# CSE 5370: Bioinformatics
# Homework 1

Due Monday, February 6th, 2023 at 9:00am CST

In this homework your will be writing code to conduct a mini Genome Wide Association Study (GWAS). The designed time to completion is 5 hours. This assignment is due at 9:00am CST on Monday, February 6th, 2023.

## Logistics, Expectations, & Extra Credit

### Assignment Submission & Specifications

Submissions must be a *single* `.zip` file named in the following format and submitted to Canvas:

`[Student Last Name]_[UTA ID Number]_HW1_CSE5370.zip`

If you last name contains white spaces/hyphens/etc., you have to enter it without those special characters. For example, student Jack Jones-Doe with student ID 1009999999 should have the following file uploaded on Canvas:

`JonesDoe_1009999999_HW1_CSE5370.zip`

This `.zip` file must contain all of the following items ***on its root directory***:

1. Any number of items requested from you by the problem set (e.g. output files, source codes, etc.). If you have used a piece of code to derive these requested items, you should include that too.
2. A *single* text write up page in `.pdf` format. It can be a word document converted to `.pdf`, the output of the LaTeXproject, or any other typesetting solutions you prefer. Non-typeset submissions will not receive credit.
3. If your common sense tells you that the TA/Grader would have difficulties running your code(s), a *single* `readme.txt` file explaining the guideliness to run your code(s) is encouraged to be included in the `.zip` file.

It is your responsibility to ensure that files are not corrupted (It is recommended to make sure that you test your submission to ensure that it can be decompressed) and that your code compiles/runs.

- **IMPORTANT:** Your assignments will be graded automatically. Any submission that does not follow the naming convention and formatting guideline presented above would receive a **zero** for this assignment.
- **IMPORTANT:** Any corrupted files or code that does not compile or run will not be given credit.

## Extra Credit

An extra 2.5% credit will be given to assignments where the text write up page is typeset with LaTeX(must include both the .tex and .pdf file inside your .zip file for credit). Assignments where all code is submitted as a Jupyter Notebook (one .ipynb file or a link to a Google Colab notebook where the link gives permission to view the notebook) will receive an extra 2.5% credit.

## Academic Honesty & Office Hours

Many of the answers on CHEGG and similar sites that appear similar to questions on this assignment have incorrect answers. Students are encouraged to refer back to lecture recordings/slides and come to office hours before the assignment is due if they are struggling.

## Group Work

This is an individual assignment and group work is not permitted. Some coding problems require an individual submission based on a individualized data set generated randomly from your UTA ID. Every person will have their assignment graded individually.

## StackOverflow.com & Similar Sites

Use of stackoverflow.com and other sites is explicitly allowed (industry researchers and academic labs use these sites frequently). However, for this course you must include a comment in your code with the link to the page you referenced whenever these sites influence your own code writing. For example, when writing this homework assignment I forgot how to insert code into LaTeXdocuments and recalled how to after visiting stackoverflow.com. If I were submitting this as an assignment, I would want to include a comment like the below example in my code submission:

```
% When writing this homework assignment, I did not recall how to
% insert code in a nice looking way into LaTeX documents,
% so I referred to this page on stackoverflow for help:
% https://stackoverflow.com/questions/3175105

\usepackage{minted}
```

```
\begin{minted}[mathescape, linenos]{python}
Code To Insert in \LaTeX...
```

It is academic dishonesty to copy code from sites like stackoverflow without attribution like this, but is fine as long as you include attribution.

### Office Hours

Please start working on assignments in advance. Office hours are offered Monday, Tuesday, Wednesday, and Thursday by the course staff (Dr. Luber and TAs) for at least an hour each day. The Tuesday before each homework is due, lecture will be a flipped classroom where students will have the opportunity to work on code and ask the course staff questions. Office hours are not offered on Friday and the course staff will not respond to emails/canvas messages/TEAMS/other communication about homework assignments on Friday evenings, Saturdays, Sundays, and Monday mornings before 9:01am unless the email is a follow-up to an office hours visit or previous question asked during the previous flipped classroom session (in these cases the course staff will make their best effort to respond).

### Late Submission Policy

All homework assignments are graded out of 100 points. Assignments submitted late will be penalized, at a rate of 4 penalty points per hour. The submission time will be the time shown on Canvas. Any assignment submitted more than 25 hours late will receive no credit for the assignment.

Exceptions to late submission penalties will only be made for emergencies documented in writing, in strict adherence to UTA policy. For all such exception requests, the student must demonstrate that he or she made all efforts to notify the instructor as early as possible.

Computer crashes, network crashes, software or hardware failure, temporary Canvas failure, email failure, will NOT be accepted as justification for late submissions. If you want to minimize chances of a late submission, aim to submit early.

## 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 [10 points]

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 [your UTA ID]
```

Running the program will create a file named `[your UTA ID].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.

Generate your unique data set and include it in your submission (please note that changing the data set generator script or failing to generate your own unique data set will result in a grade of 0 for the assignment). The automated grading script will run "diff" in bash against all pairwise combinations of submitted data sets that should be unique based on your UTA ID; data sets from different submissions that "diff" flags as the same will automatically result in a 0 for both submissions. [10 points]

## 1.2 Fisher's Exact Test [20 points]

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 $5 \times 10^{-8}$, 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 (See Fig. 1). You should also report the number of significant SNPs in your written answer. [20 points]

| | | |
|---|---|---|
| snp_1 | 6.30E-13 | TRUE |
| snp_2 | 9.81E-08 | FALSE |
| snp_3 | 4.36E-08 | TRUE |
| snp_4 | 1.80E-10 | TRUE |
| ... | ... | ... |
| snp_999 | 1.57E-19 | TRUE |
| snp_1000 | 6.38E-07 | FALSE |

Figure 1: An example of how `results.csv` should be formatted by the end of Sec. 1.2.

### 1.3 Corrected P-Values [30 points]

Assuming each association between a SNP and the phenotype is an independent hypothesis, and we want our effective p-value to be $5 \times 10^{-8}$, 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. [30 points]

### 1.4 Manhattan Plots [40 points]

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. [40 points]

An example output for a random student ID is provided on the next page.

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

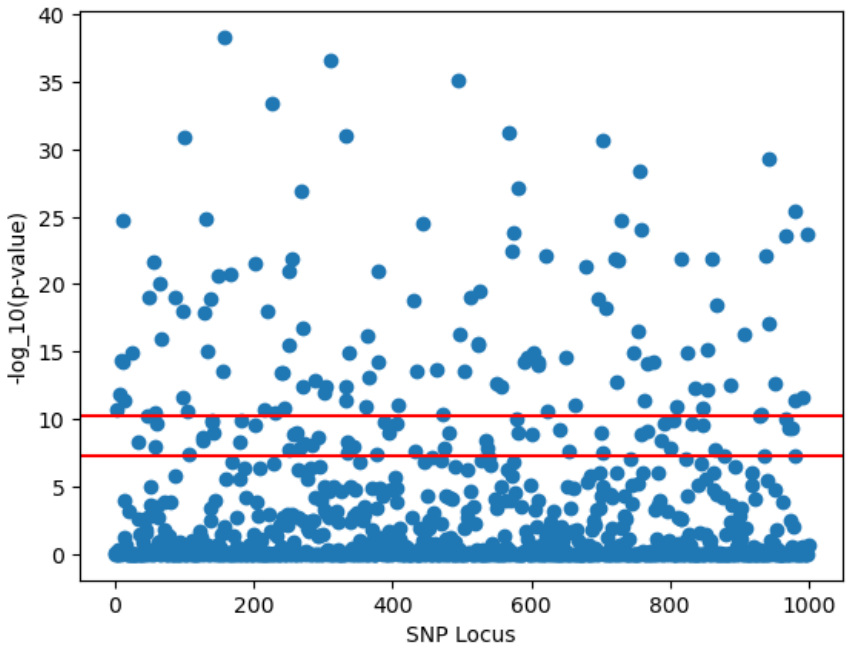


Figure 2: Sample Manhattan plot for the study.