CSE 5370: Bioinformatics Homework 1

Harshini Kandimalla(1001960046)

1 Genome Wide Association Studies (GWAS)

You are working as a population geneticist for the government of a large country trying to understand associations between a complex genetic trait (phenotype) and genetic variants in a sequencing study conducted on hundreds of volunteer participants.

In this study, there are 50 patients in the case cohort and 100 people in the control cohort. For these participants, 1000 particular SNPs (snp 1, snp 2, ..., snp 1000) are measured and reported (in a real-world study, number of SNPs tested can be several million). These SNPs are either C-alleles or T-alleles. You are required to conclude *whether there is significant evidence whether any of the C-allele SNPs contribute to a person's risk of developing the complex trait* (Note: this question may be challenging to complete prior to the walk through lecture).

1.1 Generating Your Own Unique Data

You are provided with a python script named datasetGenerator.py. This program will take in your UTA student ID as an argument and generates a unique artificial dataset of the mentioned study. To run the code, simply run:

```
>> python3 datasetGenerator.py --ID 1001960046
```

Running the program will create a file named **1001960046.csv** in the same directory this program is located in. This data set has 1000 rows representing each SNP 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.

1.2 Fisher's Exact Test

In this scenario, you can represent the data as contingency tables and the effect sizes as odds ratios (please refer to the walk through lecture and slides). For each SNP, if there is significant evidence that the odds ratio for allele C is higher than 1, you can conclude that allele C is among the causes of the complex genetic trait.

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. Using the scipy's fisher exact function, find the p-value associated with each SNP for your data set. Assuming an effective p-value of 5×10^{-8}, which SNPs can be considered statistically significant regarding the complex genetic trait? Based on the documentation of fisher exact function, you need to explain what the null hypothesis of this test is and what it means. Also you need to choose to explain how you choose the alternative argument in this function. You have to provide a file named results.csv containing the name of the SNPs in the first column, p-values for each SNP in the second column, and whether the SNP is significant as a Boolean variable in the third column. You should also report the number of significant SNPs in your written answer.

what the null hypothesis of this test is and what it means

- Null Hypothesis (H_0) is Allele C and it is among the causes of the complex genetic trait
- Alternate Hypothesis is Null Hypothesis (H_1) as Allele C is not the complex genetic trait and it is a cause only when Allele C is greater than 1.
- Since the odds ratio of the underlying population is less than one we use the alternate argument of fisher's exact. Here the alternate hypothesis is Null Hypothesis H1 which means Allele C is not the cause of complex genetic trait.
- The number of significant SNPs is **170**.

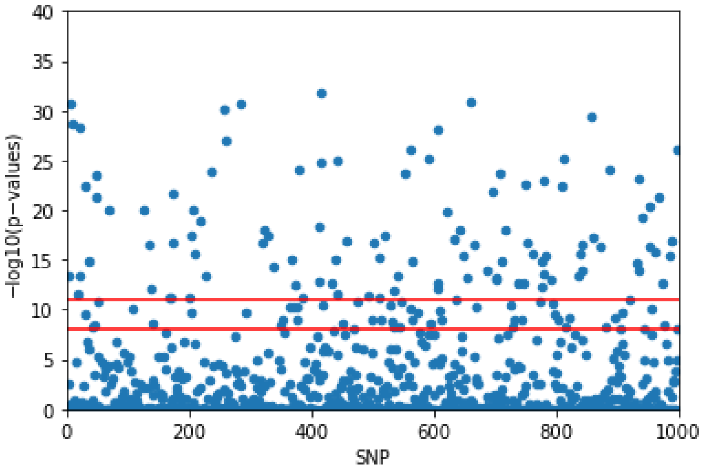
1.3 Corrected P-Values

Assuming each association between a SNP and the phenotype is an independent hypothesis, and we want our effective p-value to be 5×10^{-8}, what is our Bonferroni-corrected p-value? How many SNPs are significant under the corrected p-value? You should also include the SNPs that are significant under the corrected threshold in the fourth column of results.csv as a Boolean variable.

- Bonferroni-corrected p-value is $\frac{\alpha}{n} = \frac{5 \times 10-8}{1000} = 5 \times 10^{-11}$
- The number of corrected p-values is **118**

1.4 Manhattan Plots

Generate a psuedo-Manhattan plot of the $-log_{10}\,(p - values)$ with the original and corrected p-value thresholds illustrated and distinguished (Note that in the example below, these thresholds are only illustrated but not distinguished). This can be done by plotting the thresholds with different colors and adding a legend to distinguish them. Include a paragraph describing what the Manhattan plot shows.



The X axis in the plot represents the SNPs and Y axis represents the associated p values. The blue dots in the plot are the scattered p values for the respective SNPs. The plots between lines represents the threshold.

2 Difficulty Adjustment

Your answers to this section will be used to adjust the difficulty of future assignments in the class.

- How long did this assignment take you to complete?

  It took me around 12 hours to complete.

- If the assignment took you longer than the 10 hours, which parts were overly difficult?

To understand the assignment, it took me 5 hrs. The assignment is challenging yet interesting.