1. INTRODUCTION

The given dataset has 1000 SNPs from 50 patients in the case cohort and 100 people in the control cohort and 5 columns representing the name of the SNP, number of C-alleles in the case cohort, number of T-alleles in the case cohort, number of C-alleles in the control cohort, and number of T-alleles in the control cohort.

1.1 Generating Your Own Unique Data

I have used the ***python3 datasetGenerator.py --ID 1002104402*** to generate the dataset.

1.2 Fisher's Exact Test

I have taken the null hypothesis as "SNP is not significant". I had found out that the number of significant SNPs by checking whether the p-value of the SNP is less than the effective p-value ($5\times10-8$). I performed the Fisher's exact test on C-allele and T-allele to find out the number of significant SNPs. If it is less, we conclude the SNP is significant. The number of significant SNPs in our dataset were **166**.

The alternative argument for the *fisher_exact* function is chosen as "greater" as it is the odds ratio of allele C which is higher than 1, allele C is one among the causes of the complex of the genetic trait.

1.3 Corrected P-Values

Before this I had taken the first p-values and checked for significance of the SNP. Then it taken the Bonferroni-corrected ($/n$)p-value. Bonferroni-corrected p-value is the original p-value divided by the number of comparison of the SNP's of 1000. Therefore, the corrected p-value is $5 \times 10 -11$. The number of SNPs that are significant under the corrected p-value is **114**.

1.4 Manhattan Plot

The Manhattan Plot prompts the scattering of the p-values of all the SNPs in the dataset. I have also illustrated the threshold of the first p-value and corrected p-values. While seeing that the major p value beneath the threshold value. Even the few p-values are in between the first p-value threshold and the corrected one. The point which are above the threshold are the ones where the null hypothesis is rejected.

2. DIFFICULTY ADJUSTMENT

The Level and Complexity of the assignment is a bit tough as I worked on this assignment for 20+ hours, but it was interesting working on various topics like n no of datasets, fisher exact test, finding p value, manhattan plotting etc.