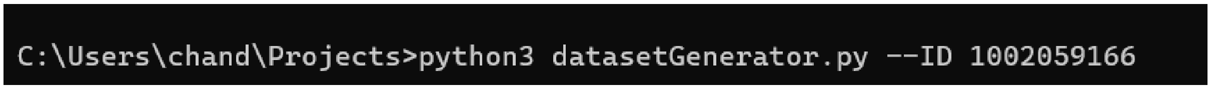
CSE 5370: BIO-INFORMATICS

HOMEWORK - 1

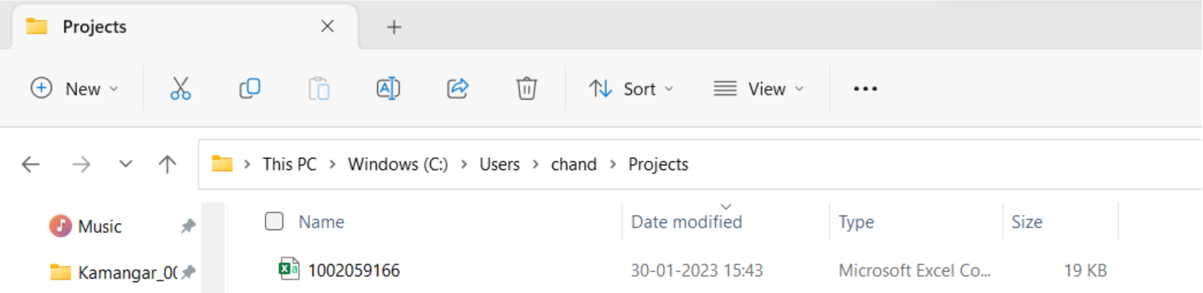
Genome Wide Association Study (GWAS)

Generating Your Own Unique Data

Ran the "datasetGenerator.py" script with my UTA ID and generated the 1002059166.csv file as shown below.

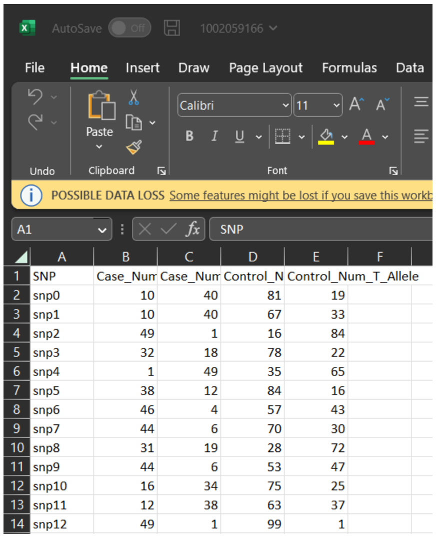


```
C:\Users\chand\Projects>python3 datasetGenerator.py --ID 1002059166
```



Name	Date modified	Type	Size
1002059166	30-01-2023 15:43	Microsoft Excel Co...	19 KB

The Above 1002059166.csv has 1000 rows and 5 columns, which is randomly generated according to the script.



	A	B	C	D	E	F
1	SNP	Case_Num	Case_Num	Control_N	Control_Num_T_Allele	
2	snp0	10	40	81	19	
3	snp1	10	40	67	33	
4	snp2	49	1	16	84	
5	snp3	32	18	78	22	
6	snp4	1	49	35	65	
7	snp5	38	12	84	16	
8	snp6	46	4	57	43	
9	snp7	44	6	70	30	
10	snp8	31	19	28	72	
11	snp9	44	6	53	47	
12	snp10	16	34	75	25	
13	snp11	12	38	63	37	
14	snp12	49	1	99	1	

Fisher’s Exact Test

Imported the required libraries: pandas for data handling, fisher_exact from scipy.stats for calculating p-values, numpy for numerical computation, and matplotlib.pyplot for plotting.

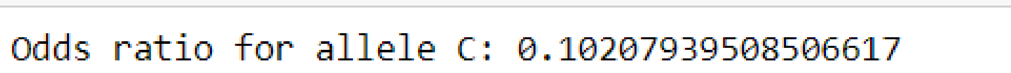
Fisher's Exact Test is a statistical test that determines whether the observed frequencies of a categorical variable in a contingency table differ significantly from expected frequencies based on a null hypothesis. The test is used to determine the association between two categorical variables and to calculate the odds ratio, which is a measure of the strength of association between the variables. The test is commonly used in genetic association studies to assess the association between a genetic variant (SNP) and a disease outcome.

a) The null hypothesis of the fisher_exact function assumes that there is no association between the presence of a SNP and the likelihood of developing a complex trait. In other words, the odds of having the complex trait in individuals with the C-allele is equal to the odds of having the complex trait in individuals with the T-allele. The purpose of the Fisher's exact test is to determine if the data provides evidence to reject the null hypothesis.

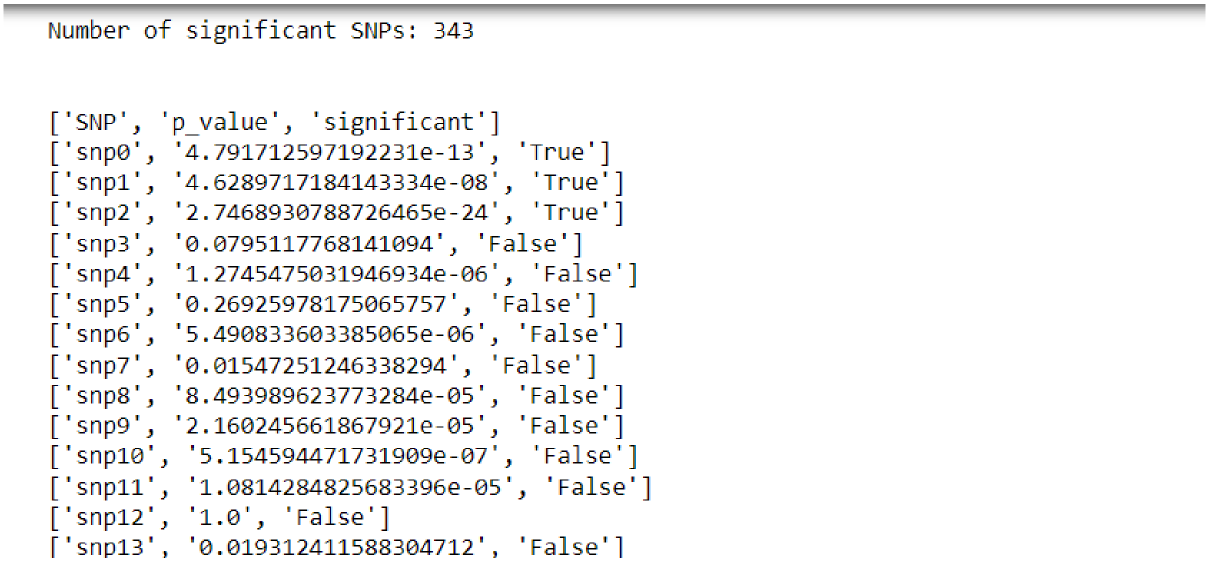
b) The "alternative" argument in the fisher_exact function determines the alternative hypothesis being tested. By default value of "two-sided" is being used. In a two-sided test, the null hypothesis is tested against the alternative that the odds ratio is not equal to 1 (i.e. the proportions of C-allele and T-allele are not equal between the case and control groups) and odds ratio for Allele C is 0.102079 as shown below. So choose two-sided.

c) Printing the SNP name, P-value, Significance in the first 3 columns of the results.csv file. Stored the results in a new dataframe results, which includes columns for the SNP name, the calculated p-value, and a Boolean variable indicating whether the SNP is significant under the original p-value threshold (p-value < 5e-8).

d) And printing the number of significant SNPs as 343(means P-value less than 5e-8).



```
Odds ratio for allele C: 0.10207939508506617
```



```
Number of significant SNPs: 343


['SNP', 'p_value', 'significant']
['snp0', '4.791712597192231e-13', 'True']
['snp1', '4.6289717184143334e-08', 'True']
['snp2', '2.7468930788726465e-24', 'True']
['snp3', '0.0795117768141094', 'False']
['snp4', '1.2745475031946934e-06', 'False']
['snp5', '0.26925978175065757', 'False']
['snp6', '5.490833603385065e-06', 'False']
['snp7', '0.01547251246338294', 'False']
['snp8', '8.493989623773284e-05', 'False']
['snp9', '2.160245661867921e-05', 'False']
['snp10', '5.154594471731909e-07', 'False']
['snp11', '1.0814284825683396e-05', 'False']
['snp12', '1.0', 'False']
['snp13', '0.019312411588304712', 'False']
```

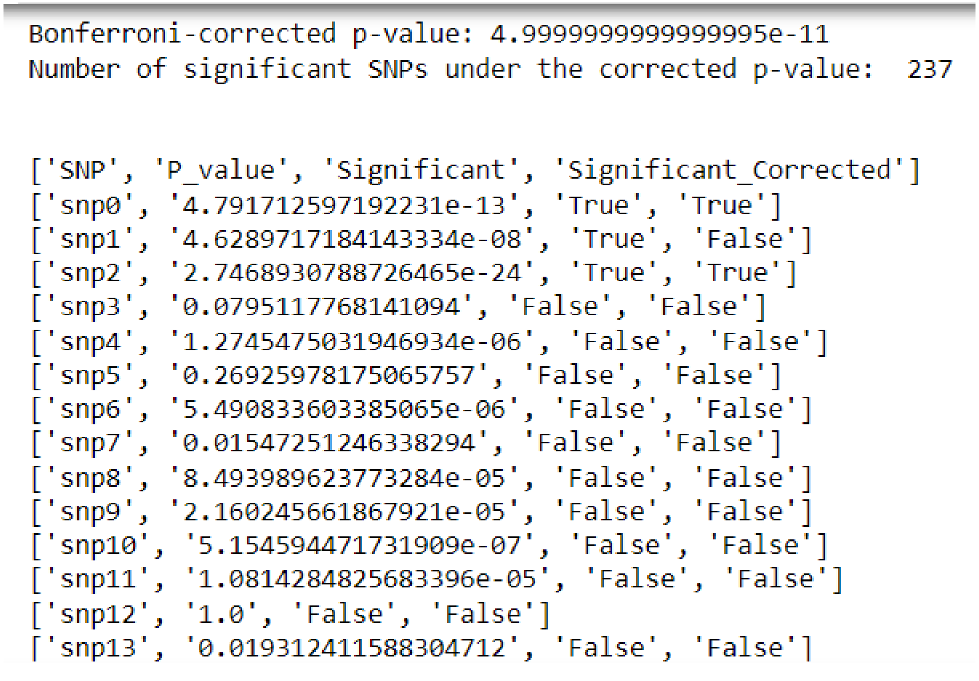
Corrected P-Values

Calculating the Bonferroni-corrected p-value by dividing the original p-value threshold (5e-8) by the number of SNPs.

a) Bonferroni-Corrected P-value is 4.999e-11

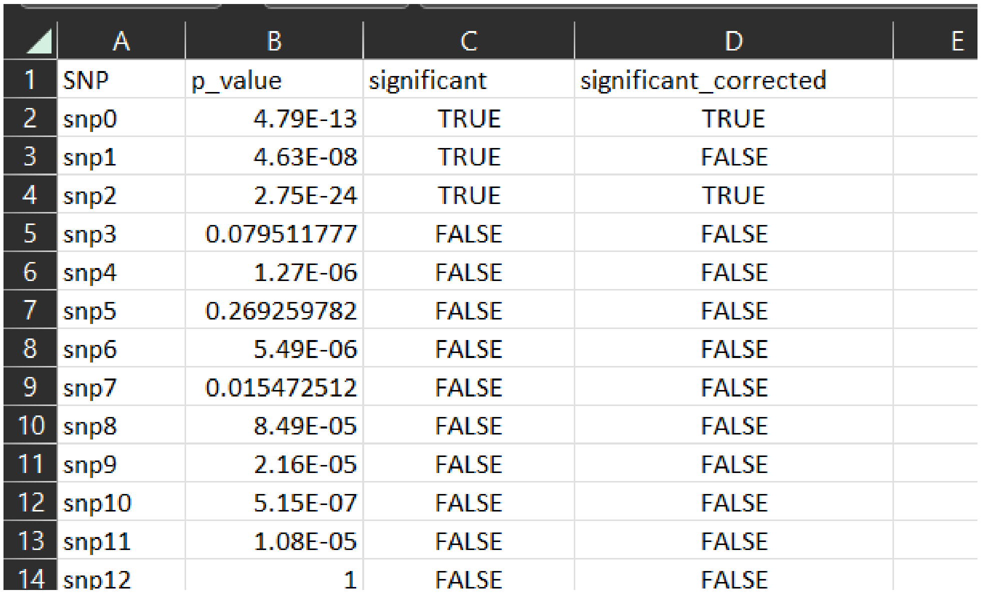
b) Printing the Corrected Significant P-value, Number of significant SNPs under the corrected p-value: 237. The number of significant SNPs (237) under the corrected p-value does not allow us to definitively conclude whether any of the C-allele SNPs contribute to a person's risk of developing the complex trait. Further analysis, such as evaluating the effect size and functional significance of the SNPs, is needed to establish a causal relationship between the SNPs and the complex trait. Additionally, replication of the findings in independent studies is important to validate the results and increase confidence in the conclusions.

c) Printing the Significant_Corrected P-value in the 4rth column of results.csv file.



```
Bonferroni-corrected p-value: 4.9999999999999995e-11
Number of significant SNPs under the corrected p-value:  237


['SNP', 'P_value', 'Significant', 'Significant_Corrected']
['snp0', '4.791712597192231e-13', 'True', 'True']
['snp1', '4.628971718414334e-08', 'True', 'False']
['snp2', '2.7468930788726465e-24', 'True', 'True']
['snp3', '0.0795117768141094', 'False', 'False']
['snp4', '1.2745475031946934e-06', 'False', 'False']
['snp5', '0.26925978175065757', 'False', 'False']
['snp6', '5.490833603385065e-06', 'False', 'False']
['snp7', '0.01547251246338294', 'False', 'False']
['snp8', '8.493989623773284e-05', 'False', 'False']
['snp9', '2.160245661867921e-05', 'False', 'False']
['snp10', '5.154594471731909e-07', 'False', 'False']
['snp11', '1.0814284825683396e-05', 'False', 'False']
['snp12', '1.0', 'False', 'False']
['snp13', '0.019312411588304712', 'False', 'False']
```



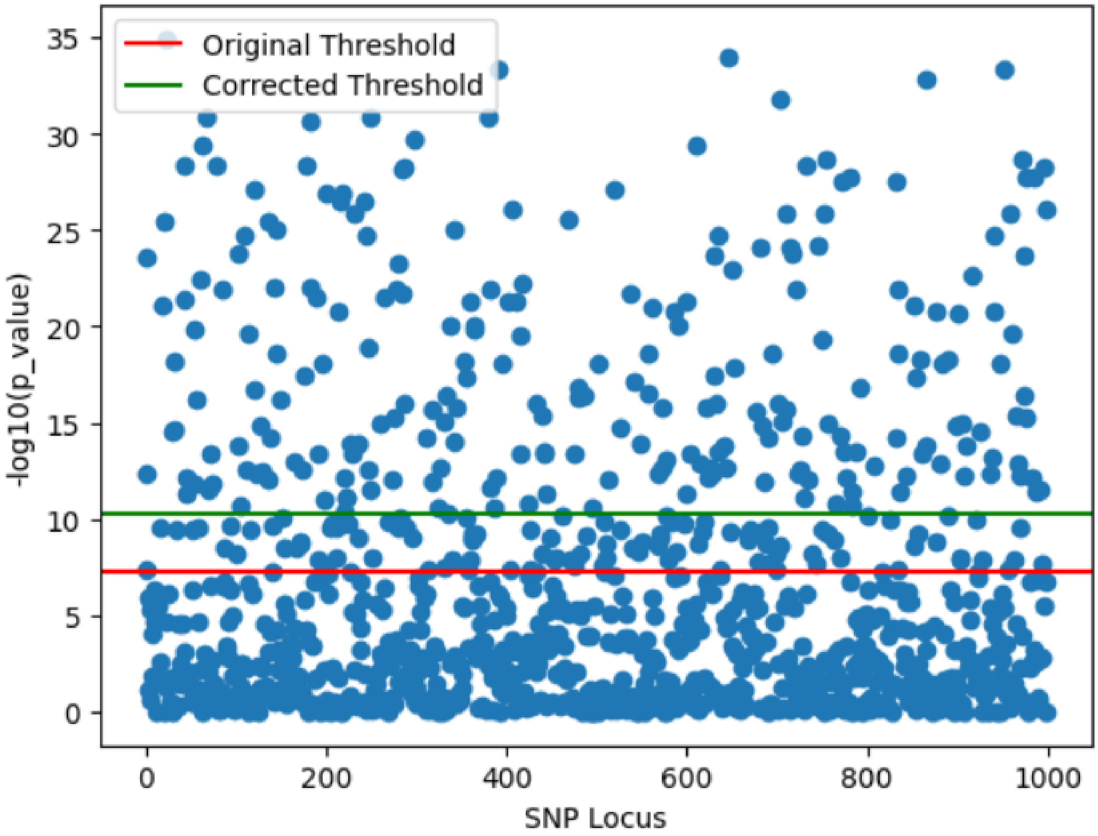
	A	B	C	D	E
1	SNP	p_value	significant	significant_corrected	
2	snp0	4.79E-13	TRUE	TRUE	
3	snp1	4.63E-08	TRUE	FALSE	
4	snp2	2.75E-24	TRUE	TRUE	
5	snp3	0.079511777	FALSE	FALSE	
6	snp4	1.27E-06	FALSE	FALSE	
7	snp5	0.269259782	FALSE	FALSE	
8	snp6	5.49E-06	FALSE	FALSE	
9	snp7	0.015472512	FALSE	FALSE	
10	snp8	8.49E-05	FALSE	FALSE	
11	snp9	2.16E-05	FALSE	FALSE	
12	snp10	5.15E-07	FALSE	FALSE	
13	snp11	1.08E-05	FALSE	FALSE	
14	snp12	1	FALSE	FALSE	

Manhattan Plots

Manhattan plot, which is a graphical representation of the association between a complex trait and SNPs along a chromosome. The x-axis represents the SNP locus and the y-axis represents the negative logarithm of the P-value, which is a measure of the statistical significance of the association between the SNP and the trait.

The scatter plot shows the -log10(P-value) for each SNP, with higher values indicating more significant associations. The red line represents the original P-value threshold of 5e-8, while the green line represents the corrected P-value threshold obtained through the Bonferroni correction.

SNPs with -log10(P-value) values greater than the corrected threshold are considered statistically significant, meaning they are likely to contribute to the risk of developing the complex trait.



Difficulty Adjustment

In total assignment took around 7 hours to complete, initially to understand the concept took some time. But after getting a clear understanding on concept, coding didn't took much time.

Initially bit confused about Bonferroni Corrected P- values.