Homework-1

Sanjana Reddy Thigulla
1002142811

1. Executing the given "datasetGenerator.py" script on terminal with the UTA ID: 1002142811 has generated a unique dataset with name 1002142811.csv. It is included in the zip file.

2. - The fisher's exact test is a statistical test performed on the contingency tables and tests whether the odds ratio of the underlying populations are close to 1 or not
   - The odds ratio is defined as the ratio of the odds of event A in the presence of B and the odds of A in the absence of B. it quantifies the strength of the association between two events.
   - Oddsratio and p value have been calculated using fisher exact function from scipy module.
   - Null hypothesis: C-allele or T-allele SNPs don't contribute to person's risk of developing the complex trait. (independent)
   - Alternative hypothesis: C-allele or T-allele SNPs contribute to person's risk of developing the complex trait. (dependent)
   - The fisher exact function takes 2 inputs, the 2×2 matrix/table and alternative. By default alternative is two-sided that is the odds ratio is not one.
   - The total number of significant p values is 299.
   - The fisher exact function gives 2 values, oddsratio and p-value. This p-value is then compared with effective p-value to find out whether the SNP is significant.

3. The assumed p-value is 5×10^{-8} . Bonferroni corrected p-value is given by

$$\begin{aligned} &= \frac{(\text{original } p \text{ value})}{(\text{no of tests performed})} \\ &= \frac{5 \times 10^{-8}}{1000} \\ &= 5 \times 10^{-11} \end{aligned}$$

   Therefore, the Bonferroni corrected p-value is 5×10^{-11}.

   After Bonferroni correction, Total number of new significant p values is 209.

   The *results.csv* file depicts the results of the fisher exact test after comparing with the effective p value and corrected pvalue.

4. Using the negative logarithm of (p value) a pseudo Manhattan plot (Figure 1) has been generated.

   - This plot depicts the SNPs and their p values; all of the significant SNPs over the threshold indicate that the C alleles will have an impact on the complex phenotype.
   - The upper red line represents the standard p-value threshold, and the lower redline represents the Bonferroni-corrected threshold value.

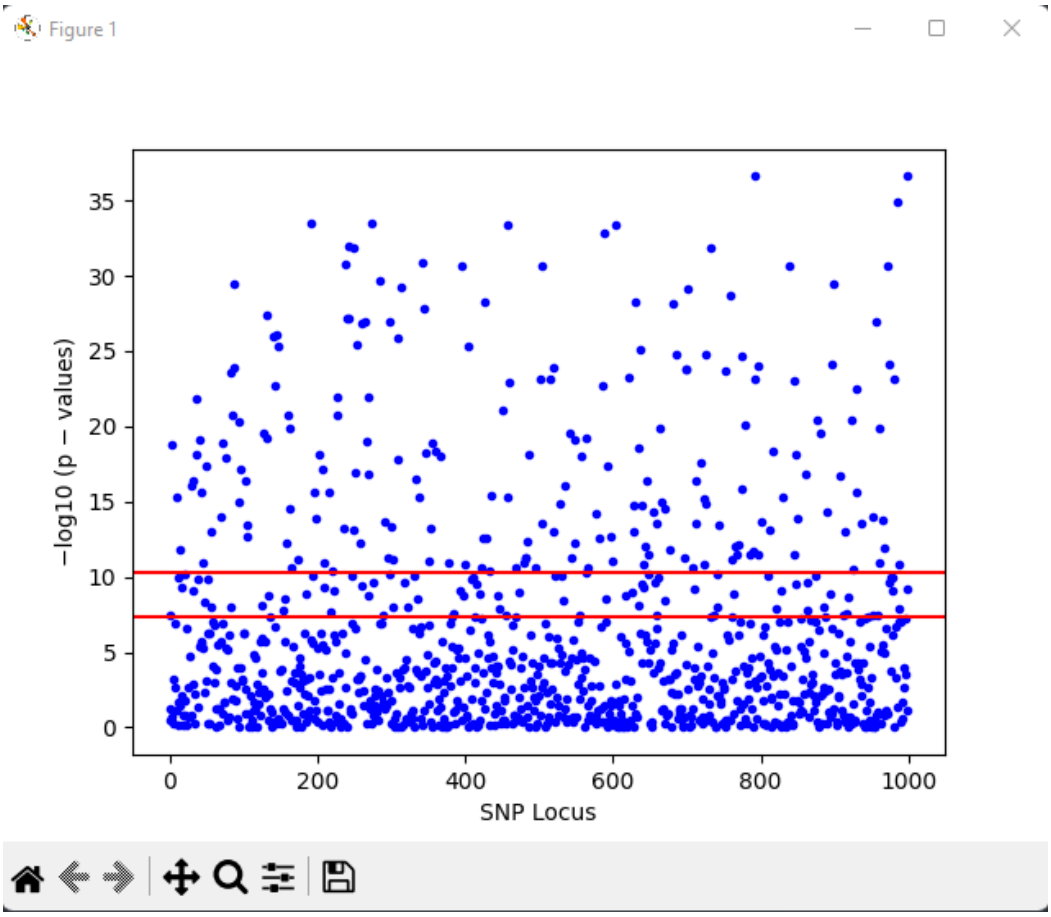

Figure 1: Pseudo Manhattan plot