General Concepts

1. What is TCGA and why is it important?
The Cancer Genome Atlas (TCGA), a joint effort between the National Cancer Institute and the National Human Genome Research Institute, is a program that characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. This allowed researchers to access breakthrough amounts of data and conduct large analyses to understand cancer better. TCGA offers a more comprehensive database and allowed for improved cancer diagnosis (such as through biomoarkers) and treatment.

2. What are some strengths and weaknesses of TCGA?
TCGA has a large amount of publicly available multi-omic data, which is a large strength. However, some clinical variables are missing, such as data about if a patient died, when they died, and so on. This may be because a lot of the data isn't collected longitudinally, as, due to HIPPA and other privacy law, it's difficult to follow a patient through their lives and cancer journey.

Coding Skills
1. What commands are used to save a file to your GitHub repository?
add file to GitHub: git add <file>
commit: git commit -m "commit message"
push to branch: git push
2. What command(s) must be run in order to use a package in R?
install.packages("name of package")
library("name of package")
3. What command(s) must be run in order to use a Bioconductor package in R?
install.packages("BiocManager")
library()
4. What is boolean indexing? What are some applications of it?
Boolean indexing selects items from a data frame based on a statement that returns T or F. This allows us to use Boolean masks, so we can subset into dataframes or vectors to remove data that doesn't follow the logical statement and keep data that does.
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.
dataframe:

	col 1	col_2	col 3	col 4
row 1	2	5	92	3
row 2	1	2	3	4

a. an ifelse() statement
ifelse(dataframe[2,2] == “g”, print(“hello”), print(“dog”))
if ran: prints “dog” because dataframe at row 2, column 2 does not equal “g”
thus, does second statement, which is print(“dog”)
b. boolean indexing
ifelse(dataframe$col_2 > 4, T, F)
this goes through the column “col_2” in the dataframe, checking if the values are greater than 4. if they are, the vector value is T. otherwise, the vector value is F.