QBIO 490: Directed Research - Multi-omic Analysis

Spring 2023 Midterm

Project Due: Friday, March 10th (11:59 pm). Submit your GitHub link to Blackboard, with all your code (as R script files), review questions answers (as PDF), and report (as PDF) in a folder called midterm_project_lastname. Also share your report with Wade and Kayla + Nicole + TAs on Google Drive. Extension requests due Wednesday, March 8th 11:59 pm. Email Wade and Kayla and cc Nicole + TAs.

Purpose: This midterm project is meant to recap the analyses we've performed so far in R. It's also intended to introduce you to scientific writing and communication. For this project, please do your own work and submit your own written report, but you are more than encouraged to discuss ideas and debug code in groups! Note there are three parts to this assignment.

Part 1: Review Questions

General Concepts

1. What is TCGA and why is it important?

TCGA stands for the Cancer Genome Atlas. It is a cancer genomics program hosted by the National Human Genome Research Institute and the National Cancer Institute. This program makes genomic, epigenomic, transcriptomic, and proteomic data on different cancer types available to the public, so any researcher can access these data.

2. What are some strengths and weaknesses of TCGA?

TCGA has more comprehensive data and larger sample size that is not usually available in individual research. Also, the data are standardized.

A weakness of TCGA is that the data is not collected from patients from all areas, so the conclusion may not be representative for all populations of cancer patients since geographical factors can be influential.

Coding Skills

1. What commands are used to save a file to your GitHub repository?

```
git status # not required
git add [name of file]
git commit -m [comment]  # comment is optional
git push
```

2. What command(s) must be run in order to use a package in R?

```
Install.packages()
library()
```

3. What command(s) must be run in order to use a Bioconductor package in R?

```
BiocManager::install() # only the first time
library()
```

4. What is boolean indexing? What are some applications of it?

Boolean indexing involves creating an array of booleans and applying it on a dataframe. It's most commonly used in filtering data based on the values of each entry (ie. boolean masking).

5. Draw a mock up (just a few rows and columns) of a sample dataframe. Show an example of the following and explain what each line of code does.

a. an ifelse() statement

b. boolean indexing

6	8	5	7
10	3	8	2
0	4	6	4

# Assume the matrix is called df

a.
```
greater_mask  <- ifelse(df[2, ] > 5, T, F) # creates a mask for any value in the second row
                                             of df that's greater than 5
                                           # result would be TFTF
```

b.
```
row_mask <- FTF # create a mask for rows of df
col_mask <- TFFT # create a mask for cols of df
new_df <- df[row_mask, col_mask] # apply the masks; result would be 10, 2
```