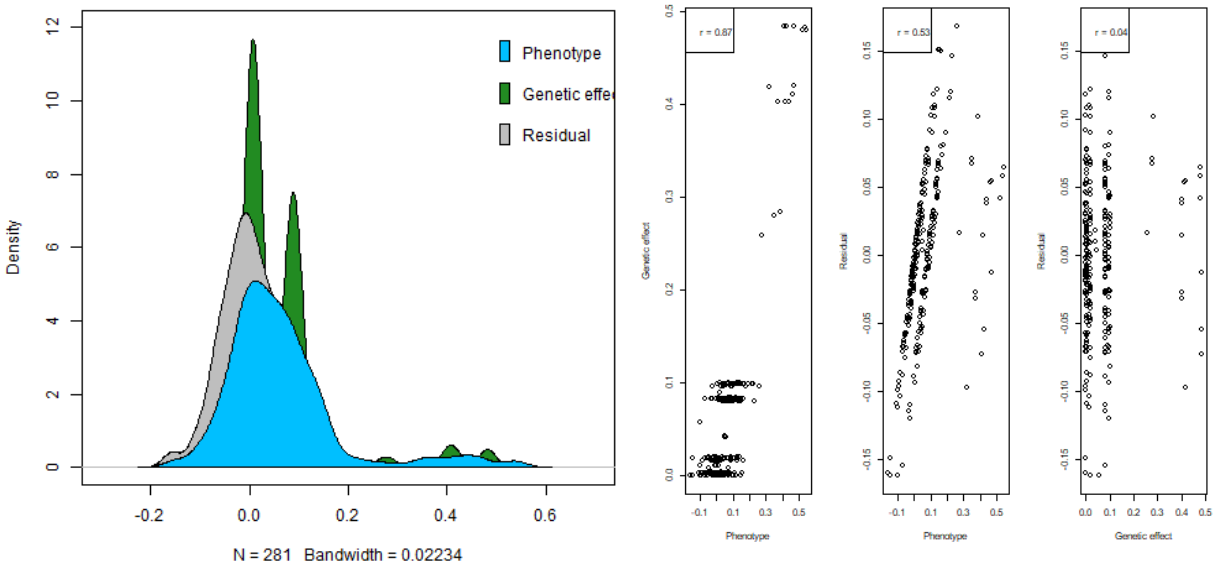


Figure 6. (a) Density plots of the phenotype, genetic effect, and residual and (b) complete correlation comparisons between the phenotype, genetic effect, and residual.